IBM and Cisco
Together for a World Class Data Center

- Read about how IBM and Cisco are collaborating in the data center
- Learn about the technologies and products that are leveraged
- See practical use case examples

Pall Beck
Peter Clemens
Santiago Freitas
Jeff Gatz
Michele Girola
Jason Gmitter
Holger Mueller
Ray O'Hanlon
Veerendra Para
Joe Robinson
Andy Sholomon
Jason Walker
Jon Tate

Redbooks

ibm.com/redbooks

International Technical Support Organization

IBM and Cisco: Together for a World Class Data Center

July 2013

SG24-8105-00

Note: Before using this information and the product it supports, read the information in “Notices” on page xiii.

First Edition (July 2013)

This edition applies to the products described herein and that were available at the time of writing in October 2012.

Contents

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at http://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

Active Cloud Engine™
Active Memory™
AIX®
BladeCenter®
BNT®
Cognos®
DataPower®
DS8000®
Easy Tier®
EnergyScale™
Enterprise Storage Server®
FICON®
GDPS®
Global Technology Services®
GPFS™
Green Sigma™
HACMP™
IBM Flex System™
IBM Flex System Manager™
IBM SmartCloud™
IBM Systems Director Active Energy Manager™
IBM Watson™
IBM®
iDataPlex®
Linear Tape File System™
Lotus Notes®
Lotus®
Maximo®
MVS™
Netcool®
NetView®
Notes®
OS/390®
POWER Hypervisor™
Power Systems™
POWER6®
POWER7+™
POWER7®
PowerLinux™
PowerPC®
PowerVM®
POWER®
PR/SM™
PureApplication™
PureData™
PureFlex™
PureSystems™
RACF®
RackSwitch™
Real-time Compression™
Real-time Compression Appliance™
Redbooks®
Redpapers™
Redbooks (logo) ®
RS/6000®
SiteProtector™
Smarter Analytics™
Storwize®
System i®
System p®
System Storage®
System x®
System z9®
System z®
System/390®
Tivoli®
VMready®
WebSphere®
X-Architecture®
X-Force®
XIV®
xSeries®
z/OS®
z/VM®
z9®
zEnterprise®

The following terms are trademarks of other companies:

Algo, and Ai logo are trademarks or registered trademarks of Algorithmics, an IBM Company.

Intel Xeon, Intel, Intel logo, Intel Inside logo, and Intel Centrino logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

ITIL is a registered trademark, and a registered community trademark of The Minister for the Cabinet Office, and is registered in the U.S. Patent and Trademark Office.

IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency which is now part of the Office of Government Commerce.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Linear Tape-Open, LTO, Ultrium, the LTO Logo and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.

Microsoft, Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Java, and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Preface

This IBM® Redbooks® publication is an IBM and Cisco collaboration that articulates how IBM and Cisco can bring the benefits of their respective companies to the modern data center.

It documents the architectures, solutions, and benefits that can be achieved by implementing a data center based on IBM server, storage, and integrated systems, with the broader Cisco network.

We describe how to design a state-of-the art data center and networking infrastructure combining Cisco and IBM solutions. The objective is to provide a reference guide for customers looking to build an infrastructure that is optimized for virtualization, is highly available, is interoperable, and is efficient in terms of power and space consumption. It will explain the technologies used to build the infrastructure, provide use cases, and give guidance on deployments.

Authors

This book was produced by a team of specialists from around the world working on behalf of the International Technical Support Organization and were based at the Cisco campus in Raleigh, North Carolina, to whom thanks must go to for their support.

Pall Beck is a SAN Technical Lead in IBM Nordic. He has 15 years of experience working with storage, both for dedicated clients as well as very large shared environments. Those environments include clients from the medical and financial sector including some of the largest shared SAN environments in Europe. He is a member of the SAN and SVC best practice community in the Nordics and in EMEA. In his current job role, he is the SCE+ Storage Deployment Lead, were he is responsible for SCE+ storage deployments world wide. Pall is also a member of a team assisting in critical situations and doing root cause analysis. He is co-author of the books, *Implementing SVC 5.1, SVC Advanced Copy Services 4.2,* and *Introduction to SAN and System Networking*. Pall has a diploma as an Electronic Technician from Odense Tekniske Skole in Denmark and IR in Reykjavik, Iceland and he is an IBM Certified IT Specialist.

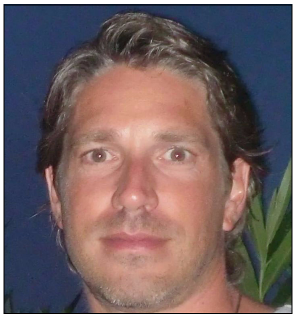
Peter Clemens works as an IT Architect for IBM Business Services GmbH in IBM Germany in the internal IBM account. Before joining the internal account team in 2007, he worked as a networking instructor, consultant, project manager, and Cisco specialist in numerous client engagements. Peter has 16 years of experience in project management and all areas of networking.

Santiago Freitas is a Consulting Systems Engineer for Cisco's IBM Global Team. He is based in London, United Kingdom and supports IBM globally. He has over 10 years of experience with enterprise networks and data centers. He is specialized in architecting and designing complex data center LAN and SAN solutions, as well as data center interconnection technologies. His is recognized within Cisco and IBM as an expert on next generation data centers and cloud computing. Santiago holds two Cisco CCIE certifications, Routing & Switching and Service Provider, and is a regular speaker at Cisco Live! events around the world.

Jeff Gatz is a Systems Engineer with Cisco currently dedicated to IBM worldwide. He has spent the last 16 years specializing in data center supporting servers, network, and storage. Over the years he has held many certifications from Microsoft, McData, and Cisco. Over the past two years he has become a subject matter expert on Cisco's DCNM management solution for both LAN and SAN. Jeff is a frequent speaker at IBM conferences on the subject of data centers.

Michele Girola, at the time of writing, was a part of the ITS Italy Network Delivery team. His areas of expertise include network integration services, data center networking, and wireless networking solutions. Since 2010, Michele was the Global Service Product Manager for GTS Network Integration Services. Prior to joining IBM, Michele worked for NBC News and Motorola. He holds an M.S. degree "Laurea" in Telecommunications Engineering from the Politecnico di Torino, and an M.B.A. from the MIP - Politecnico di Milano. Michele has since left IBM.

Jason Gmitter is a Data Center Technical Solutions Architect for Cisco. Jason has 14 years of experience in architecting, designing, and deploying IP networks, storage area networks, and networking services (WAN optimization, firewalls, SLB). Jason is also a 10 Year CCIE (R&S).



Holger Mueller, at the time of writing, was a Data Center Network Architect for IBM SO Delivery and worked for the Integrated Communications & Networking Speciality Services Area with an emphasis on data center network virtualization and cloud networking. Before joining the IC&N SSA organization, he was a Global Data Center Network Architect for the IBM Internal Account. Holger has 14 years of experience in the networking and is a Cisco Certified Network Professional. He is also a voting member of the global Network Service Architecture and Control Board (NS-ACB). Holger has since left IBM.

Ray O'Hanlon is a Solutions Architect at Cisco responsible for cloud and data center solution design. Based in the UK, Ray has over 30 years of experience in the networking industry and during his 15 years at Cisco has specialized in DC architecture, storage networks, HPC, automation tools, compute platforms, and applications.

Veerendra Para is a Senior Advisory IT Specialist at the IBM India Software Lab (ISL) in Bangalore, India. He has 12 years of experience in the IT Industry and has been directly involved with data center infrastructures for last four years. His current job role is to transform ISL traditional server rooms into energy efficient smarter data centers. Introducing innovative green solutions, strategies, architectures, and timely executing assessment of relevant efficiency elements with reference to industry standard models. Veerendra also shares the success stories with IBM clients via briefings, demos, and education sessions. In past, he has worked in the systems management field, mainly on IBM System i, IBM POWER Systems, and IBM AIX. Prior to IBM Software Group, Veerendra worked for IBM Global Technology Services. He also has technical disclosures that have been published.

Joe Robinson is an IBM Distinguished Engineer for IBM Global Technology Services delivery organization, a group that delivers services for IBM Strategic Outsourcing customers. He is the Chief Architect for Enterprise Network Solutions and responsible for the design, development, and operations of network solutions. Joseph has 29 years of experience at IBM and had many leading technical roles from Development Engineering, Midrange Systems Engineer, Advanced Technical Support of sales, Delivery Manager, and Lead Network Architect for North America.

Andy Sholomon CCIE No. 15179, works as a Solutions Architect in Cisco's Advanced Services team. He routinely designs and tests networks for many of Cisco's largest customers. In his nine years at Cisco, Andy has been involved in both planning and deploying some of the largest data centers in the United States. Before joining Cisco, Andy worked as a Network Engineer in the global financial industry. Besides the CCIE, Andy holds multiple other industry certifications.

Jason Walker is a Solutions Architect for Cisco's Global Enterprise Account organization. Jason has 16 years of experience at Cisco, with 21 years of total experience in Information Technology. Although his experience spans many facets of IT, his focus in recent years has been primarily on data center and cloud architectures. Among his technology expertise areas are data center switching fabrics, SAN, unified I/O, and data center infrastructure virtualization. Jason holds a Cisco CCIE Certification and is a regular speaker at Cisco Live! Events.

Jon Tate is a Project Manager for IBM System Storage® SAN Solutions at the International Technical Support Organization, San Jose Center. Before joining the ITSO in 1999, he worked in the IBM Technical Support Center, providing Level 2 support for IBM storage products. Jon has 27 years of experience in storage software and management, services, and support, and is both an IBM Certified IT Specialist and an IBM SAN Certified Specialist. He is also the UK Chairman of the Storage Networking Industry Association.

Thanks to the following people for their contributions to this project:

Kevin Barnes
Cheryl Gera
Sangam Racherla
Lorinda Schwarz
IBM Redbooks, International Technical Support Organization

Cathy Lemeshewsky
IBM System Networking, San Jose

Christian Allan Schmidt
IBM Denmark

Mikio Itaba
IBM Japan

Vikul Banta
George C Mihalachioaie
Wayne Riley
IBM US

Special thanks also go to the Cisco sponsors who have supported this project since its conception: John McKibben, Don Hansen, Ted Odgers, Joe Aulizia, Will Ochandarena, Yousuf Khan, and Tuqiang Cao.

Thanks also go to the Cisco leadership team that supported and encouraged their employees to participate on this project: Nick Robinson, Chris Larkin, Charles Zillmann, and Tony Volle.

Now you can become a published author, too!

Here’s an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an ITSO residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

We want our books to be as helpful as possible. Send us your comments about this book or other IBM Redbooks publications in one of the following ways:

- Use the online **Contact us** review Redbooks form found at:

  ibm.com/redbooks

- Send your comments in an email to:

  redbooks@us.ibm.com

- Mail your comments to:

  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400

Stay connected to IBM Redbooks

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks

- Follow us on Twitter:

  http://twitter.com/ibmredbooks

- Look for us on LinkedIn:

  http://www.linkedin.com/groups?home=&gid=2130806

- Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm

- Stay current on recent Redbooks publications with RSS Feeds:

  http://www.redbooks.ibm.com/rss.html

1

The modern data center

Most data center infrastructures were not built to support the explosive growth in computing capacity and information that we see today. Many data centers have become highly distributed and somewhat fragmented. As a result, they are limited in their ability to change quickly and support the integration of new types of technologies or to easily scale to power the business as needed. A more integrated IT approach is needed. IT service delivery needs to be changed to help move beyond today's operational challenges to a data center model that is more efficient, service-oriented, and responsive to business needs; a model with improved levels of economies, rapid service delivery, and one that can provide tighter alignment with business goals.

Globalization and knowledge-based economies are forcing companies to embrace new business models and emerging technologies to stay competitive. The global integration is changing the corporate model and the nature of work itself. The reality of global integration can be seen, for example:

- Tight credit markets and limited access to capital
- Economic difficulties and uncertainty about the future
- Energy shortfalls and erratic commodity prices
- Information explosion and risk and opportunity growth
- Emerging economies
- New client demands and business models

In this chapter, we examine the key technology and business trends leading to an approach to implement a modern data center.

1.1 Data center drivers and trends

There is a monumental shift in market dynamics and client demands, which directly impacts the data center landscape of the future. The modern data center environment needs to adapt to be more flexible and cost-effective to meet the unpredictable needs of business and IT. Dynamic market pressures require data centers to assimilate to business and technology growth. Modularity will occur at multiple levels and the facilities and the IT integration will grow in importance.

The right data center strategy is imperative to meet the future business requirements with the optimal portfolio of data center assets. The following drivers are pervasive and directly impact the evolution of the data center landscape into a new “modern data center” era:

- Reduce operational costs
- Reduce operational complexity and improve adoption of new virtualization technologies
- Meet business continuity requirements through reduction of outages and data loss
- Design for flexibility to meet dynamic business requirements
- Reduce time to deploy new technology

The next sections will shed more light on the key trends impacting IT and the data center.

1.1.1 Key operational challenges

The delivery of business operations impacts IT in terms of application services and infrastructure. Business drivers also impose key requirements for data center networking:

- Support cost-saving technologies such as consolidation and virtualization with network virtualization
- Provide for rapid deployment of networking services
- Support mobile and pervasive access to corporate resources
- Align network security with IT security based on enterprise policies and legal issues
- Develop enterprise-grade network design to meet energy resource requirements
- Deliver a highly available and resilient networking infrastructure
- Provide scalability to support the need of applications to access services on demand
- Align network management with business-driven IT Service Management in terms of processes, organization, Service Level Agreements, and tools

When equipped with a highly efficient, shared, and dynamic infrastructure, along with the tools needed to free up resources from traditional operational demands, IT can more efficiently respond to new business needs. As a result, organizations can focus on innovation and on aligning resources to broader strategic priorities. Decisions can be based on real-time information. This new environment creates an infrastructure that provides automated, process-driven service delivery and is economical, integrated, agile, and responsive.

A holistic approach is required to transform the IT landscape. A clear vision and strategy is essential for the future of enterprise computing based on leading practices and technologies. This approach enables businesses to better manage costs, improve operational performance and resiliency, and quickly respond to business needs. IT professionals have continued to express concerns about the magnitude of the operational issues they face. These operational issues are described below.

Costs and service delivery

The daily expense of managing systems and networks is increasing, along with the cost and availability of skilled labor. Meanwhile, there is an explosion in the volume of data and information that must be managed, stored and shared. These pressing issues result in growing difficulty for IT departments to deploy new applications and services. Here are some facts to consider:

- Server management and administration costs grew four-fold between 1996 and 2012,
- Data volumes and network bandwidth consumed are doubling every 18 months, with devices accessing data over networks doubling every 2.5 years.
- 37% of data is expired or inactive.
- There are 32.6 million servers worldwide:
  - 85% computer capacity is idle.
  - 15% of servers run 24/7 without being actively used on a daily basis.
- 1.2 zetabytes (1.2 trillion gigabytes) exist in the “digital universe”
  - 50% YTY growth
  - 25% of data is unique; 75% a copy
- Internet connected devices are growing 42% per year.
- Between 2000 and 2010:
  - Servers grew 6x.
  - Storage grew 69x.
- Virtual machines are growing 42% a year.

The growth listed above must be considered in relation to the fact that the IT budgets are expected to grow less than 1% a year.

Fore more information, use the following link:

http://www-05.ibm.com/si/ibmforum/presentations/2011/1day_ModrejsaInfrastruktura_1735_1815.pdf

Energy efficiency

The larger a company grows, the greater its need for power and cooling. But with power at a premium (and in some areas, capped), organizations are forced to become more energy efficient.

These trends are forcing technology organizations to control costs while developing a flexible foundation from which to scale. In fact, data centers have doubled their energy use in the past five years and the increase in data center energy costs is projected with 18%.[1]

[1] IBM Smarter Computing, Eric Schnatterly: http://www-05.ibm.com/si/ibmforum/presentations/2011/1day_ModrejsaInfrastruktura_1735_1815.pdf

Business resiliency and security

Global expansion has increased the need for tighter security measures. Users require real-time access to confidential, critical data. Enterprise risk management is now being integrated into corporate ratings. At the same time, companies are demanding that users have instantaneous access to this information, putting extra, and often conflicting, pressure on the enterprise to be available, secure, and resilient.

Consider these facts:

- The average US legal discovery request can cost organizations from $150,000 to $250,000[2].
- Downtime costs can amount to up to 16% of revenue in some industries.
- The year 2012 shows an upward trend in overall vulnerabilities with a possibility of an all-time high by year end[3].

Accelerated pace of business and technology innovations

Nearly all analysts have the same business and technology innovations listed in their outlook for the next decade. The following list shows the pervasive trends:

- Cloud computing
- Virtualization
- Big Data
- Analytics
- Mobility, Mobile computing, and Social Communication
- Consumerization of IT
- Human/Computer Interaction
- Business Intelligence
- Security

Some of the trends and innovations listed above are more mature in their adoption and development than others, however, these are the key trends that the IT leaders and CIOs will face in the next years.

A major shift has taken place in the way people connect, not only to each other but also to information, services, and products. The actions and movements of people, processes, and objects with embedded technology are creating vast amounts of data, which consumers use to make more informed decisions and drive action.

[2] IBM's Vision for the new enterprise data center:
ftp://public.dhe.ibm.com/software/uk/itsolutions/optimiseit/energy-efficiency-solutions/1-new-enterprise-data-centre-ibm-s-vision-nedc.pdf

[3] IBM X-Force® update, September 2012:
http://www.smartplanet.com/blog/bulletin/-8216record-year-for-corporate-security-breaches-looms-ibm-x-force-update/728

Figure 1-1 shows how the third generation of computing platform, the third phase of the Internet, and the explosion of information are colliding to form a perfect storm of disruption and transformation.

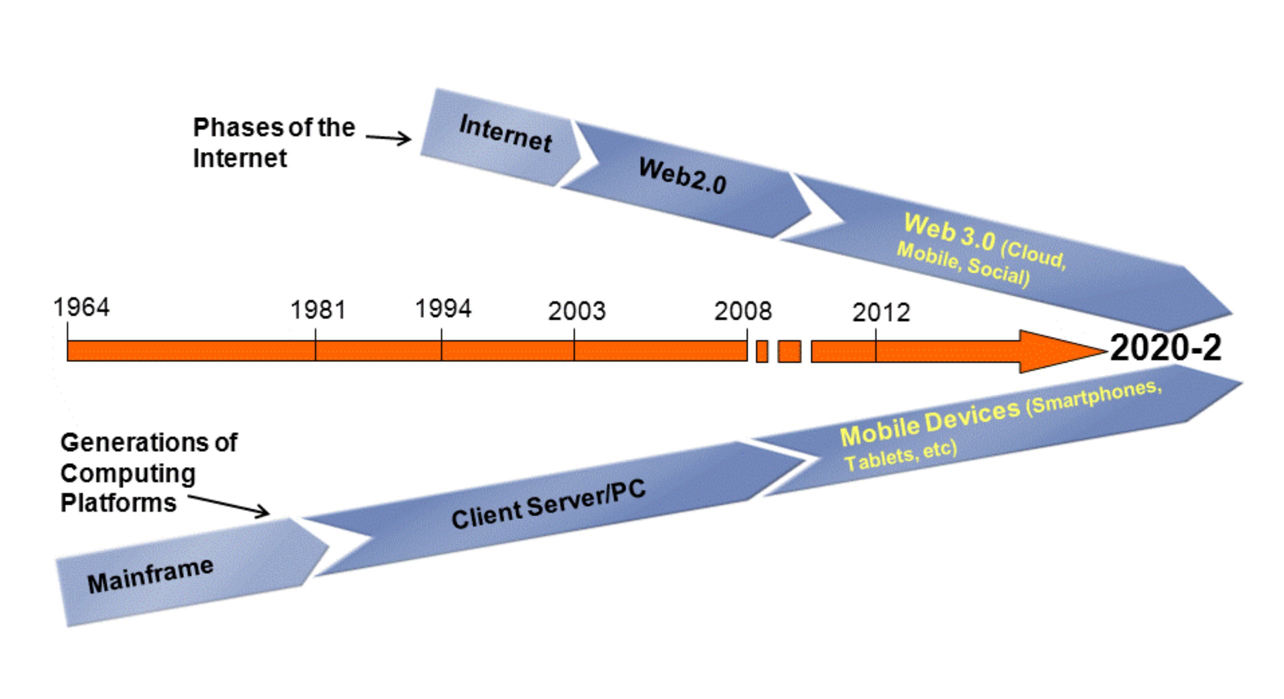

Phases of the Internet
Internet
Web2.0
Web 3.0 (Cloud, Mobile, Social)
1964
1981
1994
2003
2008
2012
2020-2
Generations of Computing Platforms
Mainframe
Client Server/PC
Mobile Devices (Smartphones, Tablets, etc)

Figure 1-1 Trigger for Transformation[4]

The dashed line in the time-line around 2008-11 is meant to reflect the economic disruption that has been happening worldwide. Many times during periods of disruption, new innovation can occur.

The ascending arrow blocks are representing the generations of computing platforms. As depicted, we are entering a new phase where the focus is on mobile platforms, mobile technology, and mobile applications.

The top arrow blocks diagonally going down right indicate the various phases of the Internet. We have gone from a stagnant Internet where the first websites were informational, but not interactive to the Web 2.0 phase where users could participate by adding comments and feedback. We are entering a third phase and the Internet of tomorrow is being built on technologies like cloud computing, social business, and mobile. This third phase of the Internet will handle billions and billions of connected computing devices as well as be able to enable the future of communication between humans and computers.

Furthermore, mobile devices now give us the ability to transport information, or access it online, nearly anywhere. Today, the people at the heart of this technology acceleration, Generation Y, cannot imagine a world without the Internet. They are the ones entering the workforce in droves, and they are a highly attractive and sought-after consumer segment. Because of this, business models are no longer limited to business-to-business or business-to-consumer. Instead, these new generations of technology, and people, have created a bidirectional business model that spreads influential content from consumer-to-consumer to communities-to-business.

[4] HorizonWatching, Eleven Technology Trends to watch in 2012, Bill Chamerlin
http://www.slideshare.net/HorizonWatching/horizon-watching-2012-technology-trends-12jan2012

In a nutshell, the IT landscape from today with a somehow decoupled view of IT and business aspects will evolve over time into a business model, where IT is the function of the business. This transformation of IT from a cost center to a profit center does not happen overnight. It requires a well thought out strategy and implementation, in which the data center plays the most important role.

In this context, network planning for a modern data center must take a holistic, long-term view that considers the network, servers, storage, and applications as well as their end-to-end manageability.

1.1.2 Attributes of the modern data center

The business requirements and trends that we have identified place specific demands upon data center infrastructures. The data center needs to provide services that leverage data center characteristics as listed next.

Simplification

Over the past few years there have been significant developments that are designed to reduce the complexity of the infrastructure and reduce the operational management burden of the data center.

These simplifications typically reduce the data center components that in turn reduce the failure domain of the system. Alternatively, fewer management “touch points” will ease troubleshooting and provisioning.

All of these innovations are designed to increase uptime and reduce the operational management burden.

Greater visibility

Hardware and software vendors such as IBM and Cisco are constantly trying to improve the instrumentation capabilities of their products. Greater visibility into traffic flows, utilization, error thresholds will provide more information to predict and avoid system or component outages.

This increase in management data can be used in many ways such as SLA reporting or automated reconfiguration to avoid problems. Additionally, there are great improvements being made on the DC toolsets to monitor, collect and correlate this data and present in the most efficient way

Efficient facilities

The “Green Agenda” has gathered momentum and legislation and is now playing a part in data center design. All the players in the data center are now obliged to make their equipment as energy efficient as possible. This ranges from the building power and cooling infrastructure, to cabling, servers, storage, and networks.

The data center vendors are working hard to drive out inefficiencies and optimize the available power. Ultimately, the reduced power consumption will reduce the carbon footprint as well as the costs for running the data center.

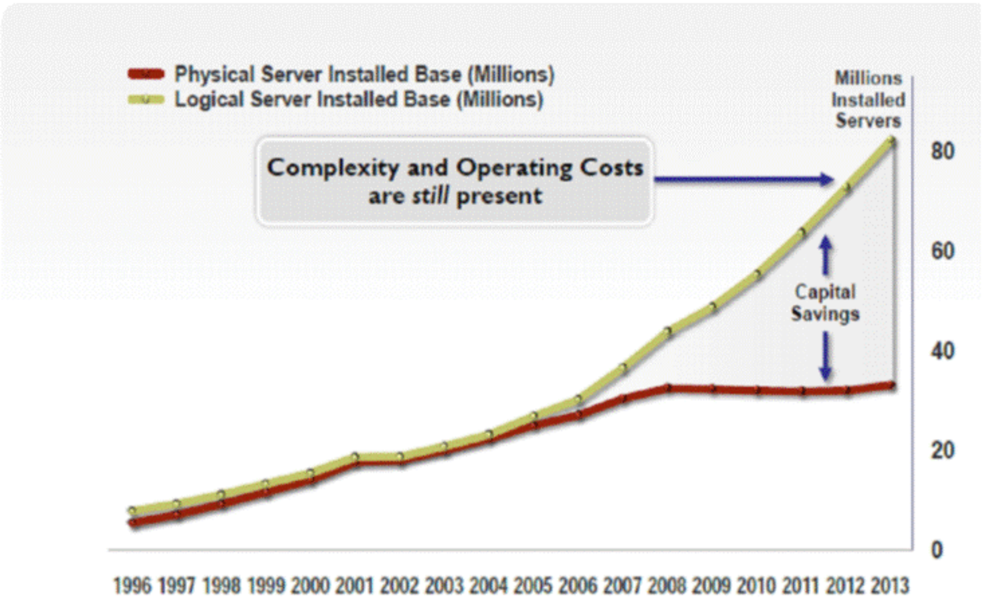
Other efficiencies can be gained from virtualization and convergence which are discussed in detail in this redbook. The consolidation of physical appliances into virtual services (servers, networks, and so on) has delivered the most significant benefits, and this evolution is continuing at a rapid pace.

Virtualization

Classic examples of virtualization are network VLANs, server consolidation (IBM PowerVM®, VMware vSphere) and storage partitioning and VSANs.

Virtualization for servers is a mature technology and is perceived to be one of the biggest impacts to IT in the last 20 years. The consolidation of many physical devices into fewer boxes with multiple logical entities is now spreading from the network and server world into storage and other domains.

Figure 1-2 shows the constant increase of virtualized servers.



Source: IDC

Figure 1-2 Server virtualization trend

Virtualization has driven knock-on innovations that address security and availability concerns. Secure Multi-Tenancy is now available for the most demanding customers and applications.

While reducing hardware and facilities, virtualization has typically added more complexity to the management of the infrastructure. The software toolsets to manage virtualized environments have matured significantly to mitigate these administrative challenges.

Virtualization is a technology enabler that delivers multiple IT benefits:

- Power and cooling savings
- Reduced data center footprint
- Faster provisioning
- Increased uptime and better disaster recovery
- Maximized asset utilization
- Investment protection
- Workload mobility and cloud empowerment

Secure multi-tenancy

Data centers of all sizes typically accommodate multiple communities that require isolated environments. The definition of a "tenant" will differ between data centers, but the most common environments include these:

- Applications: Regulatory compliance may dictate the secure boundaries for an application. One example is payment card industry (PCI) compliance that specifies which applications may talk to each other and which applications must be "ring fenced."
- Customers: In service provider data centers, there may be many paying customers sharing the same physical infrastructure. Each customer must be secured to avoid leakage between tenants. The isolation technique between customers will depend upon the customer specific requirements. This may be as simple as firewalled VLANs, or may be as extreme as "air gaps" for severe requirements (such as defence industries).
- Line of Business: Many government and enterprise customers have departmental boundaries that dictate the isolation of IT. Examples of LoB separation might be different divisions within the same enterprise (for example, sales and engineering).

The benefits of multi-tenancy include asset optimization, reduced costs and sometimes data sharing. The increased demand for data center multi-tenancy has driven further technology innovation, especially regarding high availability (such as in-service software upgrades).

When organizations share the same physical environment, the infrastructure architect must be careful to consider the quality of service requirements for each entity as well as the interdependencies such as firmware release management.

This increased focus on robust HA is even driving the change in network topology. It is now popular to spread the risk of failure across multiple boxes as opposed to focusing on component uptime. Historically it was typical to see just two devices providing core network availability, but there is a growing appeal of spreading the risk of failure across many "spine" devices.

Integrated systems

Integrated systems deliver a turnkey solution for application hosting, whereby the storage, compute and network access are combined in a single architecture. This unified architecture can offer greater flexibility and agility than the legacy data center infrastructure.

Sometimes known as a Point of Delivery (PoD) architecture, the integrated system ensures a smaller span of control. The systems can be managed as a single entity and service provisioning tasks are combined to accelerate delivery. Ideally the PoD and has a single pane of glass for management and provisioning. The objective of the PoD is to provide a single pool of "IT" as opposed to a network pool, a compute pool and a storage pool.

The deployment of integrated systems ensures standardization of infrastructure across data centers which offers benefits such as efficient remote management of lights-out data centers. Furthermore, restricting the infrastructure components options and their variability simplifies the automation of provisioning tasks without the need to develop custom scripts that depend on individual infrastructure components and that are difficult and costly to maintain.

It is typical to see several "Points of Delivery" in a data center where the network needs to provide connectivity to external services such as WAN, Management, and Internet connectivity in a modern data center. These PoD interconnects will drive specific connectivity requirements to ensure appropriate bandwidth and security controls. The adoption of integrated systems alongside the legacy data center architecture is changing the traditional model of infrastructure topology.

Software tools to manage the PoD environment are unified to ease delivery and enable automation of tasks, however, it is impractical in most cases to build a parallel management stack for specific environments. Careful consideration must be taken when deploying management tools to ensure that they will work with the old and the new.

There are operational advantages in greater collaboration between server, network, storage, facilities and virtualization teams. Integrated Systems drive this collaboration and often accelerate the convergence of the IT silos, thus reducing delays in operations.

Automation

Automation in the data center dramatically reduces operations costs and the potential for misconfiguration while optimizing capacity utilization and accelerating service delivery. Data center automation can be an enabler for self service provisioning of dynamic IT resources.

Efficient automation of service delivery requires orchestration of the entire infrastructure. In recent years there have been huge leaps forward in capabilities of automation software stacks. Vendors of hardware and software platforms for data center infrastructure are delivering new capabilities to enhance and accelerate the automated data center.

In the network space, examples of this innovation include Software Defined Networking, Open APIs, Overlay Networks, and virtualized services.

Workload mobility

Workload mobility is the ability to move a set of applications over a distance without effort and disruption. There are many benefits to workload mobility, such as these:

- Moving applications to optimize hardware assets and balance utilization
- Moving workloads between or within data centers for higher availability
- Performing graceful migration to new data center facilities
- Retiring hardware without application downtime

Workload mobility has introduced new stresses for the network. Moving a VM requires up to 40 MB of data transfer, so traffic management techniques must ensure that there is sufficient bandwidth to migrate the machine in a timely fashion without adversely impacting other traffic flows.

Additionally, a multitude of network services need to be adapted for the users to locate and access the VM. The migrated machine must retain the same security policy, management, backup and availability characteristics after the move.

Graceful scaling

The explosive growth of data and multi tiered applications is putting increased demands upon the data center facility. Network specific challenges include these:

- Increase bandwidth and reduce latency across the data center
- Increase port count while improving device density and efficiency
- Use more logical partitions and associated resources (such as VLAN, addressing space)
- Reduce number of management interfaces and overall complexity of the network
- Make a shift from physical to logical network services
- Change traffic patterns from “north-south” to “east-west”
- Have any-to-any IP connectivity with seamless connectivity at Layer 2 or Layer 3
- Accommodate legacy IT alongside the new modular data center network design concepts

- Convergence of Ethernet and Fibre Channel
- No-disruptive network changes to ensure an “always on” service
- A demand to optimize all resources in an active/active manner (links, and so on)

The list above grows longer every day and network administrators are struggling to balance the rapid pace of technology against the uptime demands of the business. The modern data center must be flexible enough to grow without impacting the services it delivers today.

Network design has shifted in recent times from static rigid architectures to a more elastic environment that can flex rapidly with the business needs of the data center. This book shows the reader some best practices to deliver a scalable and agile data center.

Efficient Data Center Interconnect

Data Center Interconnect (DCI) is an essential characteristic of most data centers today. The requirement for high availability is driving a model towards application hosting across multiple physical facilities.

The connectivity options between data centers vary between customers, applications and availability requirements. The geographic location and external service provider capabilities often influence the choice of DCI connectivity.

Network solution providers such as Cisco have many options to ensure that the business and application needs between facilities are provided in the most efficient manner. For example, the network requirements for a “failover facility” are significantly different from an “active/active” distributed data center. Likewise, the requirements to connect to an external cloud service provider may be drastically different than between enterprise data centers.

The ideal DCI connectivity technology is influenced by many factors. These include;

- Distance between facilities
- Local service provider capabilities
- Application latency and bandwidth requirements
- Layer 2 or Layer 3 communication
- Security concerns (transport, firewall, regulatory compliance, and so on)
- Cost (OPEX and CAPEX)
- Workload and hypervisor capabilities or hypervisor agnostic
- Future growth predictions
- Restoration of Service limitations (RTO, RPO)

This book is not specifically focused on DCI, but the reader can learn more from this site:

http://www.cisco.com/go/dci

Cost reduction

There are many factors that impact the total cost of data centers. The cost of maintaining the data center infrastructure over many years far outweighs the capital costs to purchase the infrastructure.

In the current economic climate, there is a pervasive need to drive cost out of the business. Enterprise and service providers are keen to reduce the costs of IT and data centers are typically one of the largest budget items.

CIOs are reviewing the efficiency of all their operations and are typically introducing the following considerations:

- Consolidation of many data center facilities into fewer data centers
- Rationalization of applications
- Virtualization of server, storage, network, and facilities
- Outsourcing or offshoring IT functions to external providers or cloud
- Streamlining operations to reduce full time employees (FTEs)
- Introducing IT automation
- Optimizing physical data center facilities (power, cooling, cabling, and so on)
- Introducing IT standardization to reduce IT sprawl and minimize operations
- Considering open source alternatives
- Changing employee behavior through new technology (BYOD, home working, and so on)

1.1.3 Impact of virtualization in the modern data center

It is, meanwhile, a fairly common approach to consolidate and virtualize the IT resources; servers, storage, applications, networks, and even desktops. Virtualization decouples hardware and software resources from their physical implementation. It enables organizations to make IT resources available to applications as needed rather than requiring a dedicated server for each application.

For example, applications can call on additional server processing power or storage capacity to meet changing demands. And with the addition of advanced, automated provisioning capabilities to this virtualized environment, resources can be dynamically applied, providing organizations with an even more efficient, responsive and flexible infrastructure. These characteristics underpin the modern data center approach.

To understand how the network must support the dynamic and virtualized infrastructure underpinning the modern data center, it is useful to look at how the IT infrastructure is evolving from the static, one-application-per-device past to the virtualized and dynamic infrastructure of the future. This evolution is depicted in Figure 1-3.

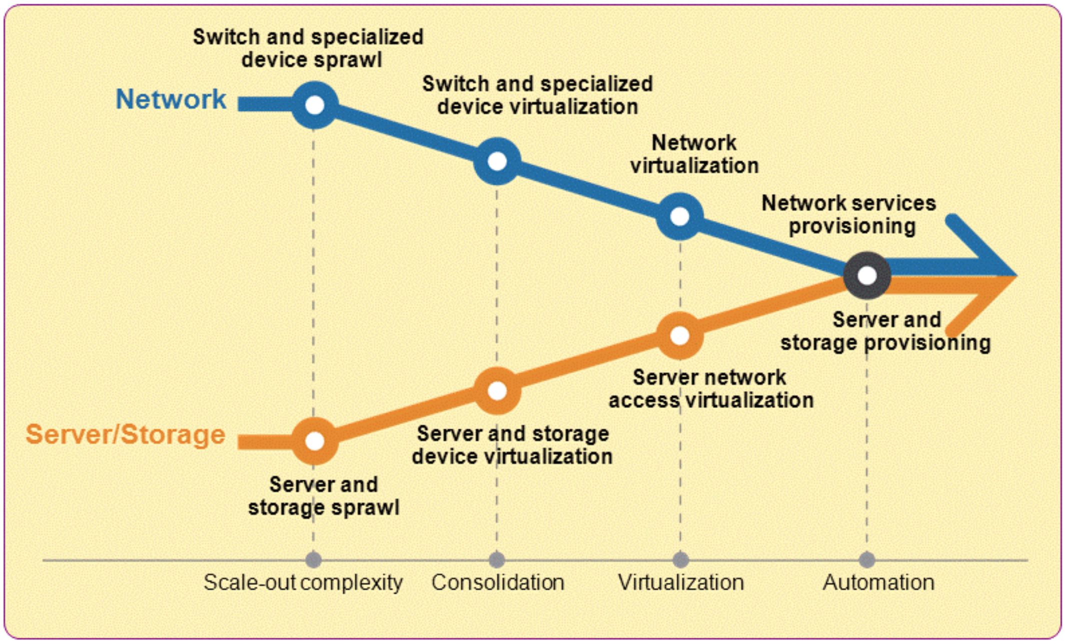

Switch and specialized device sprawl
Network
Switch and specialized device virtualization
Network virtualization
Network services provisioning
Server and storage provisioning
Server network access virtualization
Server/Storage
Server and storage device virtualization
Server and storage sprawl
Scale-out complexity
Consolidation
Virtualization
Automation

Figure 1-3 Impact of Virtualization

Phase 1: Scale-out complexity: In distributed computing, each environment is managed independently. This approach relies on a relatively simple data center network design that includes a wide array of switches and specialized devices that are static and endpoint agnostic. The result is growth in the number of specialized network, server and storage devices that are expensive, vastly under-utilized and difficult to maintain.

Phase 2: Consolidation: In this phase of the evolution, IT organizations physically consolidate data centers, servers, storage and network devices and take advantage of device-specific virtualization to increase utilization and simplify management. The interrelationship between the network and the server and storage infrastructure grows in importance as focus shifts to efficiency and cost-effectiveness.

Phase 3: Virtualization: During this phase, network flexibility and responsiveness are critical as applications are virtualized across multiple platforms on virtual machines and data and storage adapters are converged. The network must support the multiple layers of virtualization, connect virtual server and storage environments, and support platform-specific network requirements.

Phase 4: Automation: The final phase is to realize the promise of dynamic network, server and storage resources across the infrastructure. In this phase, the network, server and storage resources are virtualized, integrated, scalable and secure. Provisioning is automated through service management, and multiple resource options are supported. The infrastructure also supports cross-platform virtual machine mobility.

Throughout the evolution of the IT infrastructure, you can see the increasing importance of stronger relationships between infrastructure components that were once separately planned and managed. In a modern data center, the applications, servers, storage, and network must be considered as a whole and managed and provisioned jointly for optimal function.

Security integration is at every level and juncture to help provide effective protection across your infrastructure, and across your business. Rapid innovation in virtualization, provisioning and systems automation necessitates expanding the considerations and trade-offs of network capabilities. Additionally, the ultimate direction for dynamic provisioning of server, storage and networking resources includes automatic responses to changes in business demands, such as end-user requests, business continuity and energy constraints, so your current network design decisions must be made within the context of your long-term IT and business strategies.

Impact of server and storage virtualization trends on the data center network

The current landscape in the data center is characterized by several business issues and leading edge technology trends. From a business perspective:

- New economic trends necessitate cost-saving technologies, such as consolidation and virtualization, like never before.
- Mergers and acquisition activities bring pressure for rapid integration, business agility, and fast delivery of IT services.
- The workforce is getting increasingly mobile; pervasive access to corporate resources becomes imperative.
- Data security is becoming a legal issue. Companies can be sued and fined, or incur other financial damages for security breaches and for non-compliance.
- Energy resources are becoming a constraint and should be considered a requirement for an enterprise-grade network design.
- Users' dependency on IT services puts increasing pressure on the availability and resilience of the infrastructure, and competition and on demand for real-time access to services and support.

Those business drivers are causing new trends to develop in data center architectures and the network must adapt with new technologies and solutions, as follows:

- The new landscape brings the focus back to the centralized environment. The distributed computing paradigm may not always match the needs of this dynamic environment. Virtualization and consolidation together put more traffic on single links since the “one machine-one application” correlation is changing. This can cause bottlenecks and puts pressure on the access layer network.

  For example, the End-of-Row (EoR) model may be unfit to sustain this new traffic pattern and so the Top-of-rack (ToR) model is rapidly becoming attractive. The access switch sprawl can be effectively managed by consolidating several access nodes into a bigger, virtual one.
- Virtualization brings a new network element into the picture, the virtual switch (vSwitch). access layer networking has evolved from a simple physical NIC-to-physical switch connection to a complex virtual network infrastructure living within the physical server in the form of vNICs and vSwitches and associated technologies, a trend that is likely to continue.

Figure 1-4 shows the new network elements in the virtualized servers.

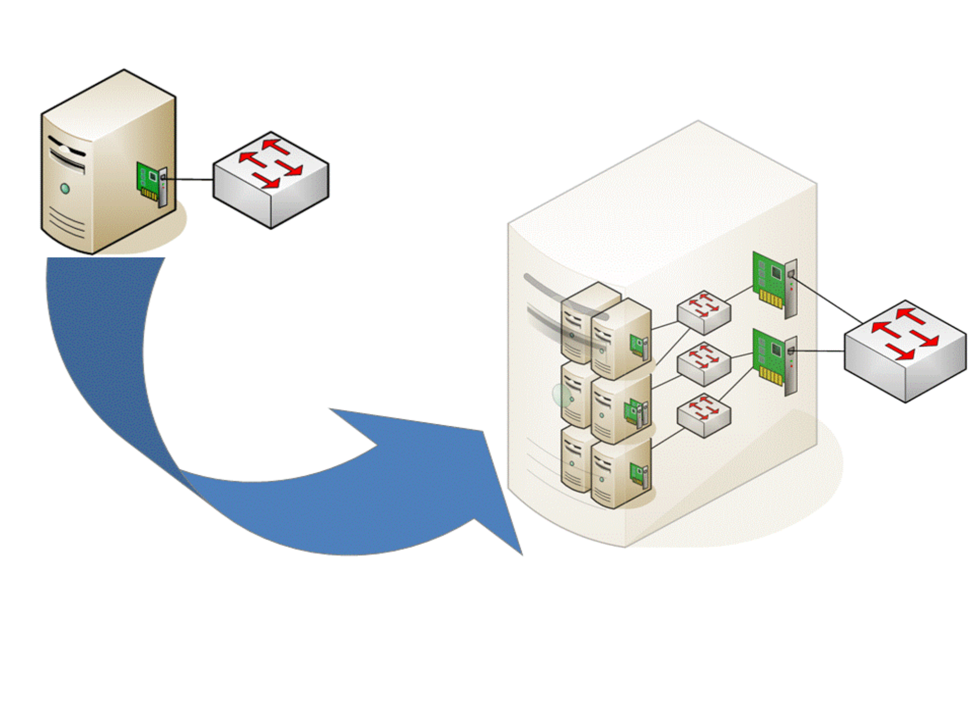


Figure 1-4 Server virtualization

- The 10 GE and 40 GE technology is reaching a price point for attractiveness. This increased capacity makes it possible to carry the storage traffic on the same Ethernet cable. This is also an organizational issue because the storage and the Ethernet networks are usually designed by different teams.

Figure 1-5 shows storage and data convergence and its impact on the network infrastructure.

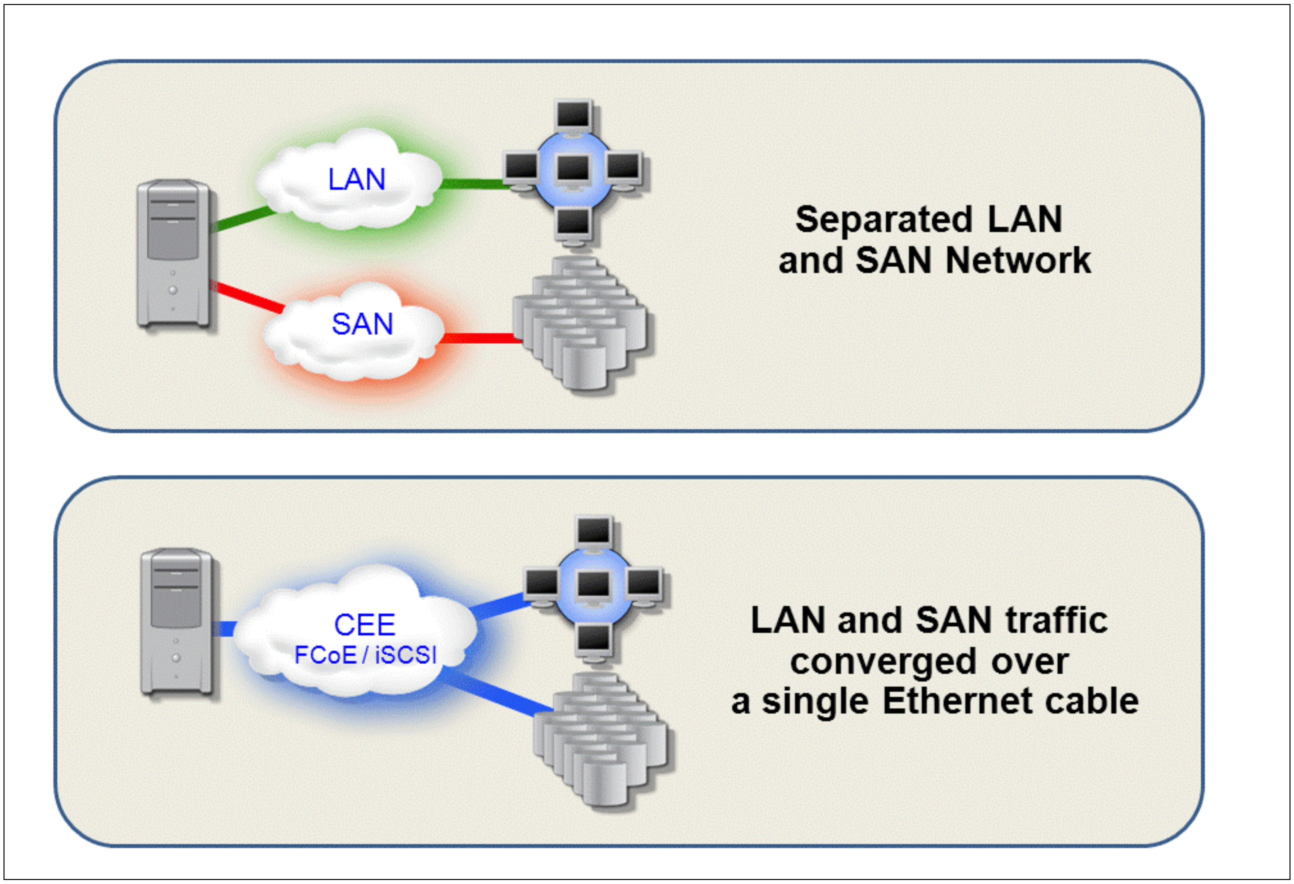

LAN
SAN
Separated LAN and SAN Network
CEE
FCoE / iSCSI
LAN and SAN traffic converged over a single Ethernet cable

Figure 1-5 LAN and SAN convergence

- The consolidation of the distributed IT environment into single distributed data centers struggles with an increasingly mobile workforce and distributed branch offices. Distance, and thus latency, increases for remote access users. New technologies are available to address this issue and bring better response times and performance to this scattered user base.

  This situatioin becomes even more critical for applications that exchange significant amounts of data back and forth, because both bandwidth and latency must be optimized on an application and protocol level, but optimizing network utilization is not usually a priority during application development.

- Designing a traditional data center network usually implied deploying firewalls between the users and the applications, but this is rapidly changing because network and security must be designed together with a comprehensive architecture in order to ensure that security appliances do not decrease end-to-end performance. The high-level design should include secure segmenting based on business needs and priorities. Security services should be coupled with the applications even if these are mobile across data centers.

  Additional technologies can be used to provide this secure segmentation, for example virtual routers. At the same time remote users must be authenticated and access is to be controlled and virtualized to enter the IP core and use the centralized, virtualized IT resources.

Figure 1-6 illustrates access control and defined policies, as well as support for remote and mobile application access.

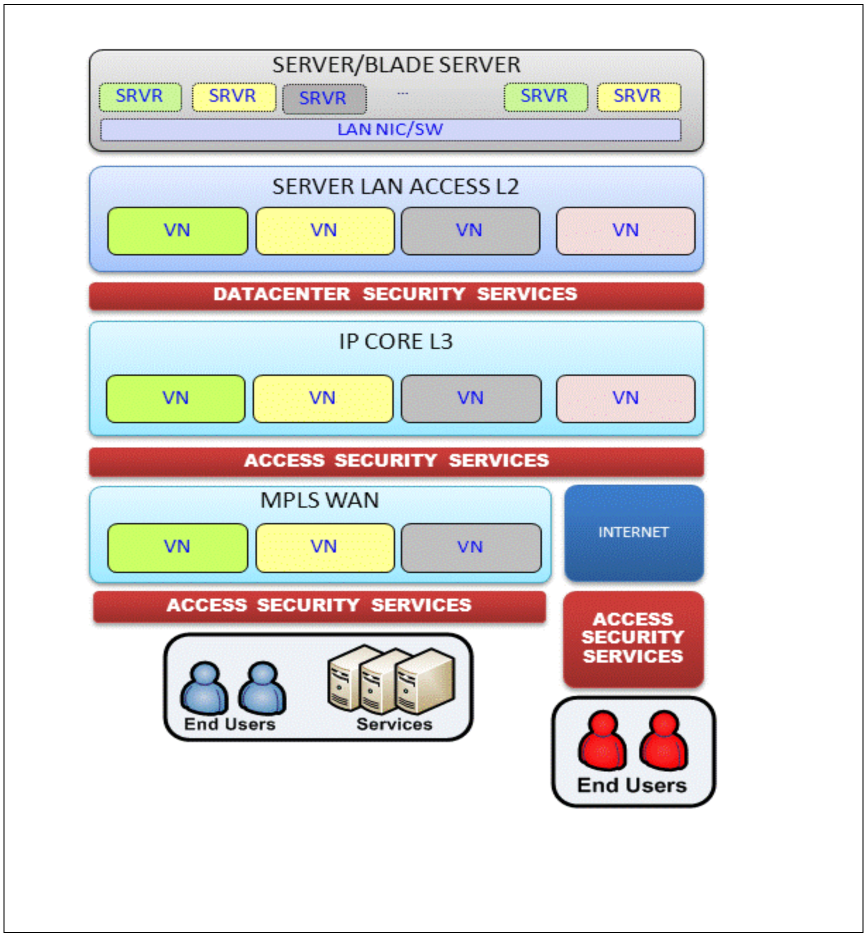


Figure 1-6 Access control and defined policies

1.1.4 The journey towards cloud computing

In the last few years, IT has embarked on a new paradigm: cloud computing. Although cloud computing is only a different way to deliver computer resources, rather than a new technology, it has sparked a revolution in the way organizations provide information and service.

Originally IT was dominated by mainframe computing. This sturdy configuration eventually gave way to the client-server model. Contemporary IT is increasingly a function of mobile technology, pervasive or ubiquitous computing, and of course, cloud computing. But this revolution, like every revolution, contains components of the past from which it evolved.

Thus, to put cloud computing in the proper context, keep in mind that in the DNA of cloud computing is essentially the creation of its predecessor systems. In many ways, this momentous change is a matter of “back to the future” rather than the definitive end of the past. In the brave new world of cloud computing, there is room for innovative collaboration of cloud technology and for the proven utility of predecessor systems, such as the powerful mainframe. This veritable change in how we compute provides immense opportunities for IT personnel to take the reins of change and use them to their individual and institutional advantage.

What is cloud computing?

Cloud computing is a comprehensive solution that delivers IT as a service. It is an Internet-based computing solution where shared resources are provided like electricity distributed on the electrical grid. Computers in the cloud are configured to work together and the various applications use the collective computing power as if running on a single system.

The flexibility of cloud computing is a function of the allocation of resources on demand. This facilitates the use of the system's cumulative resources, negating the need to assign specific hardware to a task. Before cloud computing, websites and server-based applications were executed on a specific system. With the advent of cloud computing, resources are used as an aggregated virtual computer. This amalgamated configuration provides an environment where applications execute independently without regard for any particular configuration.

Why the rush to the cloud?

Cloud computing has already passed the peak of inflated expectations in the Gartner Technology Hype Cycle from July 2011[5], but is still one of the prevailing IT trends as we head into the new decade. Cloud computing has already moved from a talking point to another way to deliver IT as one of the key transformation technologies in the marketplace. By providing greater levels of automation, orchestration, provisioning, and deployment, cloud computing can help organizations become more nimble, reduce operating costs, improve application performance, and better allocate their compute resources.

Benefits of the cloud

Cloud computing fundamentally changes the way that IT services are delivered to organizations. Instead of both owning and managing IT services for themselves, or using an outsourcing approach built around dedicated hardware, software, and support services, organizations can use cloud computing to meet their IT requirements using a flexible, on-demand, and rapidly scalable model that requires neither ownership on their part, nor provisioning of dedicated resources.

Some of the benefits that cloud computing brings are as follows:

- Reduced cost: Cost is a clear benefit of cloud computing, both in terms of capital expense (CAPEX) and operational expense (OPEX). The reduction in CAPEX is obvious because an organization can spend in increments of required capacity and does not need to build infrastructure for maximum (or burst) capacity. For most enterprises, OPEX constitutes the majority of spending; therefore, by utilizing a cloud provider or adopting cloud paradigms internally, organizations can save operational and maintenance budgets.
- Flexibility: Flexibility benefits derive from rapid provisioning of new capacity and rapid relocation or migration of workloads. In public sector settings, cloud computing provides agility in terms of procurement and acquisition process and timelines.
- Improved automation: Cloud computing is based on the premise that services can not only be provisioned, but also de-provisioned in a highly automated fashion. This specific attribute offers significant efficiencies to enterprises.
- Focus on core competency: Government agencies can reap the benefits of cloud computing in order to focus on its core mission and core objectives and leverage IT resources as a means to provide services to citizens.
- Sustainability: The poor energy efficiency of most existing data centers, due to poor design or poor asset utilization, is now understood to be environmentally and economically unsustainable. Through leveraging economies of scale and the capacity to manage assets more efficiently, cloud computing consumes far less energy and other resources than a traditional IT data center.

[5] Gartner Hype Cycle 2011
http://www.gartner.com/it/page.jsp?id=1763814

But cloud computing does not come without its challenges. In particular, organizations with traditional infrastructures may find that their networks are not set up to take full advantage of the cloud, and they may suffer from poor application performance or expose themselves to security risks when migrating to the cloud.

The spectrum of cloud service models

As prominent as cloud computing has already become in today's enterprises, you can realize that the number of cloud service models are increasing. There is meanwhile a formal interpretation of the role services play in the cloud, by offering a new term for the concept called XaaS (anything as a service). XaaS is a collective term said to stand for a number of things including "X as a service," "anything as a service" or "everything as a service". The acronym refers to an increasing number of cloud services.

The most common examples of XaaS are:

- Software as a Service (Saas)
- Platform as a Service (PaaS)
- Infrastructure as a Service (Iaas)

The combined use of these three is sometimes referred to as the SPI (SaaS, PaaS, Iaas). Other examples of XaaS include Storage as a Service (SaaS), Communications as a Service (CaaS), Network as a Service (NaaS) and Desktop as a Service (DaaS).

The most common cloud services can be described in the following ways.

Infrastructure as a Service (IaaS)

The infrastructure is the foundation of the cloud. It consists of the physical assets: servers, network devices, storage disks, and so on. IaaS you do not actually control the underlying infrastructure, but you do have control of the operating systems, storage, deployment applications, and, to a limited degree, control over select networking components.

Platform as a Service (PaaS)

Platform as a Service (PaaS) provides the application infrastructure and access to operating systems and associated services. It provides a way to deploy applications to the cloud using programming languages and tools supported by the provider. You do not have to manage or control the underlying infrastructure, but you do have control over the deployed applications and, to some degree over application hosting environment configurations.

Software as a Service

Software as a service (SaaS) is a software model with applications centrally hosted in a cloud computing environment and accessed by users over the Internet. Many business applications use SaaS as a common delivery model because of its ability to help reduce costs and simplify deployment.

Figure 1-7 demonstrates the three most common cloud services.

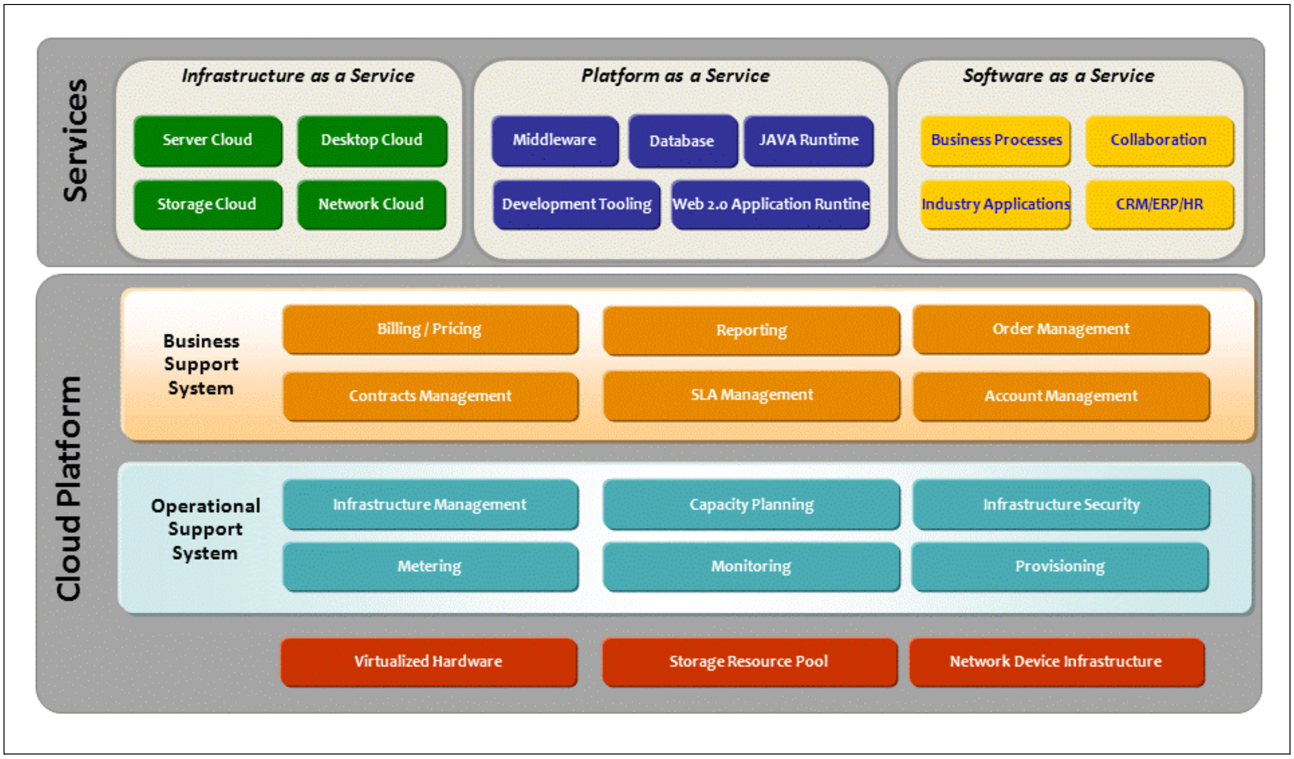


Figure 1-7 Simplified version of cloud services

The distinction between the five categories of cloud offering is not necessarily clear-cut.

Cloud formations

There are three types of cloud delivery models: private (on premise), public, and hybrid:

- Public clouds are available to the general public or a large industry group and are owned and provisioned by an organization selling cloud services. A public cloud is what is thought of as the cloud in the usual sense; that is, resources dynamically provisioned over the Internet using web applications from an off-site third-party provider that supplies shared resources and bills on a utility computing basis.
- Private clouds exist within your company's firewall and are managed by your organization. They are cloud services you create and control within your enterprise. Private clouds offer many of the same benefits as the public clouds; the major distinction being that your organization is in charge of setting up and maintaining the cloud.
- Hybrid clouds are a combination of the public and the private cloud using services that are in both the public and private space. Management responsibilities are divided between the public cloud provider and the business itself. Using a hybrid cloud, organizations can determine the objectives and requirements of the services to be created and obtain them based on the most suitable alternative.

1.2 Architecture evolution

The characteristics and requirements of the modern data center identified in previous sections are changing the way that data centers are built. We explore and contrast these changes in this section, comparing the traditional data center against recent developments and needs.

1.2.1 Traditional data center network architecture

Hierarchical network design has been commonly used in enterprise networking for many years. This model uses redundant switches at each layer of the network topology for device-level failover that creates a highly available transport between end nodes using the network. data center networks often require additional services beyond basic packet forwarding, such as server load balancing, firewalls, or intrusion prevention. These services might be introduced as modules populating a slot of one of the switching nodes in the network, or as standalone appliance devices. Each of these service approaches also supports the deployment of redundant hardware to preserve the high availability standards set by the network topology.

A structured data center environment uses a physical layout that correlates tightly to the hierarchy of the network topology. Decisions on cabling types and the placement of patch panels and physical aggregation points must match the interface types and densities of the physical switches being deployed. In a new data center build-out, the two can be designed simultaneously, also taking into consideration the constraints of power and cooling resources. When seeking to avoid significant new investment within an existing data center facility, the pre-existing physical environment of cabling, power, and cooling can strongly influence the selection of switching platforms. Careful planning in conjunction with networking requirements and an eye toward flexibility for the future is critical when designing the physical data center environment. Taking a modular approach to data center design provides flexibility and scalability in both network topology design and utilization of physical resources.

Figure 1-8 illustrates the primary network switching layers of the hierarchical network design reference model for the data center environment. The overall hierarchical model is similar to the reference topology for enterprise campus design, but the term *aggregation layer* replaces the term *distribution layer*. This denotes the fact that the data center network is less concerned with distributing network access across multiple geographically disparate wiring closets and is more focused on the aggregation of server resource, and providing an insertion point for shared data center services.

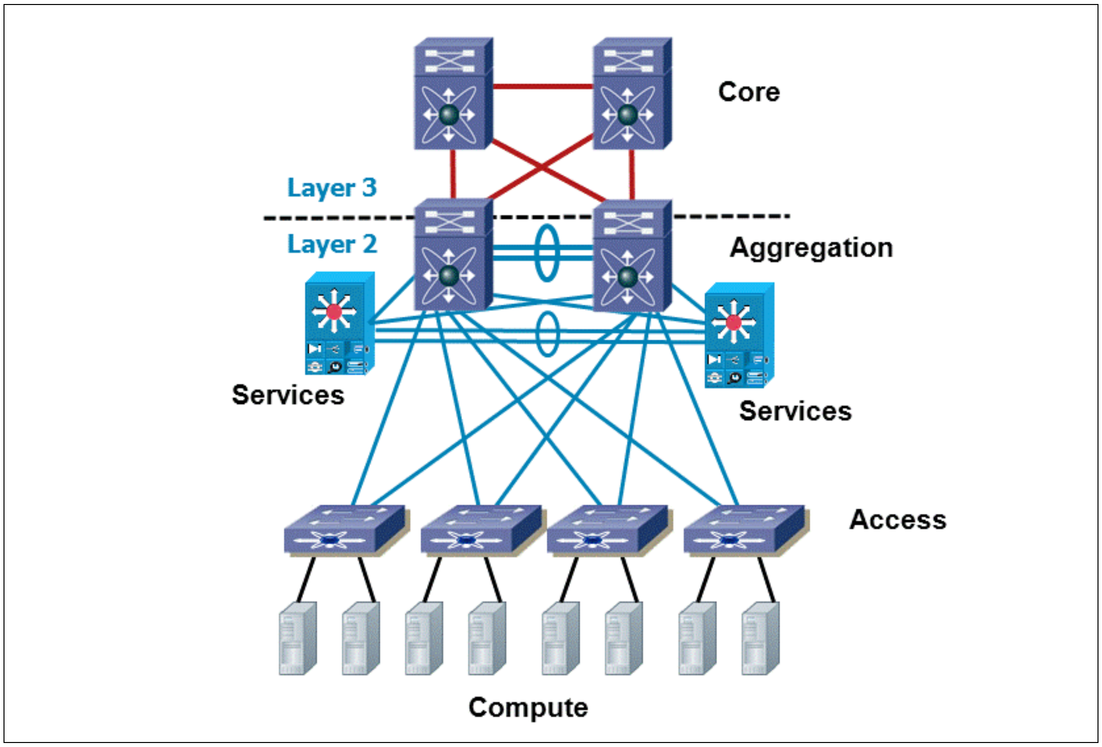


Figure 1-8 Hierarchical network design

Figure 1-8 illustrates the boundary between Layer-3 routed networking and Layer 2 Ethernet broadcast domains at the aggregation layer. Larger Layer 2 domains increase the physical flexibility of the data center, providing the capability to manually or virtually relocate a server to a different physical rack location with less chance of requiring a change of IP addressing to map to a specific subnet. This physical flexibility comes with a trade-off. Segregating the network into smaller broadcast domains results in smaller spanning tree domains and failure domains, which improve network stability, reduce convergence times and simplify troubleshooting. When determining how to scale Layer 2 domains, the network architect must consider many factors including the access switching model in use and the nature of the underlying applications being serviced.

Features such as bridge assurance and dispute mechanism incorporated into switching products allow greater scalability of Layer 2 domains with increased stability of the STP. These features are discussed in detail later in this book. Many data center networks prior to the evolution of server virtualization were built with different requirements and made the best of the technologies available at the time.

More detail on the functional network layers is provided in the Cisco paper, *Data Center Design—IP Network Infrastructure:*

http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DC_3_0/DC-3_0_IPInfra.html

The traditional 3-tier network architecture has served the industry very well over many years and is still the dominant structure for large data centers. Classic core-aggregation-access architecture has been the prevalent topology for large scale data centers from the mid-1990s for at least the following 10-15 years. It provided exceptional scalability and availability characteristics, given the technologies available at the time.

The 3-tier approach consisting of the access, aggregation, and core layers permit flexibility in the following areas:

- Layer 2 domain sizing: When there is a requirement to extend a VLAN from one switch to another, the domain size is determined at the aggregation layer. If the access layer is absent, the Layer 2 domain must be configured across the core for extension to occur. Extending Layer 2 has the risk of uncontrollable broadcast issues related to extending Layer 2 domains, and therefore should be avoided.
- Network Services: An aggregation plus access layer solution enables services to be shared across the entire access layer of switches. This lowers TCO and lowers complexity by reducing the number of components to configure and manage.
- NIC teaming and HA clustering support: Supporting NIC teaming with switch fault tolerance and high availability clustering requires Layer 2 adjacency between NIC cards, resulting in Layer 2 VLAN extension between switches. This would also require extending the Layer 2 domain through the core, which was not recommended, especially where spanning tree protection mechanisms were the only option for L2.

Recent times have presented new features, capabilities, as well as new demands such as virtualization. Virtualization has introduced workload mobility, Layer 2 adjacency and a new tier of switching in within the host; the virtual switch. The virtual switch resides within the hypervisor and the management and capabilities of each hypervisor switch are inconsistent.

The adoption of blade systems for high density computing environments introduced the concept of embedded switches which added even more complexity to the data center topology. The original 3-tier model for data center connectivity was in danger of growing to 4 or 5 tiers, which would be operationally impractical and suboptimal for most environments.

Feature parity and management tools are difficult to achieve in such a complex environment, which negatively affects network uptime and flexibility. Troubleshooting challenges in highly complex network environments was demanding new thinking with regard to the data center topology.

The traditional architecture was originally designed for “north-south” client-server traffic, which means that the majority of the traffic was between the access and core layers. As we describe later, new paradigms in the data center are driving traffic patterns to a more “east-west” model which refers to traffic moving across the access layers.

1.2.2 Workload trends affecting data center architecture

This section discusses some of the applications that are changing the way data centers are built. Specifically, we cover some of the new computing paradigms that are impacting the network infrastructure design.

Server virtualization

Server virtualization has created an inflection point in data center network design. In particular, the following trends have impacted best practices:

- Live migration techniques requiring Layer 2 adjacency for application mobility within and between data centers. Typically a connection of 1 Gbps between hosts is recommended using the same subnet and broadcast domain.
- Common, shared storage across the cluster means additional connectivity to replicate and synchronize the data stores. I/O latencies and data volume will impact the data center architecture.

- VM disaster recovery techniques between data centers such as VMware Site Recovery Manager. Typically the round trip latency should be under 5ms and distance should be no more than 400km (typically much shorter).
- VM and storage migration creates large volumes of data which can cause performance problems for other applications, so careful consideration of Quality of Service (QoS) needs to be applied.
- Hypervisor switching that shifts the network access layer into the host which impacts the end-to-end network manageability, feature and security consistency. The expansion of VM switching impacts the diameter and scale of the network and causes significant troubleshooting challenges.
- MAC address scalability is a key concern when server consolidation is implemented. A high density of VMs can overload the MAC tables in network equipment.
- Movement of application workload requires the network to direct traffic to a new location. Load balancing, firewall and other services need to be associated and coordinated with the application so they need to be flexible or portable.
- Cloud computing models which require connectivity for VM mobility between public and private data centers.
- Server virtualization has accelerated the adoption or IP network attached storage such as NAS or iSCSI. Careful placement of disk arrays on the network is imperative to ensure adequate performance. More recently, the storage industry is adopting a “unified storage” model whereby file and block storage is converged within an array. This is driving up the adoption of unified interfaces such as FCoE or iSCSI which changes the network termination interface and services.
- Dense computing requirements dictating the deployment of efficient platforms such as blade systems with embedded network access layers. Pass-thru modules can reduce the complexity and span of the network but dilute some of the cabling consolidation value of blade systems.
- PoD based architectures that house VM clusters are driving “hot spots” for network bandwidth. Server consolidation increases the bandwidth requirements and reduces the latency toleration from a physical machine. VM Clusters are often co-located in PoDs with high bandwidth requirements within the PoD and for PoD access. This is changing the traditional model of oversubscription at the network access layer.

Massively scalable data centers

Massively scalable data centers (MSDCs) are large data centers, with at least a physical thousand servers (sometimes hundreds of thousands), that have been designed to scale in size and computing capacity with little impact on the existing infrastructure. Often the MSDC hosts large distributed applications or a distributed platform for other applications.

Some of the reasons for the increased interest in MSDCs around the world are data center consolidation, server virtualization, and security. MSDCs are built on three main concepts:

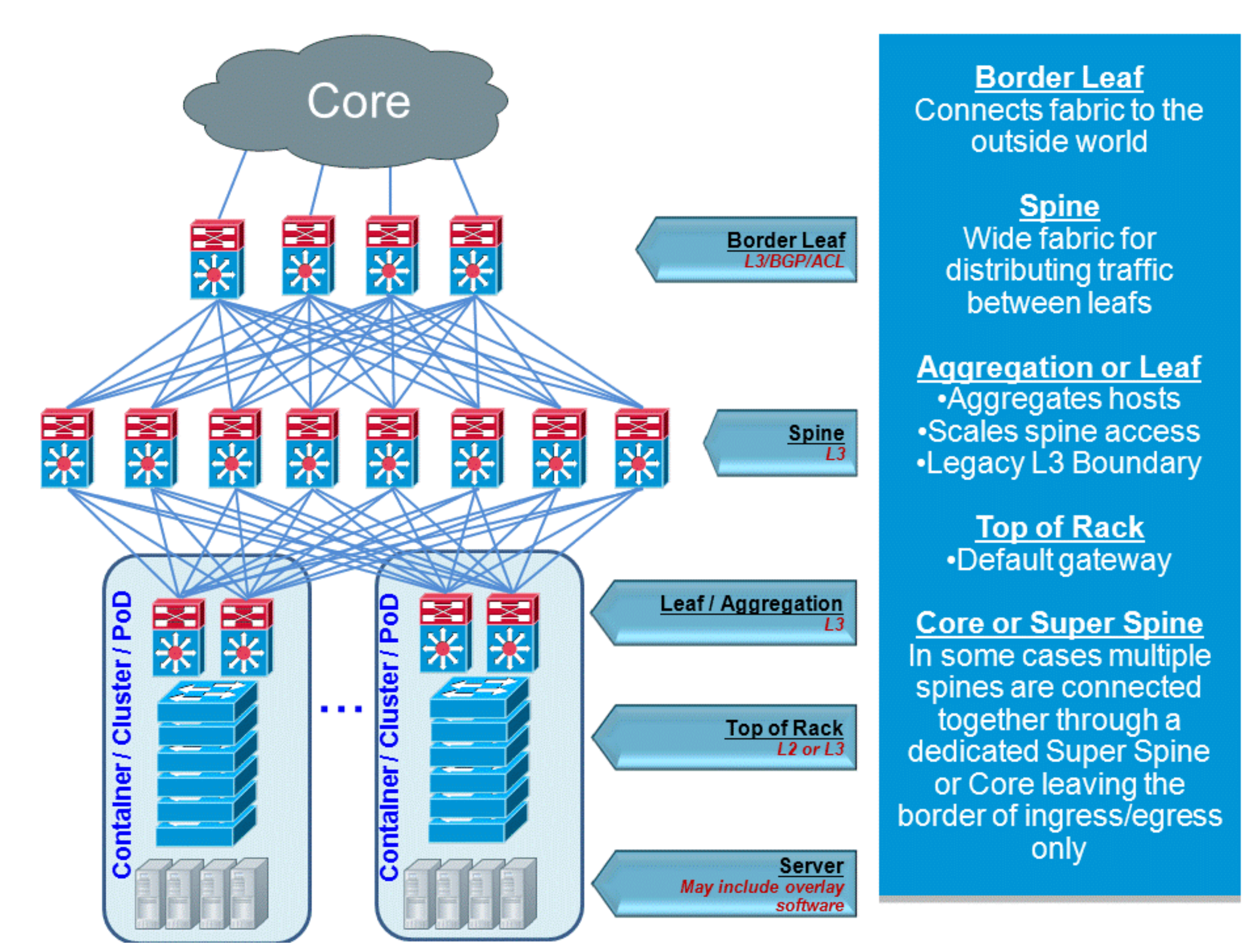
- Optimized power and cooling
- PoD-based building blocks
- Zero-touch provisioning and management

MSDCs introduce some significant challenges for networks as they require different characteristics than a general purpose “enterprise data center”. The first challenge is the sheer number of networking devices required to connect a very high number of servers.

Automation of this huge network infrastructure is imperative to ensure efficient operations of the data center. Network availability is typically achieved by spreading the load across multiple devices. In Figure 1-9, you will see that a Clos network architecture is used, which used a Spine-Leaf topology.

Multiple Spine switches are used to achieve scale and distribute the failure domain. If we lose one spine it is not a big deal because the other devices will take up the load. Switch redundancy is not the highest priority, system/topology redundancy is critical.

Interior Gateway Protocols such as OSPF or eBGP are often deployed across the DC, all the way to Top-of-Rack (ToR) switches to ensure rapid updates. MAC address scalability of the MSDC typically has to accommodate millions of addresses.



Border Leaf
Connects fabric to the outside world

Spine
Wide fabric for distributing traffic between leafs

Aggregation or Leaf
•Aggregates hosts
•Scales spine access
•Legacy L3 Boundary

Top of Rack
•Default gateway

Core or Super Spine
In some cases multiple spines are connected together through a dedicated Super Spine or Core leaving the border of ingress/egress only

Figure 1-9   MSDC network architecture

Another trend in MSDC is the adoption of network overlays; dynamic, elastic services on top of static network infrastructure. This allows VM mobility without worrying about L2 domains and reducing pressure on (and cost of) MAC and FIB tables. Examples of network overlays are VXLAN or NVGRE which are discussed later in this book.

Despite the appeal of hypervisor technology, many of the MSDC applications are run on bare metal systems without a hypervisor.

One of the most important innovations coming from the MSDC space is network programmability. One of the biggest challenges to MSDC operations is predicting change and forecasting growth. The vision for many MSDC customers is that the network should provision itself so these customers have been scripting network changes using tools such as Python and levering Cisco developments like PoAP (Power on Auto Provisioning) and Cisco EEM (Embedded Event Manager).

EEM details;

http://www.cisco.com/en/US/products/ps6815/products_ios_protocol_group_home.html

PoAP details;

http://www.cisco.com/en/US/docs/switches/datacenter/nexus3000/sw/fundamentals/503_U3_1/b_Nexus_3000_Fundamentals_Guide_Release_503_U3_1_chapter_0111.html

More recently, this notion of network has matured from several communities into “Software Defined Networking” (SDN). Cisco is providing an API to support SDN initiatives for each of its primary network operating systems (IOS, NXOS and IOS-XR) for developers to leverage.

Finally, the MSDC community is one of the driving forces of Open Source solutions for their large, expensive environments. Along with cloud service providers (CSPs), there is a shift towards open source tools and protocols to control and orchestrate the network services. Typical areas of interest are OpenStack and OpenFlow.

You can learn more about Software Defined Networks, Open Source networking and Virtual Network Overlays in Chapter 11 or at the following website:

http://www.cisco.com/go/one

This document is primarily focused on enterprise data center design, but many of the concepts introduced by Yahoo, Google, Apple, Microsoft, Facebook and Amazon are creeping into every day designs. The adoption of large distributed parallel job processing for enterprise applications is driving a trend towards MSDC like data center architectures.

You can learn more about MSDC from the Cisco paper:

http://www.cisco.com/en/US/prod/collateral/switches/ps5718/ps6021/at_a_glance_c45-652675.pdf

Ultra low latency and HPC

High Performance Computing (HPC) is often assumed to require very low latency between compute nodes for efficient operations. This is true for “tightly coupled” HPC applications, whereas “loosely coupled” applications are rarely impacted by latency considerations.

Low latency is certainly a consideration for many environments in the data center, but it is certainly not a requirement for most applications. Network vendors such as Cisco will typically offer different switching platforms that address different latency requirements.

Recent adoption of High Frequency Trading (HFT) techniques in the financial community have put the spotlight on network latency for certain niche applications. Financial institutes are building faster trading systems to gain competitive advantage to increase profitability.

In the case of HFT, nanoseconds can make a huge financial difference. Accurate timestamp is also a mandatory requirement for compliance; network and applications need accurate timing information to correlate all the events such as network diagnostics, application transactions, digital forensics, data and event-log analysis.

With transactions happening at nanoseconds, network devices need to have ultra low latency (measures in nanoseconds) and an efficient timing mechanism. PTP (Precision Timing Protocol), defined in IEEE1588, distributed time synchronization protocol for packet networks can provide nanosecond accuracy.

Big Data

Big Data refers to diverse data sets that are too large to manage and analyze with traditional IT techniques in a reasonable time-frame and cost. Recent technological advancements in distributed computing now make Big Data processing and analysis a reality and bring significant competitive advantage to businesses.

Examples of Big Data services and applications are IBM Watson™, Hadoop for data storage and analytics, or MySQL for real-time capture, read, and update operations. These applications generate lots of server-server traffic, so the “east-west” traffic requirements tend to be considerable.

The new Big Data computing model is influencing the network architecture in several key ways;

- Availability
- Buffering
- Data Node Speed (1G versus 10G)
- Oversubscription
- Latency

The Big Data compute platform is typically rack servers with plenty of onboard storage, so centralised NAS/SAN is typically not required for the cluster. The cluster is often 50-150 nodes, but can scale to 1000+ nodes. As with HPC, the term Big Data does not necessarily dictate high I/O or low latency requirements. Each Big Data application should be considered on its own merits, with input from the application teams, however it is quite common to see Hadoop clusters for multi-use.

Since resiliency is provided within the application, node redundancy is not the highest priority depending upon the use case and criticality of the job. Node failure still has the biggest impact on job completion time, so redundant links are recommended. Often redundant load sharing links will also provide increased bandwidth, availability and response time.

Some Big Data applications such as HDFS/MapReduce operations are very bursty in nature and a network that cannot handle bursts effectively will drop packets, so optimal buffering is needed in network devices to absorb bursts. Modern switches should not only provide appropriate buffering, but also buffer depth monitoring through an external API.

Non-blocking infrastructure is not typically a requirement for Big Data environments and small ratios of oversubscription are typical. Dual attached 10G Ethernet attached servers will typically provide the consistent job completion time, buffer utilization, reduced bursts and future proofing.

More information about IBM Watson is available here:

http://www.ibm.com/innovation/us/watson/

You can learn more about Big Data networking here:

http://www.cisco.com/go/bigdata

1.2.3 Multiple types of data center

The application environments listed previously are driving the adoption of multiple types of data centers.

Figure 1-10 summarizes the data center types, their requirements, and suggested Cisco networking platforms.

The advent of virtualization coincided with new developments in network availability and flexibility.

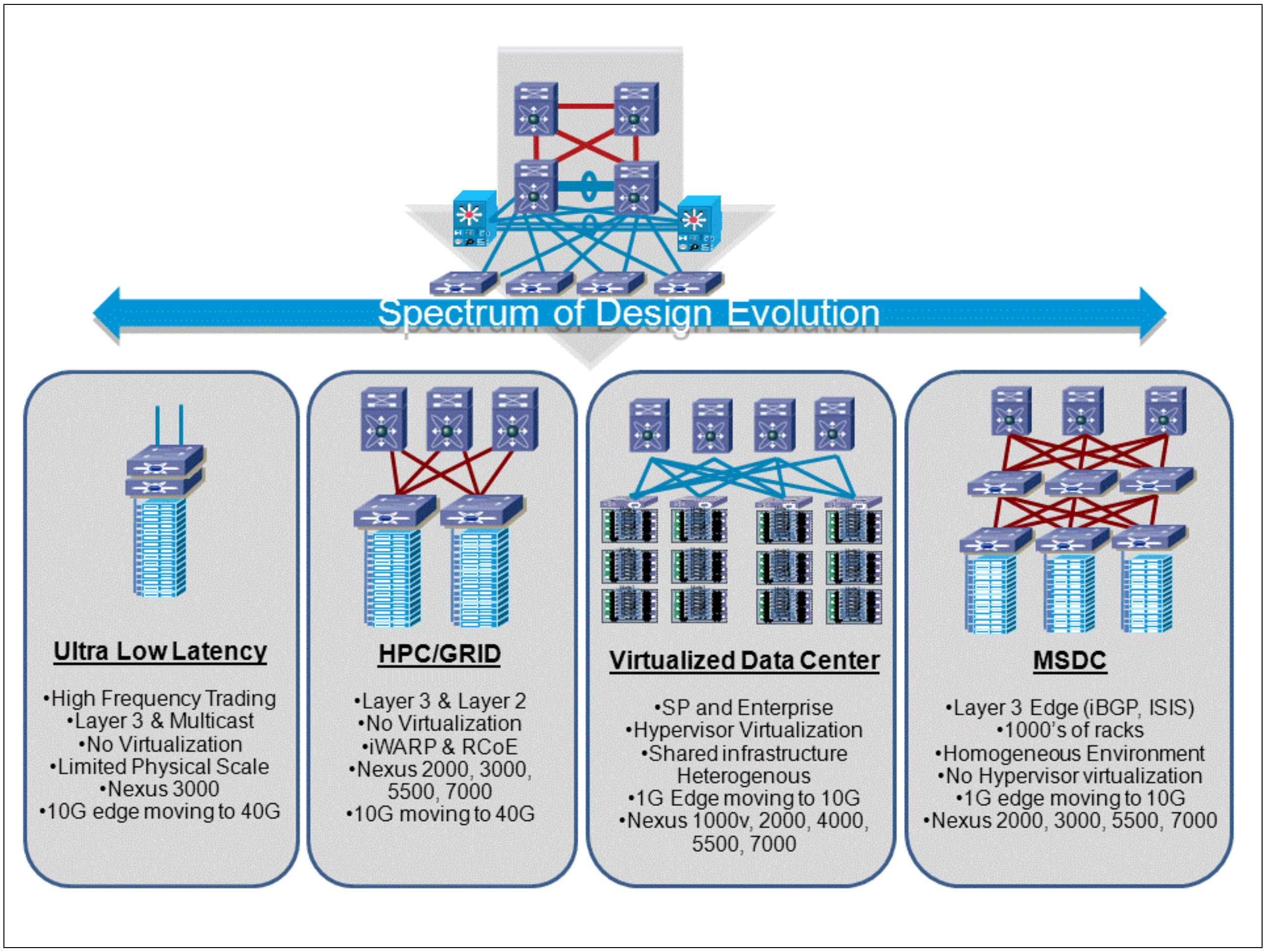


Figure 1-10 Spectrum of data center types

This book mainly focuses on the enterprise virtualized data centers, even if the lines between different models tend to blur.

2

IBM and Cisco building blocks for the data center

The way the world works is changing; demand on the IT infrastructure is increasing as consumers are looking for instant answers and ubiquitous access from smart devices. They are responsible for and driving billions of conversations on global social networks. Businesses selling goods and services to consumers are utilizing newly available information and instantaneous communication capabilities to more effectively market and today, organizations of all sizes can participate in the data-centric revolution, even while struggling with rising costs and limited IT resources.

Cloud computing, green IT, and energy efficient data centers require an IT infrastructure that is a perfect match to each other. Additionally, in this smarter computing era for business and IT, forward-thinking companies are considering more than server performance, existing skills and ease of management when choosing a platform for new application workloads. Some organizations are building new data centers and others are consolidating to participate in this revolution or to adopt these new trends.

On the other hand, traditional network architectures have been useful in providing reasonable availability for static applications, wherein the latest technologies such as virtualization and distributed applications require greater bandwidth scalability to support any type of communication with high quality and that are economically attractive.

In this chapter, we introduce you to some of the major building blocks from IBM and CISCO for next generation data centers. Whether it is consolidation or building a new data center, these new generation IT infrastructure which are more reliable and energy efficient perfectly match these new trends.

2.1 IBM systems

Across the globe, IT support organizations are playing a key role in shaping the way organization functions, however IT support capabilities are not keeping pace with today's immense demands on the IT infrastructure. Our focus in this topic is to introduce you to a new breed of IBM systems along with the IBM proven, robust yet innovative Systems and Storage family members, which can address such issues and help organizations to build next generation data centers.

2.2 IBM PureSystems

IBM PureSystems™ are the newest addition to the IBM diverse list of systems and solutions. They combine the flexibility of a general purpose system, the elasticity of cloud and the simplicity of an appliance. Integrated by design and built in expertise gained from decades of experience delivers a simplified IT experience, freeing IT to play a major role in true real business innovation.In other words, IBM PureSystems are the major building blocks of the next generation data centers. In this section, we briefly introduce you to the following types of IBM PureSystems and IBM Flex System™ Networking:

- IBM PureFlex™ System
- IBM PureApplication™ System
- IBM PureData™ System
- IBM Flex System Networking

Tip: The Flex System family of compute, network, storage (internal and external), and systems management products can be ordered as either a fully integrated PureFlex system or a customized build-to-order (BTO) Flex System configuration. However, both solutions leverage the common Flex System architecture. The significant difference between is being additional flexibility of component level choice while configuring BTO. Although PureFlex comes integrated with IBM virtual fabric technologies, you can order Flex System with your choice of Ethernet switches and pass-thru I/O modules, which provides a seamless integration into the smarter data center infrastructure.

2.2.1 IBM PureFlex System

The PureFlex System is built from no-comprise building blocks based on reliable IBM technology that supports open standards and offers clear and consistent road-maps. PureFlex System is designed for multiple generations of technology, supporting different kinds of workload today and ready for the future demands of businesses. With built-in patterns of expertise for deployment, management and optimization including cloud capabilities to simplify key tasks across all of IT resources, the system is designed to deploy quickly and be easy to manage.

IT organizations can select the capabilities based on target workload and environment. Each configuration is available with choice of IBM POWER7®-based or Intel processor-based compute nodes. Organizations can choose from the following variants of configurations of IBM PureFlex system to perfectly match with their business needs.

Express

The Express configuration is designed for small and medium businesses and is the most affordable entry point into a PureFlex System.

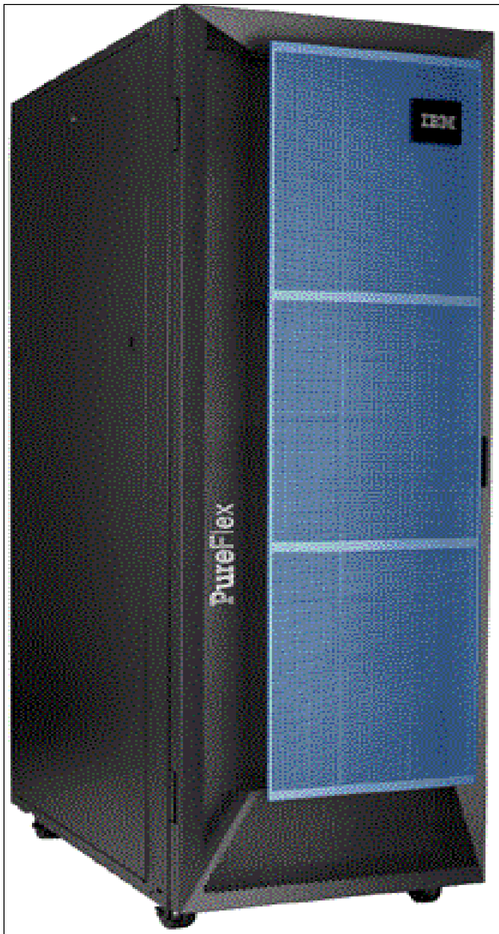
Standard

The Standard configuration is optimized for application servers with supporting storage and networking and is designed to support the key ISV solutions.

Enterprise

The Enterprise configuration is optimized for scalable cloud deployments and has built-in redundancy for highly reliable and resilient operation to support the critical applications and cloud services.

Figure 2-1 shows the IBM PureFlex System with its expertise.



Simplicity:

Organizations need less complex environments. Patterns of expertise help in easy consolidation of diverse servers, storage, and applications onto an easy-to-manage, integrated system.

Efficiency:

To reduce costs and conserve valuable resources, the IT staff needs to get the most out of from systems with energy efficiency; simple management and fast, automated response to problems. Built-in expertise optimizes the critical business applications to get the most out of the investments.

Control:

Optimized patterns of expertise accelerate cloud implementations to lower risk by improving security and reducing human error.

Agility:

As companies look to innovate to bring products and services to market faster, they need fast time-to-value. Expertise built into a solution eliminates manual steps, automates delivery, and supports innovation.

Figure 2-1 IBM PureFlex System and expertise

Table 2-1 explains the IBM PureFlex enterprise system's features and respective benefits.

Table 2-1 IBM PureFlex Enterprise system feature and benefits

Feature	Benefits
Enterprise configuration	► Optimized for scalable cloud deployments ► Built-in redundancy for highly reliable and resilient operation to support critical applications and cloud services.
Deeply integrated compute, storage, and networking resources	► Integration by design ► Deploy in hours instead of days ► Accelerate time to value
Automated management and deployment expertise for physical and virtual resources	► So that experts can focus on high value business tasks ► Automate repetitive tasks

2.2.2 IBM Flex System Manager

IBM Flex System Manager™ is designed to get the most out of an IBM PureFlex System while automating repetitive tasks. Flex System Manager can significantly reduce the number of manual navigational steps for typical management tasks. From simplified system set-up procedures with wizards and built-in expertise to consolidated monitoring for all of an organization's physical and virtual resources, compute, storage, and networking Flex System Manager provides core management functionality along with automation so that the IT team can focus on business innovation. From a single user interface, IT support teams will get:

- Intelligent automation
- Resource pooling
- Improved resource utilization
- Complete management integration
- Simplified setup

For more information on IBM PureSystems, see this website

http://www.ibm.biz/Bdx2eJ

2.2.3 IBM Flex System nodes

IBM Flex System is the element that makes up the IBM PureFlex System, and it allows organizations to build their own system to meet unique IT requirements with a set of no-compromise components including compute, storage, networking and systems management and ideally suited for data center environments that require flexible, cost-effective, secure, and energy-efficient hardware.

The innovative design features of the IBM Flex System products make it possible to configure totally integrated, customized, secure solutions that meet any kind of data center needs of and provides flexible expansion capabilities for future needs. The scalable hardware features and the unprecedented cooling capabilities of the IBM Flex System products helps in optimized hardware utilization, minimizes cost, and simplifies the overall management of the data center.

Figure 2-2 shows the IBM Flex chassis front and rear view.

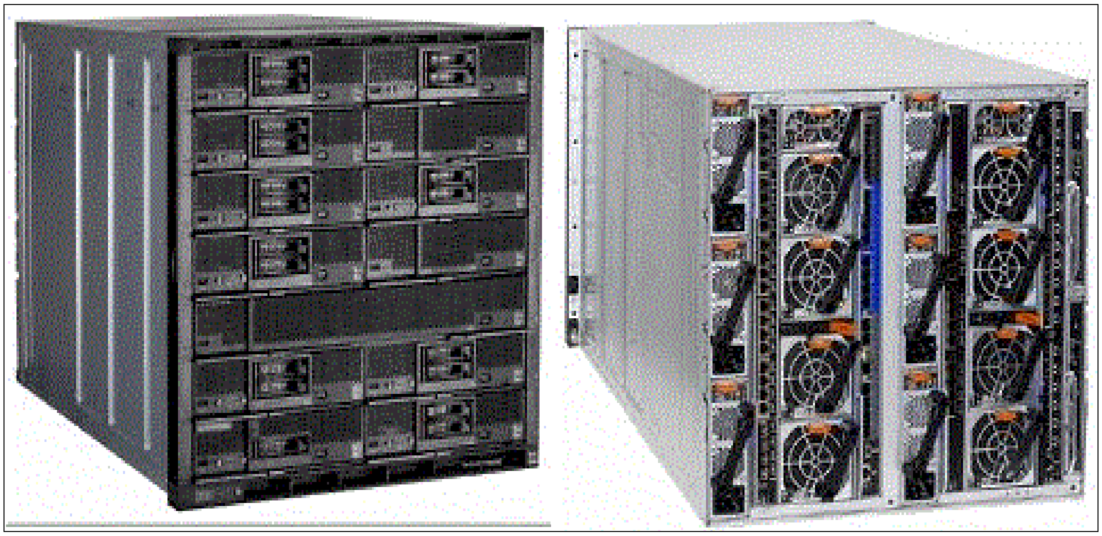


Figure 2-2 IBM Flex chassis front and rear view

2.2.4 IBM PureApplication System

Users of IT systems now consume and create their own functionality. This behavior requires significantly more elasticity on the part of the IT infrastructure and drives an enormous acceleration in the iteration of solutions that lie far beyond traditional demand cycles. Multiple approaches exist today as IT and business leaders shift to smarter computing. As IT teams take on the risk of pioneering new approaches within industry, they still need to know that any services will be delivered as promised. This typically requires expertise that goes beyond the experience that resides within the organization.

IBM PureApplication platform systems include middleware and expertise for deploying and managing application platforms and is designed and tuned specifically for transactional web and database applications. This workload aware, flexible platform is designed to be easy to deploy, customize, safeguard and manage. Whether in a traditional or private cloud environment, this IBM solution can provide businesses with superior IT economics. With the PureApplication System, organizations can provision its own patterns of software, middleware and virtual system resources. Provisioning these patterns are within an innovative framework that is shaped by IT best practices and industry standards that have been culled from many years of IBM experience with its clients and from a deep understanding of smarter computing. These IT best practices and standards are infused throughout the system.

2.2.5 IBM PureData System

As today's Big Data challenges increase, the demands on data systems have never been greater. PureData System, the newest member of the PureSystems family, is optimized exclusively for delivering data services to today’s demanding applications with simplicity, speed and lower cost.

PureData System is optimized exclusively for delivering data services to today's demanding applications. Like each of the IBM PureSystems, it offers built-in expertise, integration by design, and a simplified experience throughout its lifecycle.

Figure 2-3 shows the IBM PureData System along with its key features.

Built-in expertise

Codified data management best practices are provided for each workload. The PureData System delivers automated pattern-based deployment and management of highly reliable and scalable database services.

Integration by design

Hardware, storage, and software capabilities are designed and optimized for specific high performance data workloads such as patented data filtering using programmable hardware (FPGAs) for ultrafast execution of analytic queries without the need for indices.

Simplified experience

The PureData System provides single part procurement with no assembly required (ready to load data in hours), open integration with 3rd party software, integrated management console for the entire system, single line of support, integrated system upgrades, and maintenance.

Figure 2-3 IBM PureData System

The new PureData System comes in different models that have been designed, integrated, and optimized to deliver data services to next generation demanding applications with simplicity, speed, and lower cost.

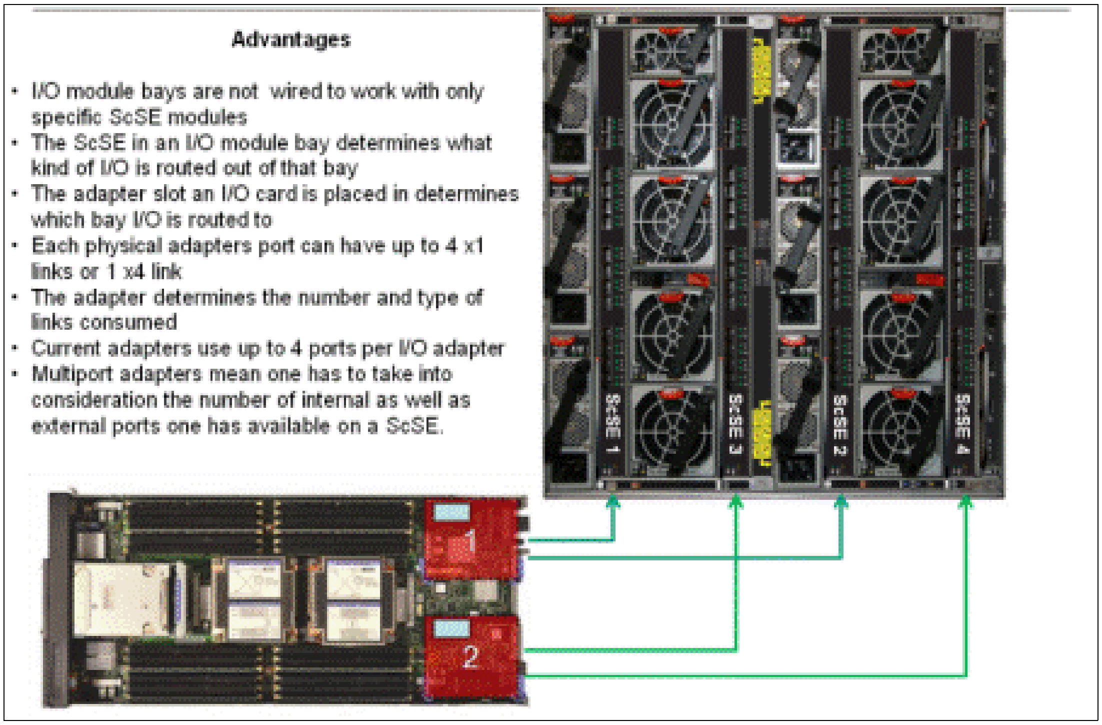
2.2.6 IBM Flex System Networking

Complexity of modern application infrastructures requires networks of low latency at high bandwidth and due to virtualization and the adoption of cloud concepts the physical network infrastructure is merging with logical/virtual networking environment. IBM Flex System Networking portfolio modules are built using industry standards and enterprise class features thereby offers the seamless interoperation with existing infrastructures while also providing capabilities for the future.

IBM® favors a standards-based approach to networking and works closely with its partners, the IEEE, and other leading standards organizations. This approach complements the Open Choice design of the IBM Flex System™ and PureFlex™ System, which gives customers the ability to configure their system with a wide variety of elements from IBM and its partners such as Cisco.

The Ethernet networking I/O architecture for the IBM Flex System Enterprise Chassis includes an array of connectivity options for server nodes installed in the enclosure. Users can decide to use a local switching model that provides superior performance, cable reduction and a rich feature set, or use pass-thru technology and allow all Ethernet networking decisions to be made external to the Enterprise Chassis. By far, the most versatile option is to use modules that provide local switching capabilities and advanced features that are fully integrated into the operation and management of the Enterprise Chassis.

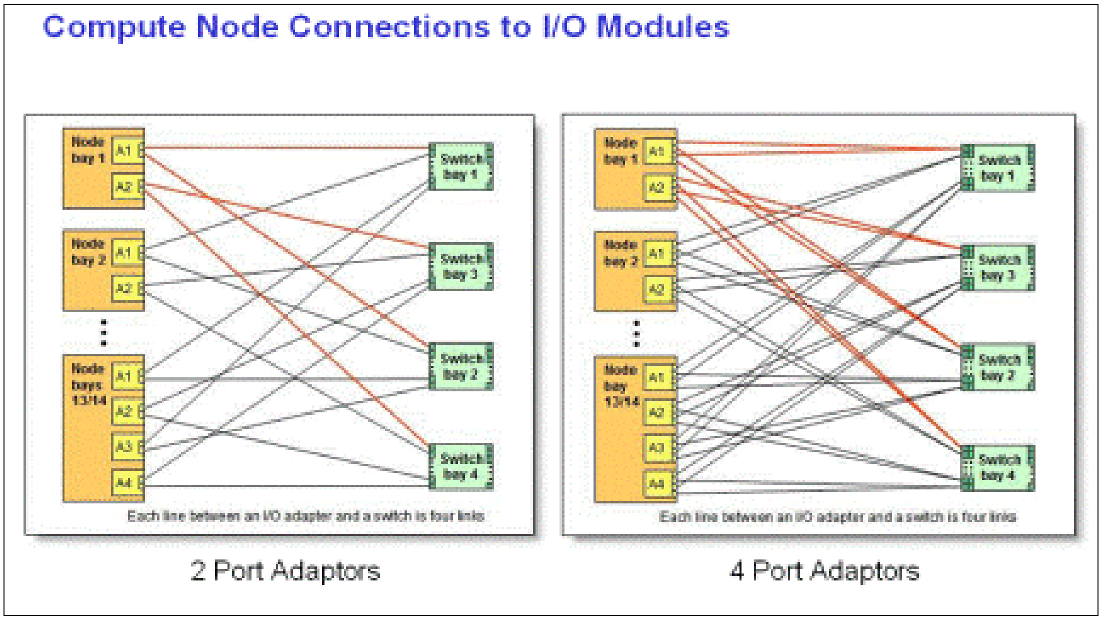
Figure 2-4 and Figure 2-5 show the IBM Flex System enterprise chassis physical I/O layout and compute node connections to the I/O module.



Advantages

- I/O module bays are not wired to work with only specific ScSE modules
- The ScSE in an I/O module bay determines what kind of I/O is routed out of that bay
- The adapter slot an I/O card is placed in determines which bay I/O is routed to
- Each physical adapters port can have up to 4 x1 links or 1 x4 link
- The adapter determines the number and type of links consumed
- Current adapters use up to 4 ports per I/O adapter
- Multiport adapters mean one has to take into consideration the number of internal as well as external ports one has available on a ScSE.

Figure 2-4   I/O Module physical lay out with advantages



Compute Node Connections to I/O Modules

Each line between an I/O adapter and a switch is four links

Each line between an I/O adapter and a switch is four links

2 Port Adaptors

4 Port Adaptors

Figure 2-5   Compute node connections to I/O modules

In the following section, we discuss some of the IBM Flex Systems high performance Ethernet and converged networking switches and adapters.

2.2.7 IBM Flex System Fabric EN4093/EN4093R 10 Gb Scalable Switch

IT departments always face challenges in preparing for the future. Some of those key areas are around virtualization and reducing cost and complexity, while automating things like VM mobility and in parallel these departments are looking for technologies that are more tightly integrated, while providing investment protection for the transition to cloud computing.

The IBM Flex System Fabric EN4093 and EN4093R 10 Gb Scalable Switches provide unmatched scalability, interoperability, and performance, while also delivering innovations to help address a number of networking concerns today and providing capabilities that will help you prepare for the future. IBM designed these switches based on industry standards, for seamless interoperability into Cisco and other networks.

These switches are capable of supporting up to sixty-four 10 Gb Ethernet connections while offering Layer 2/3 switching. They are designed to install within the I/O module bays of the IBM Flex System Enterprise Chassis. These switches can help in migrating to a 10 Gb or 40 Gb Ethernet infrastructure and offer virtualization features like Virtual Fabric and IBM VMready®, plus the ability to work with IBM Distributed Virtual Switch 5000V.

Figure 2-6 shows the IBM Flex System Fabric EN4093/EN4093R 10 Gb Scalable Switch.

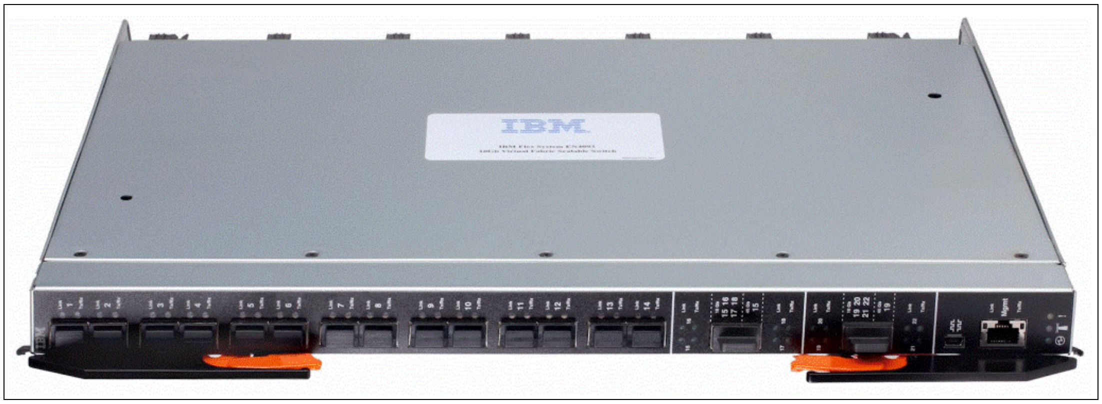


Figure 2-6 IBM Flex System Fabric EN4093/EN4093R 10 Gb Scalable Switch

The Flex System Fabric EN4093R simplifies deployment and growth by using its innovative scalable architecture. The extreme flexibility of the switch comes from the ability to turn on additional ports as required, both down to the server and for upstream connections (including 40bE). LAN, and SAN, the Flex System Fabric EN4093R supports the newest protocols, including Data Center Bridging/Converged Enhanced Ethernet (DCB/CEE) that you can use in either an iSCSI, Fibre Channel over Ethernet (FCoE), or NAS type converged environment.

Additionally, network management is simplified by features like network convergence (FCoE) and stacking, where multiple EN4093R switch modules can be managed as one logical switch reducing system complexity.In addition, the EN4093R’s stacking capabilities simplify management for clients by stacking up to eight switches that share one IP address and one management interface together. Plus with Switch Partition (SPAR), clients can virtualize the switch with partitions that isolate communications for multi-tenancy environments.

The IBM Flex System Fabric EN4093/EN4093R has the following key performance features:

- 100% line rate performance
- 1.28 Tbps non-blocking switching throughput (full duplex)
- 960 Mpps
- Stackable FCoE Transit Switch capability

Table 2-2 provides a quick view of hardware summary of IBM Flex System Fabric EN4093R 10 Gb Scalable Switch.

Table 2-2 IBM Flex System Fabric EN4093R 10 Gb Scalable Switch's hardware summary

Models and upgrades	Stacking	Power and heat dissipation
Base configuration ► 14 x 10 Gb server ports & 10 x 10 Gb uplinks Upgrade options (must be done in order): ► Upgrade 1 Total: 28* x 10 Gb server ports & 10 x 10 Gb plus 2 x 40 Gb uplinks (*Requires 4-port adapter) ► Upgrade 2 Total: 42**x 10 Gb server ports & 14 x 10 Gb plus 2 x 40 Gb uplinks (**Requires adapter able to support 6-ports like CN4058)	Stacking available on either 40 Gb and 10 Gb server ports	Typical power consumption of 95 Watts and 1127 BTU/hour (typical)

Note: For details on the Flex System Fabric EN4093R comprehensive software feature list, see the IBM Networking Operating System datasheet:

```
http://www-03.ibm.com/systems/networking/software/networkingos/index.html
```

For details about IBM Flex system switch interoperability with Cisco, read the Tolly Interoperability Report# 212134 for IBM Flex System with Cisco at:

```
http://www.tolly.com/DocDetail.aspx?DocNumber=212134
```

or:

```
https://ibm.biz/BdxpNj
```

2.2.8 IBM Flex System EN2092 1 Gb Ethernet Scalable Switch

The IBM Flex System™ EN2092 1 Gb Ethernet Scalable switch provides outstanding flexibility allowing you to buy one switch today and enhance its functionality in the future. The scalable architecture allows you to support 2-port or 4-port 1 Gb adapters with this switch. The EN2092 supports up to 28 server ports and up to twenty 1 Gb uplinks and four 10 Gb uplinks. Designed with top performance in mind, the EN2092 provides high availability with legendary IBM quality and switch failover capability. This switch is an exceptionally flexible integrated switch with extreme scalability and performance, while also delivering best-in-class networking innovations to help you address today's networking requirements. It also provides advanced capabilities to address future needs.

Figure 2-7 shows the IBM Flex System™ EN2092 1 Gb Ethernet Scalable Switch.

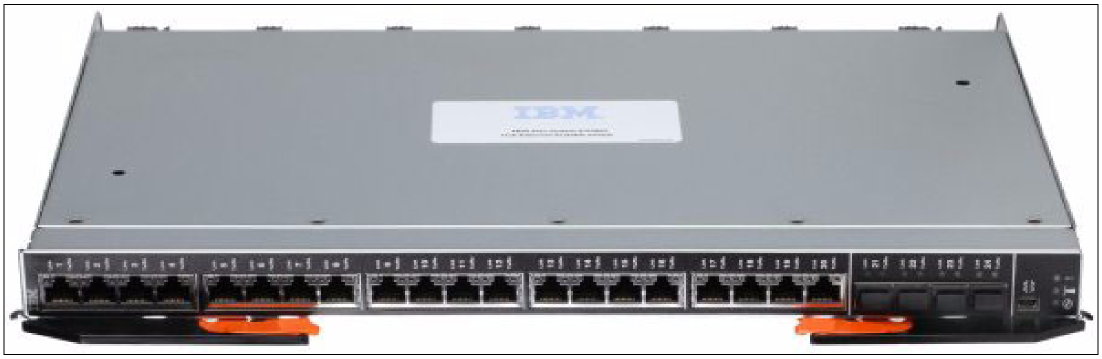


Figure 2-7 IBM Flex System EN2092 1 Gb Ethernet Scalable Switch

Simplified scalability

The Flex System EN2092 simplifies deployment and growth via its unique scalable architecture, which helps increase ROI for today's needs, while providing investment protection with easy enhancements that add additional port capacity and bandwidth for today's and tomorrow's data centers. The Flex System EN2092's exceptional flexibility enables you to turn on additional 1 Gb ports per server (up to two per switch) and add 1 Gb or 10 GbE uplink bandwidth for upstream connections. You no longer need to purchase new switches or go through the process of recertifying or testing; you can simply purchase a license key to activate the additional ports.

Unified network management

Aligning the system network with the requirements of Big Data, cloud and optimized workloads is a key challenge for IT organizations. The EN2092 helps address this problem with its integrated network management solution which allows net-work administrators to manage servers, storage, and networks together as one logical unit. The EN2092 is tightly integrated and managed via the Flex System Manager, which allows you to simplify network management via port profiles, topology views and virtualization management.

For more advanced levels of management and control, IBM System Networking Element Manager (SNEM) can significantly reduce deployment and daily maintenance requirements, while providing in-depth visibility into network performance and operations of the IBM network. Together with tools including IBM System Director and VMware Virtual Center or vSphere clients, IBM System Networking Element Manager provides additional integration for better optimization.

Optimized and automated network virtualization

With the growth of virtualization and emerging cloud-based applications, demand for intelligence and performance at the edge of the network is more important than ever. Many organizations are looking for high performance and awareness of virtual machines (VMs) at the network level. Using the IBM innovative VMready virtualization-aware networking technology, you can simplify virtualization management.

VMready automatically discovers virtual machines and allows unique networking parameters to be configured for each virtual machine and if a virtual machine migrates, it is tracked by VMready and its network configurations are migrated in real-time using VMready NMotion® technology to protect the virtual machine, wherever it moves. VMready works with all major hypervisors. You can further optimize and automate virtualization management using the new IBM Distributed Virtual Switch 5000V. The IBM System Networking Distributed Virtual Switch 5000V is an advanced, feature-rich distributed virtual switch for VMware environments with policy-based VM connectivity that enables network administrators to manage and provision network settings at the virtual machine level.

Increased network performance

For customers standardized on 1 Gb networks, this switch offers ample room for future growth. You can start with 10 Gb uplink bandwidth and grow up to 60 Gb uplink bandwidth. In addition, the switch also provides a rich, robust and mature set of Layer 2/3 features, making it ideal for virtually any data center network topology or configuration requirement.

Table 2-3 shows a summary of the EN2092 models and power.

Table 2-3 IBM Flex System Fabric EN2092 1 Gb Scalable Switch

Models and upgrades	Power and heat dissipation
IBM Flex System EN2092 1 Gb Ethernet Scalable Switch (upgrades not required in order; 1 Gb ports are RJ45 connectivity): ► Base configuration (49Y4294) 14 × 1 Gb server port and 10 × 1 Gb uplinks ► Switch upgrade 1 (90Y3562) Additional 14 × 1 Gb server ports and 10 × 1 Gb uplinks ► Switch upgrade 2 (49Y4298) Enables 4 × 10 Gb uplinks	Typical power consumption of 40-52 Watts and 177.4 BTU/hour (typical)

For more information, see these websites:

http://www.ibm.com/systems/flex/networking/bto/ethernet/en2092_1gb_e/index.html

http://www.redbooks.ibm.com/abstracts/tips0861.html?Open

2.2.9 IBM Flex System EN4091 10 Gb Ethernet Pass-thru Module

The IBM Flex System™ EN4091 10 Gb Ethernet Pass-thru Module offers easy connectivity of the IBM Flex System Enterprise Chassis to any external network infrastructure. This unmanaged device enables direct Ethernet connectivity of the compute node in the chassis to an external Top-of-Rack data center switch. This module can function at both 1 Gb and 10 Gb Ethernet speeds. It has fourteen internal 1 Gb or 10 Gb links, and fourteen external 1 Gb or 10 Gb SFP+ uplinks.

Figure 2-8 shows the IBM Flex System™ EN4091 10 Gb Ethernet Pass-thru Module.

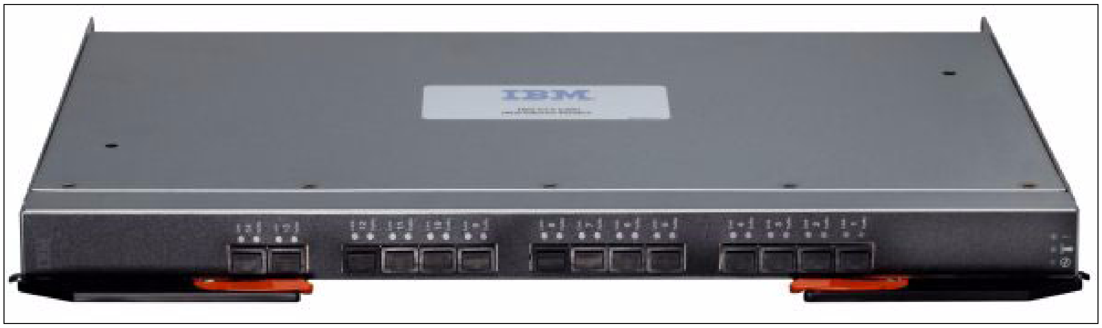


Figure 2-8 IBM Flex System™ EN4091 10 Gb Ethernet Pass-thru Module

With the flexibility of the IBM Flex System EN4091 10 Gb Ethernet Pass-thru Module, clients can take advantage of the technologies that they require for multiple environments:

- For 1 GbE links, you can use SFP transceivers plus RJ-45 cables or LC-to-LC fiber cables, depending on the transceiver.
- For 10 GbE, you can use direct-attached copper (DAC, also known as Twinax) cables, which come in lengths between 1 m and 5 m. These DAC cables are a cost-effective and low-power alternative to transceivers, and are ideal for all 10 Gb Ethernet connectivity to the Top-of-the-Rack (TOR) switches. For longer distances, you can use SFP+ transceivers (SR or LR) plus LC-to-LC fiber optic cables.

Benefits

The IBM Flex System EN4091 10 Gb Ethernet Pass-thru Module is particularly suited for the following clients:

- Clients who require direct connectivity of the compute nodes in the chassis to an external TOR data center switch
- Clients who are implementing a converged environment
- Clients who want to reduce total cost of ownership (TCO) and improve performance, while maintaining high levels of availability and security
- Clients who want to avoid oversubscription, which can result in congestion and loss of performance

Features and specifications

The IBM Flex System EN4091 10 Gb Ethernet Pass-thru Module has the following features and specifications:

- Internal ports: Fourteen internal full-duplex Ethernet ports that can operate at 1 Gb or 10 Gb speeds.
- External ports: Fourteen ports for 1 Gb or 10 Gb Ethernet SFP+ transceivers (support for 1000BASE-SX, 1000BASE-LX, 1000BASE-T, 10GBASE-SR, or 10GBASE-LR) or SFP+ copper direct-attach cables (DAC). SFP+ modules and DAC cables are not included and must be purchased separately.
- An unmanaged device that has no internal Ethernet management port, but is able to provide its vital product data (VPD) to the secure management network in the Chassis Management Module.

For more information, see these websites:

http://www.ibm.com/systems/flex/networking/bto/ethernet/en4091_10gb_e/index.html

http://www.redbooks.ibm.com/abstracts/tips0865.html?Open

2.2.10 IBM Flex System Fabric CN4093 10 Gb Converged Scalable Switch

The IBM Flex System Fabric CN4093 10 Gb Converged Scalable Switch provides unmatched scalability, performance, convergence, and network virtualization. The switch offers full Layer 2/3 switching as well as FCoE Full Fabric and Fibre Channel NPV Gateway operations to deliver a truly converged integrated solution, and it is designed to install within the I/O module bays of the IBM Flex System Enterprise Chassis.

The switch can help in migrating to a 10 Gb or 40 Gb converged Ethernet infrastructure and offers virtualization features like Virtual Fabric and VMready, plus the ability to work with IBM Distributed Virtual Switch 5000V. Also, it offers up to 12 external programmable IBM Omni Ports, which provide extreme flexibility with the choice of SFP+ based 10 Gb Ethernet connectivity or 4/8 Gb Fibre Channel connectivity, depending on the SFP+ module used. The CN4093 is also designed, like other IBM Flex system switches, for standards-based interoperability into Cisco and other vendor networks.

Figure 2-9 shows the IBM Flex System Fabric CN4093 10 Gb Converged Scalable Switch.

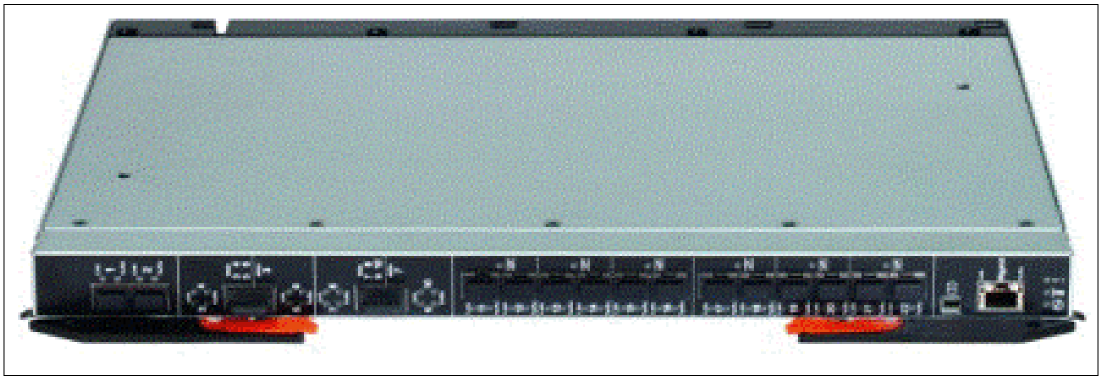


Figure 2-9 IBM Flex System Fabric CN4093 10 Gb Converged Scalable Switch

In addition, Logical Network Profiles (part of Flex System Manager) delivers advanced virtualization that simplifies management and automates VM mobility. IBM VMready makes the network VM aware and works with all major hyper-visors. For clients using VMware, the IBM System Networking Distributed Virtual Switch 5000V enables network administrators to simplify virtual switch management by having a consistent virtual and physical networking environment.Also, the IBM Flex System Fabric CN4093 is tightly integrated and managed through the Flex System Manager. This provides an additional option to manage the switch directly through CLI commands. Network management can be simplified using port profiles, topology views, and virtualization management.

For more advanced levels of management and control, IBM offers IBM System Networking Element Manager (SNEM) which can significantly reduce deployment and day to day maintenance times, while providing in-depth visibility into the network performance and operations of IBM switches. Plus, SNEM provides additional integration when leveraging tools like VMware Virtual Center or vSphere. The IBM Flex System Fabric CN4093 10 Gb has the following key performance features:

- 100% line rate performance
- 1.28 Tbps non-blocking switching throughput (full duplex)

Table 2-4 shows a summary of the IBM Flex System Fabric CN4093 10 Gb hardware.

Table 2-4 IBM Flex System Fabric CN4093 hardware summary

Models	Interface option	Power
► Base configuration 00D5823/ESW2 14 x 10 Gb downlink ports, plus 2 SFP+ and 6 OmniPorts uplinks Upgrade options (supported in any order): ► Upgrade 1 - 00D5845/ESU1 Turns on additional 14 x 10 Gb downlink ports and 2 x 40 Gb uplinks. ► Upgrade 2 - 00D5487/ESU2 Turns on additional 14 x 10 Gb downlink ports and six OmniPorts)	2 SFP+ (1/10 GbE), 2 QSFP+ (40 Gb or break-out to 4x10 Gb) and 12 OmniPorts (10 GbE, 4/8 Gb FC)	Typical power consumption of 141 Watts

For more information, see this website:

http://www.ibm.com/systems/networking/software/networkingos/index.html

2.2.11 IBM Flex System CN4054 10 Gb Virtual Fabric Adapter

The IBM Flex System CN4054 10 Gb Virtual Fabric Adapter is a 4-port 10 Gb converged network adapter (CNA) for Intel processor-based compute nodes that can scale to up to 16 virtual ports and that supports Ethernet, iSCSI, and FCoE. The adapter supports up to eight virtual NIC (vNIC) devices, where each physical 10 GbE port can be divided into four virtual ports with flexible bandwidth allocation. The CN4054 Virtual Fabric Adapter Upgrade adds FCoE and iSCSI hardware initiator functionality.

The IBM Flex System EN4054 4-port 10 Gb Ethernet Adapter is based on similar hardware but simply operates as a straight 4-port 10 GbE adapter. The EN4054 it is supported by POWER7 processor-based compute nodes.

CN4054 are based on industry-standard PCIe architecture and offers the flexibility to operate as a Virtual NIC Fabric Adapter or as a quad-port 10 Gb or 1 Gb Ethernet device. Because this adapter supports up to 16 virtual NICs on a single quad-port Ethernet adapter, organization can see the benefits in cost, power/cooling, and data center footprint by deploying less hardware. Compute nodes like the x240 support up to two of these adapters for a total of 32 virtual NICs per system.

Both the IBM Flex System CN4054 10 Gb Virtual Fabric Adapter and the EN4054 4-port 10 Gb Ethernet Adapter have following features:

- Four-port 10 Gb Ethernet adapter
- Dual-ASIC Emulex BladeEngine 3 (BE3) controller
- Connection to either 1 Gb or 10 Gb data center infrastructure (1 Gb and 10 Gb auto-negotiation)
- PCI Express 3.0 x8 host interface (The p260 and p460 support PCI Express 2.0 x8.)
- Full-duplex (FDX) capability
- Bus-mastering support
- Direct memory access (DMA) support
- Preboot Execution Environment (PXE) support
- IPv4/IPv6 TCP, UDP checksum offload:
  - Large send offload (LSO)
  - Large receive offload
  - Receive side scaling (RSS)
  - IPv4 TCP Chimney Offload
  - TCP Segmentation Offload
- VLAN insertion and extraction
- Jumbo frames up to 9000 bytes
- Load balancing and failover support, including adapter fault tolerance (AFT), switch fault tolerance (SFT), adaptive load balancing (ALB), teaming support, and IEEE 802.3ad
- Enhanced Ethernet (draft):
  - Enhanced Transmission Selection (ETS) (P802.1Qaz)
  - Priority-based Flow Control (PFC) (P802.1Qbb)
  - Data Center Bridging Capabilities eXchange Protocol, CIN-DCBX and CEE-DCBX (P802.1Qaz)
- Supports Serial over LAN (SoL)
- Total Max Power: 23.1 W

The CN4054 10 Gb Virtual Fabric Adapter also supports the following additional features:

- Operates either as a 4-port 1/10 Gb Ethernet adapter or supports up to 16 vNICs.
- In virtual NIC (vNIC) mode, it supports:
  - Virtual port bandwidth allocation in 100 Mbps increments.
  - Up to 16 virtual ports per adapter (four per port).
  - With the CN4054 Virtual Fabric Adapter Upgrade, 90Y3558, four of the 16 vNICs (one per port) support iSCSI or FCoE.
- Supports for two vNIC modes: IBM Virtual Fabric Mode and Switch Independent Mode.
- Wake On LAN support.
- With the CN4054 Virtual Fabric Adapter Upgrade, 90Y3558, the adapter adds FCoE and iSCSI hardware initiator support:
  - iSCSI support is implemented as a full offload.
  - It presents an iSCSI adapter to the operating system.
- TCP Offload Engine (TOE) support with Windows Server 2003, 2008, and 2008 R2 (TCP Chimney) and Linux:
  - Connection and its state are passed to the TCP offload engine.
  - Data transmit and receive are handled by adapter.
  - It is supported with iSCSI.

2.2.12 IBM Flex System EN4132 2-port 10 Gb Ethernet Adapter

The IBM Flex System EN4132 2-port 10 Gb Ethernet Adapter delivers high-bandwidth and industry-leading Ethernet connectivity for performance-driven server and storage applications in enterprise data centers, high-performance computing (HPC), and embedded environments. Clustered databases, web infrastructure, and high frequency trading are just a few applications that achieve significant throughput and latency improvements, resulting in faster access, real-time response, and more users per server.

Based on Mellanox ConnectX-3 EN technology, this adapter improves network performance by increasing available bandwidth while decreasing the associated transport load on the processor, also Mellanox networking adapters deliver industry-leading bandwidth with ultra low, sub-microsecond latency for performance-driven server clustering applications. Combined with the IBM Flex System Fabric EN4093 10 Gb Scalable Switch, IT organization can achieve efficient computing by off-loading the processor protocol processing and data movement overhead, such as RDMA and Send/Receive semantics, allowing more processor power for the application.

The IBM Flex System EN4132 2-port 10 Gb Ethernet Adapter has the following features:

RDMA over converged Ethernet

ConnectX-3 uses IBTA RoCE technology to provide efficient RDMA services, delivering low-latency and high-performance to bandwidth and latency sensitive applications. With link-level interoperability in existing Ethernet infrastructure, network administrators can use existing data center fabric management solutions.

Sockets acceleration

Applications using TCP/UDP/IP transport can achieve industry-leading throughput over 10 GbE. The hardware-based stateless off-load and flow steering engines in ConnectX-3 reduce the processor overhead of IP packet transport, freeing more processor cycles to work on the application. Sockets acceleration software further increases performance for latency sensitive applications.

I/O virtualization

ConnectX-3 EN provides dedicated adapter resources and guaranteed isolation and protection for virtual machines within the server. ConnectX-3 EN gives data center managers better server utilization and LAN and SAN unification while reducing costs, power, and complexity.

Precision data centers

ConnectX-3 EN IEEE 1588 precision time protocol circuitry synchronizes the host clock to the data center master clock for accurate data delivery time stamping and data center SLA measurements. The hardware-based mechanisms ensure high accuracy and low jitter.

Storage acceleration

A consolidated compute and storage network achieves significant cost-performance advantages over multifabric networks. Standard block and file access protocols can use RDMA for high-performance storage access.

Quality of Service

Resource allocation per application or per VM is provided and protected by the advanced QoS supported by ConnectX-3 EN. Service levels for multiple traffic types can be based on IETF DiffServ or IEEE 802.1p/Q, allowing system administrators to prioritize traffic by application, virtual machine, or protocol. This powerful combination of QoS and prioritization provides fine-grained control of traffic, ensuring that applications run smoothly.

2.2.13 IBM Flex System EN2024 4-port 1 Gb Ethernet Adapter

The IBM Flex System EN2024 4-port 1 Gb Ethernet Adapter is a quad-port Gigabit Ethernet network adapter. When it is combined with the IBM Flex System EN2092 1 Gb Ethernet Scalable Switch, organization can leverage an end-to-end 1 Gb solution on the IBM Flex System Enterprise Chassis. The EN2024 adapter is based on the Broadcom 5718 controller and offers a PCIe 2.0 x1 host interface with MSI/MSI-X. It also supports I/O virtualization features like VMware NetQueue and Microsoft VMQ technologies. This EN2024 is based on the industry-standard PCIe architecture and is ideal for using Gigabit Ethernet in network infrastructure. The IBM Flex System compute nodes support up to two of these adapters for a total of eight NICs per system.

The IBM Flex System EN2024 4-port 1 Gb Ethernet Adapter has the following features:

- Dual Broadcom BCM5718 ASICs
- Quad-port Gigabit 1000BASE-X interface
- Two PCI Express 2.0 x1 host interfaces, one per ASIC
- Full-duplex (FDX) capability, enabling simultaneous transmission and reception of data on the Ethernet network
- MSI and MSI-X capabilities, up to 17 MSI-X vectors
- I/O virtualization support for VMware NetQueue, and Microsoft VMQ
- Seventeen receive queues and 16 transmit queues
- Seventeen MSI-X vectors supporting per-queue interrupt to host
- Function Level Reset (FLR)
- ECC error detection and correction on internal SRAM
- TCP, IP, and UDP checksum offload
- Large Send offload, TCP segmentation offload
- Receive-side scaling
- Virtual LANs (VLANs): IEEE 802.1q VLAN tagging

- Jumbo frames (9 KB)
- IEEE 802.3x flow control
- Statistic gathering (SNMP MIB II, Ethernet-like MIB [IEEE 802.3x, Clause 30])
- Comprehensive diagnostic and configuration software suite
- ACPI 1.1a-compliant: multiple power modes
- Wake-on-LAN (WOL) support
- Preboot Execution Environment (PXE) support
- RoHS-compliant

2.3 IBM Distributed Virtual Switch 5000V

The IBM System Networking Distributed Virtual Switch 5000V is an advanced, feature-rich distributed virtual switch for VMware environments with policy-based virtual machine (VM) connectivity.

2.3.1 Benefits

The IBM Distributed Virtual Switch (DVS) 5000V enables network administrators familiar with IBM System Networking switches to manage the IBM DVS 5000V just like IBM physical switches using advanced networking, troubleshooting and management features so the virtual switch is no longer hidden and difficult to manage.

Support for Edge Virtual Bridging (EVB) based on the IEEE 802.1Qbg standard enables scalable, flexible management of networking configuration and policy requirements per VM and eliminates many of the networking challenges introduced with server virtualization. The IBM DVS 5000V works with VMware vSphere 5.0 and beyond and interoperates with any 802.1Qbg-compliant physical switch to enable switching of local VM traffic in the hypervisor or in the upstream physical switch.

2.3.2 Highlights

Here we list the highlights of this product:

- Large-scale server virtualization by providing enterprise-level switch functionality in the hypervisor
- Advanced networking features not available through base vSwitch
- Mobility of VM security and network properties
- 802.1Qbg standards-based unified management of VM network policies across the virtual and physical network
- Network administrators can manage and provision network settings at the virtual machine level
- Flexible and scalable to a large number of ports

For more information, see this website:

http://www.ibm.com/systems/networking/switches/virtual/dvs5000v/

2.4 IBM System Networking

IBM System Networking solutions enable organizations to deploy a high-speed, low-latency and unified fabric architecture. Its advanced virtualization awareness reduces the cost and complexity of deploying physical and virtual data center infrastructure. In addition, it provides the simplified management of data center networks by delivering the foundation of consolidated operations for dynamic infrastructure management. Fixed form factor or Top-of-Rack switches, called IBM RackSwitch, share a common operating system with the IBM Flex System and IBM BladeCenter embedded switches. IBM Networking OS enables provides seamless, standards-based interoperability with Cisco networks.

2.4.1 IBM RackSwitch G8264

G8264 is an enterprise-class and full-featured data-center Top-of-Rack switch that delivers line rate, high-bandwidth switching, filtering, and traffic queuing without delaying data and is a key piece when connecting multiple chassis together with Pure System Solutions. The G8264 supports IBM VMready technology, an standards-based solution to manage virtual machines. Also supports the newest protocols including Data Center Bridging/Converged Enhanced Ethernet (DCB/CEE) for support of Fibre Channel over Ethernet (FCoE), in addition to iSCSI and NAS. Is ideal for latency-sensitive applications such as high performance computing clusters and financial applications.

The G8264 supports stacking for up to eight switches for simplified switch management as well as IBM Virtual Fabric to diminish the number of I/O adapters buy creating virtual NICs out of a 10 Gb adapter, helping reduce cost and complexity. In addition, G8264 also offers the benefit of OpenFlow that enables the network administrator to easily configure and manage virtual networks that control traffic on a “per-flow” basis. It creates multiple independent virtual networks and related policies without dealing with the complexities of the underlying physical network and protocols. The redundant power and fans along with numerous high availability features equips the switch for business-sensitive traffic. The highlights are as follows:

- Optimized for applications requiring high bandwidth and low latency
- Supports Virtual Fabric and OpenFlow Up to sixty-four 10 Gb SFP+ ports in a 1U form factor
- Future-proofed with four 40 Gb QSFP+ ports
- Stacking: A single switch image and configuration file can be used for up to eight switches, sharing only one IP address and one management interface

Figure 2-10 shows the IBM RackSwitch™ G8264.

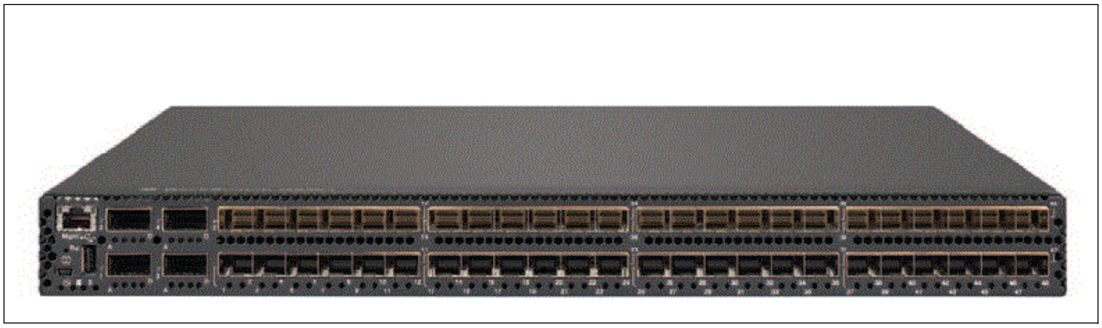


Figure 2-10 IBM RackSwitch G8264

2.4.2 IBM System Networking RackSwitch G8264CS

The IBM System Networking RackSwitch G8264CS is an enterprise-class switch that offers high bandwidth performance with thirty six 1/10 Gb SFP+ connections, twelve Omni Ports that can be used for 10 Gb SFP+ connections, 4/8 Gb Fibre Channel connections or both, plus four 40 Gb QSFP+ connections.

It simplifies the deployment with its innovative IBM Omni Port technology (providing compatibility with any Cisco storage). and offers the flexibility to choose 10 Gb Ethernet, 4/8 Gb Fibre Channel or both for upstream connections and in FC mode, Omni Ports provide convenient access to FC storage.

G8264CS provides 100% line rate performance with low latency and 1.28 Tbps non-blocking switching throughput (full duplex) on Ethernet ports, making it an optimal choice for managing dynamic workloads across the network and provides a rich Layer 2 and Layer 3 feature set that is ideal for many of today’s data centers.

Also, its Omni Port technology helps in consolidating the enterprise storage, networking, data, and management onto a simple to manage single fabric and reduces costs associated with energy and cooling, management and maintenance, and capital costs.

Figure 2-11 shows the IBM System Networking RackSwitch G8264CS.

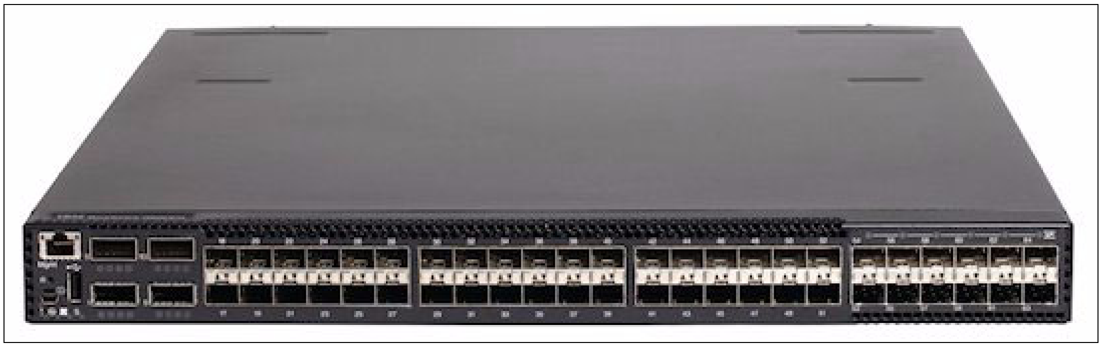


Figure 2-11 IBM System Networking RackSwitch G8264CS

Two models are available, either rear-to-front air flow or front-to-rear air flow:

- Rear-to-front model: Ports are in the rear of the rack and are suitable for use with System x, IBM Power Systems, IBM Flex System™ or IBM PureSystems™, and IBM BladeCenter® designs.
- Front-to-rear model: Ports are in the front of the rack and are suitable for use with IBM iDataPlex® systems.

These key features are included:

- Omni Ports provide flexibility of using 10 Gb Ethernet and/or 4/8 Gb Fibre Channel connections 36 × 10 GbE SFP+ ports, 12 Omni Ports and 4 × 40 GbE QSFP+ ports in a 1U form factor
- G8264CS is optimized for High Performance Computing and other applications requiring high bandwidth and low latency
- Hot-swappable redundant power supplies and fans
- Software is based on Internet standards for optimal interoperability with Cisco or other vendors’ networks
- Two airflow options to manage hot- and cold-aisle applications
- VMready and Virtual Fabric for virtualized networks

For more information, see these websites:

http://www.ibm.com/systems/networking/switches/converged/g8264cs/index.html

http://www.redbooks.ibm.com/abstracts/tips0970.html?Open

2.4.3 IBM RackSwitch G8264T

The IBM RackSwitch G8264T provides flexible connectivity for high-speed server and storage devices. This switch comes with forty-eight 10GBase-T/1000Base-T RJ45 connections plus four 40 Gb QSFP+ connections in a 1U form factor.

The G8264T offers a flexible low-cost connectivity option for customers who want to reuse existing copper cabling while upgrading to 10 Gigabit Ethernet, in environments supporting distances up to 100m. Also, G8264T supports IBM VMready technology, and OpenFlow1; the first protocol designed specifically for software defined networking (SDN), providing high-performance, traffic control, improved automation and centralized management across multiple network devices. The G8264T is ideal for organizations looking to expand into 10 Gb environments without compromising its data center's overall TCO. IBM RackSwitch G8264T has the following key performance features:

- 100% Line rate performance
- 1280 Gbps non-blocking switching throughput (full duplex)
- 3.2 microseconds latency
- 960 Mpps

Table 2-5 shows the G8264T features along with its hardware summary.

Table 2-5 IBM RackSwitch G8264T features and hardware summary

G8264T features	G8264T hardware summary
G8264T is a low cost, flexible connectivity option for high-speed servers and storage devices	Forty-eight 10GBase-T/1000Base-T RJ45 ports, 4 × 40 GbE QSFP+ ports in a 1U form factor
Software is based on Internet standards for optimal interoperability with Cisco or other vendors' networks	Up to 16x 10 GbE SFP+ connections using optional breakout cables
VMready for virtualized networks	Hot-swappable redundant power supplies and fans
OpenFlow enabled, making user-controlled virtual networks easy to create, optimizing performance dynamically and minimizing complexity	Two airflow options to manage hot- and cold-aisle applications

Further reading: Visit the following link for further information on IBM System Networking Solutions:

http://www-03.ibm.com/systems/networking/switches/rack.html

2.4.4 IBM System Networking RackSwitch G8316

The IBM® System Networking RackSwitch G8316 provides low latency, lossless performance, and a feature-rich design with key virtualization features such as Converged Enhanced Ethernet (CEE) / Data Center Bridging (DCB), high availability, and enterprise class Layer 2 and Layer 3 functionalities. In addition, the G8316 also delivers excellent cost savings as you consider acquisition costs, energy costs, operational expense, and ease of use and management for a 40 Gb class switch. The IBM System Networking RackSwitch G8316 supports both 10 Gb and 40 Gb Ethernet and is suitable for clients using 10 Gb or 40 Gb connectivity (or both).

Figure 2-12 shows the IBM System Networking RackSwitch G8316.

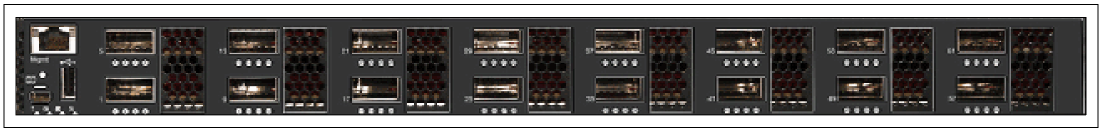


Figure 2-12 IBM System Networking RackSwitch G8316

The G8316 Ethernet aggregation switch enables customers to own an end-to-end flat 2-tier network based on industry standards using IBM switches and servers. For example, the G8316 is an ideal tier two switch to use when connecting a number of our RackSwitch G8264 40 Gb uplink ports at the access layer. Other clients like the G8316 as an investment protection switch, as they can leverage it with their 10 Gb Ethernet environments today, but can also leverage it in the future as they move to 40 Gb Ethernet.

With latency below 1 microsecond, the G8316 is an industry leader. This and the 1.28 Tbps throughput makes the G8316 an ideal offering for latency-sensitive applications such as high-performance computing, financial applications, hosting environments, and cloud designs. In addition, this switch has the ability to also support IBM Virtual Fabric, which helps clients significantly reduce cost and complexity when it comes to the I/O requirements of most virtualization deployments today. With Virtual Fabric, IT can allocate and dynamically change bandwidth per vNIC, making room for the ability to adjust over time without downtime. In addition, IBM VMready® with Virtual Vision enables the network to be Virtual Machine (VM) aware, and provides the capability to have a virtualization environment that is simpler, less expensive, and provides exceptional performance.

Benefits

The IBM System Networking RackSwitch G8316 offers the following benefits:

- High performance: This 10/40 Gb Low Latency (880 nanoseconds) Switch with 1.28 Tbps throughput provides the best combination of low latency, non-blocking line-rate switching, and ease of management. The G8316 is also a single ASIC design, which promises consistent lower port-to-port latency compared with other vendors with multiple chip designs, which causes port-to-port latency to be inconsistent and unpredictable.
- Lower power and better cooling: The RackSwitch G8264 uses as little as 330 W of power, which is a fraction of the power consumption of many competitive offerings. The front-to-rear or rear-to-front cooling design reduces data center air conditioning costs by having airflow match the servers in the rack. In addition, variable speed fans assist in automatically reducing power consumption.
- High availability: The G8316 also comes standard with hot-swap redundant power supplies and fans, making the switch highly reliable out-of-the-box and easy to service in the unlikely event of a failure.

- Virtual Fabric: This can help customers address I/O requirements for multiple NICs, while also helping reduce cost and complexity. Virtual Fabric for IBM allows for the carving up of a physical NIC into multiple virtual NICs (between 2 - 8 vNIC) and creates a virtual pipe between the adapter and the switch for improved performance, availability, and security, while reducing cost and complexity.
- VM-Aware Networking: VMready software on the switch simplifies configuration and improves security in virtualized environments. VMready automatically detects virtual machine movement between physical servers and instantly reconfigures each VM's network policies across VLANs to keep the network up and running without interrupting traffic or impacting performance. VMready works with all leading VM providers, such as VMware, Citrix, Xen, Microsoft Hyper V RedHat KVM, and IBM PowerVM®.
- Layer 3 functionality: The IBM switch includes Layer 3 functionality, which provides security and performance benefits, as inter-VLAN traffic stays within the switch. This switch also provides the full range of Layer 3 protocols from static routes for technologies such as Open Shortest Path First (OSPF) and Border Gateway Protocol (BGP) for enterprise customers.
- Seamless interoperability: IBM switches interoperate seamlessly with other vendors' upstream switches.
- Fault tolerance: These IBM switches learn alternate routes automatically and perform faster convergence in the unlikely case of a link, switch, or power failure. The switch uses proven technologies like L2 trunk failover, advanced VLAN-based failover, VRRP, and HotLink™.
- Multicast: This supports IGMP Snooping v1, v2, and v3 with 2 K IGMP groups, as well as Protocol Independent Multicast such as PIM Sparse Mode or PIM Dense Mode.
- Converged fabric: The IBM switch is designed to support CEE and Data Center Bridging, which is ideal for storage connectivity (NAS, iSCSI or FCoE). CEE will help enable clients to combine storage, messaging traffic, VoIP, video, and other data on a common data center Ethernet infrastructure. Data Center Bridging helps with iSCSI and FCoE with features like Priority-based Flow Control, Enhanced Transmission Selection, and Congestion Notifications. FCoE will help enable highly efficient block storage over Ethernet for consolidating server network connectivity. As a result, clients can deploy a single server interface for multiple data types, which can simplify both deployment and management of server network connectivity, while maintaining the high availability and robustness required for storage transactions.

For more information, see these websites:

http://www.ibm.com/systems/networking/switches/rack/g8316/index.html

http://www.redbooks.ibm.com/abstracts/tips0842.html?Open

2.4.5 IBM System Networking RackSwitch G8052

The IBM® System Networking RackSwitch G8052 is a Top-of-Rack data center switch that delivers unmatched line-rate Layer 2/3 performance at a very attractive price. It has forty-eight 10/100/1000BASE-T RJ45 ports and four 10 Gigabit Ethernet SFP+ ports, and includes hot-swap redundant power supplies and fans as standard, minimizing your configuration requirements. The G8052 has rear-to-front or front-to-rear airflow that matches server airflow.

Figure 2-13 shows the IBM System Networking RackSwitch G8052.

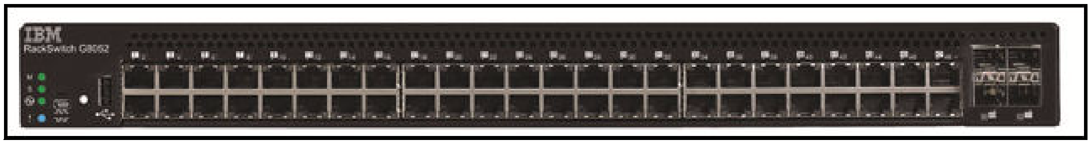


Figure 2-13 IBM® System Networking RackSwitch G8052

The IBM System Networking RackSwitch G8052 provides the following benefits:

- High performance: The RackSwitch G8052 provides up to 176 Gbps throughput and supports four SFP+ 10 G uplink ports for a very low oversubscription ratio, in addition to a low latency of 1.8 microseconds.
- Lower power and better cooling: The RackSwitch G8052 typically consumes just 130 W of power, a fraction of the power consumption of most competitive offerings. Unlike side-cooled switches, which can cause heat recirculation and reliability concerns, the G8052's rear-to-front or front-to-rear cooling design reduces data center air conditioning costs by matching airflow to the server's configuration in the rack. Variable speed fans assist in automatically reducing power consumption.
- VM-Aware Network Virtualization: IBM VMready software on the switch simplifies configuration and improves security in virtualized environments. VMready automatically detects VM movement between physical servers and instantly reconfigures each VM's network policies across VLANs to keep the network up and running without interrupting traffic or impacting performance. VMready works with all leading hypervisors, such as VMware, Citrix Xen, Red Hat KVM, Microsoft Hyper V, and IBM PowerVM® hypervisor.
- Layer 3 Functionality: The RackSwitch G8052 includes Layer 3 functionality, which provides security and performance benefits plus the full range of Layer 3 static and dynamic routing protocols, including Open Shortest Path First (OSPF) and Border Gateway Protocol (BGP) for enterprise customers at no additional cost.
- Fault tolerance: These switches learn alternate routes automatically and perform faster convergence in the unlikely case of a link, switch, or power failure. The switch uses proven technologies such as L2 trunk failover, advanced VLAN-based failover, VRRP, HotLink, Uplink Failure Detection (UFD), IGMP V3 snooping, and OSPF.
- Seamless interoperability: IBM switches interoperate seamlessly with other vendors' upstream switches.

Key features are as follows:

- Designed for line-rate throughput and low latency less than 2 microseconds.
- Includes redundant and hot-swappable power supplies and fans.
- RackSwitch G8052 is designed specifically for the data center environment with server-matching airflow, high-availability hardware and software features, rich Layer 2/3 functionality, and ease of management.

- IBM VMready® switch-resident software reduces the complexity of configuring and managing virtual machines throughout the network, making it VM-aware.
- IBM Networking Operating System software features deliver seamless, standards-based integration into existing upstream switches.
- 48 × 1 GbE RJ45 ports, 4 × 10 GbE SFP+ ports.
- Standard redundant power and fans.
- Low 130 W power consumption.
- Simplified configuration with ISCLI.
- Server-like airflow for hot- and cold-aisle applications.
- VMready for virtualized networks.

For more information, see these websites:

http://www.ibm.com/systems/networking/switches/rack/g8052/index.html

http://www.redbooks.ibm.com/abstracts/tips0813.html?Open

2.4.6 IBM System Networking RackSwitch G8000

The IBM System Networking RackSwitch™ G8000 is an Ethernet switch specifically designed for the data center, providing a virtualized, cooler, and simpler network solution.

The G8000 is virtualized for the first time providing rack-level virtualization of networking interfaces for a rack full of server and storage systems decoupling the scaling of networking and computing capacity via on-switch VMready software. VMready enables the movement of virtual machines, providing matching movement of VLAN assignments, ACLs, and other networking and security settings. VMready works with all leading VM providers.

The G8000 is cooler implementing a choice of directional cooling options to maximize data center layout and provisioning. Its superior airflow design complements the hot-aisle and cold-aisle data center cooling model.

The G8000 is easier with server-oriented provisioning via point-and-click management interfaces, along with optional IBM System Networking Element Manager for updating large groups of switches.

The IBM RackSwitch G8000 offers 48 x 1 Gigabit Ethernet ports and up to 4 x 10 Gigabit Ethernet (GbE) ports in a 1U footprint. Redundant power supplies and fans, along with numerous high-availability features, mean that the G8000 is always available for business-sensitive traffic. Figure 2-14 shows the IBM RackSwitch G8000.

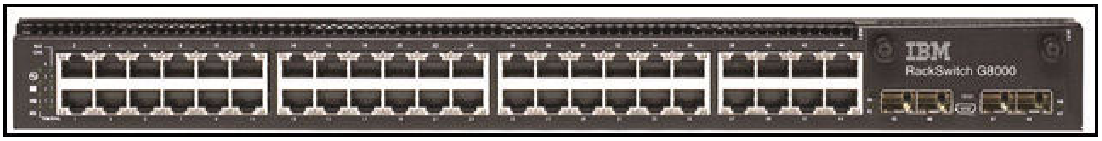


Figure 2-14 IBM RackSwitch G8000

The IBM System Networking RackSwitch G8000 offers the following benefits:

- High performance: The G8000 provides up to 176 Gbps throughput and supports up to four 10 G uplink ports for a very low oversubscription ratio.
- Lower power and better cooling: The G8000 uses approximately 120 W, which is a fraction of the power consumption of most competitive offerings. Unlike side-cooled switches, which can cause heat recirculation and reliability concerns, the G8000 rear-to-front/front-to-rear cooling design reduces data center air conditioning costs by matching airflow to the server's configuration in the rack. Variable speed fans assist in automatically reducing power consumption.
- VM-Aware network virtualization: VMready software on the switch reduces configuration complexity while significantly improving security levels in virtualized environments. VMready automatically detects virtual machine movement from one physical server to another, and instantly reconfigures each VM's network policies across VLANs to keep the network up and running without interrupting traffic or impacting performance. VMready works with all leading hypervisors, such as VMware, Citrix Xen, RedHat KVM, Microsoft Hyper V, and IBM PowerVM® hypervisor.
- Layer 3 functionality: The switch includes Layer 3 functionality, which provides security and performance benefits, plus static and dynamic routing protocols including Open Shortest Path First (OSPF) and Border Gateway Protocol (BGP) for enterprise customers.
- USGv6 Testing Program Report (USGv6-V1):

  http://public.dhe.ibm.com/common/ssi/ecm/en/qcl12356usen/QCL12356USEN.PDF
- Fault tolerance: These switches learn alternate routes automatically and perform faster convergence in the unlikely case of a link, switch, or power failure. The switch uses proven technologies such as L2 trunk failover, advanced VLAN-based failover, VRRP, HotLink™, Uplink Failure Detection (UFD)™, IGMP V3 snooping, and OSPF.
- Seamless interoperability: Switches interoperate seamlessly with other vendors' upstream switches. See the following resources:
- Tolly Functionality and Certification: RackSwitch G8000, G8100, and G8124 and Cisco Catalyst Interoperability Evaluation:

  http://www.bladenetwork.net/userfiles/file/PDFs/Tolly209116BladeRackSwitchInteroperability.pdf

  Tolly functionality and certification and Cooperative Interoperability Evaluation with the Cisco Nexus 5010:

  http://public.dhe.ibm.com/common/ssi/ecm/en/qcl12346usen/QCL12346USEN.PDF
- Stacking: Cluster virtualization, single IP, and stacking of up to six virtual switches allow configuration of multiple switches as one, saving valuable time, IP addresses, and resources.

For more information, see these websites:

http://www.ibm.com/systems/networking/switches/rack/g8000/index.html

http://www.redbooks.ibm.com/abstracts/tips0786.html?Open

2.5 IBM Power Systems

Any large or medium business relies on the systems that provide enterprise-wide information. This information is derived from financial data, customer data, and enterprise resource data, and is held across multiple lines of business. To transform the management of critical information and processes, businesses rely on enterprise systems, the servers, storage, and software at the core of an enterprise IT infrastructure. And, everyone knows what "performance" meant for IT in the past; it was based on processing power and benchmarks.

However, in this new smarter computing era for business and IT, forward-thinking companies consider more than server performance, existing skills, and ease of management when choosing a platform for new application workloads. They also evaluate how well the platform will help them achieve three core business objectives: delivering new services faster, with higher quality, and superior economics.

The IBM Power Systems™ platforms are designed specifically to meet increasing workloads of new era computing and, at same to provide a comprehensive approach to protect the continuity and availability of business-critical applications and services. Deploying integrated disaster recovery allows companies to deliver new services faster, with higher quality, and with superior economics. The IBM Power Systems have always focused on delivering outstanding performance along with the business needs of application choice, enterprise integration, IT efficiency, and data availability and security, thus enabling business to achieve outstanding business performance and compute performance.

In addition, today's businesses depend on Big Data, cloud, mobile, and social business applications along with their traditional and new business applications. An IBM Power systems portfolio with its capability of managing new mix of business applications and information centric workloads can deliver these benefits:

- Dynamic efficiency, making IT a driver of business growth with intelligent, dynamic resource allocation and rapid response to changes in application and workload requirements
- Business analytics, to unlock the value of data by providing business insight with servers optimized for Big Data, compute intensive applications, and built-in availability
- Enhanced compliance, by automating infrastructure security and enabling policy-based compliance without compromising system performance

Power architecture: IBM POWER® processor technology is a reduced instruction-set (RISC) architecture that spans applications from consumer electronics to supercomputers. POWER processor technology is built on an open architecture, making it an open ecosystem that supports freedom of design. This provides a foundation for designing workload optimized systems in conjunction with software and expert domain knowledge. You can learn more about the world of POWER at the following website:

`http://www.power.org`

2.5.1 IBM Power models

In this section, we introduce you some of the family members from IBM Power Enterprise, Express editions, and IBM High Performance computing, which are fundamental building blocks for any kind of next generation data centers. We briefly discuss the PowerVM feature and Power Systems management solution.

IBM Power 770 Enterprise system: A reliable and secure system

The emerging measures of IT performance are around agility and the ability to help the business capitalize on new opportunities. IT is measured on providing an infrastructure that can handle rapid growth and manage business risk while meeting higher required service levels. And of course it is expected that new services will be delivered within tighter budget constraints with IT expected to do more with less and find the most efficient solutions possible. IBM Power Enterprise models delivers on performance, availability, efficiency and virtualization in a way that is unique in the industry and Power VM enables continuous, dynamic resource adjustments across all partitions and operating environments to optimize performance and enable higher utilization levels while optimizing energy usage.

IBM Power 770 provides a unique combination of performance across multiple workloads and availability features to keep businesses running. In addition, PowerVM virtualization that helps to maximize the efficiency and non-disruptive growth options are designed to keep costs in line with business. For transaction processing workloads, the IBM Power 770 server delivers outstanding performance, mainframe-inspired reliability, modular non-disruptive growth and innovative virtualization technologies. These features are integrated to enable the simplified management of growth, complexity and risk. For database serving, the Power 770 provides a system designed for demanding, critical, cloud-enabled workloads. Additional benefits include increased application performance, operational availability and security that can only come from IBM Power 770.

Table 2-6 explains some of the IBM Power 770 features and their benefits.

Table 2-6 IBM Power 770 features and benefits

Feature	Benefits
Industry-leading IBM POWER7+™ performance	► Better customer satisfaction due to improved response time to customers ► Infrastructure cost savings from a reduction in the number of servers and software costs ► Improved efficiency in operations from consolidating multiple workloads on fewer
Exceptional PowerVM virtualization capability	► Improves system efficiency, which lowers operational expense ► Provides flexibility in responding to changing business requirements ► Enables energy savings and maintains application availability ► Provides ability to handle unexpected workload peaks by sharing resources ► Enables consolidation of multiple AIX®, IBM i and Linux workloads
Mainframe-inspired availability features	► Better customer satisfaction due to improved application availability ► Get more work done with less disruption to business ► Faster repair when required due to sophisticated system diagnostics
Nondisruptive growth options	► Enables system to change with business without forcing everything to stop ► Aligns expense with usage without sacrificing performance or future growth options
Frugal Energy Scale technology	► Helps lower energy costs without sacrificing performance or business flexibility ► Allows business to continue operations when energy is limited

Feature	Benefits
Innovative IBM Active Memory™ Expansion	► Enables more work to be performed with existing resources ► Partition level control enables flexibility and optimization for workload
Broad business application support	► Allows clients the flexibility to select the right application to meet their needs ► Helps keep in the mainstream and off the bleeding edge

Figure 2-15 shows the IBM Power 770 server.

Figure 2-15 IBM Power 770

IBM Power 730 Express: A high-performance and reliable system

Organizations of all sizes are struggling with rising energy prices and limited resources. They are running out of space and exceeding power thresholds in computer rooms and data centers. And, they are increasingly being asked to *do more with less.* IBM Power 730 Express is designed with innovative workload-optimizing and energy management technologies to help organizations to get the most out of their systems; that is, running applications fast and energy efficiently to conserve energy and reduce infrastructure operational costs.

The IBM Power 730 Express server delivers the outstanding performance of the POWER7 processor in a dense, rack-optimized form factor and is ideal for running multiple application and infrastructure workloads in a virtualized environment. Advantages of the Power 730 Express servers are scalability and capacity, as it is leveraging the IBM industrial strength PowerVM technology to fully utilize the servers' capability.

PowerVM offers this capability to dynamically adjust system resources to partitions based on workload demands, enabling a dynamic infrastructure that can dramatically reduce server sprawl via consolidation of applications and servers.

The Power 730 Express server is a two-socket high-performance server supporting up to sixteen POWER7 cores and a choice of AIX, IBM i, or Linux operating systems. Combining the outstanding performance with PowerVM and the workload-optimizing capabilities of POWER7 and business, you are ready to run multiple application and infrastructure workloads in a virtualized environment while driving higher utilization and reducing costs.

Table 2-7 explains the features of the IBM Power 730 Express Server and their benefits.

Table 2-7 iBM Power 730 Express Server features and benefits

Feature	Benefits
Leadership POWER7 performance	► Accesses data faster and improve response time ► Does more work with fewer servers and benefit from infrastructure cost savings from a reduction in the number of servers and software licenses
IBM Systems Director Active Energy Manager™ with IBM EnergyScale™ Technology	► Dramatically and dynamically improves energy efficiency and lower energy costs with innovative energy management capabilities. ► Enables businesses to continue operations when energy is limited.
PowerVM Virtualization	► Easily adds workloads as business grows ► Utilizes the full capability of the system to reduce infrastructure costs by consolidating workloads running the AIX, IBM i or Linux operating ► Provides ability to handle unexpected workload peaks by sharing resources
Intelligent Threads	► Optimizes performance by selecting the suitable threading mode for varies applications
Active Memory Expansion	► Enables more work to be done with existing server resources
RAS Features	► Keeps applications up and running so IT organization can focus on growing business.
Light Path Diagnostics	► Easily and quickly diagnoses hardware problems, reducing service time.

The Power 730 Express is designed with the capabilities to deliver leading-edge application availability and allow more work to be processed with less operational disruption. RAS capabilities include recovery from intermittent errors or fail-over to redundant components, detection and reporting of failures and impending failures, and self-healing hardware that automatically initiates actions to effect error correction, repair or component replacement. In addition, the Processor Instruction Retry feature provides for the continuous monitoring of processor status with the capability to restart a processor if certain errors are detected. If required, workloads are redirected to alternate processors, all without disruption to application execution.

Figure 2-16 shows the IBM Power 730 Express rack-mount server.

Figure 2-16 IBM Power 730 Express rack-mount server

The Power 730 Express implements Light Path diagnostics, which provide an obvious and intuitive means to positively identify failing components. This allows system engineers and administrators to easily and quickly diagnose hardware problems. Such failures might have taken hours to locate and diagnose, but can now be detected in minutes, avoiding or significantly reducing costly downtime.

> **Tip:** For more technical information, see the *IBM Power 710 and 730 (8231-E2B) Technical Overview and Introduction*, REDP-4636.

IBM BladeCenter Express servers series

Built on the proven foundation of the IBM BladeCenter® family of products that are easy-to-use, integrated platforms with a high degree of deployment flexibility, energy efficiency, scalability and manageability the BladeCenter PS700, PS701 and PS702 Express are the premier blades for 64-bit applications. The new POWER7 processor-based PS blades automatically optimize performance and capacity at either a system or virtual machine level and benefits from the new POWER7 processor, which contains innovative technologies that help maximize performance and optimizes energy efficiency. They represent one of the most flexible and cost-efficient solutions for UNIX, IBM i and Linux deployments available in the market. Further enhanced by its ability to be installed in the same chassis with other IBM BladeCenter blade servers, the PS blades can deliver the rapid return on investment that businesses demand.

IBM BladeCenter PS700 blade server

The PS700 blade server (8406-70Y) is a single socket, single wide 4-core 3.0 GHz POWER7 processor-based server. The POWER7 processor is a 64-bit, 4-core with 256 KB L2 cache per core and 4 MB L3 cache per core. The PS700 blade server has eight DDR3 memory DIMM slots. The industry standard VLP DDR3 Memory DIMMs are either 4 GB or 8 GB running at 1066 MHz. The memory is supported in pairs, thus the minimum memory required for PS700 blade server is 8 GB (two 4 GB DIMMs). The maximum memory that can be supported is 64 GB (eight 8 GB DIMMs). It has two Host Ethernet Adapters (HEA) 1 GB integrated Ethernet ports that are connected to the BladeCenter chassis fabric (midplane).

The PS700 has an integrated SAS controller that supports local (on-board) storage, integrated USB controller and Serial over LAN console access through the service processor, and the BladeCenter Advance Management Module. It supports two on-board disk drive bays. The on-board storage can be one or two 2.5-inch SAS HDD. The integrated SAS controller supports RAID 0, RAID 1, and RAID 10 hardware when two HDDs are used. The PS700 also supports one PCIe CIOv expansion card slot and one PCIe CFFh expansion card slot.

Figure 2-17 shows the following blades:

- IBM BladeCenter PS700: Single-wide blade with a single-socket 4-core processor
- IBM BladeCenter PS701: Single-wide blade with a single-socket 8-core processor
- IBM BladeCenter PS702: Double-wide blade with two single-socket 8-core processors

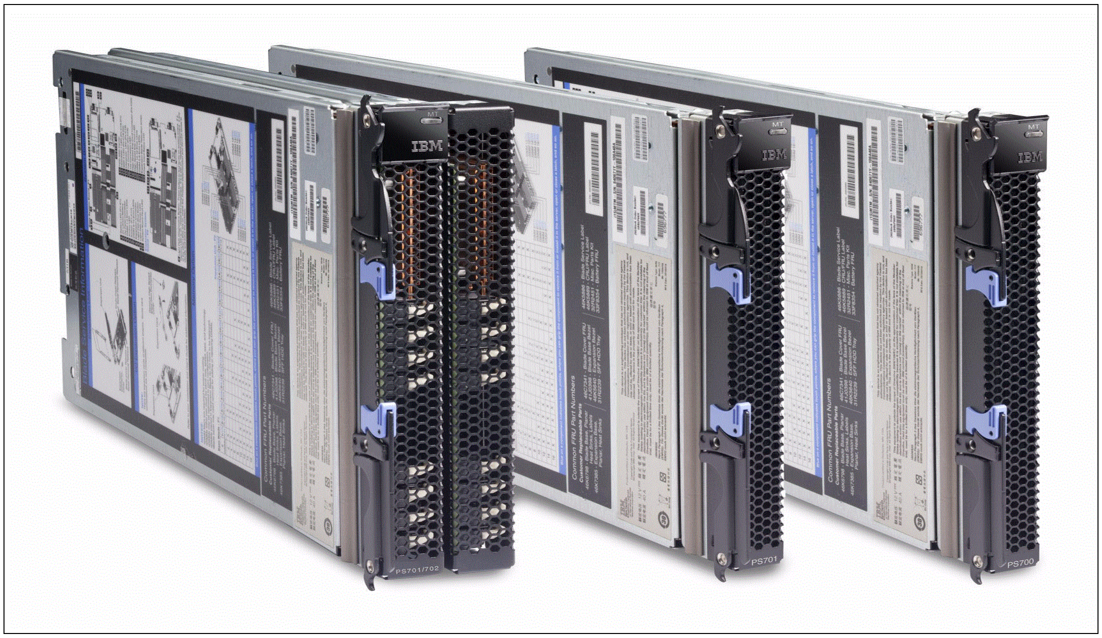


Figure 2-17 The IBM BladeCenter PS702,BladeCenter PS701 and a BladeCenter PS700

IBM BladeCenter PS701 blade server

The PS701 blade server (8406-71Y) is a single socket, single-wide 8-core 3.0 GHz POWER7 processor-based server. The POWER7 processor is a 64-bit, 8-core with 256 KB L2 cache per core and 4 MB L3 cache per core. The PS701 blade server has 16 DDR3 memory DIMM slots. The industry standard VLP DDR3 memory DIMMs are either 4 GB or 8 GB running at 1066 MHz. The memory is supported in pairs, thus the minimum memory required for PS701 blade server is 8 GB (two 4 GB DIMMs). The maximum memory that can be supported is 128 GB (16x 8 GB DIMMs). The PS701 blade server has two Host Ethernet Adapters (HEA) 1 GB integrated Ethernet ports that are connected to the BladeCenter chassis fabric (midplane).

The PS701 also has an integrated SAS controller that supports local (on-board) storage, integrated USB controller and Serial over LAN console access through the service processor, and the BladeCenter Advance Management Module. The PS701 has one on-board disk drive bay. The on-board storage can be one 2.5-inch SAS HDD. The PS701 also supports one PCIe CIOv expansion card slot and one PCIe CFFh expansion card slot.

Table 2-8 explains some of the key features of IBM BladeCenter PS700, PS701, and PS702 Express servers and their benefits.

Table 2-8 IBM BladeCenter Express servers features and their benefits

Feature	Benefits
World's first scalable POWER7-based blade servers	► Maximize performance and minimize costs; consolidate workloads and virtualize on an energy-efficient platform that supports the latest POWER7 processor technology ► Pay as grow; start with a 1-processor chip (8 core) blade and upgrade to a 2-processor chip (16 core) blade when ready without scrapping the initial investment ► Save time and money; standardize on a single blade platform for both 1- and 2-processor chip server application needs
Highly efficient and flexible design of IBM BladeCenter	► Densely pack more servers in a smaller space ► Tailor system to meet varied business requirements with a choice of BladeCenter chassis ► Lower acquisition cost and energy consumption versus traditional 1U or 2U rack servers ► Integrate networking switch infrastructure for improved cabling and data center maintenance ► Deploy in virtually any office environment for quiet, secure and contaminant-reduced operation
Pioneering EnergyScale technology and IBM Systems Director Active Energy Manager software	► Generate less heat by managing application utilization and server energy consumption ► Use less energy to cool the system
Industry-leading IBM PowerVM virtualization technology	► Reduce infrastructure costs by running more workloads with fewer servers ► Simplify IT operations by virtualizing storage, network and computing resources ► Manage risk and minimize downtime with rapid provisioning and improved resilience
Innovative reliability features and systems management	► Expedite hardware repairs and reduce service time ► Enable scheduled maintenance with proactive monitoring of critical system components to help reduce unplanned failures
Choice of AIX, IBM i or Linux operating systems	► Standardize on a single platform that runs the large and varied portfolio of applications that support the business ► Take advantage of the power of the IBM industry-leading UNIX operating system, AIX ► Utilize the Linux for Power operating system to access the breadth of open source applications ► Exploit the simplicity of the integrated IBM i operating environment

IBM BladeCenter PS702 blade server

The PS702 blade server (8406-71Y +FC 8358) is a two socket, double-wide 16-core 3.0 GHz POWER7 processor-based server. The POWER7 processor is a 64-bit, 8-core with 256 KB L2 cache per core and 4 MB L3 cache per core. The PS702 combines a single-wide base blade (PS701) and an expansion unit (feature 8358), referred to as double-wide blade, which occupies two adjacent slots in the IBM BladeCenter chassis.

The PS702 blade server has 32 DDR3 memory DIMM slots. The industry standard VLP DDR3 memory DIMMs are either 4 GB or 8 GB running at 1066 MHz. The memory is supported in pairs, thus the minimum memory required for PS702 blade server is 8 GB (two 4 GB DIMMs). The maximum memory that can be supported is 256 GB (32x 8 GB DIMMs).

The PS702 blade server has four Host Ethernet Adapter 1 GB integrated Ethernet ports that are connected to the BladeCenter chassis fabric (midplane). The PS702 also has an integrated SAS controller that supports local (on-board) storage, integrated USB controller and Serial over LAN console access through the service processor, and the BladeCenter Advance Management Module. The PS702 blade server has two disk drive bays, one on the base blade and one on the expansion unit. The on-board storage can be one or two 2.5-inch SAS disk drives. The integrated SAS controller supports RAID 0, RAID 1 and RAID 10 hardware when two HDDs are used. The PS702 supports two PCIe CIOv expansion card slot and two PCIe CFFh expansion card slots.

To know more about IBM Power Blade models, see the *IBM BladeCenter PS700, PS701, and PS702 Technical Overview and Introduction*, REDP-4655.

> **Note:** Red Hat and Novell/SUSE Linux for Power operating systems may be ordered from IBM and select Linux distributors.

IBM PowerLinux server: A high-performance server for Linux

The IBM PowerLinux™ 7R2 server is designed specifically as an economical foundation for emerging and traditional scale-out workloads. IBM PowerLinux workload optimized solutions, each tuned to a specific task, are affordable for businesses of all sizes. A simpler PowerLinux-based IT environment with fewer servers to manage helps reduce infrastructure costs and frees IT staff to focus on bringing innovative products and services to market faster.

Figure 2-18 shows the IBM PowerLinux 7R2 rack-mountable server.

Figure 2-18 IBM PowerLinux 7R2 rack-mount server

The IBM PowerLinux 7R2 is designed with capabilities to deliver leading-edge application availability and allow more work to be processed with less operational disruption. RAS capabilities include recovery from intermittent errors or failover to redundant components, detection and reporting of failures and impending failures, and self-healing hardware that automatically initiates actions to effect error correction, repair or component replacement. In addition, the Processor Instruction Retry feature provides for the continuous monitoring of process or status with the capability to restart a processor if certain errors are detected. If required, workloads are redirected to alternate processors, all without disruption to application execution.

Table 2-9 lists the IBM PowerLinux servers features and their respective benefits.

Table 2-9 IBM Power Linux server features and their respective benefits

Feature	Benefits
Leadership POWER7 performance	► Access data faster and improve response time ► Do more work with fewer servers and benefit from infrastructure cost savings from a reduction in the number of servers and software licenses
IBM Systems Director Active Energy Manager with EnergyScale Technology	► Dramatically and dynamically improve energy efficiency and lower energy costs with innovative energy management capabilities ► Enables businesses to continue operations when energy is limited
IBM PowerVM for IBM PowerLinux	► Easily add workloads as business grows ► Utilize the full capability of the system to reduce infrastructure costs by consolidating workloads onto the Linux operating system ► Provides ability to efficiently handle unexpected workload peaks by sharing resources
Intelligent Threads	► Optimize performance by selecting the suitable threading mode for application
Active Memory Sharing	► Do more with less memory by dynamically allocating memory to virtualized workloads as required vs. inefficient fixed allocations
RAS Features	► Keep applications up and running so focus can be on growing the business
Light Path Diagnostics	► Easily and quickly diagnose hardware problems, reducing service time
Industry standard Linux from Red Hat and SUSE Operating System support	► Access thousands of applications available from the Linux and open source community, ISVs and IBM Software ► Take advantage of widely available skills and collaboration across the Linux community ► Choice of Linux operating system from Red Hat and SUSE

With the Live Partition Mobility feature of PowerVM for PowerLinux, running workloads can be moved between servers to eliminate planned downtime. The PowerLinux 7R2 implements Light Path diagnostics, which provide an obvious and intuitive means to positively identify failing components. This allows system engineers and administrators to easily and quickly diagnose hardware problems. Hardware failures that may have taken hours to locate and diagnose can now be detected in minutes, avoiding or significantly reducing costly downtime.

Heat: The heat produced by server chips is typically removed by using massive copper plates. IBM actively researches this field and has current solutions that allow for direct water cooling at the chip's backside. After the heat is transferred to water, it can be handled more efficiently. In this chapter we discuss some innovative IBM water cooled servers too.

IBM Power 775 Supercomputer: High performance computing

The IBM Power 775 super-computing server is designed for organizations that require a highly scalable system with extreme parallel processing performance and dense, modular packaging.

You can use it in clustered configurations of as few as 512 processor cores or in world-class supercomputer configurations of hundreds of thousands of processors. Combined with specialized software from IBM, this system is designed to perform and represents the latest Power technology available

Ideal workloads for this system include high performance computing (HPC) applications such as weather and climate modeling, computational chemistry, physics, computer-aided engineering, computational fluid dynamics and petroleum exploration that require highly intense computations where the workload is aligned with parallel processing methodologies.

The Power 775 is 100% water cooled and powered from either 3-phase AC or DC to enable maximum data center energy efficiency. It also takes advantage of advanced IBM EnergyScale technology. It is designed to deliver high performance, scalability, reliability, and manageability for demanding computing applications.

The system racking sequence places the water cooling units in the bottom of the rack, followed by the Central Electronic Complex (CEC) drawers which are also populated from the bottom up. A 2U tall, 30-inch wide, 256-core POWER7 CEC drawer houses the system planars, redundant 350 V DC power supplies, water-cooled cold plates and manifolds.

This radical new approach to HPC represented by the Power 775 system marks another step in the evolution of modular clusters designed to solve the world's greatest problems. This ultra-dense packing of 256 cores in a 2U drawer coupled with new water cooling technology enables the POWER7 processors of the Power 775 to satisfy extremely demanding High Performance Computing (HPC) applications in a very small physical footprint while effectively removing excess heat.

Operating environment

- Inlet water continuous operating temperature: 2 to 20 degree C (35 to 68 degree F)
- Inlet air continuous operating temperature: 5 to 40 degree C (41 to 104 degree F)
- Relative humidity: Non-condensing 20% to 80%
- Maximum dew point: 17 degrees

Power requirements

- AC operating voltages (3-phase V ac) at 50/60Hz: 200 to 240V; 380 to 440V; 480V
- DC operating voltage (5 -wire, V dc); 380 V dc; 520 V

Highlight

The dense. modular and highly integrated design enables data center consolidation when space is at a premium.

Figure 2-19 IBM Power 775 Supercomputer

2.5.2 Power VM technology

IBM PowerVM virtualization technology is a combination of hardware and software that supports and manages virtual environments on IBM POWER5, POWER5+, IBM POWER6®, and POWER7 processor-based systems. These systems are available on IBM Power Systems and IBM BladeCenter servers as optional editions, and are supported by the IBM AIX, IBM i, and Linux operating systems. With this set of comprehensive systems technologies and services, IT organizations can aggregate and manage resources with a consolidated, logical view.

By deploying PowerVM virtualization and IBM Power Systems, organizations can take advantage of the following benefits:

- Lower energy costs through server consolidation
- Reduced cost of the existing infrastructure
- Better management of the growth, complexity, and risk of the infrastructure

See the *IBM PowerVM Getting Started Guide*, REDP-4815 for assistance with the quick and easy installation and configuration of new PowerVM server in a virtualized environment.

2.5.3 IBM Systems Management Solutions for Power Servers

IT staff today have more flexibility in deploying new and expanded business workloads due to proven technologies like virtualization. However, virtualization alone, without effective systems management tools and processes can increase the workload on an already-overburdened IT staff, inhibiting their ability to deliver innovative services and drive business growth.Orgnizations need to address today's challenges and tomorrow's opportunities, ensure high availability, security, and quality of existing services, while meeting customer expectations for new services that are delivered in the shortest possible time.

IBM Systems Director Editions for Power Servers, a web based, point and click offers comprehensive solutions that enable IT staff to improve service delivery by providing a consolidated view of physical and virtual resources running any workload combinations on AIX, Linux, and IBM i operating environments. Enterprise service management solutions for AIX workloads helps more closely link managing IT operations with meeting business requirements.

With platform management technologies on Power Systems, businesses not only get a complete picture of their systems and how well they are operating, but also the tools to deploy, optimize and maintain these systems at maximum effectiveness and efficiency.

These are a few key highlights for IBM System Director for Power Systems:

- Optimize virtualization to more quickly respond to changing business requirements
- Visualize and monitor physical and virtual servers at the platform, system, and enterprise level
- Increase operational efficiency by unifying essential management of servers, storage, networks, applications and energy with a consistent look and feel
- Reduce energy costs and usage through monitoring and analysis
- Easy integration with enterprise service management tool from IBM Tivoli® as well as third-party providers

For more information on deployment of IBM System Director at your environment, see the *IBM Systems Director Management Console: Introduction and Overview*, SG24-7860 and *Implementing IBM Systems Director 6.1*, SG24-7694.

2.6 IBM System z

IBM System z®, popularly known as IBM mainframes, are systems designed to provide the most reliable environment for high-value critical enterprise applications, with the goal of reaching near zero service downtime. In this section we discuss the newest IBM zEnterprise® systems such as z196 and z114.

The IBM zEnterprise System (zEnterprise) offers a revolutionary system design that addresses the complexity and inefficiency in today’s multi-architecture data centers. zEnterprise extends the strengths and capabilities of the mainframe such as security, fault tolerance, efficiency, virtualization and dynamic resource allocation to other systems and workloads running on AIX on POWER7, Linux on IBM System x®, and now Microsoft Windows fundamentally changing the way data centers can be managed.

zEnterprise is a workload-optimized, multi-architecture compute system capable of hosting many workloads integrated together, and efficiently managed as one single entity. It is designed to deploy and intelligently manage workloads across both mainframe and distributed technologies with the same tools, techniques and a single management interface.

These are hybrid computing solutions, which can utilize a zEnterprise BladeCenter Expansion (zBX) unit to support up to 8 BladeCenters. This expansion increases flexibility as each BladeCenter can be equipped with Optimizers to improve performance or with IBM Power7 or System x blade servers for a heterogeneous workload. Both the z196 and z114 are designed to work seamlessly with system software, middleware and storage to be the most robust, cost effective, efficient, and reliable data serving and transaction processing environment. Also we introduce you to an IBM z/VM® feature that is key to the software side of virtualization on the mainframes along with linux on System z.

Figure 2-20 shows the zEnterprise System with its management capabilities.

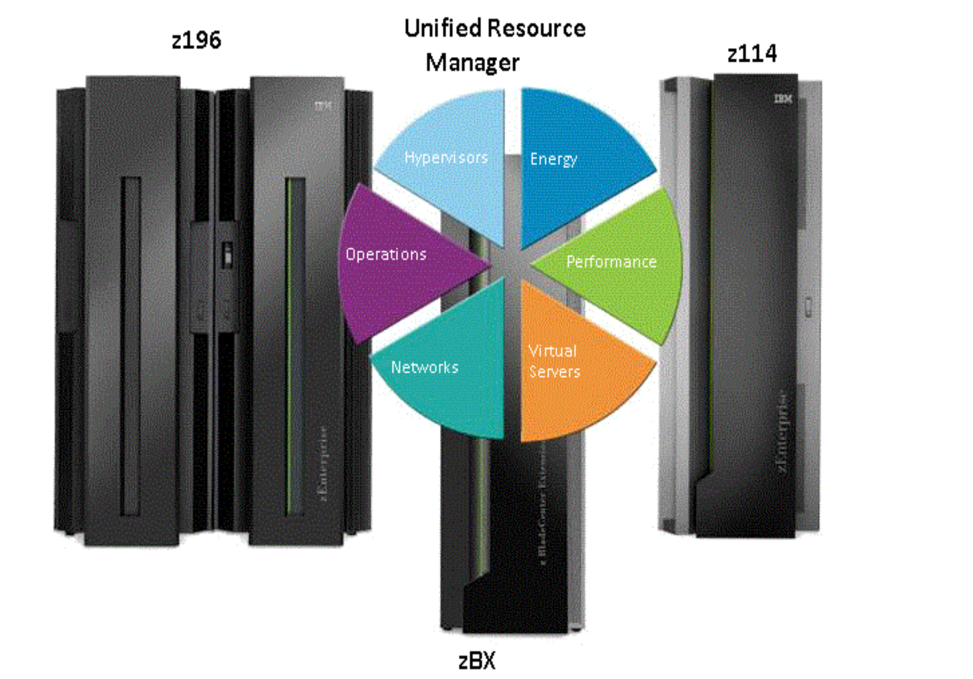


Figure 2-20 zEnterprise System and management capabilities

2.6.1 IBM zEnterprise 114: A system of systems

Designed specifically as a midrange zEnterprise offering with extensive growth options, industry-leading virtualization and consolidation capabilities, the z114 continues the heritage in mainframe qualities of service with a more granular cost structure and significant improvements in packaging, performance and total system scalability over prior generations. And as a highly workload-optimized system, the z114 extends the unique value of hybrid computing capabilities to businesses of all sizes for integrated, centrally managed workload deployment all within an environmentally friendly footprint. With improved processor performance, increased capacity, new hybrid computing capabilities, and significant power, space and cooling benefits, the z114 is now a genuine “data-center in a box” solution and a perfect fit for infrastructure simplification and true cloud computing.

These exclusive hybrid capabilities are designed to address the complexity and inefficiency of today’s multi-architecture data centers. The z114 can extend the strengths and capabilities of the mainframe as in security, fault tolerance, efficiency, extreme virtualization and dynamic resource allocation to other systems and workloads running on AIX, Linux on System x, configured with a zEnterprise BladeCenter Extension (zBX), one can combine System z, UNIX and Intel server technologies into a single unified system integrating workloads with affinity to mainframe applications and data and manage it all with the same tools, techniques and resources for consistent, automated, and reliable service delivery.

This capability provides the unique ability to extend the strategic role of mainframe across business and an opportunity to simplify and reduce the range of skills necessary for managing the IT infrastructure. This revolutionary yet simplified multi-architecture design helps drive operational efficiencies in deployment, integration and management. It is designed to enhance the attractiveness of hosting applications on the mainframe by providing the freedom and ease of selecting the right architecture and operating system for application deployment within a common management infrastructure for both mainframe and distributed-system resources.

The virtualization capabilities of the z114 can support an average of 30 distributed servers or more on a single core depending on the workload, or up to hundreds in a single foot-print, delivering a virtual Linux server for under US$1.45 day.[1] And, for ease of installation, the z114 is a single frame, air cooled system that now supports either top or bottom exit I/O and power; raised floor and non-raised floor options and high-voltage DC power, providing increased flexibility to accommodate small data center installations and support for future data center design and efficiencies. IBM system z114 is the probable answer for whether to deploy new applications quickly, growing the business without growing IT costs or consolidating existing infrastructure for reduced complexity.

Note: To learn more about the IBM system z114, see this website:

http://www.ibm.biz/Bdx2Pa

2.6.2 IBM zEnterprise 196

The demands of a fast moving market are stretching the limits of today's data centers. Add to this the management and integration challenges that data centers face as they invest in the next generation of smart applications and it is clear something new is needed. Smarter computing systems that raise the bar on efficiency, performance and cost savings while lowering management complexity. The zEnterprise System is a revolutionary combination of mainframe and distributed computing designed in one system. The zEnterprise 196 is one of the industries fastest and most scalable enterprise systems combined with the zEnterprise BladeCenter Extension (zBX), and the unique zEnterprise Unified Resource Manager, brings integration of governance to manage risk across the infrastructure, integration that will help accelerate insight for the business, integration of processes to increase business agility, and integration of people to enable new business innovation.

The zEnterprise 196 (z196) is a workload optimized system, that can scale up (over 52,000 MIPS in a single footprint), scale out (80 configurable cores) and scale within (specialty engines, cryptographic processors, hypervisors) executing in an environmentally friendly footprint. And the z196 is designed to work together with system software, middleware and storage to be the most robust and cost effective data server. The z196 offers an industry standard PCIe I/O drawer for IBM FICON® and OSA-Express multi-mode and single mode fiber optic environments for increased capacity, infrastructure bandwidth, and reliability.

z196 highlights

The major z196 improvements over its predecessors include the following features:

- Increased total system capacity in a 96-way server (with 80 characterizable PUs) and additional sub-capacity settings, offering increased levels of performance and scalability to help enable new business growth
- Quad-core 5.2 GHz processor chips that can help improve the execution of processor-intensive workloads
- Implementation of out-of-order instruction execution
- Cache structure improvements and larger cache sizes that can benefit most production workloads

[1] Linux on System z virtual servers can be: Less than $1545 for three years; Less than $515 per year; Less than $1.45 per day. Based on US Enterprise Linux Server pricing. Pricing may vary by country. Model configuration included 10 IFL cores running a mixed workload averaging 31 virtual machines per core with varying degrees of activity. Includes zEnterprise hardware and z/VM virtualization software. Does not include Linux OS or middleware software.

- Improved availability in the memory subsystem with redundant array of independent memory (RAIM)
- Up to 3 TB of available real memory per server for growing application needs (with up to 1TB real memory per logical partition)
- Just-in-time deployment of capacity resources, which can improve flexibility when making temporary or permanent changes, and plan-ahead memory for nondisruptive memory upgrades
- A 16 GB fixed hardware system area (HSA) that is managed separately from customer-purchased memory
- Exploitation of InfiniBand technology, also introducing enhanced HCA3 fan-outs with up to 40% better service time
- Improvements to the I/O subsystem and new I/O features
- Additional security options for the CP Assist for Cryptographic Function (CPACF)
- A HiperDispatch function for improved efficiencies in hardware and IBM z/OS® software
- Hardware decimal floating point on each core on the processor unit (PU)
- Server Time Protocol (STP) enhancements for time accuracy, availability, recovery, and systems management with message exchanges using ISC-3 or 1x InfiniBand connections

In all, these enhancements provide options for continued growth, continuity, and ability to upgrade.

Models: The z196 is offered in five models: M15, M32, M49, M66, and M80. The model determines the maximum number of processor units (PUs) available for characterization. PUs are delivered in single-engine increments. The first four models use 20-PU multi-chip modules (MCM), of which 15 to 17 PUs are available for characterization. The fifth model, M80, uses 24-PU MCMs to provide a maximum of 80 configurable PUs.

For an in-depth discussion of the zEnterprise 196 functions and features see the *IBM zEnterprise 196 Technical Guide*, SG24-7833.

IBM has made tremendous progress in energy efficiency. The z196 system can now be ordered with a water cooling option, which can result in large savings on energy costs to cool the system. Redundant water cooling units transfer the chilled water from the data center to a closed loop internal water system. The Multichip Modules (MCM) are cooled through cold plates, as the rest of the components are cooled through heat exchangers on the back of the system.

zBX

The zEnterprise BladeCenter Extension (zBX) is the new infrastructure for extending System z governance and management capabilities across a set of integrated, fit-for-purpose POWER7 and IBM System x compute elements in the zEnterprise System. The new zBX Model 003 is configured to attach to the zEnterprise EC12 (zEC12). The zBX Model 002 is configured with either the zEnterprise 196 (z196) or zEnterprise 114 (z114). Attachment to the zEnterprise is via a secure high-performance private network. The zBX can also house the IBM WebSphere® DataPower® Integration Appliance XI50 for zEnterprise (DataPower XI50z).

IBM zEnterprise Unified Resource Manager

All enterprise IT infrastructures today consist of multiple platforms to host the individual components of an end-to-end workload or business process. These platforms are connected but are often not well integrated, flexible, or efficiently aligned to customer service level objectives. The zEnterprise System changes all that. The system delivers end-to-end platform integration and resource optimization with the IBM zEnterprise Unified Resource Manager (zManager). zManager delivers an integrated management fabric that brings consistent, automated and reliable service delivery with a single system view.

zManager brings mainframe governance to the multi-platform environment. It offers automatic sensing and definition of new resources, patch management, dynamic defines, repair verification and monitoring. These capabilities help the installation to make good use of its resources and maintain the highest possible throughput. Best of all, zManager introduces "workload context" that can be used to identify and optimize the physical and virtual system resources that support an application. This allows the zManager to have workload awareness the ability to inspect, report, and manage all connected resources, used in the service of the defined workload. Taking on this management helps to mitigate the risks and provide controls previously not available in the combined mainframe and distributed environment.

With zManager, the system can auto-discover new server, network and storage resources, load the virtualization environments, and prepare system resources for use. Updates to the hypervisor and firmware become the responsibility of the zManager, and can be treated uniformly with a managed rollout. Support and maintenance for the entire system, including the blades, are delivered 24×7 by an IBM System z Support Representative. The blades benefit from the robust System z maintenance management environment with functions like first-failure data capture, "call home," and guided repair.

Application programming interfaces (APIs) allow integration between zManager and the broader ecosystem of management tools. This capability will allow service automation tools to be able to gain access to functions such as discovery, monitoring and provisioning for the heterogeneous resources owned by zEnterprise. The capabilities available with zManager provide value well beyond connectivity for heterogeneous workloads. It brings the ability to manage the IT infrastructure and business applications as an integrated whole.

Helping to manage energy usage

Power and cooling discussions continue to be part of any IT budget planning. As energy prices have risen and utilities have restricted power usage for some, organizations would need to review the role of the server in balancing IT spending. A zEnterprise System can help take better control of energy usage in the data center. Unified Resource Manager will monitor and provide trend reporting of energy efficiency for the entire heterogeneous infrastructure.

A static power savings mode allows for turning off engines that are not being used. The query max potential power mode helps when doing total data center energy use management. There is value in reducing wattage and power across the entire data center, and zEnterprise offers solutions that can help. There is an option for high-voltage DC, which can eliminate the need for a Universal Power Supply (UPS) inverter and Power Distribution Unit (PDU). Top exit I/O cabling can improve flexibility in the data center by helping to increase air flow in a raised-floor environment. The z196 offers a water cooling option that does not increase the system footprint and offers energy savings without compromising performance, and the zBX has an optional rear door heat exchanger to reduce energy consumption.

> **Note:** To learn more about IBM Rear Door Heat eXchanger and its usage with the IBM system z platform, visit this website:
>
> http://www.ibm.biz/Bdx2yH

2.6.3 z/VM

z/VM is key to the software side of virtualization on the mainframe. It provides each user with an individual working environment known as a virtual machine (VM). The virtual machine uses virtualization to simulate the existence of a real machine by sharing resources of a real machine, which include processors, storage, memory, and input/output (I/O) resources. Operating systems and application programs can run in virtual machines as guests. For example, multiple Linux and z/OS images can run on the same z/VM system that is also supporting various applications and users. As a result, development, testing, and production environments can share a single physical platform.

Figure 2-21 shows an example of a configuration of an LPAR with z/VM. A first-level z/VM means that it is the base operating system that is installed directly on top of the real hardware. With a second-level system, a user is brought up in z/VM where an operating system can be executed upon the first-level z/VM. In other words, a first-level z/VM operating system sits directly on the hardware, but the guests of this first-level z/VM system are virtualized. By virtualizing the hardware from the first level, as many guests as needed can be created with a small amount of actual real hardware.

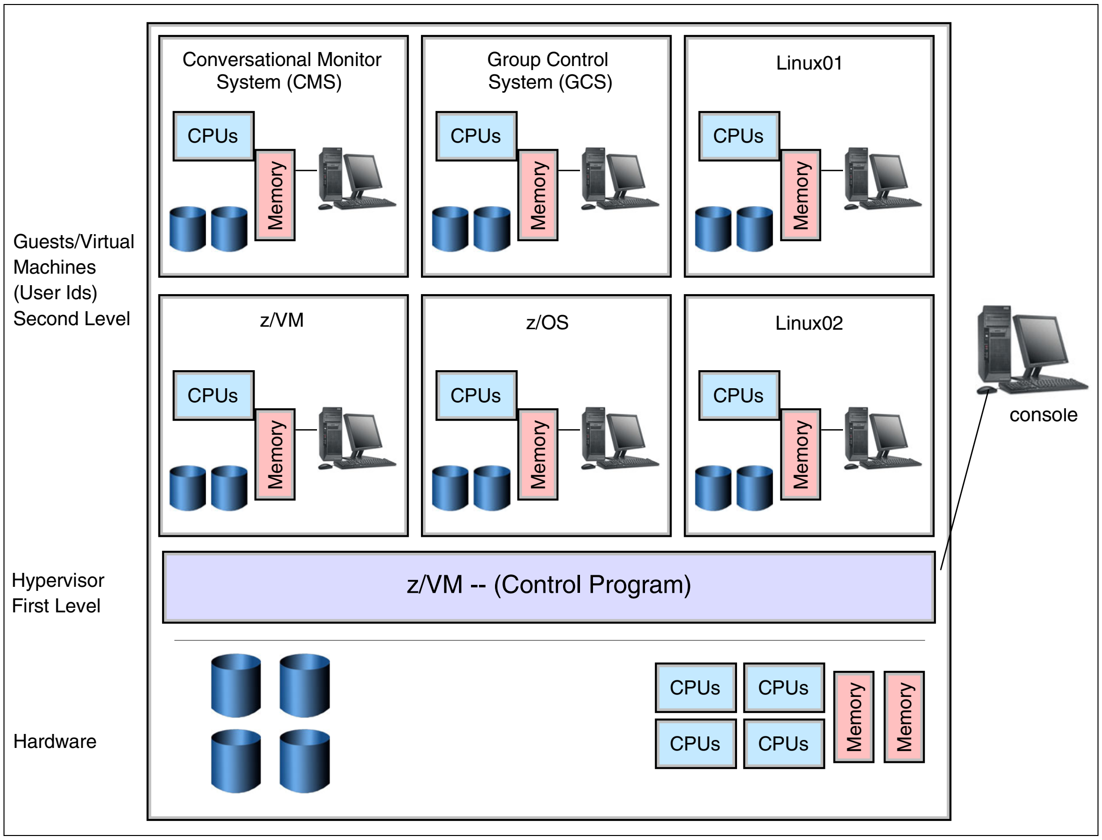

Guests/Virtual Machines (User Ids) Second Level
Conversational Monitor System (CMS)
Group Control System (GCS)
Linux01
z/VM
z/OS
Linux02
CPUs
Memory
console
Hypervisor First Level
z/VM -- (Control Program)
Hardware
CPUs
CPUs
CPUs
CPUs
Memory
Memory

Figure 2-21 Sample configuration for an LPAR

z/VM consists of many components and facilities that bring the reliability, availability, scalability, security, and serviceability characteristics of System z servers, such as these:

- TCP/IP for z/VM brings the power and resources of the mainframe server to the Internet. Using the TCP/IP protocol suite of TCP/IP for z/VM, multiple vendor networking environments can be reached from the z/VM system.

- Applications can be shared transparently across z/VM, Linux, and other environments. Users can send messages, transfer files, share printers, and access remote resources with a broad range of systems from multiple vendors.
- Open Systems Adapter-Express (OSA-Express), Open Systems Adapter Express2 (OSA-Express2), and Open Systems Adapter Express3 (OSA-Express3) are integrated hardware features that enable the System z platform to provide industry-standard connectivity directly to clients on Local Area Networks (LAN) and Wide Area Networks (WAN).
- The Resource Access Control Facility (IBM RACF®) Security Server for z/VM is a security tool that works together with existing functions in the z/VM base system to provide improved data security for an installation. RACF protects information by controlling access to it. RACF also controls what can be done on the operating system and protects the resources. It provides this security by identifying and verifying users, authorizing users to access protected. resources, and recording and reporting access attempts.

Contrasted with a discrete server implementation, z/VM-based System z solutions are designed to provide significant savings, which can help lower total cost of ownership (TCO) for deploying new business and enterprise application workloads on a mainframe.

2.6.4 Linux on System z

With the ability to run Linux on System z, IT departments can simplify their processes by consolidating their multiple server farms down to a single System z running z/VM and Linux on System z. This can lead to both hardware savings and power savings, and could also simplify the management of the infrastructure. Using virtualization, bringing up Linux systems takes a matter of minutes rather than days waiting for new hardware and the installation process to complete. Because each Linux image is running in its own virtual system, the Linux images do not affect the other systems around them.

By leveraging this option, organizations can consolidate hundreds or thousands of Linux instances onto a single IBM system z and considerable amount of operational cost around energy, cooling and the premium data center space.

Tip: Virtually any application that is portable to Linux can be consolidated to Linux on System z.

Integrated Facility for Linux

Integrated Facility for Linux (IFL) is a central processor (CP) that is dedicated to Linux workloads. IFLs are managed by IBM PR/SM™ in logical partitions with dedicated or shared processors. The implementation of an IFL requires a logical partition (LPAR) definition, following the normal LPAR activation procedure. An LPAR defined with an IFL cannot be shared with a general purpose processor. IFLs are supported by z/VM, the Linux operating system and Linux applications, and cannot run other IBM operating systems.

Tip: A Linux workload on the IFL does not result in any increased IBM software charges for the traditional System z operating systems and middleware. For more information on Linux on System z, see *z/VM and Linux Operations for z/OS System Programmers*, SG24-7603.

2.6.5 Tivoli Performance Modeler for z/OS

IBM Tivoli Performance Modeler for z/OS is a systems management tool for performance modeling and capacity planning which runs under Microsoft Windows on a PC and is designed for system management on the IBM System z and IBM System/390®. It can model the performance characteristics for an individual workload or many workloads on z/OS or IBM OS/390®. Systems programmers or operations professionals can use the performance modeling or capacity planning tools to simulate the actual performance behavior of the mainframe computer. They can predict the output of mainframe systems and also the impact of changing hardware or software. It can model up to 40 workloads within a single LPAR, model processors with up to 64 central processors, model the IBM System z9® family of processors, help extract zAAP specific information, help gather, edit, and report on day-to-day workload data, model “what-if” versus actual scenarios, and provide multiple image modeling capability and it includes zIIP support.

Following is a list of scenarios where this tool is useful:

- Help optimize mainframe system resource
- Help deliver optimum service for user satisfaction
- Enter a proactive mode to help with planning and predicting
- Accommodate additional or new workloads
- Help gather and report on day-to-day workload data
- Model “what-if” versus actual scenarios
- Provide multiple image modeling capability
- Help edit workload and LPAR data
- Help plan for future upgrades and other long-term system capacity requirements

2.7 IBM iDataPlex

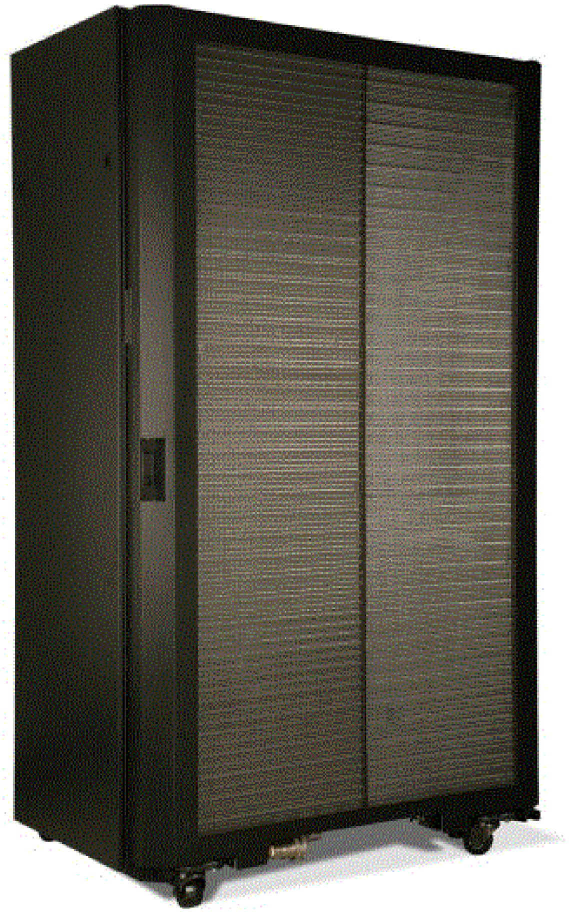
IBM iDataPlex® Rack is built with industry-standard components to create flexible configurations of servers, chassis, and networking switches that integrate easily. Using technology for flexible node configurations, iDataPlex can configure customized solutions for different types of applications to meet specific business needs for computing power, storage intensity, and the right I/O and networking.

In addition to flexibility at the server level, iDataPlex offers flexibility at the rack level. It can be cabled either through the bottom, if it is set on a raised floor, or from the ceiling. Front access cabling and Direct Dock Power allow making changes in the networking, power connections, and storage quickly and easily.

The rack also supports multiple networking topologies including Ethernet, InfiniBand, and Fibre Channel. With the optional IBM Rear Door Heat eXchanger as part of an iDataPlex solution, one can have a high-density data center environment that can alleviate the cooling challenges. With further adjustments, the Rear Door Heat eXchanger can help cool the room, helping reduce the need for air conditioning in the data center:

- Innovative blade-like design saves power and cooling costs for more affordable computing.
- Unparalleled efficiency and easy deployment help organizations get up and running quickly.
- Flexible node and rack configuration enables better customization for expanding data centers.
- Complete front access and intelligent rack components simplify serviceability and manageability.

Figure 2-22 shows the latest innovation, iDataPlex: a flexible and power efficient design that enables massive scale-out deployments with affordable and customized value.



Features:

- Efficient power and cooling
- Flexible, integrated configurations
- New levels of density
- Cost-optimized solution
- Single-point management

IBM Rear Door Heat Exchanger:

- Increase data center density by eliminating hot/cold aisles
- Eliminate rack heat exhaust
- Same dimensions as standard iDataPlex rear door 4" deep
- Liquid cooling at the rack is 75%-95% more efficient than air cooling by a CRAC
- No electrical or moving parts
- No condensation
- Uses Chilled water

Figure 2-22 IBM System x iDataPlex with Rear Door Heat Exchanger

2.8 IBM System x

IBM System x advancements in scalability, reliability, and performance can help organizations to take business where it wants to go. In addition, managing energy in the data center is a growing concern due to the increasing numbers of servers, the incremental heat they generate, and ultimately, the rising cost of energy. IBM System x servers broad portfolio (servers such as x3550-M3, x3650 M3, x3850-X5, x3690-X5 and eX5) not only helps increase performance per watt, but also helps in budget, planning, and control power usage.

And, by consolidating and virtualizing on IBM System x and BladeCenter servers, organizations can increase the utilization of hardware and decrease the number of physical assets on which IT staff need to watch over. With the IBM Director extension and IBM Virtualization Manager, managing the virtualized environment is even easier, a single interface to manage physical and virtual systems from almost anywhere. In this section, we briefly introduce you to IBM System x3690 X5, which is a ground-breaking two-socket server, delivering outstanding performance, memory, and storage.

2.8.1 IBM System x3690 X5

When it comes to business productivity, more is always better, especially when enterprises can achieve it without using more money, space, or energy. Innovative and cost-efficient, the IBM System x3690 X5 delivers four-processor performance, memory capacity, and reliability features in a breakthrough two-processor system that incorporates Intel Xeon processors and fifth-generation IBM X-Architecture® (eX5) technology. The x3690 X5 offers leadership performance, including the ability to better handle database transactions than industry-standard two-processor servers[2].

Figure 2-23 shows the IBM x3690-X5 rack-mount server.

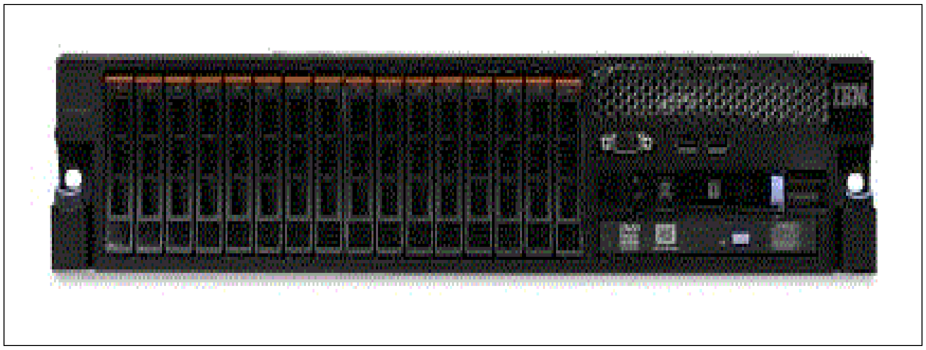


Figure 2-23 IBM x3690-X5 rack-mount server

The x3690 X5 configurations are designed to meet the needs of small and mid-sized business and it features greater flexibility to help organizations changing workload demands as business grows. With MAX5, users can double the memory capacity of two-processor system to 64 DIMM slots. And the x3690 X5 can be tailored with a wide range of options to meet any kind of business needs.

Table 2-10 lists the IBM X3690-X5 systems features and their respective benefits.

Table 2-10 IBM x3690-X5 systems features and their respective benefits

Feature	Benefits
Industry-leading enterprise X-Architecture performance, now available in a 2-processor system	► Greater performance in a denser form factor for a lower purchase price and total cost of ownership ► Ample flexibility to meet changing requirements
Fifth-generation IBM X-Architecture chipset design combined with next-generation Intel Xeon processor E7 product line	► Increased productivity for the investment with higher utilization and greater throughput and bandwidth ► Optimized for databases, enterprise applications and virtualized server workloads ► Deploy more virtual machines than ever in a 2-processor system and double the number of virtual machines on that system with MAX5 ► Higher database transactions processed per minute[2]

[2] Based on IBM comparison between a two-socket IBM System x3690 X5 without MAX5 and a two-socket x3690 X5 with MAX5

Feature	Benefits
Hot-swap, redundant, rear-access power supplies	► Increase system and application uptime by eliminating another single point of failure with a second hot-swap power supply ► Improved usability with rear chassis accessibility, allowing IT team to remove the power supplies without having to pull the system from the rack
Predictive Failure Analysis	► Detects when components or hard drives are operating outside of set thresholds or are approaching historical failure thresholds ► Helps in increase of uptime by enabling the proactive alerts in advance of failure via IBM Systems Director
Hot-swap redundant components	► Allow the server to continue operating in case of fan, power supply or disk drive failure ► Replace components without taking server offline or out of the rack
Integrated dual Gigabit Ethernet	► Provides increased network throughput and redundancy with efficient slot-saving integration ► Optional Emulex 10 GbE Virtual Fabric Adapter
OS flexibility	► Choice of Microsoft Windows 64-bit, Red Hat Enterprise Linux and SUSE Enterprise Linux, (Server and Advanced Server), VMware vSphere Hypervisor
IBM Systems Director and Active Energy Manager (AEM)	► Increased uptime, reduced costs and improved productivity via advanced server management capabilities ► Simplified management including around-the-clock remote management capability ► Helps improve physical and virtual server availability by continuously monitoring the system and notifying of potential system failures or changes ► Advanced power notification, control and monitoring to help achieve lower heat output and cooling costs

2.9 IBM Smarter Storage

In this era of smarter computing, the amount of information that most organizations deal with on a daily basis has exploded by as much as a doubling every 18 to 24 months. At the same time, Big Data projects can quickly double or triple this amount of information. And information is becoming more valuable, with the ability to analyze it counting as a competitive advantage. To stay ahead, organizations need faster access to more data. Another trend, cloud computing, and most of the world wide organizations are either looking for adopting this trend or deploying. And, this growth and transformation is happening at a time when IT budgets are flat. Inflexible infrastructures and increasing complexity can create significant obstacles for organizations trying to take advantage of these tremendous opportunities.

These growing demands require a new approach to storage a more intelligent, more efficient, more automated approach that fundamentally changes the way we think about it.

In this section we introduce you to another IBM building block for next generation data centers, IBM Smarter Storage, a new strategy for storage design and deployment. IBM Smarter Storage enables organizations to take control of their data so they can focus on gaining more valuable insights from this information and delivering more value to the business.

2.9.1 Efficient by design

Efficiency features are designed into IBM storage systems from the beginning. Advanced features are pre-installed, ready to deploy, and operate consistently. Storage efficiency technologies, like real-time data compression, can significantly reduce the amount of data that has to be stored because they support all data, including active primary data.

Self-optimizing

Self-optimizing storage operates at the speed of business much faster and more accurately than manual performance tuning. Self-optimizing storage can increase throughput for critical applications while reducing the cost of storing less active data.

Cloud-agile

Cloud-agile storage has virtualization built in. Virtualization increases agility by enabling online data migration and simplified storage provisioning. Cloud-agile storage enables immediate self-service and automates most manual storage administration tasks, improving the ability to manage larger, more unpredictable workloads.

Through the following sections, we focus on IBM Midrange/Enterprise storage products.

2.10 Disk systems and virtualization

IBM disk storage systems provide storage efficiency solutions such as IBM Real-time Compression™, virtualization, and thin provisioning

2.10.1 IBM System Storage SAN Volume Controller

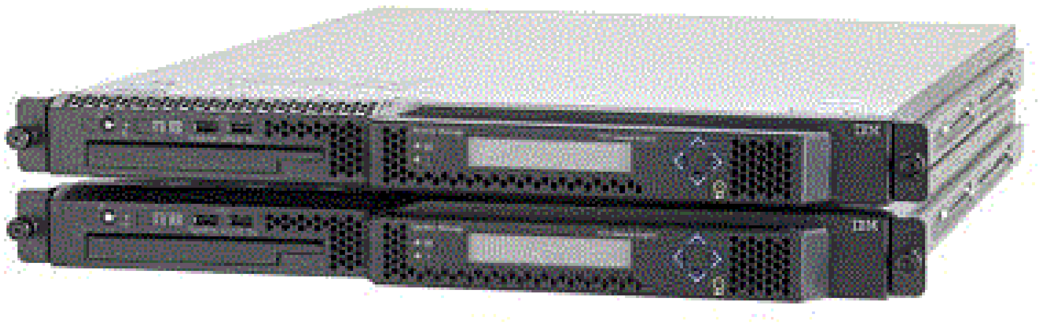
Today’s businesses are facing explosive data growth every day. Every moment or action is a transaction, which creates data that is stored, copied, analyzed, classified, and audited. Therefore, IT infrastructure has a new challenge: Store more with limited or less resources. This challenge means growing capacity without complexity while controlling capital and operational expenses and improving the efficiency of the storage environment. Toward that end, many businesses have pursued strategies such as consolidation, tiering, and virtualization to optimize their resources, simplify the environment, and scale to support information growth. These strategies help to get the most from the storage resources, and achieve a simpler, more scalable and cost- efficient IT infrastructure that aligns more flexibly with the business goals.

Originating at IBM some 40 years ago, virtualization has taken on new life in a variety of contexts: Virtual servers to virtual storage, optimized networks, workstations in virtualized environments and application virtualization. The potential benefits are far reaching, ranging from increased utilization, business flexibility and improved productivity to lower total costs of computing and improved reliability. Depending on the starting point, type and extent of the virtualization implemented, clients can quickly achieve many of these benefits.Storage and server virtualization are complementary technologies that help enables to build a completely virtualized infrastructure. When used together, server and storage virtualization are intended to enable and derive greater benefit from each technology than if deployed alone.

SAN Volume Controller is a storage virtualization system that enables a single point of control for storage resources to help support improved business application availability and greater resource utilization. The objective is to manage storage resources in the IT infrastructure and to make sure they are used to the advantage of the business and do it quickly, efficiently, and in real time, while avoiding increases in administrative costs.

SAN Volume Controller supports attachment to servers using iSCSI protocols over IP networks at 1 Gbps or 10 Gbps speeds, which can help reduce costs and simplify server configuration. SAN Volume Controller also supports FCoE protocol, enabling use of converged data center infrastructures; this new capability simplifies data center and eases management by using the same network for storage, WAN, and LAN.

Figure 2-24 shows the IBM SAN Volume Controller and its highlights.



Highlights

- Manages heterogeneous storage from a single point of control
- Moves data among virtualized storage systems without disruptions
- Stores up to five times1 as much active data in the same physical disk space using IBM Real-time Compression
- Simplifies storage infrastructure with support for Fibre Channel over Ethernet (FCoE) protocol
- Optimizes solid-state storage deployments automatically with IBM System Storage Easy Tier
- Allows for non-disruptive scalability from the smallest configuration to the largest
- Implements stretched configurations for high availability and data mobility between data centers

Figure 2-24 IBM SAN Volume Controller with its highlights

SAN Volume Controller systems with 10 Gbps Ethernet ports now support attachment to next-generation Converged Enhanced Ethernet networks using Fibre Channel over Ethernet. This support enables customers to connect SAN Volume Controller to servers and to other SAN Volume Controller systems for clustering or mirroring using Fibre Channel or FCoE interfaces using these networks. The same ports may also be used for iSCSI server connections.

Improve energy efficiency

Many data centers today are focusing on reducing their energy usage to reduce costs and demonstrate concern for the environment. SAN Volume Controller can be a key tool to improve the energy efficiency of data centers. This solution can help in improving the energy efficiency in several ways. For example, IBM Real-time Compression with SAN Volume Controller can help by significantly increasing the effective capacity of storage and reduce requirements for additional storage in the future. This can help reduce the total amount of physical storage required by up to 80 percent, helping reduce energy use. The thin provisioning and snapshot functions are designed to extend this benefit even further.

Simplify management

SAN Volume Controller uses a completely new graphical user interface modeled on the IBM XIV® Storage System, which has been very well received by customers. The user interface is designed to be easy to use and includes many built-in IBM recommendations to help simplify storage provisioning so new users can get started quickly. At the same time, the new interface preserves access to all the functional richness that experienced users have come to expect from SAN Volume Controller.

2.10.2 IBM Storwize V7000 and Storwize V7000 Unified

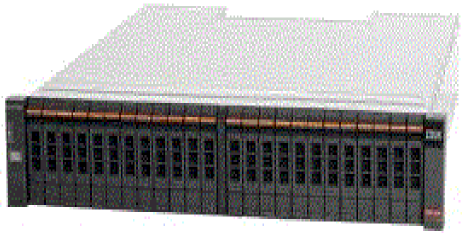
To stand up to challenges and allow businesses to respond to a rapidly changing marketplace, IBM Storwize® V7000 and Storwize V7000 Unified are virtualized storage systems to complement virtualized server environments. They provide unmatched performance, availability, advanced functions, and highly-scalable capacity never seen before in midrange disk systems. As members of the IBM Storwize family, Storwize V7000 and Storwize V7000 Unified are powerful midrange disk systems that have been designed to be easy to use and to enable rapid deployment without additional resources. Storwize V7000 supports block workloads, whereas Storwize V7000 Unified consolidates block and file workloads into a single storage system for simplicity of management and reduced cost.

Storwize V7000 and Storwize V7000 Unified are virtual storage systems that offer greater efficiency and flexibility through built-in SSD optimization and thin-provisioning technologies. Integrated IBM Real-time Compression enhances efficiency even further by enabling storing up to five times as much active primary data in the same physical disk space[3]. Storwize V7000 and Storwize V7000 Unified advanced functions also enable nondisruptive migration of data from existing storage, simplifying implementation and minimizing disruption to users. Finally, these systems also enable organizations to virtualize and reuse existing disk systems, supporting a greater potential return on investment (ROI).

The Storwize V7000 system is packaged in 2U rack-mountable enclosures that house up to twenty-four 2.5-inch drives or up to twelve 3.5-inch drives. Control enclosures contain drives, redundant dual-active intelligent controllers and dual power supplies, batteries and cooling components. Expansion enclosures contain drives, switches, power supplies, and cooling components. And, they can attach up to nine expansion enclosures to a control enclosure supporting up to 240 drives. Four control enclosures (each with up to nine expansion enclosures) can be clustered together in a single system for even greater capacity and performance growth potential. Other components and characteristics of the system include:

- Internal storage capacity: Up to 36 TB of physical storage per enclosure using twelve 3 TB near line SAS disk drives or up to 24 TB of physical storage per enclosure using twenty-four 2.5-inch 1 TB near-line SAS disk drives
- Disk drives: SAS disk drives, near-line SAS disk drives, and SSDs; intermix of these drive types within the Storwize V7000 control and expansion enclosures adds flexibility
- Cache memory: 16 GB cache memory (8 GB per internal controller) as a base feature designed to improve performance and availability

[3] IBM lab measurements



Highlights

- Delivers sophisticated, enterprise-class storage functionality for businesses
- Supports growing business requirements while controlling costs
- Provides up to 200 percent performance improvement with automatic migration to high-performing solid-state drives (SSDs)[4]
- Enables storing up to five times more active primary data in the same physical disk space using IBM Real-time Compression[4]
- Consolidates block and file storage for simplicity, greater efficiency and ease of management
- Enables near-continuous availability of applications through dynamic migration
- Easy-to-use data management designed with a graphical user interface and point-and-click system management capabilities

Figure 2-25 IBM Storwize with its highlights

Next-generation networking with IBM Storwize

As organizations evolve towards a dynamic infrastructure, they need new ways to reduce the complexity of their environments. To address this challenge, clients are turning to Converged Enhanced Ethernet (CEE) networks, which help enable them to combine storage, messaging traffic, VoIP, video, and other data on a common data center Ethernet infrastructure.In this environment, Fibre Channel over Ethernet (FCoE) helps enable highly efficient block storage over Ethernet for consolidating server network connectivity.

As a result, IT staff can deploy a single server interface for multiple data types, which can simplify both deployment and management of server net-work connectivity, while maintaining the high availability and robustness required for storage transactions.Storwize V7000 and Storwize V7000 Unified systems with 10 Gbps Ethernet ports now support attachment to next- generation CEE networks using FCoE. This support enables the connecting the systems to servers and to other Storwize V7000 or Storwize V7000 Unified systems for clustering or mirroring using Fibre Channel or FCoE interfaces using these networks. The same ports may also be used for iSCSI server connections.

Integrated management

Storwize V7000 and Storwize V7000 Unified provide a tiered approach to management designed to meet the diverse needs of different organizations. The system's management interface is designed to give administrators intuitive control of the system and provides a single, integrated approach for managing both block and file storage requirements in the same system.

A recent study proved the effectiveness of the user interface, finding that tasks are 47 percent less time-consuming and 31 percent less complex than managing a competitor's system. For organizations looking to manage both physical and virtual server infrastructures and the storage they consume (including provisioning and monitoring for higher availability, operational efficiency and infrastructure planning), Storwize V7000 and Storwize V7000 Unified are integrated with IBM Systems Director Storage Control and IBM Flex System Manager. A single administrator can manage and operate IBM servers.

For more information on IBM Storwize implementation, see *Implementing the IBM Storwize V7000 Unified*, SG24-8010 and *Implementing the IBM Storwize V7000 V6.3*, SG24-7938.

2.10.3 XIV

IBM XIV is a high-end disk storage system helping thousands of enterprises meet the challenge of data growth with hotspot-free performance and game-changing ease of use. Simple scaling, high service levels for dynamic, heterogeneous workloads, and tight integration with hypervisors and the OpenStack platform enable optimal storage agility for cloud environments.

Figure 2-26 shows the front side of the IBM XIV and its highlights.

Born optimized with inherent efficiencies that simplify storage, XIV delivers the benefits of IBM Smarter Storage for Smarter Computing, empowering organizations to take control of their storage and to extract more valuable insights from their data. XIV extends ease of use with integrated management for large and multi-site XIV deployments, reducing operational complexity and enhancing capacity planning.

Highlights

- A revolutionary, proven, high-end disk storage system designed for data growth and unmatched ease of use
- Consistent high performance without hotspots, enabled through massive parallelism and self-tuning
- Extra performance boost option through management-free solid-state drive (SSD) caching
- High reliability and availability via full redundancy, self-healing and unprecedented rebuild speed
- Low total cost of ownership (TCO) enabled by high-density storage, simplified planning, cost-free features and low-touch management
- Virtualized storage, easy provisioning and extreme agility for optimized virtual environments and cloud services

Figure 2-26 IBM XIV inside and highlights

The XIV Storage System provides fully virtualized storage designed to eliminate the need for performance tuning, planning for capacity and performance growth, and numerous other storage management activities. The highly intuitive XIV graphical user interface (GUI) and built-in management tools make administrative tasks easy and efficient, with little training or expertise required, from provisioning volumes to monitoring multiple systems. XIV Multi-System Manager enables one-stop, centralized, and integrated administration of XIV systems across the enterprise, driving down operational complexity and simplifying capacity management. A powerful command line interface (CLI) supports complex scripting. The exceptional flexibility of the XIV Storage System extends to mobile devices and tablets with the XIV Mobile Dashboard, empowering IT staff with on-the-go monitoring of system performance, capacity, and health from their smart phone[4].

Grid: XIV storage is ideally suited for cloud storage due to its grid architecture, which features a deep affinity for dynamic, heterogeneous workloads; eliminates hot spots; and removes the need for manual tuning, which is typically complex and time-consuming.

2.10.4 IBM Scale Out Network Attached Storage

The demand to manage and store massive amounts of file data continues to challenge data centers. IBM Scale Out Network Attached Storage (SONAS) is designed to embrace and deliver cloud storage in the petabyte age. SONAS can meet today's and next generations storage challenges with quick and cost-effective IT-enabled business enhancements designed to grow with unprecedented scale. SONAS can also deliver storage services that make the supporting technology almost invisible. It allows applications and services to be uncoupled from the underlying infrastructure, enabling businesses to adjust to change quickly. As a result, SONAS can easily integrate with organization's strategies to develop a more dynamic enterprise.

SONAS is an ideal solution for cloud storage implementations as well. It is designed to consolidate files that are scattered in multiple storage locations and allow them to be efficiently shared and managed. The IBM Active Cloud Engine™ is a unique capability of SONAS that enables this and many more management abilities. The IBM Active Cloud Engine is a suite of capabilities specifically designed to manage files in an automated, scalable manner. It creates the appearance of a single system despite geographic, physical, media or other discrepancies that may exist in the physical world. It is designed to put the right file in the right place at the right time, and to give users the fastest possible access with the same view of their data no matter where they are. It enables ubiquitous access to files from across the globe quickly and cost effectively.

The IBM Active Cloud Engine offers clients not only a high-performance file serving function, but also reduced network costs. It localizes file data where it is needed. It can pre-populate files to remote sites in advance of that file request coming in, which allows for high availability and very fast access of those files at remote sites. This capability allows for file sharing and content distribution while it shields applications from fluctuating wide area network (WAN) latencies (performance fluctuations and outages during data access across a WAN). In addition, it eliminates unnecessary file replication to remote sites, thereby significantly lowering network and storage costs.

When organizations are managing millions to billions of files, the protection and availability of these files are critical. SONAS is a reliable, scalable storage solution with built-in high availability and redundancy. The IBM Active Cloud Engine within SONAS provides automated policy-based file management that enables backups, restores, snapshots and remote replication. SONAS has a specific exploitation and integration with IBM Tivoli Storage Manager to provide very efficient and extremely fast backup and restore processes, and enables movement of files to an external disk or tape.

[4] Currently Apple iPhone and iPad are supported.

In addition, SONAS provides support for the Network Data Management Protocol (NDMP) “at scale” to provide both full and incremental file backup as well as restoration of these files and related file system data if needed. Support for the NDMP enables SONAS to be backed up with independent software vendor (ISV) backup applications like Symantec, CommVault, Legato and others over the local area network. The SONAS policy engine increases the efficiency of backing up billions of files by scanning their metadata in just minutes.

SONAS provides central deployment and management of automated tiered storage, with high performance and high storage utilization. Based on a Standard Performance Evaluation Corporation System File Server benchmark using a realistic, standard configuration, SONAS demonstrates excellent near-linear performance for a single file system without compromising capacity or cost.

Protecting business information from malware and viruses is a business imperative. Integration with antivirus software is designed to provide the ability to isolate or delete compromised files, and protect data from risks due to viruses that can potentially compromise SONAS data.

SONAS provides ease of use for administrators through an intuitive graphical user interface (GUI). This user interface is similar to those used in XIV or Storwize V7000 systems. The highly intuitive GUI and built-in policy-driven management tools through IBM Active Cloud Engine make administrative tasks easy and efficient. Administrators can also leverage IBM Tivoli Productivity Center version 5.1 to manage SONAS, providing discovery and visualization, health status monitoring and capacity usage reporting. The new release of Tivoli Productivity Center 5.1 also enables applications to perform SONAS provisioning,

2.10.5 IBM DS8000 Series

Virtually every organization struggles to manage the exponential growth of stored data and the increasing complexity of its IT environments. Managing this extraordinary growth is complicated by the variety of server environments spread across data centers, and each server environment can have a distinct storage system that must be managed separately. As the number of servers grows, managing these disparate storage systems can significantly increase storage costs.

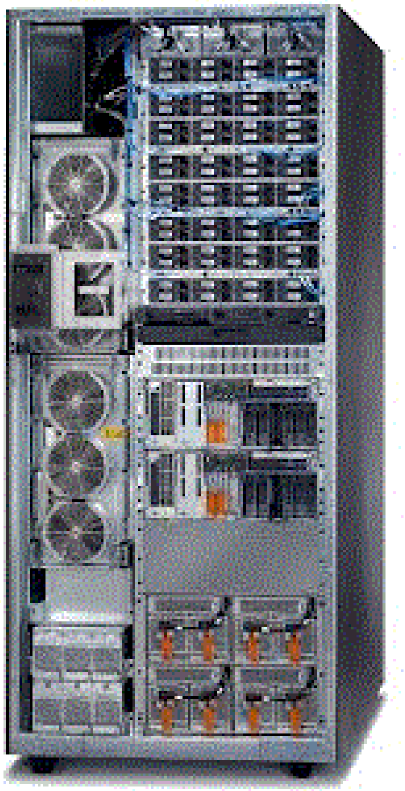
IBM Smarter Storage, such as the IBM System Storage® IBM DS8000® series, enables organizations to take control of their storage so they can focus on gaining more valuable insights from their data and delivering more value to their business.

Through its extraordinary flexibility, the DS8000 Series is designed to effectively and efficiently manage a broad range of storage workloads that exist in today’s complex data centers. This flagship IBM disk system can bring simplicity to storage environment by supporting a mix of random and sequential I/O workloads as well as interactive and batch applications, regardless of whether they are running on one of today’s popular distributed server platforms or on the mainframe. With the ability to manage all these workloads simultaneously, this dynamically self-tuning system can greatly reduce the challenge of managing the increasingly complex storage environment as requirements are getting changed.

The DS8000 Series delivers these capabilities with a range of features that provide self-tuning performance through auto-mated storage tiering, automated quality-of-service (QoS) management, as well as broad server support and outstanding scalability. The I/O Priority Manager feature, for example, can dynamically manage the performance of the various applications running on the system through a very easy and intuitive approach. This is especially helpful when organizations want to consolidate storage workloads and at the same time, need to ensure that system resources are aligned to the priority of the applications.

These advanced features can help businesses to manage more data for more applications in a smaller footprint, without the constant oversight and manual performance tuning that other high-end storage vendors require.

Figure 2-27 shows highlights of the IBM DS8000.



Highlights

- Provides extraordinary system availability with full hardware redundancy built on the market-proven IBM Power Systems architecture.
- Dynamically optimizes performance for single- and multi-tiered systems with IBM System Storage IBM Easy Tier® and other self-tuning features.
- Enables systems to scale to more than 2.3 petabytes (PB) and support multiple drive tiers.
- Optimizes the broad scope of storage workloads that exist in today's complex data centers.
- The new IBM System Storage DS8870 delivers up to three times performance improvement by leveraging IBM POWER7 technology and up to 1 terabyte (TB) system cache.

Figure 2-27 IBM DS8000 internals and highlights

The DS8000 Series includes powerful management capabilities that can help IT administrators more effectively control their storage environments as capacity grows. The DS8000 graphical user interface (GUI) enables easy and effective management of the system and its advanced features. Inspired by the acclaimed IBM XIV Storage System GUI, the DS8000 GUI includes redesigned navigation for more intuitive use, streamlined configuration processes, and embedded links to relevant videos for quick access to useful DS8000 Series information. In addition, it provides dynamic and customizable views, as well as interactive menus, tabs, charts, and more.

IBM Tivoli Storage Productivity Center has also adopted the advanced IBM storage administration interface and provides advanced capabilities to manage the entire storage environment by supporting a variety of storage systems and devices from IBM and other vendors. Clients have the option of deploying the comprehensive Tivoli Storage Productivity Center Standard Edition or IBM System Storage Productivity Center appliance, which includes Tivoli Storage Productivity Center Basic Edition software. Both versions provide single sign-on capability for the various devices they manage and a comprehensive view of the storage topology that enables administrators to inspect the real-time health of the environment at an aggregate or in-depth view.

2.11 Tape systems

IBM invented the concept of magnetic tape computer storage back in 1952. Since then, IBM has delivered many innovations in tape storage, and that innovation continues today to solve, for the next generation of data centers, more mysteries surrounding tape storage.

2.11.1 TS1140

Business demands require applications that rely on high capacity and fast access to data as well as long-term data retention. The IBM System Storage TS1140 Tape Drive (TS1140 tape drive) features storage capabilities to help organizations to establish easy, rapid access to data, improve security, provide long-term retention, and help maintain data governance and regulatory compliance. The TS1140 tape drive offers high-performance flexible data storage with support for data encryption. For support of the heterogeneous server environment, the TS1140 offers multi-platform support and can scale from midrange to enterprise environments. The TS1140 tape drive supports the System Storage TS3500 Tape Library and IBM racks that enable stand-alone installation.

The TS1140 features three options for Type C media. The 3592 Advanced data tape cartridge, JC, provides up to 4.0 TB native capacity, and up to4.0 TB are provided by the 3592 Advanced WORM cartridge, JY. A limited capacity of up to 500 GB Economy cartridge, JK, offers fast access to data. The TS1140 tape drive can also read and write on previous media, type B (JB and JX), and read only on type A (JA, JW, JJ and JR).

To help optimize drive utilization and reduce infrastructure requirements, the TS1140 tape drives can be shared among supported open system hosts on a Storage Area Network (SAN) or between IBM FICON mainframe hosts when attached to a System Storage Tape Controller Model C07 for IBM System z.

Table 2-11 explains the features and benefits of the IBM TS1140.

Table 2-11 IBM TS1140 features and benefits

Feature	Benefits
Up to 4 TB native capacity on a 3592 JC/JY cartridge	► High capacity helps reduce the need for additional tape drives and cartridges ► Use fewer drives to reduce complexity of tape infrastructure
Data transfer rate up to 650 MBps with compression	► Improves performance of tape operations to help maintain business performance
3592 media	► Media options include short and long lengths, re-writable and WORM formats and media partitioning ► Flexible ordering options
Supports encryption and key management	► Supports a single encryption key management approach, which may help reduce audit and compliance costs ► Helps to avoid the need for host-based encryption of data or the use of specialized encryption appliances
Media partitioning	► Improves data management and can provide faster data access. ► Allows updating tape partitions independently
Small form factor	► Improves space efficiency of the tape infrastructure

Feature	Benefits
Virtual backhitch and high resolution directory	► Improves access to data ► Reduces wear and tear on media
Compatibility with existing IBM tape automation installations	► Protects the existing tape investments while improving performance with next generation technology
IBM Power Systems, IBM System i®, IBM System p®, System z and System x support	► Supports the heterogeneous infrastructure ► Supports flexibility for growth as well as change in the data centers

2.11.2 TS3500

The IBM System Storage TS3500 Tape Library is designed to provide a highly-scalable, automated tape library for mainframe and open systems backup and archive that can scale from midsize to large enterprise environments and it continues to lead the industry in tape drive integration with features such as persistent worldwide name, multipath architecture, drive/media exception reporting, remote drive/media management and host-based path failover.

The L23 and D23 frames support the IBM System Storage TS1140, TS1130, TS1120 or IBM 3592 J1A tape drives. The L53 and D53 frames support LTO Ultrium 6 tape drives as well as previous-generation LTO Ultrium tape drives. L-frame models support improved cartridge handling, hot-swappable drive packaging and the option of an additional 16-slot I/O station. The TS3500 model D23 and D53 frames can be attached to existing model L22 or D52 frames. Mixed media is supported by combining LTO Ultrium tape drives and the TS1140, TS1130, TS1120 or 3592 J1A tape drives within the TS3500 library, frame by frame.

It supports IBM System z when used with the IBM 3953 Tape System, the IBM Virtualization Engine TS7740 or the IBM System Storage Tape Controller for System z with its embedded library manager. These systems enable System z hosts to access the TS3500 Tape Library cartridge inventory and allow connection to TS1140, TS1130, TS1120 and IBM 3592 Model J1A tape drives.

The TS3500 Tape Library can support up to four 3953 tape systems, up to eight Virtualization Engine TS7740 subsystems per physical library, and up to 16 System Storage tape controllers for System z per logical library. The System Storage Tape Controller for System z offers IBM FICON attachments of TS1140 tape drives in a TS3500 Tape Library or rack, and reduces hardware and infrastructure requirements by sharing tape drives with FICON hosts. It can also add tape drives non-disruptively, which helps enhance configuration flexibility and availability.

Figure 2-28 shows front view (opened) of the IBM TS3500 Tape Library and its highlights.

Highlights

- Supports highly-scalable, automated data retention on tape using Linear Tape-Open (LTO) Ultrium and IBM 3592 tape drive families
- Delivers extreme scalability and capacity, growing from one to 16 frames per library and from one to 15 libraries per library complex
- Provides up to 900 PB of automated, low-cost storage under a single library image, improving floor space utilization and reducing storage cost per TB with IBM 3592 JC Enterprise Advanced Data Cartridges
- Enables data security and regulatory compliance via support for tape drive encryption and write-once-read-many (WORM) cartridges

Figure 2-28 The IBM TS3500 front view and highlights

Features next-generation tape-drive support

TS1140 tape drives support IBM 3592 JC and JY advanced, standard and WORM cartridges that feature four times more capacity than the previous generation up to 4 TB native or 12 TB compressed with more than 50 percent faster data transfer rates at 250 megabits per second (Mbps) native, TS1140 tape drives also support the IBM 3592 JK economy cartridge with a capacity of up to 500 gigabytes (GB). TS1140 tape drives can read and write IBM 3592 JB and JX cartridges with improved performance.LTO Ultrium 6 2.5 TB cartridges (standard and WORM) and LTO Ultrium 6 tape drives are supported and designed to provide 67 percent more tape cartridge capacity and greater than 12.5 percent improvement in tape drive performance over the fifth-generation LTO Ultrium drives.

They can also support media partitioning and the IBM Linear Tape File System™ to help improve data management on tape media. The sixth-generation LTO Ultrium is designed to support up to 160 Mbps native data transfer rates. LTO Ultrium 6 tape drives can read and write LTO Ultrium 5 cartridges and read LTO Ultrium 4 cartridges at original capacities with improved performance. The TS1140 and LTO Ultrium 6 tape drives and cartridges (standard or WORM) can reside in the same TS3500 Tape Library string with previous generations of 3592 or LTO Ultrium tape drives and cartridges to help protect existing investments.

Integrates powerful software support

The power of the TS3500 Tape Library can be enhanced by integrating it with industry-leading storage management solutions, such as IBM Tivoli Storage Manager and a wide array of other storage software.

2.11.3 TS3310 Tape Library

The IBM System Storage TS3310 Tape Library is a modular scalable tape library designed to address the tape storage needs of rapidly growing companies who find themselves space constrained and resource constrained with tape backup and other tape applications.

Designed around a 5U high modular base library unit, the TS3310 can scale vertically with expansion for Linear Tape-Open (LTO) cartridges, drives and redundant power supplies. The base library module, model L5B, is the entry point for the product family. It contains all of the necessary robotics and intelligence to manage the 5U high library system, which houses up to 41 cartridges 35 storage slots and six input/output (I/O) slots; and two LTO generation 5, generation 4 and generation 3 tape drives. The TS3310 model L5B can be expanded with the addition of expansion units, the model E9U.

Highlights

- Provides a modular, scalable tape library designed to grow as business needs grow
- Features desktop, desk-side and rack-mounted configurations
- Delivers optimal data storage efficiency with high cartridge density using standard or WORM LTO data cartridges
- Offers hot-swap tape drives and power supplies
- Includes redundant power and host path connectivity failover options
- Enables remote web-based management and Storage Management Initiative Specification (SMI-S) interface capabilities

Figure 2-29 IBM TS3310 Tape library with its highlights

Capacity on demand: TS3310 Tape Library's Capacity on Demand (CoD) capability allows the system to scale as needs grow. In the initial shipped configuration, an E9U has half of its storage cells enabled.

Building on the success of the IBM patented multipath architecture, the TS3310 can be divided into one logical library per installed tape drive. These logical libraries can be simultaneously connected to a wide variety of different servers, running different operating systems and tape applications.Designed for the ultimate in system availability, the optional data path and control path feature can help support ongoing host connectivity under a variety of adverse conditions.

To help measure and evaluate media and drive performance in the library, an advanced reporting feature can be activated by license key. This feature enables drive resource utilization reporting and media integrity analysis reporting functions. These user-configurable reporting functions provide graphical displays for both diagnostic and trend analysis data to help evaluate tape drive as well as media performance and drive utilization parameters. The drive resource utilization report can help identify drives or groups of drives that are nearing100 percent utilization. It will also help identify drive resources that are not being fully utilized and provide information necessary to determine if additional drives are necessary.

2.11.4 TS7700: Enterprise Virtual Tape Library

Tape is a key part of both the backup and archive life cycle of data. Tape is the storage option with the lowest total cost of ownership (TCO) for protecting data, for securely backing up and storing data both short term and for long-term archives, and for record keeping and disaster recovery. As the need for data storage grows, data protection can become more complex, involving continuous backup, multiple sites, off-site recovery, disaster recovery testing and many other processes to insure enterprise data is protected. This growth can lead to increased backup and restore times as well as higher data management overhead and costs.

The IBM TS7700 Virtualization Engine is a family of mainframe virtual tape solutions that optimize data protection and business continuance for System z data. Through the use of virtualization and disk cache, the TS7700 family operates at disk speeds while maintaining compatibility with existing tape operations. The TS7700 family of System z tape virtualization is offered in two models: the TS7720, which uses only virtual tape drives and tape cartridge volumes, and the TS7740, which not only stores the data on virtual tape drives, but also writes data by policy to physical tape. This fully integrated tiered storage hierarchy of disk and tape takes advantage of both tape and disk technologies to deliver performance for active data and best economics for inactive data. With the TS7740, tape and disk are combined to provide a flexible, policy-managed configuration that offers the benefits of both types of storage. Deploying these innovative models can help reduce batch processing time, TCO and management overhead.

The TS7740 Virtualization Engine attaches to and leverages the performance and capacity of the IBM System Storage TS1140, TS1130 and TS1120 tape drives or the IBM TotalStorage 3592 Model J1A Tape Drive installed in an IBM System Storage TS3500 Tape Library. Support for the TS1140 Tape Drive can reduce the number of cartridges and the size of the library by supporting the storage of up to 12 TB on a single 3592 JC cartridge, (assuming 3:1 compression). The TS7720 Virtualization Engine provides high capacity for workloads that are not required to additionally be stored on physical tape. The TS7720 Virtualization Engine features high capacity cache using 3-TB SATA disk drives with RAID 6 to allow clients to scale their solution to meet the needs of growing workloads without affecting application availability. The TS7700 Virtualization Engine family has been designed to help enhance performance and provide the capacity needed for today’s tape processing requirements.

Ensuring data security and regulatory compliance

Designed to keep data more secure and help meet regulatory guidelines, the TS7700 Virtualization Engine provides end-to-end data encryption. Encryption starts from the point where data is stored in the TS7700 system cache, goes across IP connections to remote TS7700 systems (through Internet Protocol Security), and ends with data stored in the cache of other nodes (TS7720 and TS7740) or tape cartridges (TS7740 only). With this feature, the TS7700 Virtualization Engine offers industry-standard encryption of data transferred between TS7700 systems. To help keep information confidential if backup tapes are lost or compromised, the TS7740 Virtualization Engine supports TS1140, TS1130 and TS1120 tape drive encryption capabilities.

The TS1140, TS1130, and TS1120 tape drives include data encryption capabilities in the drives, helping to avoid the need for host-based encryption of data and the concurrent drain on host performance and resources or the additional expense of specialized encryption appliances. The IBM Tivoli Key Lifecycle Manager can generate and manage encryption keys for TS1140, TS1130 and TS1120 tape drives across the enterprise. Tivoli Key Lifecycle Manager delivers advanced, federated, cross-domain key management designed to help lock down organizational data more comprehensively and easily than ever. To help support the long-term retention of reference data and meet the requirements of regulatory bodies worldwide, microcode capabilities enable the TS7700 Virtualization Engine to support a virtual equivalent of write-once-read-many (WORM) tape media.

2.12 IBM Real-time Compression

The newest IBM storage efficiency technology is IBM Real-time Compression. Available in both block and file-based storage arrays, this state of the art innovation is designed to improve efficiency by compressing data as much as 80 percent, enables the user to store up to five times as much data in the same physical disk space. Unlike other approaches to compression, IBM Real-time Compression is designed to be used with active primary data such as production databases and e-mail applications, which dramatically expands the range of candidate data that can benefit from compression. As its name implies, IBM Real-time Compression operates immediately as data is written to disk so no space is wasted storing uncompressed data awaiting post-processing. The benefits are described in Table 2-12.

Table 2-12 Benefits of IBM Real-time Compression

Efficient by design	Just turn it on, it is 100% transparent for systems, storage, and applications. Delivers reduced storage requirements for online, active primary data with no application performance degradation.
Patented technology	Able to compress data when it is first stored so it can send less physical data to storage systems. Transparent to business applications and IT processes, it works with any data type that is not already compressed even complementing deduplication, virtualization and high availability solutions.
Real-time efficient operation	Supports the performance and accessibility requirements of business-critical applications because data is compressed up to 80% in real time, allows to store up to 5 times more data in the same physical space.
Transparency	RtC is 100% transparent to systems, storage, and applications.
Less Cost, Greener	Cost reduction benefits carry throughout the storage lifecycle: less storage to power, cool, and manage means less cost and a greener data center.

2.12.1 IBM Real-time Compression Appliance STN6800 and STN6500

IBM Real-time Compression Appliance™ STN6800 and STN 6500 both effectively increases the capacity of the existing storage infrastructure to meet the demands of rapid data growth while also enhancing storage performance and utilization. Both models have a two rack unit form factor and support the following capabilities. See Figure 2-30.

Management software

- Intuitive web GUI
- Command line interface (CLI) for management tasks
- Comprehensive SNMP MIB provides statistics information and alerts
- Active Directory integration supports external Syslog server for sending notifications and audit information

High availability

- Transparent path failover, when deployed in pairs
- Predictive Failure Analysis for hardware components
- Link aggregation (IEEE 802.3ad)
- Ethernet trunking (Cisco EtherChannel)

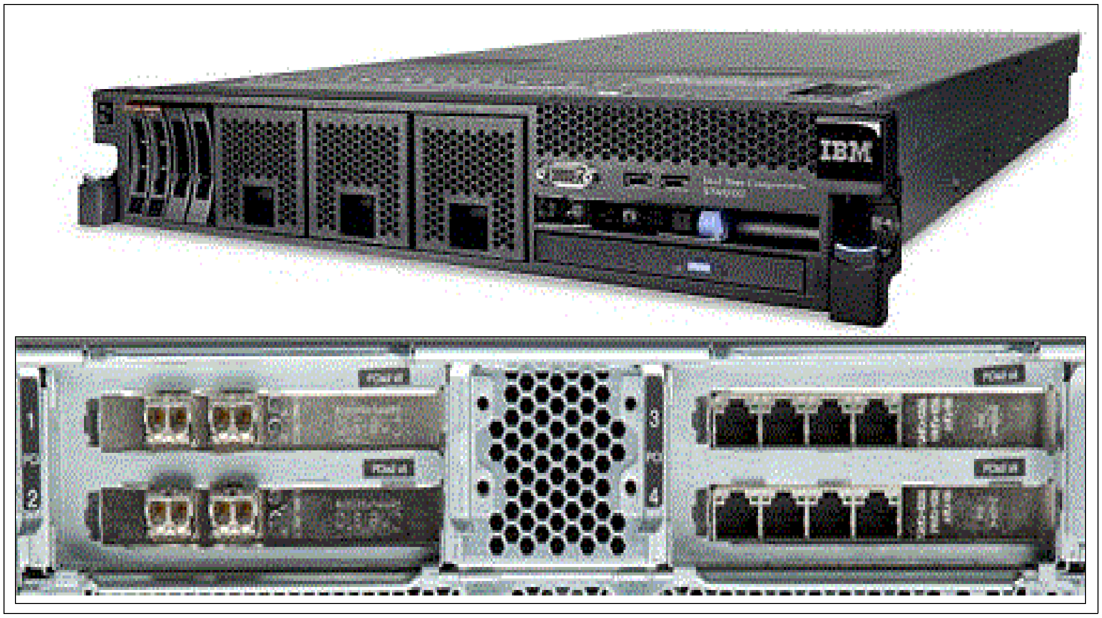


Figure 2-30 IBM Real-time Compression Appliance STN 6800 front and back view

From the downstream operational efficiency perspective, both appliance models deliver the following benefits.

- Shortened backup/restore windows
- Reduced data replication bandwidth and cost
- Improved efficiency for organizations using data deduplication

Note: Appliance model STN6800 has connectivity option of 10 GbE and 1 GbE, while appliance model STN6500 supports only 1 GbE connectivity.

Both appliance models are enterprise ready models. STN6500 has sixteen 1 Gb Ethernet ports, which supports eight connections between NAS systems and network switches and STN6800 supports multiple 10 GbE and 1 GbE connections for high throughput environments. The appliance design offloads compression processing from both application servers and storage systems, helps in improving efficiency and performance at the same time. Both appliance sits between NAS systems using UNIX (NFS) and Microsoft Windows (CIFS) protocols and the network switches, for completely transparent operations.

2.13 IBM Tivoli Storage Productivity Center

As the amount of data that IT staff needs to store and retain grows exponentially, IT teams must find better ways to control the cost of storage. In addition, managing the storage infrastructure has become increasingly complex as businesses continue to acquire new storage infrastructures or inherit a mix of multivendor storage assets due to acquisitions or company mergers. You need to be able to identify, evaluate, control and predict the growth of data through its life cycle in order to meet storage service levels in accordance with IT Infrastructure Library (ITIL) and data retention requirements. Both requirements, managing storage infrastructure and the data that resides there, are highly labor intensive. Tivoli Storage Productivity Center is a storage infrastructure management tool that is easy to deploy and use, and can help in reducing the complexity of managing the storage environments by centralizing, simplifying, and automating storage tasks associated with storage systems, storage networks, replication services, and capacity management.

Tivoli Storage Productivity Center is a comprehensive, end- to-end storage infrastructure management software solution that provides an easy-to-use, next-generation user interface to monitor, report, and manage storage capacity, storage networks, storage systems, and replication services. The Tivoli Storage Productivity Center family of products includes Tivoli Storage Productivity Center and Tivoli Storage Productivity Center Select. Tivoli Storage Productivity Center provides a single point of control to manage every aspect of the storage network, between the hosts, through the fabric, to the physical disks in a multi-site storage environment. Tivoli Storage Productivity Center is designed to provide the foundation for storage service-level management by offering storage system and SAN performance and availability management.

Tivoli Storage Productivity Center is designed to help control and reduce total cost of ownership in the following ways:

- Improving storage performance
- Reducing administrative costs through centralized management of heterogeneous storage
- Significantly reducing planned downtime
- Configuring storage systems the best-practices way
- Reducing backup and restore costs
- Simplifying overall IT operations

2.13.1 Heterogeneous storage environment management

Tivoli Storage Productivity Center provides device-level, integrated storage infrastructure management capabilities for managing both IBM System Storage products and other vendor storage systems. It is by design that Tivoli Storage Productivity Center provides comprehensive device, capacity, availability and performance management features for IBM platforms, as well as significant capabilities for managing multivendor storage systems. It has advanced, native integration for IBM System Storage DS8000, IBM Storwize V7000, IBM System Storage SAN Volume Controller, and IBM XIV Storage System.

Two key areas of focus are performance management of storage systems and storage networking platforms, and capacity and operational management of IBM and heterogeneous storage systems. Tivoli Storage Productivity Center can help storage administrators more efficiently manage large and complex heterogeneous storage environments by simplifying day-to-day operations, reducing management complexities and improving system availability. Tivoli Storage Productivity Center can help with faster problem determination and resolution to reduce downtimes. Advanced disaster recovery capabilities allows you to plan for replication while storage is being provisioned, ensuring data protection and business continuity.

IBM storage systems support includes operational control and performance reporting of XIV Storage System, System Storage IBM Enterprise Storage Server® Models 800 and F20, System Storage DS3000, DS5000 and DS8000, Storwize V7000, and SAN Volume Controller. Heterogeneous storage support is offered via the Storage Networking Industry Association's Storage Management Initiative Specification Versions 1.0.2 and 1.1 (or later). Management capabilities include operational control, as well as provisioning of heterogeneous storage platforms, including storage devices from HDS, HP, EMC, NetApp, LSI, and Engenio. Tivoli Storage Productivity Center also provides performance reporting at the port and switch levels for platforms from Cisco Systems and Brocade Communications.

2.13.2 Simplified management for storage efficiency and reduced costs

As a management option, Tivoli Storage Productivity Center is available with IBM System Storage hardware. This tool provides storage administrators a simple way to conduct device and performance management for multiple storage arrays and SAN fabric components from a single integrated administrative console. Tivoli Storage Productivity Center provides:

- Web-based management console
- Reporting and analytical capabilities
- Device management
- SAN fabric management
- Replication services management

Tivoli Storage Productivity Center also offers an easy upgrade to IBM SmartCloud™ Virtual Storage Center which is designed to help create private storage clouds. Tivoli Storage Productivity Center provides a fully integrated, web-based user interface that is designed to greatly improve the user experience. Through its centralized user interface, Tivoli Storage Productivity Center enables a number of storage infrastructure management tasks:

- Monitoring the infrastructure, accessing health at a glance and creating logs
- Planning for problem areas and performance trends
- Configuration by providing allocation, zoning and masking
- Reporting capacity, utilization and performance statistics
- Providing problem determination, that is, aggregated status, drill down, identification of impacted resources and recommendations to fix the problem
- Enabling disaster recovery by simplifying advanced copy services (replication) tasks on IBM storage devices

2.13.3 Reporting and analytical capabilities

Tivoli Storage Productivity Center offers improved reporting and analytical capabilities through its integration with IBM Cognos® Business Intelligence reporting and modeling. Data abstraction and ad-hoc reporting enables the creating high-quality reports and charts, and provides out-of-the-box reporting capabilities for quicker time to value. Here we list some key benefits of Cognos-based reporting capabilities:

- Simplified report creation and customization
- Quicker time to value with out-of-the-box reports
- Intuitive drag-and-drop function for rapid reports
- Easy drill-up and drill-down functions for both reporting and charting
- Reports on schedule and/or on demand in multiple distribution formats
- Enhanced visibility into performance to help users make the right decisions

2.14 Cisco Local Area Networking Hardware

The Cisco Local Area Networking Hardware switching portfolio is made up of two main product lines:

- Cisco Catalyst product line: For example, the Cisco Catalyst 4500 or the Cisco Catalyst 6500. These will not be explored in detail in this publication; however, we introduce the Catalyst blade switching module in “Catalyst 3110 Series switches for IBM BladeCenters” on page 112.
- Cisco Nexus product line: These platforms will be discussed in detail in this section, as these represent the fundamental building blocks for the modern enterprise data center. We discuss the following Cisco products:
  - Modular (Chassis-based) switching platforms will be discussed in 2.14.1, “Modular switching platforms”
    - Cisco Nexus 7000
  - Fixed form switching platforms will be discussed in 2.15, “Fixed form factor switching platforms”
    - Cisco Nexus 6000
    - Cisco Nexus 5500
    - Cisco Nexus 3500
    - Cisco Nexus 3000
    - Cisco Nexus 2200
  - Blade switching platforms will be discussed in 2.16, “IBM BladeCenter Switching platforms” on page 110
    - Cisco Nexus 4000

Additional relevant Cisco building blocks for modern data center networks are as follows:

- Cisco Storage Area Network Solutions will be discussed in 2.17, “Cisco Storage Area Network and FCoE Hardware” on page 114.
- Cisco Data Center network services appliances for load balancing, security, and application acceleration will be described in Chapter 6, “Application networking solutions and data center security, physical and virtual” on page 287.
- Cisco Virtual Networking solutions will be described in Chapter 5, “Cisco virtual networking solutions” on page 253.
- Cisco Enterprise and Service Provider routing platforms are beyond the intended scope of this publication.

2.14.1 Modular switching platforms

Modular switching platforms are based on a chassis that can host different types of line cards that share power and cooling resources. The main type of modules that can be hosted on a modular switching platform are as follows:

- Supervisor modules - two for redundancy
- Switching Fabric modules
- I/O modules
- Fan modules - at least two for redundancy
- Power modules - at least two for redundancy

2.14.2 Nexus 7000

Cisco Nexus 7000 Series switches provide the foundation for Cisco Unified Fabric. They are a modular data center-class product line designed for highly scalable 1/10/40/100 Gigabit Ethernet networks with a fabric architecture that scales beyond 17 terabits per second (Tbps). Designed to meet the requirements of the most mission-critical data centers, the switches deliver continuous system operation and virtualized, pervasive services. The Cisco Nexus 7000 Series is based on the proven Cisco NX-OS Software operating system, with enhanced features to deliver real-time system upgrades with exceptional manageability and serviceability.

The first in the next generation of switch platforms, the Cisco Nexus 7000 Series provides integrated resilience combined with features optimized specifically for the data center for availability, reliability, scalability, and ease of management.

The Nexus 7000 family is made up of four different chassis options, as depicted in Figure 2-31 on page 96. The most noticeable difference between models is the number of slots. All chassis support online insertion and removal (OIR) of all redundant components: supervisor modules, power supplies, and fan trays. Four chassis options are available:

- **Nexus 7004**: 7RU, 4-slot switch with 1.44 billion of packets per second IPv4 unicast performance:
  - It supports up to 96 x 1 and 10 Gigabit Ethernet ports, 12 x 40 Gigabit Ethernet ports and 4 x 100 Gigabit Ethernet ports, meeting the needs of small to medium-size data centers, co-locations, access- and aggregation-layer deployments, high-speed core deployments, and smaller operation zones.
  - It is different from other Nexus 7000 chassis in that there are no fabric modules in the Nexus 7004, the local I/O module fabrics are connected back-to-back to form a two-stage crossbar that interconnects the I/O modules and the supervisor engines. The backplane capacity is determined by the installed I/O modules.
  - Does not support F1 cards and Sup1 cards.
  - It supports up to 4 power supplies, both AC and DC.
  - The fan tray with built-in fan and controller redundancy helps ensure reliability of the system and support for hot swapping of fan trays. The fan tray is on the top side of the chassis and draws the air from the right side of the chassis through the line card and supervisor slots and propagates it through the empty space on the left side of the chassis. The air then flows up to the fan tray on the top side and finally flows out from the vent holes on the back side of the chassis.
  - All modules, including power supplies and the fan tray, are accessible from the front.
  - The minimum NXOS release level supported on this platform is 6.1.(2).
- **Nexus 7009**: 14RU, 9-slot switch with 5.04 billion of packets per second IPv4 unicast performance:
  - It supports up to 336 x 1 and 10 Gigabit Ethernet ports, 42 x 40 Gigabit Ethernet ports, and 14 x 100 Gigabit Ethernet ports, meeting the demands of mission-critical campus core and data center deployments.
  - It supports all Supervisor cards and all I/O cards except for the F1 modules.
  - It supports up to 2 power supply slots.
  - Independent variable-speed system and fabric fans provide efficient cooling capacity to the entire system. Fan-tray redundancy features help ensure reliability of the system and support for hot swapping of fan trays.

  - I/O modules, supervisor modules, and fabric modules are accessible from the front. Power supplies and fan trays are accessible from the back.
  - The minimum NXOS release level supported on this platform is 5.2.(1)
- **Nexus 7010**: 21RU, 10-slot switch with 5.76 billion of packets per second IPv4 unicast performance:
  - It supports up to 384 x 1 and 10 Gigabit Ethernet ports, 48 x 40 Gigabit Ethernet ports, and 16 x 100 Gigabit Ethernet ports, meeting the demands of large data center deployments.
  - It supports all Supervisor cards and all I/O cards.
  - It supports up to 3 power supply modules.
  - Front-to-back airflow helps ensure that use of the Cisco Nexus 7000 10-Slot chassis addresses the requirement for hot-aisle and cold-aisle deployments without additional complexity.
  - The system uses dual system and fabric fan trays for cooling. Each fan tray is redundant and composed of independent variable-speed fans that automatically adjust to the ambient temperature, helping reduce power consumption in well-managed facilities while helping enable optimum operation of the switch. The system design increases cooling efficiency and provides redundancy capabilities, allowing hot swapping without affecting the system; if either a single fan or a complete fan tray fails, the system continues to operate without a significant degradation in cooling capacity.
  - The system supports an optional air filter to help ensure clean air flow through the system. The addition of the air filter satisfies Network Equipment Building Standards (NEBS) requirements.
  - I/O modules and supervisor modules are accessible from the front, and fabric modules, power supplies, and fan trays are accessible from the back.
  - The minimum NXOS release level supported on this platform is 4.0.
- **Nexus 7018**: 25RU, 18-slot switch with 11.5 billion of packets per second IPv4 unicast performance.
  - It supports up to 768 x 1 and 10 Gigabit Ethernet ports, 96 x 40 Gigabit Ethernet ports, and 32 x 100 Gigabit Ethernet ports, meeting the demands of the largest data center deployments.
  - It supports all Supervisor cards and all I/O cards.
  - It supports up to 4 power supply modules.
  - Side-to-side airflow increases the system density in a 25RU footprint, optimizing the use of rack space. The optimized density provides more than 16RU of free space in a standard 42RU rack for cable management and patching systems.
  - Independent variable-speed system and fabric fans provide efficient cooling capacity to the entire system. Fan-tray redundancy features help ensure reliability of the system and support for hot swapping of fan trays.
  - I/O modules and supervisor modules are accessible from the front, and fabric modules, power supplies, and fan trays are accessible from the back.
  - The minimum NXOS release level supported on this platform is 4.1(2).

Figure 2-31 shows an overview of the Nexus 7000 Product line. From left to right, the picture represents the 7009, 7010, 7004, and 7018, and they host the line cards horizontally.

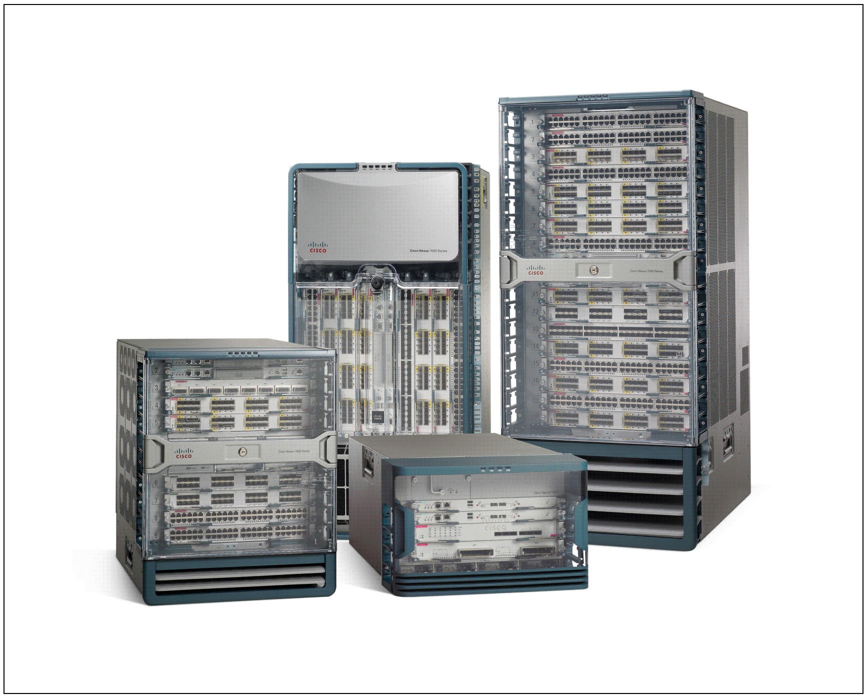


Figure 2-31 Cisco Nexus 7000 Product Line

For additional information on the Nexus 7000 Series switches, visit this website:

http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9402/ps9512/Data_Sheet_C78-437762.html

> **Note:** For each Nexus 7000 Series chassis, two slots must be reserved for the supervisor cards so a Nexus 7010 will have 8 available slots for I/O cards. Fans, power supplies, and switching fabric modules have dedicated slots so they will not consume I/O and supervisor slots.

Table 2-13 summarizes the main features and benefits of the Nexus 7000 Series switches. These features will be explored in detail later.

Table 2-13 Features and benefits of the Nexus 7000 Series switches

Feature	**Benefit**
In Service Software Upgrade (ISSU)	Hitless software upgrade with zero packet loss
Netflow	Visibility and network monitoring
Multi-hop FCoE	Director-class Fibre Channel over Ethernet
OTV and LISP	Seamless workload mobility
MPLS L3 VPN	Multi-tenant segmentation
Virtual Device Context (VDC)	Consolidation and hardware virtualization
FabricPath/TRILL	Highly scalable data center networks

Key design points of the Nexus 7000

These are the key design points of the Nexus 7000 platform:

- Scalability and investment protection: High density 10 GbE, 40 GbE and 100 GbE ready platform
- Operational continuity and high availability: Built-in fault tolerance and high availability for zero downtime environments
- Flexibility and support for virtualization: VM-aware, Unified I/O and cutting edge technologies such as FCoE, FabricPath, and LISP

Most of the line cards can be reused across different Nexus 7000 chassis, with some restrictions. The available line cards at the time of the writing are described in the next sections.

Nexus 7000 supervisor modules

The Nexus 7000 supervisor modules are designed to deliver scalable control plane and management functions for the Cisco Nexus 7000 Series chassis. They are based on dual core processors that scales the control plane by harnessing the flexibility and power of the dual cores. The supervisors control the Layer 2 and 3 services, redundancy capabilities, configuration management, status monitoring, power and environmental management, and more. They also provides centralized arbitration to the system fabric for all line cards.

Two supervisor modules are required in a Cisco Nexus 7000 Series system for high availability with active-standby redundancy, with one supervisor module operationally active, and the standby device serving as a hot backup. Both supervisors in the redundancy configuration must be of the same type.

There are three available Supervisor modules for the Nexus 7000:

- **Supervisor 1 module**: The Nexus 7000 Series Sup1 module is based on a dual-core Intel Xeon processor with 4 GB of memory. The minimum software version is Cisco NX-OS Software Release 4.0.
- **Supervisor 2 module**: The Nexus 7000 Series Sup2 module is based on a quad-core Intel Xeon processor with 12 GB of memory and scales the control plane by harnessing the flexibility and power of the quad cores. A faster CPU and expanded memory compared to the first-generation Sup1 module enable better control-plane performance and an enhanced user experience. The minimum software version is Cisco NX-OS Software Release 6.1(1).
- **Supervisor 2E module**: The Nexus 7000 Series Sup2E Module is based on a two-quad-core Intel Xeon processor with 32 GB of memory that scales the control plane by harnessing the flexibility and power of the two quad cores. With four times the CPU and memory power of the first-generation Cisco Nexus 7000 Series Supervisor1 (Sup1) Module, the Sup2E module offers the highest control-plane performance and scalability of the modules: for example, supporting more VDCs and fabric extenders. The minimum software version is Cisco NX-OS Software Release 6.1(1).

See Table 2-14 for a feature comparison of the Nexus 7000 supervisor modules.

Table 2-14 Nexus 7000 supervisor comparison

	Supervisor 1	**Supervisor 2**	**Supervisor 2E**
CPU	Dual-Core Intel Xeon	Quad-Core Intel Xeon	2 x Quad-Core Intel Xeon
SPEED	1.66 GHz	2.13 GHz	2.13 GHz
MEMORY	8G	12G	32G
EXTERNAL FLASH MEMORY	2 x External Compact Flash memory slots: Log 8GB Expansion 2GB	2 x External USB slots Log 8GB Expansion 2GB	2 x External USB slots Log 8GB Expansion 2GB
CMP	Yes	No	No
SIGNED IMAGES	No	Yes	Yes
NX-OS RELEASE	4.0 or later	6.1 or later	6.1. or later
POWER	190W Typical Power Draw 210W Max	130W Typical Power Draw 190W Max	170W Typical Power Draw 265W Max
CONSOLE	2 (console and modem)	1 (console)	1 (console)
KERNEL	32 Bit Kernel Space 32 Bit User Space	64 Bit Kernel Space 32 Bit User Space	64 Bit Kernel Space 32 Bit User Space
VDC CPU SHARES	No	Yes	Yes
TOTAL VDCs	4	4 + 1	8 + 1

Nexus 7000 switching fabric modules

These separate fabric modules provide parallel fabric channels to each I/O and supervisor module slot. The fabric module provides the central switching element for fully distributed forwarding on the I/O modules. There are two main types of switching fabric modules (each type is specific to one chassis type, except the 4-slot chassis, which has integrated switching fabric modules as described earlier):

- **Fabric1 Module**: Up to five simultaneously active Fabric1 modules work together delivering up to 230 Gbps per slot. Through the parallel forwarding architecture, a system capacity of more than 8 Tbps is achieved with the five fabric modules. Here are the available options:
  - N7K-C7010-FAB-1 (for a 10 Slot Chassis): Minimum software release is NXOS 4.0(1)
  - N7K-C7018-FAB-1 (for an 18 Slot Chassis): Minimum software release is NXOS 4.1(2)
- **Fabric2 Module**: Up to five simultaneously active Fabric2 modules work together delivering up to 550 Gbps per slot. Through the parallel forwarding architecture, a system capacity of more than 15 Tbps is achieved with the five fabric modules.

  These are the available options:
  - N7K-C7009-FAB-2 (for a 9 Slot Chassis): Minimum software release is NXOS 5.2(1)
  - N7K-C7010-FAB-2 (for a 10 Slot Chassis): Minimum software release is NXOS 6.0(1)
  - N7K-C7018-FAB-2 (for an 18 Slot Chassis): Minimum software release is NXOS 6.0(1)

Nexus 7000 I/O modules

The I/O modules are the line cards that provide different port speeds and capabilities for the switch. There are four main types of I/O modules for the Nexus7000:

- **M1 modules:** With 80G backplane capacity. Non XL modules are End of Sale from December, 18th 2012. The Nexus 7000 M-Series XL I/O modules with an "L" suffix are a group of modules with an enhanced forwarding engine that supports large IPv4 and IPv6 forwarding tables and expanded TCAM memory for ACL and QoS. These modules can be operated in either standard or large mode. The "L" suffix refers to the series of modules that have the larger forwarding table capacity. All M1 XL 10G modules have 80 Gbps of fabric bandwidth (this is the amount of traffic a module can send on the switch fabric at one time) and the 1 GE modules have 40 Gbps. The available M1 XL modules, as of the writing of this book, are as follows:
  - M132XP-12L - 32 ports 10 GE module - Minimum software release is NXOS 5.1(1)
  - M148GS-11L - 48 ports 1 GE SFP module - Minimum software release is NXOS 5.0(2)
  - M148GT-11L - 48 ports 10/100/1000 copper module - Minimum software release is NXOS 5.1(2)
  - M108X2-12L - 8 ports 10 GE module (line rate) - Minimum software release is NXOS 5.0(2)
- **M2 modules:** With 240 Gbps switch fabric capacity, the following modules are available:
  - M224XP-23L - with 24 port 10 GE module (line rate) - Minimum software release is NXOS 5.0(2)
  - M206FQ-23L - 6 port 40 GE module (line rate) - Minimum software release is NXOS 6.1(1)
  - M202CF-22L - 2 port 100 GE module (line rate) - Minimum software release is NXOS 5.1(1)
- **F1 modules:** With 230 Gbps switch fabric capacity and Layer 2 only System on Chip (SOC) design. The Cisco Nexus F1 module has a powerful forwarding engine referred to as the System on Chip. Each SoC has two front panel 10 GE ports and a 23G interface to the Fabric ASIC:
  - N7K-F132XP-15 - 32 ports SFP+ 1/10 GE module - Minimum software release is NXOS 5.1(1)
- **F2 modules:** With 480 Gbps switch fabric capacity and L2/L3 System on Chip (SOC). The Cisco Nexus 7000 F2-Series Module is built with switch-on-chip (SoC) architecture, in which a single ASIC implements all the module functions: from ingress buffering, to forwarding lookups and access control lists (ACLs) and quality-of-service (QoS) tables, to fabric interfaces and virtual output queuing (VOQ). Each SoC manages four front-panel interfaces. The F2 module allows for line rate forwarding for all 48 10 G ports:
  - F248XP-25 - 48 ports SFP+ 1/10 GE module - Minimum software release is NXOS 6.0(1)
  - F248XT-25 - 48 ports copper 1/10 GE module - Minimum software release is NXOS 6.1(2)

For a complete list of modules, supported transceivers, and required software versions, see the release notes available at this link:

http://www.cisco.com/en/US/products/ps9402/prod_release_notes_list.html

Table 2-15 summarizes the main differences among the Nexus 7000 Series I/O modules.

Table 2-15 Nexus 7000 Series I/O modules comparison

Feature	M1 Series	M2 Series	F1 Series	F2 Series
Line Rate 10 GE (9/10/18 slot chassis)	56/ 64/ 128	168/192/384	161/184/ 368 (23 ports out of 32 because of 230 Gbps fabric)	336/ 384/ 768
Latency	~20 µs	~15 µs	~5 µs	~5 µs
Power / Line rate 10 GE port	~80 Watts/port	~32 Watts/port	~10 Watts/port	~10 Watts/port
VOQ Buffer per Card	32MB	128MB	48MB	72MB
VLANs	16K	16K	4K	4K
L2 MAC table	128K	128K	16K/SoC	16K/SoC
L3 IPv4 routes	128K-1M	1M	N/A (L2 only)	32K
L3 IPv6 routes	64K-300K	300K	N/A (L2 only)	16K
Adjacency table	1M	1M	N/A (L2 only)	16K
ECMP	1...16	1...16	1...16	1...16 (up to 128)
Netflow	Full/Sampled	Full/Sampled	N/A	Sampled
ACL entries	64K-128K	128K	1K/SoC	16K/SoC
SPAN/ERSPAN sessions	2	16	16	16
IEEE 1588 Time Stamp	No	Yes	Yes	Yes
MPLS	Yes	Yes	No	No
LISP	Yes	No	No	No
FabricPath	No	No	Yes	Yes
FCoE	No	No	Yes	Yes (required Sup2)
OTV	Yes	Yes	No	No

2.15 Fixed form factor switching platforms

Fixed form factor switches are rackable appliance-type switches that can occupy 1 RU or 2 RU with a moderate degree of modularity. Typically fans and power supplies are modular and hot swappable while the port configuration will be typically fixed, although there are some noticeable exceptions such as the Nexus 5500 daughter cards and expansion modules.

In this section, we describe the following Cisco Nexus fixed form factor switches and fabric extenders:

- Nexus 6000
- Nexus 5500
- Nexus 3500
- Nexus 3000
- Nexus 2200

2.15.1 Nexus 6000

The Cisco Nexus 6000 Series is a part of the Cisco Unified Fabric offerings and it expands architectural flexibility and scale, for agile virtualized and cloud deployments. The Cisco Nexus 6000 Series brings the highest 10 and 40 Gigabit Ethernet density in energy-efficient compact form factor switches. With a robust integrated Layer 2 and Layer 3 feature set, the Cisco Nexus 6000 Series provides a versatile platform that can be deployed in multiple scenarios: direct-attach 10 and 40 Gigabit Ethernet access and high-density Cisco Fabric Extender (FEX) aggregation deployments, leaf and spine architectures, or compact aggregation to build scalable Cisco Unified Fabric in the data centers. It has series-based architectures and can adapt to increasing bandwidth demands with low power and a compact space profile. Cisco Nexus 6000 Series products use the same set of Cisco application-specific integrated circuits (ASICs) and a single software image across the products within the family, thereby it offers feature consistency and operational simplicity. In the following section, we describe Nexus 6004 and Nexus 6001 in detail.

Nexus 6004

The Cisco Nexus 6004, the first switch in the family of Cisco Nexus 6000 Series, is a critical component of the Cisco Unified Data Center architecture, complementing the existing Cisco Nexus family of switches. The Cisco Nexus 6004 platform extends the industry-leading innovations and versatility of the Cisco Nexus 5000 Series purpose-built 10 Gigabit Ethernet data center-class switches. In addition, the Cisco Nexus 6004 offers the industry's highest 40/1Gigabit Ethernet and Fibre Channel over Ethernet (FCoE) port density in a compact energy- efficient form factor with integrated Layer 2 and Layer 3 features at wire speed, and low latency of ~1 microsecond for any packet size. With a choice of front-to-back or back-to-front airflow options, the switch is designed for a broad range of traditional data center and large-scale virtualized cloud deployments. It runs the industry-leading Cisco NX-OS Software operating system, offering features and capabilities that are widely deployed worldwide.

The Cisco Nexus 6004 is a 4-rack-unit (4RU) 10 and 40 Gigabit Ethernet switch offering wire-speed performance for up to ninety-six 40 Gigabit Ethernet ports or three hundred eighty-four 10 Gigabit Ethernet ports (using QSFP breakout cables) for Ethernet and FCoE traffic with an overall throughput of 7.68 Tbps. The Cisco Nexus 6004 offers 48 fixed ports of 40 Gigabit Ethernet with the base of the chassis and 4 line-card expansion module (LEM) slots. Although the base chassis has 48 fixed 40 Gigabit Ethernet ports, it can further be customized for convenient twenty-four 40 Gigabit Ethernet port entry option to address the small-scale deployments.

With two additional 12-port 40 Gigabit Ethernet software licenses, the remaining 24 ports can be enabled to use all 48 ports of 40 Gigabit Ethernet on the base chassis. Each LEM supports 12 ports of 40 Gigabit Ethernet in a QSFP form factor. For example, by adding four LEMs to the base chassis, 48 additional 40 Gigabit Ethernet ports can be achieved. Each 40 Gigabit Ethernet port can be split into four 10 Gigabit Ethernet ports using QSFP breakout cables. Thus Cisco Nexus 6004 chassis along with the LEMs offer incredible flexibility to scale the deployments, making it the only fully extensible fixed 10 and 40 Gigabit Ethernet platform in the industry.

The Cisco Nexus 6004 supports four expansion modules that are all hot-swappable. One of the expansion modules provides 12 ports of 40-Gbps Gigabit Ethernet and FCoE ports using a QSFP interface. With all four expansion modules installed, the Cisco Nexus 6004 delivers 96 ports of QSFP or 384 ports of Small Form-Factor Pluggable Plus (SFP+) using the breakout cables.

Nexus 6001

The Cisco Nexus 6001 is a 1RU 10 and 40 Gigabit Ethernet switch offering wire-speed performance for up to sixty- four 10 Gigabit Ethernet ports (using Quad Small Form-Factor Pluggable [QSFP] breakout cables) for Ethernet and FCoE traffic, with an overall throughput of 1.28 terabits per second (Tbps). The Cisco Nexus 6001 offers 48 fixed 10 Gigabit Ethernet Enhanced Small Form-Factor Pluggable (SFP+) ports and four 40 Gigabit Ethernet QSFP+ ports. Each 40 Gigabit Ethernet port can be split into four 10 Gigabit Ethernet ports using a QSFP breakout cable. The Cisco Nexus 6001 delivers low port-to-port latency of approximately 1 microsecond and low jitter independent of packet size using cut-through switching architecture and with features enabled.

The Cisco Nexus 6001 can be deployed in multiple scenarios (direct-attach 10 and 40 Gigabit Ethernet server access and high-density fabric extender aggregation deployments, leaf, and spine designs, and compact aggregation) to build scalable Cisco Unified Fabric across a diverse set of physical and virtual server environments in the data center.

2.15.2 Nexus 5500

With low latency and choice of front-to-back or back-to-front cooling, copper or fiber access ports, and rear-facing data ports, the Cisco Nexus 5000 Series is designed for a broad range of physical, virtual, storage access, and high-performance computing environments, thus giving customers the flexibility to meet and scale their data center requirements in a gradual manner and at a pace that aligns with their business objectives.

The switch series, using cut-through architecture, supports line-rate 10 Gigabit Ethernet on all ports while maintaining consistently low latency independent of packet size and services enabled. It supports a set of network technologies known collectively as Data Center Bridging (DCB) that increases the reliability, efficiency, and scalability of Ethernet networks. These features allow the switches to support multiple traffic classes over a lossless Ethernet fabric, thus enabling consolidation of LAN, SAN, and cluster environments. Its ability to connect Fibre Channel over Ethernet (FCoE) to native Fibre Channel protects existing storage system investments while dramatically simplifying in-rack cabling.

In addition to supporting standard 10 Gigabit Ethernet network interface cards (NICs) on servers, the Cisco Nexus 5000 Series integrates with multifunction adapters called converged network adapters (CNA) that combine the functions of Ethernet NICs and Fibre Channel host bus adapters (HBAs) using FCoE (Fibre Channel over Ethernet) technology, making the transition to a single, unified network fabric transparent and consistent with existing practices, management software, and OS drivers. The switch series is compatible with integrated transceivers and Twinax cabling solutions that deliver cost-effective connectivity for 10 Gigabit Ethernet to servers at the rack level, eliminating the need for expensive optical transceivers. The Cisco Nexus 5000 Series portfolio also provides the flexibility to connect directly to servers using 10GBASE-T connections or fiber with Enhanced Small Form-Factor Pluggable (SFP+) transceivers. In addition, the Cisco Nexus 5500 platform supports 1 Gigabit Ethernet connectivity options using 1GBASE SFP modules, 8/4/2 -Gbps Fibre Channel SFP+ and 4/2/1-Gbps Fibre Channel SFP interfaces are supported with expansion module options.

The Nexus 5500 Series switches are the second generation of the Nexus 5000 platform. The Cisco Nexus 5500 platform extends the industry-leading versatility of the Cisco Nexus 5000 Series purpose-built 10 Gigabit Ethernet data center-class switches and provides innovative advances toward higher density, lower latency, and multilayer services. The Cisco Nexus 5500 platform is well suited for enterprise-class data center server access-layer deployments across a diverse set of physical, virtual, storage-access, and high-performance computing (HPC) data center environments.

Another innovation of the Nexus 5500 series of switches is the concept of unified ports. Unified ports allow any capable port to take on the character of 1 and 10 Gigabit Ethernet, SAN and LAN shared on 10 Gigabit Ethernet, or 8/4/2/1-Gbps Fibre Channel. Unified ports give the user flexibility in choosing SAN and LAN port options consistent with the virtualized data center and offer a migration path to FCoE for those users not yet ready to make the move from native Fibre Channel. This means that any physical port on a Nexus 5500 U Series switch can be configured as a native Fibre Channel port for example.

The Nexus 5500 can be configured with front-to-back or back-to-front airflow to allow for data center placement flexibility.

There are four main types of Nexus 5500 switches:

- **Nexus 5548P** - N5K-C5548P-FA - minimum NXOS release 5.0(2):
  - A one-rack-unit (1RU) 10 Gigabit Ethernet and FCoE switch offering up to 960-Gbps throughput and up to 48 ports. The switch has 32 1/10-Gbps fixed SFP+ Ethernet and FCoE ports and one expansion slot.
- **Nexus 5548UP** - N5K-C5548UP-FA - minimum NXOS release 5.0(3):
  - A 1RU 10 Gigabit Ethernet, Fibre Channel, and FCoE switch offering up to 960 Gbps of throughput and up to 48 ports. The switch has 32 unified ports and one expansion slot.
- **Nexus 5596 UP** - N5K-C5596UP-FA - minimum NXOS release 5.0(3):
  - A 2RU 10 Gigabit Ethernet, Fibre Channel, and FCoE switch offering up to 1920 Gbps of throughput and up to 96 ports. The switch has 48 unified ports and three expansion slots.
- **Nexus 5596T** - N5K-C5596T-FA - minimum NXOS release 5.2(1):
  - A 2RU form factor, with 32 fixed ports of 10GBASE-T and 16 fixed ports of SFP+. The switch also supports up to three expansion slots. The switch supports 10 Gigabit Ethernet (fiber and copper), Fibre Channel, and FCoE, offering up to 1920 Gbps of throughput and up to 96 ports. The switch supports unified ports on all SFP+ ports. The hardware for the 10GBASE-T ports is capable of supporting FCoE.

The Cisco Nexus 5500 platform is equipped with expansion modules that can be used to increase the number of 10 Gigabit Ethernet and FCoE ports or to connect to Fibre Channel SANs with 8/4/2/1-Gbps Fibre Channel switch ports, or both.

The Cisco Nexus 5548P and 5548UP supports one expansion module, and the Cisco Nexus 5596UP and 5596T support three expansion modules from the following options:

- Ethernet module that provides sixteen 1/10 Gigabit Ethernet and FCoE ports using the SFP+ interface
- Fibre Channel plus Ethernet module that provides eight 1/10 Gigabit Ethernet and FCoE ports using the SFP+ interface, and eight ports of 8/4/2/1-Gbps native Fibre Channel connectivity using the SFP+/SFP interface
- Unified port module that provides up to sixteen 1/10 Gigabit Ethernet and FCoE ports using the SFP+ interface or up to sixteen ports of 8/4/2/1-Gbps native Fibre Channel connectivity using the SFP+ and SFP interfaces; the use of 1/10 Gigabit Ethernet or 8/4/2/1-Gbps Fibre Channel on a port is mutually exclusive but can be selected for any of the 16 physical ports per module

Note: Given that the price is the same, we always recommend the Unified ports as expansion modules, as these give more flexibility to a data center network deployment.

In addition to these expansion modules, the Cisco Nexus 5548P and 5548UP support a Layer 3 daughter card that can be ordered with the system or as a spare (field upgradable). This daughter card provides up to 160 Gbps of Layer 3 forwarding capability (240 million packets per second [mpps]) that can be shared by all 48 ports in the chassis. The Layer 3 daughter card does not take up one of the expansion slots on the rear of the chassis, but instead is installed by replacing the Layer 2 I/O module (N55-DL2) that is located on the front of the chassis:

- In addition to three expansion modules, the Cisco Nexus 5596UP and 5596T support a Layer 3 module that provides up to 160 Gbps of Layer 3 forwarding capability (240 mpps) that can be shared by all the I/O ports in the chassis.
- Version 2 of the Layer 3 daughter cards and expansion modules have enhanced hardware capabilities that increase the host table size from 8000 to 16,000 entries, or multicast routes from 4000 to 8000 (enabled in a future software release).
- The Layer 3 module requires a base LAN license (N55-BAS1k9=),which is free of charge, to be installed. This is included in the hardware purchase, however the license would need to be manually installed.

Figure 2-32 shows an overview of the Nexus 5000 switches. From top to bottom the picture shows the Nexus 5596 UP, the 5548UP, the 5548P and the older Nexus 5010 and 5020 products. The expansion modules are visible at the right hand side of each switch.

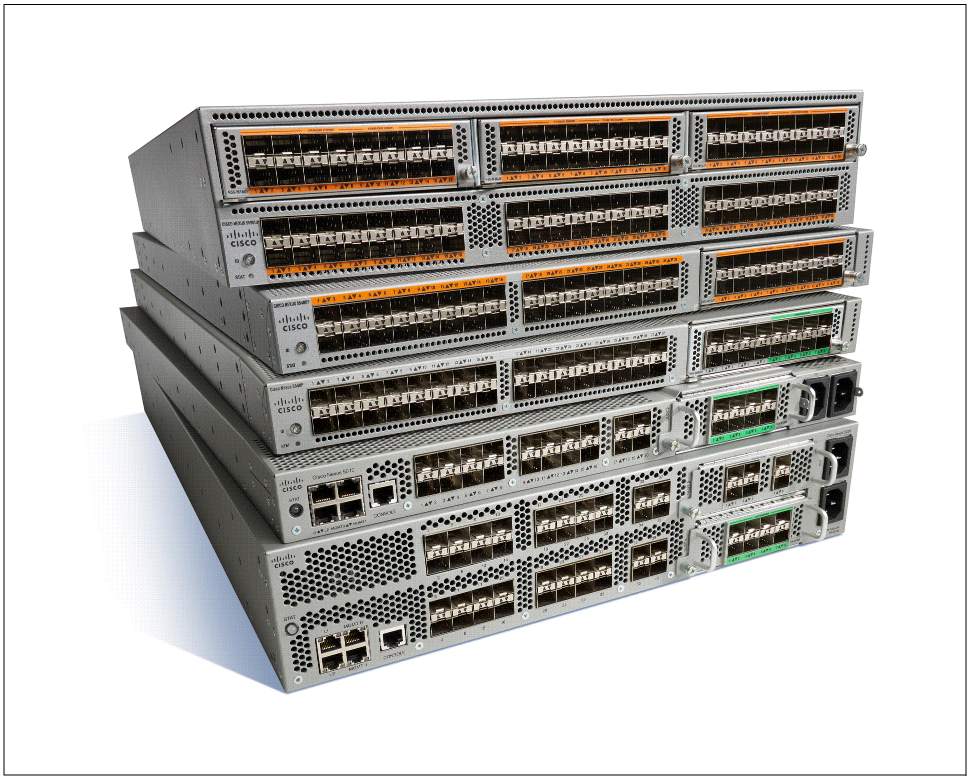


Figure 2-32 Nexus 5000 Series switches product line

For additional detailed information on the Nexus 5000 switching platform, see this website:

http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9670/data_sheet_c78-618603.html

2.15.3 Nexus 3000

The Cisco Nexus 3000 Series switches are a comprehensive portfolio of 1, 10, and 40 Gigabit Ethernet switches built from a switch-on-a-chip (SoC) architecture. Introduced in April 2011, this series has established itself as a leader in high frequency trading, high-performance computing, and Big Data environments by pairing high performance and low latency with innovations in performance visibility, automation, and time synchronization. This switch is based on merchant silicon from Broadcom, while the 3500 Series is built on Cisco internally developed ASICs.

The Cisco 3000 Series includes these main features:

- Ultra-Low Latency:
  - Line-rate Layer 2/Layer 3 switching at ultra-low latencies for High-Performance Trading workloads
- Mission-Critical Features:
  - Network Address Translation
  - Virtual Port Channel
  - IEEE-1588 Precision Time Protocol (PTP)
- Performance Visibility:
  - Buffer monitoring
  - Embedded Remote SPAN with PTP time-stamping
- Easy Deployment and Configuration:
  - Modular and resilient Cisco NX-OS operating system
  - Robust and proven software that is deployed in thousands of data centers worldwide
  - Simplified management with support for Python scripting, EEM, and other XML manageability tools

The Cisco Nexus 3000 Series switches come in three types:

- **Cisco Nexus 3064** switches:
  - The Cisco Nexus 3064-X and 3064-T Switches are high-performance, high-density, ultra-low-latency Ethernet switches that are part of the Cisco Nexus 3000 Series switches portfolio. These compact one-rack-unit (1RU) form-factor 10 Gigabit Ethernet switches provide line-rate Layer 2 and 3 switching. They run the Cisco NX-OS Software operating system. They support both forward and reverse airflow schemes with AC and DC power inputs. The Cisco Nexus 3064-X and 3064-T Switches are well suited for financial co-location and Web 2.0 deployments that require comprehensive unicast and multicast routing protocol features at ultra-low latencies:
    - Cisco Nexus 3064-X: This 10-Gbps Enhanced Small Form-Factor Pluggable (SFP+)-based Top-of-Rack switch has 48 SFP+ ports and 4 Quad SFP+ (QSFP+) ports. Each SFP+ port can operate in 100-Mbps, 1-Gbps, or 10-Gbps mode, and each QSFP+ port can operate in native 40-Gbps or 4 x 10-Gbps mode. This switch is a true phy-less switch that is optimized for low latency and low power consumption.
    - Cisco Nexus 3064-T: This 10GBASE-T switch has 48 10GBASE-T RJ-45 ports and 4 QSFP+ ports. This switch is well suited for customers who want to reuse existing copper cabling while migrating from 1-Gbps to 10-Gbps servers.

- **Cisco Nexus 3048** switch:
  - The Cisco Nexus 3048 Switch is a 48 ports line-rate Gigabit Ethernet Top-of-Rack (ToR) switch and is part of the Cisco Nexus 3000 Series switches portfolio. The Cisco Nexus 3048, with its compact one-rack-unit (1RU) form factor and integrated Layer 2 and 3 switching, complements the existing Cisco Nexus family of switches. This switch runs the Cisco NX-OS Software operating system. The Cisco Nexus 3048 is ideal for Big Data customers that require a Gigabit Ethernet ToR switch with local switching that connects transparently to upstream Cisco Nexus switches, providing an end-to-end Cisco Nexus fabric in their data centers. This switch supports both forward and reversed airflow schemes with AC and DC power inputs.
- **Cisco Nexus 3016** switch:
  - The Cisco Nexus 3016 Switch is a 16 port 40 Gigabit Ethernet switch platform. It is a high-performance, ultra-low-latency Ethernet switch providing line-rate Layer 2 and 3 switching in a compact one-rack-unit (1RU) form factor. The switch runs the Cisco NX-OS Software operating system. The line-rate Layer 2 and 3 switching at ultra-low latencies along with the serialization benefits of 40 Gigabit Ethernet switching make the Cisco Nexus 3016 an ideal switch platform for financial co-locations. This switch supports both forward and reversed airflow schemes with AC and DC power inputs.

For more information on the Cisco 3000 product line, see this website:

http://www.cisco.com/en/US/products/ps11541/products_data_sheets_list.html

Figure 2-33 shows the Nexus 3000 Series switches. From top to bottom, the picture shows the Nexus 3548P (which will be discussed in the next section), the Nexus 3064T, the Nexus 3016, the Nexus 3064X, and the Nexus 3048.

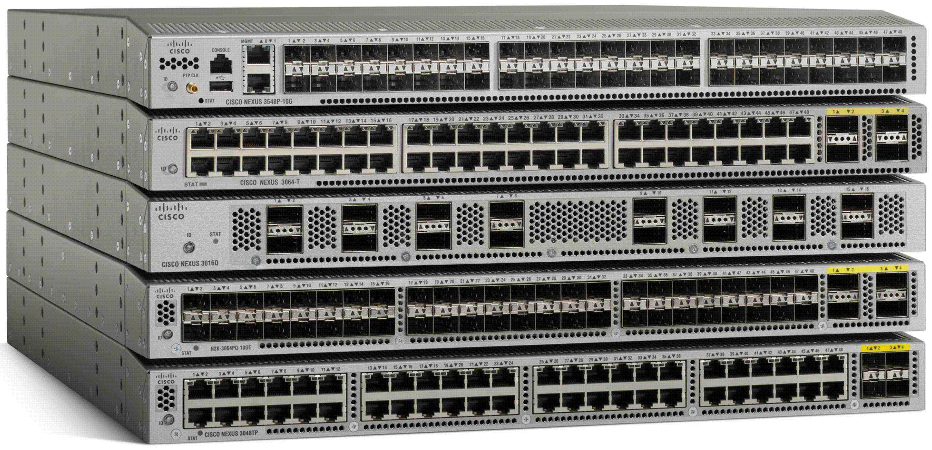


Figure 2-33 Nexus 3000 Series switches

2.15.4 Nexus 3500

The Cisco Nexus 3500 platform further extends the leadership of the Cisco Nexus 3000 Series by including the innovative Algorithm Boost (or Algo® Boost) technology. Algo Boost technology, built into the switch application specific integrated circuit (ASIC), allows the Nexus 3548 to achieve exceptional Layer 2 and Layer 3 switching latencies of less than 200 nanoseconds (ns). In addition, Algo Boost contains several innovations for latency, forwarding features, and performance visibility.

The main features of the Nexus 3500 are as follows:

- Hitless Network Address Translation (NAT): In many financial trading environments, trade orders must be sourced from the IP space of the provider, requiring NAT at the border between networks. The Cisco Nexus 3500 platform can perform NAT for IPv4 unicast routed packets without incurring any additional latency.
- Active buffer monitoring: Even on the lowest-latency switches, data packets can incur a millisecond or more of latency during periods of congestion. Today's switches do not adequately inform administrators about the presence of this congestion, leaving them unaware and hindered in their ability to address the conditions causing suboptimal performance. Previous buffer utilization monitoring techniques were based entirely on software polling algorithms with polling intervals higher than 100ms, which can miss important congestion events. In contrast, Algo Boost accelerates the collection of buffer utilization data in hardware, allowing sampling intervals of 10 ns or less.
- Advanced traffic mirroring: The Algo Boost technology on the Cisco Nexus 3500 platform facilitates not only network troubleshooting by supporting Switched Port Analyzer (SPAN) and Encapsulated Remote SPAN (ERSPAN) technologies, but also in-service network monitoring with enhancements including the capability to perform these operations:
  - Apply user-configurable filters to reduce the amount of captured traffic to a specified flow or protocol
  - Capture a sample of eligible packets, such as one out of every thousand
  - Truncate packets after a user-defined threshold
  - Insert a nanosecond-level timestamp in the ERSPAN header of captured packets (requires ERSPAN and Precision Time Protocol [PTP])
- IEEE 1588 PTP with pulse-per-second* (PPS) output:
  - The capability to build and maintain a synchronized, accurate timing solution is the basis for successful provisioning and management of high-performance trading networks and applications. Using IEEE 1588 PTP, Cisco Nexus 3000 Series switches deliver highly accurate precision time synchronization to applications within existing network infrastructure with no need to invest in and deploy a separate timing network.
  - Network administrators deploying IEEE 1588 PTP often find it challenging to measure the accuracy to which each device is synchronized. To assist this effort, the Cisco Nexus 3500 platform includes a 1-PPS output port that can be used to measure timing drift from the grandmaster clock.

The Cisco Nexus 3548 Switch, the first member of the Cisco Nexus 3500 platform, is a compact one-rack-unit (1RU) form-factor 10 Gigabit Ethernet switch providing line-rate Layer 2 and 3 switching at ultra low latency. The switch runs the Cisco NX-OS Software operating system. The Cisco Nexus 3548 contains no physical layer (PHY) chips, allowing low latency and low power consumption. This switch supports both forward and reversed airflow schemes and both AC and DC power inputs.

The Cisco Nexus 3548 has the following hardware configuration:

- 48 fixed Enhanced Small Form-Factor Pluggable (SFP+) ports (1 Gbps or 10 Gbps)
- Dual redundant, hot-swappable power supplies
- Four individual, redundant, hot-swappable fans
- One 1-PPS timing port, with the RF1.0/2.3 QuickConnect connector type

For more information on the Cisco 3500 switch, see this website:

http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps11541/ps12581/data_sheet_c78-707001.html

2.15.5 Nexus 2200

The Cisco Nexus 2000 Series Fabric Extenders comprise a category of data center products designed to simplify data center access architecture and operations. The Cisco Nexus 2000 Series uses the Cisco Fabric Extender architecture to provide a highly scalable unified server-access platform across a range of 100 Megabit Ethernet, Gigabit Ethernet, 10 Gigabit Ethernet, unified fabric, copper and fiber connectivity, rack, and blade server environments. The platform is ideal to support today's traditional Gigabit Ethernet while allowing transparent migration to 10 Gigabit Ethernet, virtual machine-aware unified fabric technologies.

The Cisco Nexus 2000 Series Fabric Extenders behave as remote line cards for a parent Cisco Nexus switch. The fabric extenders are essentially extensions of the parent Cisco Nexus switch fabric, with the fabric extenders and the parent Cisco Nexus switch together forming a distributed modular system. This means that the fabric extenders are completely managed from the parent switch and appear as physical ports on that switch. The fabric extenders get their NX-OS code from the parent switch and supports ISSU for code upgrades and downgrades.

This architecture enables physical topologies with the flexibility and benefits of both Top-of-Rack (ToR) and End-of-Row (EoR) deployments. The Cisco Nexus 2000 Series Fabric Extenders behave like remote line cards for a parent Cisco Nexus 5000 or 7000 Series Switch. Working in conjunction with Cisco Nexus switches, the Cisco Nexus 2000 Series Fabric Extenders extend the capabilities and benefits offered by the parent Cisco Nexus switch while providing flexible, scalable, and cost-effective server access.

For more information on how to attach Nexus 2000 to different parent switches, see Chapter 4, “Data center physical access layer evolution” on page 191.

The Cisco Nexus 2000 Series architecture provides the following benefits:

- Architecture flexibility: Common, scalable, and adaptive architecture across data center racks and points of delivery (PoDs) that supports various server options, connectivity options, physical topologies and evolving needs.
- Highly scalable server access: Scalable Gigabit and 10 Gigabit Ethernet server access with no reliance on Spanning Tree.
- Simplified operations: One single point of management and policy enforcement using upstream Cisco Nexus switches eases the commissioning and decommissioning of server racks through zero-touch installation and automatic configuration of fabric extenders.
- Increased business benefits: Consolidation, cabling reduction, rack space reduction, reduced power and cooling, investment protection through feature inheritance from the parent switch, and the capability to add functions without the need for a major equipment upgrade of server-attached infrastructure all contribute to reduced operating expenses (OpEx) and capital expenditures (CapEx).

The Nexus 2000 Series platforms come in different types:

- The Cisco Nexus **2224TP**, **2248TP**, and **2248TP-E** provide port density options for highly scalable 100-Mbps and Gigabit Ethernet connectivity.
- The Cisco Nexus **2248TP-E** Fabric Extender is a general-purpose 1 Gigabit Ethernet Fabric Extender with enhancements that target workloads such as large-volume databases, distributed storage, and video editing. Just like the Cisco Nexus 2248TP, the Cisco Nexus 2248TP-E supports 48 100/1000BASE-T host facing ports and four 10 Gigabit Ethernet fabric interfaces. It also supports 32MB shared buffers.

- The Cisco Nexus **2232PP** 1/10 GE Fabric Extender is the ideal platform for migration from Gigabit Ethernet to 10 Gigabit Ethernet and unified fabric environments. It supports FCoE and a set of network technologies known collectively as Data Center Bridging (DCB) that increase the reliability, efficiency, and scalability of Ethernet networks. These features allow the switches to support multiple traffic classes over a lossless Ethernet fabric, thus enabling consolidation of LAN, SAN, and cluster environments.
- The Cisco Nexus **2232TM** 1/10GBASE-T Fabric Extender supports scalable 1/10GBASE-T environments, ease of migration from 1GBASE-T to 10GBASE-T, and effective reuse of existing structured cabling. It comes with an uplink module that supports eight 10 Gigabit Ethernet fabric interfaces. The Nexus 2232TM supports DCB.
- The Cisco Nexus 2248PQ 1/10 Gigabit Ethernet Fabric Extender supports high-density 10 Gigabit Ethernet environments and has 48 1/10 Gigabit Ethernet SFP+ host ports and 4 QSFP+ fabric ports (16 x 10 GE fabric ports). QSFP+ connectivity simplifies cabling while lowering power and solution cost. The Cisco Nexus 2248PQ 10 GE Fabric Extender supports FCoE and a set of network technologies known collectively as Data Center Bridging (DCB) that increase the reliability, efficiency, and scalability of Ethernet networks. These features allow support for multiple traffic classes over a lossless Ethernet fabric, thus enabling consolidation of LAN, storage area network (SAN), and cluster environments.
- The Cisco Nexus **2232TM-E** 1/10GBASE-T Fabric Extender is a super-set of the Cisco Nexus 2232TM with the latest generation of 10GBASE-T PHY, enabling lower power and improved bit error rate (BER).

Table 2-16 Nexus 2000 Series feature comparison table

Feature	**Nexus 2148T**	**Nexus 2224TP**	**Nexus 2248TP**	**Nexus 2248TP-E**	**Nexus 2232PP**	**Nexus 2232TM**	**Nexus 2232TM-E**	**Nexus 2248PQ**
Port speed	1 G	1 G	1 G	1 G	1/10 G	1/10 G	1/10 G	1/10G
Host interfaces	48	24	48	48	32	32	32	48
Uplinks	4	2	4	4	8	8	8	4*QSFP (10G*16)
Fabric speed	80 Gbps full duplex	40 Gbps full duplex	80 Gbps full duplex	80 Gbps full duplex	160 Gbps full duplex	160 Gbps full duplex	160 Gbps full duplex	320 Gbps full duplex
Oversubscription	1.2: 1	1.2: 1	1.2: 1	1.2: 1	4: 1	4: 1	4: 1	3:1
FCoE	No	No	No	No	Yes	No		Yes
Parent switch	Nexus 5000	Nexus 5000 / 7000	Nexus 5000 / 7000	Nexus 5000	Nexus 5000 / 7000	Nexus 5000	Nexus 5000	Nexus 6000
Minimum NXOS software release	5k 4.0(1)	5k 4.2(1) 7k 5.2	5k 4.2 7k 5.1	5k 5.1(3) 7k 6.1(1)	5k 4.2 7k 5.2	5k 5.0(3) 7k 6.1(1)	5k 5.2(1)	6k 6.0(2)

Figure 2-34 shows the Nexus 2000 product line. From bottom left to top right: Cisco Nexus 2148T, 2224TP GE, 2248TP GE, 2248TP-E, 2232TM 10GE,2232PP 10 GE and 2248PQ; cost-effective fabric extender transceivers for Cisco Nexus 2000 Series and Cisco Nexus Parent Switch Interconnect are in front of the fabric extenders.

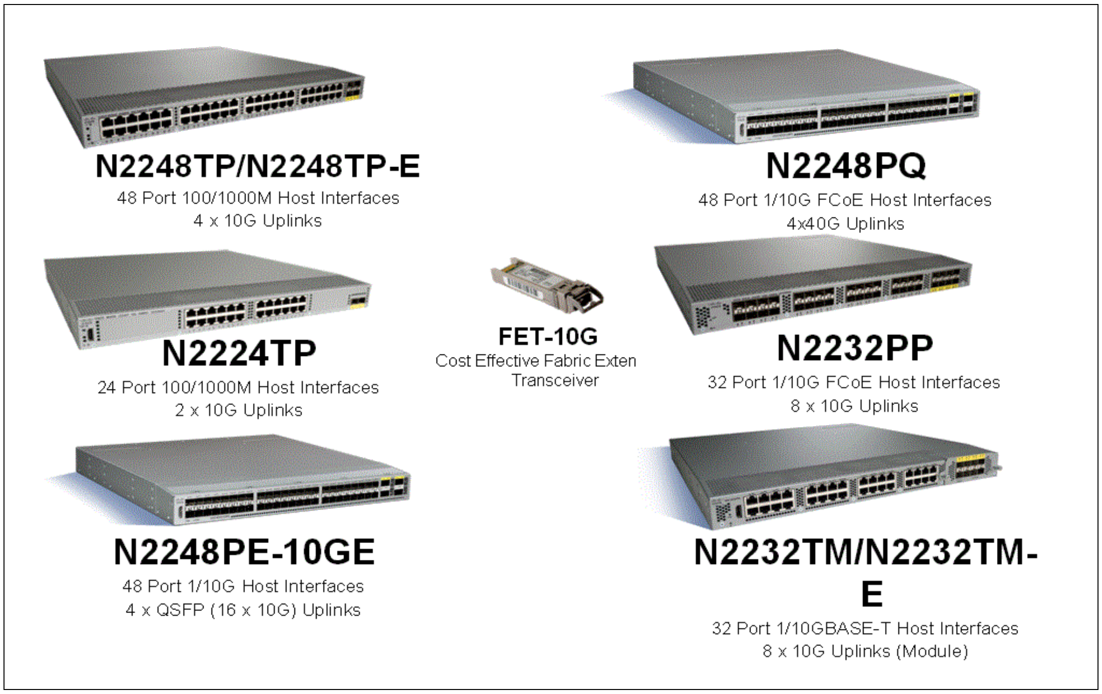

N2248TP/N2248TP-E
48 Port 100/1000M Host Interfaces
4 x 10G Uplinks
N2248PQ
48 Port 1/10G FCoE Host Interfaces
4x40G Uplinks
N2224TP
24 Port 100/1000M Host Interfaces
2 x 10G Uplinks
FET-10G
Cost Effective Fabric Exten
Transceiver
N2232PP
32 Port 1/10G FCoE Host Interfaces
8 x 10G Uplinks
N2248PE-10GE
48 Port 1/10G Host Interfaces
4 x QSFP (16 x 10G) Uplinks
N2232TM/N2232TM-E
32 Port 1/10GBASE-T Host Interfaces
8 x 10G Uplinks (Module)

Figure 2-34 Nexus 2000 Series

For more information on the Nexus 2000 Series products, see this website:

http://www.cisco.com/en/US/partner/prod/collateral/switches/ps9441/ps10110/data_sheet_c78-507093.html

2.16 IBM BladeCenter Switching platforms

The following section describes the Cisco Ethernet switching modules that are available for IBM BladeCenters. We introduce the following modules:

- The Nexus 4000 Series switching module
- The Catalyst 3110 Series switching modules

2.16.1 Nexus 4000 Series switches for IBM BladeCenters

The Cisco Nexus 4001I Switch Module is a blade switch solution for the BladeCenter H and HT chassis providing the server I/O solution required for high-performance, scale-out, virtualized, and non-virtualized x86 computing architectures. It is a line rate, extremely low-latency, non-blocking, Layer 2, 10 Gigabit Ethernet blade switch that is fully compliant with Fibre Channel over Ethernet (FCoE) and IEEE Data Center Bridging standards.

The Cisco Nexus 4001I enables a standards-based, high-performance Unified Fabric running over 10 Gigabit Ethernet in the blade server environment. This Unified Fabric enables consolidation of LAN traffic, storage traffic (IP-based such as iSCSI, NAS, and Fibre Channel SAN), and high-performance computing (HPC) traffic over a single 10 Gigabit Ethernet server network.

The Cisco Nexus 4001I Switch Module provides the following benefits:

- Lower total cost of operation (TCO): Deployment of Unified Fabric with the Nexus 4001I on the blade server access leverages a significant reduction in the number of switches, network interface cards (LAN and SAN), ports, optic modules, and cables. This consolidation of server access network elements significantly reduces the overall capital and operation costs of the data center network through the reduction of network elements to purchase, manage, power, and cool.
- High performance: The Nexus 4001I is a feature-rich, extremely low-latency switch capable of enabling server access migration from 1 GbE to 10 GbE to lossless 10 GbE, as well as supporting the demanding latency requirements of High Performance Compute clusters or high-frequency trading applications.
- Enhanced server virtualization: Utilizing Unified Fabric on the server access with Nexus 4001I provides uniform interfaces, simplified cabling, and consistent server access design required to leverage the advantages of automated virtual machine mobility. Using the Nexus 4001I in conjunction with the Nexus 1000V delivers the most operationally consistent and transparent server access design for virtual machine deployments, substantially reducing the overhead to configure, troubleshoot, and repair the server access link between the vNIC, virtual switch, and blade switch.
- Increased resilience: The Nexus 4001I extends NX-OS to blade server access, providing a fault-tolerant network with a single modular operating system across the data center.

The Cisco Nexus 4001I Switch Module includes the following features and functions:

- Six 10 Gb SFP+ ports operating at wire speed. Also designed to support 1 Gb SFP if required, with the flexibility of mixing 1 Gb/10 Gb. Table 2 lists supported transceivers and cables.
- One 10/100/1000 Mb copper RJ-45 used for management or data.
- Fourteen internal auto-negotiating ports: 1 Gb or 10 Gb to the server blades.
- One internal full-duplex 100 Mbps port connected to the management module.
- Non-blocking architecture with wire-speed forwarding of traffic and full line rate performance of 400 Gbps full duplex.
- Forwarding rate of 300 million packets per second (mpps).
- Low, predictable, and consistent latency of 1.5 microseconds regardless of packet size, traffic pattern, or enabled features on 10 Gigabit Ethernet interface.
- EtherChannels and LACP (IEEE 802.3ad) link aggregation, up to 60 Gb of total uplink bandwidth per switch, up to seven trunk groups, and up to six ports per group.
- Up to 512 VLANs supported per switch; VLAN numbers ranging from 1 to 4000 with 802.1Q VLAN tagging support on all ports.
- Support for T11-compliant FCoE on all 10-Gigabit Ethernet interfaces.
- FCoE Initialization Protocol (FIP): Converged Enhanced Ethernet Data Center Bridging Exchange (CEE-DCBX) protocol supports T11-compliant Gen-2 CNAs.
- Priority-based flow control (IEEE 802.1Qbb) simplifies management of multiple traffic flows over a single network link and creates lossless behavior for Ethernet by allowing class-of-service (CoS)-based flow control.
- Enhanced Transmission Selection (IEEE 802.1Qaz) enables consistent management of QoS at the network level by providing consistent scheduling of different traffic types (IP, storage, and so on).
- Data Center Bridging Exchange (DCBX) Protocol (IEEE 802.1AB) simplifies network deployment and reduces configuration errors by providing auto-negotiation of IEEE 802.1 DCB features between network interface card (NIC) and the switch and between switches.

Figure 2-35 shows the Cisco Nexus 4001I for IBM BladeCenters.

Figure 2-35 Cisco Nexus 4001I for IBM BladeCenter

For more information on the Cisco Nexus 4000 switch, see this website:

http://www.redbooks.ibm.com/abstracts/tips0754.html

2.16.2 Catalyst 3110 Series switches for IBM BladeCenters

The Cisco Catalyst Switch Module 3110G and 3110X are Gigabit Ethernet Switch Modules in a standard switch-bay form-factor for use in all BladeCenter chassis. These stackable switches are full wire-rated, non-blocking switches for use with high performance servers. The 3110G offers four external RJ-45 Gigabit Ethernet connections and the 3110X offers one external 10 Gb Ethernet slot (for use with an X2 transceiver module) for making 10 Gb uplinks to backbone switches or routers.

The Cisco Catalyst Switch Module 3110 has a unique technology called Virtual Blade Switch (VBS). Much like server virtualization technology, this switch virtualization technology treats the individual physical switches within a rack as a single logical switch. As with server virtualization technology, this innovation allows the switches to deliver better utilization, increased performance, and greater resilience while simplifying operations and management.

A VBS stack has the following major characteristics:

- Up to nine physical switches can be combined in one stack: one master switch, and up to eight member switches.
- Two dedicated connectors on a switch are used to create VBS, and switches are connected in a ring. Stack connections form two counter-rotating unidirectional links, with up to 32 Gbps of raw capacity per stack member (16 Gbps per ring link).

- Single point of management.
- All switches in a stack are managed through a single IP address on the master switch.
- Single configuration file per stack.
- Single software upgrade for entire stack.
- Single Spanning Tree Protocol (STP) instance for Layer 2 networks.
- Single router for Layer 3 networks.
- Consolidation of uplink ports (or uplink sharing).

Built upon Cisco's hardware and IOS software, the switches are designed to deliver scalable, high performance, highly resilient connectivity while reducing server infrastructure complexity.

The supported features and specifications for the Cisco Catalyst Switch Modules 3110G and 3110X are as follows:

- 3110G: Four external RJ-45 1000BASE-T connectors for making 10/100/1000 Mbps connections to a backbone, end stations, and servers.
- 3110X: One external 10 Gb Ethernet Module slot for forming 10 Gb uplinks to backbone switches or routers. This module slot operates at full-duplex and uses hot-swappable Cisco X2 transceiver modules. The transceiver module is not included and must be ordered from a Cisco Systems reseller.
- Two external high-speed StackWise Plus ports for switch module stacking to support Virtual Blade Switch (VBS) technology. Each 3110G switch module ships with one 1-meter StackWise Plus cable. Other cables are available for order from Cisco Systems resellers.
- 14 internal full-duplex Gigabit ports, one connected to each of the blade servers in the BladeCenter unit.
- One internal full-duplex 100 Mbps port connected to the management module.
- 3110G: Auto-sensing of speed on the 10/100/1000 ports and auto-negotiation of duplex mode on the ports for optimizing bandwidth.
- 3110X: Fixed 10 Gbps speed on external 10 Gb Ethernet port for maximum uplink bandwidth.
- Up to 64 Gbps of throughput in a switch stack.
- Gigabit EtherChannel (3110G) or 10 Gb EtherChannel (3110X) for enhanced fault tolerance and to provide up to 8 Gbps (3110G) or 80 Gbps (3110X) of bandwidth between switches, routers, and servers.
- Forwarding of Layer 2 frames and Layer 3 packets at 1 Gbps line rate across switches in stack.
- Web Cache Communication Protocol (WCCP) for redirecting traffic to wide area application engines, for enabling content requests to be fulfilled locally, and for localizing Web traffic patterns in the network (supported by IP Services feature set only).
- Support for 1005 total VLANs. These VLANs can be any VLAN ID from 1–4094, except 1001–1005, which are reserved by Cisco.
- Layer 3 features such as Cisco’s own HSRP for Layer 3 router redundancy, RIP Versions 1 and 2, OSPF (IP services feature set is required), Enhanced IGRP (EIGRP) (IP services feature set is required), Border Gateway Protocol (BGP) Version 4 (IP services feature set is required), Static IP routing for manually building a routing table of network path information and IPv6 support.

Figure 2-36 shows the 3110X module.

Figure 2-36 Cisco Catalyst 3110X module for IBM BladeCenter

For more information on the Cisco 3110 Series switches switch, see this website:

http://www.redbooks.ibm.com/Redbooks.nsf/RedbookAbstracts/tips0752.html

2.17 Cisco Storage Area Network and FCoE Hardware

This section provides a high level description of the Cisco MDS Product Line for Storage Area Networking. The MDS product line is also OEM'd by IBM within the Systems Storage portfolio and is available using IBM part numbers.

In this section, we cover the following Cisco Storage Area Networking platforms:

- Cisco MDS 9124 - fixed form factor
- Cisco MDS 9148 - fixed form factor
- Cisco MDS 9222i - fixed form factor
- Cisco MDS 9506, 9509, and 9513 - modular platforms
- Cisco MDS 9710 - modular platform

Cisco Storage Area Networking switches are also available for the IBM BladeCenter as blade switches form factor modules at 4 Gb in a 20-port and a 10-port versions. These items will not be covered in detail in this publication. For additional information, see this website:

http://www.redbooks.ibm.com/abstracts/tips0753.html?Open

Figure 2-37 shows an overview of the Cisco MDS Product Line, which is covered in detail in the next subsection. From top left, the modular, chassis-based platforms are shown: the MDS 9506, the 9513, 9509, and the 9710. From the bottom left, the fixed form factor switches are shown: the 9124, the 9248, and the 9222i at the far right.

Figure 2-37 Cisco MDS product line

The Cisco MDS product line shares the following design points across the entire portfolio:

Cisco Storage Solutions include these benefits and features:

- Unified OS: The NX-OS (since version 4) runs on all Cisco MDS 9000 family switches, from multilayer fabric switches to multilayer directors. Using the same base system software across the entire Cisco Data Center product line enables Cisco Systems to provide an extensive, consistent, and compatible feature set across the Cisco MDS 9000 family and the Cisco Nexus family.
- Multiprotocol storage networking support: In addition to supporting Fibre Channel Protocol (FCP), the NX-OS also supports IBM Fibre Connection (FICON), Small Computer System Interface over IP (iSCSI), and Fibre Channel over IP (FCIP) in a single platform. Native iSCSI support in the Cisco MDS 9000 family helps customers to consolidate storage for a wider range of servers into a common pool on the SAN. Native FCIP support allows customers to take advantage of their existing investment in IP networks for cost-effective business-continuance solutions for both Fibre Channel and FICON environments. With NX-OS multiprotocol support, customers can better use their enterprise resources, thereby lowering costs.
- Virtual Storage Area Networking and Virtual SANs: Virtual SAN (VSAN) technology partitions a single physical SAN into multiple VSANs. The Cisco MDS 9000 family switches are the first SAN switches on the market with VSAN support built into the switch hardware. VSAN capabilities allow the NX-OS to logically divide a large physical fabric into separate isolated environments to improve Fibre Channel SAN scalability, availability, manageability, and network security. For FICON, VSANs help ensure that there is true hardware-based separation of FICON and open systems.
- High availability features:

  Nondisruptive Software Upgrades**:** NX-OS provides nondisruptive software upgrades for director-class products with redundant hardware and minimally disruptive upgrades for the fabric switches that do not have redundant supervisor engine hardware.

  Stateful Process Failover**:** The NX-OS automatically restarts failed software processes and provides stateful supervisor engine failover to help ensure that any hardware or software failures on the control plane do not disrupt traffic flow in the fabric.
- Management: Cisco DCNM can be leveraged also to manage the SAN environment together with the Data Center LAN environment in a single pane of glass view. For more information on DCNM, see Chapter 10, “Data center network management and automation” on page 499.

For more information on Fibre Channel and Fibre Channel over Ethernet, see Chapter 7, “Convergence of LAN and SAN: Fibre Channel over Ethernet” on page 369.

For more detailed information on the common features of the MDS product line, see this website:

http://www.cisco.com/en/US/prod/collateral/ps4159/ps6409/ps5989/ps6217/product_data_sheet09186a00801bcfd8_ps4358_Products_Data_Sheet.html

In Storage Area Networking, it is very important to qualify the hardware for compatibility with third party server, OS, and storage platforms. For detailed information on the compatibility matrix for the Cisco MDS platforms with IBM Systems, see this website:

http://public.dhe.ibm.com/common/ssi/ecm/en/ts312350usen/TS312350USEN.PDF

2.17.1 Cisco MDS 9124

The MDS 9124 is a 1 RU Fibre Channel switch with 24 ports capable of speeds of 4, 2, and 1 Gbps. It supports the same industry-leading Cisco MDS 9000 NX-OS Software that is supported on the entire Cisco MDS 9000 family of products.

The following additional capabilities are included:

- Buffer credits: Up to 64 for a group of four ports, with a default of 16 buffer credits per port.
- Ports per chassis: Up to 24 (base configuration with eight ports; additional ports in eight-port increments with the port activation license).
- PortChannels: Up to 16 ports per PortChannel.
- Offers nondisruptive software upgrades, optional dual-redundant, hot-swappable AC or DC power supplies (with integrated fans), VSANs for fault isolation, and optional dual-redundant VSANs for fault isolation, and PortChannels for Inter-Switch Link (ISL) resiliency.
- The Cisco MDS 9124 supports RADIUS and TACACS+, port security, fabric binding, Fibre Channel Security Protocol (FC-SP) host-to-switch and switch-to-switch authentication, Secure FTP (SFTP), Secure Shell Version 2 (SSHv2) and Simple Network Management Protocol Version 3 (SNMPv3) implementing Advanced Encryption Standard (AES), VSANs, hardware-enforced zoning, broadcast zones, and per-VSAN role-based access control (RBAC).

While the focus of the Cisco MDS 9124 is as the foundation of medium-sized SMB SANs, it can also be configured to participate as a component of a tiered enterprise SAN with other members of the Cisco MDS 9000 family. Cisco MDS 9000 SAN-OS firmware provides enterprise-class capabilities such as virtual SANs (VSANs), Port Channels, quality of service (QoS) and security for deployment in core-edge enterprise SAN solutions. These capabilities help provide investment protection as SAN requirements evolve and grow over time.

Fabric connectivity capabilities can be the basis for infrastructure simplification solutions for IBM System i, System p, and System x server and storage consolidation and high-availability server clustering with IBM System Storage disk storage arrays. Business continuity capabilities can help businesses protect valuable data with IBM System Storage tape libraries and devices and IBM Tivoli Storage Manager data protection software.

More information on this product can be found here:

http://www-03.ibm.com/systems/networking/switches/san/ctype/9124/

2.17.2 Cisco MDS 9148

The MDS 9148 is a 1 RU Fibre Channel switch with 48 line-rate 8 Gbps ports, also supporting 1/2/4Gbps connectivity. It supports the same industry-leading Cisco MDS 9000 NX-OS Software that is supported on the entire Cisco MDS 9000 family of products.

The following additional capabilities are included:

- Each port group consisting of 4 ports has a pool of 128 buffer credits, with a default of 32 buffer credits per port. When extended distances are required, up to 125 buffer credits can be allocated to a single port within the port group. This extensibility is available without additional licensing.
- PortChannels allow users to aggregate up to 16 physical Inter-Switch Links (ISLs) into a single logical bundle, providing load-balancing bandwidth use across all links. The bundle can consist of any port from the switch, helping ensure that the bundle remains active even in the event of a port group failure.
- Ports per chassis: On-Demand Port Activation license allows the addition of eight 8-Gbps ports. Customers have the option of purchasing preconfigured models of 16, 32, or 48 ports and upgrading the 16- and 32-port models on-site all the way to 48 ports by adding these license
- The Cisco MDS 9148 provides a comprehensive framework to protect highly sensitive data crossing today's enterprise networks. Included as standard with every Cisco MDS 9148 are features such as VSAN isolation, role-based access control (RBAC), secure FTP (SFTP), intelligent packet inspection at the port level, hardware zoning using access control lists (ACLs), extended broadcast zoning, secure shell (SSH), control-plane security, simple network management protocol (SNMP), and FC-SP switch-to-switch plus host-to-switch access authentication.
- The Cisco MDS 9124 supports RADIUS and TACACS+, port security, fabric binding, Fibre Channel Security Protocol (FC-SP) host-to-switch and switch-to-switch authentication, Secure FTP (SFTP), Secure Shell Version 2 (SSHv2) and Simple Network Management Protocol Version 3 (SNMPv3) implementing Advanced Encryption Standard (AES), VSANs, hardware-enforced zoning, broadcast zones, and per-VSAN role-based access control (RBAC).

With the ability to expand from 16 to 48 ports in eight-port increments, the Cisco MDS 9148 can be used as the foundation for small, stand-alone SANs, as a Top-of-Rack switch, or as an edge switch in larger, core-edge storage area network (SAN) infrastructures.

The Cisco MDS 9148 Multilayer Fabric Switch is designed to support quick configuration with zero-touch plug-and-play features and task wizards that allow it to be deployed quickly and easily in networks of any size. Powered by Cisco MDS 9000 NX-OS Software, it includes advanced storage networking features and functions, and is compatible with Cisco MDS 9000 Series Multilayer Directors and Switches, providing transparent, end-to-end service delivery in core-edge deployments.

Fabric connectivity can be the foundation for infrastructure simplification solutions with IBM System i, IBM System p and IBM System x servers. It also supports storage consolidation and high-availability clustering with System Storage disk arrays. Business continuity capabilities can help businesses protect valuable data with System Storage tape subsystems and IBM Tivoli Storage Manager data protection software.

More information on this product can be found here:

http://www-03.ibm.com/systems/networking/switches/san/ctype/9148/

2.17.3 Cisco MDS 9222i

The Cisco MDS 9222i provides unique multilayer and multiprotocol functions in a compact three-rack-unit (3RU) form factor. It provides 18 fixed 4-Gbps Fibre Channel ports for high-performance SAN connectivity and 4 fixed Gigabit Ethernet ports for Fibre Channel over IP (FCIP) and Small Computer System Interface over IP (iSCSI) storage services. It supports the same industry-leading Cisco MDS 9000 NX-OS Software that is supported on the entire Cisco MDS 9000 family of products.

The following additional capabilities are included:

- High-performance ISLs: Cisco MDS 9222i supports up to 16 Fibre Channel inter-switch links (ISLs) in a single PortChannel. Links can span any port on any module in a chassis for added scalability and resilience. Up to 4095 buffer-to-buffer credits can be assigned to a single Fibre Channel port to extend storage networks over very long distances.
- The Cisco MDS 9222i supports the Cisco MDS 9000 4/44-Port 8-Gbps Host-Optimized Fibre Channel Switching Module in the expansion slot, with four of the Fibre Channel ports capable of running at 8 Gbps. Thus, the Cisco MDS 9222i cost-effectively scales up to 66 ports for both open systems and IBM Fiber Connection (FICON) mainframe environments. The Cisco MDS 9222i also supports a Cisco MDS 9000 4-Port 10-Gbps Fibre Channel Switching Module in the expansion slot.
- Intelligent application services engines: The Cisco MDS 9222i includes as standard a single application services engine in the fixed slot that enables the included SAN Extension over IP software solution package to run on the four fixed 1 Gigabit Ethernet storage services ports.
- The Cisco MDS 9222i can add a second services engine when a Cisco MDS 9000 18/4-Port Multiservice Module (MSM) is installed in the expansion slot. This additional engine can run a second advanced software application solution package such as Cisco MDS 9000 I/O Accelerator (IOA) to further improve the SAN Extension over IP performance and throughput running on the standard services engine.
- If even more intelligent services applications need to be deployed, the Cisco MDS 9000 16-Port Storage Services Node (SSN) can be installed in the expansion slot to provide four additional services engines for simultaneously running application software solution packages such as Cisco Storage Media Encryption (SME).
- Overall, the Cisco MDS 9222i switch platform scales from a single application service solution in the standard offering to a maximum of five heterogeneous or homogeneous application services, depending on the choice of optional service module that is installed in the expansion slot.
- Advanced FICON services: The Cisco MDS 9222i supports System z FICON environments, including cascaded FICON fabrics, VSAN-enabled intermix of mainframe and open systems environments, and N-port ID virtualization (NPIV) for mainframe Linux partitions. IBM Control Unit Port (CUP) support enables in-band management of Cisco MDS 9200 Series switches from the mainframe management console. FICON tape acceleration reduces latency effects for FICON channel extension over FCIP for FICON tape read and write operations to mainframe physical or virtual tape. This feature is sometimes referred to as tape pipe-lining. The Cisco MDS 9000 XRC Acceleration feature enables acceleration of dynamic updates for IBM z/OS Global Mirror, formerly known as XRC.
- In Service Software Upgrade (ISSU) for Fibre Channel interfaces: Cisco MDS 9222i promotes high serviceability by allowing Cisco MDS 9000 NX-OS Software to be upgraded while the Fibre Channel ports are carrying traffic.

- Comprehensive network security framework: The Cisco MDS 9222i supports RADIUS and TACACS+, Fibre Channel Security Protocol (FC-SP), Secure File Transfer Protocol (SFTP), Secure Shell (SSH) Protocol, Simple Network Management Protocol Version 3 (SNMPv3) implementing AES, VSANs, hardware-enforced zoning, ACLs, and per-VSAN role-based access control (RBAC). Additionally, the Gigabit Ethernet ports offer IP Security (IPsec) authentication, data integrity, and hardware-assisted data encryption for FCIP and iSCSI.
- IP Version 6 (IPv6) capable: The Cisco MDS 9222i supports IPv6 as mandated by the U.S. Department of Defense (DoD), Japan, and China. IPv6 support is provided for FCIP, iSCSI, and management traffic routed in-band and out-of-band.

The Cisco MDS 9222i is designed to address the needs of medium-sized and large enterprises with a wide range of storage area network (SAN) capabilities. It can be used as a cost-effective high performance SAN extension over IP router switch for midrange midsize customers in support of IT simplification, replication, disaster recovery and business continuity solutions. It can also provide remote site device aggregation and SAN extension connectivity to large customer data center directors.

While the Cisco MDS 9222i is the foundation of mid-sized business continuity solutions, it can also be configured as a component of global business continuity solution for remote site device aggregation for consolidated wide area network (WAN) connectivity to data center enterprise directors. Cisco MDS 9000 SAN-OS firmware provides enterprise-class capabilities such as virtual SANs (VSANs), PortChannel, quality of service (QoS) and security for deployment in this type of enterprise SAN solution. The Cisco MDS 9222i includes SAN Extension over IP and FCIP Inter-VSAN routing firmware, which is designed to isolate remote site and WAN link events from local site operations.

More information on this product can be found here:

http://www-03.ibm.com/systems/networking/switches/san/ctype/9222/

2.17.4 Cisco MDS 9250i

The Cisco MDS 9250i offers up to forty 16-Gbps Fibre Channel ports, two 1/10 Gigabit Ethernet IP storage services ports, and eight 10 Gigabit Ethernet Fibre Channel over Ethernet (FCoE) ports in a fixed two-rack-unit (2RU) form factor. The Cisco MDS 9250i connects to existing native Fibre Channel networks, protecting current investments in storage networks. The Cisco SAN Extension over IP application package license is enabled as standard on the two fixed 1/10 Gigabit Ethernet IP storage services ports, enabling features such as Fibre Channel over IP (FCIP) and compression on the switch without the need for additional licenses. Also, using the eight 10 Gigabit Ethernet FCoE ports, the Cisco MDS 9250i platform attaches to directly connected FCoE and Fibre Channel storage devices and supports multitiered unified network fabric connectivity directly over FCoE.

The following additional capabilities are included:

- SAN consolidation with integrated multiprotocol support: The Cisco MDS 9250i is available in a base configuration of 20 ports of 16-Gbps Fibre Channel for high-performance SAN connectivity, 2 ports of 10-Gigabit Ethernet for FCIP and Small Computer System Interface over IP (iSCSI) storage services, and 8 ports of 10 Gigabit Ethernet for FCoE connectivity.

- High-density Fibre Channel switch with 16-Gbps connectivity: The Cisco MDS 9250i scales up to 40 ports of 16-Gbps Fibre Channel in a fixed configuration switch. The base configuration comes with 20 ports of 16-Gbps Fibre Channel enabled for high-performance SAN connectivity, and it can be upgraded onsite to enable an additional 20 ports of 16-Gbps Fibre Channel by adding the Cisco MDS 9250i On-Demand Port Activation license. Additionally, the Cisco MDS 9250i cost-effectively scales up for IBM Fibre Connection (FICON) mainframe environments.
- Intelligent application services engine: The Cisco MDS 9250i includes as standard a single application services engine that enables the included Cisco SAN Extension over IP software solution package to run on the two fixed 1/10 Gigabit Ethernet storage services ports. The Cisco SAN Extension over IP package provides an integrated, cost-effective, and reliable business-continuance solution that uses IP infrastructure by offering FCIP for remote SAN extension, along with a variety of advanced features to optimize the performance and manageability of FCIP links.
- Hardware-based virtual fabric isolation with virtual SANs (VSANs) and Fibre Channel routing with Inter-VSAN Routing (IVR): VSANs and IVR enable deployment of large-scale multisite and heterogeneous SAN topologies. Integration into port-level hardware allows any port in a system or in a fabric to be partitioned into any VSAN. Included in the optional Cisco MDS 9000 Enterprise advanced software package, IVR provides line-rate routing between any of the ports in a system or in a fabric without the need for external routing appliances.
- Remote SAN extension with high-performance FCIP:
  - Simplifies data protection and business continuance strategies by enabling backup, remote replication, and other disaster-recovery services over WAN distances using open-standards FCIP tunneling.
  - Optimizes utilization of WAN resources for backup and replication by enabling hardware-based compression, hardware-based encryption, FCIP write acceleration, and FCIP tape read and write acceleration; up to 16 virtual Inter-Switch Link (ISL) connections are provided on the two 10 Gigabit Ethernet ports through tunneling.
  - Preserves Cisco MDS 9000 family enhanced capabilities, including VSANs, IVR, advanced traffic management, and network security across remote connections.
- Cost-effective iSCSI connectivity to Ethernet-attached servers:
  - Extends the benefits of Fibre Channel SAN-based storage to Ethernet-attached servers at a lower cost than is possible using Fibre Channel interconnect alone.
  - Increases storage utilization and availability through consolidation of IP and Fibre Channel block storage.
  - Through transparent operation, preserves the capability of existing storage management applications.
- Advanced FICON services: The Cisco MDS 9250i will support FICON environments, including cascaded FICON fabrics, VSAN-enabled intermix of mainframe and open systems environments, and N-port ID virtualization (NPIV) for mainframe Linux partitions. IBM Control Unit Port (CUP) support enables in-band management of Cisco MDS 9200 Series switches from the mainframe management console. FICON tape acceleration reduces latency effects for FICON channel extension over FCIP for FICON tape read and write operations to mainframe physical or virtual tape. This feature is sometimes referred to as tape pipelining. The Cisco MDS 9250i will also support the IBM Extended Remote Copy (XRC)3 Acceleration feature that enables acceleration of dynamic updates for IBM z/OS Global Mirror, formerly known as XRC.

- Cisco Data Mobility Manager (DMM) as a distributed fabric service: Cisco DMM is a fabric-based data migration solution that transfers block data nondisruptively across heterogeneous storage volumes and across distances, whether the host is online or offline.
- Platform for intelligent fabric applications: The Cisco MDS 9250i provides an open platform that delivers the intelligence and advanced features required to make multilayer intelligent SANs a reality, including hardware-enabled innovations to host or accelerate applications for data migration, storage backup, and data replication. Hosting or accelerating these applications in the network can dramatically improve scalability, availability, security, and manageability of the storage environment, resulting in increased utility and lower total cost of ownership (TCO).
- In Service Software Upgrade (ISSU) for Fibre Channel interfaces: Cisco MDS 9250i promotes high serviceability by allowing Cisco MDS 9000 NX-OS Software to be upgraded while the Fibre Channel ports are carrying traffic.
- Intelligent network services: Cisco MDS 9250i uses VSAN technology for hardware-enforced, isolated environments within a single physical fabric, access control lists (ACLs) for hardware-based intelligent frame processing, and advanced traffic management features such as fabric wide quality of service (QoS) to facilitate migration from SAN islands to enterprise wide storage networks.
- High-performance ISLs: Cisco MDS 9250i supports up to 16 Fibre Channel ISLs in a single PortChannel. Links can span any port on any module in a chassis for added scalability and resilience. Up to 256 buffer-to-buffer credits can be assigned to a single Fibre Channel port to extend storage networks over long distances.
- Comprehensive network security framework: The Cisco MDS 9250i supports RADIUS and TACACS+, Fibre Channel Security Protocol (FC-SP), Secure File Transfer Protocol (SFTP), Secure Shell (SSH) Protocol, Simple Network Management Protocol Version 3 (SNMPv3) implementing Advanced Encryption Standard (AES), VSANs, hardware-enforced zoning, ACLs, and per-VSAN role-based access control (RBAC). Additionally, the 10 Gigabit Ethernet ports offer IP Security (IPsec) authentication, data integrity, and hardware-assisted data encryption for FCIP and iSCSI.
- IP Version 6 (IPv6) capable: The Cisco MDS 9250i supports IPv6 as mandated by the U.S. Department of Defense (DoD), Japan, and China. IPv6 support is provided for FCIP, iSCSI, and management traffic routed in band and out of band.
- FIPS compliance: The Cisco MDS 9250i will be FIPS 140-2 compliant as mandated by the U.S. federal government.
- Sophisticated diagnostics: The Cisco MDS 9250i provides intelligent diagnostics, protocol decoding, and network analysis tools as well as integrated Cisco Call Home capability for added reliability, faster problem resolution, and reduced service costs.

2.17.5 Cisco MDS 9500 Series Multilayer Directors

Cisco MDS 9500 Series Multilayer Directors are high-performance, protocol-independent director-class SAN switches. These directors are designed to meet stringent requirements of enterprise data center storage environments.

Common capabilities across the entire MDS 9500 family include these:

- Offers scalability to 192, 336 and 528 maximum Fibre Channel port count, depending on the specific MDS 9500 platform, at 1, 2, 4, 8 and 10 Gbps Fibre Channel speed.
- Multilayer architecture transparently integrates Fibre Channel, Fibre Channel over Ethernet (FCoE), IBM FICON, Internet Small Computer System Interface (iSCSI), and Fibre Channel over IP (FCIP) in one system.

- 32- and 48-port 8 Gbps Advanced Fibre Channel switching modules designed to allow a port to be configured as either 1, 2, 4, 8 or 10 Gbps, consolidating all ports into the same Fibre Channel switching module.
- High-performance Inter-Switch Links (ISLs) that provide additional availability at the fabric level; PortChannel capability allows users to aggregate up to 16 physical links into one logical bundle.
- Supports all generations of Cisco MDS 9000 family switching modules, providing outstanding investment protection.

The Cisco MDS 9500 Series Multilayer Directors are director-class storage area networking (SAN) switches designed for deployment in scalable enterprise and service provider clouds to enable flexibility, resiliency and reliability. Layering a comprehensive set of intelligent features onto a high-performance, protocol-independent switch fabric, the MDS 9500 Series Multilayer Directors address the critical requirements of large virtualized data center storage environments such as high availability, security, scalability, ease-of-management, and simple integration of new technologies for extremely flexible data center SAN solutions.

Sharing the same operating system and management interface with other Cisco data center switches, the MDS 9500 Directors can help enable smooth deployment of unified fabrics with high-performance Fibre Channel and FCoE connectivity for low total cost of ownership (TCO).

The MDS 9500 Series support different type of modules with different purposes. The following modules are available:

- Supervisor modules - at least two for redundancy
- Power supplies - at least two for redundancy
- Fan modules
- Switching I/O modules
- Crossbar modules - with dedicated slots on the 9513 only
- Clock modules - with dedicated slots on the 9513 only

There are many different options in terms of I/O modules, or line cards, including these:

- Cisco MDS 9000 10-Gbps 8-Port FCoE Module (Sup2A required)
- Cisco MDS 9000 16-Port Storage Services Node
- Cisco MDS 9000 18/4-Port Multiservice Module
- Cisco MDS 9000 family 4-Port 10-Gbps Fibre Channel Switching Module
- Cisco MDS 9000 family 8-Gbps Advanced Fibre Channel Switching Modules
  - 32-Port 8-Gbps Advanced Fibre Channel Switching Module
  - 48-Port 8-Gbps Advanced Fibre Channel Switching Module
  - 32-Port 8-Gbps Advanced Fibre Channel Switching Module
- Cisco MDS 9000 family 8-Gbps Fibre Channel Switching Modules
  - 24-Port 8-Gbps Fibre Channel Switching Module
  - 48-Port 8-Gbps Fibre Channel Switching Module
  - 4/44-Port 8-Gbps Host-Optimized Fibre Channel Switching Module

There are three main types within the MDS 9500 Series multilayer directors:

- **Cisco MDS 9506:** A 7 RU, 6 slots modular platform (2 supervisor slots and 4 I/O slots plus 2 power supply bays). It combines nondisruptive software upgrades, stateful process restart/failover, and full redundancy of all major components for best-in-class availability. Supporting up to 192 Fibre Channel ports in a single chassis and up to 1152 Fibre Channel ports in a single rack, the Cisco MDS 9506 is designed to meet the requirements of large data center storage environments.

It supports the following interfaces:

  - 1/2/4/8-Gbps and 10-Gbps Fibre Channel: The Cisco MDS 9506 supports new Cisco MDS 9500 8-Gbps Advanced Fibre Channel switching modules as well as existing 10-Gbps and 8-Gbps MDS Fibre Channel switching modules for deployment in both open systems and FICON environments.
  - 10-Gbps Multihop FCoE: The Cisco MDS 9506 supports multihop FCoE, extending connectivity from FCoE/Fibre Channel fabrics to FCoE/Fibre Channel storage devices. With 10-Gbps Multihop FCoE switching modules, the Cisco MDS 9506 supports extension of Fibre Channel SAN to devices that are connected using FCoE protocol over Ethernet, thereby extending the powerful capabilities of intelligent services to unified fabric deployments.
  - 1/2/4/8-Gbps FICON: The Cisco MDS 9506 supports advanced FICON services including cascaded FICON fabrics, VSAN-enabled intermix of mainframe and open systems environments, and N_Port ID virtualization (NPIV) for mainframe Linux partitions. Control Unit Port (CUP) support enables in-band management of Cisco MDS 9000 family switches from mainframe management applications.

- **Cisco MDS 9509:** A 14 RU, 9 slots modular platform (2 supervisor slots and 7 I/O slots plus 2 power supply bays). It combines non-disruptive software upgrades, stateful process restart/failover, and full redundancy of all major components for best-in-class availability. Supporting up to 336 Fibre Channel ports in a single chassis and up to 1008 Fibre Channel ports in a single rack, the Cisco MDS 9509 is designed to meet the requirements of large data center storage environments. It supports the following interfaces:
  - 1/2/4/8-Gbps and 10-Gbps Fibre Channel: The Cisco MDS 9509 supports new Cisco MDS 9500 8-Gbps Advanced Fibre Channel switching modules as well as existing 10-Gbps and 8-Gbps MDS Fibre Channel switching modules for deployment in both open systems and FICON environments.
  - 10-Gbps Multihop FCoE: The Cisco MDS 9509 supports multihop FCoE, extending connectivity from FCoE/Fibre Channel fabrics to FCoE/Fibre Channel storage devices. With 10-Gbps Multihop FCoE switching modules, the Cisco MDS 9509 supports extension of Fibre Channel SAN to devices that are connected using FCoE protocol over Ethernet, thereby extending the rich capabilities of intelligent services to unified fabric deployments.
  - 1/2/4/8-Gbps FICON: The Cisco MDS 9509 supports advanced FICON services including cascaded FICON fabrics, VSAN-enabled intermix of mainframe and open systems environments, and N_Port ID virtualization (NPIV) for mainframe Linux partitions. Control Unit Port (CUP) support enables in-band management of Cisco MDS 9000 family switches from mainframe management applications.
- **Cisco MDS 9513:** A 14 RU, 13 slots modular platform (2 supervisor slots and 11 I/O slots plus 2 power supply bays). It combines non-disruptive software upgrades, stateful process restart and failover, and full redundancy of all major components for best-in-class availability. With 8.4 terabits per second (Tbps) of system bandwidth and up to 528 1/2/4/8-Gbps auto-sensing Fibre Channel ports in a single chassis or up to 1584 Fibre Channel ports in a single rack, the Cisco MDS 9513 leads the industry in scalability and is designed to meet the requirements of the largest data center storage environments.

  It supports the following interfaces:
  - 1/2/4/8-Gbps and 10-Gbps Fibre Channel: The Cisco MDS 9513 supports both 1/2/4/8-Gbps and 10-Gbps ports on the new 8-Gbps Advanced Fibre Channel switching modules as well as existing 10-Gbps and 8-Gbps Cisco MDS 9000 family Fibre Channel switching modules for deployment in both open systems and FICON environments.

- 10-Gbps Multihop FCoE: The Cisco MDS 9513 supports multihop FCoE, extending connectivity from FCoE/Fibre Channel fabrics to FCoE/Fibre Channel storage devices. With 10-Gbps Multihop FCoE switching modules, the Cisco MDS 9513 supports extension of a Fibre Channel SAN to devices that are connected using FCoE protocol over Ethernet, thereby extending the rich capabilities of intelligent services to unified fabric deployments.
- 1/2/4/8-Gbps and 10-Gbps FICON: The Cisco MDS 9513 supports advanced FICON services including cascaded FICON fabrics, VSAN-enabled intermix of mainframe and open systems environments, and N_Port ID virtualization (NPIV) for mainframe Linux partitions. Cisco Control Unit Port (CUP) support enables in-band management of Cisco MDS 9000 family switches from mainframe management applications.

More information can be found here:

http://www.ibm.com/systems/networking/switches/san/ctype/9500/

2.17.6 Cisco MDS 9710 Series Multilayer Director

The Cisco® MDS 9710 Multilayer Director is a director-class SAN switch designed for deployment in large-scale storage networks to enable enterprise clouds and business transformation. Layering a comprehensive set of intelligent features onto a high-performance, protocol-independent switch fabric, the Cisco MDS 9710 addresses the stringent requirements of large virtualized data center storage environments: uncompromising high availability, security, scalability, ease of management, and transparent integration of new technologies for extremely flexible data center SAN solutions. Sharing the same operating system and management interface with other Cisco data center switches, the Cisco MDS 9710 enables seamless deployment of fabrics with high-performance Fibre Channel, IBM Fibre Connectivity (FICON), and Fibre Channel over Ethernet (FCoE) connectivity to achieve low total cost of ownership (TCO).

The following additional capabilities are included:

- Outstanding SAN performance: The combination of the 16-Gbps Fibre Channel switching module and the Fabric-1 crossbar switching modules enables up to 1.5 Tbps of front-panel Fibre Channel throughput between modules in each direction for each of the eight Cisco MDS 9710 payload slots. This per-slot bandwidth is twice the bandwidth needed to support a 48-port 16-Gbps Fibre Channel module at full line rate. Cisco MDS 9710 architecture, based on central arbitration and crossbar fabric, provides 16-Gbps line-rate, non-blocking, predictable performance across all traffic conditions for every port in the chassis.
- High availability: The Cisco MDS 9710 provides outstanding availability and reliability. The Cisco MDS 9710 enables redundancy on all major components, including the fabric card. It provides grid redundancy on the power supply and 1+1 redundant supervisors. Users can add an additional fabric card to enable N+1 fabric redundancy. The Cisco MDS 9710 combines non-disruptive software upgrades, stateful process restart and failover, and full redundancy of all major components for best-in-class availability.
- Scalability: With up to 24 terabits per second (Tbps) of Fibre Channel system bandwidth and 384 2/4/8-Gbps, 4/8/16-Gbps or 10-Gbps full line-rate autosensing Fibre Channel ports in a single chassis or up to 1152 Fibre Channel ports in a single rack, the Cisco MDS 9710 is designed to meet the requirements of the largest data center storage environments.

- Intelligent network services: VSAN technology, access control lists (ACLs) for hardware-based intelligent frame processing, and fabric wide quality of service (QoS) enable migration from SAN islands to enterprise-wide storage networks.
  - Integrated hardware-based VSANs and Inter-VSAN Routing (IVR): Integration of VSANs into port-level hardware allows any port in a system or fabric to be partitioned to any VSAN. Integrated hardware-based IVR provides line-rate routing between any ports in a system or fabric without the need for external routing appliances.
  - Intelligent storage services: The Cisco MDS 9710 interoperates with intelligent service capabilities on other Cisco MDS 9000 family platforms and the intelligent services switch to provide services such as acceleration of storage applications for data replication and backup and data migration to hosts and targets attached to the Cisco MDS 9710.
  - Smart Zoning: When the Smart Zoning feature is enabled, Cisco MDS 9700 Series multi-protocol Director fabrics provision the hardware access control entries specified by the zone set more efficiently, avoiding the superfluous entries that would allow servers (initiators) to talk to other servers, or allow storage devices (targets) to talk to other storage devices. This feature makes larger zones with multiple initiators and multiple targets feasible without excessive consumption of hardware resources. Thus, smart zones can correspond to applications, application clusters, hypervisor clusters, or other data center entities, saving the time that administrators previously spent creating many small zones, and enabling the automation of zoning tasks.
- Virtual machine transparency: The Cisco MDS 9700 Series provides deterministic hardware performance and a comprehensive feature set that allows virtual machines to have the same SAN attributes as a physical server. On a per-virtual machine basis, the Cisco NX-OS Software offers VSANs, QoS policies, access control, performance monitoring, and data protection to promote the scalability and mobility of virtual machines. Cisco Prime Data Center Network Manager provides end-to-end visibility all the way from the virtual machine down to the storage, with resource allocation, performance measurements, and predictions available on a per-virtual machine basis to enable rapid troubleshooting in mission-critical virtualized environments.
- Comprehensive security: In addition to support for services such as VSANs, hardware-enforced zoning, ACLs, per-VSAN role-based access control (RBAC), and Cisco TrustSec® Fibre Channel link encryption, the Cisco MDS 9700 Series supports a comprehensive security framework consisting of RADIUS and TACACS+, Fibre Channel Security Protocol (FC-SP), Secure File Transfer Protocol (SFTP), Secure Shell (SSH) Protocol, and Simple Network Management Protocol Version 3 (SNMPv3). Cisco TrustSec Fibre Channel link encryption delivers transparent, hardware-based 16-Gbps line-rate encryption of Fibre Channel data on 16-Gbps Fibre Channel switching modules in addition to 10-Gbps line-rate encryption.
- SAN management: The Cisco MDS 9700 Series includes built-in storage network management with all features available through a command-line interface (CLI) or Cisco Prime Data Center Network Manager (DCNM), a centralized management tool that simplifies management of unified fabrics. Cisco DCNM supports integration with third-party storage management applications to allow seamless interaction with existing management tools. Cisco DCNM supports federation of up to 10 Cisco DCNM servers to manage up to 150,000 devices using a single management pane.

- Sophisticated diagnostics: The Cisco MDS 9710 provides intelligent diagnostics, protocol decoding, and network analysis tools as well as integrated Cisco Call Home capability for added reliability, faster problem resolution, and reduced service costs. Starting with Cisco MDS 9000 NX-OS 6.2, the Cisco Generic Online Diagnostics (GOLD) framework replaces the Cisco Online Health Management System (OHMS) diagnostic framework on the new Cisco MDS 9700 Series Multilayer Director chassis. Cisco GOLD is a suite of diagnostic facilities to verify that hardware and internal data paths are operating as designed. Boot-time diagnostics, continuous monitoring, standby fabric loopback tests, and on-demand and scheduled tests are part of the Cisco GOLD feature set. This diagnostics subsystem enables the rapid fault isolation and continuous system monitoring critical in today's continuously operating environments.
- Multiprotocol architecture: The multilayer architecture of the Cisco MDS 9700 Series enables a consistent feature set over a protocol-independent switch fabric. The Cisco MDS 9710 transparently integrates Fibre Channel, FCoE1 and FICON1.
  - 2/4/8-Gbps, 4/8/16-Gbps, 10-Gbps Fibre Channel and 10 Gigabit Ethernet: The Cisco MDS 9710 supports 2/4/8/16-Gbps and 10-Gbps ports on the Cisco MDS 9700 48-Port 16-Gbps Fibre Channel Switching Module for deployment in both open systems and FICON environments. The Cisco MDS 9710 also supports 10 Gigabit Ethernet clocked optics carrying Fibre Channel traffic.
  - 2/4/8-Gbps, 4/8/16-Gbps and 10-Gbps FICON1: The Cisco MDS 9710 is mainframe-ready, with full support for IBM System z FICON and Linux environments.
  - Multihop FCoE: The Cisco MDS 9710 supports multihop FCoE, extending connectivity from FCoE and Fibre Channel fabrics to FCoE and Fibre Channel storage devices.

More information can be found here:

http://www.cisco.com/en/US/prod/collateral/ps4159/ps6409/ps12970/data_sheet_c78-727769.html

2.18 Power and cooling requirements

This section provides additional information on the power and cooling requirements of Cisco LAN and SAN hardware.

2.18.1 Power and cooling for the Nexus 7000

Because of its modular design the Nexus 7000 can have many power combination depending of the chassis size, types of modules and number of power supplies. Cisco has created an online power calculator tool to help administrators define their power and cooling needs. The Cisco Power Calculator enables data center planer to calculate the power supply requirements for a specific switch configuration. The results will show output current, output power, and system heat dissipation. The Cisco Power Calculator supports the following Cisco product switching and routing platforms: Cisco Nexus 7000, Cisco Catalyst 6500 Series, Catalyst 4500-E/4500 Series, Catalyst 3750-E/3750 Series, Catalyst 3560-E/3560 Series, Catalyst 2960 Series, Catalyst 2975 Series, Catalyst Express 500 Series, and the Cisco 7600 Series Router. It can be found at:

http://tools.cisco.com/cpc/

The power calculator requires a CCO login.

2.18.2 Power and cooling for the Nexus 5500

Table 2-17 has the typical power and cooling requirements for the Nexus 5500 product lines.

Table 2-17 Nexus 5500 power and cooling requirements

AC power supply properties	Nexus 5548P and 5548UP	Nexus 5596UP	Nexus 5596T
Typical operating power	390W	660W	900W
Maximum power (Layer 2)	600W (without Layer 3 daughter card)	882W (without Layer 3 daughter card)	1050W (with 3x 10G BASE-T (Cu) expansion module)
Maximum power (Layer 3)	730W (with Layer 3 daughter card)	882W (with Layer 3 daughter card)	1050W (with 3x 10G BASE-T (Cu) expansion module)
Input voltage	100 to 240 VAC	100 to 240 VAC	100 to 240 VAC
Frequency	50 to 60 Hz	50 to 60 Hz	50 to 60 Hz
Efficiency	95 to 98% (50 to 100% load)	95 to 98% (50 to 100% load)	95 to 98% (50 to 100% load)
RoHS compliance	Yes	Yes	Yes
Hot swappable	Yes	Yes	Yes
Heat dissipation	1998 BTU/hr (600W)	3010 BTU/hr (882W)	3583 BTU/hr (1050W)
Front-to-back air flow power supply	Yes	Yes	Yes
Back-to-front air flow power supply	Yes (only on 5548UP; need fan tray as well)	Yes	Yes

2.18.3 Power and cooling for the Nexus 3000

Table 2-18 has the typical power and cooling requirements for the Nexus 3000 product lines.

Table 2-18 Nexus 3000 power and cooling requirements

AC power supply properties	Nexus 3016	Nexus 3048	Nexus 3064T	Nexus 3064X
Typical operating power	172 watts (W; with Twinax at 100% load; 2 power supply units [PSUs]) 174W (with short-reach optics at 100% load; 2 PSUs)	120 watts (W; (48p of 1G and 4p of 10G/SR at 100% load, with 2 power supply units [PSUs]**)**	362W (48p with 3m cables; 4 SR4 at 100% load)	143 watts (W; 64p with Twinax at 100% load; 2 power supply units [PSUs]) 177W (64p with SR optics at 100% load; 2 PSUs)
Maximum power	227W	124W	455W	199W
Input voltage	100 to 240 VAC -40 to -72 VDC	100 to 240 VAC -40 to -72 VDC	100 to 240 VAC -40 to -72 VDC	100 to 240 VAC -40 to -72 VDC
Frequency (AC)	50 to 60 Hz	50 to 60 Hz	50 to 60 Hz	50 to 60 Hz
Max current (DC)	33A	33A	33A	33A
Efficiency	89 to 91% at 220V	89 to 91% at 220V	89 to 91% at 220V	89 to 91% at 220V
RoHS compliance	Yes	Yes	Yes	Yes
Hot swappable	Yes	Yes	Yes	Yes
Heat dissipation	587 BTU/hr (16p with Twinax at 100% load; 2 PSUs) 594 BTU/hr (16p with SR4 optics at 100% load; 2 PSUs)	409 BTU/hr (48p of 1G and 4p of 10G/SR at 100% load, with 2 PSUs)	1235 BTU/hr (48p with 3m cables; 4 SR4 at 100% load)	488 BTU/hr (64p with Twinax at 100% load; 2 PSUs) 605 BTU/hr (64p with SR optics at 100% load; 2 PSUs)
Front-to-back air flow power supply	Yes	Yes	Yes	Yes
Back-to-front air flow power supply	Yes	Yes	Yes	Yes

2.18.4 Power and cooling for the Nexus 3500

Table 2-19 has the typical power and cooling requirements for the Nexus 3500 product lines.

Table 2-19 Nexus 3500 Power and Cooling Requirements

AC power supply properties	**Nexus 3548**
Typical operating power	152W (48p with Twinax at 100% load; 2 power supply units (PSUs) at 25C) 180W (48p with SR optics at 100% load; 2 power supply units (PSUs) at 25C)
Maximum power	265W
Input voltage	100 to 240 VAC -40 to -72 VDC
Frequency (AC)	50 to 60 Hz
Max current (DC)	33A
Efficiency	89 to 91% at 220V
RoHS compliance	Yes
Hot swappable	Yes
Heat dissipation	519 BTUs per hr (48p with Twinax at 100% load; 2 power supply units (PSUs) at 25C) 614 BTUs per hr (48p with SR optics at 100% load; 2 power supply units (PSUs) at 25C)
Front-to-back air flow power supply	Yes
Back-to-front air flow power supply	Yes

2.18.5 Power and cooling for the Nexus 2000

Table 2-20 has the typical power and cooling requirements for the Nexus 2000 product lines.

Table 2-20 Nexus 2000 Power and Cooling Requirements

AC power supply properties	Nexus 2224TP	Nexus 2248TP	Nexus 2248TP-E	Nexus 2232PP
Typical operating power	80W	95W	95W	210W
Maximum power	95W	110W	110W	270W
Input voltage	100 to 240 VAC	100 to 240 VAC	100 to 240 VAC	100 to 240 VAC
Frequency	50 to 60 Hz	50 to 60 Hz	50 to 60 Hz	50 to 60 Hz
Efficiency	95 to 98% (50 to 100% load)	95 to 98% (50 to 100% load)	95 to 98% (50 to 100% load)	95 to 98% (50 to 100% load)
RoHS compliance	Yes	Yes	Yes	Yes
Hot swappable	Yes	Yes	Yes	Yes
Heat dissipation	201/282 BTU/hour (typical/ maximum)	322/403 BTU/hour (typical/ maximum)	322/403 BTU/hour (typical/ maximum)	806/1330 BTU/hour (typical/ maximum)
Front-to-back air flow power supply	Yes	Yes	Yes	Yes
Back-to-front air flow power supply	Yes	Yes	Yes	Yes

For power and cooling information on other Nexus 2000 Fabric Extenders, visit this website:

http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps10110/data_sheet_c78-507093.html

3

Designing networks for the modern data center

In this chapter we address critical network implications, which means that we describe specific network issues and requirements in virtualized data centers and elaborate on those issues by providing guidance and exploring possible solutions.

The process to develop the network design can no longer be done in isolation. It has to include key requirements from the other areas of the data center infrastructure including server, storage, security, and applications. The input can be gathered across six key areas of architectural considerations:

- Design scope: Means developing the central design principles based on present and planned application workloads, server and storage platform virtualization capabilities, IT services for internal and external customers, and anticipated schedules for growth and investment.
- Organization: Means addressing the existing boundaries between the network, servers and storage to establish better interfaces for operational teams to work effectively in virtualized and dynamic environments; creating a comprehensive design by requiring the network team to develop a greater operational awareness of server and storage virtualization technologies and capabilities.
- Technology: Means choosing from current and emerging networking technology options to make the best decision for executing your IT strategy while minimizing business disruption.
- Management: Means deciding on the right level of integration with the IT management environment to both enhance current operations and provide a path to emerging, policy-driven network configuration and provisioning standards.
- Security: Means balancing the introduction of evolving capabilities, such as virtualization and provisioning, to provide the optimal relationship between business demands and risk tolerance. Best practices for security management and integration can create a comprehensive view of the physical and virtualized infrastructure to provide robust security, visibility and segmentation.
- Automation: Means the automated approach to network and security configurations with the provisioning and orchestration of servers, storage, and facilities.

The use of virtualized servers implies change in some characteristics of the data center network. Beyond much more complex scalability issues due to a much higher number of servers (now virtualized), there is also the need to have mechanisms to assist the network in adapting dynamically to VM migration between different physical servers, which might be placed in geographically dispersed data centers.

Issues in the Layer 3 domain for highly virtualized data center networks also exist and need to be addressed. For example, due to the larger Layer 2 domains, the placement of Layer 3 needs to be carefully designed.

In this chapter, the following design considerations are addressed in more detail:

- The modular data center network design concept
- Data center fabric characteristics and solutions
- IPv6 and the data center
- Data center network throughput considerations
- Multi-tenancy
- Virtualized workload mobility between data centers
- High availability and fast failover
- First hop redundancy protocols
- Network services such as load balancing, firewall, WAN acceleration, and traffic analysis
- IP-based storage requirements
- Regulatory requirements
- Consistency between physical and virtual network
- Differentiated service levels on a shared infrastructure (QoS)
- Data center cabling

3.1 The modular data center network design concept

The concept of a modern data center network uses a modular and hierarchical topology to facilitate the following capabilities:

- Modular deployment of IT services using common, repeatable storage, network, and compute building blocks that can be added as needed
- Flexibility in end-system placement and connectivity (racking, cabling, port speed, number of ports, LAN/SAN convergence)
- Scalability of bandwidth and resilience
- Combination of clear aggregation points and segmentation allows for easy management
- Supporting services can be layered onto the data center design to support required common services (firewall/security, load balancing, instrumentation/monitoring, application delivery control and so on) where needed without redesigning the complete data center network architecture.
- Overlay of operational functionalities to deliver scalability, isolation, capacity management, and availability by segmentation
- Adaptable architecture that can adjust as technology changes over the next 5 to 10 years.
- Flexibility of architecture and the ability to absorb emerging technologies like Unified I/O (LAN/SAN convergence)
- Segmentation that is not bound to the physical data center infrastructure, enabling pervasive virtualization.
- Support of multi-tenancy via virtualization technologies without increasing the capital expense (CAPEX)
- Elasticity in a way that resources can scale up or down in a short period based on service-level agreements (SLAs). Elasticity enables resources on demand and scale resource utilization as needed.
- Automation management and provisioning of services hosted in the modular data center without jeopardizing the stability of the infrastructure

We use a modular data center network design based on the following units:

- Point of Delivery (PoD): A group of cabinets with a common set of network, compute, and potential co-located storage equipment including the structured cabling components.
- Service Zone: A logical group of PoDs, defined by fault and scalability domain, with common services/capabilities per Service Zone.

The following sections describe the characteristics of the units and illustrate why they are introduced as important design criteria.

3.1.1 The PoD concept

The introduction of a Point of Delivery (PoD) in relation to a modern data center design is something that can be observed more and more, but the term PoD is often used in different variations with different meanings.

A PoD is, in general, a physical repeatable pattern, and its components maximize the modularity, scalability, and manageability of data centers. The concept of a design pattern is nothing new in computer or network architecture, but the term PoD is often used in different variations, especially in combination with cloud services. As the scope of this book is far more than designing a data center network for cloud services, the definition of a PoD in this book might differ to documents which are referencing a virtualized, multi-tenant and fully automated data center design for specific services such as cloud services or massively scalable data centers.

We define a PoD as a physical, modular building block in a data center with the following characteristics:

- Collection of cabinets with a common set of network equipment
- Server and network hardware are co-resident in the same PoD
- Network functionality is typically Layer 2 only in the form of network access/aggregation
- Cabinets in the PoD are typically contiguous but are not limited to this due to special site characteristics
- All server NIC to network port connections stay within the PoD to optimize cabling
- All PoDs are connected to a data center network core

The characteristics for a PoD defined before are the common ones. It is recommended to define further possible and more granular characteristics, but this should be defined on a case by case basis depending on the data center specifics. The following ones should be considered additionally:

- Use of Top-of-Rack (ToR), Middle-of-Rack (MoR), or Bottom-of-Rack (BoR) installation of network devices
- Use of low cost copper cables for 10 GE server ports
- PoD internal placement of storage and storage network devices
- Use and placement of cable management patch panels
- Rack power and weight requirements
- Network device airflow should comply with data center hot/cold isle requirements
- Server types and connectivity in regards to the number of physical 1 GE and 10 GE interfaces to allow an efficient use of network resources

The following examples show different types of PoDs for illustration. All have the same common characteristics in slightly different deployment scenarios. For simplification, the storage shown in the pictures is utilizing Ethernet, but not native Fibre Channel.

The first example, in Figure 3-1, shows a PoD in which the storage devices are implemented in each PoD. Four physical network devices are installed ToR to provide redundant 10 GE and 1 GE ports for the servers.

These are the benefits of this approach:

- Servers and storage can be expanded by a rack
- Cables and storage traffic are closed in the PoD

These are the drawbacks of this approach:

- May not make efficient use of storage
- Distributed storage (data store) management

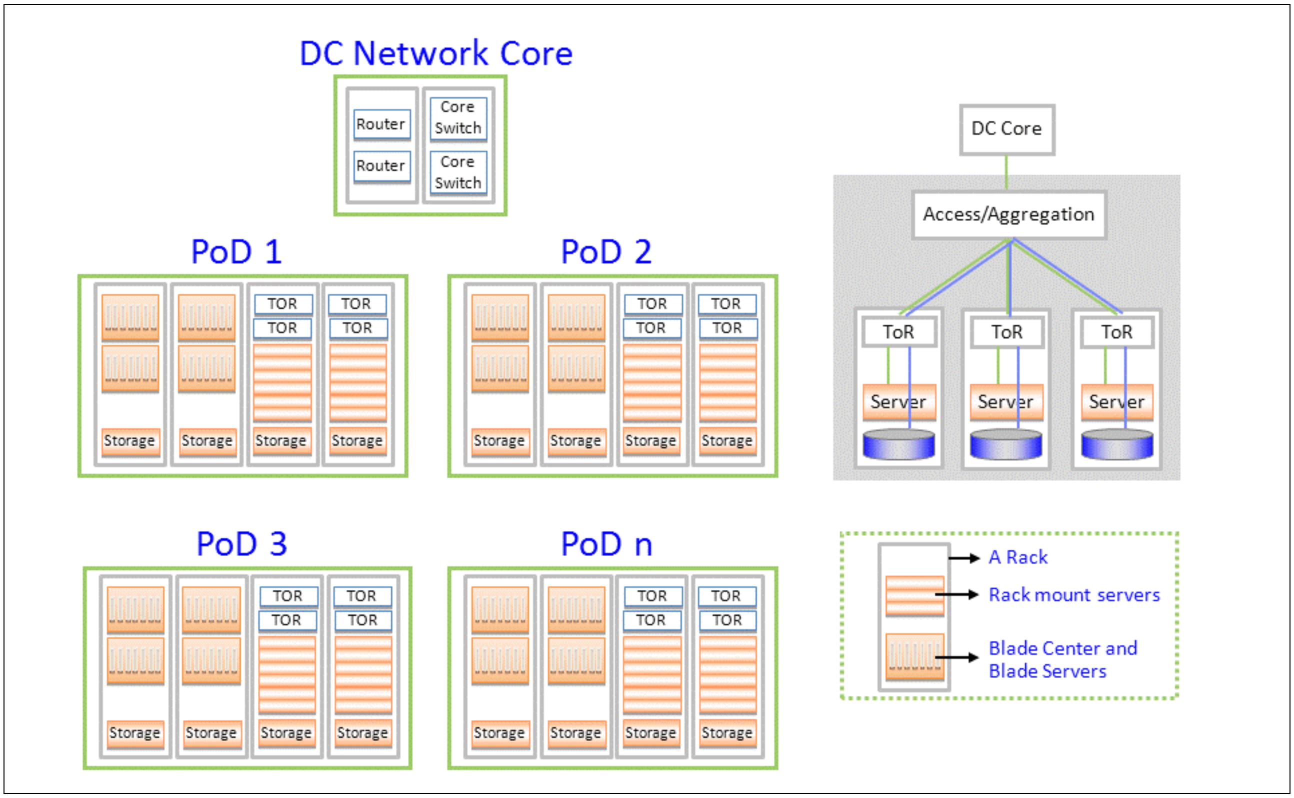


Figure 3-1 PoD example with integrated storage

The second example in Figure 3-2 shows the storage consolidated in a separate storage block. Four physical network devices are installed ToR to provide redundant 10 GE and 1 GE ports for the servers.

These are the benefits of this approach:

- Scale out storage easily
- Integrated storage management

These are the drawbacks of this approach:

- More cables are required
- More switch ports may be required
- Points of failure increase

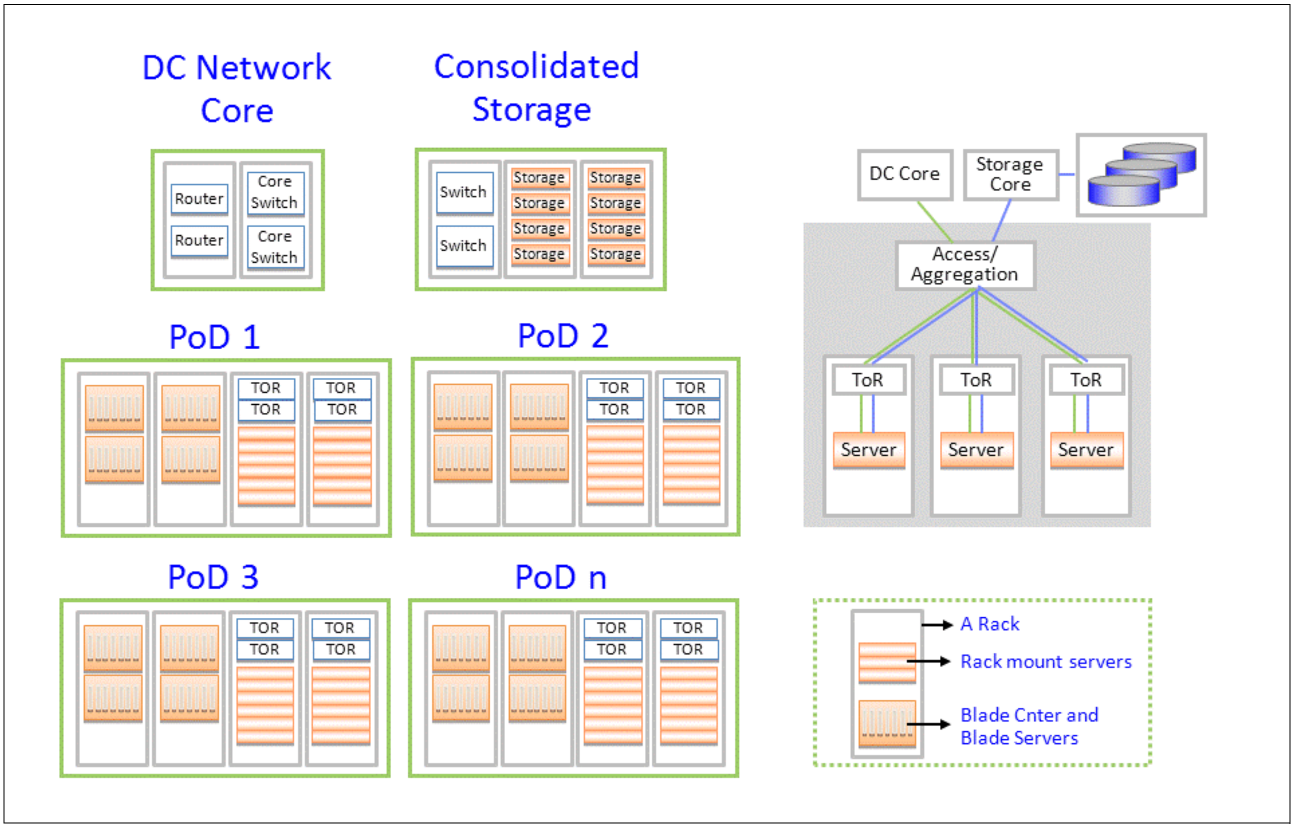


Figure 3-2 PoD example with separated storage block

The use of Top-of-Rack (ToR), Middle-of-Rack (MoR), or Bottom-of-Rack (BoR) switch placement lowers overall data center network costs and has the following additional benefits:

- It facilitates copper patch cables directly from servers to network switches within PoDs. This is becoming even more interesting and cost effective due to the 10GBase-T LAN on Motherboard (LoM) server shipment.
- Typically it reduces the need for centralized cable racks and patch panels.
- It significantly eliminates under floor copper cabling across the whole data center and supports the deployment of a modular cabling system.
- It provides improved airflow under raised floors.
- It reduces the amount of time to build the network.
- It simplifies troubleshooting of the physical infrastructure by demarcation.
- It enables graceful scaling of storage with modern converged storage platforms, simplifying backup and recovery architectures.

Initially, you implement the data center using the data center network core and consolidated storage block and add the PoDs over time to expand the data center depending on the demand of additional resources. This modular PoD-based architectures provide predictable resource pools, power, and space consumption.

The number of supported PoDs is among others dependent on the scalability limits of the data center network core. A network fabric design helps to increase the scalability limits, which is discussed in more detail later in this chapter.

3.1.2 The Service Zone concept

The introduction of a modular PoD concept in the data center addresses mainly physical design patterns, but there are other design requirements and use cases that have to be addressed in a modern data center design. The optional Service Zone concept is introduced to address these additional requirements and use cases, which may not be covered by the PoD concept. In some cases, the PoD will support some of the characteristics of a Service Zone, but IBM and Cisco have abstracted the Service Zone concept to ensure efficient scaling of the modular data center.

Figure 3-3 represents the interaction between the Service Zone and the PoD concept.

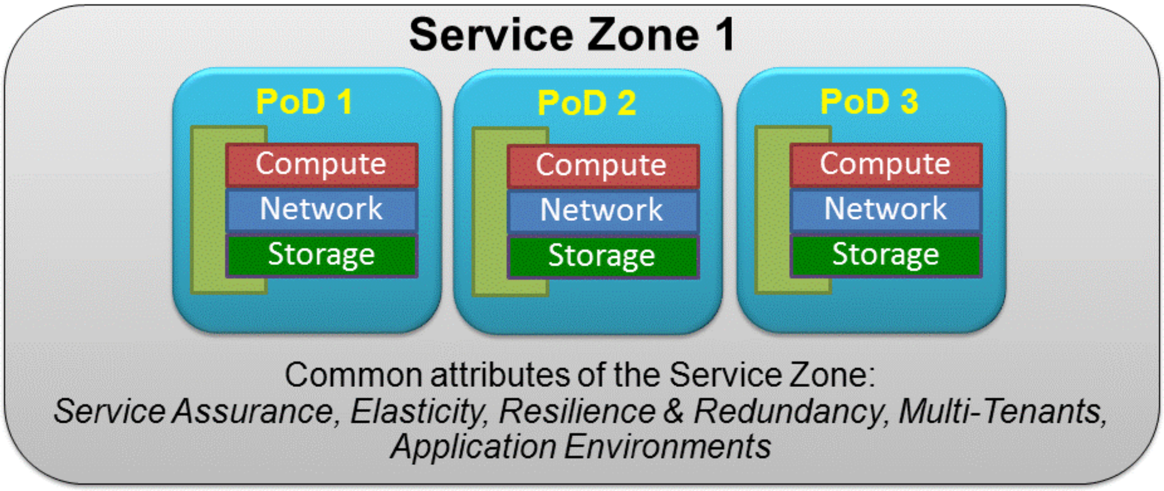

Service Zone 1
PoD 1
Compute
Network
Storage
PoD 2
Compute
Network
Storage
PoD 3
Compute
Network
Storage
Common attributes of the Service Zone:
Service Assurance, Elasticity, Resilience & Redundancy, Multi-Tenants, Application Environments

Figure 3-3 PoDs and Service Zones

The Service Zone concept applies to distinct application environments through a modular approach to building the physical, network, compute and storage infrastructure in a predictable and repeatable manner. It allows organizations to plan the rollout of distinct compute environment as needed in a shared physical data center using pay as you go model.

A Service Zone is a discrete, homogeneous, modular unit of data center components, which do not have to be physically placed together in the data center. The main purpose of the Service Zone is to support for incremental build-out of the data center that address environmental, physical, logical, operational and application requirements. In some cases, a multi-tier Layer 2 access model (or virtualization features within the access layer) may define the boundaries of a Service Zone differently. Many possible variations exist for using the Service Zone concept to scale the data center topology.

For example, Service Zones could be categorized as follows;

- Cloud Service Zone: This is a fully virtualized environment focused on workload mobility and application flexibility.
- Integrated Systems Service Zone: This is typically an “intelligent PoD” that includes a unified, distinct infrastructure platform that delivers compute, storage and network access. It may have a unified management or even automation stack.
- General Purpose Service Zone: This is an enterprise computing environment hosting multi-tiered applications (web, app, DB) and typically dedicated to a single customer with stateful services.

Next, we list some additional design requirements that need to be considered:

Service Assurance: This is a generally a defined set of procedures intended to optimize performance and provide service management guidance. Service assurance involves quality assurance, quality control and service level management processes. Quality assurance and control processes insure that a product or service meet specified requirements, adhering to a defined set of criteria which fulfill customer or client requirements. Service level management involves the monitoring and management of key performance indicators of a product or service. The fundamental driver behind service assurance is to maximize customer satisfaction.

Service Assurance has also to cope with maintenance tasks, lifecycle activities, capacity management and disaster recovery testing to ensure the stability and availability of the service. All these considerations have to be made in alignment with the data center design. With a focus on network characteristics, the following considerations have to be made:

- Synchronized maintenance windows
- In service software upgrades
- Quality of service

Scalability: This is a multi-dimensional attribute that refers to the ability to provision, manage, and operate the system. Elasticity is the ability to scale resources up or down in a short period based on service-level agreements (SLAs). Elasticity enables resources on demand and scale resource utilization from a Pod or collection of Pods as needed.

Capacity planning is an important aspect of the modular data center design. As no customer can build resources at infinite scale, the optimal and cost-effective solution adds resource pools to existing data centers based on projected usage without disrupting the existing network. The resource container concept, called a Service Zone, achieves elasticity, simplifies capacity planning, and does not disrupt the existing environment.

High availability: This is defined as the probability that a service or network is operational and functional as needed at any point in time. A highly available data center infrastructure is the foundation of SLA guarantee and successful infrastructure deployment. The Service Zone is typically governed by a common or group of common SLAs.

An end-to-end, highly available network infrastructure design provides predictable operational continuity. As organizations must satisfy SLAs made for business application uptime, they cannot afford to lose any connectivity due to an equipment failure. Therefore, the Service Zone design must ensure that a single failure either in hardware or software in the infrastructure does not affect the service. The following characteristics describe the design considerations that a modern data center design needs to leverage to enable highly available and resilient end-to-end infrastructure:

- Physical redundancy
- Network Layer high availability
- Compute Layer high availability
- Storage Layer high availability

Secure tenant separation: This calls for common resource pools to be efficiently shared among multiple tenants. For the purposes of this book, we collect a group of tenants together under that Service Zone concept.

Traditionally, IT departments deployed dedicated infrastructure for each tenant. This approach poses serious limitations on management of costs and complexity, and inefficiency in use of resources. Deploying multiple tenants in a common infrastructure yields more efficient resource use and lower costs. However, each tenant requires isolation for security and privacy from others sharing the common infrastructure. Therefore, secure separation is a fundamental requirement in multi-tenant environments. To achieve secure tenant separation and path isolation, we configure end-to-end virtualization of network, compute, and storage resources.

In a private cloud, a tenant is an enterprise department. In a public cloud, a tenant is an individual customer or enterprise who subscribes to cloud resources. In both public and private cloud scenarios, multi-tenancy provides better resource utilization but requires designing of secure tenant separation to ensure end-to-end security and path isolation within the shared infrastructure.

Additional to the compute separation, storage separation and application tier separation, the following network aspects have to be considered to ensure an end-to-end separation.

- Network Layer 2 separation
- Network Layer 3 separation
- Network services separation (firewall and load balancing services)
- Tenant path isolation
- Client-server traffic separation

Network Services: Figure 3-4 presents the access/aggregation-layer network, compute, and storage (file and/or block) resources in one data center Service Zone. Complementary network services are also shown in the picture, but not each and every Service Zone requires to have additional services inside the Service Zone. There could also be a special purpose Zone delivering network services for all Service Zones in the data center. The constitution of a Service Zone must be considered cross competency (server, network and storage) to meet the defined use cases and functional requirements.

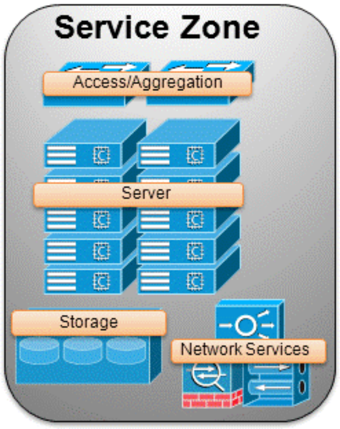


Figure 3-4 Service Zone

In deployment scenarios, where all Service Zones deliver the same type of services, the Service Zone resources must be consistently repeated to achieve predictable performance. We recommend consistent and balanced sizing of the network, compute, and storage resources within a Service Zone to achieve predictable performance. If a capacity threshold of >70% is breached, Service Zones could be extended by expanding the available resources (such as adding PoDs) or by creating a new Service Zone. Do not extend the resources in the second Service Zone relative to the first.

Multiple Service Zones can co-exist in the same data center. All Service Zones can either provide the same service or each and every Service Zone is providing different services. Common for all Service Zones is that they are all connected to the same Network Core infrastructure and rely on one or multiple physical PoDs. Figure 3-5 shows multiple Service Zones in a single room data center.

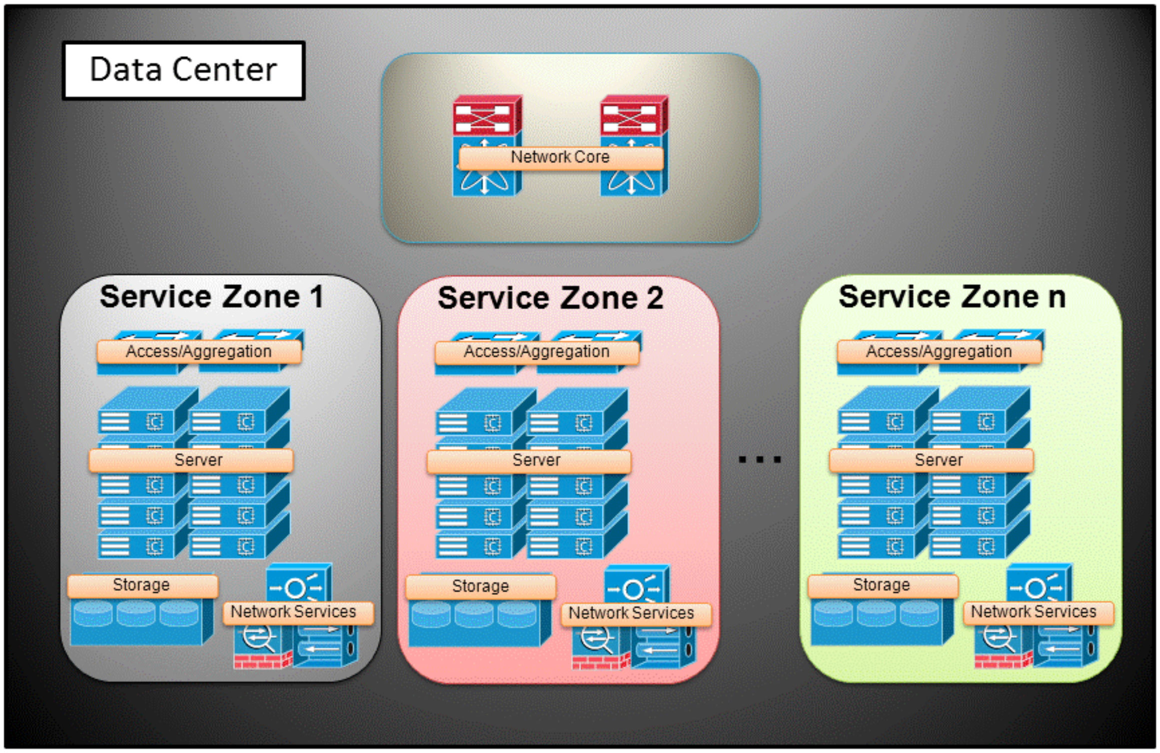


Figure 3-5 Multiple Service Zones in the same data center room

Another example shown in Figure 3-6 identifies that one data center could be split in either two separated rooms or even two data center Sites. Dark fiber is typically available to provide the interconnectivity between the two rooms or sites. Such a data center setup delivers a better resiliency in regards to complete room or site failures especially for power or cooling outages. The number of supported Service Zones compared to the single room data center is the same, but the key difference is that the Service Zone is now split between both rooms or sites in order to facilitate the give data center setup.

The network design is still following the design concept of a single data center, as we have only one network core which exists out of multiple core devices to avoid massive inter-room/site connectivity. Such a Service Zone design is often called Split-Service Zone.

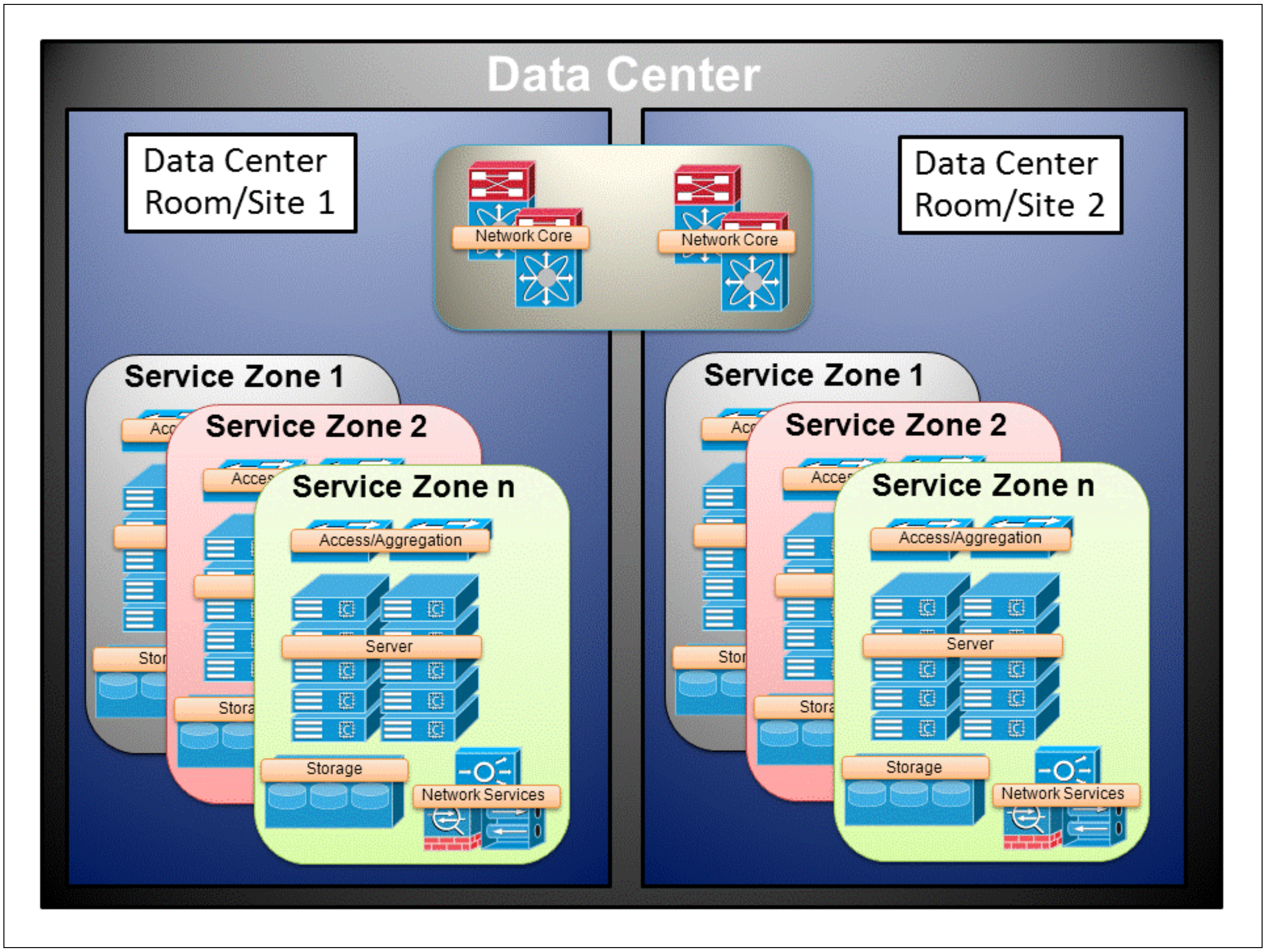


Figure 3-6 Split Service Zone in dual room/site data center

In cases where a Split-Service Zone must be stretched among two discrete data centers (that is, in order to meet data center disaster recovery requirements), Data Center Interconnect (DCI) technologies are required to span VLANs across the Wide Area Network.

Figure 3-7 shows that a Service Zone can also be stretched in a dual data center setup utilizing DCI technologies.

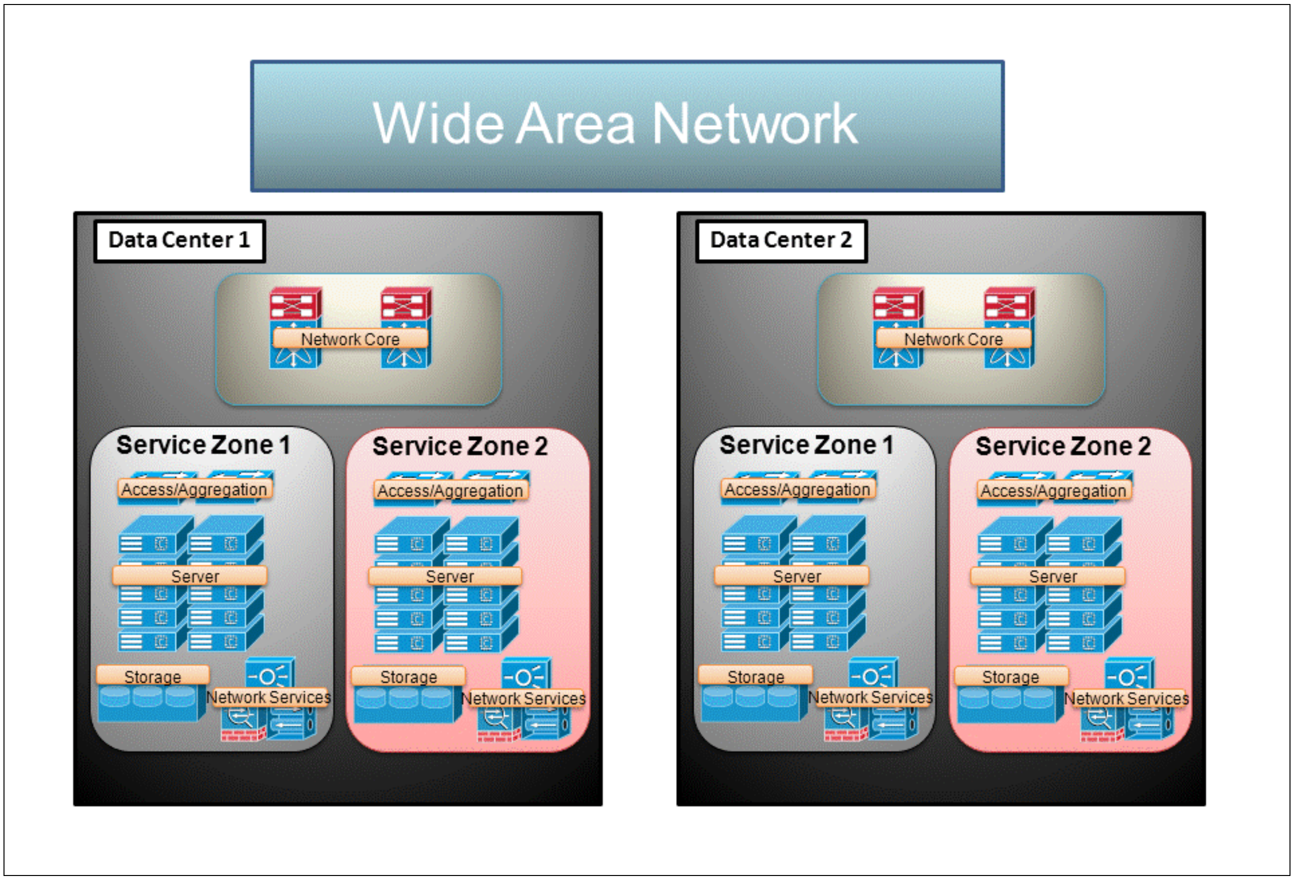


Figure 3-7 Split Service Zones in dual data center setup

3.2 Data center network design considerations

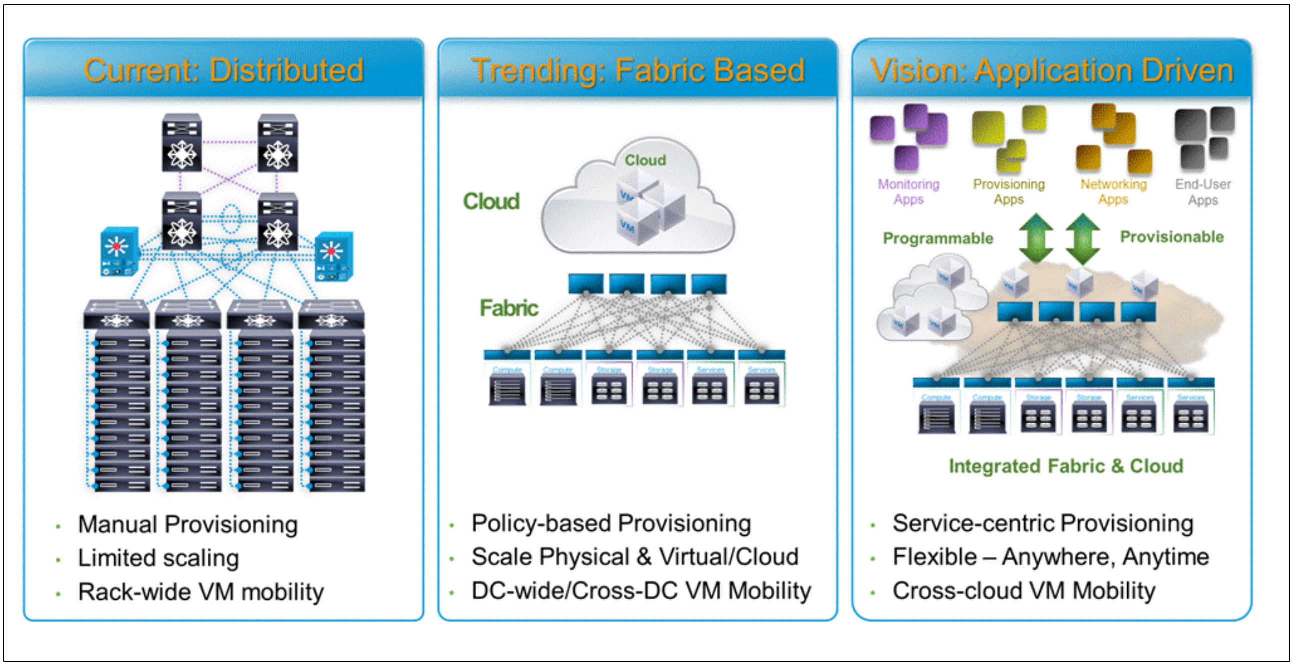
The drivers and trends have a direct impact on the network design considerations. In this section, we describe the most important design characteristics and requirements of the modern data center. It is quite unlikely that all of the mentioned characteristics and considerations are of equal importance in every data center design, but relevance and priorities will be dictated by service requirements.

3.2.1 Data center fabric characteristics and solutions

The term “DC Fabric” is replacing the traditional “DC network” as the requirements of modern virtualized data centers drive network innovation, increased flexibility, and new services that were not prevalent in the traditional three tiered hierarchical DC networks. The new paradigm for the data center attempts to deliver “any workload, anywhere” across a more flexible DC fabric.

Additionally, dynamic provisioning and simplified management are design goals for the new DC infrastructures, and we distinguish this new network infrastructure as a “DC fabric”

Figure 3-8 shows the evolution of data center networks.


Current: Distributed

- Manual Provisioning
- Limited scaling
- Rack-wide VM mobility

Trending: Fabric Based

- Policy-based Provisioning
- Scale Physical & Virtual/Cloud
- DC-wide/Cross-DC VM Mobility

Vision: Application Driven

- Service-centric Provisioning
- Flexible – Anywhere, Anytime
- Cross-cloud VM Mobility

Figure 3-8 Data center network evolution

At the end of the day, the data center network strategy and design is really a balancing act: balancing costs, availability, flexibility, growth, and capacity to meet business objectives. In fact, initiatives aimed at satisfying one set of requirements impact on other requirements:

- IT demand: In order to keep up with IT demand, typical initiatives in the data center network domain are aimed at enhancing the scalability of the data center network to accommodate a growing number of server and storage end points. Enhancing the consumability of the data center itself while providing access to very different end user devices is from similar importance. At the same time, greater scale and more endpoints cannot come at the expense of the overall network manageability in terms of simplification of network control and integration with server and storage management and provisioning.
- Cost reduction: In order to address cost reduction pressures, typical data center network initiatives are aimed at optimizing the existing through new architectural approaches or leveraging network consolidation and virtualization techniques. At the same time, security policy enforcement approaches need to adapt to be effective in securing virtualized and consolidated environments.
- Delivery flexibility: In order to deliver flexibility, typical data center network initiatives are aimed at introducing network automation. At the same time, the traditional availability requirement is still in place and redundancy and fault tolerance need to be carefully addressed.

Data center fabric solutions enable enhanced scalability with simplified management and increased capacity, but also address the following data center network drivers mentioned in chapter 1:

- Enterprises and IT Service Providers need data center architectures that accommodate increased application level flexibility, which translates into a more scalable, flatter, dynamic, highly available network infrastructure at the hardware layer, so that resources can be pooled and virtual machines freely moved.
- Shift in traffic patterns from primarily client-server or as commonly referred to as north-to-south flows, to a combination of client-server and server-server or east-to-west plus north-to-south streams.
- 10 GE LAN on Motherboard (LoM) interfaces on servers will increase the traffic generation at the server level and storage traffic integration will require more non-blocking capacity with predictable performance

Traditionally, network architects in search of high scalability and reliability have utilized Layer 3 or IP addressing to segment and scale networks. With the drive towards Layer 2 and Layer 3 agnostic network fabrics, the design decisions will be simplified by having the appropriate technology “everywhere.”

Scalable data center fabrics

In the following section, we discuss a new architecture for the data center fabric, but before we do, a word about the real world. A three-tier network architecture is the dominant structure in data centers today, and will likely continue as the standard design for many networks. By three tiers, we mean access switches/Top-of-Rack (ToR) switches, or modular/End-of-Row (EoR) switches that connect to servers and IP based storage.

As shown in Figure 3-9, these access switches are connected via Ethernet to aggregation switches. The aggregation switches are connected into a set of core switches or routers that forward traffic flows from servers to an intranet and internet, and between the aggregation switches. It is common in this structure to over-subscribe bandwidth in the access tier, and to a lesser degree, in the aggregation tier, which can increase latency and reduce performance, but more on this later.

It is common that VLANs are constructed within access and aggregation switches, while Layer 3 capabilities in the aggregation or core switches, route between them. Within the high-end data center market, where the number of servers is in the thousands to tens of thousands and east-west bandwidth is significant, and also where applications need a single Layer 2 domain, the existing Ethernet or Layer 2 capabilities within this tiered architecture do not meet emerging demands.

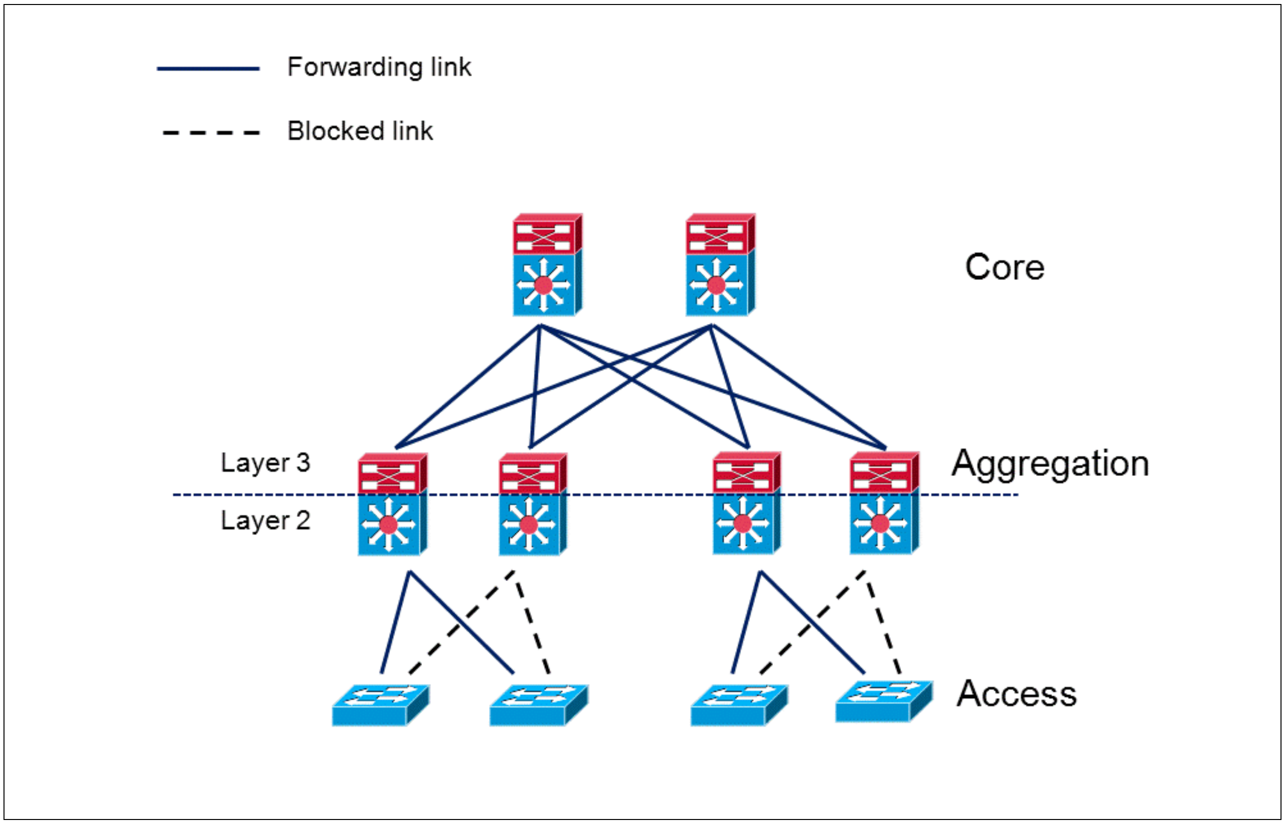


Figure 3-9 Traditional three-tier Network Fabric Architecture

An emerging way to design a scalable data center fabric is often called a “fat-tree” and has two kinds of switches; one that connects end devices and the second that connect switches creating a non-blocking, low latency fabric. As shows in Figure 3-10, we use the terms “leaf” switch to denote server connecting switches and “spine” to denote switches that connect leaf switches. Another design is to connect every switch together in a full mesh, with every server being one hop away from each other.

Together, a leaf and spine architecture creates a scalable data center fabric and has multiple advantages such as spreading risk from two core devices to multiple devices. The Leaf/Spine architecture is constructed so that failure of a spine switch will not disrupt traffic due to the use of ECMP. Additional advantages of this architecture are discussed later.

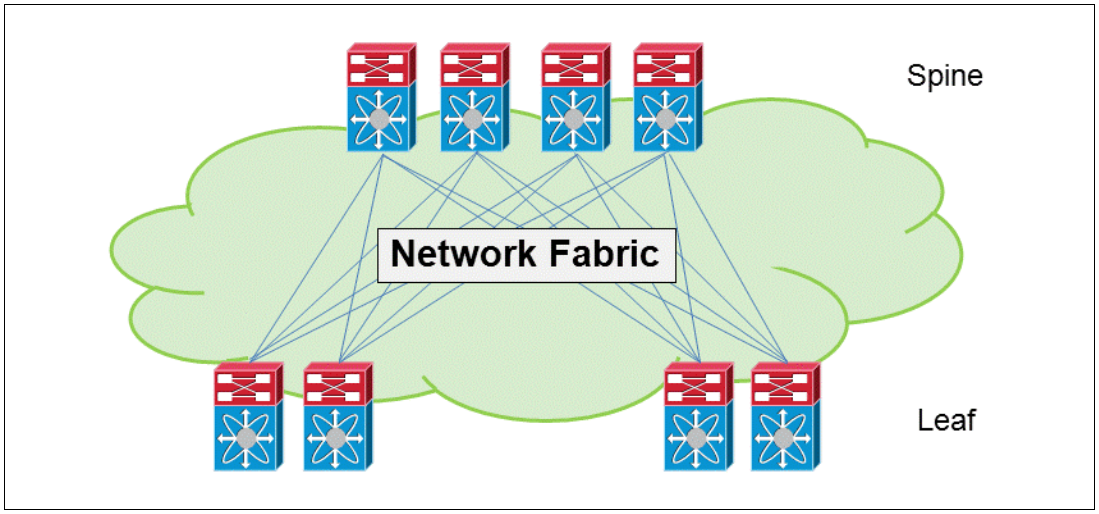


Figure 3-10 Two-tier Leaf-Spine Network Fabric Architecture

A unified fabric

The concept of a unified fabric is to virtualize data center resources and connect them through a high bandwidth network that is very scalable, high performance and enables the convergence of multiple protocols onto a single physical network. These resources are compute, storage and applications, which are connected via a network fabric. In short, the network is the unified fabric and the network is Ethernet.

The industry tends to focus on storage transport over Ethernet as the main concept behind a unified fabric with technologies such as Fibre Channel over Ethernet or FCoE, iSCSI over Ethernet, iWARP over Ethernet and even InfiniBand over Ethernet. However, this is a narrow view of a unified fabric, which is being expanded thanks to continual innovation of Ethernet by the vendor community and standards organizations such as the IEEE and IETF. Ethernet innovations such as FCoE, Data Center Bridging or DCB, Cisco's VN-Link, FEX-Link and virtual PortChannel or vPC have enhanced Ethernet networking to support a wide range of new data center fabric design options. In addition to these protocol enhancements, the availability of 40 Gb and 100 Gb Ethernet is significantly increasing Ethernet’s ability to scale bandwidth. To demonstrate how Ethernet is evolving to be the unified fabric for high-end data centers, we explore Cisco’s FabricPath innovation.

Cisco FabricPath

FabricPath is a Cisco proprietary protocol that provides a new level of bandwidth scale to connect Nexus switches and delivers a new fabric design option with unique attributes for IT architects and designers. Cisco FabricPath is an NX-OS innovation, meaning its capabilities are embedded within the NX-OS network OS for the data center. Cisco FabricPath is essentially multipath Ethernet; a scheme that provides high-throughput, reduced and more deterministic latency, and greater resiliency compared to traditional Ethernet.

Cisco FabricPath combines today's Layer 2 or Ethernet networking attributes and enhances it with Layer 3 capabilities. In short, Cisco FabricPath brings some of the capabilities available in routing into a traditional switching context. For example, Cisco FabricPath offers the benefits of Layer 2 switching such as low cost, easy configuration and workload flexibility. What this means is that when IT needs to move VMs and/or applications around the data center to different physical locations, it can do so in a simple and straightforward manner without requiring VLAN, IP address and other network reconfiguration. In essence, Cisco FabricPath delivers plug and play capability, which has been an early design attribute of Ethernet.

Further, large broadcast domains and storms inherent in Layer 2 networks that occurred during the mid-1990s have been mitigated with technologies such as VLAN pruning, reverse path forwarding, time-to-live, and so on.

The Layer 3 capabilities of Cisco FabricPath deliver scalable bandwidth allowing IT architects to build much larger Layer 2 networks with very high cross-sectional bandwidth eliminating the need for oversubscription. In addition, Cisco FabricPath affords high availability as it eliminates the Spanning Tree Protocol (STP), which only allows one path and blocks all others, and replaces it with multiple active paths between endpoints within the data center. This offers increased redundancy as traffic has multiple paths in which to reach its final destination.

Cisco FabricPath employs routing techniques such as building a route table of different nodes in a network. It possesses a routing protocol, which calculates paths that packets can traverse through the network. What is being added to Cisco FabricPath is the ability for the control plane or the routing protocols to know the topology of the network and choose different routes for traffic to flow. Not only can Cisco FabricPath choose different routes, it can use multiple routes simultaneously so traffic can span across multiple routes at once. These Layer 3 features enable Cisco FabricPath to use all links between switches to pass traffic as STP is no longer used and would shut down redundant links to eliminate loops. Therefore, this would yield incremental levels of resiliency and bandwidth capacity, which is paramount as compute and virtualization density continue to raise driving scale requirements up.

Basically, here are the benefits of deploying an Ethernet fabric based on Cisco FabricPath:

- Simplify and reduce the cost of deployment and operation: Cisco FabricPath provides plug-and-play simplicity with little configuration. Application deployment becomes simpler, and troubleshooting remains straightforward and transparent.
- Increase flexibility: Cisco FabricPath enables simplified workload mobility and network clustering applications and generally removes the design constraints associated with spanning tree protocol.
- Increase available bandwidth: With capabilities such as equal-cost multipath (ECMP) and multi-topology forwarding, Cisco FabricPath enables you to use all available bandwidth in the network.
- Increase availability: Cisco FabricPath provides fast re-convergence and fault-domain isolation, insulating end users and applications from the network.

In other words, while Cisco FabricPath benefits the server and application teams by providing a transparent network fabric that breaks down application silos, permits workload mobility, and provides a high level of deployment flexibility, it also benefits the network operations team by reducing inter-team dependencies, streamlining deployment and configuration, and simplifying network maintenance and troubleshooting.

New network design enabled by Cisco FabricPath

In essence, Cisco FabricPath is an enhanced form of link aggregation. As mentioned before, when using STP between two upstream switch chassis, one path is active and one is standby. To get around this limitation, companies such as Cisco offered vPC, which allows link aggregation between two chassis with both links active. What Cisco FabricPath's multipathing allows is to scale link aggregations up to 16 chassis. This is significant as network design completely changes when link aggregation scales from 1-to-2 then to 16 links.

When network design was limited to one or two aggregated links, the three-tier structure of access switch, aggregation, and core was required for scalability reasons. In all of these different tiers, planners would build in a certain level of oversubscription based on north-to-south and/or east-to-west traffic patterns.

With multipathing to 16 ways, IT architects now have an opportunity to build very large, broad, scalable, and data center wide topologies, without having to build multiple tiers. What this enables in HPC, cloud computing, or hosting environments is that the network is transformed from a hierarchical architecture to a flat topology where only one hop separates any node in the data center. This completely changes the economics and dynamics with which IT designers build data center fabrics by affording a simpler and very scalable network.

As shown in Figure 3-11, there are other data center designs possible using the Cisco FabricPath technology. The design flexibility is a major benefit of the technology, because it allows a nearly seamless transition from the traditional three-tier architecture to a data center wide fabric with all its benefits.

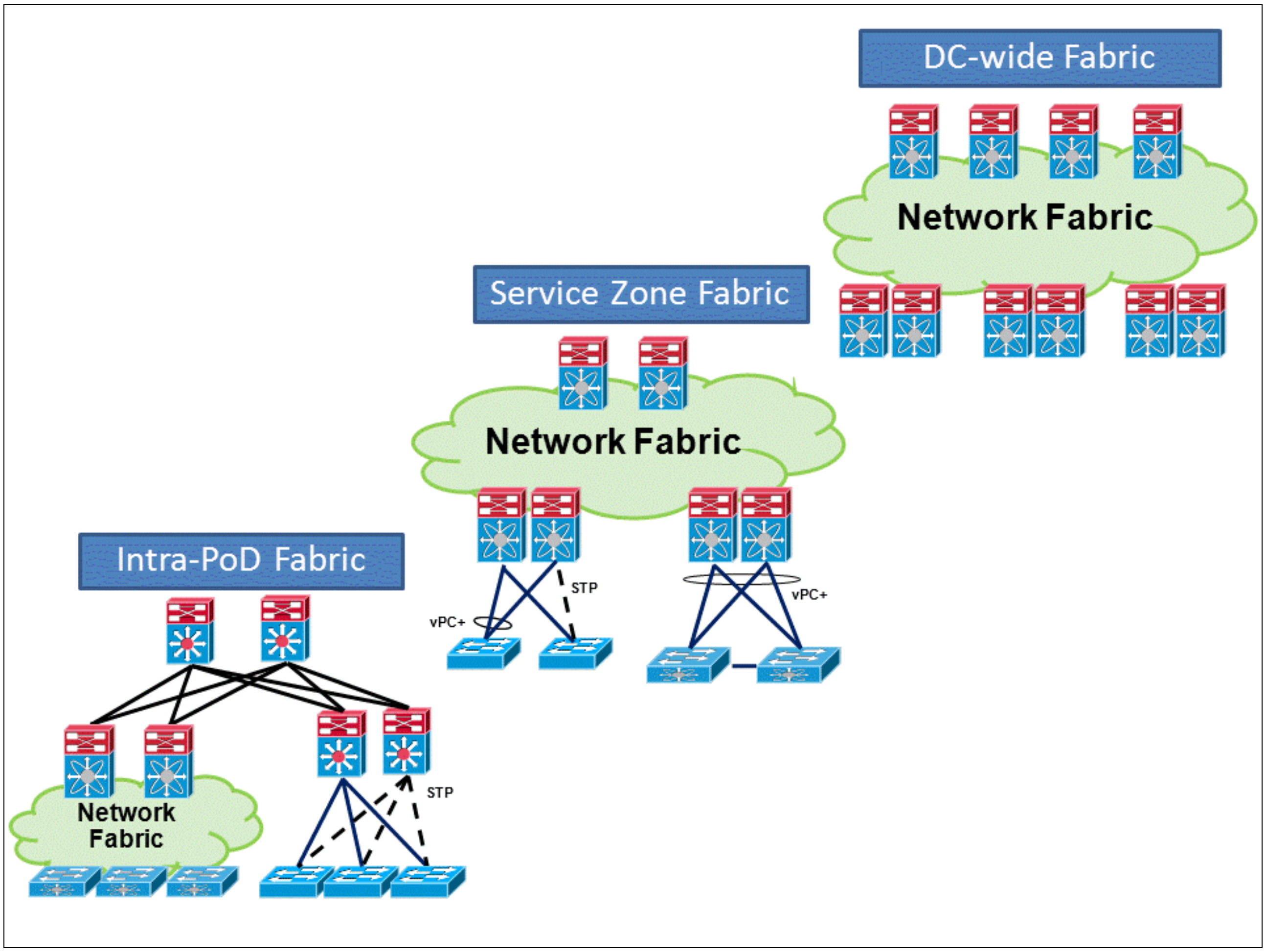


Figure 3-11 Network Fabric solution designs and possible migration path

Workload mobility using Cisco Fabric Path

To highlight the ability to move data center resources freely, we focus on how a flatter network fabric changes application silos, which have traditionally been built through segmentation to isolate various applications. For scale considerations, many IT architects segment applications such as web servers, application services, and database services by placing them in different subnets or different IP address domains; in effect siloing applications via subnet domains.

This approach increases cross-sectional bandwidth, thanks to Layer 3 demarcations between silos, but is very inflexible. For example, each subnet is usually a silo made up of a physical rack with multiple VLANs. Consider a common scenario when incremental server capacity is needed in subnet one, but there is no more physical space. There may be space within subnet two, but unfortunately the IT architect cannot pool physical space across subnets, because of segmentation.

Consider how this approach changes with Cisco FabricPath. First, a single domain Layer 2 fabric connects all servers, offering a much more scalable and flexible network fabric. Since all servers reside within one common layer two network, IT operations can move servers around as needs require, meaning that there is no movement constraints. The level of flexibility and scale is just significantly enhanced. Cisco FabricPath translates into simpler and lower operational costs and increased flexibility in terms of deployment. When Cisco FabricPath is introduced, IT architects scan scale more broadly with subnet barriers eliminated, allowing a VLAN to span across the entire physical infrastructure. Even though the physical infrastructure is still protected by Cisco FabricPath, IT architects no longer have to separate application groups in to logical silos as shown in Figure 3-12.

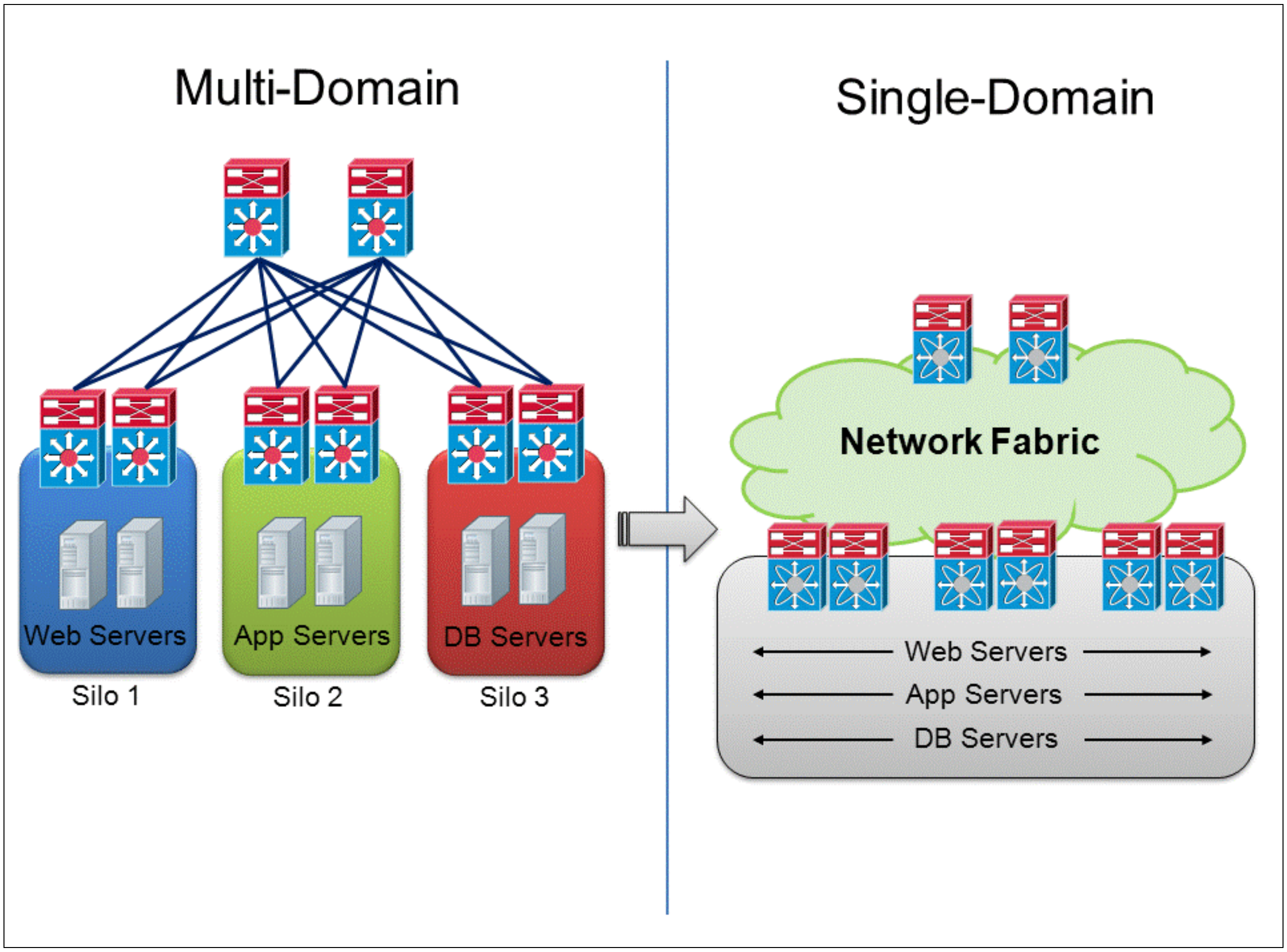


Figure 3-12 Workload mobility with Cisco FabricPath

3.2.2 IPv6 and the data center

The public Internet and most enterprise networks route traffic based on the IPv4 protocol, developed by the Internet Engineering Task Force (IETF) in 1981. The address space available in IPv4 is being outstripped by a spiraling number of servers and other network-attached devices that need or will need addresses; traditional computer and networking equipment, smart phones, sensors in cars and trucks and on outdoor light and utility posts, and more.

A newer protocol, IPv6, also developed by the IETF, is available. IPv6 offers a bigger address space to accommodate the increasing need for new IP addresses. IPv6 also promises improvements in security, mobility, and systems management. While IPv4 and IPv6 can coexist, ultimately the network of every organization that uses the public Internet will need to support IPv6. Once the available IPv4 address space for the world is exhausted, the ability to route network traffic from and to hosts that are IPv6-only will become a requirement. No matter the reason why a company introduces IPv6, deployment of IPv6 will take focus, planning, execution, and testing and require operational changes, all of which will impact servers, other devices in the data center, and applications, in addition to the network itself.

On 3 February 2011, the Internet Assigned Numbers Authority (IANA) issued the remaining five IPv4 /8 address blocks in the global free pool equally to the five Regional Internet Registries (RIRs), and as such, the American Registry for Internet Numbers (ARIN) is no longer able to receive additional IPv4 resources from the IANA. Each RIR community has different consumption rates, different reserves of address space, and different run-out policies. Therefore, depletion of each RIR free pool will occur at different times, ranging from perhaps as soon as year 2012 in the case of APNIC, to as late as 2015, in the case of AfriNIC. Once the RIR free pool is exhausted, the Internet will still run IPv4, but no more IPv4 address space will be available from the RIRs and new allocations will be for IPv6 addresses only.

Network equipment providers ship new equipment that is IPv6-ready. Telecommunications carriers can, in many cases, handle IPv6 traffic or have backbone upgrade projects underway. However, for a specific enterprise, the network, network-attached devices, the tools used to manage data centers and other IT resources, and the applications that are used to run the business need to be checked for their readiness to support IPv6 and plans must be made and executed for any upgrades. Also, each enterprise must determine its strategy for the coexistence of IPv4 and IPv6; tunnelling, translation, dual stack, or a combination, because both protocols will need to be supported and interoperate until IPv4 can be retired.

Depending on the geography, the IP address range allocated to an organization, and any workarounds in place to conserve IP addresses, such as Network Address Translation (NAT) or Classless Inter-Domain Routing (CIDR), planning and readiness for IPv6 may be considered prudent or essential. However, it is risky to wait until IP addresses and workarounds are exhausted, or important customers or supply chain partners or government bodies require IPv6 as a condition of doing business, and implementation must be hurried or performed in a makeshift fashion.

Impact on the enterprise data center

Typically, IPv4 address exhaustion is not an immediate issue for the enterprise data center. Most data centers are employing RFC 1918 private IP addressing. Those that are not, often have enough public address space to handle their existing needs. With healthy address conservation and reclamation, it is possible that IPv4 could remain viable for a long while within an enterprise. But IPv6 in data centers must be planned early enough, because data centers have limited change windows and cannot afford unplanned downtime or performance degradation.

Nevertheless, there remain many good reasons to move to IPv6, including these:

- Virtualization: IPv6 can deliver the quantity of IP addresses required to support highly virtualized environments. IPv6’s address flexibility could support flexible new VM deployments.
- Cloud computing: Large quantities of globally-routable IPv6 addresses allow direct communication with large numbers of cloud nodes without requiring NAT to be set up for each.
- External access: By using globally-routable IPv6 addresses on all devices, external access could be provisioned to any device in the data center. (Firewalls would still be used to prevent unauthorized access to globally-routable IPv6 devices.)
- Visibility and control of IPv6 traffic: All modern OS have IPv6 enabled by default. Typically they attempt to communicate over IPv6 if nodes are Layer 2 adjacent. Many existing intrusion detection and firewall systems might not understand this tunnelled traffic as IPv6, might misinterpret the payload, and might not control the traffic in a manner that corresponds with the established security policies.
- Easier mergers and acquisitions: When organizations use globally-unique IPv6 prefixes, merging networks does not create IP addressing plan overlaps.
- Ubiquity of access: As more and more of the world's Internet access happens over IPv6, it will become important to provide access to IPv6 services.
- Separation of managing services from managed services: The Service Provider potentially use RFC 1918 private IP addresses internally to manage the services for the customers. If customers use the same internal IP addresses, the service provider has to handle the IP address overlapping conflict. This can be avoided if the service provider uses globally-routable IPv6 addresses for the internal managing services.

However, many enterprises note that these motivations do not necessarily form a compelling business case for action. Yet, with the exhaustion of public IPv4 addresses, it is now likely that some form of IPv6 deployment cannot be avoided in the enterprise. This is for the simple reason that most enterprises cannot afford to voluntarily cut themselves off from the IPv6-only connected Internet that is arising. Each enterprise will have to evaluate their IPv6 stance in light of the new realities of Internet addressing.

3.2.3 Multi-tenancy

New data center network designs require a new way of thinking that is different from the traditional design. Main drivers are virtualization technologies and the increasing demand for cloud services.

Regardless of whether the focus is on cloud services or traditional data center services, these pursuits of multi-tenancy share many objectives:

- Increase operational efficiency through cost-effective use of expensive infrastructure.
- Drive up economies of scale through shared resources.
- Provide rapid and agile deployment of customer environments or applications.
- Improve service quality and accelerate delivery through standardization.
- Promote green computing by maximizing efficient use of shared resources, lowering energy consumption.

Achieving these goals can have a profound, positive impact on profitability, productivity, and product quality. However, leveraging shared infrastructure and resources in a data center architecture introduces additional challenges such as security isolated customers, business units, or multi-tier application services in conjunction with highly flexible management.

As enterprise IT environments have dramatically grown in scale, complexity, and diversity of services, they have typically deployed application and customer environments in silos of dedicated infrastructure. These silos are built around specific applications, customer environments, business organizations, operational requirements, and regulatory compliance (Sarbanes-Oxley, HIPAA, PCI), or to address specific proprietary data confidentiality. Here are some examples:

- Large enterprises need to isolate HR records, finance, customer credit card details, and so on.
- Resources externally exposed for outsourced projects require separation from internal corporate environments.
- Healthcare organizations must insure patient record confidentiality.
- Universities need to partition student user services from business operations, student administrative systems, and commercial or sensitive research projects.
- Telcos and service providers must separate billing, CRM, payment systems, reseller portals, and hosted environments.
- Financial organizations need to securely isolate client records and investment, wholesale, and retail banking services.
- Government agencies must partition revenue records, judicial data, social services, operational systems, and so on

Enabling enterprises or service providers to migrate such environments to a shared infrastructure architecture demands the capability to provide secure isolation. External cloud service providers must enable all customer data, communication, and application environments to be securely isolated from other tenants. With external clouds, the separation must be so complete and secure that the tenants can have no visibility of each other. Internal cloud service providers must deliver the secure separation required by their organizational structure, application requirements, or regulatory compliance.

Secure tenant separation

In traditional data center deployments, environments that needed additional security and isolation required a separate set of equipment and infrastructure. While this provided the required separation, efficiency dropped significantly due to lower utilization, while costs increased due to higher management and capital expenditures. This will remain a valid design for security separation and might be a requirement from the customer, but in cases where efficiency and cost reductions are key drivers like in cloud services, resource pools must be efficiently shared among multiple tenants.

In a private cloud, a tenant is an enterprise department. In a public cloud, a tenant is an individual customer or enterprise who subscribes to cloud resources. In both public and private cloud scenarios, multi-tenancy provides better resource utilization but requires designing of secure tenant separation to ensure end-to-end security and path isolation within the shared infrastructure. Secure separation is a fundamental requirement in multi-tenant environments. To achieve secure tenant separation and path isolation, the data center design must be capable to virtualize the network, compute, and storage resources end-to-end.

The following design considerations provide secure tenant separation and path isolation:

- Network separation
- Compute separation
- Storage separation
- Application tier separation
- VM security
- Fibre Channel separation and isolation

Network separation

End-to-end virtualization of the network requires separation at each network layer in the architecture:

- Network Layer 3 separation
- Network Layer 2 separation
- Network services separation (firewall and load balancing services)
- Tenant path isolation
- Application traffic separation

Additionally, virtual private remote access in the form of SSL and IPSec or MPLS VPNs serves to further secure tenant traffic over the internet, shared public or private cloud.

Network Layer 3 separation

As shown in Figure 3-13, an end-to-end VRF-Lite deployment can provide Layer 3 separation among tenants.

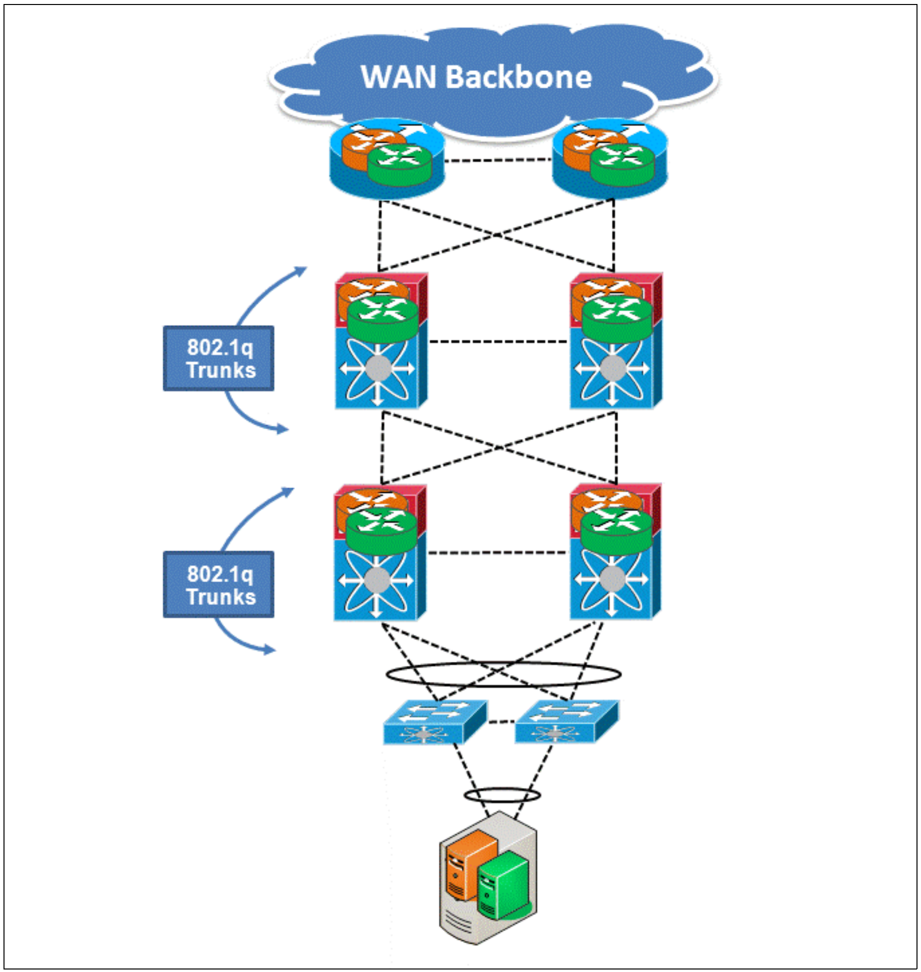


Figure 3-13 VRF-Lite end-to-end

A VRF-Lite instance represents a tenant container. Depending on the service tier, the container could include services such as load balancer, firewall, IPSec/SSL off loading; VLANs representing the application tiers (Web, application, database).

A VRF instance consists of the following elements:

- An IP routing table
- A derived forwarding table
- A set of interfaces that use the forwarding table
- A set of rules and routing protocols that determine what goes into the forwarding table

As shown in Figure 3-14 VRF-Lite uses a Layer 2 separation method to provide path isolation for each tenant across a shared network link. Typically, a dot1q tag provides the path separation.

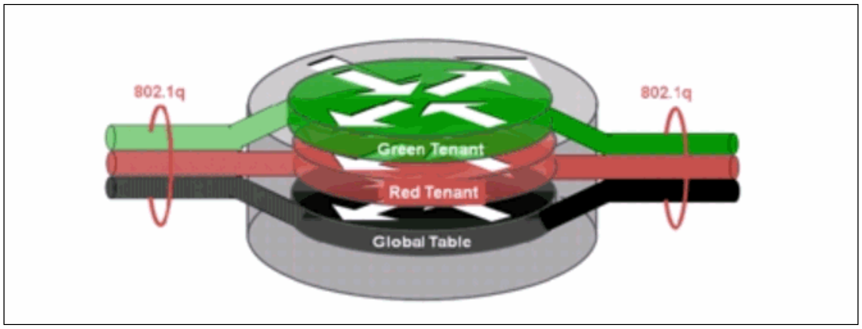


Figure 3-14 Network device virtualization with VRF-

In a multi-tenant environment, Cisco VRF-Lite technology offers the following benefits:

- True routing and forwarding separation: Dedicated data and control planes are defined to handle traffic belonging to groups with various requirements or policies. These groups represent an additional level of segregation and security as no communication is allowed among devices belonging to different VRFs unless explicitly configured.
- Uniform deployment across Cisco Nexus and Catalyst products: VRF-Lite is natively supported in hardware across specific Cisco Nexus and Catalyst switches to achieve higher throughput and performance.
- Virtualized physical infrastructure: Each virtual network represents a replica of the underlying physical infrastructure. The end result is a physical infrastructure shared among multiple tenants.

Network Layer 2 separation

Network separation at Layer 2 is accomplished using VLANs. While VRFs are used to identify a tenant, VLAN-IDs are used to provide isolation at Layer 2.

In addition to the tenant VLAN-ID, there my be more granular isolation requirements such as application and service tier segregation. In large deployments, you may need to enable the extended range of VLANs.

In Figure 3-15, a tenant using VRF Red is assigned three VLANs (21,22,23) for its web application as per the Gold service tier while VRF Green is assigned two VLANs as per the Silver service tier. This design assumes that there is no need to connect the tenant VRFs.

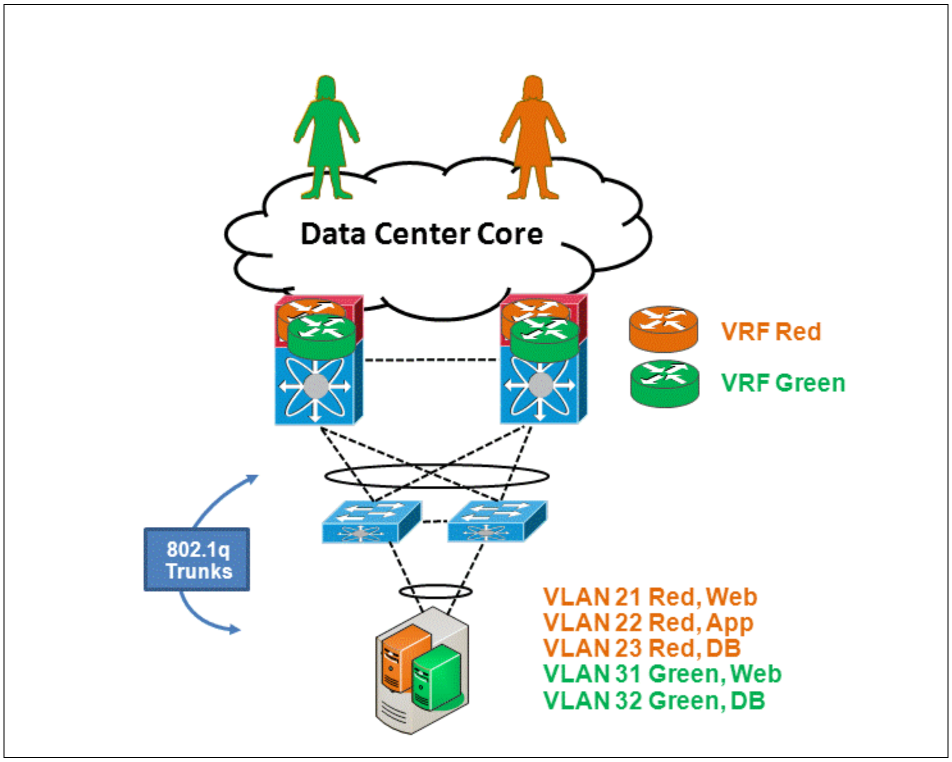


Figure 3-15 VLANs to VRF mapping

Network services virtualization (firewall and load balancing services)

Data center networks also require intelligent services, such as firewall and load balancing of servers and hosted applications. To achieve secure separation across the network, the services layer must also be virtualized. Security contexts and/or zones with independent policies are used to virtualize services. The contexts are stitched to a VRF via VLANs to provide services.

A classic design dedicates a services layer to isolate intelligent services into their own layer. This allows changes or enhancements of the network services without impacting all other layers. Using the virtualization features of the firewalls or load balancers, you can create separate contexts that represent separate virtual devices.

Tenant path isolation

As the modern data center requirements shifted from dedicated infrastructure for each tenant to a shared resource pool model, the network architects began to consider alternative technologies that supported sharing the infrastructure and provided the same level of tenant isolation as a dedicated infrastructure. Path isolation techniques logically divide a shared infrastructure into multiple virtual networks.

Path isolation defines independent, logical traffic paths over a shared physical network infrastructure. These paths use VRFs, VLANs and contexts/zones as resources to provide tenant isolation.

The goal of segmenting the network is to support multi-tenancy on the physical infrastructure; however, it must ensure resiliency, scalability, and security.

A hierarchical network combines Layer 3 (routed), services (firewall and server load balancing), and Layer 2 (switched) domains. Therefore, the three types of domains must be virtualized, and the virtual domains must be mapped to each other to enable end-to-end traffic segmentation. Finally, the server and storage virtualization is integrated with this virtualized path.

The following levels of virtualization are required and implemented in this design:

- Device virtualization: The virtualization of the network device, which includes all forwarding/routing processes, databases, tables and interfaces within the device
- Data path virtualization: The virtualization of the interconnection between devices. This interconnection can be a single or multi-hop. For example, an Ethernet link between two switches provides a single-hop interconnection that can be virtualized using 802.1q VLAN tags.

Figure 3-16 shows how each tenant can be logically segmented using end-to-end path isolation. In a virtualized, multi-tenant network design, path isolation provides these features:

- Tenant Isolation and boundary
- Per-tenant security and policy control

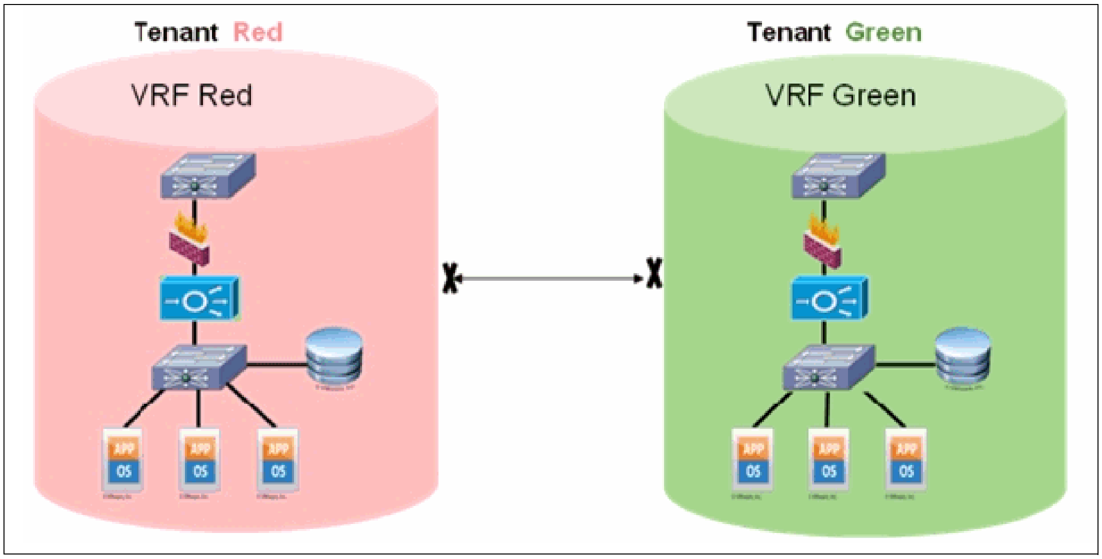


Figure 3-16 Tenant path isolation

End user security and managing traffic separation

The path isolation provided by VRFs helps secure tenants against overlapping network architectures. However, the managing services must be protected against security exploit attempts by external users who access the managed service. Consider the case where a web client accesses the web services inside a cloud. For a given web application, we must secure the server farm and the communication between the web server and back-end servers, such as middleware and database servers.

So additional to the tenant security aspects, the firewall model has to cope with managing services security separation. It must be ensured, that the required service administration flows are access restricted and additionally any kind of tenant specific installation, backup and restore traffic is fully security controlled and security separated from each other. The increasing number of required firewall instances due to the multi-tenant design can be addressed by the firewall virtualization functionality. This helps to reduce the costs.

Compute separation

Virtualization introduces new security challenges and concerns. Traditionally, security policies were applied at the physical server level. However, as physical hosts can now contain multiple logical servers, policy must be applied the VM level. Also, new technologies, such as vMotion, introduced VM mobility within a cluster, where policies follow VMs as they are moved across switch ports and among hosts.

Virtual computing continues to aggregate higher densities of VMs. This high-density model forces us to reconsider firewall scale requirements at the aggregation layer of the network. The result is that high-density compute architectures may require the distribution of security policies to the access tier.

To address some of these new security challenges and concerns, virtual firewalls may be deployed at the access layer of the data center infrastructure to create intra-tenant policy zones. Firewalls at the aggregation layer of the network enable one to define perimeter and per-tenant policies. Like firewalling at the aggregation layer, access layer firewalling can enforce security among the tiers of an application, as described in the next section, Application Tier Separation.

Application tier separation

Many applications follow a three-tiered component architecture. The tiers are typically web, application, and database. The web, or customer facing, tier communicates with the application and database tiers to process transactions. The application and database tiers serve as the middleware and back-end components for the web tier. To ensure a higher degree of security, Enterprises deploy firewall security in front of the web tier and to inspect intra-application inter-tier communications. For such requirements, this design proposes using virtual firewall appliances.

The virtual multi-tenant data center network architecture proposes VLAN separation as the first degree of security in application tier separation. The design suggests that each application resides in separate VLANs within a tenant VRF. If communication must occur between tiers of an application, the traffic is routed through the default gateway where security access lists can enforce traffic inspection and access control. At the access layer of the network, the virtual firewall appliance monitors and restricts inter-VM traffic within and between VMs.

Storage separation

To extend secure separation to the storage layer, considerations have to be made about isolation mechanisms available in a SAN environment. Fibre Channel fabrics have the capability to be devised in a similar way to VLAN by using Virtual SAN (VSAN) technology, as well as the traditional zoning capabilities. For more information related to storage separation, see *Introduction to Storage Area Networks and System Networking*, SG24-5470.

3.2.4 Virtualized workload mobility between data centers

The ability to move workloads between physical locations within the virtualized data center (one or more physical data centers used to share IT assets and resources) has been a goal of IT departments since the introduction of virtualized server environments. One of the main drivers for the continued increased virtualized server concept being adopted is really the control it gives the IT department because now the IT department has the flexibility and agility that comes with decoupling the hardware and software from one another. Ultimately, this means new use cases can be developed addressing specific requirements and adapted to a specific environment. In this virtualized environment, the ability to dynamically and manually redistribute virtual machines to new locations provides the opportunity for value-add and operationally efficient new use cases.

In the virtualized environment, many different use cases have been discussed at times, including these:

- Data center capacity expansion and/or consolidation:

  Applications need to be migrated from one data center to another without business downtime as part of a data center migration, maintenance or consolidation efforts. Virtual machines need to be migrated to a secondary data center as part of data center expansion to address power, cooling, and space constraints in the primary data center.

- Virtualized server resource distribution over distance:

  Virtual machines need to be migrated between data centers to provide compute power from data centers closer to the clients (follow the sun) or to load-balance across multiple sites.

- Disaster planning strategies, including disaster avoidance scenarios:

  Data centers in the path of natural calamities (such as hurricanes) needs to proactively migrate the mission-critical application environment to another data center.

Individually, each of these use cases is addressing specific needs of the business. Taken together, these use cases aggregate into one clearly identifiable use case: Virtualized Workload Mobility (VWM).

Virtualized Workload Mobility is a vendor and technically agnostic term that describes the ability to move virtualized compute resources to new, geographically disperse, locations. For example, a host Operating System (as examples: Red Hat Linux or Windows 7) that has been virtualized on a virtualization host Operating System (as examples: VMware ESXi, Windows Hyper-V, or Red Hat KVM) can be transported over a computer network to a new location that is in a physically different data center.

Additionally, Virtualized Workload Mobility refers to active, hot, or live workloads. That is, these virtualized workloads are providing real-time compute resources before, during and after the virtualized workload have been transported from one physical location to the other. Note that this live migration characteristic is the primary benefit when developing the specific use case and technical solution and therefore must be maintained at all times.

Virtualized Workload Mobility use case

Virtualized Workload Mobility use case requirements include these features:

- Live migration: The live migration consists of migrating the memory content of the virtual machine, maintaining access to the existing storage infrastructure containing the virtual machine’s stored content, and providing continuous network connectivity to the existing network domain.

- Non-disruptive: The existing virtual machine transactions must be maintained and any new virtual machine transaction must be allowed during any stage of the live migration.
- Continuous data availability: The virtual machine volatile data will be migrated during the live migration event and must be accessible at all times. The virtual machine non-volatile data, the stored data typically residing on a back-end disk array, must also be accessible at all times during the live migration. The continuous data availability use case requirements have a direct impact on the storage component of Data Center Interconnect for the Virtualized Workload Mobility use case.

With the Virtualized Workload Mobility use case described and the main use case requirements identified, there is also a need to identify the challenges, concerns and possible constraints that need to be considered and how they impact the solution development and implementation.

Use cases: Challenges, concerns, and constraints

Use case challenges, concerns, and constraints are derived by a systematic review of the use case requirements.

- Live migration: With the live migration requirement, a virtualization host operating system must be identified and integrated into the solution. Vendors that provide virtualization host operating systems include IBM, VMware, Red Hat, and Microsoft, and they all have their own technical and supportability requirements. This means that the solution will have to consider the selected vendors requirements, which may impact the design and implementation of the Virtualized Workload Mobility solution. Virtualization Host OS selection becomes a critical component in the development.
  - Criteria 1: Live migration 5 millisecond round trip timer requirement
- Non-disruptive: The non-disruptive requirement means that we have to consider both existing server transactions and new server transactions when designing the architecture. Virtual machine transactions must be maintained so the solution must anticipate both obvious and unique scenarios that may occur.
  - Criteria 2: Maintain virtual machine sessions before, during and after the live migration.
- Continuous data availability: The non-volatile memory (storage) must be accessible wherever the virtual machine is located in the virtualized data center. In the event of a virtual machine migration, stored content must continuously be accessible and this impacts the Virtualized Workload Mobility solution with respect to how the storage infrastructure is specifically designed and integrated into the solution. Additionally, there are many different storage topologies and products to choose from, so the solution must be flexible and adaptable to different storage scenarios:
  - Criteria 3: Maintain storage availability and accessibility with synchronous replication
  - Criteria 4: Develop flexible solution around diverse storage topologies

Data Center Interconnect solution overview

The term Data Center Interconnect (DCI) is relevant in all scenarios where different levels of connectivity are required between two or more data center locations in order to provide flexibility for deploying applications and resiliency schemes.

Figure 3-17 shows the different levels of connectivity.

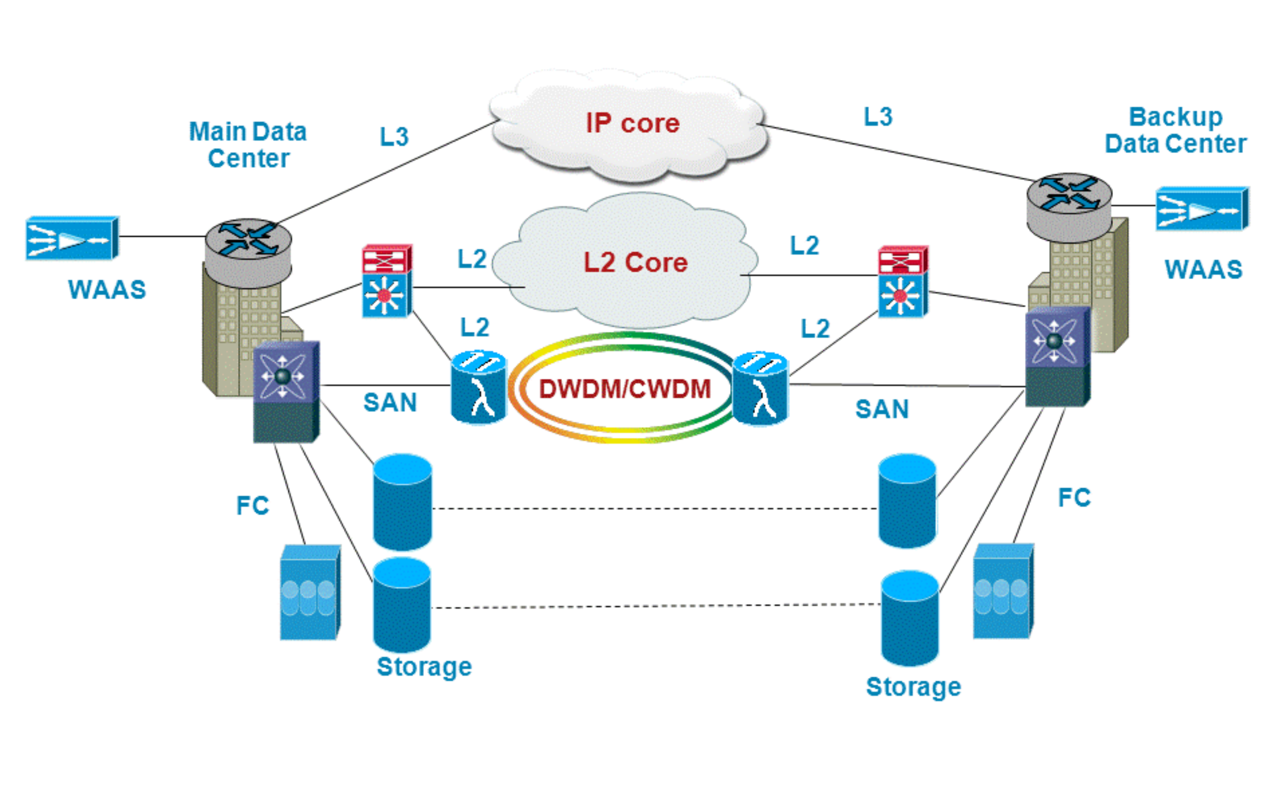


Figure 3-17 Data Center Interconnect connections

Data Center Interconnect components

Four critical functional components are addressed in this DCI solution, as follows:

- LAN extension: Provides a single Layer 2 domain across data centers. The data center applications are often legacy or use embedded IP addressing that drives Layer 2 expansion across data centers. Layer 2 Extension provides a transparent mechanism to distribute the physical resources required by some application frameworks such as the mobility of the active machine (virtual or physical).
- Path optimization: Deals with the fact that every time a specific VLAN (subnet) is stretched between two (or more) locations that are geographically remote, specific considerations need to be made regarding the routing path between client devices that need to access application servers located on that subnet. The same challenges and considerations also apply to server-to-server communication, especially for multi-tier application deployments. Path Optimization includes various technologies that allow optimizing the communication path in these different scenarios.
- Layer 3 extension: Provides routed connectivity between data centers used for segmentation/virtualization and file server backup applications. This may be Layer 3 VPN-based connectivity, and may require bandwidth and QoS considerations.
- SAN extension: Presents different types of challenges and considerations because of the requirements in terms of distance and latency and the fact that Fibre Channel cannot natively be transported over an IP network.

In addition to the four functional components listed, a holistic DCI solution usually offers a couple more functions. First, there are Layer 3 routing capabilities between data center sites. This is nothing new, since routed connectivity between data centers used for segmentation/ virtualization and file server backup applications has been developed for quite some time. Some of the routed communication may happen leveraging the physical infrastructure already used also to provide LAN extension services; at the same time, it is common practice to have also L3 connectivity between data centers through the MAN/WAN routed network.

The second aspect deals with the integration of network services (as FW, load-balancers). This also represents an important design aspect of a DCI solution, given the challenges brought up by the usual requirement of maintaining stateful services access while moving workloads between data center sites.

Virtualized Workload Mobility solution overview

DCI enables the Virtualized Workload Mobility use case. The DCI system has four main components that leverage existing products and technologies and is modular in design, allowing for DCI system to support new and evolving use cases. For the Virtualized Workload Mobility use case, the solution components are shown in Figure 3-18.

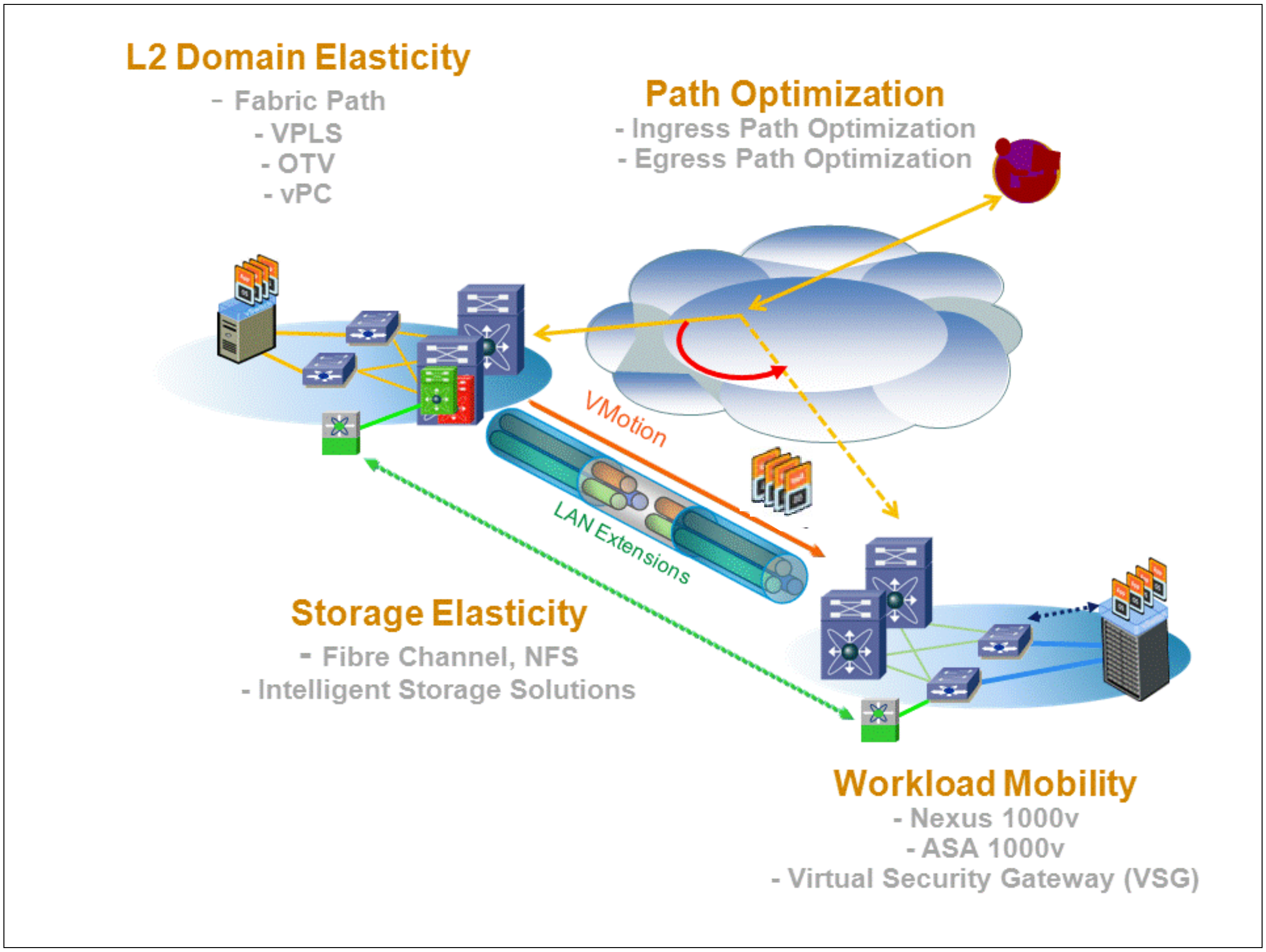


Figure 3-18 Virtualized Workload Mobility Solution Components

A key point to recognize is that the solution components are designed to be independent from one another. That is, new technologies and products can replace existing products and technologies within a specified component.

More information is available in the *Cisco Data Center Interconnect Design and Implementation Guide* PDF at the following link:

http://www.cisco.com/en/US/solutions/collateral/ns340/ns517/ns224/ns949/ns304/ns975/data_center_interconnect_design_guide.pdf

3.2.5 Data center network throughput considerations

In networking, throughput is defined as the average amount of traffic that can be transmitted between two given nodes. Throughput is different than capacity because capacity defines the theoretical capacity of the communication link while throughput defines the actual traffic that can be transmitted over the link.

Usually data center networks communication paths (either client to server, server to server or server to storage) have several intermediate nodes (also called hops) and links in between the two end nodes where throughput is measured. If this is the case the achievable throughput will be equal to the minimum link speed or the minimum capacity that can be processed by the intermediate node, if this processes the flow. A typical example of a node that can affect overall throughput even if the links are all high speed is a firewall node because this typically processes traffic flows and if the maximum processing speed for the firewall is 1 Gbps, then having 10 Gbps links in the communication path will not allow you to have an end-to-end throughput higher than 1 Gbps.

Factors affecting throughput in data center networks are as follows:

- Number of links in the communication path: If intermediate nodes are in the path, the end-to-end throughput will be equal to or lower than the lowest link capacity.
- Capacity of the links: These can be at different speeds at different hierarchical layers in the data center network. Here are typical access speeds for data center network links:
  - Access layer: The 10/100/1000 Mbps access speeds are very typical at the server connectivity layer, but 10 GE is becoming mainstream and getting more traction.
  - Aggregation/core layer: This layer capacity needs to be able to handle traffic from multiple access switches so it is typically deployed with link speeds that are one order of magnitude greater than the access layer speed. For example, 1 GE access is handled with 10 GE links from access to aggregation/core. The adoption of 10 GE at the access layer drives the need for 40 GE links between Access and aggregation/core, even if typically to keep up with capacity demand additional links can be bundled in a single logical EtherChannel, as long as access and aggregation layer ports are available and as long as the maximum number of physical links in an EtherChannel limit is not reached.
  - WAN/ISP links: These links connect the data center to the outside world. Branch offices, campus networks, extranets, and the Internet are the typical external networks that a data center network would be attached to. Usually an increased capacity of these links come at a price because they are usually procured from a Service Provider so their actual speed requirements should be carefully planned from an end-to-end point of view. Typically these kind of links tend to be 1 Gbps or lower, and WAN optimization and acceleration techniques can mitigate the performance challenges of “thin” links.

  - DC to DC links: These links are used to interconnect data centers together in order to provide high availability and disaster recovery capabilities. These links can be procured from a Service Providers or deployed through a Private Network. A common example is the usage of optical links via dark fiber and DWDM. Usually these links carry various channels which are mapped to optical wavelengths which tend to be very high capacity.Because of this 100 GE can be an option, especially if Storage traffic is carried on the links for asynchronous/synchronous replication and remote backups.
  - Oversubscription: The ratio between the available capacity between two points of the network. It is a very common practice in multiple access (shared) communications networks to oversubscribe networks because not all the traffic sources will be transmitting at the same time and designing a network for 1:1 oversubscription tends to be very costly. Oversubscription in data center networks can be measured in different points but typically it is measured between the access layer and the core layer. Obviously the oversubscription ratio will influence the actual throughput of a given data center network flow because links in the middle (in the aggregation and core layers) are shared among various traffic sources and also network congestion can occur.
  - Error rates: If there are transmission errors on the communication path, the overall throughput will be affected.

After having defined throughput and the factors affecting it from a data center networking point of view, let us now look at some of the trends that modern data centers are experiencing regarding these factors. We then point out which IT trends are driving this change in data center networking:

- The widespread adoption of *server virtualization* techniques has generated a higher demand for 10 GE server access because now multiple logical hosts (VMs) will share the same access layer port. This generates the need for 10 GE access ports for server connectivity and also higher capacity at the aggregation layer; 40 GE interfaces are now available to handle an increased traffic at the access layer. Also the usage of virtual switches introduce a new oversubscription measurement point as VMs on the same physical hosts will be contending the same physical port through the vSwitch. As 10 GE LAN-on-motherboard (LOM) become standard this will drive an even higher demand for 10 GE access ports and 40 GE ports in aggregation/core. Also as server technology evolves it is expected that in the future hundreds of VMs could share the same physical host, driving also the need for 40 GE at the access layer and 100 GE in the core. Most networking manufacturers are preparing for this and launching new hardware which is 100 GE ready. For example, the Cisco Nexus 7000 hardware is 100 GE ready.
- Increased *server to server traffic* (driven by SOA and cloud paradigms) together with the convergence of storage protocols with a non-dropping behavior on the same network link affects the tolerable oversubscription rate at the core of the data center network. This drives the need not just for higher capacity but also the need for non-blocking data center network as, for example, it is not tolerable to discard Fibre Channel traffic due to congestion on the communication path.Network vendors are addressing this by launching products to build non-blocking data center networks such as Cisco FabricPath**.**
- *Virtual Machine mobility and Big Data* are driving the need for very high speed connectivity between data centers to cope with an increased capacity from disaster recovery and backup strategies. 100 Gbps and greater connectivity between sites, for example, may become a mainstream requirement for optical interconnects between data centers. Optical networking manufacturers, such as Cisco, are currently testing Terabit/s optical networks to prepare for future requirements.

3.2.6 High availability and fast failover

Data center high availability (HA) is a large subject that would typically focus on application uptime. The purpose of this section is to concentrate on the specific features of the network to enhance application availability. The highest available application environments are typically spread between data centers to address localized or external environmental disasters. In the interest of brevity, we focus on the HA characteristics of a network within a single data center.

HA is typically measured as a percentage of overall uptime and Service Level Agreements (SLA) are associated with this measurement. Other closely related concepts include these:

- RCA: Root Cause Analysis. Understanding “what went wrong” is vital to reduce future outages and learn from past mistakes. RCA is used to understand why a failure happened (HW, SW, link, power/environment, change, design) and how it could be prevented (people, process, tools or technology). Organizations should adopt a culture that understands the RCA process and benefits from RCA output.
- MTTR: Mean Time to Repair. This is the measurement of how long things take to fix and can be impacted by poor documentation or overly complex environments. MTTR is heavily influenced by culture, tools, and process.
- MTBF: Mean Time between Failures. Calculated by measuring the average time between failures on a device. MTBF should be as high as possible for critical components in the network and “system” MTBF is increased by parallel (redundant) devices that can share the total load.
- RTO: Recovery Time Objective. This is the total time required for a planned outage or the time required to fully recover from an unplanned outage.
- RPO: Recovery Point Objective. This defines the maximum amount of data that your organization is willing to sacrifice after a problem. A zero RPO business continuance solution can survive a disaster without any loss of data.
- Disaster Radius: The probable distribution of a disaster. Also called the disaster or threat radius, it affects the business continuance solution. The probability and extent of damage from earthquakes, floods, fires, hurricanes, cyclones, or terrorist threats varies according to the region in which the data center physically resides.
- ITIL: Information Technology Infrastructure Library. This is a set of practices for IT service management (ITSM). As part of ITIL, Availability planning is addressed through Reliability, Maintainability, Serviceability, Resilience and Security.
- Change Management: Change management refers to the consistent process of successfully managing change within an organization and this process includes Change documentation requirements, Risk level assignment, Validation and approval procedures, Change meetings, Emergency change procedures and Change management system/metrics.

The cost of downtime varies based on multiple factors, and will be different for each application and each customer in the data center. An example of this variance can be read here:

http://www.evolven.com/blog/downtime-outages-and-failures-understanding-their-true-costs.html

Two of the most important tools to maintain availability are standardization (configuration, version control, naming conventions, upgrade procedures, and solution templates) and documentation (device, link, inventory, software version, configuration, network topology, and so on).

It is important to distinguish between availability and uptime; in the event of a network outage, an application may be “up” but not available. high availability (HA) in the network is provided by a combination of architecture, system, component, and feature resilience. The data center network infrastructure includes routing, switching, load balancing, firewalls, and other services that need to maintain service and throughput despite routine maintenance, upgrades, adds, moves, and failures.

Often the configuration of network is critical to the integrity of external HA features (such as configuration to allow cluster heartbeats to transit between HA nodes) and the network design is critical to support HA features of highly virtualized data centers (such as providing mobility for VMs and ensuring they can be found after relocation)

Network high availability needs to mitigate the impact of SCHEDULED downtime, as well as unscheduled downtime. Here are some key characteristics of highly available networks:

- In service software upgrades (ISSU): Cisco ISSU provides the capability to perform transparent software upgrades on platforms with redundant supervisors, reducing downtime and allowing customers to integrate the newest features and functions with little or no negative effect on network operation.
- Stateful graceful restart: Processes are monitored and may be restarted without service interruption (such as routing daemon on an L3 switch). Sometimes known as Non Stop Forwarding (NSF), this stabilizes network routing tables through failures and component switchover.
- Link aggregation: Cisco PortChannel allows users to aggregate multiple physical links for active-active redundancy and increased bandwidth. As a result of this aggregation, a single logical link is presented to Layer 2 or Layer 3 protocols. Traffic is load-shared across the links in a PortChannel using one of the available load-balancing schemes.
- Equal-cost multipath (ECMP) forwarding model: Load balancing technology that optimizes flows across multiple IP paths. ECMP can be used by Layer 3 protocols to provide link-level and chassis-level redundancy as well as increased bandwidth.
- Process survivability: Critical processes are run in protected memory space and independently of each other and the kernel, providing granular service isolation and fault containment and enabling modular patching and upgrading and rapid restart.
- Stateful failover: Redundant components or platforms are kept synchronized at all times to enable rapid stateful failover. Sophisticated checks are in place to help ensure that the state is consistent and reliable throughout the entire distributed architecture after failover occurs. Practical examples of Stateful failover can be seen in clustered firewalls or redundant switch supervisors (SSO = Cisco Stateful Switch Over).
- Failure isolation: It is critical that the impact of failure is contained and does not propagate across the data center, or indeed to the resilient, backup data center. Specific protocols have been developed to ensure that failure domains are contained, yet critical traffic remains intact (such as OTV). Other features such as Nexus Virtual Device Contexts (VDC) are introduced to offer a “firebreak” between critical environments and contain failures.
- Storm control: A traffic storm occurs when packets flood the LAN, creating excessive traffic and degrading network performance. The traffic storm control feature (also called traffic suppression) prevents disruptions on Layer 2 ports due to a broadcast, multicast, or unknown unicast traffic storm on physical interfaces.

To limit the impact of traffic storms, the network administrator should monitor the levels of the incoming broadcast, multicast, and unicast traffic over a predefined interval, and set an appropriate threshold. When the ingress traffic for which traffic storm control is enabled reaches the traffic storm control level that is configured on the port, traffic storm control drops the traffic until the traffic storm control interval ends. Network Administrators have several optional actions that the storm threshold can trigger such as shutdown the port or generating an SNMP Trap. Storm Control can be configured granularly and have different policies for broadcast or multicast traffic

- Control-plane policing (CoPP): In critical network infrastructure the control plane policing (CoPP) feature increases security on the switch by protecting the route processing from unnecessary or Denial of Service (DoS) traffic and giving priority to important control plane and management traffic.
- Protocol fast convergence: Typically, fast convergence refers to a collection of features used for faster-than-default network convergence in response to a failure. Fast convergence features can be categorized into two general types:
  - Fast failure detection (FFD): Features that affect the speed with which a neighbor failure event can be detected.
  - Fast failure reaction (FFR): Features that affect the speed with which protocols notify the rest of the network that a change has occurred and the speed with which protocols running on network devices recalculate the new state based on an event that was received.
- Bidirectional forwarding detection (BFD): In addition to the Fast Failure options, BFD implements a lightweight hello communication mechanism and builds a neighbor relationship state. This common mechanism, standardized in RFC 5580, can be used by multiple Layer 3 protocols for FFD instead of using per-protocol hellos for this purpose.
- Configuration rollback: It is important to verify the consistency of a configuration prior to committing it, and include checkpoints, to allow operators to roll back to a known good configuration.

Modularity is one approach used to contain any failure domain, and this can range from modularity in the end-to-end architecture (such as PoD architectures), down to modularity on a single component (such as multi-threaded CPU balancing). This book discusses in detail many of the options available to network architects to build the best data center infrastructure possible. Network high availability is addressed in all of the following chapters.

Network administrators need serviceability functions to take early action based on network trends and events, enhancing network planning and improving network operations center (NOC) and vendor response times. The Cisco infrastructure incorporates troubleshooting and diagnostic features that directly reduce downtime through faster time to resolution. Examples of these features are Smart Call Home, Cisco Generic Online Diagnostics (GOLD), Embedded Event Manager (EEM).

In modern Cisco Data Center infrastructures, the NX-OS operating system is the control plane to enable and monitor many of these HA capabilities. NX-OS delivers continuous operation with failure detection, fault isolation, self-healing features, and reduced maintenance windows.

For more information on NX-OS, see Chapter 9, “NX-OS network operating system” on page 481.

Also see the following link:

http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9402/white_paper_c11-673862.html

A general trend in IT is to shift the focus from “box redundancy” to “topology redundancy” where the traffic loads are distributed across parallel devices that can cope with individual failure without application interruption. A great example of this concept is seen in applications such as Hadoop which is reducing the need to provide hot-swappable redundant fans and power supplies in servers. In the network space, analogous developments such as Cisco FabricPath are enabling “fabric redundancy,” the distribution of workloads across multiple switches.

Cisco highly recommends that hardware be configured with the maximum level of redundancy to provide the highest level of system availability.

3.2.7 First Hop Redundancy Protocols

The purpose of a First Hop Redundancy Protocol (FHRP) is to provide IP routing redundancy by allowing transparent fail-over at the first-hop IP router. There are several protocols that provide FHRP, each with important differentiating characteristics.

FHRP refers to protocols such as Hot Standby Router Protocol (HSRP), Virtual Router Redundancy Protocol (VRRP), and Gateway Load Balancing Protocol (GLBP). They are primarily designed to provide protection against the loss of a default gateway.

HSRP

HSRP was invented by Cisco and was documented in RFC2281 in 1998:

“The goal of the protocol is to allow hosts to appear to use a single router and to maintain connectivity even if the actual first hop router they are using fails. Multiple routers participate in this protocol and in concert create the illusion of a single virtual router. The protocol insures that one and only one of the routers is forwarding packets on behalf of the virtual router. End hosts forward their packets to the virtual router.

The router forwarding packets is known as the active router. A standby router is selected to replace the active router should it fail. The protocol provides a mechanism for determining active and standby routers, using the IP addresses on the participating routers. If an active router fails a standby router can take over without a major interruption in the host's connectivity.”

Over the years, HSRP has been further developed and offers extended functionality such as IPv6 support and the ability to trigger a failover if one or more interfaces on the router go down. The HSRP process monitors upstream connections and if the upstream connection of the primary router fails, the backup router would take over. HSRP is patented by Cisco and its use is licensed by Cisco, and although not an Open Standard, it has widespread adoption.

VRRP

VRRP is an Open Standard alternative to Cisco’s HSRP providing much the same functionality. VRRP (version 3) is documented in RFC5798:

“VRRP specifies an election protocol that dynamically assigns responsibility for a virtual router to one of the VRRP routers on a LAN. The VRRP router controlling the IPv4 or IPv6 address(es) associated with a virtual router is called the Master, and it forwards packets sent to these IPv4 or IPv6 addresses.”

Both HSRP and VRRP enable two or more devices to work together in a group, sharing a single IP address, the virtual IP address. The virtual IP address is configured in each end user's workstation as a default gateway address and is cached in the host's Address Resolution Protocol (ARP) cache.

In an HSRP or VRRP group, one router is elected to handle all requests sent to the virtual IP address. With HSRP, this is the active router. An HSRP group has one active router, at least one standby router, and perhaps many listening routers. A VRRP group has one master router and one or more backup routers.

GLBP

GLBP is a Cisco proprietary protocol and was developed to provide automatic, first-hop gateway load balancing, which allows for more efficient resource usage and reduced administrative costs. It is an extension of Hot Standby Router Protocol (HSRP) and specifies a protocol that dynamically assigns responsibility for a virtual IP address and distributes multiple virtual MAC addresses to members of a GLBP group.

GLBP delivers improvements to HSRP that include these:

- First-hop redundancy:

  GLBP allows traffic from a single common subnet to go through multiple redundant gateways while using a single virtual IP address. It provides the same level of first-hop failure recovery capability as provided by Hot Standby Routing Protocol (HSRP) and Virtual Router Redundancy Protocol (VRRP).

- Reduced administrative burden:

  It is possible to achieve similar performance without GLBP, but that requires supporting multiple client gateway configurations. GLBP eliminates the need to create multiple VLANs or manually divide clients for multiple gateway IP address assignment. With GLBP, all clients on a subnet can be mapped to a single VLAN and a single IP subnet in a wiring closet.

- Reduced costs:

  GLBP does not require a standby router on a dedicated data link, so enterprises no longer incur regular monthly costs for an unused backup data link or the one-time cost for the extra backup router.

  This ability to utilize multiple upstream available paths has a direct impact on the return on the investment of network resources. By increasing usable bandwidth, resources are used more efficiently: file transfer times are lowered and productivity is increase.

- Efficient resource utilization:

  GLBP makes it possible for any router in a group to serve as a backup. This eliminates the need for a dedicated backup router since all available routers can support network traffic.

  GLBP provides upstream load-sharing by utilizing the redundant up-links simultaneously. It uses link capacity efficiently, thus providing peak-load traffic coverage. By making use of multiple available paths upstream from the routers or Layer 3 switches running GLBP, output queues may also be reduced.

  Only a single path is used with HSRP or VRRP, while others are idle, unless multiple groups and gateways are configured. The single path may encounter higher output queue rates during peak times, which leads to lower performance from higher jitter rates. The impact of jitter is lessened and over performance is increased, because more upstream bandwidth is available and additional upstream paths are used.

GLBP features

Following is a list of the GLBP features:

- Fast failover: GLBP failover occurs immediately after the failure in a gateway is detected. Like HSRP, end stations and applications continue as if no failure has occurred.
- Simple configuration: Basic configuration of the protocol is very simple. Configuration is similar to that of HSRP, since many customers are already familiar with it.
- Low Overhead: The protocol uses minimal processor and bandwidth resources.
- Authentication and security: The protocol initially implements a very simple authentication scheme. An 8-character string carried in every packet is compared with a configured value, and only packets that match are accepted. MD5 authentication is planned for a future release.
- Extensibility: The protocol was designed to be very extensible due to the packet encoding. This allows for future extensions and enables retroactive compatibility with older releases of Cisco IOS Software releases.
- Flexibility: Several load-balancing algorithms are available for different network configurations and customer requirements.
- Multi-device redundancy: GLBP allows the traffic for a virtual address to be distributed over as many as four devices. If one of the devices fails, the load from that device is automatically shifted to one of the remaining devices.
- Load balancing modes: Three types of load balancing methods can be configured:
  - WeigHted Load Balancing Algorithm: The amount of load directed to an Active Virtual Forwarder is dependent upon the weighting value advertised by the gateway containing that Active Virtual Forwarder.
  - Host Dependent Load Balancing Algorithm: A host is guaranteed to use the same virtual MAC address as long as that virtual MAC address is participating in the GLBP group
  - Round Robin Load Balancing Algorithm: Each Virtual Forwarder MAC address takes turns being included in address resolution replies for the virtual IP address.
- GLBP Tracking: GLBP uses Cisco IOS Enhanced Object Tracking, a new software feature. The HSRP interface-tracking feature has been widely deployed and permits the standby router to take over in the event of a failed WAN connection (line-protocol “down”), an event more likely than a router failure.

FHRP references

HSRP RFC2281:

`http://tools.ietf.org/html/rfc2281`

VRRP RFC5798:

`http://tools.ietf.org/html/rfc5798`

First hop routing protocols:

`http://www.cisco.com/en/US/products/ps6644/products_ios_protocol_option_home.html`

GLBP data sheet:

`http://www.cisco.com/en/US/products/ps6600/products_data_sheet09186a00800b4694.html`

NX-OS configuration of GSLB:

`http://www.cisco.com/en/US/docs/switches/datacenter/sw/6_x/nx-os/unicast/configuration/guide/l3_glbp.pdf`

3.3 Network services: Load balancing, firewall, WAN acceleration, and traffic analysis

In this section we discuss the additional requirements above connectivity that enable efficient traffic flow and availability. Broadly grouped as “network services,” we provide an overview of load balancing, firewalls, WAN acceleration, and traffic analysis. Other common network services such as DNS and DHCP are not covered in this section.

In general, we can define network services as functions that perform server load balancing, network traffic control, service redundancy, resource management, encryption and security, and application acceleration and optimization. Network services help to keep the network available during routine maintenance and migrations and play a key role in enabling flexible network environments that are ideal for highly elastic and transient applications.

In recent years there has been a major transformation in the way services are delivered to the network, and as with most technologies, the trend is towards greater virtualization. In the past, network services were layered on top of the physical connectivity infrastructure with dedicated appliances. Now we are seeing the convergence of multiple functions into single devices and the transition from physical to virtual appliances. With the emergence of the cloud computing model, these services are further changing to be more mobile along with the applications they are supporting.

Network connectivity components (Switches and Routers) are typically focused on delivering the IP frames to their destination in the most efficient manner. Network Services typically operate higher in the OSI model (layer 4-7) and require greater insights into the application requirements, not just the “source/destination” connectivity.

In Figure 3-19, the relay open system represents a router that operates at Layer 3 in the OSI stack. Network Services typically use information from the transport, session, presentation, and application layers (Layer 4-7) to make their forwarding decisions.

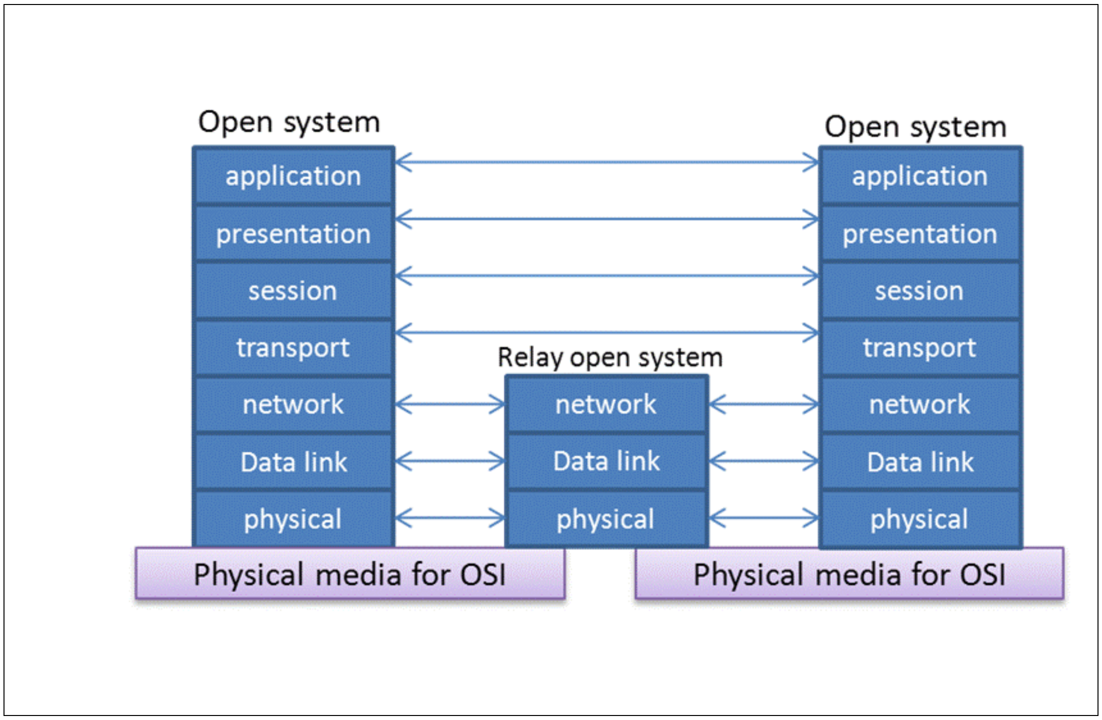


Figure 3-19 Relay open system in the OSI stack

3.3.1 Load balancing

There are many interpretations of what we mean by “load balancing” in the network, the most obvious being Server Load Balancing (SLB). However, to be precise, load balancing within the network infrastructure is a critical component of every data center and it refers to the distribution of traffic across multiple links or devices.

The most common application of “network” load balancing is the use of standard routing protocols (OSPF, RIP, and so on). This kind of load balancing is inherent to the forwarding process in network switches and routers and is automatically activated if the routing table has multiple paths to a destination. Network load balancing is also performed at Layer 2 and there have been multiple innovations to optimize path management in Layer 2 networks in recent years. Another example of network load balancing is the use of FHRP to balance traffic between uplinks in a branch office environment.

Server load balancing

Server load balancing refers to spreading a service load among multiple server computers. Balancing ensures maximum service availability by offering network traffic distribution services. SLB is the process of deciding to which server a load-balancing device should send a client request for service. By distributing user requests across a cluster of servers, SLB optimizes responsiveness and system capacity, and dramatically reduces the cost of providing Internet, database, and application services for large-scale sites as well as small- and medium-sized sites. Load balancing involves scaling the address, name space, transport, session, identity, and business logic of the application.

Key benefits include these:

- Cost savings: Efficient SLB design ensures infrastructure consolidation that saves on capital expenditures (CapEx), such as hardware, software, and IT expenditures; and operating expenses (OpEx), such as rack space, power, and cooling costs as well as the ongoing management cost of servers and infrastructure.
- Better end-user productivity through improved application availability
- Accelerated web services and application response times by up to 500* percent
- Up to 400* percent lower power and cooling expenses
- Increased application, Extensible Markup Language (XML) web services, and data center security
- Up to 75* percent more rapid application deployments and build-outs
  * denotes Cisco example numbers
- Better utilization of server resources through offloading of Secure Sockets Layer (SSL) Protocol and Transport Control Protocol (TCP) processing

SLB facilitates scalability, availability, and ease of maintenance as follows:

- Application availability through load balancing and health monitoring of the application environments
- The addition of new physical (real) servers, and the removal or failure of existing servers, can occur at any time, transparently, without affecting the availability of the virtual server.
- SLB typically allows a new server to increase its load gradually, preventing failures caused by assigning the server too many new connections too quickly.
- SLB Devices typically offer device redundancy with failover capabilities for high availability
- Scalability through virtualization should allow load balancing resources to be logically partitioned and assigned to meet specific tenant service requirements

- Improved workflow through device virtualization: With traditional application delivery solutions, application deployment often proceeds slowly because of the need for complex workflow coordination. For a new application to be deployed or for an existing application to be tested or upgraded, the application group must work with the network administrator. Coordination is required to make the desired configuration changes on the application delivery device (typically a load balancer), a process particularly problematic for the network administrator, who is responsible for helping to ensure that any change to the configuration does not impact existing services.

  Virtual partitioning with role-based access control (RBAC) in modern Application Delivery Controllers (ADC) mitigates this concern by enabling the network administrator to create an isolated configuration domain for the application group. By assigning configuration privileges within a single isolated virtual device to the application group, the network administrator can stay out of the workflow and eliminate the risk of misconfiguration of existing applications enabled in other virtual devices. This improved workflow creates a self-service model in which the application group can independently test, upgrade, and deploy applications faster than ever before.

- Role-based system administration: Multiple departments or stakeholders can independently manage appropriate role-assigned tasks, reducing interdepartmental conflict and increasing productivity.
- High availability with stateful redundancy with both active-active and active-standby designs. Stateful redundancy between resilient ADCs is crucial because it enables user sessions to continue uninterrupted during a failover event. Ideally stateful redundancy can be enabled on a per-virtual-device basis.

As an example, a client request may consist of an HTTP GET for a web page or an FTP GET to download a file. The job of the load balancer is to select the server that can successfully fulfil the client request and do so in the shortest amount of time without overloading either the server or the server farm as a whole.

Depending on the load-balancing algorithm or predictor that you configure, the SLB device performs a series of checks and calculations to determine the server that can best service each client request. The load balancer bases server selection on several factors, including the server with the fewest connections with respect to load, source or destination address, cookies, URLs, or HTTP headers.

The evolution of server load balancing has led to the further consolidation of functionality to enhance the delivery of applications. Later generations of SLB devices as shown in Figure 3-20 are known as Application Delivery Controllers (ADC) since they perform many functions that optimize the application delivery.

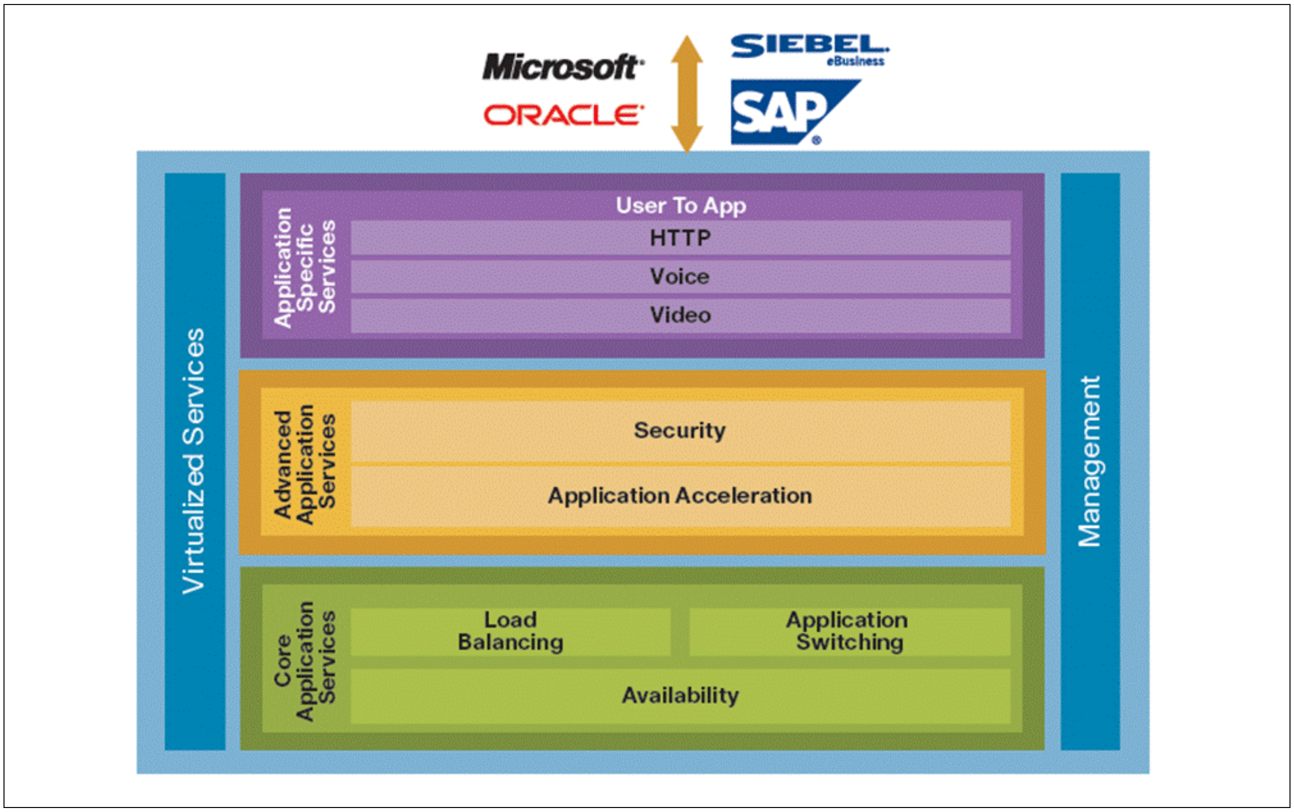


Figure 3-20 Multi-functional Application Delivery Controller

SLB devices now encompass security functions as they are often seen as a last line of server defense. They can protect against application threats and denial-of-service (DoS) attacks with features such as deep packet inspection, network and protocol security, and highly scalable access control capabilities. The device may also provide Network Address Translation (NAT) and access control list (ACL),SSL offload, and HTTP compression capability.

Following the appetite for applications to migrate to external cloud providers, applications loaded onto hosted infrastructures are still expected to deliver consistently excellent service levels even though infrastructure and operations teams do not have full control of the infrastructure or the connection. In particular, application delivery controllers (ADCs) that optimize traffic for the data center, improve customer experience, and connect the private data center hosted internally to a public infrastructure. This has advanced their capabilities and cost models significantly in recent years.

The form factor for ADCs is flexible. SLB functionality can be embedded in the device OS (as seen by Cisco IOS-SLB) or can be part of a dedicated family of Application Network Services.

Virtualized application delivery controllers (vADCs) are becoming a key feature in network infrastructure. Since physical application delivery controllers are the closest entity to the application and improve a user's application experience, virtualized application delivery controllers will allow data center teams to deploy IaaS. vADCs can easily follow the application in order to replicate the VM's networking environment in its new location. In addition, they encourage better collaboration because it is much easier and quicker to put a virtual appliance into the hands of a developer than it is to give that developer access to an expensive piece of hardware that is sitting in a production environment. Network engineers can allow developers to spin up a virtual application delivery controller right on their laptops or lab servers.

With the advent and rapid adoption of mobile computing, ADC devices have further evolved to ensure accelerate the application experience for all users, whether they are in the office or on the road. To enable optimal application performance for remote and traveling users, SLB products use a range of acceleration capabilities to improve application response time, reduce bandwidth volume, and improve the efficiency of protocols. These technologies, including hardware-based compression, delta encoding, and Flash Forward, improve performance and reduce response time by minimizing latency and data transfers for any HTTP-based application, for any internal or external end user.

Additionally, web-based applications are communications intensive, so providing LAN-like service over the web can be a challenge. To meet this challenge, ADC products use the above acceleration capabilities to boost remote end-user application response times. These functions minimize distance-imposed latency when application requests are served to remote users across a WAN and reduce the number of round-trip data transfers and messages required for any HTTP-based application.

For more detail on Cisco Application Network Services, see this link:

http://www.cisco.com/go/ans

For general information on server load balancing, see this link:

http://www.cisco.com/en/US/products/ps7027/prod_white_papers_list.html

Global Site Selectors (GSS)

Geographically dispersed data centers provide added application resiliency and workload allocation flexibility. The deployment of geographically dispersed data centers allows the IT designer to put in place mechanisms that effectively increase the availability of applications. Geographic dispersion allows flexible mobility of workloads across data centers to avoid demand hot spots and fully utilize available capacity.

In addition to the load balancing decisions exhibited inside the data center, the choice of which data center to serve the application to the users is a key decision. An extension to “load balancing” is Global Server Load Balancing (GSLB). With this technology the Internet traffic can be distributed among different data centers located at different locations of the planet. This technology is highly efficient in avoiding local downtimes.

GSLB typically includes:

- Redirection of requests to other data centers in case the usual data center or server does not respond. An algorithm would calculate the shortest distance and best data center from the user request.
- Forwarding the visitor requests to the site that is closes to the place from where the request is raised from a geographic point of view.
- Forward requests to another site if a threshold is reached at a site.

It is possible to load balance traffic between active-active data centers, or redistribute traffic between active-passive data centers with GSLB technology such as Cisco Global Site Selectors (GSS).

Cisco GSS provides connection failover between data centers and helps ensure business continuity, and these devices would interact with the SLB functions inside the data center to ensure optimal delivery end-to-end.

3.3.2 Firewall

Transforming the data center to a virtualized and private cloud environment is one of the biggest trends in IT today. This trend has a huge impact on how we work and conduct business. This new environment transforms the IT services model, business models, and new data center architectures, and offers numerous business advantages.

However, this trend is also causing an architecture evolution in which new security risks and concerns are growing. These challenges are complex because they involve not only technology issues but also substantial process changes caused by new business computing models.

Security concerns are the primary barrier to making a data center transition to virtualization or the cloud. Security should be viewed as a way to gain potential business benefits while eliminating the fear of threats, and not as a hindrance or added cost.

Key characteristics of data center security should include these capabilities:

- Have the flexibility to integrate with complex, multi-site data centers and clouds
- Offer scalability and reliability
- Enable highly secure segmentation
- Block external and internal threats
- Provide visibility throughout the network

One of the most important tools in protecting the data center is the firewall. A firewall is a device that controls the flow of communications across networks by examining the data source, destination, and type; and comparing these with predetermined lists of allowed and disallowed transactions. Firewall technology has come a long way since the early days of basic stateless packet filters operating at Layers 1-3 of the OSI Reference Model.

Current firewall technology is stateful and application aware, and is delivered in many different form factors (hardware and software). Firewalls now operate all the way to Layer 7 through deep packet inspection and filtering on a per process basis instead of filtering connections on a per port basis. Hiding the addresses of protected devices has become increasingly important through the use of Network Address Translation (NAT) and Proxy Server functionality. Firewalls help to ensure network availability and the security of key resources by protecting the network infrastructure against network-layer and application-layer attacks, viruses, and worms.

A common drawback of the firewall is slowing network performance, since it has to actively analyze and manipulate traffic passing through it. Firewall performance today is measured by aggregate throughput (Gbps), connections per second (CPS) and maximum concurrent connections (millions).

A web application firewall is a hardware appliance, server plug-in, or some other software filter that applies a set of rules to a HTTP conversation. Such rules are generally customized to the application so that many attacks can be identified and blocked.

A complete data center security model should offer threat defense, application, and content security, virtualization security, and secure access. This requires much more than traditional firewalls, such as the following features:

- Cloud web security: This allows enterprises to enforce granular web access and web application policy while providing protection from viruses and malware;
- Cisco TrustSec Security Group Tags (SGTs): These integrate security into the network fabric to extend the policy construct on the firewall. Cisco TrustSec Security Group Access classifies systems or users based on context as they connect, and then transforms the way organizations implement security policy across their data center infrastructure. This context-based classification propagates using SGTs to make intelligent policy-based forward or blocking decisions in the data center. In addition, Cisco TrustSec SGA automates firewall rules and removes the management complexity of access control administration.
- Cloud firewall: In a cloud environment, the firewall provides security to the tenant edge inside the data center, separating the compute from the virtual. The cloud firewall offers edge functionality, including site-to-site VPN, Network Address Translation (NAT), Dynamic Host Control Protocol (DHCP), and inspections into network-based attacks. A multi-tenant data center or private cloud naturally requires complete isolation of application traffic between different tenants, applications, and user groups depending on the policies that are in place. Firewalls are now integrated into the hypervisor to provide granular inter-virtual-machine security, gateway services and virtual machine context-aware and zone-based security capabilities.
- Intrusion protection system (IPS): The IPS protects infrastructure and applications from advanced persistent threats (APT) and other sophisticated attacks using state-of-the-art technologies like threat intelligence, passive OS fingerprinting, and reputation along with contextual analysis to deliver superior security protection. An IPS detects threats to intellectual property and customer data by modular inspection throughout the network stack. IPS stops sophisticated attackers by detecting behavioral anomalies, evasion, dynamic threat ratings and detailed logging.
- Security management framework: A comprehensive management solution that enables consistent policy enforcement, rapid troubleshooting of security events, and summarized reports across the security deployment. Tools that manage the security environment should provide visibility across the data center and enable information sharing with other essential network services. Ideally it also maximizes operational efficiency with a powerful suite of automated capabilities and XML API.

The network security infrastructure is increasingly required to enforce identity and role-based policies, as well as to make other contextual decisions. The capability to block traffic to an application or server in the data center or cloud can no longer be based on the typical source or destination addresses of hosts. Now it must be based on the identity or role of the user, the process, or application in the transaction.

Access can also depend on context specific attributes other than identity, including the type of device accessing the application, the location of the user, the time of the request, and more. These context-aware policies are increasingly becoming the responsibility of the data center firewall and intrusion prevention system (IPS), which have to expand their capabilities to detect and decide based on these factors, as well as to monitor for the presence of malware, unauthorized access attempts, and various attacks.

3.3.3 WAN optimization

WAN optimization systems use a variety of techniques such as traffic shaping, compression, caching, protocol spoofing, and de-duplication to effectively increase the bandwidth available and minimize the effect of latency. Application awareness can also speed WAN traffic. WAN traffic typically includes email and web browsing, Windows file shares, rich media, external and business-critical applications, and real-time applications, such as video-conferencing, all of which require prioritizing in line with their network requirements and critical delivery. Since WAN optimization understands the traffic flows, it can add security to the network by blocking specifically identified data types or by integrating encryption.

WAN optimization is deployed to meet application performance, data replication, and disaster recovery requirements by reducing data replication times, increasing bandwidth efficiency, and minimizing latency. Branch Office infrastructure can be efficiently reduced and consolidated back to the data center by deploying WAN optimization. This in turn reduces operations and support functions by eliminating upgrades and maintenance of equipment. WAN optimization is a market expected to expand tenfold from 2011 to 2016.

In December 2010, Cisco Systems commissioned Forrester Consulting to evaluate what was required to support the next-generation infrastructure. The findings were as follows:

- An architectural approach to optimizing how all applications are delivered in a distributed enterprise is only possible when optimization resides consistently throughout the network.
- A pervasive solution that addresses all parts of the network (appliance, virtual, mobile, and embedded) is required to meet the requirements of business.
- Optimization embedded in the network and compute will deliver measurable benefits, such as capex savings, opex savings, and productivity gains, to all users.
- Helping enterprises understand the benefits of deploying the technology and providing guidance on best practices can be as important as the features and functionality of application delivery networks.

In the modern data center, it is clear that a new approach needs to be considered to consolidate the tactical fixes for WAN performance issues over the past decade. For example, if users were experiencing a slow Internet connection or email response at remote offices, WAN optimization controllers were installed. If web servers were slow, then load balancers and, later, application delivery controllers, were placed in front of servers. Embedding application delivery networks is an evolution that will take infrastructure from point solutions to one that has services embedded in networking hardware or virtual instances that is dispersed over the infrastructure.

As a consequence, the evolution of WAN optimization has been dramatic. For example, Cisco WAN optimization capabilities, known as Wide Area Application Services (WAAS), includes several key characteristics:

- Flexible deployment options such as physical / virtual appliance, router-integrated service module, or router-integrated Cisco IOS Software
- Transparent and standards-based secure application delivery
- LAN-like application performance with application-specific protocol acceleration
- IT agility through reduction of branch and data center footprint
- Enterprise-wide deployment scalability and design flexibility
- Enhanced end-user experience
- Outstanding deployment scalability
- Elastic, “scale as you grow” enterprise-wide deployment model

AppNav

To continue this evolution and address new trends such as Virtual Desktop Infrastructure (VDI) and Bring Your Own Device (BYOD), Cisco introduced a technology that virtualizes WAN optimization resources in the data center into one elastic pool and binds capacity to specific applications or branches based upon business requirements. This technology, known as AppNav, has the following additional benefits:

- Efficient and cost-effective expansion of WAN optimization services with an on-demand pool of Cisco WAAS instances
- Flexibility in WAN optimization deployment by the use of flexible policy definitions that dynamically bind business logical constructs to a set of Cisco WAAS pools, making intelligent flow distribution decisions based on the state of the nodes
- Redundancy in the pool of Cisco WAAS instances by AppNav clustering
- Facilitates migration to the cloud with elastic provisioning and upcoming integration of Cisco AppNav technology with the Cisco Cloud Services Router (CSR), thus providing a transparent and optimized connection from the branch office to the cloud
- Comprehensive, simple Cisco AppNav Management and Monitoring with Cisco WAAS Central Manager
- Investment protection for existing network designs

As shown in Figure 3-21, the Cisco AppNav solution can be deployed out of the data path as well as in the data path, depending on user business and network requirements. It comes as an appliance or a module for the Cisco Wide Area Virtualization Engine (WAVE) appliances.

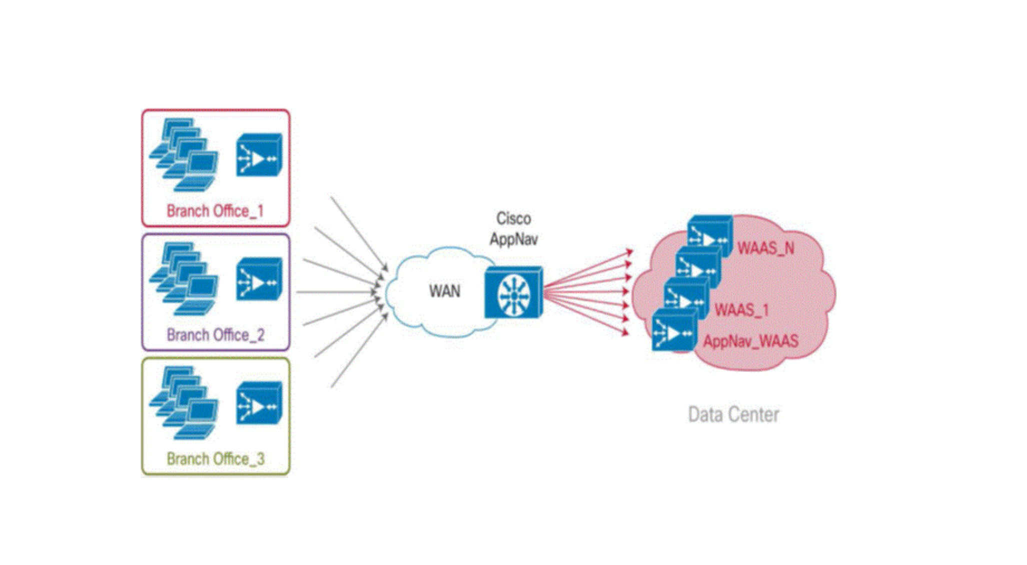


Figure 3-21 AppNav

3.3.4 Traffic analysis

As IP traffic continues its explosive growth across today's networks, enterprise and service providers must be able to characterize this traffic and account for how and where it flows. This presents business opportunities that help justify and optimize the vast investment involved in building a network, ranging from traffic engineering (to optimize traffic flow through the network) and understanding network detailed behavior. Understanding behavior allows customers to implement new IP services and applications with confidence.

Network designers need a way to properly size network capacity, especially as networks grow. Traffic theory enables network designers to make assumptions about their networks based on past experience.

Traffic is defined as either the amount of data or the number of messages over a circuit during a given period of time. Traffic also includes the relationship between call attempts on traffic-sensitive equipment and the speed with which the calls are completed. Traffic analysis enables you to determine the amount of bandwidth you need in your circuits for data, video, and voice calls.

Traffic analysis is used to collect historical data about the network traffic and create charts and graphs of network utilization over time. This can be used to charge back departments that are using the most network traffic or just monitor link utilization over time.

There are a few options to record all traffic that traverses the network links and send detailed information concerning that traffic to a collector or analysis engine. SNMP management just is not sufficient anymore because you need to see the utilization as well as the traffic that is causing the utilization.

Multiple data points are required to provide sufficient analysis that can be used for network monitoring, application monitoring, user monitoring, network planning, security analysis, accounting and billing, and network traffic data warehousing and mining.

To gain maximum visibility from the network, Cisco has introduced several tools to assist with Traffic analysis including NetFlow, SPAN, IP SLA, EtherAnalyzer and NAM.

NetFlow

Cisco IOS NetFlow technology is an integral part of Cisco IOS Software that collects and measures data as it enters specific routers or switch interfaces. By analyzing NetFlow data, a network engineer can identify the cause of congestion; determine the class of service (CoS) for each user and application; and identify the source and destination network for traffic. NetFlow allows extremely granular and accurate traffic measurements and high-level aggregated traffic collection. Because it is part of Cisco IOS Software, NetFlow enables Cisco product-based networks to perform IP traffic flow analysis without purchasing external probes, making traffic analysis economical on large IP networks.

Cisco NetFlow efficiently provides a key set of services for IP applications, including network traffic accounting, usage-based network billing, network planning, security, Denial of Service monitoring capabilities, and network monitoring. NetFlow provides valuable information about network users and applications, peak usage times, and traffic routing. Cisco invented NetFlow and is the leader in IP traffic flow technology.

NetFlow version 9, the latest Cisco IOS NetFlow innovation, is a flexible and extensible method to record network performance data. It is the basis of a new IETF standard. Cisco is currently working with a number of partners to provide customers with comprehensive solutions for NetFlow-based, planning, monitoring and billing.

Cisco NetFlow allows for extremely granular and accurate bandwidth monitoring by recording network traffic into the device cache. Since network traffic has a flow nature to it, the NetFlow accounting data that is built in the cache, characterizes the IP traffic being forwarded. As shown in Figure 3-22, Cisco NetFlow data records exported by routers and switches consist of expired traffic flows with detailed traffic statistics useful to monitor bandwidth and network traffic analysis. These flows contain information about source and destination IP addresses along with the protocols and ports used in the end-to-end conversation.

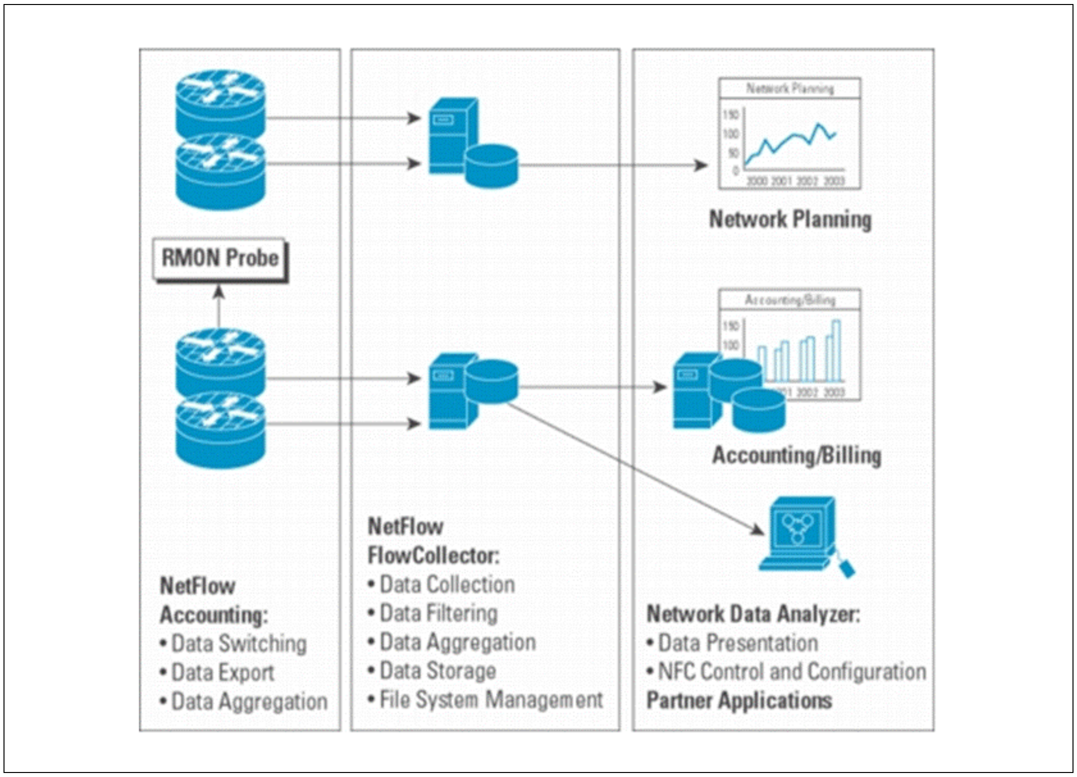


Figure 3-22 Cisco IOS Netflow Infrastructure

This exported NetFlow data is collected and analyzed by NetFlow Analyzer to generate reports on top hosts, top applications, top conversations, and top talkers using the bandwidth in your network.

Cisco NetFlow combined with NetFlow Analyzer yields valuable information about the behavior of traffic and bandwidth monitoring on your network. Armed with this information, it is easier to make critical decisions on bandwidth capacity, security, and optimal usage of network infrastructure. Cisco has developed a robust ecosystem of NetFlow partners who have developed value-add functionality and reporting specialties, including accounting, traffic analysis, security, billing, network planning, and network monitoring.

SPAN

The Switched Port Analyzer (SPAN) feature allows traffic to be mirrored from within a switch from a specified source to a specified destination. This feature is typically used when detailed packet information is required for troubleshooting, traffic analysis, and security-threat prevention.

The Switched Port Analyzer (SPAN) feature (sometimes called port mirroring or port monitoring) selects network traffic for analysis by a network analyzer. A single SPAN session can include mixed sources such as Ethernet ports, Ethernet Port-Channels, RSPAN sources, VLANs, and the CPU control-plane interface.

You can analyze network traffic passing through ports by using SPAN to send a copy of the traffic to another port on the switch that has been connected to a SwitchProbe device or other Remote Monitoring (RMON) probe or security device. SPAN mirrors received or sent (or both) traffic on one or more source ports to a destination port for analysis.

Cisco IOS IP Service Level Agreements (SLAs)

Cisco IOS IP Service Level Agreements (SLAs) enable customers to assure new business-critical IP applications, as well as IP services that utilize data, voice, and video, in an IP network. Cisco has augmented traditional service level monitoring and advanced the IP infrastructure to become IP application-aware by measuring both end-to-end and at the IP layer.

With Cisco IOS IP SLAs, users can verify service guarantees, increase network reliability by validating network performance, proactively identify network issues, and increase Return on Investment (ROI) by easing the deployment of new IP services. Cisco IOS IP SLAs use active monitoring to generate traffic in a continuous, reliable, and predictable manner, thus enabling the measurement of network performance and health.

Ethanalyzer

To help ensure serviceability for mission-critical data center environments, Cisco NX-OS provides comprehensive features, including a built-in protocol analyzer called Ethanalyzer.

Ethanalyzer provides sniffing capabilities to Cisco NX-OS within the operating system, simplifying the need for a third-party network probe to capture control traffic sourced from or destined to the switch, including Spanning Tree Protocol (STP) and Link Aggregation Control Protocol (LACP) traffic.

Ethanalyzer is a Cisco NX-OS protocol analyzer tool based on the Wireshark open source code. This tool is a command-line version of Wireshark that captures and decodes packets. You can use Ethanalyzer to troubleshoot your network and analyze the control-plane traffic.

Cisco Prime Network Analysis Module (NAM)

Cisco Network Analysis Module (NAM) products give network administrators actionable visibility to optimize network resources, troubleshoot performance issues, and ensure a consistent user experience.

With Cisco Prime NAM Software, Cisco NAM products provide quick access to critical network information that helps you meet these needs:

- Improve operational efficiency
- Enhance service levels
- Reduce total cost of ownership and simplify management

The software offers granular traffic analysis, rich application performance metrics, comprehensive voice analytics, and deep insightful packet captures.

The Cisco NAM portfolio includes integrated services modules, self-contained appliances, and virtual service blades. It offers cost-effective choices and deployment flexibility.

Cisco Prime NAM for Nexus 1100 Series delivers traffic and performance analytics that empower network administrators to more rapidly detect and resolve performance issues and improve the use of network resources in the Virtual Machine (VM) network. Integrated with Cisco Nexus 1100 Series and Nexus 1010 Virtual Services Appliances, Cisco Prime NAM offers a smaller network footprint, ease of deployment, and reduced administration cost.

With Cisco Prime NAM for Nexus 1100 Series, you can do the following tasks:

- Analyze network usage behavior by application, host/VM and conversation to identify bottlenecks that can impact performance and availability

- Troubleshoot performance issues with extended visibility into VM-to-VM traffic, virtual interface statistics, and application response times

- Improve the efficiency of your virtual infrastructure with deeper operational insight
- Reduce the risk of deploying new technologies and services in the virtualized data center

Here are some key features of Cisco Prime NAM for Nexus 1100 Series:

- Workflow-oriented user experience improves operational efficiency and user productivity.
- Historical analysis helps tackle unanticipated performance issues.
- Traffic analytics provide multifaceted insight into network behavior.
- Packet Capture, Decode, Filters, and Error Scan expedite root-cause analysis.
- Standards-based API preserves investments in existing management assets.

3.4 IP based storage requirements

IP based storage primarily refers to the use of Internet Protocol (IP) over Ethernet attached storage arrays. IP based storage is an alternative to Fibre Channel attached storage and commercial IP disk systems typically use Gigabit or 10 GE Ethernet connections.

For the purposes of this book, we discuss the Host-Disk use case requirements.

IP based storage is attractive to many organizations due to its simplicity, relatively low cost and the ubiquity of Ethernet skill and connectivity in the data center. FC transport is still the most popular transport for low latency, high bandwidth, high IOPS block data, and it is typical to find FC attached and IP attached storage in the same data center.

Storage has grown significantly over the past decade and information is currently multiplying at an unprecedented rate. This has led to the adoption of lower cost solutions and storage tiering, whereby the most critical data is stored on the best, most appropriate storage.

When we talk about IP storage arrays, we typically refer to iSCSI or NAS devices. iSCSI allows “block mode” data to be transported over TCP/IP, whereas NAS carries “file level” data to be transported over TCP/IP.

Additionally, there is a growing interest in FCoE (Fibre Channel over Ethernet), which operates directly above Ethernet in the network protocol stack. FCoE does not use IP and as such is non-routable. FCoE transport is discussed in detail later in this book.

Fibre Channel over IP (FCIP) is another way to carry block data, but this is positioned as a transport protocol for Data Center Interconnect, not a host-disk mechanism. FCIP is typically used as a SAN-SAN connection to extend the distance of Fibre Channel SANs over geographic distance using TCP/IP.

A growing trend in the storage market is the notion of Unified Storage. Unified Storage systems support block and file data with a mix of connectivity from the array (FC, NAS, iSCSI, FCoE, and so on). It is not uncommon to see disk array interfaces with 10 GE ports that support a concurrent mix of FCoE, NAS, and iSCSI

IP Storage systems require specific networking capabilities for optimal performance and availability. Incorrect configuration or attachment of an array can create catastrophic failures for multiple applications that share the storage resource. Careful consideration of the network needs to be ensured when deploying NAS/iSCSI attached arrays.

The most important considerations are listed here:

- Jumbo frames: To optimize performance, particularly with iSCSI traffic, it is typical to set the MTU size of the Ethernet interface, and the entire data path to at least 2500 bytes. Depending upon the traffic profile, an MTU of 9000 bytes could provide additional performance benefit. Collaboration between the storage and network teams is important to establish the optimal configuration
- Oversubscription and disk array placement: The physical attachment of the array is critical to avoid performance issues. Since the array is generating large traffic volumes, it is important to ensure that the IO path should not be constricted. The array should be attached to the network to avoid oversubscription.
- Localization: It is often advantageous to place the array as close to its consumers as possible. If practical, place the array within a very short IO path to the servers that use its data. Sometimes this is impractical, for example if the array is a “global” resource pool, so limiting latency (that is, reduce L2 or L3 hops) is important.
- Bandwidth: Most large NAS systems will have multiple 10 GE interfaces today. It is important that wire rate ports should be used wherever possible, as the IO tends to grow significantly over time. Reservation of spare adjacent ports may be advisable for future expansion. Some NAS vendors are already considering 40 GE Ethernet interfaces.
- Link aggregation: In some cases, the array will use multiple 1 GE or 10 GE ports to aggregate the IO bandwidth. Care should be taken to match the mechanism used on the switch and the array port (PaGP, LACP, 802.3ad, and so on). There are Parallel File Systems on the market such as IBM GPFS™. In some cases it is better to allow each parallel interface to be independent (not aggregated): collaborate with storage administration for optimal configuration.
- High availability: Typical storage attachments are using multipathing techniques (such as MPIO) to load balance traffic between the host and disk. Care should be taken to ensure that the two multipath exit points from the array are connected to separate switching resources that do not share the same failure domain.
- VLAN isolation: Many customers prefer to assign storage traffic to a separate VLAN so that they can isolate and easily identify the storage resource pool. Some iSCSI host configurations demand logical VLAN separation for correct operation over the NIC. Some NAS/iSCSI arrays can 802.1q tag the traffic on the interface.
- QoS: The NAS/iSCSI traffic should be assigned an appropriate QoS policy in the network that matches the criticality of the data.
- Full duplex and flow control: The Ethernet switch ports should be set to full duplex. Appropriate flow control on the switch should match the array capabilities. Typically this is IEEE 802.3x (PAUSE frame) or the follow-on priority-based flow control, as defined in the IEEE 802.1Qbb standard, provides a link-level flow control mechanism that can be controlled independently for each Class of Service (CoS)
- Port buffering: Bursty storage traffic can be challenged by shallow port buffers on switch interfaces. Care should be taken to connect the array to an appropriate port for maximum sustained performance. Vendor guidance should be obtained.
- Path to host: Obviously, the array needs to be reachable from the host by a routed or switched path.
- Active links: Wherever possible, the array uplinks should be forwarding and balanced.
- Monitoring: Ideally, the IO should be monitored to trigger notifications if thresholds are exceeded. As a mission critical device, this will give the administrator plenty of time for capacity planning and remediation before problems arise.
- Security: Ensure that appropriate controls are in place to protect access to this valuable resource. DoS attacks on an array could be disastrous.

3.5 Regulatory compliance

When designing networks for the data center, it is always important to ask which regulations and compliance requirements we need to meet. Certain industries have very stringent restrictions upon where data may live, and which applications may talk to each other.

Standards such as the Payment Card Industry Data Security Standard (PCI DSS) for the payment card industry and regulations like Health Insurance Portability and Accountability Act (HIPAA) for the health industry impose a series of requirements to be followed by organizations, and for which noncompliance may lead to the revocation of licenses, stiff penalties, and even legal actions. The data center design must incorporate a rich set of security best practices and functions commonly required by regulations and standards to achieve compliance.

As regulatory compliance becomes a core requirement for organizations in more industries, businesses must develop new capabilities for controlling the kinds of information traversing their network, how that information is used, and who can access it. Organizations not only face the challenge of becoming compliant, but of staying compliant as the network continuously evolves with business needs.

Some compliance examples are listed here:

- Business and Financial Services: Payment Card Industry Data Security Standard (PCI DSS), BASEL II, Federal Financial Institution Examination Council (FFIEC), Sarbanes-Oxley Act of 2002,
- Consumer Finance: Gramm-Leach-Bliley Act (GLBA)
- Government: Federal Information Security Management Act (FISMA)
- Healthcare: Health Insurance Portability and Accountability Act (HIPAA)
- Retail: Payment Card Industry Data Security Standard (PCI DSS)
- Energy and Utilities: NERC CIP compliance (CIP 009)

These regulations often call for IT functions in the areas of data confidentiality, integrity, availability, and auditability (CIAA) that older networking equipment cannot effectively support. The current Cisco portfolio of routers and switches and Cisco IOS/NX-OS Software is designed to meet this challenge, while providing benefits in the areas of performance, management, and availability.

Organizations are wrestling with information security demands that span many overarching business challenges such as complying with regulatory requirements, preventing data loss, and blocking malware. The problem is that dealing with these types of challenges requires a true security solution, not just security products. To prevent data loss alone, for example, businesses need a combination of strong perimeter defenses, malware defenses, identity services, endpoint security, policy enforcement mechanisms, and security monitoring tools, as well as a strong plan for making them all work in concert.

Each regulation will have different IT requirements. For example, the PCI compliance process requires companies to perform a thorough audit of their networks, policies, and processes;

- Install and maintain a firewall configuration to protect cardholder data
- Do not use vendor-supplied defaults for system passwords and other security parameters
- Protect stored cardholder data
- Encrypt transmission of cardholder data across open, public networks
- Use and regularly update anti-virus software or programs
- Develop and maintain secure systems and applications
- Restrict access to cardholder data on a need-to-know basis
- Assign a unique ID to each person with computer access
- Restrict physical access to network resources and cardholder data
- Regularly test security systems and processes
- Maintain a policy that addresses information security

From a network design perspective, corporate standards and compliance policy should be fully understood so that the right connectivity and processes can be implemented in the data center.

Appropriate tooling in the data center should be deployed for compliance reporting and management. For example, the Cisco Secure Access Control Server, Cisco Security Manager deliver centralized management, monitoring, and remediation.

Network Admission Control (NAC) determines which client devices are granted network access.

Some customers or governments will dictate the use of specific features, such as IPv6. It is important that the network architect is fully aware of corporate standards in IT as well as Regulatory Compliance restrictions prior to creating the design.

You can learn more about Network Design considerations for Industry Compliance here;

http://www.cisco.com/en/US/solutions/collateral/ns340/ns394/ns171/ns625/white_paper_newer_cisco_technologies_aid_customers_in_achieving_regulatory_compliance_ns431_Networking_Solutions_White_Paper.html

You can learn more about PCI compliance here;

http://www.cisco.com/web/strategy/retail/pci-compliance.html

3.6 Differentiated service levels on a shared infrastructure (QoS)

Quality of Service (QoS) refers to the capability of a network to provide better service to selected network traffic. The primary goal of QoS is to provide priority, which may include dedicated bandwidth, controlled jitter and latency, or improved loss characteristics. It is important to ensure that providing priority for one or more flows does not make other critical flows fail.

Most data centers are typically built with enough bandwidth to eradicate network bottlenecks, but nothing can ever be guaranteed over the lifespan of the data center infrastructure. Consequently, careful analysis of potential congestion points should be evaluated and balanced against the business priorities of critical applications. In reality, it is often much easier to add bandwidth in the data center to relieve congestion than it is outside the data center.

Logical QoS configuration adds complexity to the operational management of the network. Outside the walls of the data center, logical QoS is a critical service for the network.

For example, limited resources may be deployed between data centers due to service provider limitations or opex/capex restrictions. In the event of congestion on this limited resource, an analysis of the traffic over the link should be undertaken to define priorities. Real time synchronous replication could be the most critical data on the link, and network QoS could be applied to ensure that nightly backups or archives does not slow down the replication traffic, but optimizes the limited resources. Using QoS, you can limit the bandwidth consumed over a backbone link by FTP transfers or give priority to an important database access.

As previously mentioned, IP Storage arrays may require some priority handling to ensure that applications are not impacted by data sharing the links, switches or routers. In some corner cases, QoS may be a valuable asset to the data center where there may be extensive video or multimedia traffic that has the capability to saturate the fat pipes in the data center.

QoS is an integral part of FCoE as will be discussed later in the book.

3.7 Data center cabling

Data center facilities have changed significantly over recent years and it is difficult to keep track of current best practice. The mass adoption of PoD based architectures within the data center has driven a demand for more modular physical infrastructure, including cabling.

The subject of DC Facilities is a book in its own right, so this section is meant to give the reader an insight into current practices and choices for today’s DC deployments.

The factors affecting the choice of DC physical design are usually as follows:

- Media/cable cost
- Density
- Performance
- Power consumption
- Active equipment cost
- Manageability
- Future proofing (applications)
- Modularity / flexibility
- Power, cooling, loading

Many older data centers have rigid cabling infrastructures that the network architect must accommodate, however, it is typical today to see much more flexibility, modularity, and scalability in modern structured cabling systems for the data center.

Table 3-1 shows how a common rack density today might look.

Table 3-1 Rack density

6KW cooling per rack	10KW cooling per rack
24 1RU servers per rack @ 250W	40 1RU servers per rack @ 250W
48 Fibers per rack @ 2x10GE/server	96 Fibers per rack @ 2x10GE/server
96 copper cables @ 4x1GE/server	160 copper cables @ 4x1GE/server

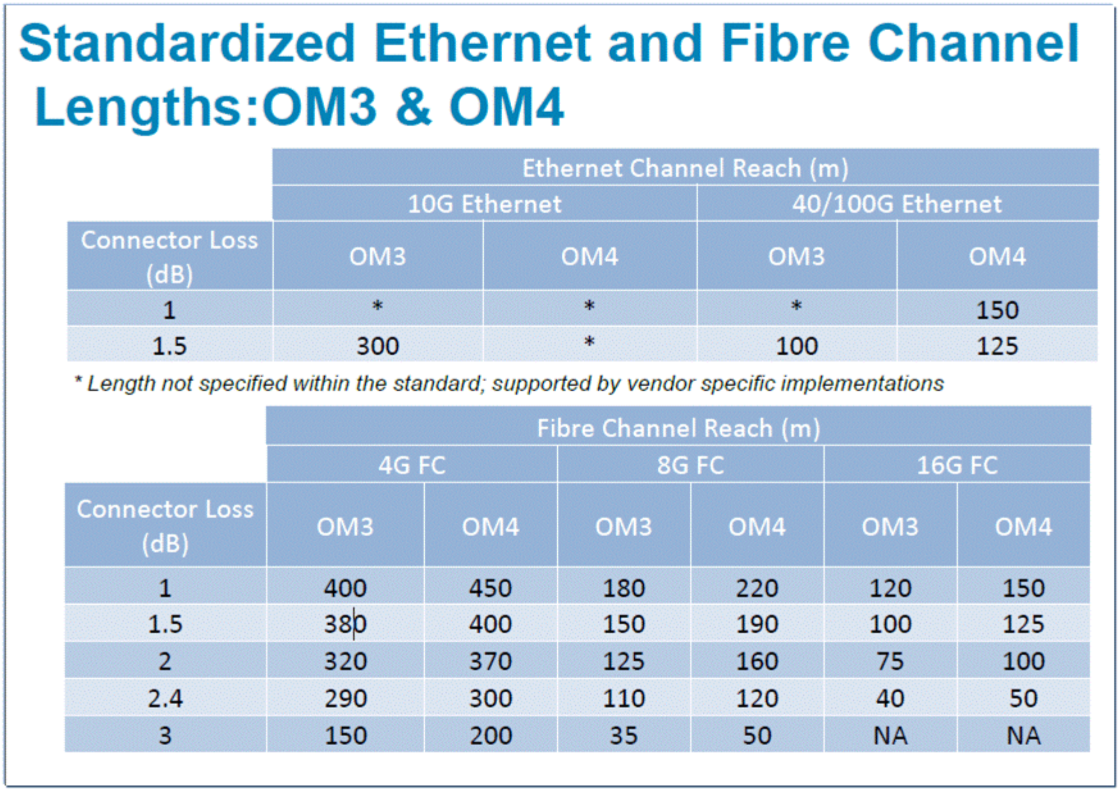
Advancements in network technologies (40 GE, 100 GE) have driven a requirement for higher performance fiber, and mass adoption of 10 GE has also driven down costs for copper cabling. The most prevalent 10 GE connector today is the SFP+ (specification SFF-8431) due to its low cost, low power and low latency characteristics.

Cabling design implications for 10/40/100 GE are as follows:

- Fiber Type: 50 micron OM3 or OM4 for 10/40/100G
- Fiber count: 2f(10G)-8f(40G) -20f(100G) per circuit
- Link Budget (channel insertion loss): 2.6dB(10G) -1.9/1.5dB
- Link Reach: 300m (10G)-100/150m (40G/100G)
- Connector Type: Duplex LC(10G): 12f MTP(40G) - 24f MTP(100G)
- Fiber Polarity: TX to RX - more complex with parallel optics
- Skew: Factor for parallel optics (16m)
- Server virtualization and data center bridging (FCoE): Increases the density of 10G connections in and to a server cabinet
- Physical architecture: Locations of X-connects
- Pathways and Spaces
  - Rack space requirements
  - Cable tray
- Patching density and management

Increased cable media options are available to meet the needs of new architectures. These new media types have been created to accommodate the new performance requirements, but also the green agenda, low power cabling solutions (CX1/10G-CR, OM4 Fiber). It is becoming less important to see 300m capable patching for 10GE, but essential that spans of 100-150m are available for much higher link speeds.

Figure 3-23 shows the standardized Ethernet and Fibre Channel Lengths for OM3 and OM4 cabling.



Standardized Ethernet and Fibre Channel Lengths:OM3 & OM4

	Ethernet Channel Reach (m)			
	10G Ethernet		40/100G Ethernet	
Connector Loss (dB)	OM3	OM4	OM3	OM4
1	*	*	*	150
1.5	300	*	100	125

* Length not specified within the standard; supported by vendor specific implementations

	Fibre Channel Reach (m)					
	4G FC		8G FC		16G FC	
Connector Loss (dB)	OM3	OM4	OM3	OM4	OM3	OM4
1	400	450	180	220	120	150
1.5	380	400	150	190	100	125
2	320	370	125	160	75	100
2.4	290	300	110	120	40	50
3	150	200	35	50	NA	NA

Figure 3-23 Cable Length for OM3 and OM4

Within the modern data center, we are seeing a drive towards pre-terminated cabling solutions to increase service agility. There is growing interest in cassette based fiber patching systems that can be deployed rapidly by general maintenance staff, without the need for expensive polish, and termination skills.

The availability of low cost twinax patching for 10 GE has driven a shift in network topology from concentrated EoR switches to distributed ToR connectivity. These CX1 integrated patch cords are typically limited to @ 15 meters. Likewise the port density on access layer devices like the Nexus 2000 Series are driving new “PoD” concepts such as “twinrack” configurations where a ToR device feeds two racks, not one, with x-patching between adjacent cabinets.

10GbaseT connections over Cat6/6A/7 are rolling out, but current limitations on Bit Error Rates (BER) mean that it is not used for FCoE.

Another innovation in the data center is a shift from hierarchical network architectures to “fat tree” or CLOS based topologies: This creates a significant change in best practice cabling, and reinforces the need for modular, flexible structured cabling. An example of a CLOS based network is the Spine-Leaf topology described earlier in this book.

Ultimately, what the network architect can do in the data center will be dictated by the facilities team governing the data center. New technologies to reduce the overall cost of cabling (such as FCoE) are proving very attractive to data center managers.

Collaboration between server, storage, network, and facilities teams is imperative to build the optimal data center infrastructure. All too often, data center architecture is designed in “silos” whereby the CIO loses out financially, being unable to benefit from converged technologies and a collaborative design team. The Cisco Data Center facility in Allen, Texas saved over $1m in cabling due to the unified fabric.

Cabling costs for the data center can be a significant part of the overall budget, and massive savings can be made by “virtualizing the wire” and consolidating multiple transport types over the same cable. This phenomenon will increase in pace as higher speed connections become available in the data center. Separate Fibre Channel networks (maximum 16G-FC today) will become less appealing as low cost 40 GE (or even 100 GE) connections become ubiquitous, with FCoE functionality.

Cabling with FCoE in mind ensures that the cabling plant is capable of delivering a “wire once” infrastructure, where the application requirements can change over time without disruption, also enabling “lights out” data center operations.

Whatever choices are made for the cabling infrastructure, it is most important to ensure that the cabling architecture chosen will scale with the life of the data center.

You can learn more about DC Facilities and Cabling solutions by reading *Design for a Productive Data Center*, available at this link:

http://www.cisco.com/web/about/ciscoitatwork/trends/tr_2007_06_article005_pdc.html

Or, you can view the CiscoLive presentation *BRKDCT-2998 Cabling for Next Generation Data Center Technologies,* available online at this link:

http://www.ciscolive365.com

4

Data center physical access layer evolution

In this chapter, we examine the evolution of the data center physical access layer over the past two decades. A traditional hierarchical network design consists of the following layers:

- Access layer
- Distribution layer
- Core layer

We consider the physical cabling and data center layout, as well as possibilities for the server architecture:

- End-of-Row (EoR)
- Middle-of-Row (MoR)
- Top-of-Rack (ToR)

Finally, we discuss a number of solutions pertaining to the overall network design.

4.1 Introduction to server access layer

Over the past two decades, server attachment has evolved considerably. In the late 1980s and early 1990s, customers found UNIX and Intel based servers attractive for department LAN segments. Over time, businesses recognized the value of the information contained on the servers and the cost of maintaining distributed servers. Enterprises decided it was best to consolidate their servers into the data center. Early servers ran multiple network protocols, IPX, NETBIOS, and TCP/IP; and there were multiple network types, Token Ring, Ethernet ARC NET, and FDDI.

In the mid-1990s, 10/100 Mbps Ethernet switching was prevalent. As the decade progressed, gigabit Ethernet and Layer 3 switching became widely available.

By the year 2000, the industry had standardized on switched Ethernet and IP only network cores. The hierarchical switched Ethernet design became the “Best Practice” data center design. Figure 4-1 shows a traditional hierarchical network design.

This model is made up of three distinct layers:

- Access layer:

  The access layer provides endpoint device connections. Servers and users are connected to the network on this layer. It is not uncommon to label the access layer as a server or user access layer. This level is normally performed by Layer 2 Ethernet switches. A ratio of many server ports to few uplink ports always exists for this layer. Typically, 48-288 servers connect to a switch with to 2-8 uplinks. It is ideal for the uplink speeds to be an order of magnitude higher than the server connections (such as 100 Mbps Servers with 1 Gbps uplinks and 1 Gbps Servers with 10 Gbps uplinks). When the servers and the uplinks are the same speeds, multiple uplinks are required to overcome the bandwidth over commitment. This book discusses the implications of over subscription in Chapter 3, “Designing networks for the modern data center” on page 131.

  Historically, the wire between the server and the access layer has been the demarcation of roles and responsibilities between the server and network support organizations.

  Quality of Service marking or access lists may be applied at the access layer.

- Distribution layer:

  The distribution layer aggregates multiple access layers, and provides Layer 3 forwarding and security services to them. Network policies and traffic manipulation are implemented on this layer. This layer is deployed with Layer 3 Ethernet switches that have multiple high speed links. These links have evolved from 1 Gbps to 10 Gbps and will continue to evolve to 40 Gbps and 100 Gbps. Hot Standby Routing Protocol (HSRP) is typically configured on the downlinks facing the access layer. Access Control List (ACL) filtering is supported by this level as well. Firewalls, load balancers, and other networking services may also be attached to this layer.

  The network administrator needs to decide if VLANs will be propagated beyond the distribution layer. Best Practice has historically made this Layer 3 routing only.

- Core layer:

  The core layer (or backbone) provides high-speed transportation services between different parts (or building blocks) of the data center network. No traffic manipulation is performed on this level. The distribution and core layers provide routing (Layer 3). A dynamic routing protocol would be used between the distributions and the core.

This hierarchical model is shown in Figure 4-1.

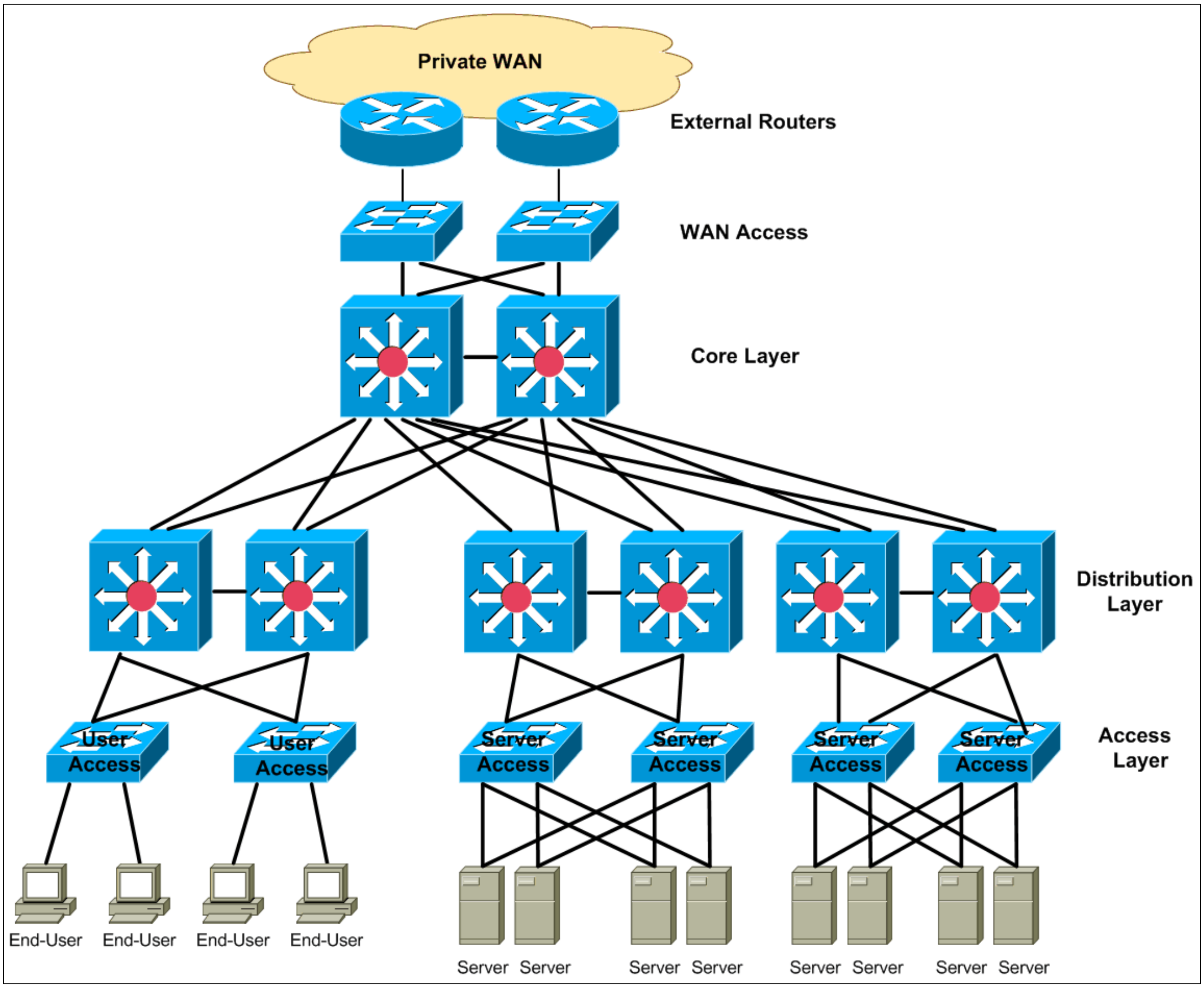


Figure 4-1 Traditional hierarchical network design

4.2 Traditional server access layer

Traditional high availability network access for servers was accomplished by putting two Network Interface Cards (NIC) in the same server. Depending on the operating system, the NICs could be configured in the same subnet/VLAN or in different subnets/VLANs.

High availability support network connectivity could be accomplished in three ways:

- Dual NICs on the subnet / VLAN were possible.
- Dual NICs on separate subnets / VLANs were also possible.
- Larger servers were capable of running a dynamic routing protocol and could support multiple NICs on multiple subnets and VLANs.

Figure 4-2 shows a server with two NICs in the same subnet and VLAN across two switches. Due to limitations of spanning tree and network virtualization one of the uplinks from each access switch was blocked by Spanning Tree Protocol (STP).

In order to efficiently utilize the uplink bandwidth multiple subnets/VLANs could be placed on the same access switch. The network administrator could utilize both uplinks by making one uplink the active path for even VLANs and the other uplink the active path for odd VLANs. Attaching a single access switch to two upstream distributions was referred to as a V-shaped configuration.

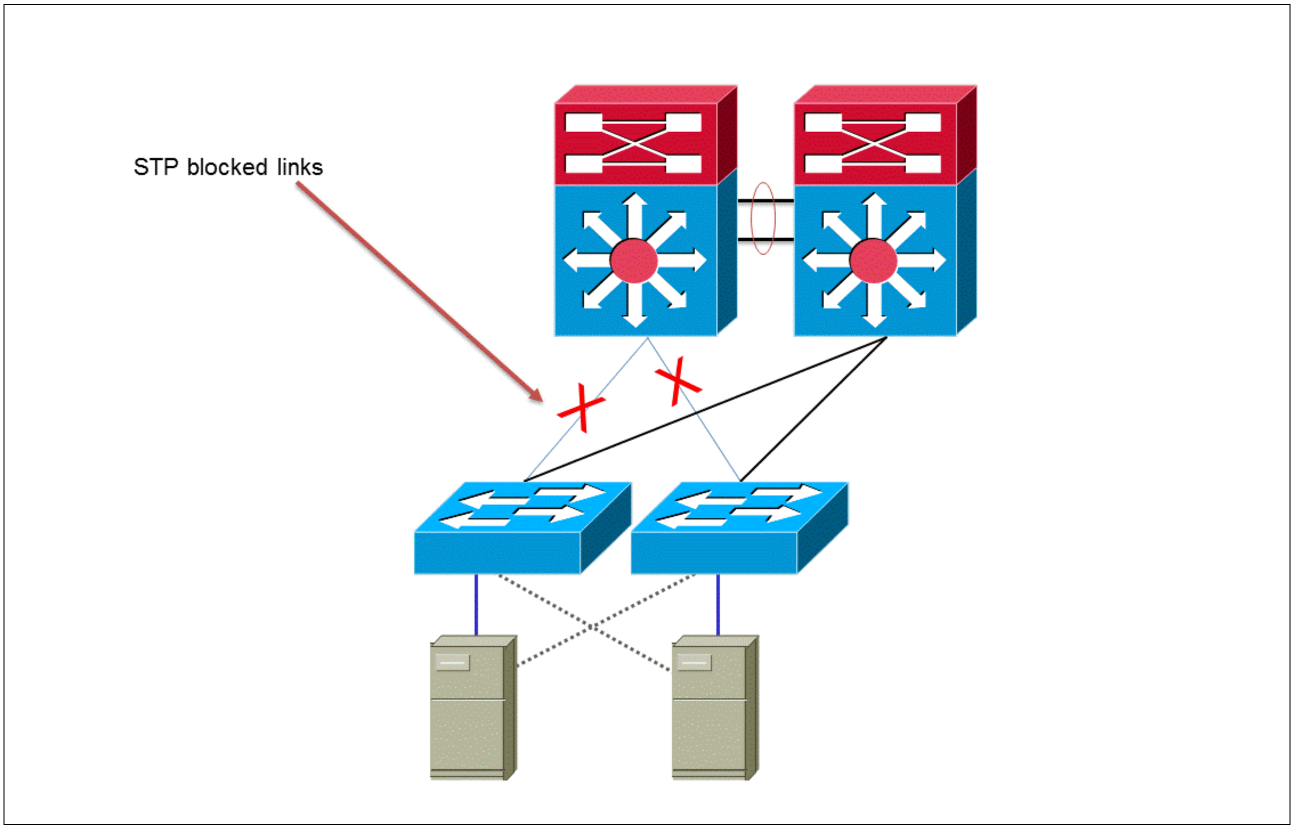


Figure 4-2   V-Shaped server access design with one uplink blocked by spanning tree

Another popular alternative was connecting two (or more) access switches together and uplinking each access switch to only one distribution switch. This was referred to as a U-shaped configuration. Figure 4-3 shows a "U" shaped network access layer with two NICs attached to two access switches on the same subnet. Notice that one server in the diagram is a single server with two NICs. The other servers in the diagram are an IBM High Availability Cluster Multiprocessing (IBM HACMP™). Each server has a single NIC and is attached to one switch. The servers are redundant, so that if one fails, the other takes over. This configuration requires a cross connect between the two servers. By only enabling Layer 3 routing between the two distribution switches, the entire U-Path is active.

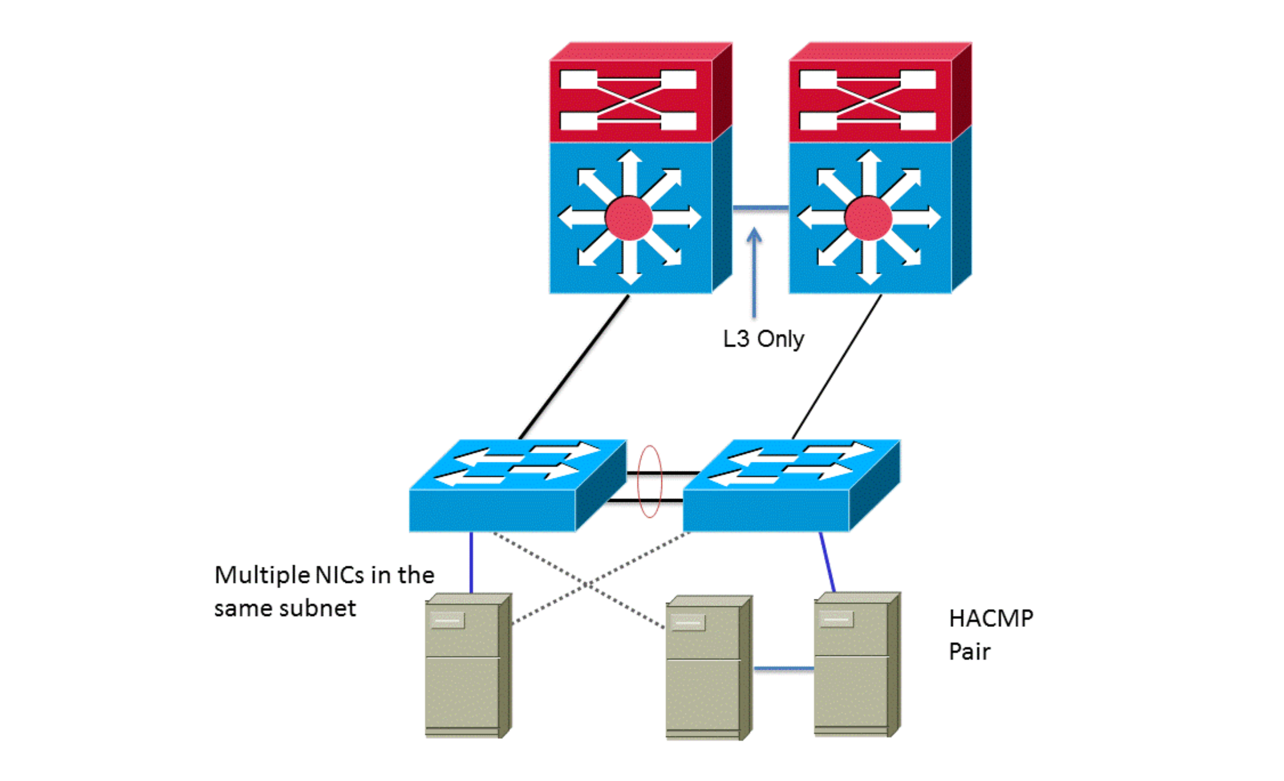


Figure 4-3 "U" Shaped access layer on the same subnet

Best Practice for server access was as follows:

- Do not mix users and servers on the same access switches.
- Do not attach the servers directly to the distribution or core switches.
- Have one subnet per VLAN. Do not exceeds 255 hosts in a subnet (/24 mask).
- Have one VLAN per switch. Do not span VLANs across multiple switches. In reality, most customers have multiple VLANs spread out across many switches.

The hierarchical design discussed so far has been end-user facing, and such networks are referred to as *front-end networks*.

Additional networks can be created that are referred to as *back-end networks*, which do not allow connections directly from users. A back-end network may be created for multiple purposes:

- Offloading data backup.
- Provide separation between application tiers; for example, traffic between the web server and its database server.

- "Private" communication mechanism. Early clusters were built with back-end networks for communicating between servers.
- Management networks where a server has a management or console network.
- Separate production traffic from other types of traffic for bandwidth and performance reasons.

The simplest way to create a back-end network is to put additional NICs in the servers and create a separate switched network for the secondary interfaces. Figure 4-4 shows a back-end network used for backup. This approach also works for SAN networks.

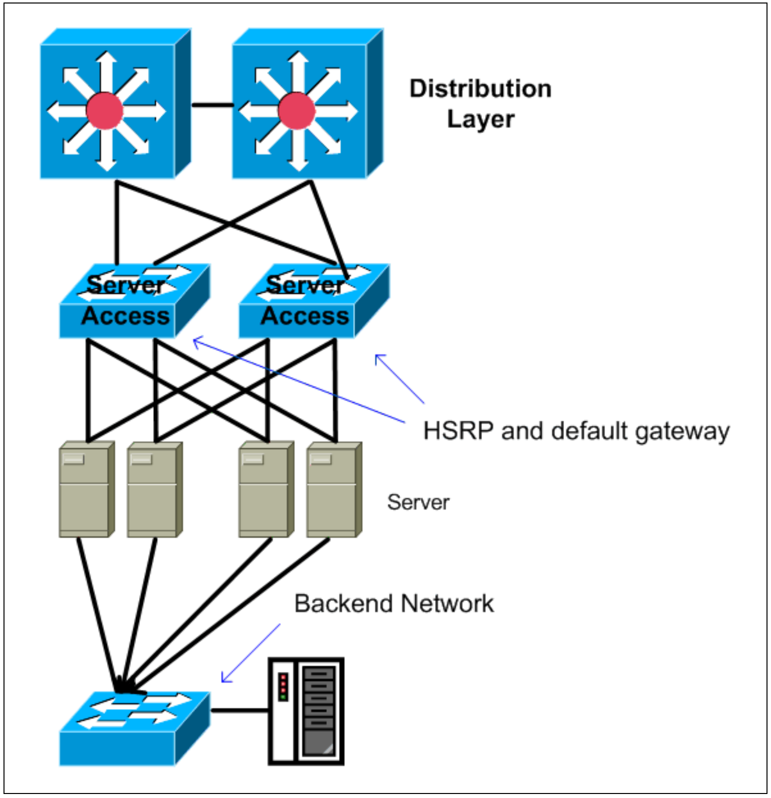


Figure 4-4 Back-end network

The traditional server access layer models worked very well for many years, and are still deployed in many data centers today. However, due to significant advancements in network technology, there are now more robust alternatives to these traditional designs that better meet the rapidly changing business demands placed on the modern data center. We explore these alternatives in the remaining sections of this chapter.

4.3 Physical cabling and data center layout

How the access layer is designed should be looked at from both a physical perspective (physical placement of access switches, what type of cabling is used, how servers are cabled to the switches, etc…) and from a logical perspective (Layer 2 vs. Layer 3, Spanning Tree versus Fabricpath, port channel or active standby NIC attachment, and so on). We look at both the physical and logical aspects of access layer design.

A traditional data center design has the server and the network devices in separate racks. Servers are all placed in the same racks and rows. Network equipment is placed in separate racks and rows. Server access switches have been centrally located in the data center. Figure 4-5 shows an example of a logical picture of a structured cable plan for centralized server access switches. Servers are attached directly to a server patch panel from the NIC. Cable runs are created from the server areas to the network switch areas. Access switches are connected to network patch panels. Jumpers are then used to make the connections from the servers to the network server access server.

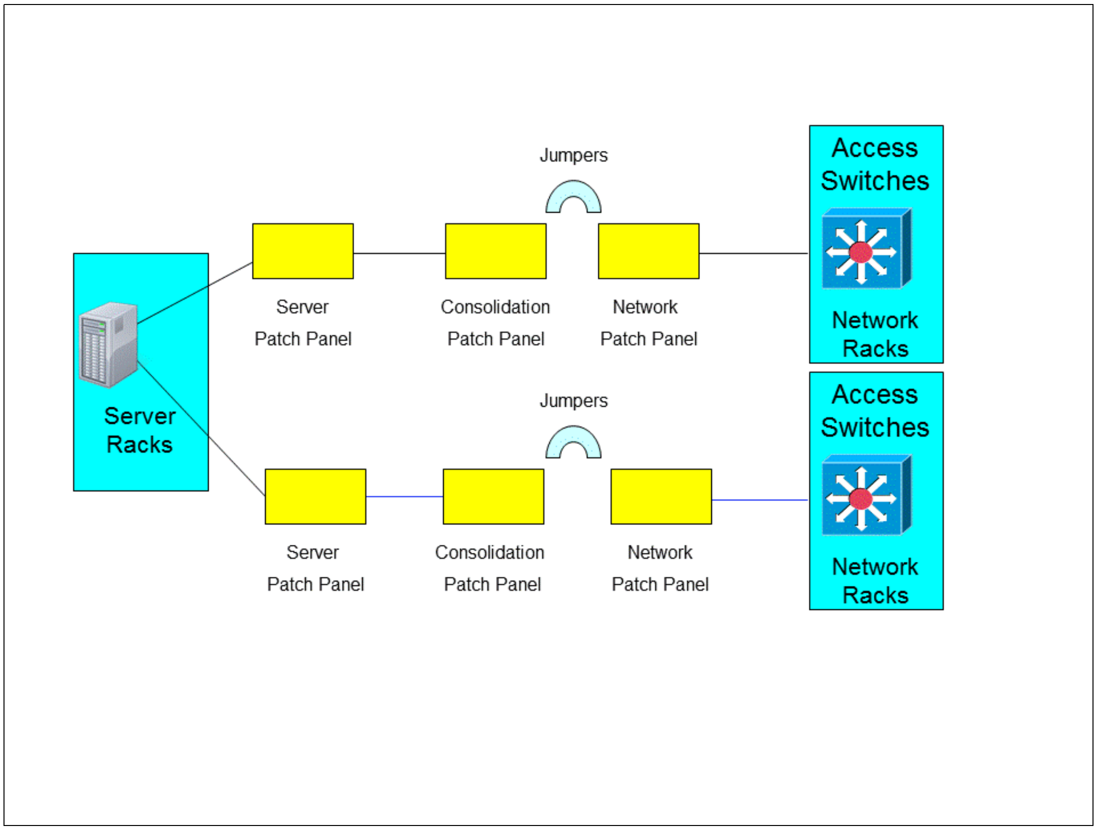


Figure 4-5 Logical drawing of server access patch cabling

Due to the extensive complexity and volume of cable runs in centralized access switch designs, labor costs to install the cabling can sometimes equal or exceed the cost of the actual cabling system hardware.

Figure 4-6 is a picture of centrally located server access switches. Although incredibly neat, it would be difficult to replace a module or a power supply.

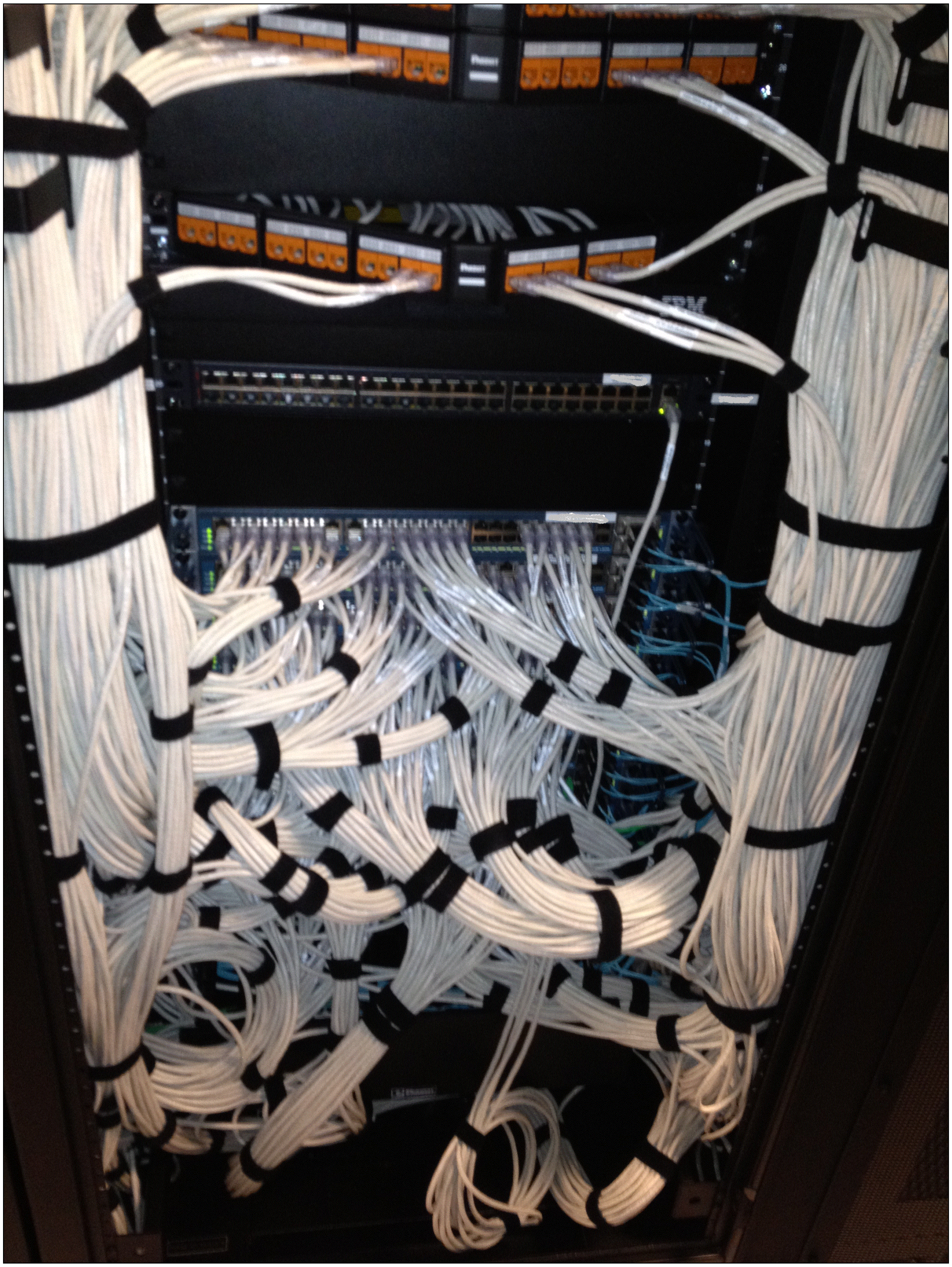


Figure 4-6 Centrally located access switches

Centrally locating server access switches also require running large amounts of copper cabling. Large bundles of server cables under the raised floor act as an air dam that can restrict air flow under the raised floor.

Figure 4-7 shows a picture of server access cable under the raised floor. Utilizing a Point of Delivery (PoD) data center layout will greatly reduce these bundles of cables.

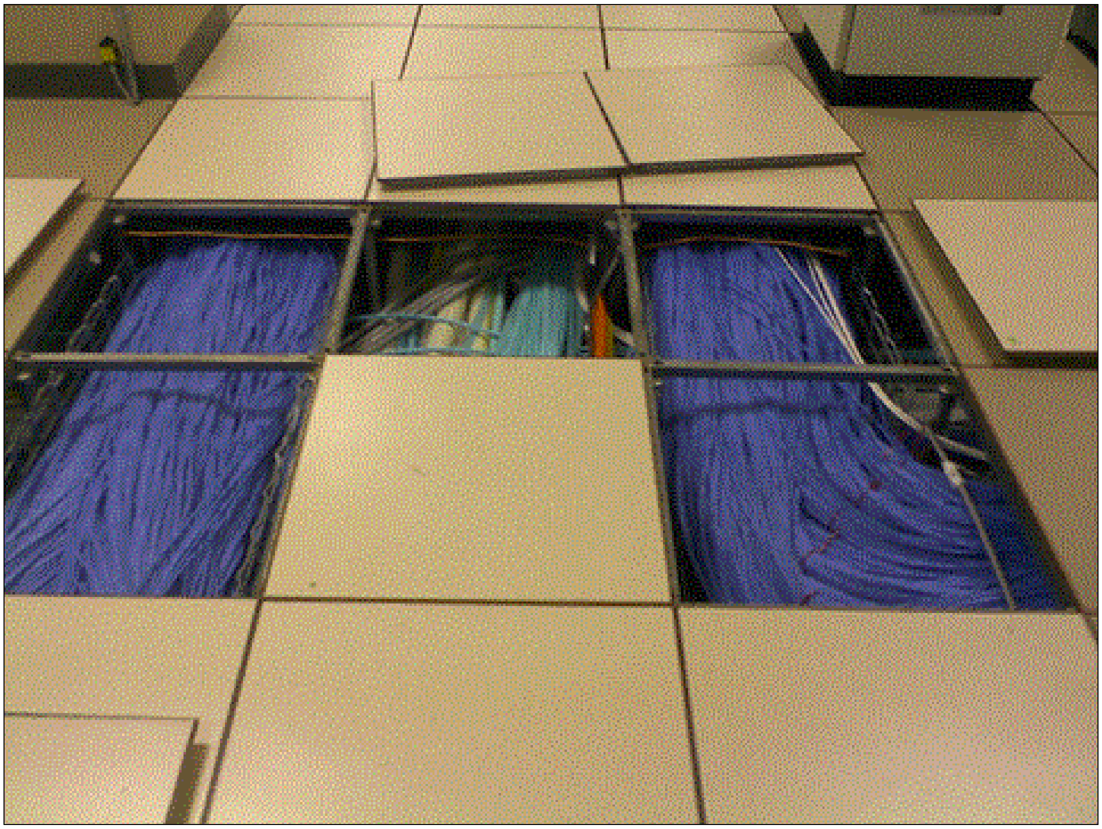


Figure 4-7 Server access cables under the raised floor

4.3.1 Modular data center infrastructure

The cabling and infrastructure design for a modern data center is governed by multiple factors. The TIA/EIA-942 Telecommunications Infrastructure Standard provides guidelines for data center cabling infrastructure that customers can adopt as a guide in the data center cabling and planning process. Other standards, such as BICSI, provide guidelines for data center cabling and implementation. The TIA/EIA-942 cabling specification considers the need for flexibility, scalability, reliability, and space management. See `http://www.tiaonline.org`. While the standard provides guidelines, specific design elements will vary with each data center. General considerations that apply to all data centers include these:

- Support for storage devices (Fibre Channel, Small Computer System Interface [SCSI] or NAS, FCoE, and so on)
- Support for convergence and unified fabric with growth factors incorporated
- Reliability, scalability, and redundancy
- High-capacity and high-density server access requirements
- Flexibility and expandability with easy access for moves, additions, and changes
- Migration from Gigabit Ethernet to 10 Gigabit Ethernet server connectivity with future support for 40 and 100 Gigabit Ethernet
- Cabling architecture balance with power, cooling, structural loading, management, and operations needs

In the context of the TIA/EIA-942 simplified logical data center layout, the Top-of-Rack (ToR) architecture maps directly to the Equipment Distribution Area (EDA) and the Horizontal Distribution Area (HDA).

The TIA/EIA-942 specification provides a simple reference for data center cabling that supports different cabling schemes, EoR or ToR, to meet differing needs from a physical and operational perspective. The ToR model defines an architecture in which servers are connected to switches that are located within the same or adjacent racks, and in which these switches are connected to aggregation switches typically using horizontal fiber-optic cabling.

4.3.2 Server cabling

The dominant server attachment technology in data centers of the past decade has been Gigabit Ethernet (GE). Cabling technologies for GE are well understood, and as such, will not be discussed here. Due to decreasing cost of 10 Gigabit Ethernet (10GE) technologies, the shift from 1 GE to 10 GE as the dominant server attachment technology is well under way. Therefore, we focus on cabling technologies and options for 10 GE in this section.

Several media and transceiver options are available today for 10 Gigabit Ethernet server connectivity. Considerations for use cases depend on variables such as cost, latency, distance, power consumption, and technology availability. See Table 4-1.

Cost tip: One of the largest sources of leveraging network cost is by choosing copper over fiber cabling for 10 Gbps Server connectivity. The cost to connect the server to the network for typical data center structured cabling can cost about $275 for fiber and $150 for copper. This amount will vary by region, but the relative costs should remain about the same. The real savings are with the fiber optics. The list price for short range 10 Gbps optics is about $1500 each and requires two per server connection (one at the server end and one at the switch end.) The list cost of a copper Direct Attach Cable (DAC) is about $400 and no transceivers are needed. Prices will vary based on your procurement methodology, but the business case for 10 Gbps copper server attachment is very compelling. Think about whether servers are connected with copper or fiber for 1 Gbps connections today. The copper or fiber decision will be based largely on the distance. Implications are discussed in the next several sections.

Cisco provides two main 10 Gb Ethernet copper server connectivity options today.

Table 4-1 10 Gbps copper options

Connector (media)	Cable	Distance	Protocol
SFP+ CU copper	Twinax	10m	10GBase-CX1
RJ45 10GBASE-T copper	Cat6, Cat6A, Cat7	100m	10GBase-T

Category 6A (Cat6a) copper cabling was developed in conjunction with the 10GBASE-T standard to achieve reliable 10 Gbps operation over 100 meter copper twisted-pair channels. 10GBASE-T has higher latency and consumes more power than other 10 Gbps Ethernet physical layers. However, in 2008 10GBASE-T silicon became available from several manufacturers with claimed power dissipation of 6W and a latency approaching 1 µs.

10GBASE-CX-1 based direct attach cooper is an alternative copper solution that uses SFP+ direct-attach Twinax cabling. Although this solution has a limited cable distance of up to 10 meters, it provides a robust, power-efficient, and cost-effective solution for 10 Gb Ethernet transmission.

Twinax Cabling is called Direct Attached Copper (DAC) and can be passive or active. Cisco offers passive Twinax cables in lengths of 1, 3, and 5 meters, and active Twinax cables in lengths of 7 and 10 meters.

The SFP+ direct-attach solution is a fully integrated SFP+ cable that is available in multiple lengths up to 10 meters. The SFP+ direct-attach solution draws 1.5 watts (W) of power per port, has a latency of 0.1 µs, and is available today.

Regardless of these variances, Cisco offers a full range of standard-compliant, high performance copper-based 10G solutions. Table 4-2 shows the cabling distances for each type of cable.

Fiber cabling is typically used for longer distance connectivity and environments that need protection from interference. Since it is very reliable and less susceptible to attenuation, it is optimum for sending data beyond 100 meters.

Fiber and copper can be used in the same data center with the choice for each connection based on the distance between devices to be connected.

Table 4-2 Ten gigabyte transceiver types

There are two different types of optical fiber: multi-mode (MM) and single-mode (SM). Cisco supports all cabling options for LAN connectivity in the data center: Device	Range	Optics (wavelength)
10GBASE-SR	300m with MMF	850nm
10GBASE-LR	10km with SMF	1310nm
10GBASE-ER	40km with SMF	1550nm
10GBASE-LRM	220m with MMF	1310nm
10GBASE-LX4	300m MMF/10km SMF	1310nm DWDM

In addition to distance, latency is also a major design point. Table 4-3 lists the latency for each 10 Gbps transceiver type.

Table 4-3 Cabling distances

Technology	Cable	Distance	Power	Transceiver latency
SFP+ CU Copper	Twinax	10m	~0.1W	~0.25 µs
SFP+ SR short reach	MM OM1 MM OM3	33km 300km	1W	~0.1µs
SFP+ LR Long Reach	SMF	10km	1W	~0.1µs
10GBASE-T -65nm	Cat6/6a/7 Cat6/6a/7	100m 30m	~6,2W ~4.5W	~3µs ~3µs
10GBASE-T -40nm	Cat6/6a/7 Cat6/6a/7	100m 30m	~3.9W ~3.3W	~3µs ~3µs

Twinax cables are suitable for very short distances and offer a highly cost-effective way to connect switches and servers within racks and across adjacent racks. Cisco offers passive Twinax cables for distances less than 5 meters and active Twinax cables for 7 and 10 meters. Make sure the NIC is compatible with the vendor cabling.

To enable the use of twinax cables, the servers need to be close to the access switch. The PoD design described next in 4.4, "End-of-Row (EoR), Middle-of-Row (MoR), and Top-of-Rack (ToR)" explains several techniques for designing the access switch near the server.

4.4 End-of-Row (EoR), Middle-of-Row (MoR), and Top-of-Rack (ToR)

In recent years, network data center access switches have been shifting to small 1 RU and 2 RU appliances. Placing 1 RU and 2 RU switches in the same rack as the servers can greatly reduce the cabling requirements for servers. When server access switches are connected to servers in the same row, multiple racks can be grouped into a server access block or PoD. See 3.1.1, "The PoD concept" on page 134.

The three variations for building physical server pods are as follows:

- Top-of-Rack (ToR) placement means physically placing the network access switch in the top of a server rack and cabling the servers directly to the switch. This requires one or two switches per server rack. Top-of-Rack (ToR) or Bottom-of Rack (BoR) installations facilitate the server/network connectivity due to the fact, that the distance between the server and the network is limited. This distance limit enables, where supported, a better usage of copper interfaces and cabling, which is cheaper than fiber connectivity.

  Furthermore, only a limited number of fiber cabling is required between the server racks and the aggregation/core network racks. The physical switches that provide both 10G and 1G interfaces are placed in either a Top-of-Rack (ToR) or Bottom-of-Rack (BoR) configuration. The exact configuration is dependent on the cable management. If the inter rack cabling is preferred via the raised floor, BoR is the preferred option, else ToR should be used.

Figure 4-8 shows Top-of-Rack cabling. The access switch could also be Middle-of-Rack or Bottom-of-Rack.

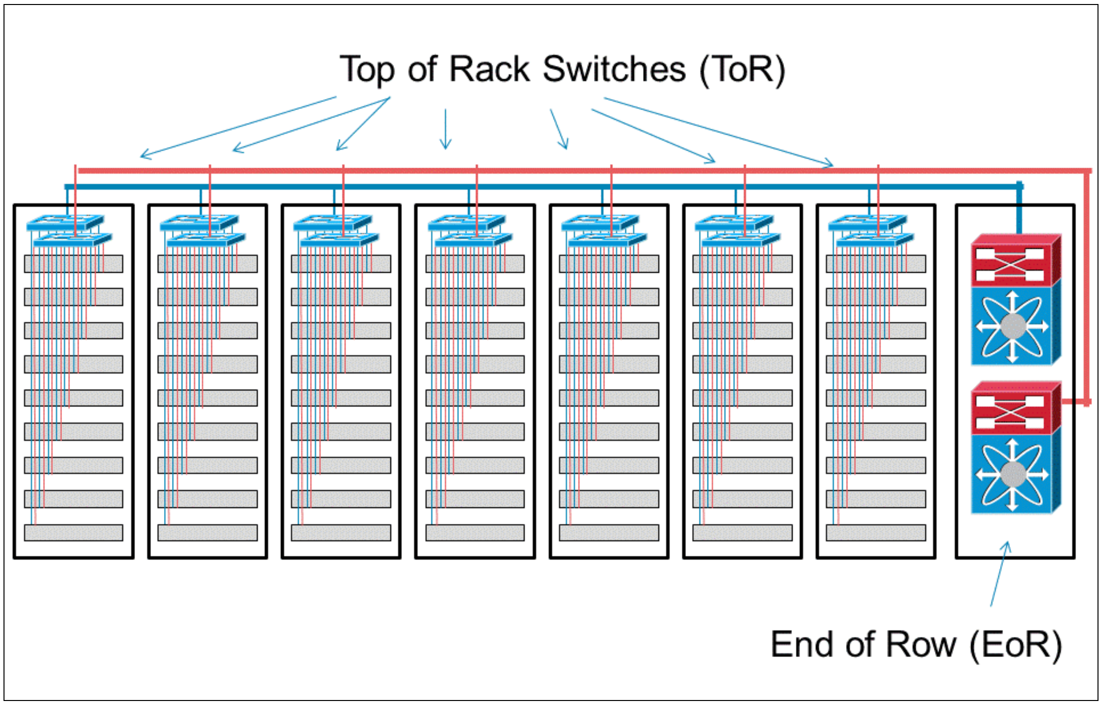


Figure 4-8 Top-of-Rack / End-of-Row cabling

The traditional ToR model has the advantages of an efficient, low cost cabling scheme and in most cases, lower equipment cost than EoR models using modular switching platforms. Disadvantages of the traditional ToR model include a proliferation of managed devices and reduced system availability due to a less robust switching platform hardware.

- End-of-Row (EoR) means placing the server access switches at the end of a row of server racks. End-of-Row placement allows the use of a chassis switch or small RU access switches.
- Middle-of-Row (MoR) means placing the server access switches in the middle of a row of servers rather than at the end of a row. This allows the network switches to be equal distance from the furthest racks.

4.4.1 Why use Top-of-Rack architecture?

Rapidly changing business requirements impose a corresponding need for flexibility and mobility in data centers. Because of the significant cost of building a new data center, designing an infrastructure that provides the flexibility to meet business objectives while increasing ROI is an IT imperative. By building the data center infrastructure (power and cooling, cabling, and so on) in a modular fashion, data center flexibility can be increased, which in turn improves business agility.

Many organizations are now deploying modular data centers. IT departments are increasingly deploying not just servers but racks of servers at a time. Racks of servers, blade systems, and integrated rack-and-blade systems are often purchased in pre-configured racks with power, network, and storage cabling pre-installed so that racks can be commissioned within hours, not days, from the time they arrive on the loading dock. While server form factors are evolving, and some racks can host up to 96 independent computer resources, the rack form factor remains constant, making it the standard unit of deployment in many data centers.

ToR solutions complement rack-at-a-time deployment by simplifying and shortening cable runs and facilitating the replication of rack configurations. This rack-and-roll deployment model offers a solution by placing switching resources in each rack so that server connectivity can be aggregated and interconnected with the rest of the data center through a small number of cables connected to End-of-Row (EoR) access- or aggregation-layer switches. In reality the switch can be placed in the middle of the rack or the bottom rack as well. The switch placement depends on the type of data center construction and policies for the data center.

In order to understand the benefits of a End-of-Row (EoR), Middle-of-Row (MoR), and Top-of-Rack (ToR) design, it is helpful to elaborate on the definition of a server access block or Server Access PoD. The term Server Access PoD and Server Access Block are used inter-changeably. From the network perspective, a Server Access PoD is a series of server racks connected to the same access/aggregation switches. A PoD can be comprised of 1-x racks in a row (where x is typically 5-8).

A server access PoD should be built to minimize cabling that leaves the PoD. Ideally, all copper cabling will stay within the PoD. Only the fiber uplinks from network devices to its upstream device should leave the PoD.

A legacy design would have copper run from the servers to the centrally located access switches. A Catalyst 6509s would typically have 5-7x 48 port line cards. It would not be uncommon for a large server farm to have 2500 copper 1G connections. Figure 4-7 on page 199 shows the results of running several hundred cat 6E cables under the floor. These cables are expensive are and cause an air dam under the floor.

By utilizing server PoDs, the copper cables can be kept within the PoD. Depending on the construction of the data center and the data centers cabling policy, patch cables or structured cabling can be used within the PoD. Within a rack patch cables are used. Rack-to-rack (also referred to as cross-rack) cabling is kept within a row.

The ToR cabling and network architecture optimizes the requirement for horizontal cabling from the server rack by placing the ToR aggregation device at the top of the server rack. The actual placement of the ToR device may vary based on customer requirements (for example, in or above the server rack or ToR aggregation per two or three server racks) and seeks to optimize customer requirements for density, cabling, and design methodology.

Figure 4-9 depicts a three rack PoD utilizing Top-of-Rack switches. The connection of the ToR devices to the aggregation/core switches will utilize the HDA cabling (based on the specific cabling model deployed). The ToR cabling design model follows a logical network layer construct in which the server network connectivity is aggregated at the rack level in the network access layer. The access layer is in turn connected to the network aggregation layer.

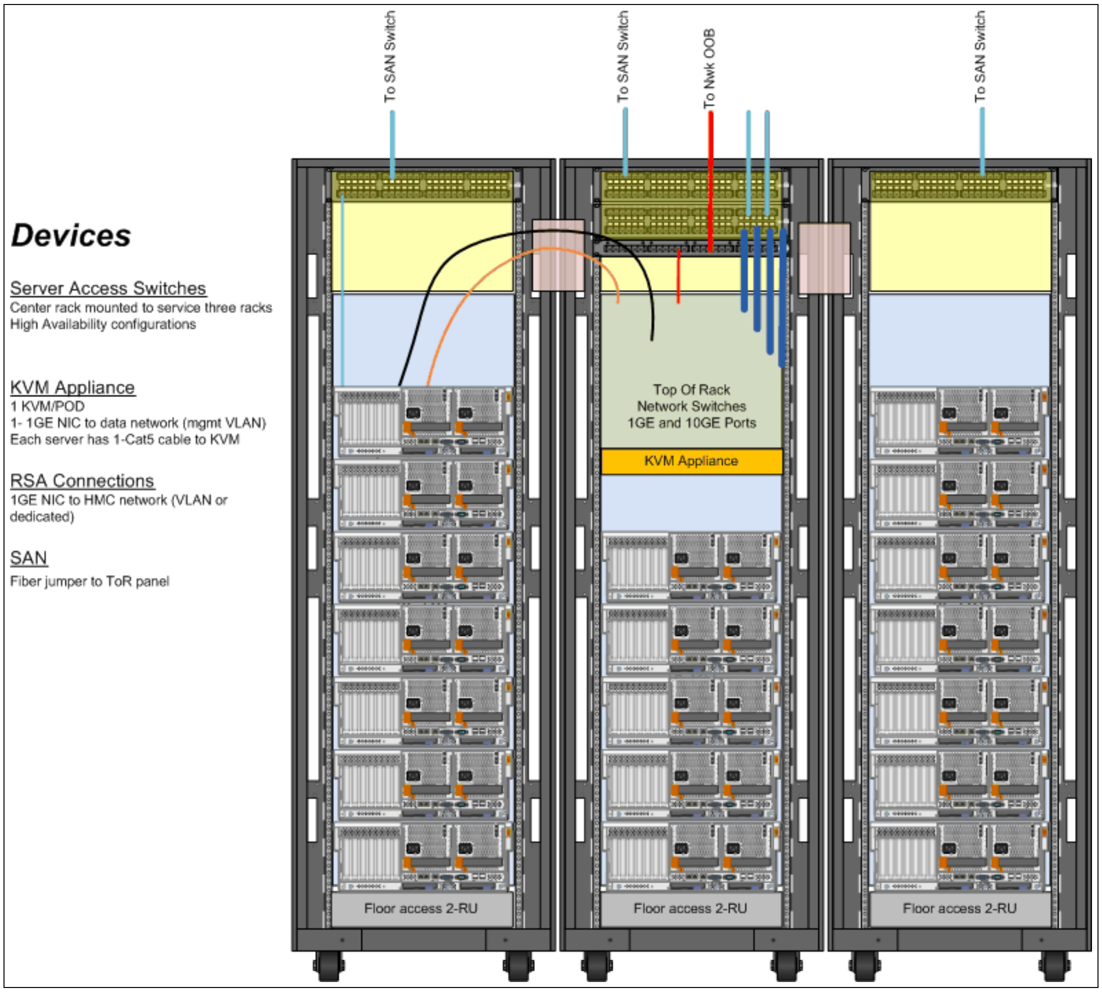


Figure 4-9 Three Rack xSeries PoD

The following benefits are available from using a physical server PoD:

- Eliminating under-floor custom cabling that spans the data center saves cost and time of cabling.
- Having co-located network and server equipment saves cost of network floor space and of network cabinets.
- Limiting the under-floor cabling improves airflow under raised floors.
- Reducing custom cabling reduces the amount of time to build a network.
- A PoD design is re-usable.
- It enables 10G copper server connections:
  - Typically 7 meters

ToR switching allows oversubscription to be handled at the rack level, with a small number of fiber cables providing uniform connectivity to each rack. The advantage of this solution is that horizontal fiber can support different I/O connectivity options, including Gigabit Ethernet and 10 Gigabit Ethernet as well as Fibre Channel. The use of fiber from each rack also helps protect infrastructure investments as evolving standards, including 40 and 100 Gigabit Ethernet, are more likely to be implemented using fiber before any other transmission mechanism.

By limiting the use of copper to within racks, the ToR model isolates the cabling that changes most often, to the parts of the data center that change most frequently, the racks themselves. The use of fiber running from racks provides a flexible data center cabling infrastructure that supports the transition from Gigabit Ethernet to 10 Gigabit Ethernet now, while enabling transition to 40 and 100 Gigabit Ethernet in the future.

4.5 Fabric Extender (FEX) solutions for the access layer

Utilizing FEX technology can provide an efficient and flexible alternative to the traditional models. Using Cisco Nexus 2000 Fabric Extenders (FEX) technology, you can design a ToR access layer model that also incorporates the advantages of the traditional EoR/MoR models. (See Chapter 2, "IBM and Cisco building blocks for the data center" on page 29 for a description of the Nexus family of products). This is accomplished by placing the Nexus 2000 in the same rack or in an adjacent rack to each server (minimizing cable costs), and managing the FEX from the parent switch (Nexus 5000 or Nexus 7000).

With the ToR deployment model supported by the Cisco Nexus 2000 Series Fabric Extenders, all cabling within racks can use inexpensive copper cabling. For 100-Mbps and Gigabit Ethernet, standard patch panels with RJ-45 connectors can be installed. For 10 Gigabit Ethernet, CX-1 cabling with built-in transceivers is an extremely cost-effective and low-latency alternative to fiber within racks. Figure 4-10 depicts physical access layer design with Top-of-Rack Fabric Extenders.

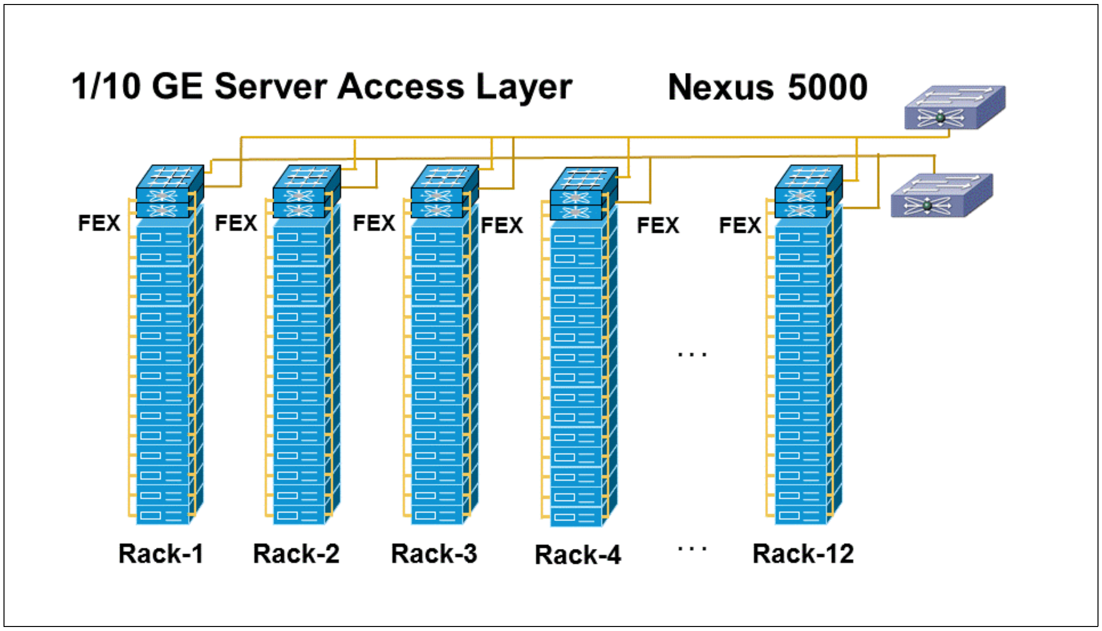


Figure 4-10 Physical access layer design with Top-of-Rack Fabric Extenders

For cabling between fabric extenders and Cisco Nexus 5000 Series switches, fiber is one of the best long-term investments. It is the most flexible transmission medium, and advancements in Ethernet technology tend to be available first on fiber. Today, fiber can be used to support 1/10 Gigabit Ethernet and 1/2/4/8-Gbps Fibre Channel. In addition, 40G and 100G Ethernet standards are now supported on fiber. Fiber in overhead cable trays running to each rack position supports a "rack- and- roll" model, in which any rack with Top-of-Rack switching can be rolled into any position. For PoD-based deployments, using fabric extenders in every rack and locating the parent switches at the middle of the row, CX-1 cabling offers a lower-cost alternative to fiber.

Forward-looking IT departments are preparing their data centers for the future by integrating support for 10 Gigabit Ethernet and a unified network fabric into their switching and cabling strategies. Since the typical data center lifecycle is 10 to 15 years, cabling architectures have a tremendous effect on the data center's ability to adapt to network architecture changes, evolving bandwidth needs, and technology moves, additions, and changes. Cabling architectures, if not chosen correctly, could force an early replacement of the cabling infrastructure to meet connectivity requirements as the network and computer technologies evolve.

Today's data centers deploy a variety of cabling models and architectures. With the migration from Gigabit Ethernet to 10 Gigabit Ethernet, the cabling and network switching architectures are being re-evaluated to help ensure a cost-effective and smooth data center transition. The choice of cabling architecture will affect throughput, expandability, sustainability, optimum density, energy management, total cost of ownership (TCO) and return on investment (ROI). Anticipating growth and technological changes can be difficult, but the data center should be able to respond to growth and changes in equipment, standards, and demands while remaining manageable and reliable.

This section examines the use of the Top-of-Rack (ToR) cabling and Fabric Extender (FEX) Architecture for next-generation data center access layer infrastructure. It explores current 10 Gigabit Ethernet cabling choices and provides a solution architecture based on FEX to address architectural challenges. Data center managers and facilities administrators will choose cabling architectures based on various factors. The FEX model offers a clear access-layer migration path to an optimized high-bandwidth network and cabling facilities architecture that features low capital and operating expenses and supports a “rack-and-roll” computer deployment model that increases business agility. The data center's access layer, or equipment distribution area (EDA), presents the biggest challenge to managers as they choose a cabling architecture to support data center computer connectivity needs. The FEX architecture and cabling model proposes the use of fiber as the backbone cabling to the rack, with copper and fiber media for server connectivity at the rack level.

IT departments building new data centers, expanding existing data center footprints, or updating racks of equipment all have to design a cabling and switching architecture that supports rapid change and mobility and accommodate transitions to 10, 40, and 100 Gigabit Ethernet over time.

The main factors that IT departments must address are as follows:

- Modularity and flexibility are of paramount importance: The need to rapidly deploy new applications and easily scale existing ones has caused server-at-a-time deployment to give way to a rack-at-a-time model. Many IT departments are ordering preconfigured racks of equipment with integrated cabling and switching and as many as 96 servers per rack. The time required to commission new racks and decommission old ones is now a matter of hours rather than days or weeks. Because different racks have different I/O requirements, data center switching and cabling strategies must support a wide variety of connectivity requirements at any rack position.
- Bandwidth requirements are increasing: Today's powerful multi-socket, multicore servers, blade systems, and integrated server and rack systems, often running virtualization software, are running at higher utilization levels and impose higher bandwidth demands. Some server racks are populated with servers requiring between five and seven Gigabit Ethernet connections and two Fibre Channel SAN connections each.
- I/O connectivity options are evolving: I/O connectivity options are evolving to accommodate the need for increasing bandwidth, and good data center switching and cabling strategies need to accommodate all connectivity requirements at any rack position. Racks today can be equipped with Gigabit Ethernet or 10 Gigabit Ethernet or a unified network fabric with Fibre Channel over Ethernet (FCoE).

- Virtualization is being added at every layer of the data center: Server virtualization is promoting server consolidation and increasing the need for bandwidth and access to network-attached storage (NAS). Virtualization is one of the main areas of focus for IT decision makers. Estimates suggest that the server virtualization market will grow by 44 percent over the next 4 years. The change to virtualization can be disruptive and necessitate a redesign of the networking infrastructure to gain all the benefits of the virtualized computer platform.

The challenge facing data centers today is how to support the modularity, reduce number of management points, and increase flexibility that is needed to promote business agility and maintain a company's competitive edge. The same strategy that allows the intermixing of different rack types and I/O requirements must also support a varied set of connectivity options including Gigabit Ethernet and 10 Gigabit Ethernet as well as a unified network fabric.

FEX is based on the concept of a remote I/O module in which fabric extenders are physically separate devices that are logically part of the parent switches to which they are connected. All management, configuration, troubleshooting, and firmware upgrade is handled by the parent switch. Fabric Extender (FEX) enable the ports on an Ethernet switch to be physically distributed across many racks within a data center but logically the switching remains within a single physical switch. The key concept here is that FEX is not a switch in itself but rather a “satellite” that connects back to a parent switch, which makes all the forwarding decisions on its behalf. As such, the capabilities on FEX are a function of both the hardware of FEX itself (such as the port speeds offered) and the forwarding features offered on the platform of the parent switch.

A FEX architecture allows oversubscription to be handled at the rack level, with a small number of fiber cables providing uniform connectivity to each rack. The advantage of this solution is that horizontal fiber can support different I/O connectivity options, including 100 Megabit Ethernet, Gigabit Ethernet, 10 Gigabit Ethernet and FCoE. The use of fiber from each rack also helps protect infrastructure investments as evolving standards, including 40 and 100 Gigabit Ethernet, are more likely to be implemented using fiber before any other transmission mechanism. By limiting the use of copper to only within the PoD racks, the ToR model isolates the cabling that changes most often to the parts of the data center that change most frequently: the racks themselves. The use of fiber runs from racks provides a flexible data center cabling infrastructure that supports the transition from Gigabit Ethernet to 10 Gigabit Ethernet now, while enabling transition to 40 and 100 Gigabit Ethernet in the future.

The FEX model enables data center managers to implement a single cabling model that can support, 100 Megabit Ethernet, Gigabit Ethernet and 10 Gigabit Ethernet and unified network fabric today, while supporting future 40 and 100 Gigabit Ethernet standards as they come to market. Using cable trays for only fiber-optic cable management, data center managers have the flexibility to deploy preconfigured racks with different connectivity requirements in any rack position. For example, a rack of servers running multiple Gigabit Ethernet connections can be placed next to a rack of servers with 10 Gigabit Ethernet and FCoE connections to each server.

This distributed architecture enables physical topologies with the flexibility and benefits of both Top-of-Rack (ToR) and End-of-Row (EoR) deployments. The FEX architecture will reduce the total cost of ownership (TCO) by providing the following benefits:

Architecture flexibility:

- Unified server access architecture: Offers a highly cost-effective access-layer architecture for 100 Megabit Ethernet, Gigabit Ethernet, 10 Gigabit Ethernet, mixed Gigabit Ethernet and 10 Gigabit Ethernet servers, physical or virtual server, and rack or blade server environments.

- Flexible physical topologies: Allow decoupling of the Layer 1 and 2 topologies, therefore providing flexibility in designing physical architectures, including ToR, MoR, and EoR deployments, while allowing a quick expansion of network capacity and remote line-card portability across multiple parent switches. It is also space optimized for all these architectures. Figure 4-11 shows a flexible ToR Design with MoR/EoR.

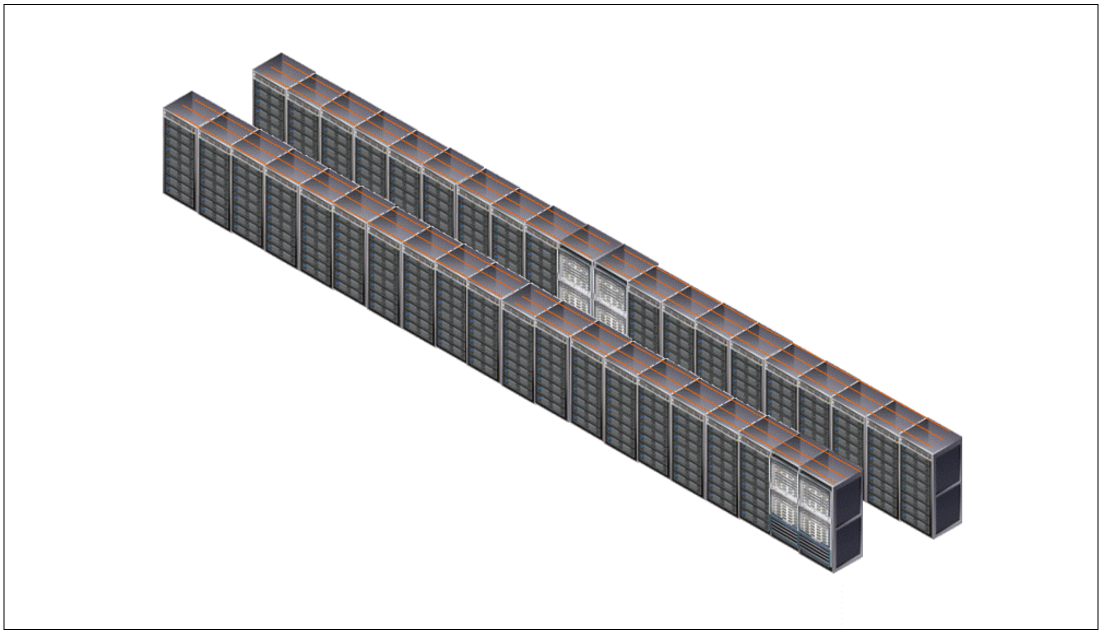


Figure 4-11 ToR and MoR/EOR Design

Highly scalable server access:

- Massive scalability: A deployment of Cisco Nexus 2000 Series Fabric Extenders connected to a Cisco Nexus 5000 or 7000 Series Switch supports highly scalable Gigabit and 10 Gigabit Ethernet environments as defined later in this section.
- Layer 2 scalability: Reliance on spanning tree protocol is eliminated between the fabric extender and the parent switch, thus enabling a large, multipath, loop-free topology. Use of a single management entity to support a large server domain allows policy to be enforced more efficiently and enhances Layer 2 data center access scalability. The use of the Virtual PortChannel (vPC) feature also allows fast convergence and effective utilization of bandwidth in Layer 2 environments.

Simplified operations:

- Single point of management: All device configurations are managed on the Cisco Nexus parent switch, and configuration information is downloaded to the Cisco Nexus 2000 Series Fabric Extender using in-band communication.
- Software maintenance simplification: The fabric extender is a plug-and-play device that automatically downloads the software image from the Cisco Nexus parent switch in the same way that a line card downloads software from the supervisor engine in a modular chassis. In-Service Software Upgrade (ISSU) on the fabric extenders provides the capability to perform transparent software upgrades, reducing downtime and allowing customers to integrate the newest features and functions with little or no effect on network operation for Ethernet, storage, and converged network environments.

- Switch feature consistency across a large number of servers: The Cisco Nexus 2000 Series forwards all traffic to the parent Cisco Nexus switch over 10 Gigabit Ethernet fabric uplinks. Passing all traffic to the parent switch allows traffic to be shaped according to policies established on the parent Cisco Nexus switch with a single point of management. Standardizing on the Cisco Nexus switches allows data centers to support the same switch features across the entire access layer with a single point of management.
- Tenfold management point reduction: The number of management points is significantly less than when discrete switches are used at the top of the rack. A traditional 12-rack design using a discrete, redundant pair of Gigabit Ethernet switches at the top of each rack has 24 management points. The equivalent architecture using the Cisco Nexus 2000 Series has only 2 management points: a tenfold reduction in management complexity.

Figure 4-12 identifies the terminology to understand FEX components that represent the architecture.

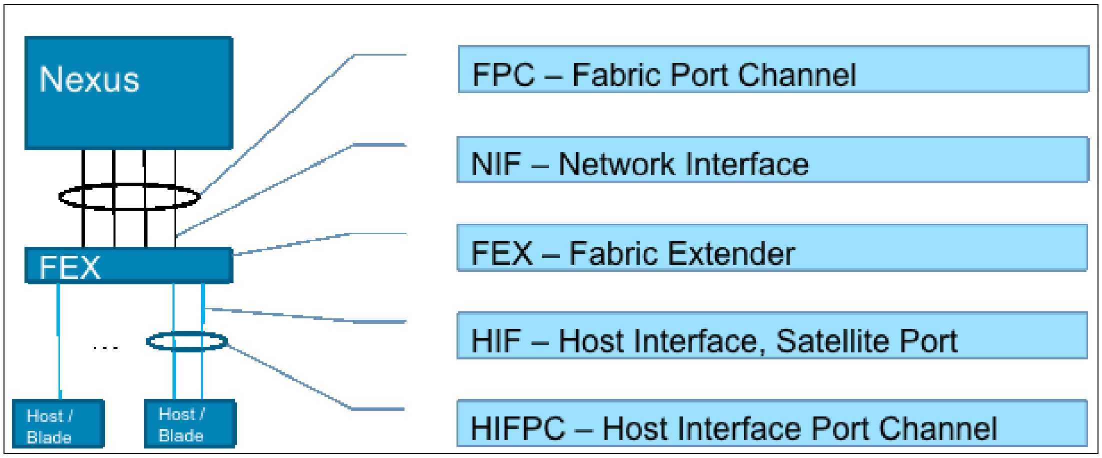

Nexus
FEX
Host / Blade
Host / Blade
FPC – Fabric Port Channel
NIF – Network Interface
FEX – Fabric Extender
HIF – Host Interface, Satellite Port
HIFPC – Host Interface Port Channel

Figure 4-12 Cisco FEX Terminology

- FEX: Fabric Extender: NEXUS 2000 is the instantiation of FEX.
- Fabric Port: Nexus side of the link connected to Fabric Extender
- Fabric Port Channel: PortChannel between Nexus and Fabric Extender
- FEX Uplink: Network facing port on the Fabric Extender (FEX side of the link connected to Nexus Parent); FEX uplink is also referenced as NIF: Network Interface
- FEX Port: Server facing port on the Fabric Extender, also referred to as server port or host port in this presentation; FEX port is also referenced as HIF: Host Interface
- Nexus: The parent switch in the Fabric Extender architecture

The combination of Cisco Nexus 5000, Nexus 6000, or Nexus 7000 Series switches and Cisco Nexus 2000 Series Fabric Extenders delivers a unified, consolidated, virtualized, and simplified access layer that addresses the challenges facing data center operators today. The product family is based on the concept of a remote I/O module in which fabric extenders are physically separate devices that are logically part of the parent switches to which they are connected. All management, configuration, troubleshooting, and firmware upgrade is handled by the parent switch.

The Cisco Nexus 2000 Series design aligns with that of Server Access PoDs. It offers front-to-back cooling, compatible with data center hot-aisle and cold-aisle designs, all switch ports at the rear of the unit in close proximity to server ports, and all user-serviceable components accessible from the front panel. It also offers back-to-front cooling, with switch ports in front of the chassis, aligned with the cold aisle, for optimized cabling in network racks. The Cisco Nexus 2000 Series is built for nonstop operation, with redundant hot-swappable power supplies and a hot-swappable fan tray with redundant fans. Its compact one-rack-unit (1RU) form factor takes up relatively little space, making it easy to incorporate into rack designs.

More information on the Nexus 2000 can be found here:

http://www.cisco.com/en/US/products/ps10110/index.html

4.5.1 FEX solutions

With a unified access layer based on Cisco Nexus 5000/6000/7000 Series switches, server racks preconfigured with fabric extenders can be rolled into position and their uplinks connected to integrate them into the data center access layer. Because all fabric extender firmware and configuration is managed by the parent switch, fabric extenders can be integrated into the access layer simply by connecting their uplinks. The simple way in which the access layer can be expanded makes rapid deployment a reality: servers can be installed into racks along with fabric extenders, rolled into position, and integrated into the access layer just by connecting the uplinks. Both Ethernet and Fibre Channel connectivity can be handled with one set of uplinks on 10-Gbps fabric extenders with unified fabric support. Within each rack, 100-Mbps Ethernet connections can be made to management consoles, Gigabit Ethernet connections can be made to older rack-mount and blade systems, and 10 Gigabit Ethernet and FCoE connections can be made to state-of-the-art systems

The Cisco Nexus Series switches and all 10 Gigabit fabric extenders support a unified fabric that brings together Ethernet and FCoE with standards-based management extensions to make the fabric even more reliable and manageable. Advanced, standards-based extensions support the lossless requirements of Fibre Channel traffic on the same physical link as standard IP traffic. When the unified fabric is accessed by servers equipped with converged network adapters (CNAs), data centers can deploy a unified fabric in a way completely transparent to the operating system and its drivers, allowing IT departments to continue using their traditional server, storage, and networking administration roles without disruption.

Fabric extenders can be deployed to meet a range of availability requirements using two different models, both of which avoid the complexity of introducing spanning tree protocol into the access layer:

- Virtual PortChannels: Support for VPC technology allows fabric extenders to connect to two different Cisco Nexus 5000 or 6000 Series switches in an efficient active-active configuration.
- Dual homing: The risk that a single fabric extender will fail can also be mitigated by dual homing each server to two Cisco Nexus Series switches through two fabric extenders using vPC technology.

The Cisco Nexus 2000 Series can be used in conjunction with a Cisco Nexus parent switch in these main design scenarios:

- Cisco Nexus 2000 Series Fabric Extenders single-connected to one upstream Cisco Nexus 5000, 6000, or 7000: This is the simplest deployment scenario.
- Series Switch: In this deployment scenario, access-layer redundancy is achieved through redundant server connections to two upstream distributed modular systems, using vPC (Cisco Nexus 5000 Series) or server NIC teaming to two Cisco Nexus 2000 Series Fabric Extenders.
- Cisco Nexus 2000 Series Fabric Extenders dual-connected to two upstream Cisco Nexus 5000 and Nexus 6000 Series switches (vPC): In this deployment scenario, access-layer redundancy is achieved through a combination of Cisco Nexus 2000 Series Fabric Extenders dual-connected to an upstream parent switch and server NIC teaming. VPC from the N2k and EVPC (enhanced VPC) will be supported on the Nexus 7k in a future software release.

Figure 4-13 depicts the Cisco FEX Distributed Modular Design.

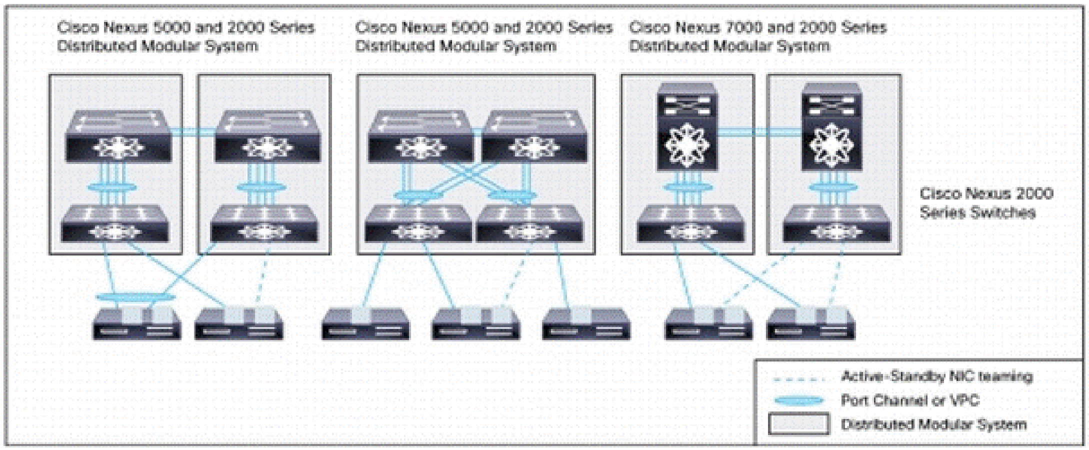
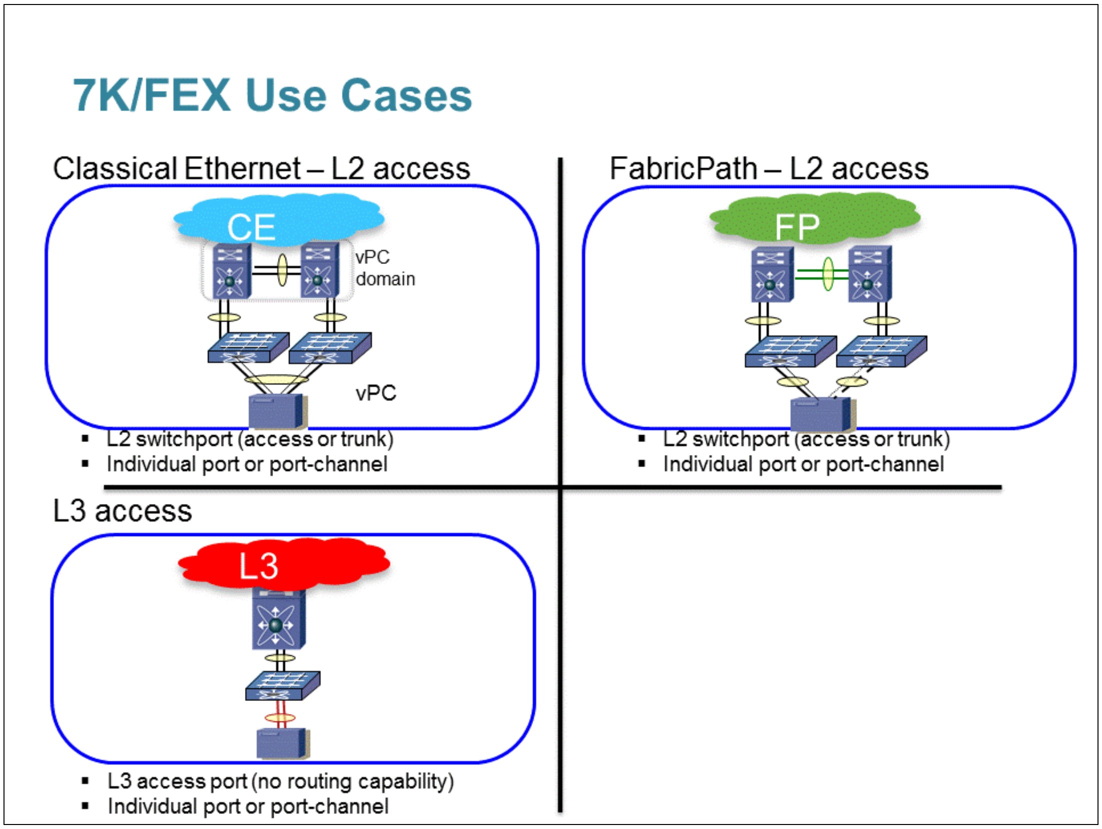

Cisco Nexus 5000 and 2000 Series Distributed Modular System
Cisco Nexus 5000 and 2000 Series Distributed Modular System
Cisco Nexus 7000 and 2000 Series Distributed Modular System
Cisco Nexus 2000 Series Switches
Active-Standby NIC teaming
Port Channel or VPC
Distributed Modular System

Figure 4-13 Cisco FEX Distributed Modular Design

4.5.2 FEX solutions with Nexus 7000

The Cisco Nexus Fabric Extender (FEX) and Nexus 7000 provide a Flexible Access and Aggregation Solution by de-coupling of the Layer 1 and Layer 2 topologies.

You can extend the Cisco Nexus 7000 Series switch architecture by connecting up to 48 FEXs as remote I/O modules with homogenous policies across 2000 1/10G/FCoE Ports with N7k Supervisor 2E. Depending on which FEX model that you connect to the switch, the FEX provides top-of-the-rack connectivity and it becomes an extension of the parent Cisco Nexus 7000 Series switch fabric, with the FEX and the switch becoming a virtual modular system. The FEX forwards all 100/1000 Gbps Ethernet or 10 Gbps Ethernet traffic from the hosts to the switch over 10 Gbps uplinks. Figure 4-14 shows Cisco 7K supporting Fabric Extenders.



7K/FEX Use Cases

Classical Ethernet – L2 access

- L2 switchport (access or trunk)
- Individual port or port-channel

FabricPath – L2 access

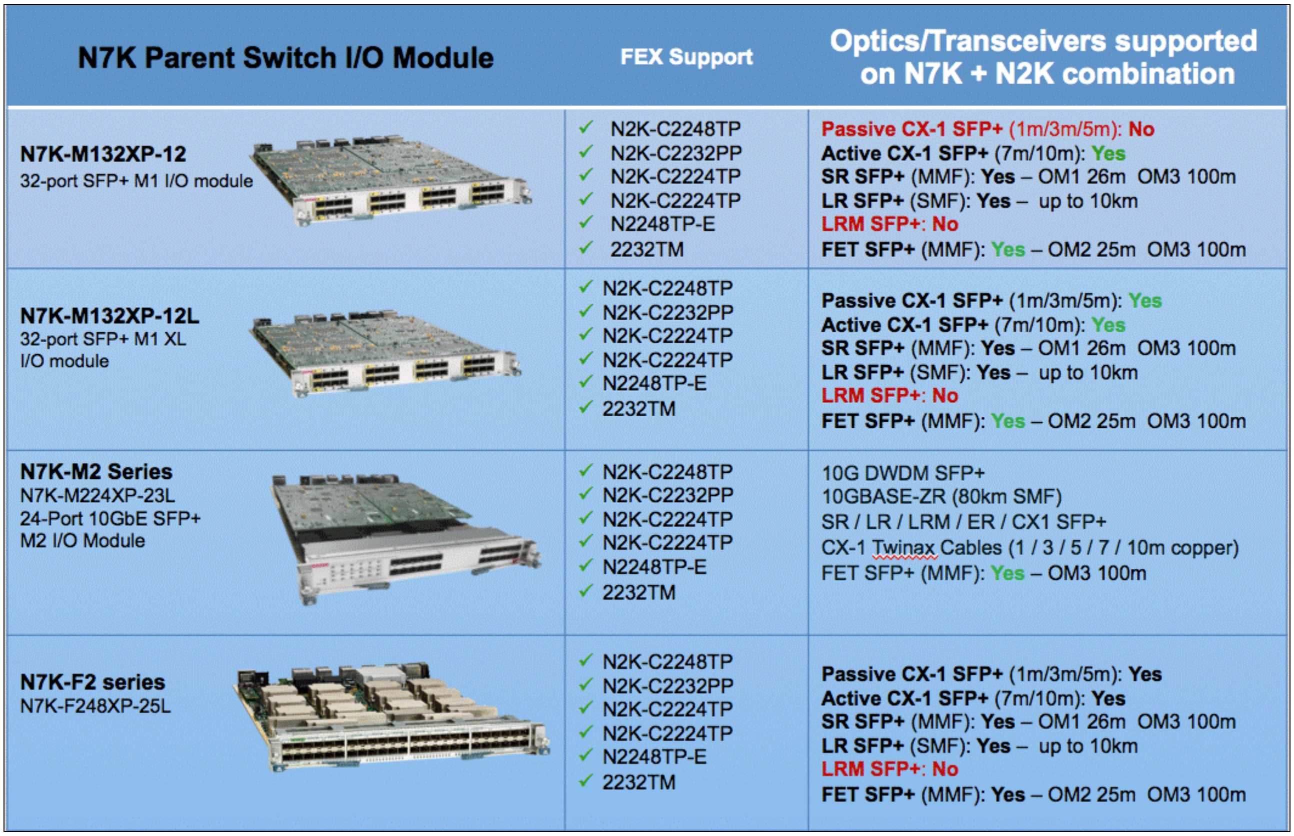
- L2 switchport (access or trunk)
- Individual port or port-channel

L3 access

- L3 access port (no routing capability)
- Individual port or port-channel

Figure 4-14   Nexus 7K with FEX

You connect a FEX uplink port to the Cisco Nexus 7000 Series switch through one of the Ethernet I/O modules as Nexus 7K I/O modules shown in Figure 4-15.



N7K Parent Switch I/O Module	FEX Support	Optics/Transceivers supported on N7K + N2K combination
N7K-M132XP-12 32-port SFP+ M1 I/O module	✓ N2K-C2248TP ✓ N2K-C2232PP ✓ N2K-C2224TP ✓ N2K-C2224TP ✓ N2248TP-E ✓ 2232TM	**Passive CX-1 SFP+** (1m/3m/5m): **No** **Active CX-1 SFP+** (7m/10m): **Yes** **SR SFP+** (MMF): **Yes** – OM1 26m OM3 100m **LR SFP+** (SMF): **Yes** – up to 10km **LRM SFP+: No** **FET SFP+** (MMF): **Yes** – OM2 25m OM3 100m
N7K-M132XP-12L 32-port SFP+ M1 XL I/O module	✓ N2K-C2248TP ✓ N2K-C2232PP ✓ N2K-C2224TP ✓ N2K-C2224TP ✓ N2248TP-E ✓ 2232TM	**Passive CX-1 SFP+** (1m/3m/5m): **Yes** **Active CX-1 SFP+** (7m/10m): **Yes** **SR SFP+** (MMF): **Yes** – OM1 26m OM3 100m **LR SFP+** (SMF): **Yes** – up to 10km **LRM SFP+: No** **FET SFP+** (MMF): **Yes** – OM2 25m OM3 100m
N7K-M2 Series N7K-M224XP-23L 24-Port 10GbE SFP+ M2 I/O Module	✓ N2K-C2248TP ✓ N2K-C2232PP ✓ N2K-C2224TP ✓ N2K-C2224TP ✓ N2248TP-E ✓ 2232TM	10G DWDM SFP+ 10GBASE-ZR (80km SMF) SR / LR / LRM / ER / CX1 SFP+ CX-1 Twinax Cables (1 / 3 / 5 / 7 / 10m copper) FET SFP+ (MMF): **Yes** – OM3 100m
N7K-F2 series N7K-F248XP-25L	✓ N2K-C2248TP ✓ N2K-C2232PP ✓ N2K-C2224TP ✓ N2K-C2224TP ✓ N2248TP-E ✓ 2232TM	**Passive CX-1 SFP+** (1m/3m/5m): **Yes** **Active CX-1 SFP+** (7m/10m): **Yes** **SR SFP+** (MMF): **Yes** – OM1 26m OM3 100m **LR SFP+** (SMF): **Yes** – up to 10km **LRM SFP+: No** **FET SFP+** (MMF): **Yes** – OM2 25m OM3 100m

Figure 4-15 Cisco Nexus 7K options

4.5.3 FEX solutions with Nexus 5000

Cisco Nexus 5000 Series switches and Cisco Nexus 2000 Series Fabric Extenders represent a virtual modular system that delivers the benefits of both fixed and modular systems (see Figure 4-16). The Cisco Nexus 2000 Series expands the management domain and logical port count of Cisco Nexus 5000 Series switches to up to 1152 Gigabit Ethernet and10 Gigabit Ethernet ports without increasing access-layer complexity.

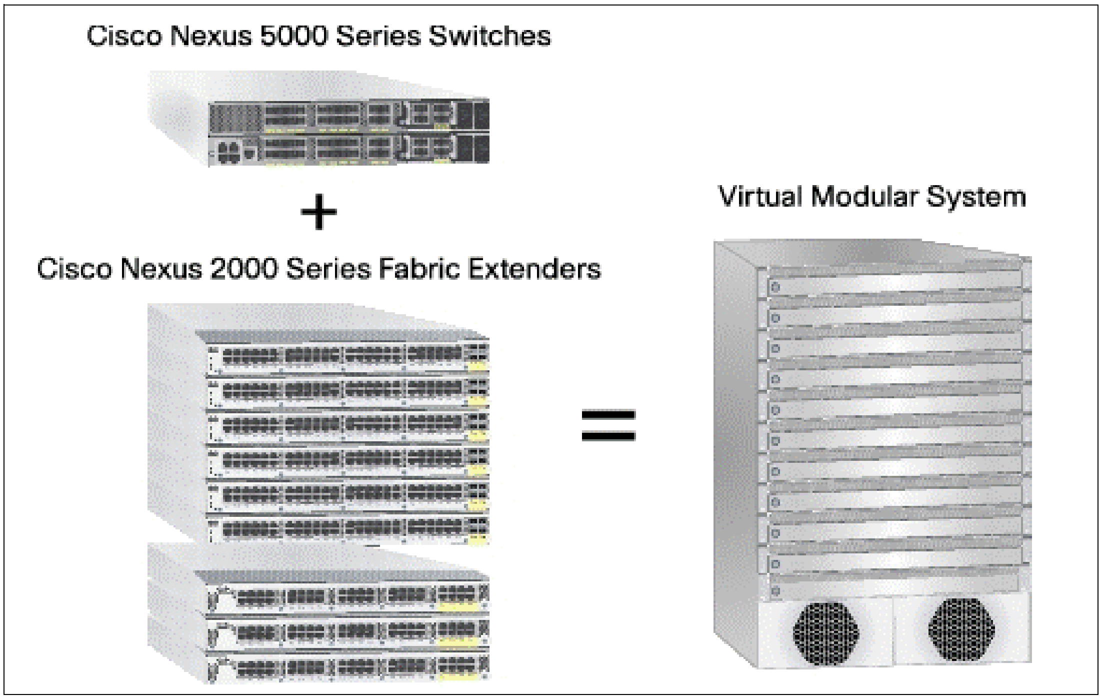

Cisco Nexus 5000 Series Switches
+
Cisco Nexus 2000 Series Fabric Extenders
=
Virtual Modular System

Figure 4-16 Cisco Nexus 5K with Fabric Extenders

From the Cisco NX-OS 5.1(3)N1(1) release and later releases, each Cisco Nexus 5500 Series device can manage and support up to 24 FEXs without Layer 3. With Layer 3, the number of FEXs supported per Cisco Nexus 5500 Series device is 16. With Enhanced vPC and a dual-homed FEX topology, each FEX is managed by both Cisco Nexus 5000 Series devices. As a result, one pair of Cisco Nexus 5500 Series devices can support up to 24 FEXs and 16 FEXs for Layer 2 and Layer 3. There are differences in scalability between the straight-through topology, the dual-homed FEX topology, and the Enhanced vPC topology. In the straight through topology, only one Cisco Nexus 5000 Series device manages each FEX and a pair of Cisco Nexus 5500 Series devices manage up to 48 FEXs. This difference is shown in Figure 4-17 and Table 4-4 for a Layer 2 scenario.

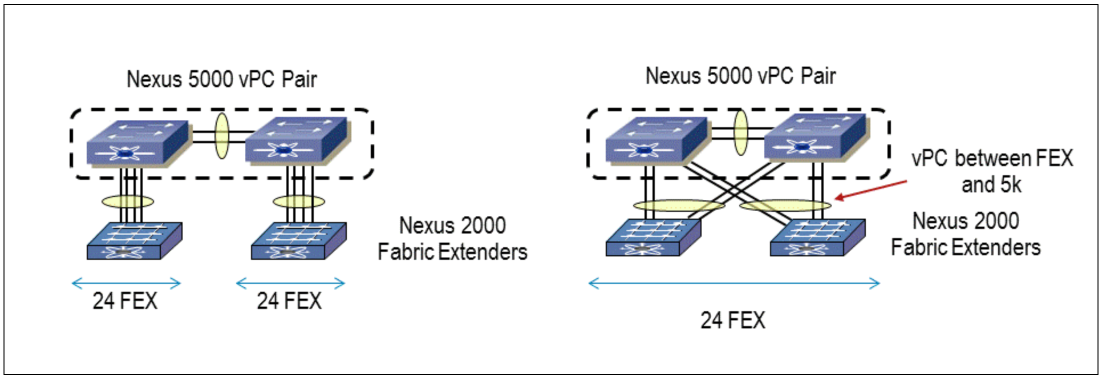


Figure 4-17 Nexus 5k scalability

Table 4-4 FEX scalability matrix for Nexus 5000 and Nexus 7000

	N5K/N2K scale	N7K/N2K scale
# of FEX	Up to 24 with Nexus 5000	48 (SUP2E)
Server Scale	1152 1GE, 768 10GE	2048 1GE, 1536 10GE
# of VLANs	4K	4K with 6.2 release
MAC table	32k with Nexus 5000	128K (M1) 16K (F2)
MST	64	90K
vPC	576	528
ACLs	2K Ingress/ 1K Egress	16K per SOC on F2, 64K-128K on M2
ECMP	16-WAY	16-way, 32-way with F2
Multicast routes	2K	32K with M2
Prefixes	8K	1M IPv4 (M1, M) 32K (F2)
Host entries	8K	1M IPv4 (M1, M) 32K (F2)

For the latest scale updates, see these Scalability Guides:

http://www.cisco.com/en/US/docs/switches/datacenter/nexus5000/sw/configuration_limits/limits_513/nexus_5000_config_limits_513.html

http://www.cisco.com/en/US/docs/switches/datacenter/sw/verified_scalability/b_Cisco_Nexus_7000_Series_NX-OS_Verified_Scalability_Guide.html

The white paper, “Data Center Top-of-Rack Architecture Design,” is available at this website:

http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9670/white_paper_c11-522337.html

4.5.4 FEX solutions with Nexus 6000

As with the Nexus 5000, the Nexus 6000 Series switches can also be used with the Nexus 2000 Series Fabric Extenders to create a virtual modular system. The Nexus 6000 can support the same number of attached FEXs (24 in Layer 2 mode, 16 in Layer 3 mode) as the Nexus 5000 at the time of this writing. The Nexus 6000 does have larger scale than the Nexus 5000 in a number of areas (interfaces, MAC addresses, and so on). See the Configuration Guides under the Support section for the Nexus 6000 platform for the most current scale information:

http://www.cisco.com/go/nexus6000

4.5.5 Layer 2 / Layer 3 design options

There is a very common requirement in modern data centers for flexible Layer 2 connectivity between many servers. This requirement is driven by the need for VM mobility, server clustering, and other applications requiring Layer 2 connectivity between servers. A Layer 2 access design makes it easier to assign multiple servers to the same VLAN, as it enables VLANs to be present on multiple access switches simultaneously. For this reason, we explore Layer 2 access design options in more detail than Layer 3 access design options.

To effectively support server connectivity models for all servers, access layer designs must provide high availability options for servers with single connections and multiple connections to the network. One of the key elements of providing high availability in the modern data center is to avoid the complexity of introducing spanning tree protocol wherever possible. In the case of the access layer, avoiding Spanning Tree is a high priority design goal. As mentioned in the previous section, fabric extenders provide two key capabilities to address high availability without relying on the spanning tree protocol:

- Virtual PortChannels (VPC): Support for vPC technology allows fabric extenders to connect to two different Cisco Nexus 5000 Series switches in an efficient active-active configuration. This approach provides continuous availability through the loss of an uplink port, cable, switch port, or switch. The active-active nature of the vPC allows efficient use of bandwidth during normal operation.
- Dual attachment: The risk that a single fabric extender will fail can also be mitigated by dual attaching each server to two Cisco Nexus 5000, 6000, or 7000 Series switches through two fabric extenders using vPC technology. In addition, this configuration mitigates NIC failures on servers. Although most server operating systems support LACP port channeling of multiple NICs, if the server operating system is unable to support port channeling, active/standby NICs can also be supported across two fabric extenders.

4.5.6 Common access layer design options with the Nexus 5000/6000

There are three common access layer design options using the Nexus 5000/6000 switching platforms. Design options will be discussed using the Nexus 5000, but the Nexus 6000 can also be used in the place of the Nexus 5000 to implement the same design options:

- Nexus 5000 as a 10G ToR switch (no fabric extenders)
- Nexus 5000 with single-attached fabric extenders
- Nexus 5000 with dual-attached fabric extenders

Option 1 is illustrated in Figure 4-18. The design involves directly connecting servers to a pair of Nexus 5000 switches via 10GE. Both active/standby and LACP port channel server NIC connectivity is illustrated. In the case of LACP port channel, the port channel to the server NICs is using vPC to support multi-chassis port channeling to the Nexus 5000 switches. Although not illustrated here, single-attached servers can also be connected to one of the Nexus 5000s, but the guidelines for connecting vPC Orphan Ports (discussed earlier) should be followed.

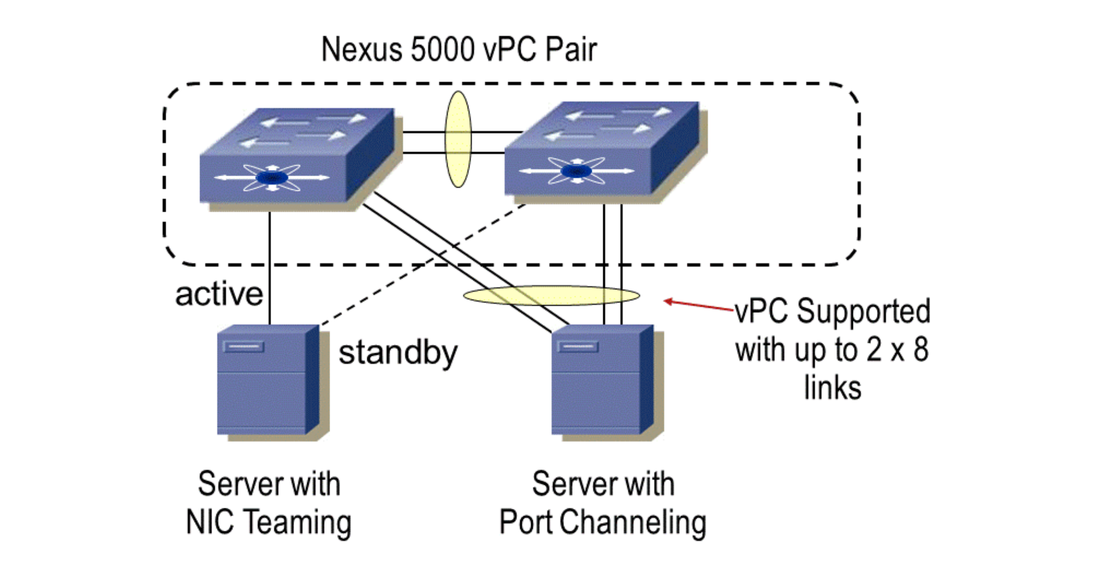

Nexus 5000 vPC Pair
active
standby
vPC Supported with up to 2 x 8 links
Server with NIC Teaming
Server with Port Channeling

Figure 4-18 Dual-attached servers direct to Nexus 5000

Option 2 is illustrated in Figure 4-19. This design includes redundant server NIC connectivity to a pair of Nexus 5000 switches using single-attached fabric extenders. Both active/standby and LACP port channel server NIC connectivity is illustrated. In the case of LACP port channel, the port channel to the server NICs is using vPC through the fabric extenders to the Nexus 5000 switches. Although not illustrated here, single-attached servers can also be connected to one of the fabric extenders, but the guidelines for connecting vPC Orphan Ports (discussed earlier) should be followed.

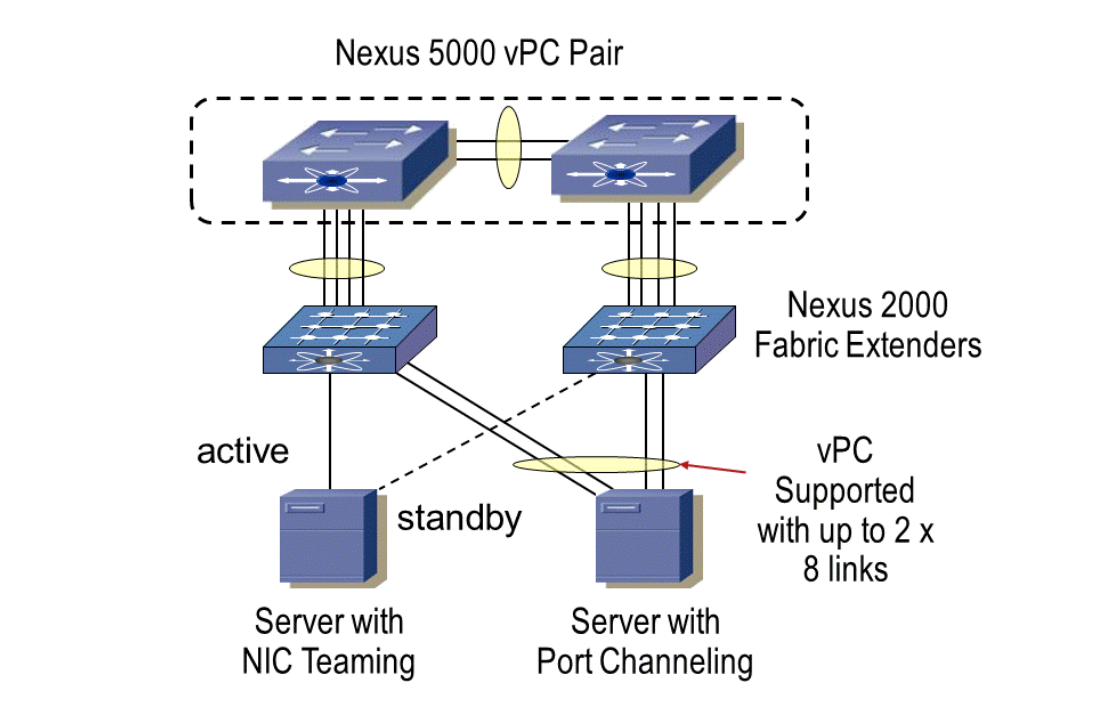


Figure 4-19 Dual-attached servers with single-homed FEX

Option 3 is illustrated in Figure 4-20. This design involves redundant server NIC connectivity to a pair of Nexus 5000 switches using dual-attached fabric extenders. The dual-attached fabric extenders are leveraging vPC for the connectivity to both Nexus 5000 switches. Both active/standby and LACP port channel server NIC connectivity is illustrated. In the case of LACP port channel, the port channel to the server NICs is using vPC to support multi-chassis port channeling up to the Nexus 5000 switches. In this configuration there are two layers of vPC: the first between the Nexus 5000 and the fabric extenders, the second between the fabric extenders and the server. This two-layer vPC configuration is referred to as Enhanced vPC. Single-attached servers can also be supported in this configuration, offering a more highly available connectivity model than a single-homed fabric extender model offers.

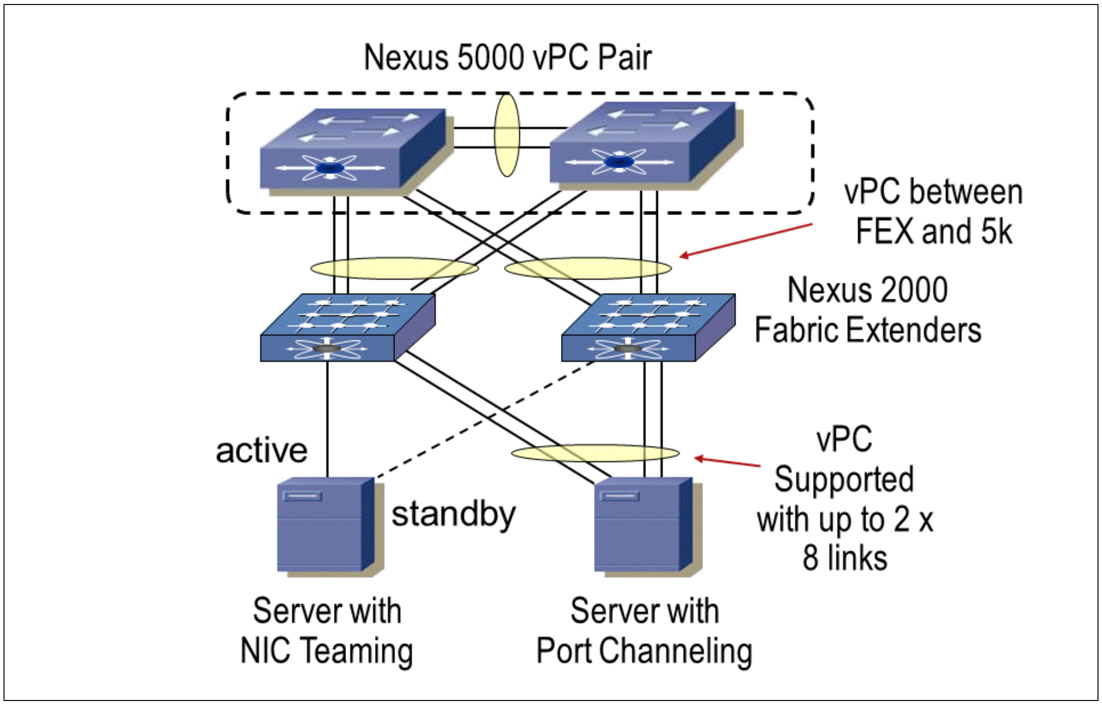

Nexus 5000 vPC Pair
vPC between FEX and 5k
Nexus 2000 Fabric Extenders
active
standby
vPC Supported with up to 2 x 8 links
Server with NIC Teaming
Server with Port Channeling

Figure 4-20 Dual-attached servers with dual-homed FEX

In all three of the foregoing designs, the Nexus 5000 access layer switches deliver L2 networking services only. The connectivity from the Nexus 5000 switches to the aggregation layer will provide trunking of the VLANs up to the aggregation layer. The aggregation layer switches will provide all L3 default gateway and routing services for each VLAN. All three of these designs can support connectivity to virtualized and non-virtualized servers.

Each Nexus 5000 supports up to 24 connected fabric extenders. Therefore, at the time of this writing, design options 1 and 2 can support a maximum of 48 fabric extenders, while design option 3 can support a maximum of 24 fabric extenders.

For the current maximum fabric extender support figures, see the *Nexus 5000 Configuration Limits Guide*:

http://www.cisco.com/en/US/docs/switches/datacenter/nexus5000/sw/configuration_limits/limits_521/nexus_5000_config_limits_521.html

Although it was not specifically discussed, the access layer design principles with the Nexus 6000 platform with FEX are identical to those of the Nexus 5000 and FEX. For more information, see Chapter 8., “Overcoming the limitations of traditional data center networks” on page 421.

Unified access layer with Cisco Nexus 5000 Series switches and Cisco Nexus 2000 Series Fabric Extenders

For more extensive configuration and design information on data center access layer designs with Nexus 5000 switches and Nexus 2000 Fabric Extenders, see the following website:

http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps10110/solution_overview_c22-588237_ps9670_Product_Solution_Overview.html

For more extensive configuration and design information on data center access layer designs with Nexus 5000 switches and Nexus 2000 Fabric Extenders, see the following website:

http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps10110/solution_overview_c22-588237_ps9670_Product_Solution_Overview.html

The Nexus 7000 platform offers an alternative access layer deployment model to the Nexus 5000. Since the Nexus 7000 is a highly available and resilient switching platform, using the Nexus 7000 in the access layer might be a option where availability or additional reduction in managed devices is required.

There are two common access layer design options using the Nexus 7000 switching platform:

1. Nexus 7000 as a 1/10 Gbps EoR or MoR switch (no fabric extenders)
2. Nexus 7000 with single-attached fabric extenders

Option 1 may be a good fit for brown-field environments where an EoR or MoR deployment is already established. However, since this is a legacy model that has significant disadvantages in cable complexity and cost, it is not recommended for greenfield deployments. For this reason, we do not discuss this design in this document.

Option 2 is illustrated in Figure 4-21. This design includes redundant server NIC connectivity to a pair of Nexus 7000 switches using single-attached fabric extenders. Both active/standby and LACP port channel server NIC connectivity is illustrated. In the case of LACP port channel, the port channel to the server NICs is using vPC through the fabric extenders to the Nexus 7000 switches. Although not illustrated here, single-attached servers can also be connected to one of the fabric extenders, but the guidelines for connecting vPC Orphan Ports (discussed earlier) should be followed.

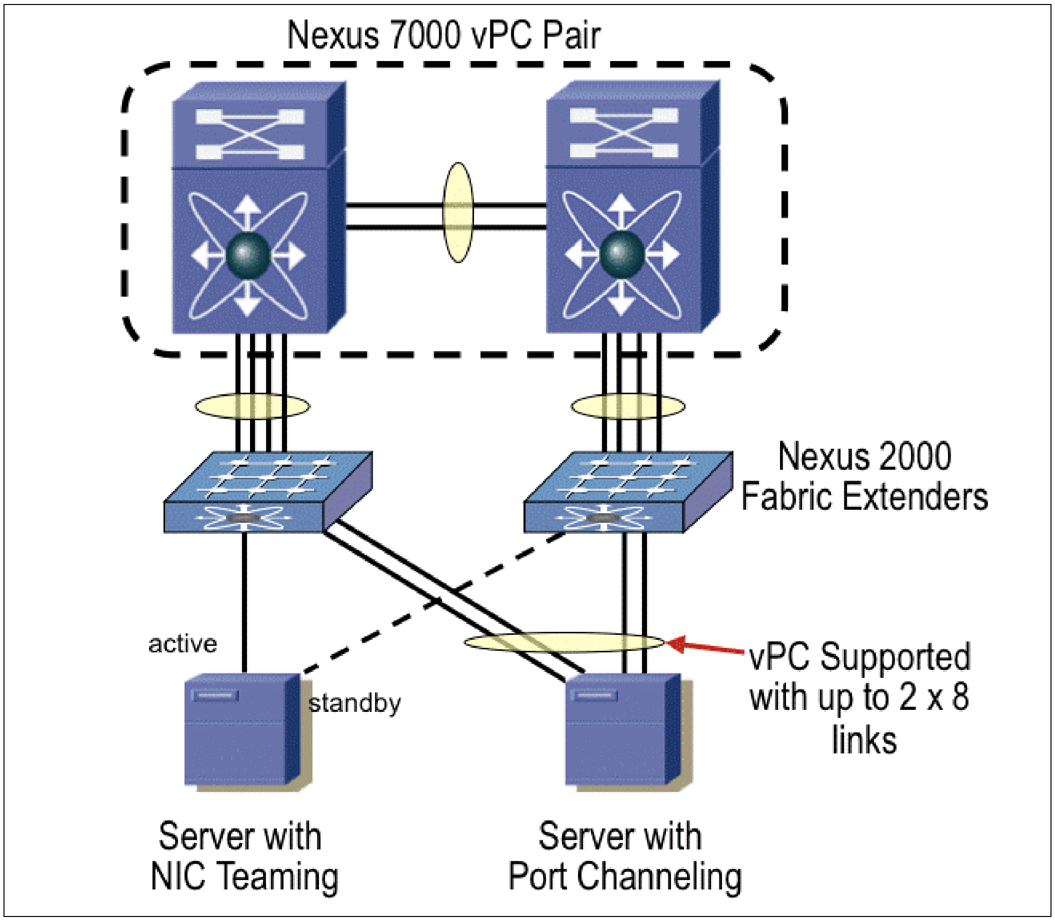

Nexus 7000 vPC Pair
Nexus 2000 Fabric Extenders
active
standby
vPC Supported with up to 2 x 8 links
Server with NIC Teaming
Server with Port Channeling

Figure 4-21 Cisco Nexus 7K VC pair with FEX for server attachment

This is a viable model for any environment, as the physical rack and cable layout is very similar to the Nexus 5000 deployment models using fabric extenders. The Nexus 7000 supports fabric extenders connected only to specific 10 GE modules (M-series and F2 series adapters at the time of this writing). Each Nexus 7000 can support up to 48 fabric extenders at the time of this writing. Therefore such a topology can support a total of 96 fabric extenders. For the current maximum fabric extender support figures, see the *Nexus 7000 Verified Scalability Guide*:

http://www.cisco.com/en/US/docs/switches/datacenter/sw/verified_scalability/b_Cisco_Nexus_7000_Series_NX-OS_Verified_Scalability_Guide.html

Connectivity of the Nexus 5000, Nexus 6000, or of the following Nexus 7000 access designs to the upstream network can be supported with any of these general methodologies:

- Connect via vPC on the Nexus 5000, 6000, or 7000 switches to the aggregation layer which is not capable of Multi-Chassis EtherChannel (vPC or VSS). This is not recommended, but is supported. Two vPCs from the Nexus 5000 switches would be configured, one going to each individual aggregation layer switch.
- Connect via vPC on the Nexus 5000, 6000, or 7000 switches to an aggregation layer that is vPC or VSS capable. **This is a recommended design**, as it requires a single vPC between the Nexus 5000 switches and the two aggregation layer switches. This option also creates a loopless model that effectively eliminates the active role of the spanning tree protocol from the aggregation layer down to the access layer.
- Connect via FabricPath on the Cisco Nexus 5500, 6000, or 7000 switches to a Nexus 5500, 6000, or 7000 Spine. This is also a recommended design, as it completely removes the spanning tree protocol from the topology, provides link load balancing, fast link failover, and many other benefits provided by Fabricpath technology (see Chapter 8, “Overcoming the limitations of traditional data center networks” on page 421 for more detail on Fabricpath).

4.5.7 Nexus L3 access layer design options

Although L2 access designs (all designs discussed thus far) are most commonly deployed in data centers, a Layer 3 access design is also a viable option in some cases. The main consideration as to whether an L3 access design is appropriate is the VLAN scope required.

If VLANs need to be extended to servers beyond the scope of a Nexus access switch pair and associated fabric extenders, L3 Access is not a viable option. However, if a given VLAN does not need to be extended beyond the Nexus access switch pair and associated fabric extenders, L3 Access is a good option. L3 access designs can be deployed with a non-redundant access layer switching model (that is, single access switch), but since these models do not support redundant L2 server NIC connectivity, we discuss only the redundant Access models in this document.

4.5.8 Nexus 5000/6000 L3 access deployment models

As with the Layer 2 access designs, the Nexus 5000 will be used to describe the design options, but the Nexus 6000 can also be deployed in place of the Nexus 5000 as the access switch.

The Nexus 5000 based L3 access design consists of a pair of Cisco Nexus 5500 switches, with or without fabric extenders, connected with Layer 3 links to an aggregation layer of L3 capable switches (such as Catalyst 6500, Nexus 7000, and so on). The Nexus 5500 and associated fabric extenders provide both L2 and L3 functions:

- Layer 2 domain for intra-PoD VM mobility, NIC teaming, and high-availability clusters
- Layer 3 routing between all local VLANs in the L2 domain and routing to/from all other resources on the network outside of the L2 domain. The L3 routing is delivered using the L3 Daughter Module in the Nexus 5500. See Nexus 5500 data sheet for more information:

  http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9670/data_sheet_c78-618603.html

Figure 4-22 shows the three previously discussed Nexus 5000 Access models, but with L3 incorporated into the Nexus 5000 switch. In particular, the Nexus 5500 switches are required, as the first generation Nexus 5010 and 5020 switches do not support the Layer 3 Daughter Module. As illustrated, the L2-L3 domain boundary is now located in the Nexus 5500 switches. All inter-VLAN routing for the VLANs in the illustrated L2 domain is handled by the Nexus 5500s. Connectivity to the rest of the data center is provided by multiple point-to-point routed 10 GE links on the Nexus 5500 switch. Equal Cost Multi-Pathing (ECMP) using OSPF or other routing protocol will provide load distribution across these L3 uplinks to the aggregation layer.

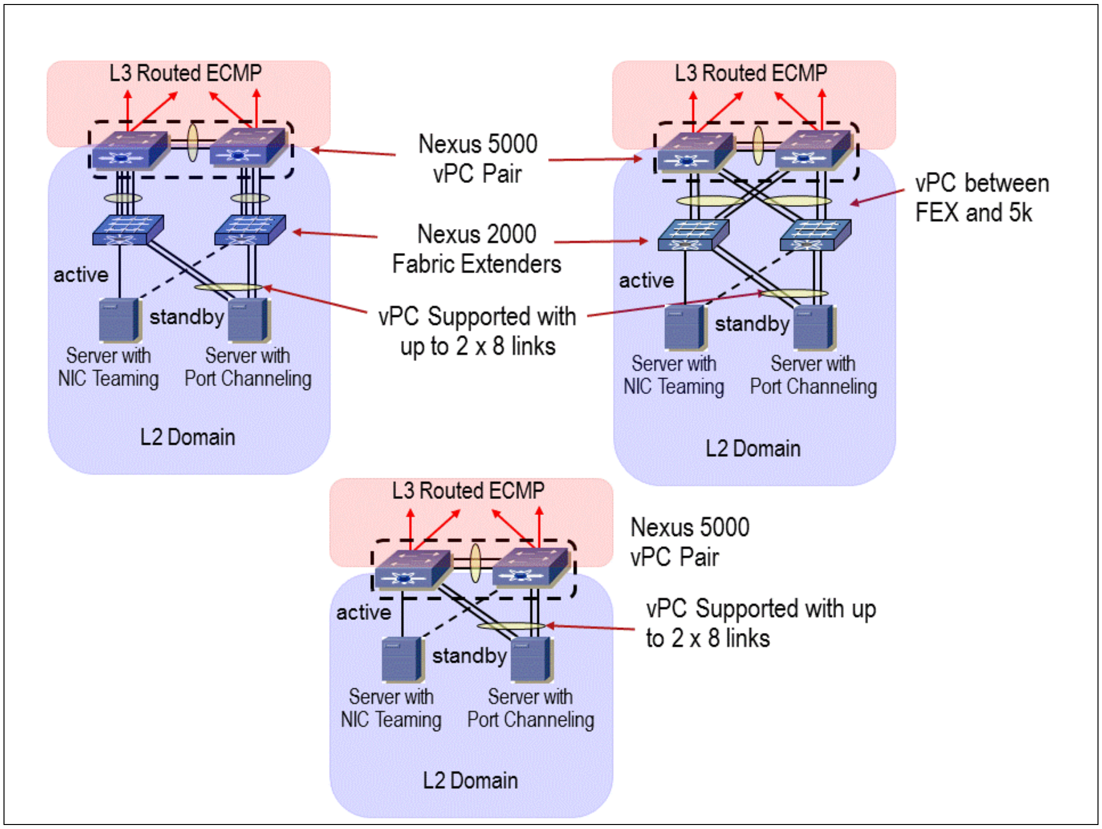


Figure 4-22   Nexus 5000 L3 access deployment models

For L3 scaling limits of the Nexus 5500 platform, see the *Nexus 5000 Configuration Limits* document at this website:

http://www.cisco.com/en/US/docs/switches/datacenter/nexus5000/sw/configuration_limits/limits_521/nexus_5000_config_limits_521.html

The Nexus 7000 based L3 access design is very similar to the Nexus 5000 based L3 design in that it provides a Layer 2 domain for VM mobility, clustering, and so on, and also provides L3 inter-VLAN routing for the L2 domain and L3 routed connectivity to the rest of the data center and broader network. As with the Nexus 7000 L2 design options, we illustrate only the Nexus 7000 with fabric extender design option, as the direct connect to Nexus 7000 design is more specific brown-field deployments and not very commonly used.

Figure 4-23 illustrates the previously discussed Nexus 7000 Access model, but with L3 incorporated into the Nexus 7000 switch. As illustrated, the L2-L3 domain boundary is now located in the Nexus 7000 switches, rather than up in the aggregation layer as with the L2 access design. All inter-VLAN routing for the VLANs in the illustrated L2 domain is handled by the Nexus 7000s. Connectivity to the rest of the data center is provided by multiple point-to-point routed 10 GE links on the Nexus 5500 switch.

Equal Cost Multi-Pathing (ECMP) using OSPF or other routing protocol will provide load distribution across these L3 uplinks to the aggregation layer. Note that either an M-series module or an F2-series module must be in the Nexus 7000 system for L3 services to be available. In addition, fabric extenders must be connected into M-series modules or F2-series modules. Consult Chapter 2 for more detail on the Nexus 7000 hardware components.

For the current feature scalability limits, see the *Nexus 7000 Verified Scalability Guide* at this website:

http://www.cisco.com/en/US/docs/switches/datacenter/sw/verified_scalability/b_Cisco_Nexus_7000_Series_NX-OS_Verified_Scalability_Guide.html

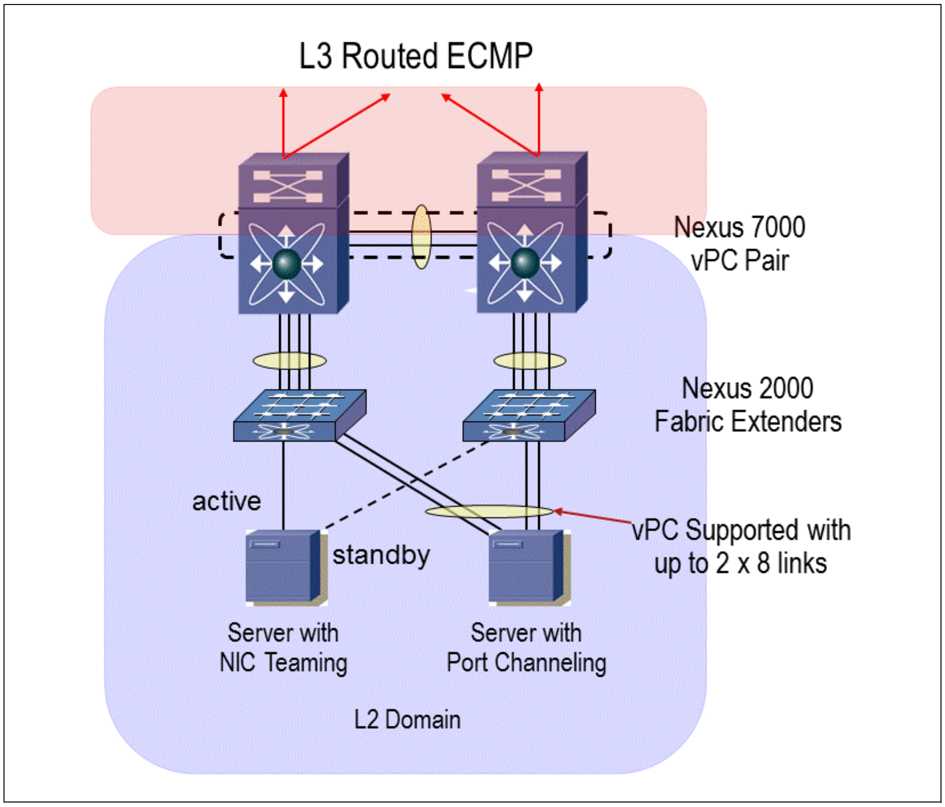


Figure 4-23 Nexus 7000 L3 access deployment model

4.6 Integrated switching: Blade switches and pass-thru modules

A BladeCenter chassis unit houses multiple server blade modules in a single enclosure. The chassis provides power for all blade servers, reduces cabling, and provides a single interface for system management.

There are five models of BladeCenter Chassis, as discussed in Chapter 2, "IBM and Cisco building blocks for the data center" on page 29. Briefly, these are the models:

- BladeCenter E (8677):

  The IBM BladeCenter Type 8677 is a high-density, high-performance, rack-mounted server system developed for medium-to-large businesses. The BladeCenter unit supports up to 14 blade servers in a 7U rack space.

- BladeCenter T (8720/8730):

  The IBM BladeCenter Type 8720/8730 is a rack-mounted, high-density, high-performance blade server system developed for NEBS telecommunications network applications and other applications requiring physical robustness. It supports up to eight blade servers in an 8U space.

- BladeCenter HT (8740/8750):

  The IBM BladeCenter HT is a telecommunications-optimized version of IBM BladeCenter H that delivers outstanding core telecom network performance and high-speed connectivity (more than 1.2 Tbps of aggregate throughput on the backplane). The BladeCenter HT supports up to 12 blade servers in a 12U space.

- BladeCenter H (8852):

  The IBM BladeCenter H Type 8852 unit is a high-density, high-performance rack-mounted server system. The BladeCenter H supports up to 14 blade servers in a 9U rack space.

- BladeCenter S (8886):

  The BladeCenter S system integrates blade servers with serial-attached SCSI (SAS) and Serial Advanced Technology Attachment (SATA) hot-swappable storage in a single chassis that is powered using up to four autoranging 110V/220V power supplies.

The five BladeCenter chassis have the following bays:

- BladeCenter S, E, and T have four standard I/O bays (1, 2, 3, and 4)
- BladeCenter H has six standard I/O bays (1, 2, 3, 4), two bridge bays (5 and 6) and four high-speed bays (7, 8, 9, and 10)
- BladeCenter HT has four standard I/O bays (1, 2, 3, 4) and four high-speed bays (7, 8, 9, and 10).

Below is a list of BladeCenter Ethernet switch modules that are optional in the standard or high-speed bays:

- IBM BladeCenter Open Fabric offers Ethernet networking options integrated into the chassis. High-speed Ethernet delivers 10 Gbps throughput to each blade server through an expansion card in the blade.

  Important: This part of the BladeCenter information center provides installation guides for I/O modules, including modules that are no longer available from IBM. To find the modules that are currently available, go to the BladeCenter Ethernet switch module Web site:

  http://publib.boulder.ibm.com/infocenter/bladectr/documentation/index.jsp?topic=/com.ibm.bladecenter.common.nav.doc/io_pm_welcome_page.html

- IBM BladeCenter 1/10 Gigabit Uplink Ethernet Switch Module:

  The IBM BladeCenter 1/10 Gbps Uplink Ethernet Switch Module serves as a switching and routing fabric for all models of BladeCenter server chassis. In addition to the Layer 2 switching capabilities, this switch supports the expanded capabilities of Layer 3 routing.

- IBM BladeCenter Virtual Fabric 10 Gbps Switch Module:

  IBM developed this 10 gigabits (Gb) I/O system for the BladeCenter as the first full 10 Gb virtual fabric I/O solution. Although 10 Gb throughput to the chassis existed previously, 10 Gb communications to blade servers did not exist before this system became available. The 10 Gbps Virtual Fabric switch module is required in each blade server that will communicate at 10 Gbps.

- IBM BladeCenter 10 Gigabit Uplink Ethernet Switch Module:

  The IBM BladeCenter 10 Gigabit Uplink Ethernet Switch Module for enabling administrators to consolidate full Layer 2 and 3 LAN switching and routing capabilities into a BladeCenter chassis.

- IBM BladeCenter Layer 2/3 GbE Switch Module (GbESM):

  The IBM BladeCenter Layer 2/3 Gigabit Ethernet Switch Module provides full Layer 2 switching and Layer 3 routing for copper and fiber.

- IBM BladeCenter Layer 2-7 GbE Switch Module:

  The IBM BladeCenter Layer 2-7 GbE Switch Module provides a full featured Ethernet switching and routing fabric environment with Layer 2-7 functionality.

- Cisco Catalyst Switch Module 3012 for IBM BladeCenter (IGESM):

  The Cisco Catalyst Switch Modules for IBM BladeCenter provide Layer 2 and basic Layer 3 Ethernet switching. The Cisco Catalyst Switch Module 3012 provides 4 external 1 Gigabits (Gb) per second ports and 14 internal 1 Gb ports.

- Cisco Catalyst Switch Module 3110X for IBM BladeCenter (IGESM):

  The Cisco Catalyst Switch Modules for IBM BladeCenter provide Layer 2 and basic Layer 3 Ethernet switching. The Cisco Catalyst Switch Module 3110X provides 1 external 10 Gigabits (Gb) per second uplink port and 14 internal 1 Gb ports.

- Cisco Catalyst Switch Module 3110G for IBM BladeCenter (IGESM):

  The Cisco Catalyst Switch Modules for IBM BladeCenter provide Layer 2 and basic Layer 3 Ethernet switching. The Cisco Catalyst Switch Module 3110G provides 4 external 1 Gigabit (Gb) per second ports and 14 internal 1 Gb ports.

- Cisco Systems Intelligent Gigabit Ethernet Switch Module (IGESM):

  The Cisco Systems IGESM is a Layer 2 switch that provides full Layer 2 LAN switching capabilities to each blade server in the chassis unit. This switch is also referred to as the Cisco Systems Gigabit Ethernet Switch Module, or the Cisco Systems GbE Switch Module.

- Cisco Systems Fiber Intelligent Gigabit Ethernet Switch Module (IGESM):

  The Cisco Systems Fiber IGESM is a fiber optic Layer 2 switch that provides full Layer 2 LAN switching capabilities to each blade server in the IBM BladeCenter chassis unit. The features and functions of this switch are equivalent to the Cisco Catalyst switches, but with enhancements and modifications for operating within the IBM BladeCenter unit. This switch is also referred to as the Cisco Systems Fiber Gigabit Ethernet Switch Module, or the Cisco Systems Fiber GbE Switch Module.

- Server Connectivity Module for IBM BladeCenter:

  The Server Connectivity Module (SCM) for IBM BladeCenter provides a simple Ethernet interface option for connecting the IBM BladeCenter system to the network infrastructure.

A representative BladeCenter chassis is shown in Figure 4-24. The chassis can have up to 14 server blades. The standard bays can accommodate 1 Gbps Ethernet switches or 1 Gbps pass-thru modules. The high speed bays can accommodate 10 Gbps switches or 10 Gbps pass-thru modules. Ethernet Switches can be from several vendors such as IBM BNT®, Brocade, or Cisco and can also be legacy Ethernet (1 Gbps or 10 Gbps) or 10-Gbps FCoE switches with up to ten 10-Gbps uplink ports each. Each H chassis, for example, can have up to four switches, which represents up to 400 Gbps full-duplex. All this hardware is standard-based.

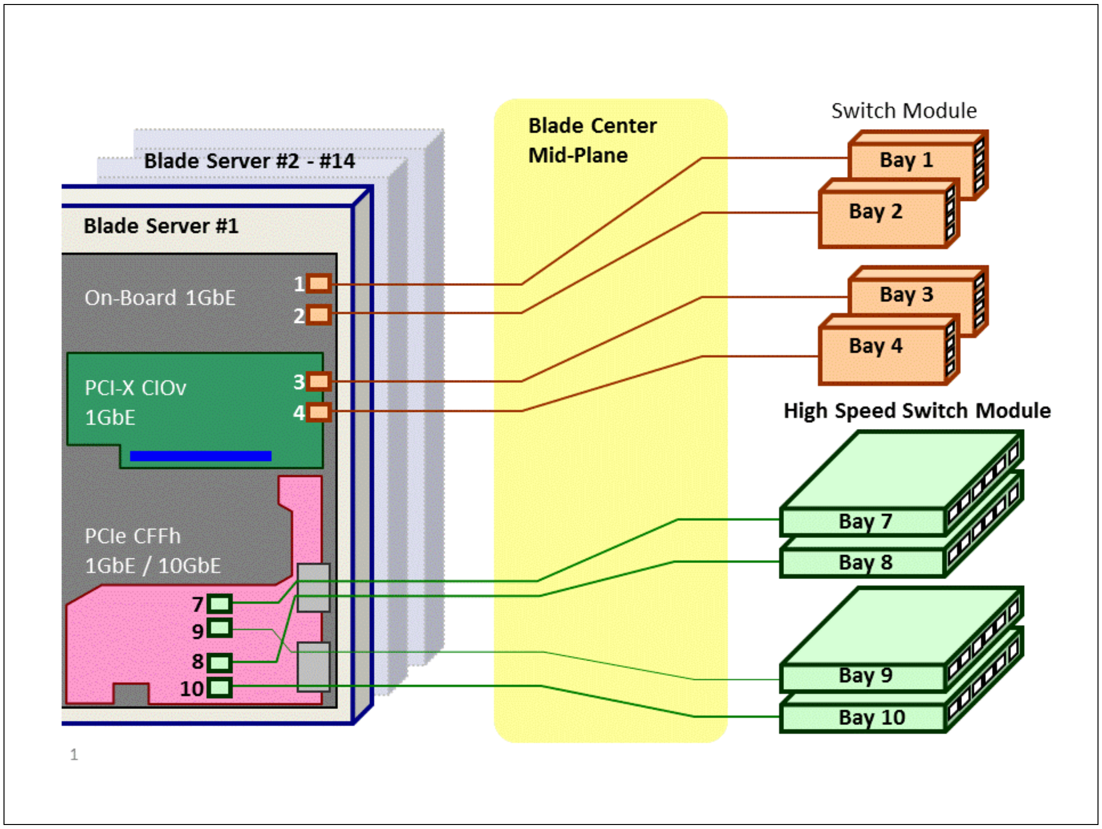


Figure 4-24 BladeCenter Chassis

The following list explains the BladeCenter extendability shown in Figure 4-24:

- Blade server on-board NIC (1GbE) x 2
- Blade server slots:
  - Combination I/O for vertical (CIOv) slot x 1
  - Combination Form Factor Horizontal (CFFh) slot x1
- On-board, CIOv, CFFh requires switch module to communicate outside network.
- By combination of adaptor and switch module, many connection options would be provided, such as Ethernet / Fibre Channel / InfiniBand / FCoE

IBM BladeCenter offers a Copper Pass-thru Module and a Fiber-Optic Pass-thru Module. The pass-thru modules connect blade servers at a data rate of 1 gigabits per second to an external Ethernet network.

Also, see the following link:

http://www.redbooks.ibm.com/technotes/tips0715.pdf

4.6.1 IBM BladeCenter 10 Gb Ethernet Pass-thru Module

The IBM 10 Gb Ethernet Pass-thru Module for IBM BladeCenter is a high-speed Ethernet pass-thru module that can be installed into a BladeCenter unit that supports high-speed I/O modules.

Note: SFP+ (small form-factor pluggable plus) transceivers are not included and must be purchased separately.

Features and specifications

The 10 Gb Ethernet Pass-thru Module includes the following features and functions:

- Single-wide high-speed switch module
- 14 internal 10 Gb ports to the server blades (no auto-negotiation)
- Up to fourteen 10 Gb SFP+ uplink ports (SFP+ modules are not included)
- Support for SR, LR, and DAC copper cables
- Direct one-to-one mappings of external and internal ports (no configuration required)

4.6.2 Intelligent Copper Pass-thru Module (ICPM) for IBM BladeCenter

The Intelligent Copper Pass-thru Module (ICPM) for IBM BladeCenter is a low-cost solution, ideal for clients who want to use their existing switching infrastructure with IBM BladeCenter. It does not require any configuration and so is very easy to use.

Pass-thru modules are I/O options that can be used for direct connections of blade ports to the external infrastructure devices such as network switches. The pass-thru module is almost like a traditional network patch-panel, it routes internal blade ports to outside the BladeCenter. Current modules (used in non-high-speed bays) use a design that links several internal blade ports to a single external port on pass-thru module, so specific “hydra” or “octopus” pass-thru cables have to be used with pass-thru modules to provide standard external connections (one connector per each blade port).

The two types of standard pass-thru modules are Copper Pass-thru Module (CPM) and Optical Pass-thru module (OPM). Both of them are capable of carrying Ethernet signals.

ICPM provides an unconfigured network connection that enables the blade servers in the BladeCenter unit to connect to an existing network infrastructure. No configuration of the Copper Pass-thru Module is required. The ICPM provides a single connection from each blade to one RJ-45 connector, which can go directly to an external switch or patch panel from each of the 14 blades in the chassis to an RJ-45 plus for connection.

ICPM features and functions include these:

- Connection with up to 14 blade servers via 14-port copper ports
- Direct one-one mapping to each blade server
- Up to 1 Gb operation on uplinks

Figure 4-25 depicts a Cisco Nexus Data Center design utilizing the BladeCenter pass-thru modules. The BladeCenters shown have 14 server blades. The example below shows two 10 Gbps pass-thru modules and two 1 Gbps pass-thru modules per chassis. Each chassis appears to have 14 servers each with two 10 Gbps Ethernet cards and two 1 Gbps Ethernet cards.

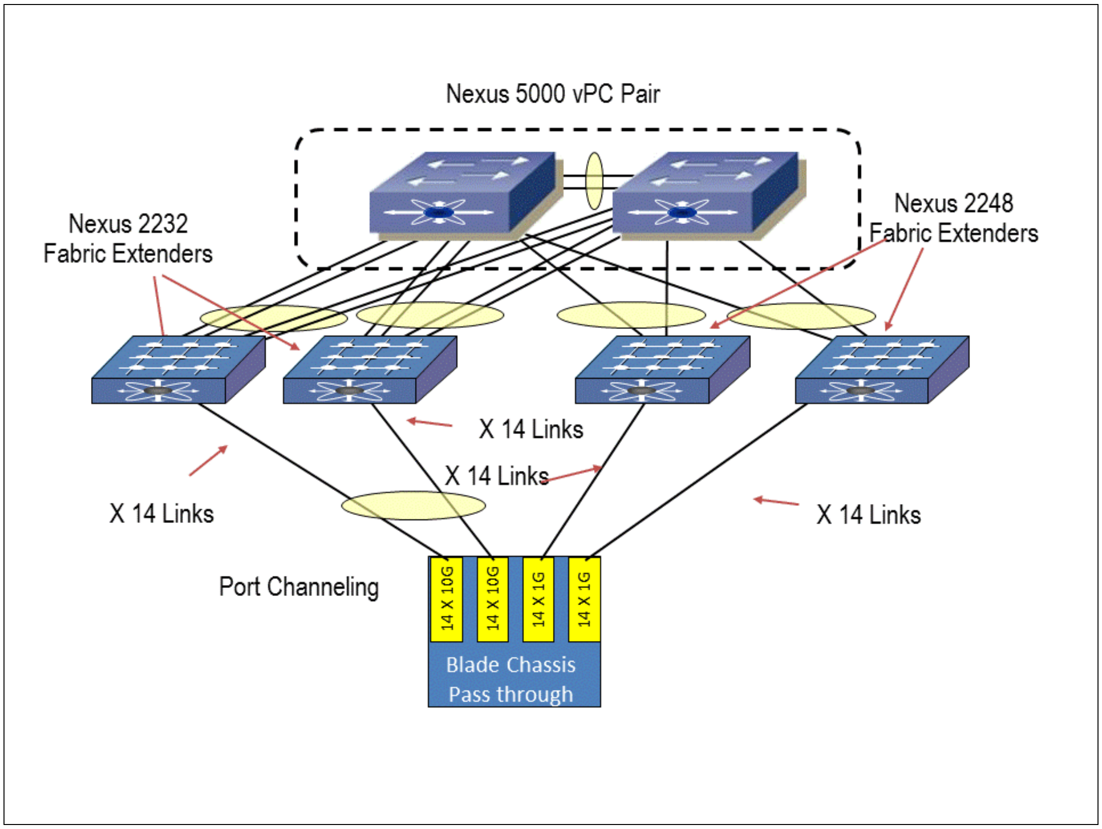


Figure 4-25 BladeCenter with pass-thru modules

Note that this pass-thru connectivity model for the IBM BC-H can also be applied to the Nexus 5000 single-attached fabric extender model, the Nexus 5000 direct attached model, or the Nexus 7000 fabric extender models discussed previously in this chapter.

4.6.3 IBM BladeCenter Network access layer with Ethernet Switch Modules

Many IBM BladeCenter deployments utilize Ethernet switches embedded in the BladeCenter chassis. The Cisco Nexus 4000 is one such blade switch that is available for IBM BladeCenter-H. See Chapter 2, “IBM and Cisco building blocks for the data center” on page 29 for a detailed overview of the Nexus 4000 switch. The recommended access design option using the Nexus 4000 switch also leverages vPC or VSS to the upstream network switches. This approach creates a loopless topology to the access layer and prevents reliance on the spanning tree protocol.

Figure 4-26 illustrates the IBM BladeCenter-H with Nexus 4000 blade switches in a vPC environment. Each BladeCenter-H chassis has two embedded Nexus 4000 blade switches. Each Nexus 4000 acts as an independent switch, with two 10G uplinks connected to each of the Nexus 5000 switches. These four total uplinks from each Nexus 4000 are configured in a vPC on the Nexus 5000 switches, thereby creating a loopless topology. Note that you can elect to use all six of the Nexus 4000 uplinks (three to each Nexus 5000) if more bandwidth is required. Since the Cisco Nexus 4000 currently does not natively run vPC, the blade server-facing ports cannot be in a vPC. Therefore, BladeCenter Network redundancy can be achieved with the Combination of Trunk failover and NIC teaming.

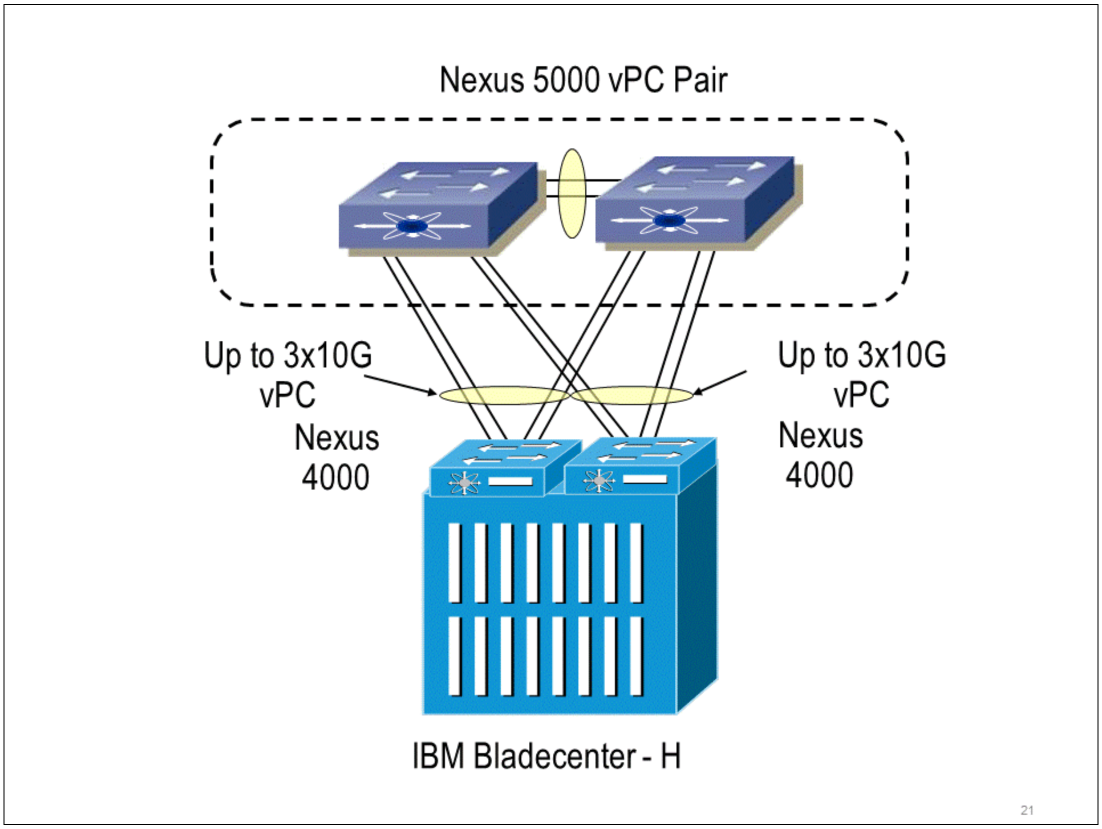


Figure 4-26 BladeCenter with Nexus 4K configured with a pair of Nexus 5K running vPC

- Trunk Failover detects failure on external network device and cable, then disables internal port on ESM.
- NIC teaming can be configured for Active-Standby configuration by server (for example, Windows, Linux) or Active-Active or Active-Standby configuration by server (for example, by port-group of Virtual Switch in VMware).

For detailed configuration and alternative design options for the Nexus 4000, see the Cisco *Nexus 4000 Series Design Guide*:

http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps10596/deployment_guide_c07-574724.html

4.6.4 Additional BladeCenter information

IBM BladeCenter Virtual Fabric home page:

http://www.ibm.com/systems/bladecenter/hardware/openfabric/virtualfabric.html

IBM BladeCenter InfoCenter, IBM Virtual Fabric 10Gb Switch Module:

http://ibm.biz/Bdxrgh

IBM Redbooks Product Guide, Emulex 10GbE VFA II and Emulex 10GbE VFA Advanced II for IBM BladeCenter HS23:

http://www.redbooks.ibm.com/abstracts/tips0875.html?Open

IBM Redbooks Product Guide, Emulex Virtual Fabric Adapter (CFFh) for IBM BladeCenter:

http://www.redbooks.ibm.com/abstracts/tips0748.html?Open

IBM Redbooks Product Guide, QLogic Virtual Fabric Extension Module for IBM BladeCenter:

http://www.redbooks.ibm.com/abstracts/tips0717.html?Open

IBM US Announcement Letter:

http://ibm.com/common/ssi/cgi-bin/ssialias?infotype=dd&subtype=ca&&htmlfid=897/ENUS109-288

IBM BladeCenter Interoperability Guide:

http://ibm.com/support/entry/portal/docdisplay?lndocid=MIGR-5073016

IBM Redbooks publication, *IBM BladeCenter Products and Technology*, SG24-7523:

http://www.redbooks.ibm.com/abstracts/sg247523.html

For further details, see this paper:

IBM Flex System Networking in an Enterprise Data Center, REDP-4834

4.7 Flex System: Integrated switches and pass-thru modules

The Ethernet networking I/O architecture for the IBM Flex System Enterprise Chassis includes an array of connectivity options for server nodes installed in the enclosure. Users can decide to use a local switching model that provides performance, cable reduction, and a rich feature set, or use pass-thru technology and allow all Ethernet networking decisions to be made external to the Enterprise Chassis.

Flex modules provide local switching capabilities and advanced features that are fully integrated into the operation and management of the Enterprise Chassis. The Enterprise Chassis has four I/O bays in the rear of the chassis.

Figure 4-27 shows the FlexSystems chassis with four Scalable switches to enable high speed connectivity; Ethernet (FCoE, iSCSI), Fibre Channel, and InfiniBand.

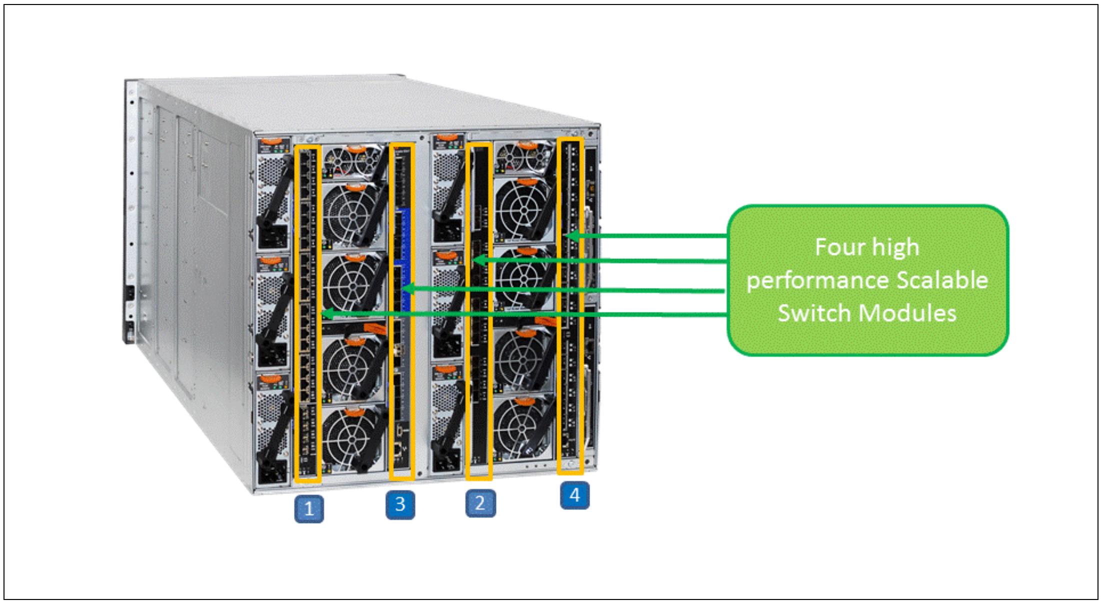


Figure 4-27 Flex Systems Chassis

The IBM Flex System Enterprise Chassis features a number of Ethernet I/O module solutions providing a combination of 1 Gb and 10 Gb ports to the servers and 1 Gb, 10 Gb, and 40 Gb for uplink connectivity to the outside upstream infrastructure. The IBM switch modules run standards-based IBM Networking OS and interoperate seamlessly with Cisco networks and the IBM Flex System Enterprise Chassis ensures that a suitable selection is available to meet the needs of the server nodes. There are four Ethernet I/O modules available for deployment with the Enterprise Chassis:

- IBM Flex System Fabric CN4093 10Gb Converged Scalable Switch
- IBM Flex System EN2092 1Gb Ethernet Scalable Switch
- IBM Flex System Fabric EN4093 10Gb Scalable Switch
- IBM Flex System EN4091 10Gb Ethernet Pass-thru

4.7.1 IBM Flex System EN2092 1 Gb Ethernet Scalable Switch

The EN2092 1 Gb Ethernet Switch is primarily a 1 Gb switch, offering up to 28 x 1 Gb downlinks to the internal nodes, with a total combination of up to 20 x 1 Gb RJ45 uplinks and four 10 Gb uplinks with “pay-as-you-grow” scalability.

The EN2092 1Gb Ethernet Scalable Switch has the following features and specifications:

Internal ports:

- Twenty-eight internal full-duplex Gigabit ports with 14 ports enabled by default; an optional Feature on Demand (FoD) capability license is required to activate the other 14 ports. Two internal full-duplex 1 GbE ports are connected to the chassis management module.

External ports:

- Four ports for 1 Gb or 10 Gb Ethernet SFP+ transceivers (support for 1000BASE-SX, 1000BASE-LX, 1000BASE-T, 10 GBASE-SR, or 10 GBASE-LR) or SFP+ copper direct-attach cables (DAC). These ports are disabled by default and an optional FoD license is required to activate them. SFP+ modules are not included and must be purchased separately.
- A total of 20 external 10/100/1000 BASE-T Gigabit Ethernet ports with RJ-45 connectors (10 ports are enabled by default; an optional FoD license is required to activate the other 10 ports).
- One RS-232 serial port (mini-USB connector) that provides an additional means to configure the switch module.

4.7.2 IBM Flex System Fabric EN4093 10 Gb Scalable Switch

The EN4093 10 Gb Scalable Switch is primarily a 10 Gb switch that can provide up to 42 x 10 Gbps server ports, and up to 14 SFP+ 10 Gb and two QSFP+ 40 Gb external upstream facing ports, depending on the applied upgrade licenses.

The IBM Flex System Fabric EN4093 10 Gb Scalable Switch has the following features and specifications:

Internal ports:

- A total of 42 internal full-duplex 10 Gigabit ports (14 ports are enabled by default; optional FoD licenses are required to activate the remaining 28 ports.)
- Two internal full-duplex 1 GbE ports connected to the chassis management module

External ports:

- A total of 14 ports for 1 Gb or 10 Gb Ethernet SFP+ transceivers (support for 1000BASE-SX, 1000BASE-LX, 1000BASE-T, 10 GBASE-SR, or 10 GBASE-LR) or SFP+ copper direct-attach cables (DAC). There are 10 ports enabled by default and an optional FoD license is required to activate the remaining four ports. SFP+ modules and DAC cables are not included and must be purchased separately.
- Two ports for 40 Gb Ethernet QSFP+ transceivers or QSFP+ DACs (these ports are disabled by default; an optional FoD license is required to activate them). QSFP+ modules and DAC cables are not included and must be purchased separately.
- One RS-232 serial port (mini-USB connector) that provides an additional means to configure the switch module.

4.7.3 IBM Flex System EN4091 10 Gb Ethernet Pass-thru Module

The EN4091 10Gb Ethernet Pass-thru Module offers a 1:1 connection between a single-node bay and an I/O module uplink. The module has no management interface and can support 1 Gb and 10 Gb dual port adapters installed on the nodes. If quad port adapters are used in a node, only the first two ports access the pass-thru modules. The necessary 1G or 10G module (SFP, SFP+, or DAC) must also be installed in the external ports of the pass-thru to support the desired speed (1 Gb or 10 Gb) and medium (fiber or copper) for adapter ports on the node.

The IBM Flex System EN4091 10 Gbps Ethernet pass-thru module includes the following features and specifications:

Internal ports:

- A total of 14 internal full-duplex Ethernet ports that can operate at 1 Gbps or 10 Gbps speeds

External ports:

- A total of 14 ports for 1 Gb or 10 Gb Ethernet SFP+ transceivers (support for 1000BASE-SX, 1000BASE-LX, 1000BASE-T, 10 GBASE-SR, or 10 GBASE-LR) or SFP+ copper direct-attach cables (DAC). SFP+ modules and DAC cables are not included and must be purchased separately.

  This device is unmanaged and has no internal Ethernet management port; however, it provides its vital product data (VPD) to the secure management network in the Chassis Management Module

For the data center network design for this book, consider FlexSystems the next generation BladeCenters.

Figure 4-28 shows IBM Flex systems attached to a Cisco Nexus access layer. The figure on the left demonstrates a Flex system with two IBM Ethernet switch modules in it. The IBM Ethernet switch modules use virtual link aggregation (vLAG), which is equivalent to Cisco vPC. The figure on the right demonstrates a Flex system with two pass-thru modules.

> **Terminology:** vPC uplinks from IBM Ethernet switch modules are called *vLAG*.

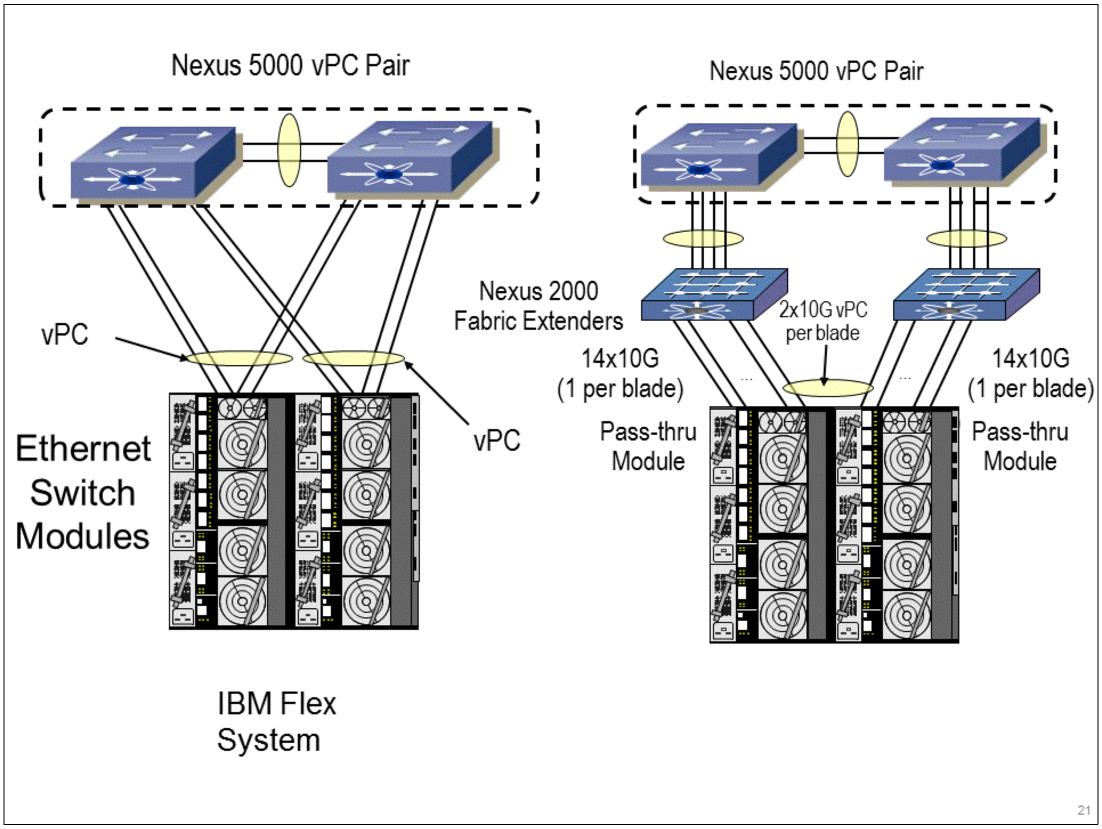


Figure 4-28 IBM Flex Systems attached to Cisco Networks

Table 4-29 shows a comparison of Flex Systems to BladeCenter.

Table 4-5 .Comparison of Flex Systems to BladeCenter

Feature	BladeCenter H	Flex System
Max 10Gb ports per server	4	8 (capable of 16)*
Max 1Gb ports per server	8	8
Max vNICs per sever	14 (HS23)	32 (capable of 64)*
Max 10Gb Ethernet + 8Gb FC per server	4(Ethernet) + 2(FC)	4(Ethernet) + 2(FC) Capable - 8(Ethernet) + 8(FC)*
10Gb Ethernet + 8Gb FC + IB	Not supported	4 CNA **+ 2 IB
Max 10Gb or FCoE per server	4X10 = 40Gb	8X10 = 80Gb (capable 160Gb)
Max Ethernet & FC throughput per server with FC	2X1Gb + 4X10Gb + 2X8Gb = 56Gb	4X10Gb + 2X16Gb = 72Gb
Max throughput possible per server	(4+8+16) PCI Gen2 X 4GT/s = 112GB/s	(16+16) PCI Gen3 X 8GT/s = 256GB/s
40Gb Ethernet Support	No	2X40Gb uplink at launch 40Gb to server Future
Max 10Gb uplink bandwidth per chassis	4x10X100Gb = 400Gb	4X10X220Gb = 880Gb

IBM Redbooks publications and IBM Redpapers™ publications: Online in draft format:

- *IBM PureFlex System Products and Technology*:

  http://www.redbooks.ibm.com/abstracts/sg247984.html?Open

- *IBM Flex System p260 and p460 Planning and Implementation Guide*:

  http://www.redbooks.ibm.com/abstracts/sg247989.html?Open

- *IBM Flex System Networking in an Enterprise Data Center*:

  http://www.redbooks.ibm.com/abstracts/redp4834.html?Open

- *Product Guides*:

  http://www.redbooks.ibm.com/portals/puresystems?Open&page=pgbycat

4.8 IBM System x (physical and virtualized)

IBM provides a full line of Intel based servers:

- Tower servers: Ideal for smaller businesses, smaller remote environments, and meeting standalone requirements.
- Rack servers: Great for small to mid-range businesses, segmented workloads, and “fit for purpose” applications.
- Blade servers: Optimized for large consolidation environments, server and infrastructure replacement, and remote environments.
- Enterprise servers: High-end rack and blade servers suited for heavy vertical workloads, virtualization, and legacy system replacements.

- Expert integrated systems: IBM PureSystems, IBM PureFlex System: The PureFlex System combines compute, storage, networking, virtualization and management into a single infrastructure system that is expert at sensing and anticipating resource needs to optimize your infrastructure.

These systems can be configured with a myriad of Network Interface Card (NIC) options.

Refer to the IBM web site for a list of currently available servers. They can be found at:

http://www-03.ibm.com/systems/x/hardware/

For currently available xSeries, see this website:

http://www.redbooks.ibm.com/redpapers/pdfs/redpxref.pdf

More information on the management interfaces can be found at:

http://www.redbooks.ibm.com/redbooks/pdfs/sg246495.pdf

The xSeries servers supports multiple Ethernet NIC cards that support both 1G and 10G Ethernet. A typical example is shown in Figure 4-29 with two 10G interfaces for production traffic and two 1G interfaces for back up. A fifth interface called the RSA or IMM interface is also shown.

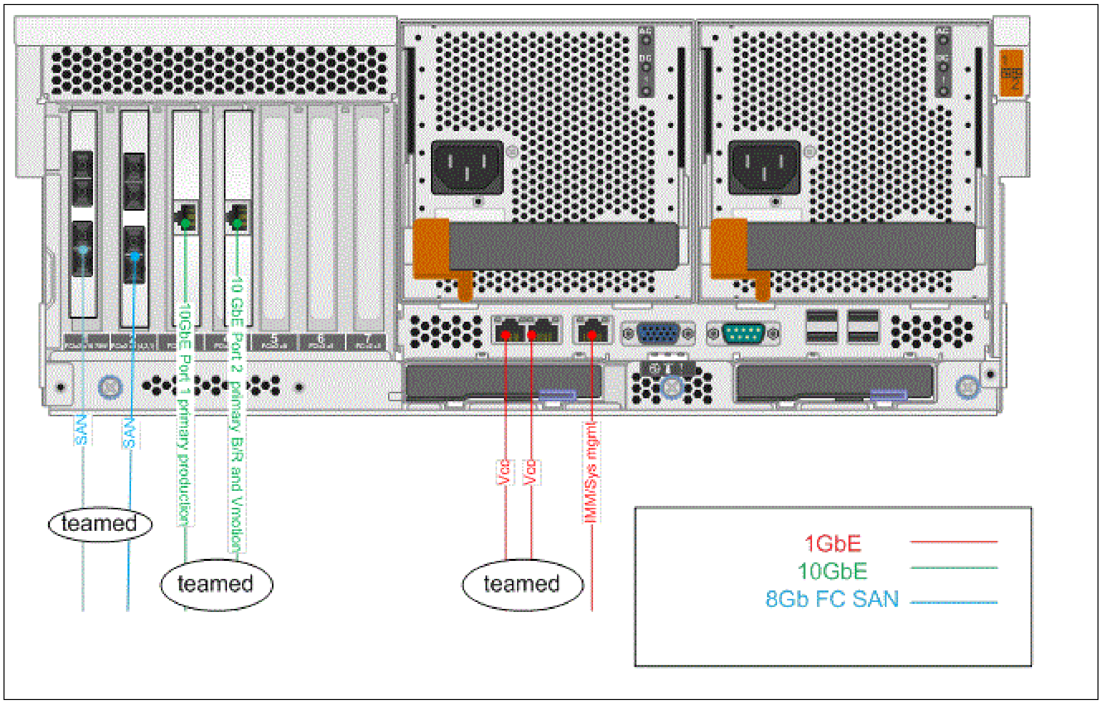

SAN
SAN
10GbE Port 1 primary production
10 GbE Port 2 primary B/R and Vmotion
teamed
teamed
Vco
Vco
IMM/Sys mgmt
teamed
1GbE
10GbE
8Gb FC SAN

Figure 4-29 Typical xSeries hardware with network and SAN interface cards

IBM xSeries® servers have an Integrated Management Module (IMM) or a Remote Supervisor Adapter (RSA) interface. IMM or RSA are systems management interfaces for xSeries. This provides many options for alerting, monitoring, and remote management of xSeries servers. The RSA or IMM can be configured for Serial or Ethernet connectivity.

This chapter only focuses on the network configurations for virtualized and non-virtualized servers.

4.8.1 Link aggregation (port trunking)

Taking multiple physical LAN ports and bundling them together to look like a single logical port is called link aggregation or port trunking. It can be run between network devices or between servers and network devices. Link aggregation accomplishes both redundancy and increasing bandwidth. It can be configured to support active-active or active-passive configurations.

Link aggregation can be accomplished with three standards:

- Static EtherChannel with no intelligent protocol maintaining integrity of the links included in the bundle.
- Port Aggregation Control Protocol (PAgP) is a Cisco proprietary link aggregation protocol sometimes used between Cisco switches.
- Link Aggregation Control Protocol (LACP) is included in the 802.3ad IEEE specification and therefore recommended for link aggregation to servers.

If one end of a link is attached to the switches or virtual switches that support link aggregation and the other end is attached to a server that supports link aggregation, the link can be configured for link aggregation. Because EtherChannel came out before the IEEE standard people use the term EtherChannel and link aggregation synonymously. In order to be safe with vendor interoperability it is always best to use the LACP 802.3ad standard. Links can typically be configured to support both VLAN trunking (802.1q) and LACP (802.3ad).

Link aggregation has been around for a long time and can be used with both virtualized and non-virtualized servers. There are also multiple other proprietary aggregations protocols not discussed here.

The other method of NIC redundancy worth mentioning here is Linux bonding, which aggregates multiple NICs into a single logical bonded interface of two or more so-called NIC slaves. This function comes as a standard with many of the current Linux implementations.

4.8.2 xSeries non-virtualized application servers

Figure 4-30 shows xSeries non-virtualized application servers. The server on the right is shown as having all 1 gigabit Ethernet interfaces. This would typically be a non-virtualized “legacy” application server. More than1 Gbps Ethernet cards may be used if higher I/O is required. However, the cross over point on cost between 1 Gbps and 10 Gbps ports tends to be at about 2 or 3 NICs.

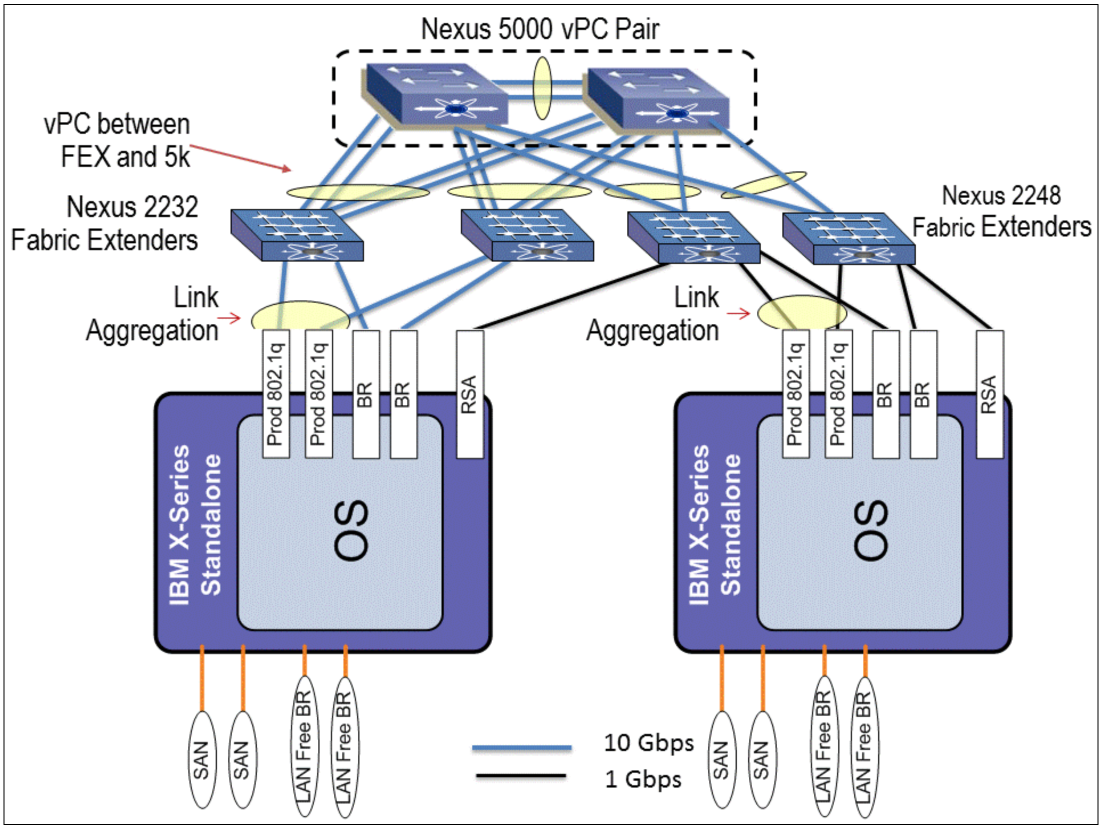


Figure 4-30 xSeries non Virtualized

The server on the left is similar in design to the application server, but shows all 10 Gbps interfaces except for the RSA. Separate logical 10 Gbps interfaces are used for production and backup recovery. Multiple 10 Gbps Ethernet NICs may be used if higher I/O is required. This would be a unique case. The RSA or IMM port is still 1 Gbps.

Table 4-6 shows the network port counts used as guidance to client architects for application servers.

Table 4-6 Server guidance for non-virtualized application servers

Platform	1 Gbps	10 Gbps	Fibre Channel 8 Gbps
Intel stand-alone servers	1 RSA/IMM	0	2 client production
Application, infrastructure	2 Production	0	2 LAN for free if large file sizes
Web servers	2 Backup/recovery	0	2 additional for high I/O environments
Total	5	0	2-6

Table 4-7 shows the network port counts used as guidance to client architects for database servers.

Table 4-7 Server guidance for non-virtualized application servers

Platform	1 Gbps	10 Gbps	Fibre Channel 8 Gbps
Intel database servers	1 RSA/IMM	0	2 client production
		2 production	2 LAN free if DB over 200 Gb
		2 backup/recovery	2 additional for high I/O environments
Total	1	4	2-6

4.8.3 xSeries virtual server

NIC teaming occurs when multiple uplinks are associated with a single vSwitch. A team can share the load of traffic, or provide passive failover in the event of a network outage.

The following load balancing options for NIC teaming are available:

- Route based on the originating port ID
- Route based on IP hash (source and destination IP address)
- Route based on source MAC hash
- Use explicit failover order

IP-based teaming requires that the physical switch be configured with link aggregation.

The following high availability options for VMware are available:

- Network vMotion
- VMware HA
- Storage vMotion
- VMware FT

Figure 4-31 shows the VMware network components.

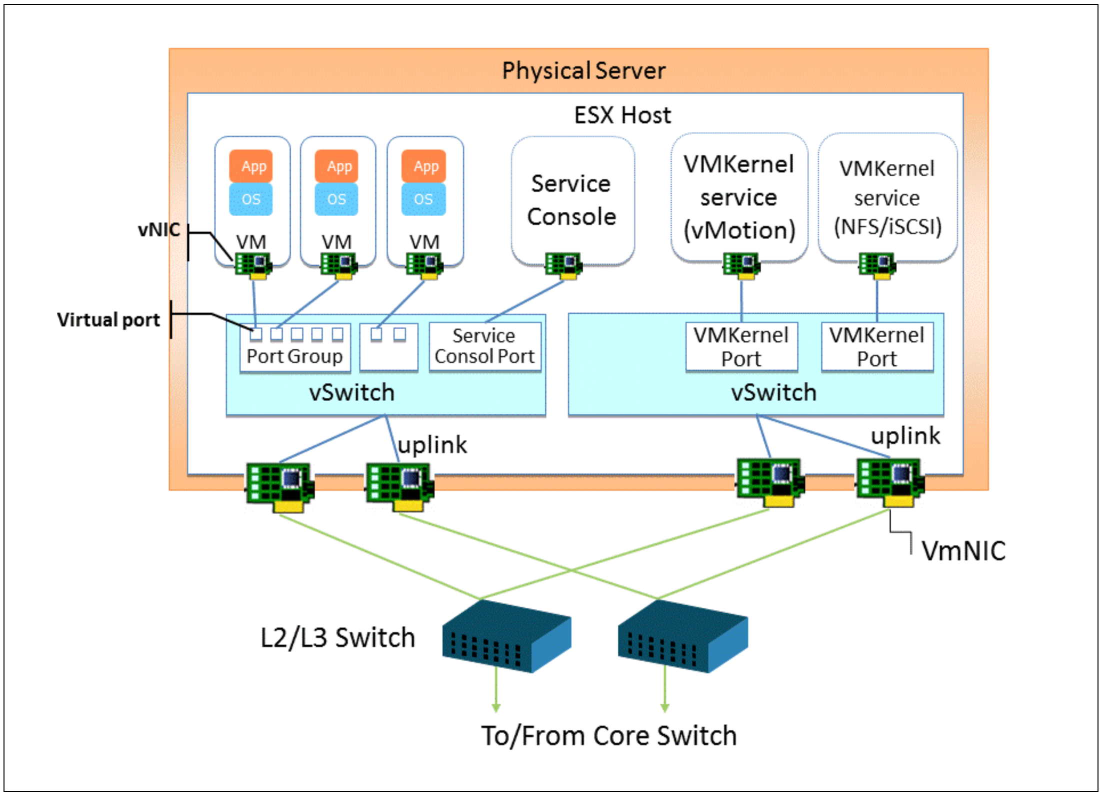


Figure 4-31 VMware Network Components

The following list describes the VMware network components:

- Service Console (also known as the Virtual Console) provides management i/f for ESX.
- VMKernel Service is the VM Kernel TCP/IP networking stack supports iSCSI, NFS, and vMotion.
- Port Group aggregates multiple virtual ports and specifies port configuration option such as VLAN, traffic shaping, and NIC teaming.
- vSwitch: Virtual Switch has two types:
  - vNetwork Standard Switch (vSS).
  - Network Distribution Switch (vDS). A vDS acts as a single vSwitch across all associated ESX hosts. Both IBM and Cisco offer VDS switches.
- vmNIC Physical Server NIC. The vmNIC cannot be shared between different vSwitches.
- vNIC: Each VM has a NIC which is Virtual. MAC addresses are automatically generated for vNICs that are used by the service console, the VMKernel, and virtual machines.

See the VMware web site for configuration maxima.

Maxima for VMware 4.1:

http://www.vmware.com/pdf/vsphere4/r41/vsp_41_config_max.pdf

Maxima for VMware 5.0:

https://www.vmware.com/pdf/vsphere5/r50/vsphere-50-configuration-maximums.pdf

IBM uses a three-interface design as a best practice to reduce network traffic interference:

- Service or Production Interface: The first NIC is carrying the actual application traffic and is the most important network interface and as such the default network interface card. Such Service Interfaces typically belong to the production segment. The production segment is provided to host internal network segments or infrastructure which require access to or from the Public Internet. Multiple production interfaces may be configured on the same servers.
- Admin Interface: The second NIC connects into an administration network. It is separate so that interference with application traffic is minimized. The Admin network (in-band management) may be routed through firewalls and proxies to management workstations of system administrators or support personal. It is also often used to separate monitoring traffic from application traffic, and then the receiving endpoint located behind the firewall is a shared management server.
- Backup Recovery Interface: The third NIC connects into a LAN based Backup & Recovery environment. The requirements here are high data transfer inside the backup window allowed by the application. So the NIC should be able to increase the payload in an IP frame, this is achieved by configuring jumbo frames on the complete path to the Backup Server.

Following is an example of an IBM System x3850x5 in a rack configuration for virtualization.

Per Server:

- 2 x 10Gb interfaces for production and production admin
- 2 x 10Gb interfaces for Backup Recovery with Jumbo frame support
- 1 x 1Gb (100Mb) interface for RSA. System port that allows basic remote tasks such as server reboot, power management and BIOS error log retrieval.
- 2 x 1Gb Admin interfaces for Management. This interface is used to communicate between vCenter and ESX hypervisor on host systems. It is also used for an administrative console via SSH for administrators to log into the ESC host operating system directly
- 2 x 1Gb interfaces for VMotion (initially only 1Gb supported from ESX due to 10Gb interface limitation - only 4x 10Gb interfaces are supported)

Figure 4-32 shows the example of the Intel xSeries virtualized server.

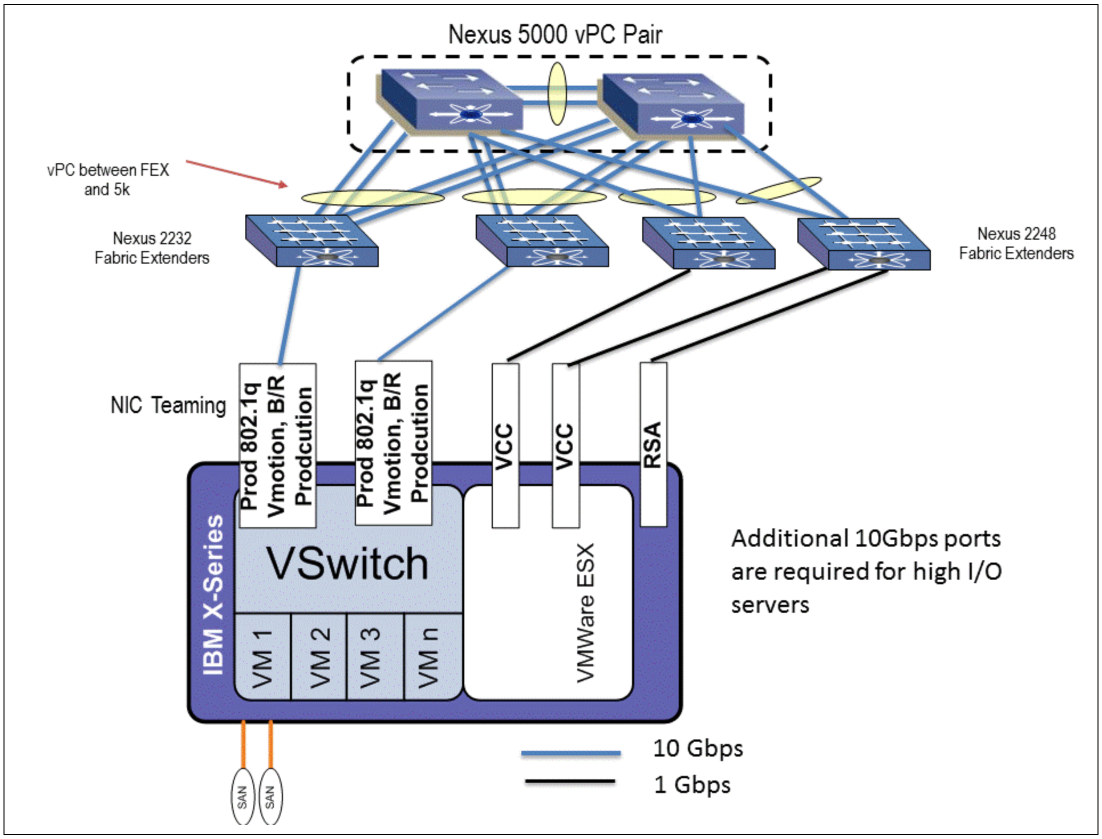

Nexus 5000 vPC Pair
vPC between FEX and 5k
Nexus 2232 Fabric Extenders
Nexus 2248 Fabric Extenders
NIC Teaming
Prod 802.1q Vmotion, B/R Prodcution
Prod 802.1q Vmotion, B/R Prodcution
VCC
VCC
RSA
IBM X-Series
VSwitch
VM 1
VM 2
VM 3
VM n
VMWare ESX
Additional 10Gbps ports are required for high I/O servers
SAN
SAN
10 Gbps
1 Gbps

Figure 4-32 Example of xSeries VmWare Server attached to the access layer.

System x Series Server Connectivity guidance:

- Since LACP was not supported on the ESX vSwitch prior to VMware 5.1, NIC teaming was typically used in IBM Data Center server builds.
- Production and production admin will be split between two physical switches
- BR interfaces will be split between two physical switches.
- VMotion and Management will be split between two physical interfaces.
- 802.1q trunking will be used to support multiple VLANs on a singleinterface.
- Jumbo frame will be used to support IBR interfaces.

Table 4-8 lists network port counts used as guidance to client architects for VmWare servers.

Table 4-8 Server guidance for virtualized server

Platform	**1 Gbps**	**10 Gbps**	**Fibre Channel 8 Gbps**
Intel virtualized	1 RSA/IMM	0	2 client production
VmWare		2 physical ports teamed together for redundancy Port 1 will be used for Client production data and backup to Port 2 Port 2 will be primary Vmotion and B/R and backup to port 1	2 LAN free if DB over 200Gb
~80 images per host	2 VC	0	2 additional for high I/O environments
Total	3	2	2-6

Following is a list of additional considerations for designing xSeries servers:

- Servers with a high number of databases may require more than the 6 Fibre Channel ports.
- Within the Intel environment, teaming should be used as a means to achieve redundancy not throughput.
- Each zone may require its own backup network.
- Physically separate cards should be used for redundant connections to separate switches.
- When teaming ports, best practice is to team onboard and PCI NICs within the same manufacturer.
- Service Level Agreements (SLAs) may drive a second HMC port requirement.
- If using LAN free, you do not need the copper backup/recovery ports.

An additional 1G bps port is required if using an Avocent KVM switch. Add two 1GigE ports for critical environments.

4.9 IBM System p

System p, formerly known as IBM RS/6000® (for RISC System/6000), is the current IBM RISC/UNIX-based server and workstation product line. It is a sub-brand of the Systems and Technology Group. System p machines mainly used the POWER5+, Power6, and Power7, but also IBM PowerPC® 970 on, for the low end and blade systems.

4.9.1 pSeries non-virtualized server

Figure 4-33 shows pSeries non-virtualized application servers. The server on the right is shown as having all 1 gigabit Ethernet interfaces. This would typically be a non-virtualized “legacy” application server. More than1 Gbps Ethernet cards may be used if higher I/O is required. Two 1 Gbps links are used for production. One 1 or 2 Gbps links is used for backup and one 1 Gbps link is used for the HMC (Host Management Console). However, the cross over point on cost between 1 Gbps and 10 Gbps ports tends to be at about 2 or 3 NICs.

The server on the left is similar in design to the application server, but shows all 10 Gbps interfaces except for the Host Management Console (HMC). Separate logical 10 Gbps interfaces are used for production and backup recovery. Multiple 10 Gbps Ethernet NICs may be used if higher I/O is required.

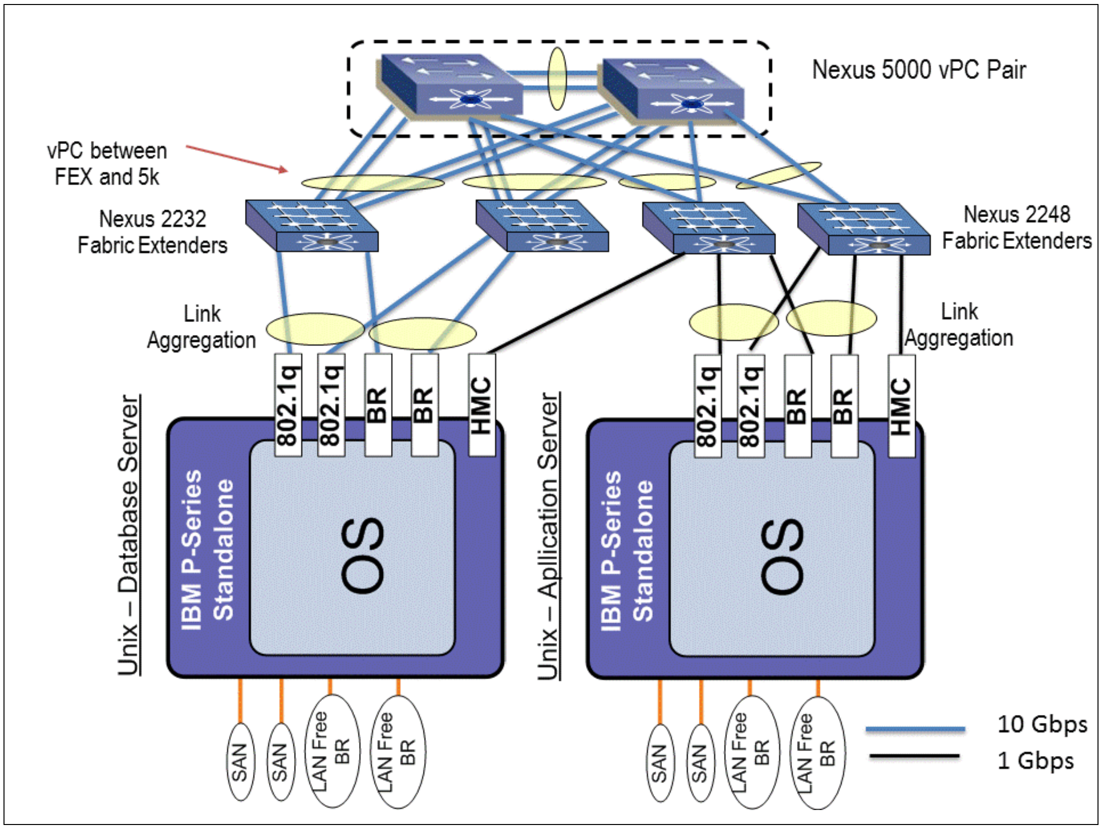


Figure 4-33 pSeries Non-Virtualized Servers

Table 4-9 lists network port counts used as guidance to client architects for application P-servers.

Table 4-9 Server guidance for non-virtualized application P-servers

Platform	**1 Gbps**	**10 Gbps**	**Fibre Channel 8 Gbps**
pSeries Stand alone servers,	1 HMC	0	2 client production
application, infrastructure	2 Production	0	
and web servers	1-2 Backup/recovery	0	2 additional for high I/O environments
Total	4-5	0	2-4

Table 4-10 lists network port counts used as guidance to client architects for database P-servers.

Table 4-10 Server guidance for non-virtualized database P-servers

Platform	**1 Gbps**	**10 Gbps**	**Fibre Channel 8 Gbps**
pSeries Database Servers	1 HMC	0	2 client production
		2 Production	2 LAN free if DB over 200Gb
		2 Backup/recovery	
Total	1	4	2-4

4.9.2 pSeries virtualized server

Power Hypervisor implements a virtual Ethernet switch to deal with traffic from virtual Ethernet adapters. Every virtual Ethernet adapter has a corresponding virtual switch port on the virtual Ethernet switch. The virtual Ethernet switch uses Virtual LANs (VLANs) to segregate traffic and the switch is consistent with IEEE 802.1q frame format.

The virtual Ethernet switch concept has the following benefits:

- Partitions “think” they have a dedicated Ethernet adapter.
- For pure inter-partition communication, no physical Ethernet adapters are needed.

Several VLANs may be defined on one virtual Ethernet adapter, one native VLAN (without a 802.1q tag) and up to 20 tagged VLANs.

The virtual Ethernet switch is not fully compliant with the 802.1Q specification since it does not support spanning-tree algorithms and will not participate in spanning-tree calculations. The hypervisor knows the MAC addresses of all virtual Ethernet adapters and thus can switch datagrams between the virtual Ethernet adapters if the adapters belong to the same VLAN.

Figure 4-34 shows the system p network components. VLANs may be configured the same as with real Ethernet switches. In Figure 4-34 packets can be exchanged between LPAR 1 and LPAR 2. LPAR 1 and LPAR 3 can also exchange packets. But LPAR 2 and LPAR 3 cannot exchange packets at the virtual Ethernet switch.

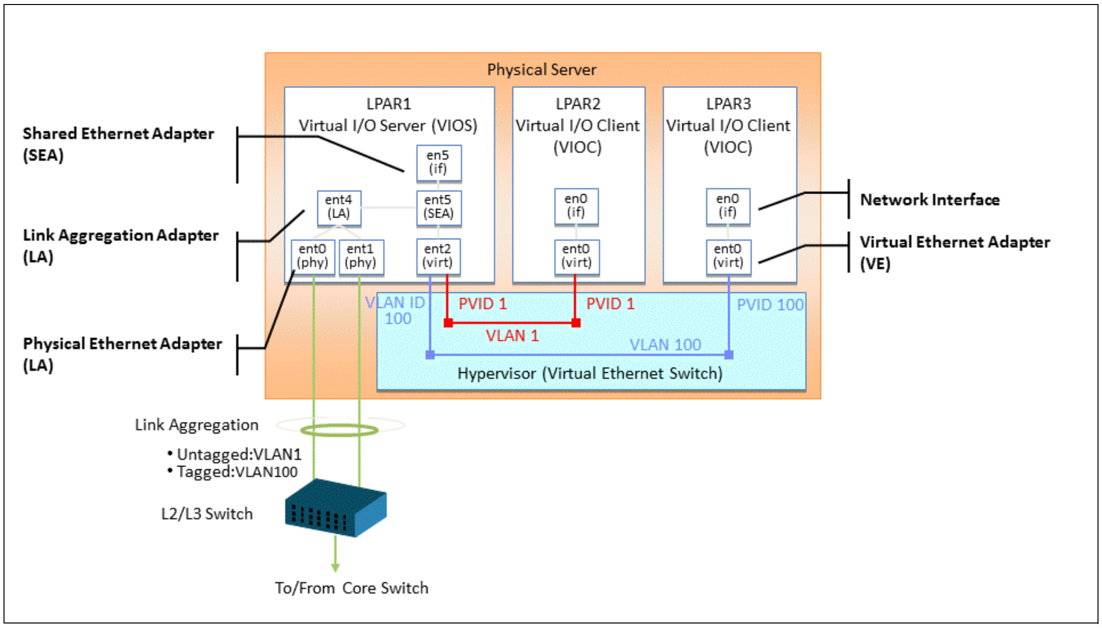


Figure 4-34 System p network components

By default, Power Hypervisor has only one virtual switch, but it can be changed to support up to 16 virtual Ethernet switches per system. Currently, the number of switches must be changed through the Advanced System Management (ASM) interface and the system must be rebooted. The same VLANs can be used on different virtual switches and traffic will still be separated by the Power Hypervisor.

To provide access for end users to the Power System, the virtual Ethernet switch needs to be connected to an external Ethernet switch. This is done through the Shared Ethernet Adapter (SEA) which provides bridging functionality between the virtual network and one or several physical adapters. The SEA provides the following functions:

- Datagrams from the virtual network can be sent on a physical network, and vice versa.
- One or more physical adapters can be shared by multiple virtual Ethernet adapters.
- Virtual Ethernet MAC is visible to outside systems.
- Broadcast and multicast datagrams are bridged.
- Both VLAN-tagged and untagged datagrams are bridged.

SEA has the following characteristics:

- A physical Ethernet adapter will not receive datagrams not addressed to itself. SEA puts the physical adapter in "promiscuous mode," where it receives all packets seen on the wire.
- Power Hypervisor will not give datagrams to a virtual Ethernet adapter if they are not addressed to its MAC address. Virtual Ethernet adapters that belong to the SEA must be created as "trunk" adapters: when the hypervisor sees a datagram destined for a MAC address it does not know about, it gives it to the trunk adapter.
- SEA does IPv4 fragmentation when a physical adapter's MTU is smaller than the size of a datagram.
- If the Virtual Input/Output Server (VIOS) fails (crashes, hangs, or is rebooted), all communication between the virtual and the physical networks ceases. SEA can be placed in a redundant failover mode.

All virtual Ethernet adapters in a Shared Ethernet Adapter (SEA) must belong to the same virtual switch, or SEA creation will fail.

To avoid single points of failure, the SEA can be configured in redundant failover mode; see Figure 4-35. This mode is desirable when the external Ethernet network can provide high availability. Also, more than one physical adapter is needed at the POWER System.

Primary and backup SEAs communicate through a control channel on a dedicated VLAN visible only internally to the IBM POWER Hypervisor™. A proprietary protocol is used between SEAs to determine which is the primary and which is the backup.

One backup SEA may be configured for each primary SEA. The backup SEA is idle until a failure occurs on the primary SEA; failover occurs transparently to client LPARs.

Failure may be due to the following conditions:

- Physical adapter failure
- VIOS crash/hang/reboot
- Inability to ping a remote host (optionally)

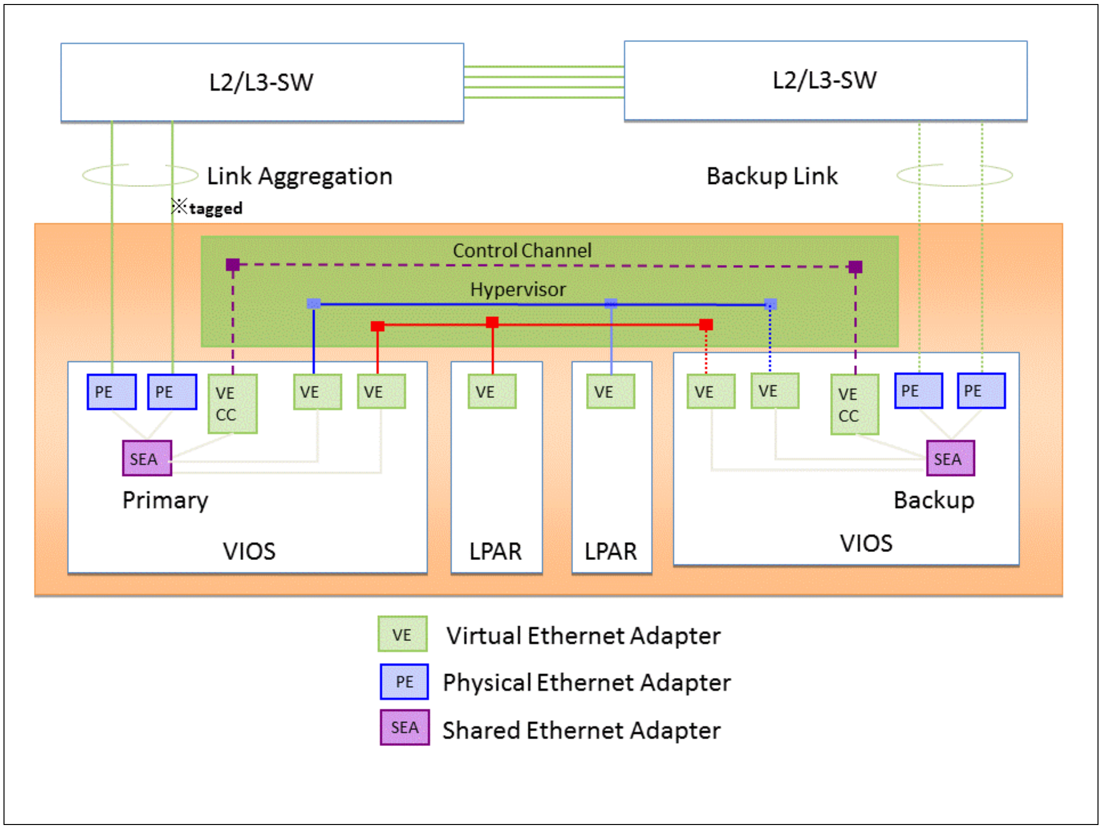


Figure 4-35 SEA in redundant failover mode

Each SEA in a failover domain has a different priority. A lower priority means that an SEA is favored to be the primary. All virtual Ethernet adapters on an SEA must have the same priority. Virtual Ethernet adapters of primary and backup SEAs must belong to the same virtual switch.

SEAs also supports EtherChannel and IEEE 802.3ad solutions, that is, several Ethernet adapters are aggregated to form one virtual adapter. The adapter aggregation is useful to address increased bandwidth demands between the server and network or in order to improve resilience with the adjacent Ethernet network, especially when a Multichassis EtherChannel solution is implemented in the Ethernet network.

SEA supports four hash modes for EtherChannel and IEEE 802.3ad to distribute the outgoing traffic:

- default mode - The adapter selection algorithm uses the last byte of the destination IP address (for TCP/IP traffic) or MAC address (for ARP and other non-IP traffic). This mode is typically the best initial choice for a server with a large number of clients.
- src_dst_port - The outgoing adapter path is selected through an algorithm using the combined source and destination TCP or UDP port values: Average the TCP/IP address suffix values in the Local and Foreign columns shown by the **`netstat -an`** command.

  Since each connection has a unique TCP or UDP port, the three port-based hash modes provide additional adapter distribution flexibility when there are several separate TCP or UDP connections between an IP address pair. Src_dst_port hash mode is recommended when possible.

- src_port - The adapter selection algorithm uses the source TCP or UDP port value. In **netstat -an** command output, the port is the TCP/IP address suffix value in the Local column.
- dst_prt - The outgoing adapter path is selected through an algorithm using the destination system port value. In **netstat -an** command output, the TCP/IP address suffix in the Foreign column is the TCP or UDP destination port value.

SEA supports a fifth hash mode, but it is only available in EtherChannel mode. When used, the outgoing datagrams are scheduled in a round-robin fashion, that is, outgoing traffic is spread evenly across all adapter ports. This mode is the typical choice for two hosts connected back-to-back (that is, without an intervening switch).

Network Interface Backup

AIX supports a backup adapter capability for EtherChannel and 802.3ad. Figure 4-36 depicts the Network Interface Backup (NIB). NIB is possible with two adapters (primary/backup). A virtual adapter is formed over NIB, and the IP address is configured on the virtual adapter.

If the primary adapter fails (for example, a cable is unplugged), NIB switches to the backup adapter, maintaining connectivity. Optionally, an IP address to ping can be specified to detect network faults further downstream.

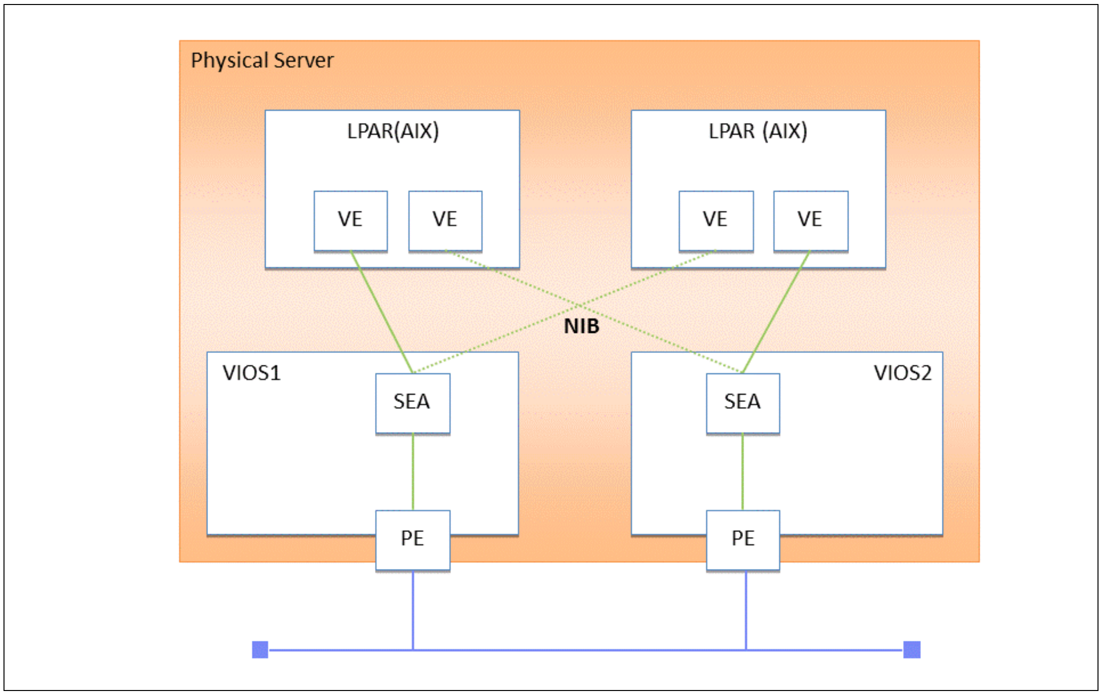


Figure 4-36 Network Interface Backup (NIB)

Link aggregation can be used with virtual network, as well. Figure 4-37 shows redundancy with Dual VIOS, network interface backup/link, and link aggregation between VIOS and external switch.

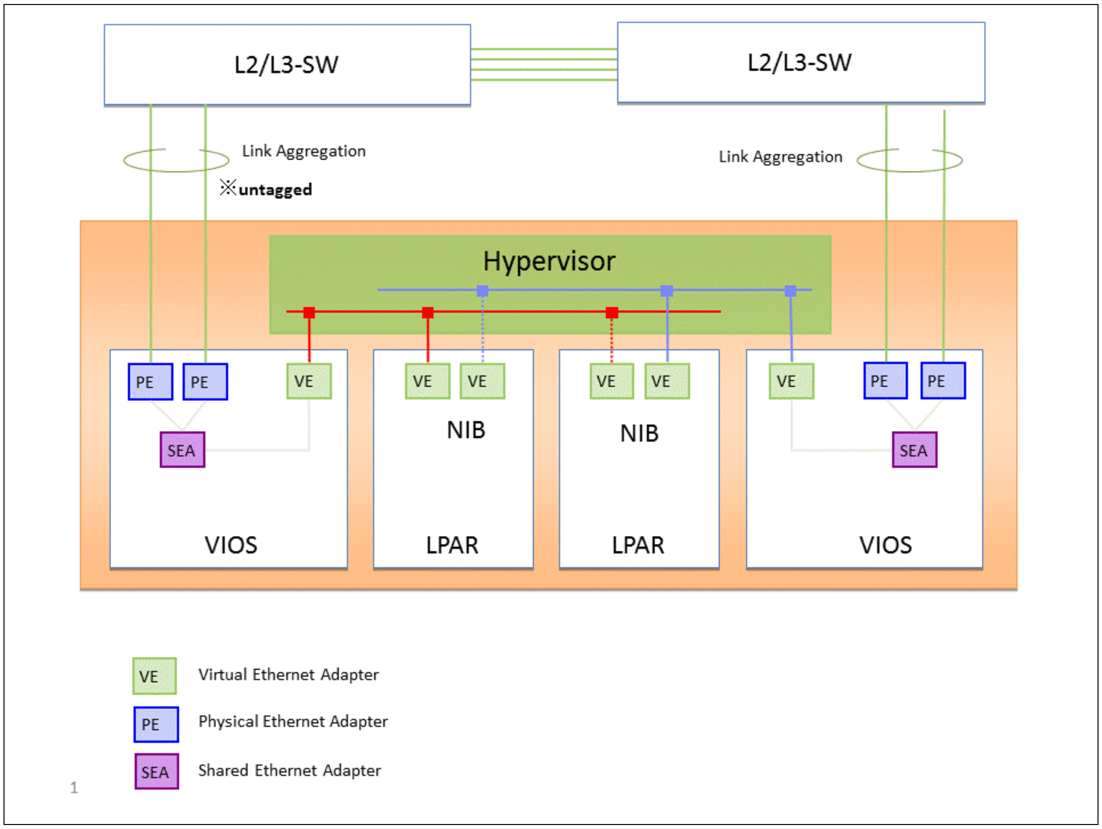


Figure 4-37   Dual VIOS/network interface backup/link aggregation

System pSeries server setup

Figure 4-38 shows an example of pSeries virtualized server physical connectivity.

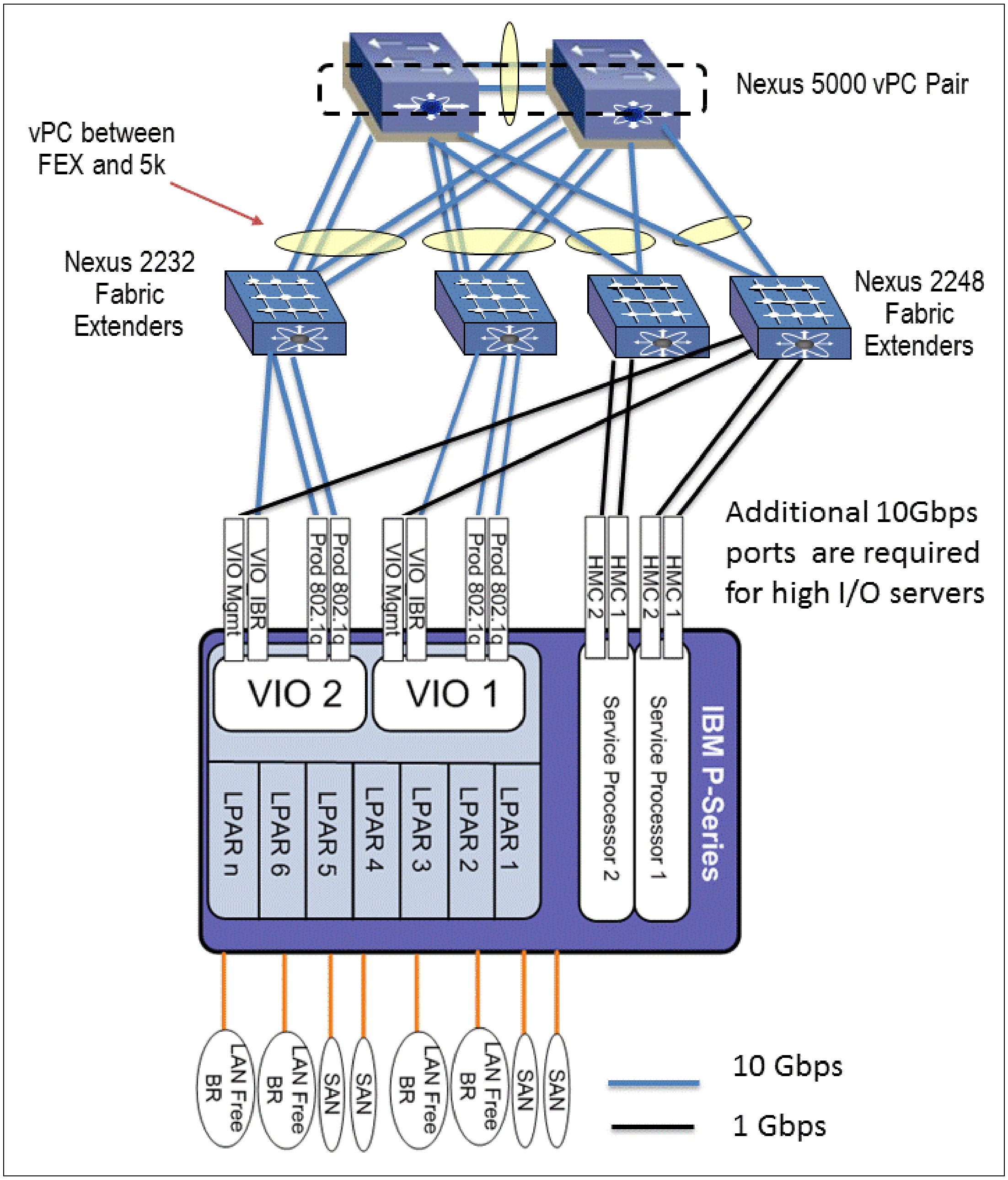


Figure 4-38 pSeries Virtualized Server Physical Connectivity

Example of two p770 in one rack as 8 Central Electronic Complexes (CeC)

- Each p770 has two VIO servers for LAN connectivity
- Each VIO server requires the following interfaces:
  - 1 x 10Gb interfaces for production / admin
  - 1 x 10Gb interface for production backup with Jumbo frame support
  - 1 x 1Gb interface for VIO management (live partition mobility)
  - 1 x 1Gb interface for VIO backup with jumbo frame support
- 4 x 1Gb service processor interfaces for redundant HMC connectivity

Table 4-11 shows the network port counts used as guidance to client architects for application P-servers.

Table 4-11 Server guidance for virtualized P-servers

Platform	1 Gbps	10 Gbps	Fibre Channel 8 Gbps
Virtualized pSeries environment ~12 LPARs per Server Database (non RAC or Purescale)	1-2 HMC	2 client production	6 client production
	1-2 VIO mgmt	2 for b/r	2 LAN free b/r for DBs over 200 GB. Not required if < 200GB
		0	
Total	2-4	4	8

Here is a list of additional considerations for designing virtualized pSeries servers:

- Servers with a high number of databases may require more than the 6 Fibre Channel ports.
- Each zone may require its own backup network.
- Physically separate cards should be used for redundant connections to separate switches.
- When teaming ports, best practice is to team onboard and PCI NICs within the same manufacturer.
- Service Level Agreements (SLAs) may drive a second HMC port requirement.
- If using LAN-free, you do not need the copper backup / recovery ports.
- Be sure to include ports for VIO servers, one per VIO server. You only need one, as the VIO servers can be reached via the HMC giving you redundant connection to the VIO servers.
- An additional 1 Gbps port is required if using an Avocent KVM switch. Add two 1 GigE ports for critical environments.

5

Cisco virtual networking solutions

This chapter explains the concept of Cisco’s virtual networking solutions. As with many of today’s virtualization techniques, this introduces additional complexity, re-definitions of borders, and new operational concepts.

5.1 The impact of server virtualization to the data center network

Server virtualization and the use of virtual machines is profoundly changing data center dynamics. Most organizations are struggling with the cost and complexity of hosting multiple physical servers in their data centers. The expansion of the data center, a result of both scale-out server architectures and traditional “one application, one server” sprawl, has created problems in housing, powering, and cooling large numbers of under-utilized servers.

In addition, IT organizations continue to deal with the traditional cost and operational challenges of matching server resources to organizational needs that seem ever changing. Virtual machines can significantly mitigate many of these challenges by enabling multiple application and operating system environments to be hosted on a single physical server while maintaining complete isolation between the guest operating systems and their respective applications. Hence, server virtualization facilitates server consolidation by enabling organizations to exchange a number of under-utilized servers for a single highly utilized server running multiple virtual machines.

5.1.1 Where is the network edge?

In data centers worldwide, the amount of data and the complexity of the infrastructure is expanding and managing the increasing data center resources has become cumbersome. Virtualization can even increase this problem, stimulating the dramatic growth in the number of server instances and adding complexity to the management of virtualized data center.

As data centers pack in more servers to support virtualization and boost application performance, they also must add more network switches. With virtualization, a physical server runs multiple virtual servers, called virtual machines (VMs); each VM requires a connection to a virtual switch (vSwitch). The virtual switches are connected to the physical network to allow communication among the VMs and the external network.

An increasingly important part of the data center infrastructure is the virtual network edge where the physical and virtual areas overlap. The virtualized network extends into the physical servers where the virtualized servers define the new demarcation between network and server. As a consequence, the traditional edge between server and network is being redefined with major impacts on the management of both infrastructures.

Figure 5-1 illustrates the network edge for a data center.

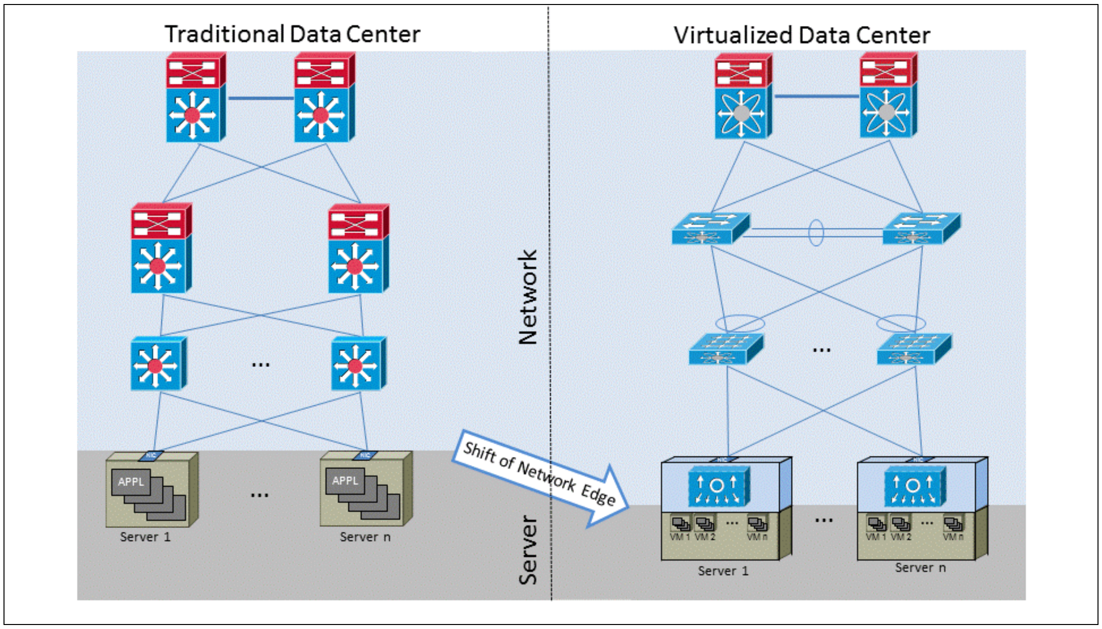


Figure 5-1 The network edge of a virtualized data center is moving into the server infrastructure

The virtual network edge architecture is composed of both server and network elements. It overlaps the server-network edge of the traditional data center. This allows for the dynamic creation, deletion, and movement of network resources in support of the dynamic creation, deletion, and movement of virtualized servers.

Traditionally, in data centers, each class of technology had been controlled by a different set of experts with different knowledge areas, each with their own processes for deploying new infrastructure or implementing changes.

Now, the introduction of server virtualization and Layer 2 bridges within the server has created the need for a higher-level coordination between server and network managers.

5.1.2 Challenges to scale server virtualization

Best-practice design, power and cooling load, and rack space typically limit server density within a data center. Although blade servers tend to increase density within the data center, virtual machines have a far more profound effect as they can be created as required with little additional “cost”. A data center that hosts 2000 physical servers, could, as a result of implementing virtual machines at a 10:1 density, radically increase the demands on the network infrastructure. Mathematically, the worst-case load is 20,000 servers. In reality, the load would be less because some physical servers will be taken offline as part of consolidation, and not all applications can be moved to a virtual machine environment. On the other hand, the availability of virtual machines and available server cycles makes additional virtual machines more likely to be brought online.

Regardless of the exact consolidation ratio, experience shows that the net server count increases. Additionally, because virtual machine creation is so easy, the rate of growth in the number of servers also increases in comparison to traditional models where procurement and provisioning requirements serve to limit the rate at which servers can be added to the data center.

Virtual machines increase the density of server identities, which places additional demands on the network fabric, especially when the dynamic creation and migration of virtual machines is considered. This stresses multiple network components in the following ways:

- Increased identity density: MAC and IP addresses and applications
- Increased control plane overhead because of increased service density
- Increased high-availability requirements because of increased risk exposure
- Increased services, such as firewalls and load balancers, which affects performance

The result is that both the growth and the rate of growth place stress on the data forwarding capacity of the network, including services modules, and also on the control-plane aspects of the network devices.

Layer 2 fabric scalability and integrity

Because consolidation and virtualization concentrates applications and their related business operations, they can significantly increase the business risk associated with a switch or data center domain failure. As noted earlier, migration to a dense virtual machine environment can stress a network in terms of both scale and volume of traffic. Thus, in dense virtual machine environments, the requirement to protect network switches from being overwhelmed by CPU bound traffic, both legitimate and malicious, is a crucial design goal for maintaining application availability and deterministic performance.

In addition to a high-speed application-specific integrated circuit (ASIC) forwarding engine, all networking devices have a central CPU for processing control functions such as network control protocols (ARP, spanning tree protocol, routing protocols, and so on), packets destined for the device (Simple Network Management Protocol [SNMP], Telnet, and so on), and non-hardware switchblade packets such as those that have exceeded the maximum time to live (TTL) or those that have IP options set.

The increase in server density within a switched domain places stress not only on the forwarding and buffering performance of the switch, but also on the switch control plane. This additional stress is the result of the increased number of servers associated with the data center network and their increased volume of traffic.

As the density of servers increases, the number of MAC and IP addresses that need to be supported increases. If a data center network of 1000 physical servers is considered, this can easily mean 24,000 to 32,000 MAC and IP addresses due to virtual machines that need to be supported. Although from a forwarding perspective, this support can be achieved easily with modern data center switches, the volume of MAC addresses and control-plane-bound traffic can still be considerable.

Because virtual machines are highly mobile, the data center network mechanisms in place must support the mobility of associated Layer 2 services. From a design perspective, as VLANs are extended between switches within the data center network, spanning tree protocol is required to help ensure that loops do not occur within the topology, and the switches must use MAC learning to determine where a particular MAC address is within the switched network. VMware ESX actively updates the Layer 2 forwarding databases of all switches on the LAN by sending gratuitous ARP whenever a VMotion event occurs.

Under most circumstances, this self-learning behavior has a fairly low overhead. However, when a spanning-tree event occurs, self-learning can introduce unique challenges in the data center aggregation switches. When a topology change occurs within a spanning-tree-based network, the core MAC addresses are typically flushed to prevent stale forwarding information from causing traffic to not be forwarded. Although flushing resolves this problem, if the switch does not know through which port the destination MAC address is available, the switch will flood all traffic associated with that address until the destination-to-port association is relearned.

As most MAC-address-to-port associations are flushed during a spanning-tree event, the rate at which MAC addresses are relearned is a critical metric, especially in dense virtual machine environments. Because potentially high volumes of unicast traffic may be flooded, the effect on network links, buffers, virtual switches, and application performance can be significant.

Because unknown destination unicast traffic is flooded on all L2 forwarding links, the links may be overwhelmed and packets buffered at egress and ingress to the switch. To protect against control traffic being dropped, which would destabilize the network, Cisco Catalyst and Nexus switches priorities control plane packets on a per-port basis.

As the flooded unicast traffic is received at the VM-enabled servers, the virtual switch needs to inspect traffic that in most cases will be dropped. This takes away from the applications both bandwidth and CPU cycles, which can cause unpredictable application response times and performance until the switch's MAC address table is re-populated. Because practically, virtual machine mobility requires Layer 2 adjacency, the switch MAC address learning rate is a critical factor in dense virtual machine environments. For example, a data center module of 1000 servers with 8 virtual machines per server will have 8000 MAC addresses. If a LAN switch can learn only 1000 MAC addresses per second, then traffic may be flooded for up to 8 seconds until all MAC-to-port associations can be relearned.

Note: VMotion itself does not require Layer 2 adjacency. However, the virtual machine being migrated should be moved within the same VLAN, unless host routes are being generated to help ensure that user sessions are reestablished to the virtual machine.

Challenges in the cloud-based data center

As virtualization continues to push the limits of the infrastructure and the tools that control and manage the devices, configurations, and services, it is apparent that a new approach is required to support the day-to-day demands of a highly virtualized dynamic network infrastructure. This need is even more evident in the cloud computing models where the demand on network infrastructure is more intensive and being driven by the "instant-on," real-time delivery of virtualized computing/storage services.

In cloud-based service delivery, in which the access model defines connectivity to the cloud as well as the security model that defines how that access and data are protected, relationships are created dynamically and deployed automatically as a "preapproved" change as the service is requested. No longer are teams of IT and network engineers able to design, test, and roll out infrastructure services within a traditional change control model process over a period of weeks or months. Cloud services are turned on and off like a light bulb based solely on end-user need. The infrastructure that provides these services, virtualized computation, storage, and the underlying network infrastructure, must be able to respond to that "flick-of-the-switch" automatically, predictably, without error, and without exposing other services or the entire enterprise to impacts or security risk.

Refer to this white paper that explains in more detail the requirements and challenges of a cloud-based data center:

http://www.cisco.com/en/US/prod/collateral/netmgtsw/ps6505/ps11636/white_paper_c11-693135.pdf

5.2 Cisco virtual networking

Cisco has extended the networking to become VM-aware so that networking services, policies, and operational models can be provided to Virtual Machines (VM) in a similar way that they are provided to physical servers.

Cisco virtual networking solutions address the unique requirements that server virtualization bring to networking by enabling network and security policies to follow VMs when they are moved between physical servers within or between data centers. It also provides to the networking team visibility at the individual VM level that otherwise would not be possible if the networking team is looking at the physical layer where the traffic from multiple VMs is presented combined as a trunk port.

One of the benefits of utilizing Cisco virtual networking solutions is that it significantly simplifies the virtual networking environment by centralizing the management and control plane while keeping a distributed data plane. It also enables the provisioning of the virtual networking to be based on pre-defined policies.

Cisco virtual networking solutions are based on the following principles:

- Supporting multi-hypervisors, including open source hypervisors. This reduces the complexity and allows a consistent end-to-end network solution even when multiple hypervisors are used on the environment.
- Offering multi-virtual network services that are easily inserted on the data path of VMs.
- Support for multi-cloud environments; private, public, and hybrid clouds.
- Offering consistency between the physical and virtual environments.
- Respect for separation of duties between server and network teams.

5.3 Nexus 1000v

The Cisco Nexus 1000v for VMware vSphere is a distributed Layer 2 virtual switch implemented based on IEEE bridging standard that is fully integrated with VMware Virtual Infrastructure, including VMware vCenter. The Cisco Nexus 1000v is an advanced software switch for VMware vSphere, providing switching for up to 64 VMware ESX hosts from a single point of management.

The Cisco Nexus 1000v manages a data center object defined by the vCenter server. This virtual data center can span across multiple physical data centers. Each server in the data center is represented as a line card in the Cisco Nexus 1000v and can be managed as if it were a line card in a physical Cisco switch.

The Nexus 1000v runs Cisco NX-OS Software and supports advanced features such as QoS, access control lists (ACLs), Private VLAN (PVLAN), Link Aggregation Control Protocol (LACP), Encapsulated Remote Switched Port Analyzer (ERSPAN), NetFlow, control-plane security features, and integration of network services to the VMs that connect to its virtual Ethernet (vEthernet) ports.

When deployed, the Cisco Nexus 1000v not only maintains the virtualization administrator's regular workflow, it also offloads the vSwitch and port group configuration to the network administrator, reducing network configuration mistakes. With the Cisco Nexus 1000v series you can have a consistent networking feature set and provisioning process all the way from the virtual machine access layer to the core of the data center network infrastructure. Virtual servers can have the same network configuration, security policy, diagnostic tools, and operational models as their physical server counterparts attached to dedicated physical network ports.

The servers that run the Cisco Nexus 1000v must be in the VMware hardware compatibility list. At time of writing the latest version of the Nexus 1000v was 4.2(1)SV1(5.2) that supports vSphere 4.1.0, 5.0.0, and 5.1.0 releases.

The Cisco Nexus 1000v is compatible with any upstream physical switch, including all Cisco Nexus and Cisco Catalyst switches as well as Ethernet switches from IBM and other vendors.

See the Cisco Nexus 1000v release notes page for information about the latest version of the Nexus 1000v:

http://www.cisco.com/en/US/products/ps9902/prod_release_notes_list.html

5.3.1 Essential Edition and Advanced Edition

Cisco Nexus 1000v Series Switch is available in an Essential Edition, which is free; and an Advanced Edition, which is paid for and includes Cisco Virtual Security Gateway.

The Essential Edition provides all the Layer 2 networking features required to connect VMs to the network, including these:

- Layer 2 switching: VLANs, private VLANs, VXLAN, loop prevention, multicast, virtual PortChannels, LACP, ACLs
- Network management: SPAN, ERSPAN, NetFlow 9, vTracker, vCenter Server plug-in
- Enhanced QoS
- Cisco vPath for services insertion

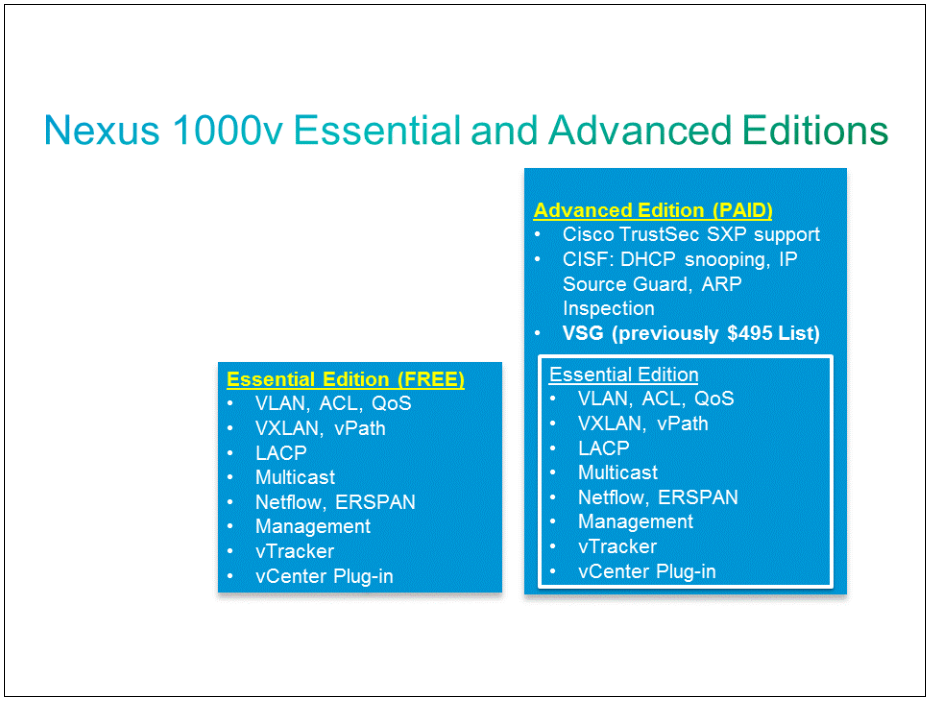
In addition to these features, Advanced Edition supports the following security features:

- DHCP snooping
- IP source guard
- Dynamic ARP inspection
- Cisco TrustSec SGA support

The Cisco Virtual Security Gateway is also included in the Advanced Edition.

Both editions integrate with Cisco ASA 1000v, Cisco vWAAS, and future virtualized networking and security services.

Figure 5-2 illustrates the difference between the two editions.



Nexus 1000v Essential and Advanced Editions

Advanced Edition (PAID)

- Cisco TrustSec SXP support
- CISF: DHCP snooping, IP Source Guard, ARP Inspection
- **VSG (previously $495 List)**

Essential Edition

- VLAN, ACL, QoS
- VXLAN, vPath
- LACP
- Multicast
- Netflow, ERSPAN
- Management
- vTracker
- vCenter Plug-in

Essential Edition (FREE)

- VLAN, ACL, QoS
- VXLAN, vPath
- LACP
- Multicast
- Netflow, ERSPAN
- Management
- vTracker
- vCenter Plug-in

Figure 5-2 Comparison of Nexus 1000v Editions

5.3.2 Nexus 1000v architecture

Cisco Nexus 1000v has two major components: the Virtual Ethernet Module (VEM), which runs inside the hypervisor, and the external Virtual Supervisor Module (VSM), which manages the VEM, as shown in Figure 5-3.

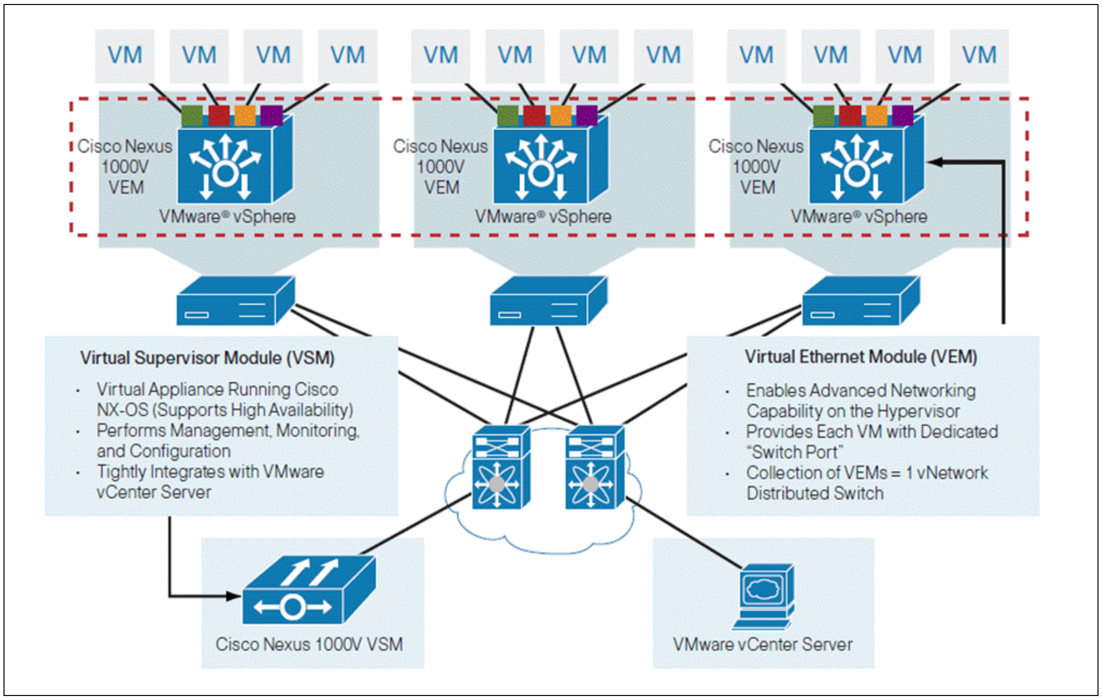

VM VM VM VM
VM VM VM VM
VM VM VM VM
Cisco Nexus 1000V VEM
VMware® vSphere
Cisco Nexus 1000V VEM
VMware® vSphere
Cisco Nexus 1000V VEM
VMware® vSphere
Virtual Supervisor Module (VSM)
Virtual Appliance Running Cisco NX-OS (Supports High Availability)
Performs Management, Monitoring, and Configuration
Tightly Integrates with VMware vCenter Server
Virtual Ethernet Module (VEM)
Enables Advanced Networking Capability on the Hypervisor
Provides Each VM with Dedicated "Switch Port"
Collection of VEMs = 1 vNetwork Distributed Switch
Cisco Nexus 1000V VSM
VMware vCenter Server

Figure 5-3 The Nexus 1000v architecture

The Nexus 1000v VEM runs as part of the VMware ESX or ESXi kernel and replaces the VMware virtual switch (vSwitch). A VEM needs to be installed on each ESX or ESXi host and each ESX/ESXi host can only have one VEM.

This level of integration helps to ensure that the Cisco Nexus 1000v series is fully aware of all server virtualization events, such as VMware vMotion and Distributed Resource Scheduler (DRS).

The VEM takes the configuration information from the VSM and performs these Layer 2 switching and advanced networking functions:

- PortChannels
- Quality of service (QoS)
- Security: Private VLAN, access control lists (ACLs), and port security
- Monitoring: NetFlow, Switch Port Analyzer (SPAN), and Encapsulated Remote SPAN (ERSPAN)

In the event of a loss of communication with the VSM, the VEM has Nonstop Forwarding (NSF) capability to continue to switch traffic based on the last known configuration.

All interfaces on the VEM are Layer 2 interfaces and include access ports, trunk ports, private VLAN, and promiscuous ports.

Virtual Ethernet interfaces (vEthernet) are configured and behave similarly to physical Ethernet interfaces, however, usually port profiles (refer to port profile section) are used to simplify their configuration.

The VMs connect to the VEM Virtual Ethernet Interfaces in the same way that a physical server connects to a Physical Ethernet interface on a physical access switch. The Virtual Ethernet (vEthernet or vEth) interfaces are the logical interfaces on the VEM that are used to connect VMs to the network.

If you do not add an interface description to the vEthernet interface, then one of the following descriptions is added at the time when the VM is initiated.

- For a VM: VM-Name, Network Adapter number
- For a VMK: VMware VMkernel, vmk number
- For a VSWIF: VMware Service Console, vswif number

Those descriptions are automatically passed from the VMware vCenter directly into the Nexus 1000v. This is only possible because of the integration between the Nexus 1000v and VMware vCenter. It allows the network administrator to quickly identify what is connected to the virtual Ethernet interface.

The following example shows how to configure a vEthernet access interface and assign access VLAN 5 for that interface:

```
n1000v# configure terminal
n1000v(config)# interface vethernet 100
n1000v(config-if)# switchport
n1000v(config-if)# switchport mode access
n1000v(config-if)# switchport access vlan 5
```

The configuration steps described above are performed by the network administrator by connecting (TELNET, SSH) to the Nexus 1000v virtual supervisor module.

The connection from the Nexus 1000v VEM (Virtual Ethernet Module) to the upstream physical network is done by using the physical Ethernet interfaces on the ESX or ESXi server.

Those physical NICs (Network Interface Cards) are presented to the Nexus 1000v as uplink ports and similarly to what is done when connecting two physical switches, the network administrator usually configure those uplinks from the Nexus 1000v VEM (Virtual Ethernet Module) as a trunk ports and enable Link Aggregation Control Protocol (LACP).

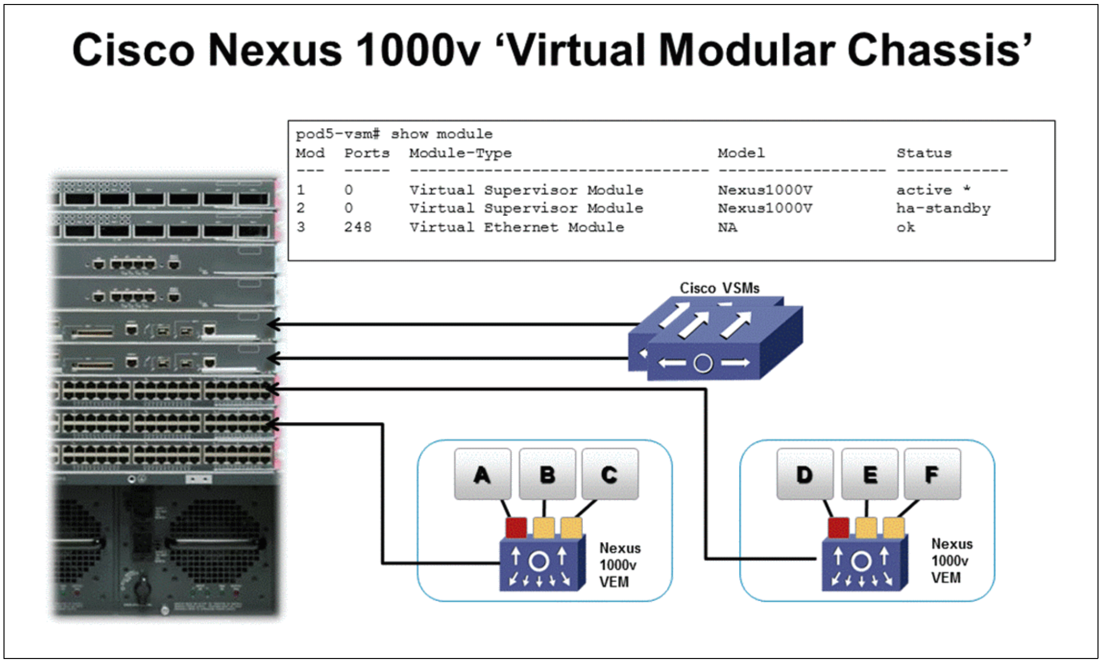
Link Aggregation Control Protocol (LACP) is a protocol used to control the bundling of several physical ports together to form a single logical port-channel. The use of LACP on the Nexus 1000v has the following advantages over the static configuration mode:

- Failover when a link fails and there is (for example) a Media Converter between the devices which means that the peer will not see the link down. With static link aggregation the peer would continue sending traffic down the link causing it to be lost, LACP would prevent this from happening.
- The Nexus 1000v and the physical upstream switch can confirm that the configuration at the other end can handle link aggregation. With static link aggregation a cabling or configuration mistake could go undetected and cause undesirable network behavior.

5.3.3 Components: Virtual Supervisor Module and Virtual Ethernet Module

The Virtual Supervisor Module (VSM) on a Nexus 1000v functions in a similar way to a Supervisor module on a modular switch (Catalyst 6500 or Nexus 7000 are examples of modular switches). It controls up to 64 Virtual Ethernet Modules (VEMs), or virtual line cards, as one logical modular switch.

Figure 5-4 illustrates how the components (VSM and VEM) of the Nexus 1000v compare with the components of a physical modular switch.



Cisco Nexus 1000v ‘Virtual Modular Chassis’

```
pod5-vsm# show module
Mod  Ports  Module-Type                      Model              Status
---  -----  -------------------------------- ------------------ ------------
1    0      Virtual Supervisor Module        Nexus1000V         active *
2    0      Virtual Supervisor Module        Nexus1000V         ha-standby
3    248    Virtual Ethernet Module          NA                 ok
```

Figure 5-4 The Cisco Nexus 1000v Virtual Modular Chassis

The configuration of the Nexus 1000v is performed through the VSM and is automatically propagated to the VEMs.

Instead of configuring virtual switches inside the hypervisor on a host-by-host basis administrators can define configurations for immediate use on all VEMs being managed by the VSM from a single interface.

This capability to centrally configure the network via the VSM, and have the network (that is, VLAN) information automatically propagated to up to 64 servers significantly simplifies the administration of the virtual network when compared with the practice commonly adopted today where the server administrator needs to configure each virtual switch (vSwitch) individually. For example, on a server farm with 32 servers, a VLAN can be centrally configured and associated with a port-profile once on the VSM in comparison with the server administrator having to configure the same port-group 32 times – one per individual server / individual vSwitch.

The VSM is configured and managed in a similar way to a Cisco physical switch and is accessible via Cisco Command-Line Interface (CLI), Simple Network Management Protocol (SNMP), XML API, CiscoWorks LAN Management Solution (LMS), and Data Center Network Manager (DCNM). The network team can use Telnet or SSH to connect to the VSM.

The VSM can be deployed as a virtual machine on a VMware ESX or ESXi host. In this case the VSM is simply a VM running NX-OS as its operating system. Alternatively, the VSM can be deployed as a Virtual Appliance on the Nexus 1010 or (refer to Nexus 1010 /1110 section). Figure 5-5 illustrates the two deployment options for the VSM.

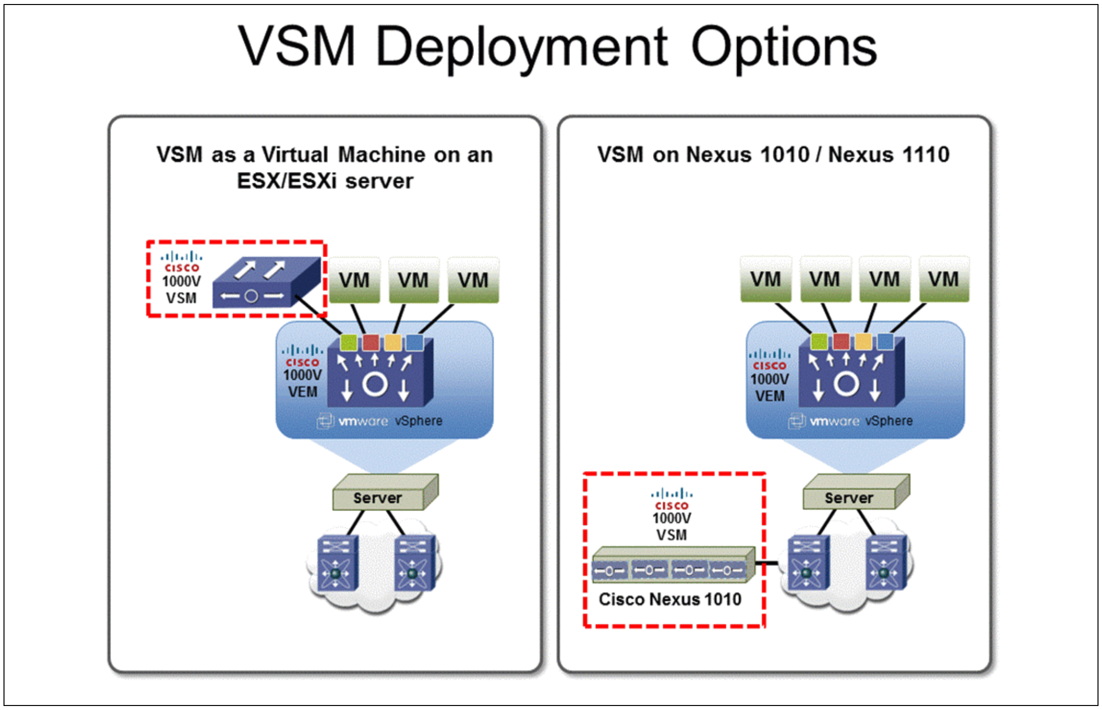


Figure 5-5 VSM deployment options

The Cisco Nexus 1000v supports redundant Virtual Supervisor Modules (VSMs). A primary and a secondary VSM are running as a high availability (HA) pair. Dual VSMs operate in an active/standby mode in which only one of the VSMs is active at any given time, while the other acts as a standby backup. The VSMs are configured as either primary or secondary as a part of the Cisco Nexus 1000v installation. The state and configuration remain constantly synchronized between the two VSMs to provide a stateful switchover if the active VSM fails. The VSMs use a VSM to VSM heartbeat to broadcast their presence, detect the presence of another VSM, and negotiate active and standby redundancy states.

5.3.4 VSM-to-VEM communication

The VSM and the VEM can communicate over a Layer 2 network or a Layer 3 network. These configurations are respectively referred to as Layer 2 or Layer 3 control mode.

Layer 3 control mode

The VEMs can be in a different subnet than the VSM and also from each other in Layer 3 control mode. Active and standby VSM control ports should be Layer 2 adjacent. These ports are used to communicate the HA protocol between the active and standby VSMs.

Each VEM needs a designated VMkernel NIC interface with an associated IP address that communicates with the VSM. This interface, which is called the L3 Control vmknic, must have a system port profile applied to it so the VEM can enable it before contacting the VSM.

In Layer 3 control mode the UDP port 4785 is used for Layer 3 communication between the VSM and VEM. If you have a firewall in your network, and are configuring Layer 3 control, then make sure that UDP port 4785 is open on your upstream switch or firewall device.

Layer 2 control mode

The VSM and VEM are in the same subnet in the Layer 2 control mode. In this deployment mode the VSM and VEM have access to a shared VLAN and utilize it to communicate directly.

Note: Layer 3 control mode is the preferred method of communication between the VSM and VEM.

Loss of communication between VSM and VEM

For the period of time that the VSM and VEM cannot communicate (headless mode of operation), the VEM has Nonstop Forwarding (NSF) capability to continue to switch traffic based on the last known configuration.

Once the connectivity is restored and the VSM-VEM communication is re-established, the system will return to its previous operational state.

LACP offload to the VEM

Traditionally, LACP is processed on the control plane of a switch (the supervisor or VSM). Because the Cisco Nexus 1000v Series has a distributed model in the virtualization environment, new problems may arise when running LACP on the VSM in this way.

For example, if the VEM operates without the presence of the VSM, also referred to as headless mode, whenever a link flap or reboot of the VEM occurs and there is no communication between the VEM and VSM, an LACP bundle will not form. The bundle will not form because the VSM is responsible for originating and processing the LACP control packets, also referred to LACP protocol data units (PDUs), needed to negotiate an LACP PortChannel.

To address the possible problem described above, by default the Nexus 1000v supports offloading the operation of the LACP protocol from the VSM to the VEMs. The origination and processing of LACP PDUs is completely offloaded to the VEM. This feature should be used so that the VEM has no dependence on VSM connectivity while negotiating an LACP PortChannel. This capability makes the overall deployment of the Cisco Nexus 1000v Series solution more robust.

5.3.5 Port profiles: Policy-based virtual machine connectivity

Port profiles are the primary mechanism by which network policy is defined and applied to the interfaces on the Nexus 1000v. Port profiles are a collection of interface configuration commands that can be dynamically applied at either physical or virtual interfaces. Any changes to a given port profile are propagated immediately to all ports that have been associated with it.

A port profile can define a sophisticated collection of attributes such as VLAN, private VLAN (PVLAN), ACL, port security, NetFlow collection, rate limiting, QoS marking, and even remote-port mirroring (through Encapsulated Remote SPAN [ERSPAN]).

An example of a port profile configuration is shown here:

```
(config)# port-profile webservers
(config-port-prof)# switchport access vlan 10
(config-port-prof)# ip access-group 500 in
(config-port-prof)# inherit port-profile server
```

The port profile can then be assigned to a given vEth interface as follows:

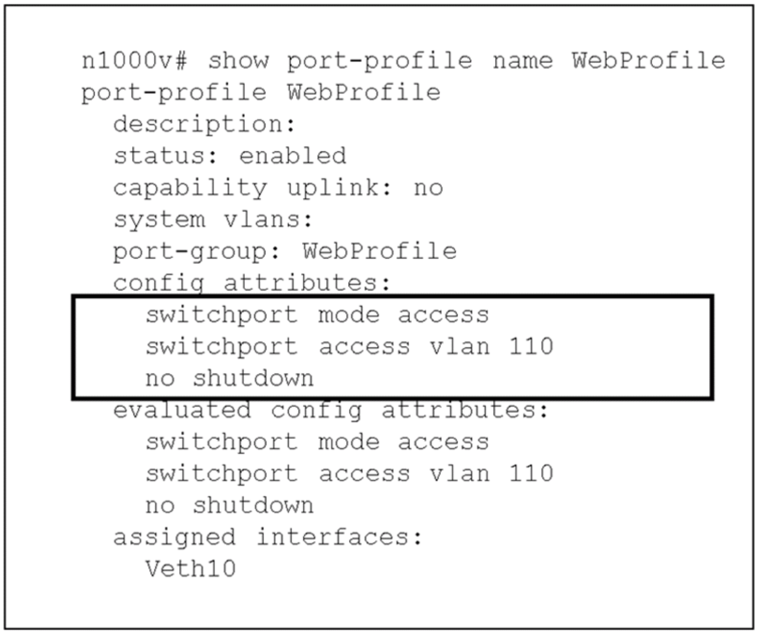
```
(config)# interface veth1
(config-if)# inherit port-profile webservers
```

Port profiles are created on the VSM and propagated to VMware vCenter server as VMware port groups using the VMware VIM API. After propagation, a port profile appears in VMware vSphere client and is available to be applied to a virtual machine's vNIC.

Figure 5-6 illustrates the network admin view of a port-profile from the Nexus 1000v VSM.

The following example illustrates the network admin view of a port-profile from the Nexus 1000v VSM **show port-profile** command output on the 1000v console.

```
your command output here
```

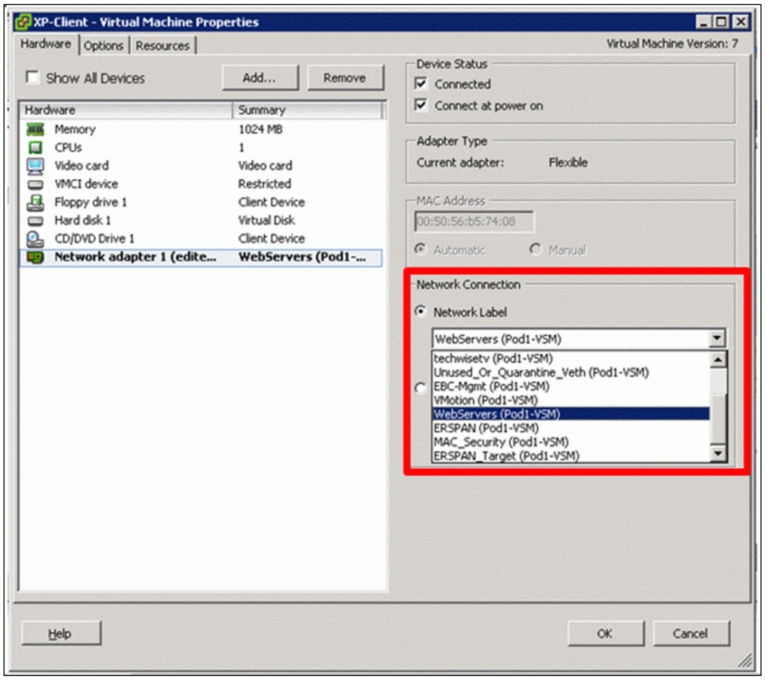


```
n1000v# show port-profile name WebProfile
port-profile WebProfile
  description:
  status: enabled
  capability uplink: no
  system vlans:
  port-group: WebProfile
  config attributes:
    switchport mode access
    switchport access vlan 110
    no shutdown
  evaluated config attributes:
    switchport mode access
    switchport access vlan 110
    no shutdown
  assigned interfaces:
    Veth10
```

Figure 5-6 Show port-profile command output on the 1000v console

When the server administrator provisions a new virtual machine, a dialog box pertaining to network configuration appears. This dialog box is consistent regardless of the presence of the Cisco Nexus 1000v Series, that is, the workflow for provisioning a new virtual machine does not change when the Cisco Nexus 1000v Series is used. The server administrator selects the port profiles to apply to each of the virtual machine's vNICs.

Figure 5-7 illustrates the server or virtualization admin view of a port-profile or a port-group in the VMware vCenter.


XP-Client - Virtual Machine Properties
Hardware
Options
Resources
Virtual Machine Version: 7
Add...
Remove
Help
OK
Cancel

□ Show All Devices

Hardware	Summary
Memory	1024 MB
CPUs	1
Video card	Video card
VMCI device	Restricted
Floppy drive 1	Client Device
Hard disk 1	Virtual Disk
CD/DVD Drive 1	Client Device
Network adapter 1 (edite...	**WebServers (Pod1-...**

Device Status
☑ Connected
☑ Connect at power on

Adapter Type
Current adapter: Flexible

MAC Address
00:50:56:b5:74:08
● Automatic ○ Manual

Network Connection
● Network Label
WebServers (Pod1-VSM)

techwisetv (Pod1-VSM)
Unused_Or_Quarantine_Veth (Pod1-VSM)
EBC-Mgmt (Pod1-VSM)
VMotion (Pod1-VSM)
WebServers (Pod1-VSM)
ERSPAN (Pod1-VSM)
MAC_Security (Pod1-VSM)
ERSPAN_Target (Pod1-VSM)

Figure 5-7 Virtual Machine Properties view in VMware vCenter

When the newly provisioned virtual machine is powered on, a vEthernet interface is created on the Cisco Nexus 1000v Series Switch for each vNIC that the virtual machine contains. The vEth interface inherits the definitions in the selected port profile. The port profile concept is new, but the configurations in port profiles use the same Cisco syntax that are used to manage physical switch ports on traditional switches.

Port profiles are not only used to manage vEth configuration; they are also used to manage the pNICs in a VMware ESX host. When a port profile is defined, the network administrator determines whether the profile will be used to manage vEth interfaces or pNICs.

Port profiles are not static entities. They are dynamic policies that can change as network needs change. Changes to active port profiles are applied to each switch port that is using the profile. This feature of port profiles is extremely useful when applying new network policies or changing existing policies.

Mobility of virtual machine security and network properties

Network and security policies defined in the port profile follow the virtual machine throughout its lifecycle, whether it is being migrated from one server to another, suspended, hibernated, or restarted. When a VM is moved, the VSM effectively moves the virtual Ethernet interface that the VM is attached to. The result is that in addition to migrating the policy (port-profile configuration), the VSM moves the virtual machine's network state, such as the port counters and flow statistics.

This is possible because a vEthernet interface has several characteristics that differ from other interface types. Besides the obvious fact that vEthernet interfaces are virtual and therefore have no associated physical components, the interface naming convention is unique. Unlike an interface on a physical switch, a vEth interface's name does not indicate the module with which the port is associated. Whereas a traditional physical switch port may be notated as GigX/Y, where X is the module number and Y is the port number on the module, a vEthernet interface is denoted as vEthY. This unique notation is designed to work transparently with VMware VMotion, keeping the interface name the same regardless of the physical server and VEM that the VM consists of. This is possible because all the VEMs are controlled by a centralized VSM and are part of a single virtual modular chassis.

Virtual machines participating in traffic monitoring activities, such as Cisco NetFlow or ERSPAN, can continue these activities uninterrupted by VMware vMotion operations because the vEthernet port and associated configuration (policy) follows the VM, see Figure 5-8.

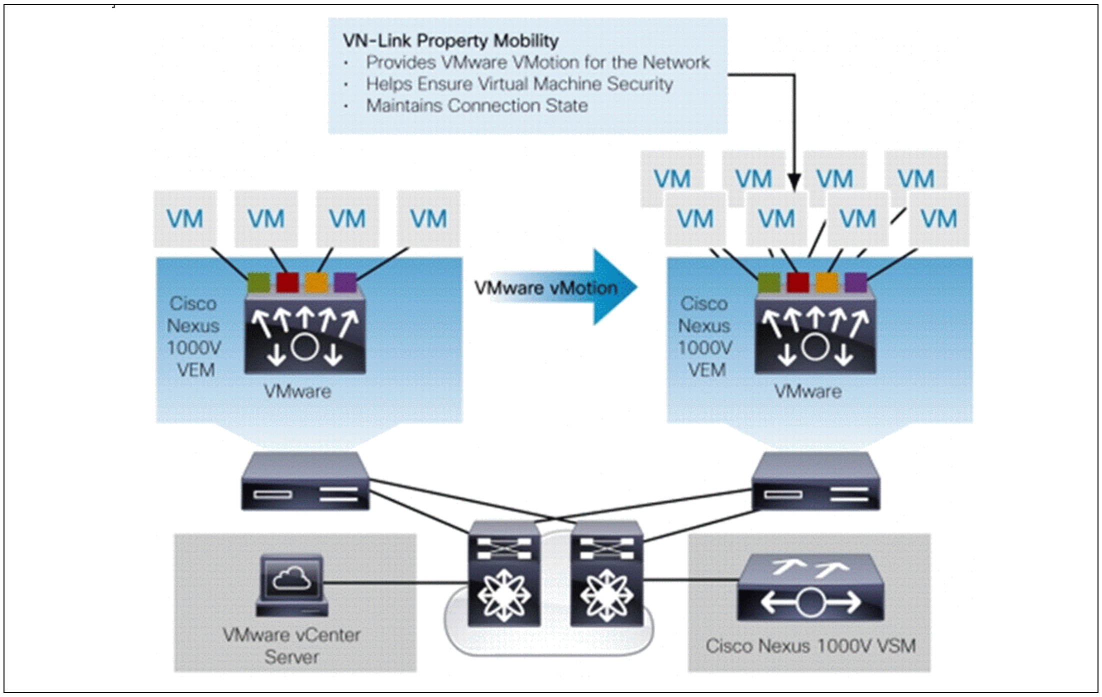

VN-Link Property Mobility
• Provides VMware VMotion for the Network
• Helps Ensure Virtual Machine Security
• Maintains Connection State
VM
VM
VM
VM
VMware vMotion
Cisco Nexus 1000V VEM
VMware
VMware vCenter Server
Cisco Nexus 1000V VSM

Figure 5-8 Mobility of network and security properties

vPath: Architecture for network services

Cisco vPath provides embedded intelligence within the Nexus 1000v Virtual Ethernet Modules (VEMs) for inserting network services such as firewalls, load balancers, and WAN optimization on the data path of traffic going to and from the VM connected to the Nexus 1000v. The concept of vPath is explained in Chapter 6, “Application networking solutions and data center security, physical and virtual” on page 287.

5.3.6 Nexus 1000v operational model

One of the challenges that the server virtualization brings to the operation of data centers is the fact that the network access layer is moved inside the virtualized server as described in 5.1.1, “Where is the network edge?” on page 254. When the server administrators are responsible for the virtual switch (virtual networking) this represents a challenge and a change on the operational practices given that a network entity (virtual switch) is controlled by the server and virtualization teams instead of the network team.

The architecture of the Nexus 1000v, as described previously, enables the operational practices to be restored and be consistent with the operational practices used to provide connectivity to physical servers.

A virtual Ethernet port on the Nexus 1000v has the same functionality as a physical Ethernet port on a physical switch. As such, it allows the network team to have full visibility of each VM because it restores the one to one mapping. One VM connects to one virtual Ethernet port and the statistics (interface counters for example) are collected on a virtual Ethernet interface representing the traffic information for a single VM.

The integration between VMware vCenter and the Cisco Nexus 1000v allows virtualization administrators to continue using VMware tools to provision virtual machines. At the same time, network administrators can provision and operate the virtual machine network by connecting to the Virtual Supervisor Module (VSM) in the same way they connect the physical network and use Cisco CLI and SNMP along with tools such as ERSPAN and NetFlow.

When restoring the operational model in which the network team provides virtual network services to the VMs and the server and virtualization administrators focus on managing VMs, the following areas are impacted:

Troubleshooting time

The network team has visibility of the individual VM level. When a problem related with network connectivity arises, for example, a VM cannot access another VM or its default gateway, the network team has end-to-end (physical and virtual) network visibility. They can check the complete data path from the core all the way to the inside of the virtualized server. This enables them to consistently correlate statistics from the virtual and physical ports. With the Nexus 1000v the networking team can also use advanced troubleshooting features, for example ERSPAN, to further investigate and resolve the problem. This reduces the total troubleshooting time and allows networking problems affecting VMs to be resolved quicker than before.

Number of outages caused by network configuration mistakes

The Nexus 1000v reduces the likelihood of outages caused by a configuration mismatch between the physical and virtual networks.

Without the Nexus 1000v, the server administrator is responsible for assigning the VLAN number to the VM, however, the server administrator has no visibility if the required VLAN is also present and allowed on the port on the physical switch. When there is a mismatch between the VLAN configured by the server administrator for the VM and the VLAN allowed on the physical port, the VM will not be able to communicate with any other VM outside the server where it exists or even with its default gateway but it will be able to communicate with other VMs in the same VLAN and in the same server. This scenario is usually very difficult to troubleshoot and to identify the root cause because there is connectivity to part of the infrastructure (other VMs in the same server and same VLAN) and no connectivity to the other parts of the infrastructure.

With the Nexus 1000v the network administrator will configure all the network policies (including VLAN number) on the port profile and given that the network team is also configuring the physical switches this significantly reduces the likelihood of a configuration mismatch between virtual and physical networks. This will reduce the number of outages caused by network configuration mistakes.

Consistency between the physical and virtual network

By running Cisco NX-OS, with similar capabilities as the physical switches, the Nexus 1000v increases consistency between the physical network and the virtual network, particularly if the physical network is based on Cisco Nexus switches that run the same operating system, NX-OS, as the Nexus 1000v. This increased feature consistency simplifies the enforcement of security policies and Quality of Service (QoS). It also allows end-to-end monitoring of the network as the Nexus 1000v can managed by the Network Management System of choice in the same way as the physical switches using Simple Network Management Protocol (SNMP) and XML.

The consistency between a physical and a virtual network also makes the response to security audits easier as the networking team can demonstrate the end-to-end path to access a VM. Furthermore, they can prove that the access to the VMs only allowed by other virtual or physical servers that comply with the required security policy. As the Nexus 1000v port profiles provide a persistent security policy even if the virtual machines move, the security audits are similar to those as for physical servers.

Performance and capacity management

When the virtual network is managed by the networking team in the same way as the physical network it becomes part of the network. The network performance and capacity management can be achieved much easier than if the virtual network is managed by the server administrators. The same tools and processes that are used to monitor the performance and the capacity of the physical network can be extended to the virtual network. By monitoring the interface statistics of the physical (uplink) and virtual Ethernet ports of the Nexus 1000v the networking team can have an accurate view of the complete network utilization. This information can also be used for billing purposes. Besides utilizing interface counters, the Nexus 1000v also supports Netflow v9.

5.3.7 Cisco Nexus 1100 Series Virtual Services Appliances

The Cisco Nexus 1100 Series Virtual Services Appliances (VSAs) are dedicated hardware platforms for hosting virtualized network services. The appliance can host a number of virtual service blades (VSBs), including the Cisco Nexus 1000v Virtual Supervisor Module (VSM), the Cisco Virtual Security Gateway (VSG), the Cisco Prime Network Analysis Module (NAM), and the Cisco Data Center Network Manager (DCNM) for both LAN and SAN management.

VSBs are created using ISO or OVA image files located in the Nexus VSA bootflash. The ISO defines the following for a VSB:

- Required number of interfaces
- Required hard disk emulations
- Disk and RAM defaults
- Type of virtual service blade
  - VSM
  - NAM
  - VSG
  - DCNM

The dedicated appliance, usually owned by the network team, allows the Nexus 1000v virtual supervisor module to be deployed in the same way as the physical switches. Network administrators can install and configure virtual access switches similar to the way they install and configure physical switches. This makes the deployment of the Nexus 1000v easier for the network administrator because it reduces the dependency and coordination with the server administrators. Hosting the VSM on the dedicated hardware (Nexus 1100) is especially helpful during the initial data center build, because there is no dependency in finding server resources for the VSM.

Figure 5-9 shows the internal architecture of the Cisco Nexus 1100 Series. The Cisco Nexus VSA Manager, based on Cisco NX-OS Software, manages VSBs, installation, and the blade configuration. The Nexus VSA Manager offers a familiar Cisco NX-OS interface for network administrators installing and configuring VSBs. The Cisco Nexus VSA Manager also supports Cisco NX-OS high availability, allowing a standby Cisco Nexus 1100 Series VSA to become active if the primary Cisco Nexus 1100 Series VSA fails.

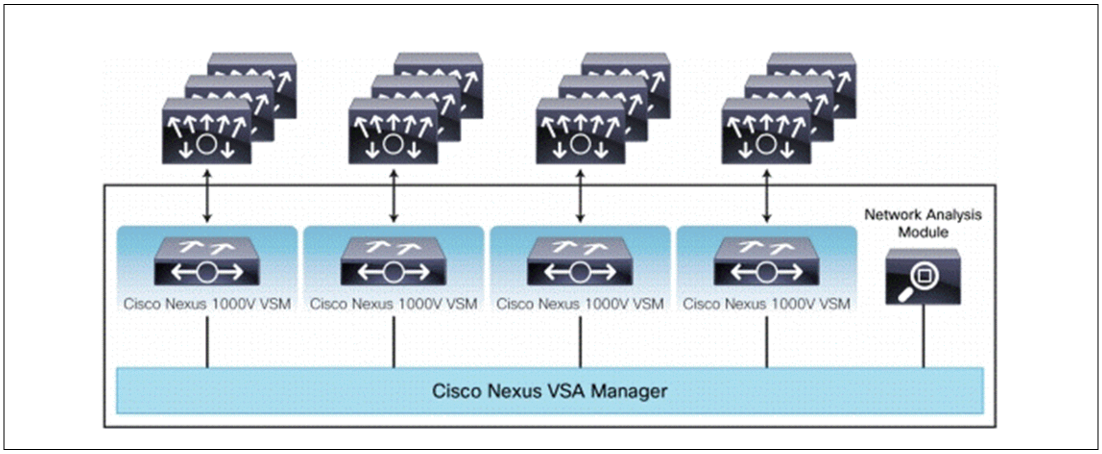


Figure 5-9 Internal Cisco Nexus 1100 Architecture

5.3.8 VSM on Nexus 1100 Series compared to VSM as a virtual machine

Table 5-1 compares deployment of a Cisco Nexus 1000v VSM as a virtual machine and on the Cisco Nexus 1100 Series. For a complete software-based deployment of the Nexus 1000v Series, deployment of the VSM as a virtual machine provides flexibility in VSM placement and even mobility with VMware vMotion. However, if the network administrators want greater control over the management of the VSM, the Cisco Nexus 1100 Series provides a NX-OS experience while installing the Cisco Nexus 1000v Series virtual access switch that is similar to installing a physical Cisco Nexus switch.

Table 5-1 Comparison of VSM on Nexus 1100 Series and VSM as a virtual machine

Feature	**VSM as virtual machine**	**VSM on Cisco Nexus 1100 Series**
Cisco Nexus 1000v Series features and scalability	Yes	Yes
VEM running on a hypervisor	Yes	Yes
Cisco NX-OS high availability of VSM	Yes	Yes
Software-only deployment (Hypervisor specific)	Yes	No
Installation like that of a standard Cisco switch	No	Yes
Network team ownership and management of the VSM	No	Yes
Support for multiple-hypervisor virtual machine traffic	No	Yes

The Cisco Nexus 1110-S VSA supports up to six VSBs, for example:

- 6 Nexus 1000v VSMs, each capable of managing 64 VMware ESX or ESXi hosts for a total of 384 VMware ESX or ESXi hosts
- 6 Cisco Virtual Security Gateway (VSG) VSBs

The Cisco Nexus 1110-X VSA supports up to 10 VSBs, for example:

- 10 Cisco Nexus 1000v VSMs, each capable of managing 64 VMware ESX or ESXi hosts for a total of 640 VMware ESX or ESXi hosts
- 10 Cisco VSG VSBs.

5.3.9 Options to connect the Nexus 1100 to the network

This section describes the options to connect the Nexus 1100 to the network. The Nexus 1100 has six 1 Gigabit Ethernet uplinks. The Nexus 1110-X has in addition two 10 gigabit Ethernet interfaces, but those are not yet enabled by the time of writing.

For the latest details on the hardware, see the Cisco Nexus 1100 Series Virtual Services Appliances Data Sheet:

http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps12752/data_sheet_c78-297641.html

It is important to understand the different classes of traffic carried on the Nexus 1100 Virtual Services Appliance uplinks.

Table 5-2 lists and describes the classes of network traffic carried on the Cisco Nexus Virtual Services Appliance uplinks.

Table 5-2 Classes of network traffic

Traffic Class	Data packets exchanged
Management	For Cisco Nexus Virtual Services Appliance and VSB management such as: ► Telnet ► SSH ► HTTP Note: If your virtual service blade uses the management class of traffic, it inherits the management VLAN from the Cisco Nexus Virtual Services Appliance.
Control	► Between the Cisco Nexus 1000v VSMs (VSBs) and VEMs. ► Between redundant Cisco Nexus Virtual Services Appliance active and standby suvervisors. ► Between redundant Cisco Nexus 1000v active and standby VSMs
Data	► VSB traffic that is not classified as either management or control ► High volume, application-specific traffic between virtual interfaces ► Traffic that is not considered management for the other VSBs should be isolated to a separate interface and classified as data. If the same interface is used for both management and data, as it the case with NAM, the traffic is classified as data. Note: Cisco Nexus 1000v VSM (VSB) traffic is not classified as data traffic

Connecting the Nexus 1100 to the network

There are five topologies (also referred to as uplink types) supported to connect the Nexus 1100 to the network. Once you configure one of the uplink type discussed below, the only way to modify it is to reload the software.

Topology / Uplink Type 1: All traffic shares a single uplink

This topology has the following characteristics:

- Management, control/packet, and data traffic will only flow over one link at any given time (ports 1 or 2)
- If interface 1 fails, traffic will fail to interface 2.
- If interface 2 fails, services will fail to the secondary Nexus 1110.
- Port-channel is not supported.
- 1 GB of throughput is shared between management, control/packet, and data traffic.
- Best practice: Interface 1 and interface 2 connect to different switches for redundancy purposes.

Figure 5-10 shows the topology with all traffic sharing a single uplink.

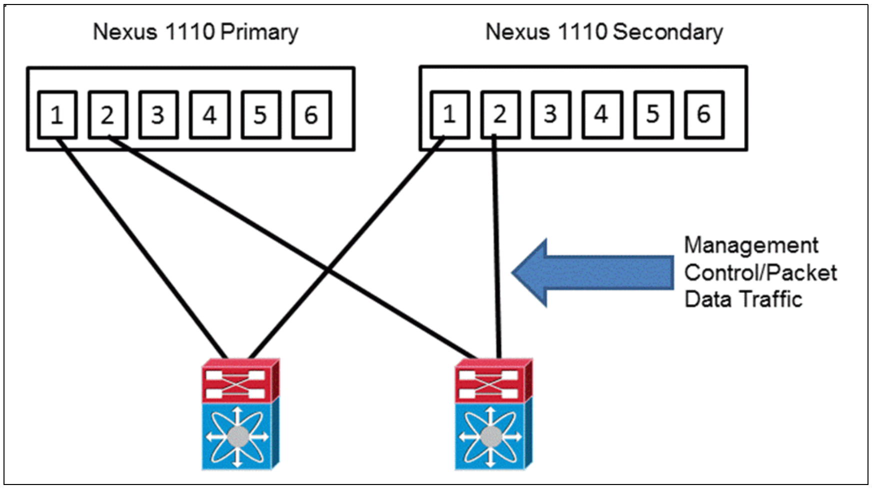


Figure 5-10 Topology / Uplink Type 1: All traffic shares a single uplink

Topology / Uplink Type 2: Management and control traffic share an uplink

This topology has the following characteristics:

Management and control traffic share an uplink - ports 1 and 2 - and has a throughput of 1 Gb.

Data traffic uses a separated uplink - ports 3-6

- Without VPC/VSS support on the switches that the Nexus 1110 connects to:
  - Eth3 and Eth5 are paired as a LACP port channel
  - Eth4 and Eth6 are paired as a LACP port channel
  - 2 GB of throughput
- With VPC/VSS support on the switches that the Nexus 1110 connects to.
  - One LACP port-channel
  - 4 GB of throughput

Figure 5-11shows the topology with the management and control traffic sharing a single uplink.

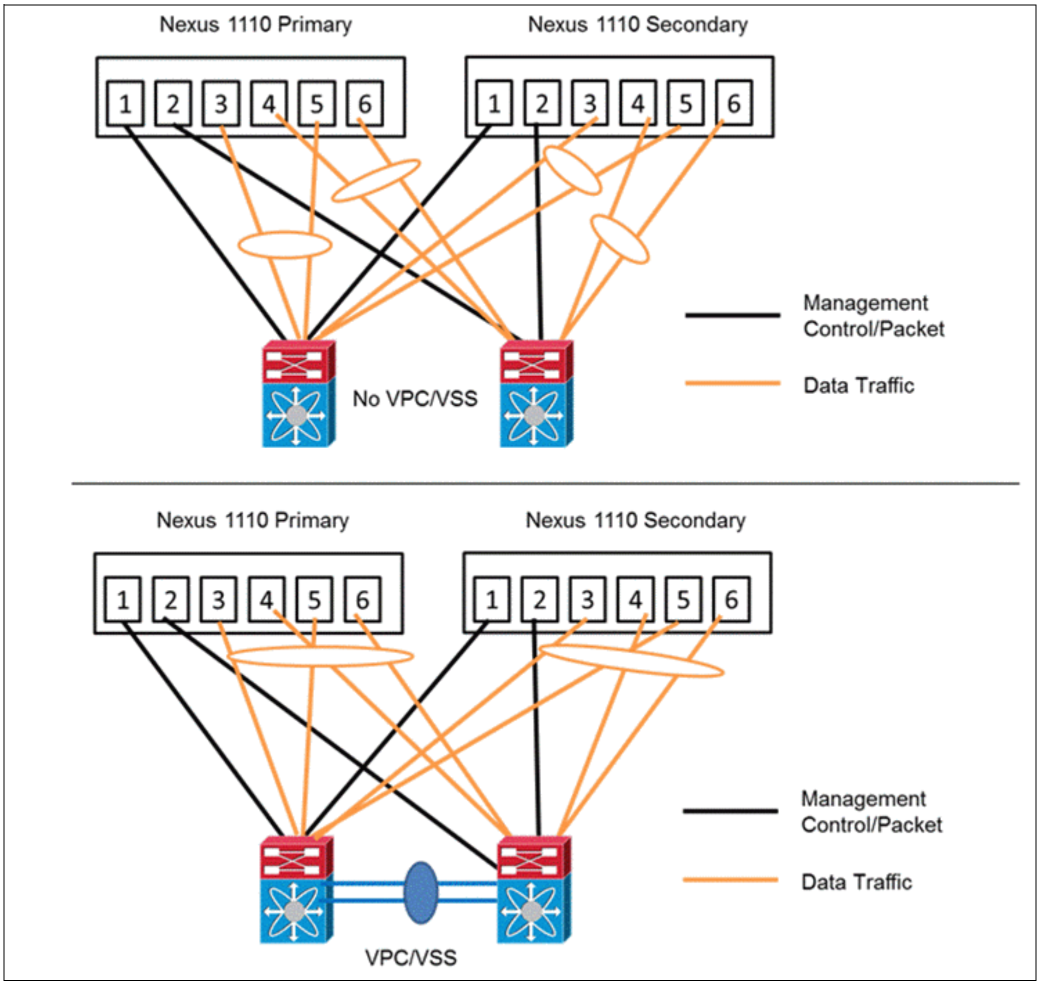


Figure 5-11 Topology / Uplink Type 2: Management and control traffic share an uplink

Topology / Uplink Type 3: Control and data traffic share an uplink

This topology has the following characteristics:

Management traffic has a dedicated interface - ports 1 and 2 - and has 1 Gb throughput. Control and data traffic share an uplink - ports 3-6.

- Without VPC/VSS support on the switches that the Nexus 1110 connects to:
  - Eth3 and Eth5 are paired as LACP port channel
  - Eth4 and Eth6 are paired as LACP port channel
  - 2 GB of throughput
- With VPC/VSS support on the switches that the Nexus 1110 connects to:
  - One big LACP port-channel
  - 4 GB of throughput

Figure 5-12 shows the topology with control and data traffic sharing a single uplink.

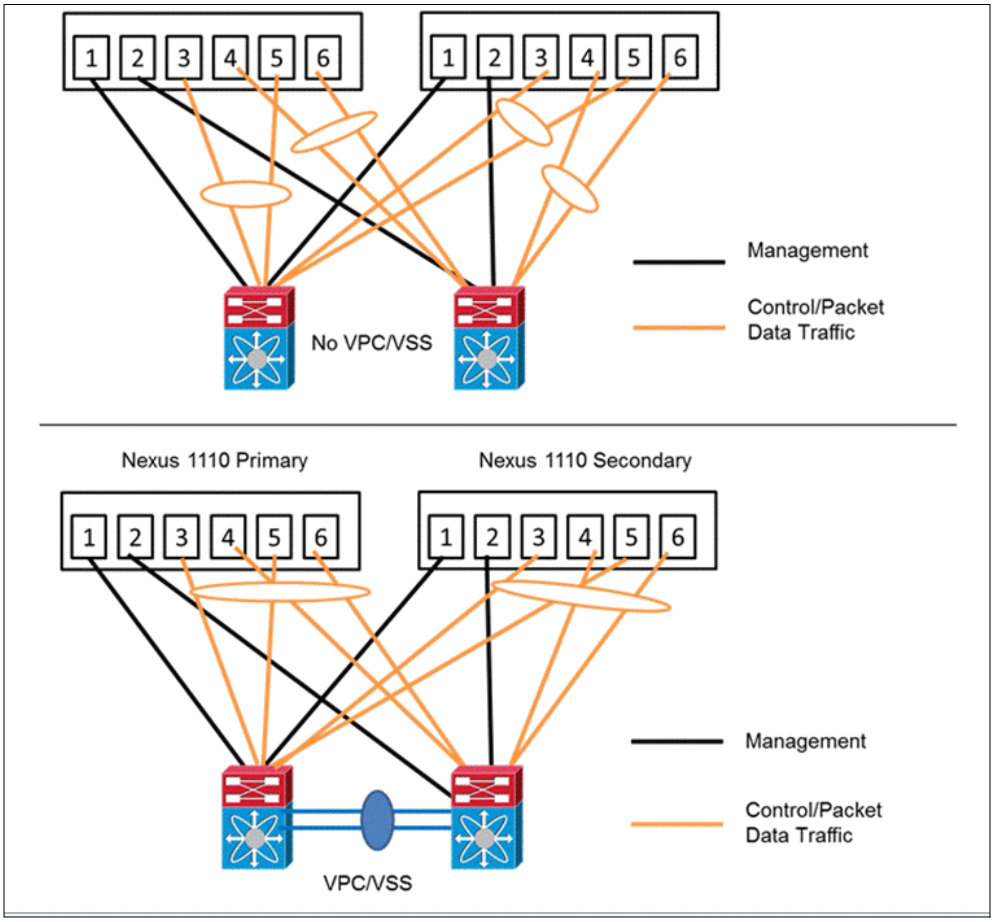


Figure 5-12 Uplink Type 3: Control and data traffic share an uplink

Topology / Uplink Type 4: Management, control, and data traffic are all on separate uplinks

This topology has the following characteristics:

- Management has dedicated interfaces:
  - Eth1 and Eth2 dedicated for management
  - 1 Gb of throughput
- Control traffic has dedicated interfaces:
  - Eth3 and Eth4 dedicated for control
  - 1 Gb of throughput
- Data traffic has dedicated interfaces:
  - Eth5 and Eth6 dedicated for data
  - 1 Gb of throughput

Figure 5-13 shows the topology with management, data, and control traffic on separate uplinks.

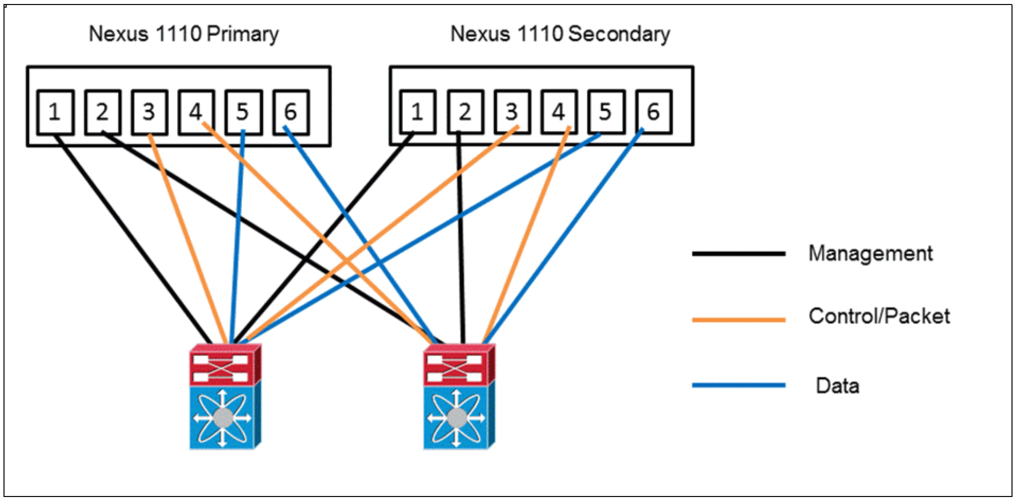


Figure 5-13 Uplink Type 4: Management, control, and data traffic are all on separate uplinks

Topology / Uplink Type 5: Flexible network uplink

A flexible network uplink configuration offers the flexibility to connect the Cisco Nexus 1110-S or Cisco Nexus 1110-X to the network, and allows flexible deployment of the VSBs on the Cisco Nexus 1110 Virtual Services Appliances.

The default flexible network uplink configuration is the basic configuration with each physical port acting as an individual uplink. This mode has the following characteristics:

- Every physical port individually forms an uplink.
- Each uplink can independently be configured.
- Ability to achieve maximum uplink of 6 Gbps.
- No default redundancy for uplinks.
- Physical ports cannot be bundled in a port channel.
- VSB traffic is segregated by default.

The VSB interface can be manually configured to share a port. Figure 5-14 shows the default flexible network uplink configuration.

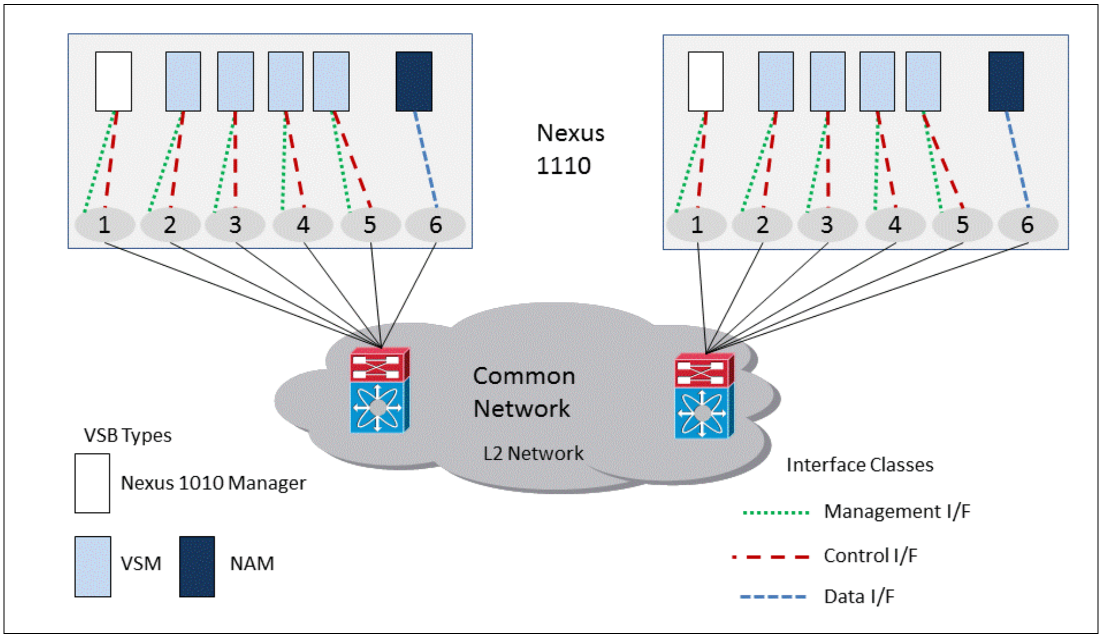


Figure 5-14 Default flexible network uplink configuration

You can then make changes to the default configuration by adding ports to a port channel or by assigning uplinks to a VSB interface. For example, you can add ports to a port channel, see Figure 5-15.

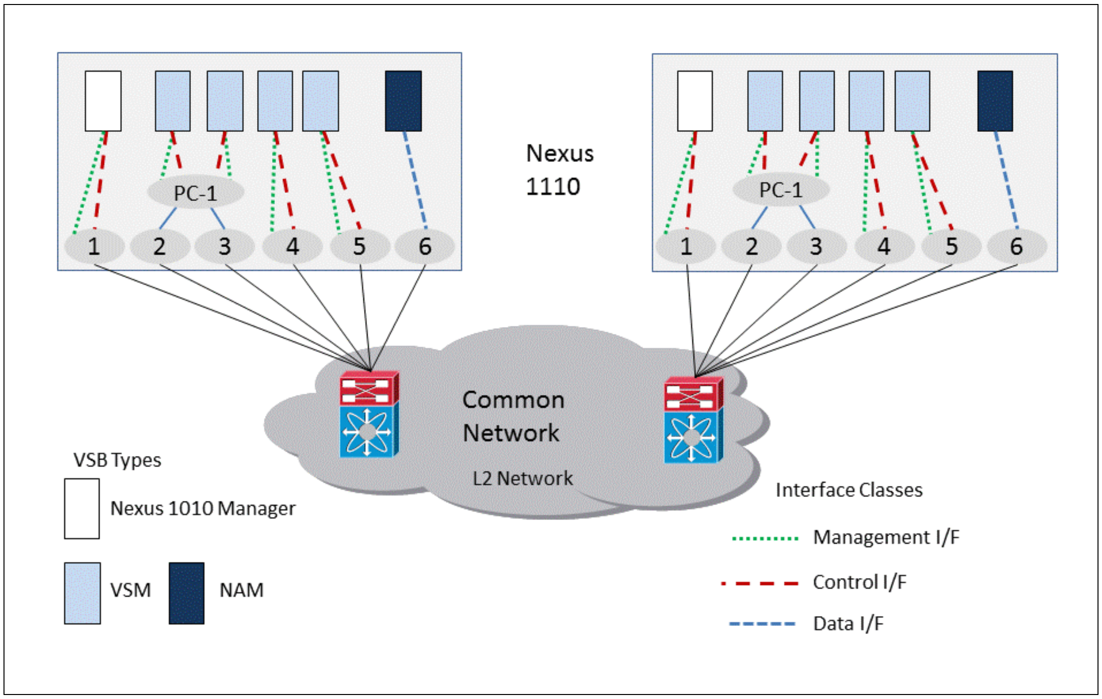


Figure 5-15 Default flexible network uplink configuration with port channel

You can also assign uplinks to a VSB interface, see Figure 5-16.

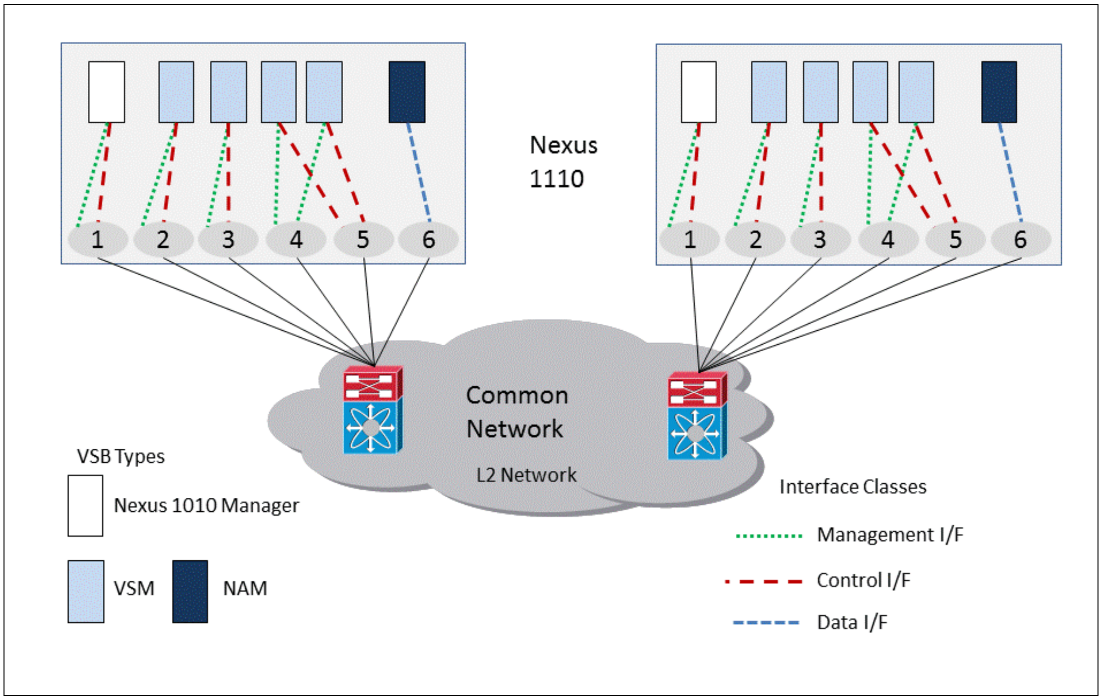


Figure 5-16 Default flexible network uplink configuration with uplinks to a VSB interface

5.3.10 VXLAN: Virtual eXtensible Local Area Network

Server virtualization has placed increased demands on the physical network infrastructure. At a minimum, there is a need for more MAC address table entries throughout the switched Ethernet network due to potential attachment of hundreds of thousands of Virtual Machines (VMs), each with its own MAC address.

Another type of demand that is being placed on data centers is the need to host multiple tenants on a shared infrastructure, each with their own isolated network domain. Each of these tenants and applications needs to be logically isolated from the others at the network level. For example, a three-tier application can have multiple virtual machines in each tier and requires logically isolated networks between these tiers. Traditional Layer 2 network isolation techniques such as IEEE 802.1Q VLAN provide 4096 LAN segments (through a 12-bit VLAN identifier) and may not provide enough segments for large cloud deployments.

5.3.11 What is a VXLAN?

VXLAN is a MAC in User Datagram Protocol (MAC-in-UDP) encapsulation technique and has a 24-bit segment identifier in the form of a VXLAN ID (Figure 5-17). The larger VXLAN ID allows LAN segments to scale to 16 million instead of 4096 VLAN identifiers. In addition, the UDP encapsulation allows each LAN segment to be extended across Layer 3 and helps ensure even distribution of traffic across PortChannel links that are commonly used in the data centers.

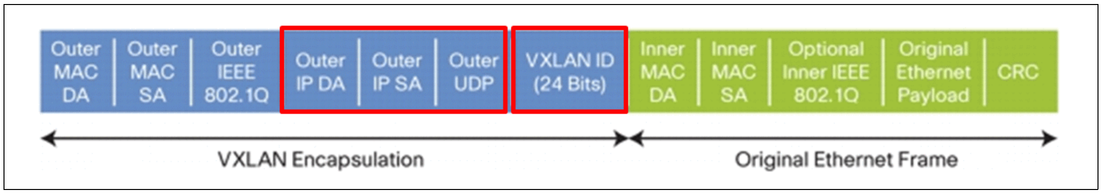


Figure 5-17 VXLAN encapsulation header

VXLAN provides the same service to End Systems as a VLAN. The VMs are not aware that they connecting to the network via a VLAN or a VXLAN.

The VM to VM traffic on different access switches (virtual or physical) is encapsulated in a VXLAN header + UDP + IP (refer to Figure 5-17).

5.3.12 Why VXLAN?

VXLAN is intended to address the problems discussed on this section. The focus is on the networking infrastructure within a data center.

Demands for more logical networks

Layer 2 data center networks use Virtual LANs (VLANs) to provide broadcast isolation. A 12-bit VLAN ID is used in the Ethernet data frames to divide the larger Layer 2 network into multiple broadcast domains. This has served well for several data centers which require fewer than 4094 VLANs. With the growing adoption of virtualization, this limit of 4094 VLANs may no longer be appropriate.

This is particularly an issue for cloud providers, due to the large number of tenants that a cloud provider might service, the 4094 VLAN limit is often inadequate.

Demands on the MAC address tables of Top-of-Rack switches in traditional Ethernet environments

Virtualized environments place additional demands on the MAC address tables of Top-of-Rack (ToR) switches that connect to the servers in a traditional Ethernet environment. Instead of just one MAC address per server link, the ToR switch now has to learn the MAC addresses of the individual VMs (which could range in the 100s per physical server). This is a requirement since traffic from/to the VMs to the rest of the physical network will traverse the link to the switch. A typical ToR switch could connect from 24 or 48 servers depending upon the number of its server facing ports. A data center might consist of several racks, so each ToR switch would need to maintain an address table for the communication of VMs across the various physical servers. This places a much larger demand on the table capacity compared to non-virtualized environments.

Cross PoD expansion

A Pod typically consists of one or more racks of servers with associated network and storage connectivity. Tenants may start off on a PoD and, due to expansion, require servers/VMs on other PoDs. This is especially the case when tenants on the other PoDs are not fully utilizing all their resources. This use case requires a stretched Layer 2 environment connecting the individual servers/VMs.

This scenario is particularly relevant for cloud providers as cloud computing involves on-demand, elastic provisioning of resources so the ability to provide service from any Pod that has available capacity is desirable.

Standardization of VXLAN

Cisco and a group of industry vendors, including VMware, Citrix, and Red Hat have submitted a proposal to IETF to standardize Virtual eXtensible Local Area Network (VXLAN):

http://tools.ietf.org/html/draft-mahalingam-dutt-dcops-vxlan-02

5.3.13 How VXLAN works

As described before, VXLAN is a Layer 2 overlay scheme over a Layer 3 network. Each overlay is termed a VXLAN segment. Only VMs within the same VXLAN segment can communicate with each other. VXLAN could also be termed as a tunneling scheme to overlay Layer 2 networks on top of Layer 3 networks where the tunnels are stateless. On the VXLAN network model, the end point of the tunnel (VTEP = VXLAN tunnel end point) is located within the hypervisor on the server that houses the VM. The VXLAN related tunnel/outer header encapsulation are known only to the VTEP; the VM never sees it.

The VXLAN network model in Figure 5-18 illustrates that the end systems (VMs) connect to the access switch, usually a virtual switch like the Nexus 1000v, which function as the VTEP.

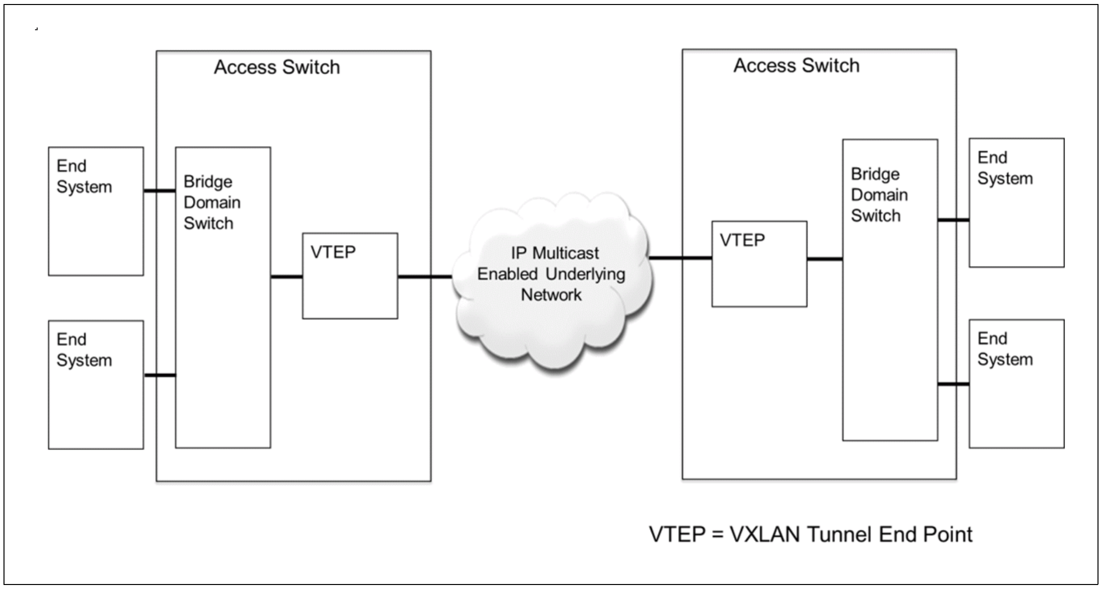


Figure 5-18 VM connection to the Access Switch using VTEP

Unicast VM to VM communication

If one considers a VM within a VXLAN segment, the VM is unaware of VXLAN. To communicate with a VM on a different host, it sends a frame destined to the target. The VTEP on the physical host looks up the VXLAN ID to which this VM is associated to. It then determines if the destination MAC is on the same segment. If so, an outer header comprising an outer MAC, outer IP address and VXLAN header are inserted in front of the original frame. The final packet is transmitted to the destination VTEP IP address as a unicast.

Upon reception, the remote VTEP (usually a remote virtual switch like the Nexus 1000v) verifies that the VXLAN ID is valid and is used by the destination VM. If so, the packet is stripped of its outer header and passed on to the destination VM. The destination VM never knows about the VXLAN ID or that the frame was transported with a VXLAN encapsulation.

In addition to forwarding the packet to the destination VM, the remote VTEP learns the inner source MAC to outer source IP address mapping from the packet it received from the source VTEP. It stores this mapping in a table so that when the destination VM sends a response packet, there is no need for an “unknown destination” flooding of the response packet.

Determining the MAC address of the destination VM prior to the transmission by the source VM is performed as with non-VXLAN environments except as described below. Broadcast frames are used but are encapsulated within a multicast packet, as detailed in the next section.

Broadcast communication and mapping to multicast

Consider the VM on the source host attempting to communicate with the destination VM using IP. Assuming that they are both on the same subnet, the VM sends out an ARP request which is a broadcast frame. In the non-VXLAN environment, this frame would be sent out as a broadcast to all switches carrying that VLAN.

With VXLAN, a header including the VXLAN ID is inserted at the beginning of the packet along with the IP header and UDP header. However, this broadcast packet is sent out to an IP multicast group. VXLAN requires a mapping between a VXLAN ID and an IP multicast group that it will use.

Based on this mapping, the VTEP can provide Internet Group Management Protocol (IGMP) membership reports to the upstream switch/router to join/leave the VXLAN related IP multicast groups as needed. This will enable pruning of the leaf nodes for specific multicast traffic addresses based on whether a member is available on this host using the specific multicast address.

The destination VM sends a standard ARP response using IP unicast. This frame will be encapsulated back to the VTEP connecting the originating VM using IP unicast VXLAN encapsulation. This is possible since the mapping of the ARP response's destination MAC to the VXLAN tunnel end point IP was learned earlier through the ARP request.

Another point to note is that multicast frames and “unknown MAC destination” frames are also sent using multicast, similar to the broadcast frames.

Summary of how VXLAN works

Here is an overview of the processing:

- VXLAN uses IP multicast to deliver broadcast, multicast and unknown MAC destinations to all access switches participating in a given VXLAN.
- VM MAC to access switch (VTEP) IP address mappings are learned by receiving encapsulated packets, similarly to Ethernet bridge flood and learn behavior.
- Known destination VM MAC addresses are carried over point-to-point tunnels between access switches (VTEPs).

5.3.14 Nexus 1000v support for VXLAN

The Nexus 1000v VEMs (Virtual Ethernet Modules) act as the VXLAN Tunnel Endpoints (VTEP) described on the previous section. The Virtual Ethernet Modules (VEMs) encapsulates the original Layer 2 frame from the virtual machines (VMs) into a VXLAN header + UDP + IP (see Figure 5-19).

Each VEM is assigned an IP address, which is used as the source IP address when encapsulating MAC frames to be sent on the network. This is accomplished by creating virtual network adaptors (VMKNICs) on each VEM (Figure 5-19). The VMKNIC is connected to a VLAN to transport the encapsulated traffic on the network.

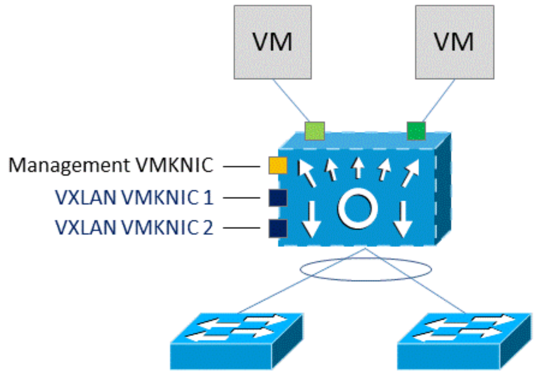


Figure 5-19 VEM VMKNIC interface with VXLAN capability

The connected VXLAN is assigned via configuration within the port profile, configuration of the vNIC and is applied when the VM connects. Each VXLAN uses an assigned IP multicast group to carry broadcast traffic within the VXLAN segment as described previously.

When a VM attaches to a VEM, if it is the first to join the particular VXLAN segment on the VEM, an IGMP join request is issued for the VXLAN's assigned multicast group. When the VM transmits a packet on the network segment, a lookup is made in the Layer 2 table using the destination MAC of the frame and the VXLAN identifier. If the result is a hit, the Layer 2 table entry will contain the remote IP address to be used to encapsulate the frame, The frame will be transmitted within an IP packet destined to the remote IP address. If the result is a miss (broadcast/multicast/unknown unicasts fall into this category), the frame is encapsulated with the destination IP address set to be the VXLAN segment's assigned IP multicast group.

Besides supporting the VXLAN encapsulation, the Nexus 1000v also has these features:

- It is fully integrated with VMware vCloud Director 1.5.
- It supports quality of service (QoS) for VXLAN.
- It extends the existing operational model to the cloud or environments where VXLAN is required.
- It supports Cisco vPath technology for virtualized network services together with VXLAN.
- It provides an XML API for customization and integration.

VXLAN-aware network services

The Cisco Nexus 1000v Series supports Cisco virtual service data path (vPath) architecture, which supports a variety of virtualized network services, such as Cisco Virtual Security Gateway (VSG) and Virtual Wide Area Application Services (vWAAS). These virtualized network services can also be applied to traffic on VXLAN.

5.3.15 Deployment considerations

This section addresses several considerations regarding deployment.

Multicast

In a typical Layer 2 network using VLANs, if a frame is received with an unknown destination MAC address, it is flooded out to every interface in that VLAN (except the one it came from). In VXLAN, multicast/broadcast (including unknown unicast) frames will be sent encapsulated with a multicast destination IP address. Ideally, each VXLAN should use a different IP multicast group address to minimize flooding frames to VEMs that do not need them. When using VXLAN encapsulation, a multicast IP address must be assigned to each VXLAN.

If the VXLAN VMKNICs on different VEMs belong to the same subnet, you only need to enable IGMP snooping on the VLAN on upstream switching to provide Layer 2 optimization for multicast traffic.

If the VXLAN VMKNICs on different VEMs are in different subnets, Layer 3 multicast routing must be configured on the upstream routers. As an alternative to enabling IP multicast routing, you can use FabricPath to extend the Layer 2 domain (place all VTEPs on the same VLAN) or you can use OTV to extend just the VXLAN transport VLAN (allowing the administrator to place all VTEPs on the same VLAN).

Proxy ARP

VXLAN VMkernel interface IP address is used to encapsulate and transport VXLAN frames in a UDP tunnel. The VMware host routing table is ignored. The VMKNIC's netmask is also ignored. The VEM will initiate an ARP for all remote VEM IP addresses, regardless of whether they are on the same subnet or not. If they are across a Layer 3 switch or router, you need to enable the Proxy ARP feature on the Layer 3 gateway so that it can respond to off-subnet ARPs.

Port channel

Port channels use different load-sharing algorithms for dividing outgoing traffic among different physical interfaces. IP encapsulation results in all outgoing VXLAN traffic carrying an outer header that has the source MAC/IP address of the VEM's VMKNIC. For optimal load balancing, you must configure a 5-tuple-based hash as the load-sharing algorithm on your switches in the data center, including the Nexus 1000v.

MTU size

VXLAN traffic is encapsulated in a UDP packet when sent out to the physical network. This encapsulation imposes the following overhead on each packet:

```
Outer Ethernet Header (14) + UDP header (8) + IP header (20) + VXLAN header (8) =
50 bytes
```

To avoid fragmentation and possible performance degradation, all the physical network devices transporting the VXLAN traffic need to handle 50 bytes greater than the maximum transmission unit (MTU) size expected for the frame. Therefore, the MTU settings has to be adjusted for all these devices, which will transport the VXLAN traffic. This includes the uplink port-profiles of Cisco Nexus 1000V Series Switch carrying the VXLAN traffic.

See the white paper, *Deploying the VXLAN Feature in Cisco Nexus 1000V Series switches*, for more details on the deployment considerations:

http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9902/guide_c07-702975.html#wp9000094

5.3.16 Working with OTV and LISP

VXLAN is intended for automated provisioning of logical networks in a cloud environment. (Overlay Transport Virtualization (OTV), using a frame format similar to that used by VXLAN, is a Data Center Interconnect technology for extending Layer 2 domains across data centers using Layer 3 protocols. However, OTV has simpler deployment requirements than VXLAN since it does not mandate multicast-enabled transport network. OTV also contains faults within a data center and provides a more robust Data Center Interconnection. Applications in a VXLAN segment are often accessed through external networks based on VLANs, and hence OTV is required to extend such applications over Layer 3 protocols so that both VXLAN and VLAN segments are extended.

The Locator ID Separation Protocol (LISP) goes a step further by providing IP address mobility between data centers with dynamic routing updates, thus providing highly efficient routing networks for LAN segment traffic stretched across Layer 3 networks. Although VXLAN, OTV, and LISP frame formats share a similar-looking packet encapsulation structure, they serve very differently networking purposes and hence, are complementary to each other.

5.3.17 Communication outside the VXLAN

If communication has to be established outside the VXLAN segment, there are two options available.

The first option is a multi-homed VM with one interface in VXLAN and one interface in a VLAN. Any communication to be established to a VXLAN from outside has to traverse by means of the multi-homed VM on the VLAN interface. For example, using a vApp with following VMs:

- Dual-homed client machine with one interface in VXLAN 4400 and one interface in VLAN 44
- Web server with one interface in VXLAN 4400 and one interface in VXLAN 4401
- Database server with one interface in VXLAN 4401

In this scenario, see Figure 5-20, one can use a remote desktop to a client machine on the interface, which is on VLAN 44, and then browse to the webserver in VXLAN 4400. Now, the webserver can communicate to the database server on VXLAN 4401, which is totally isolated, meaning that the client machine has no direct access to the database server.

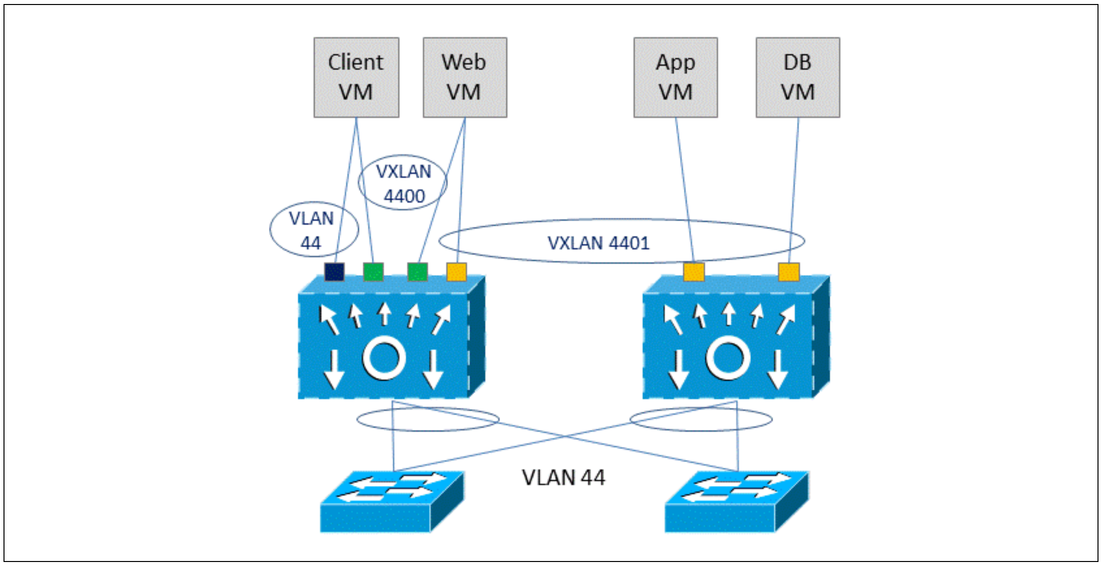


Figure 5-20 External connectivity with dual-homed VM

Another option is to have a VM (for example, a firewall VM or a router VM) to provide NAT and IP gateway functions to connect VMs on VXLAN to the network outside. The example, as in Figure 5-20, connects to an external network using firewall VM. In this case, the external interface of the firewall VM will be on VLAN 44, and the internal interface on VXLAN 4400, see Figure 5-21.

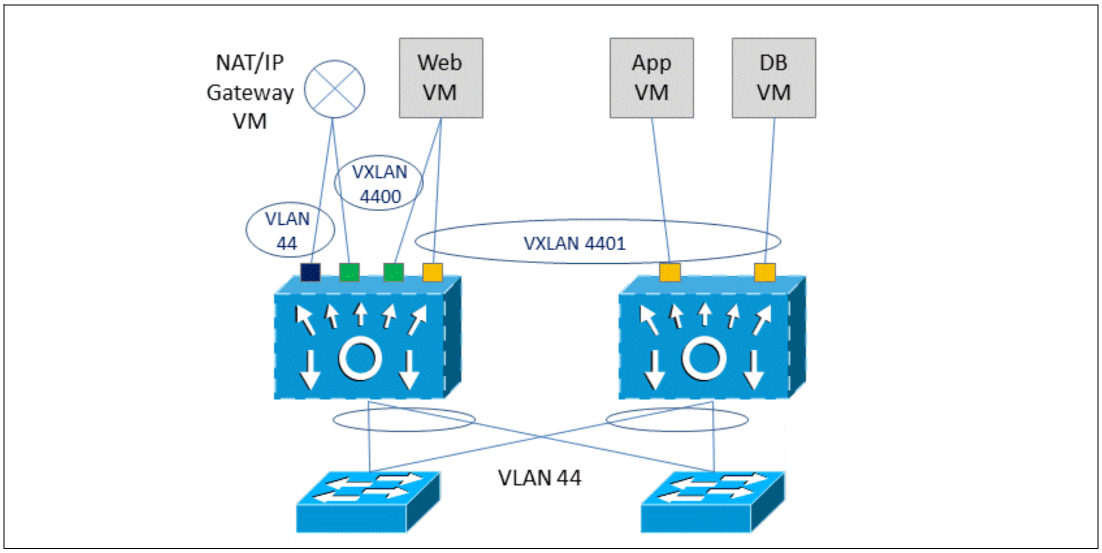


Figure 5-21 External connectivity with dual-homed VM with gateway VM

The Cisco ASA 1000V Cloud Firewall could be used as the gateway described before, to connect VMs on VXLAN to outside the network. The ASA 1000v will have its inside interface on the VXLAN segment and the outside interface on a VLAN and will provide firewall services (access control lists (ACLs), site-to-site VPN, Network Address Translation (NAT), stateful inspections, IP audit for example) to the packets from the VXLAN segment going to the outside network and vice-versa.

The Cisco Cloud Services Router 1000V is another example of a Virtualized Appliance, in this case a virtualized router, that could be used on the same function. One interface on the CSR would connect to the VXLAN segment, the traffic when arrives on this interface would be routed and could be sent out an interface that is connected to the external world via a VLAN.

The following section describes the Cisco Cloud Services Router 1000V in more detail.

5.4 Cisco Cloud Services Router 1000V Series

The Cisco Cloud Services Router (CSR) 1000V is a single-tenant router in a virtual form-factor. The CSR is deployed as a virtual machine (VM). It can run on any server that supports VMware ESXi virtualization and runs the same operating system, Cisco IOS-XE, which runs on the Cisco ASR 1000 product line.

The CSR architecture provides control plane and data plane separation and supports multi-core forwarding. IOS-XE is based on Cisco IOS software that runs on most of Cisco routers and switches, see Figure 5-22.

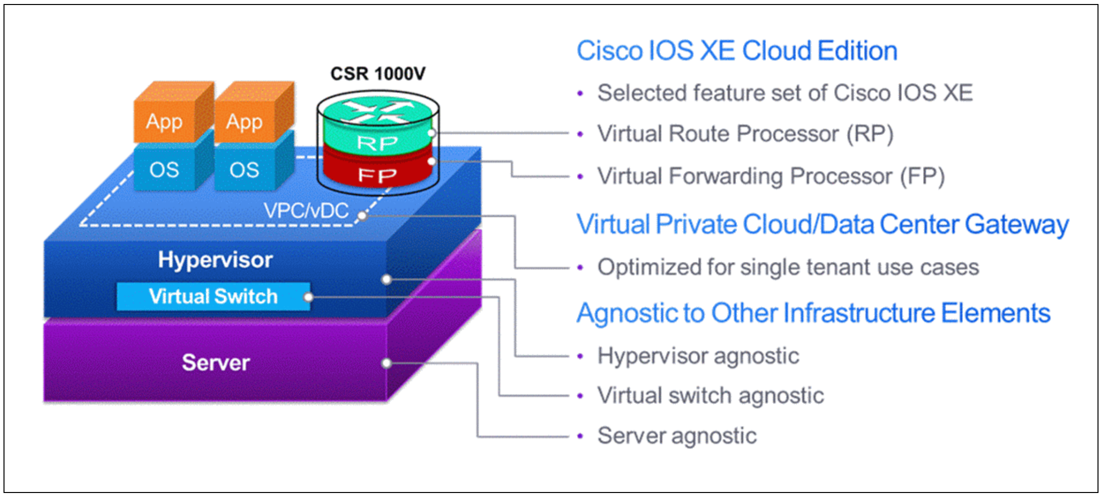

CSR 1000V
App
App
OS
OS
RP
FP
VPC/vDC
Hypervisor
Virtual Switch
Server
Cisco IOS XE Cloud Edition
• Selected feature set of Cisco IOS XE
• Virtual Route Processor (RP)
• Virtual Forwarding Processor (FP)
Virtual Private Cloud/Data Center Gateway
• Optimized for single tenant use cases
Agnostic to Other Infrastructure Elements
• Hypervisor agnostic
• Virtual switch agnostic
• Server agnostic

Figure 5-22 The Cisco Cloud Services Router 1000V

The CSR IOS-XE software has most of the networking and security features of the Cisco ASR 1000 physical router.

The CSR 1000V is intended for deployment in cloud data centers. It serves as a single-tenant router in a multi-tenant cloud. It is deployed at the edge of a tenant's network segment or a Virtual Private Cloud [VPC].

At time of writing, the CSR 1000V supports the following Cisco IOS features:

- Routing: BGP, OSPF, EIGRP, Policy-based Routing, IPv6 Forwarding, VRF-Lite, LISP
- Addressing: DHCP, NAT, LISP, 802.1Q VLAN
- VPN: IPSec VPN, DMVPN, EasyVPN, FlexVPN, GRE
- MPLS: MPLS VPN, VRF
- Security: Cisco IOS Zone-Based Firewall, ACL, AAA
- Failover: HSRP
- Traffic Redirection to Cisco WAAS Appliances: AppNav/WCCP

For more information about the Cisco Cloud Services Router 1000v, see this website:

http://www.cisco.com/en/US/prod/collateral/routers/ps12558/ps12559/data_sheet_c78-705395.html

The CSR 1000V is not related to the Cisco Nexus 1000V virtual switches. The CSR 1000V can run on a VMware ESXi hypervisor that contains the Nexus 1000V or a standard virtual switch, such as a VMware vSwitch.

Here are the primary use cases for the Cisco Cloud Services Routers:

- Secure VPN Gateway: The CSR 1000V supports route-based IPSec VPNs (DMVPN, EasyVPN, FlexVPN) along with the Cisco IOS Zone-based Firewall and access control which can be used for the enterprise to connect distributed sites directly to its cloud deployment.
- MPLS WAN Endpoint: The CSR 1000V can serve as an MPLS router that will enable a service provider to offer end-to-end managed connectivity (customer site to customer cloud deployment) with performance guarantees.
- Network Extension: The CSR 1000V offers features such as NAT and LISP that will enable an enterprise to maintain addressing consistency across premise and cloud as it moves applications back and forth or bursts compute capacity into the cloud.
- Control Point for Networking Services: The CSR 1000V can redirect traffic to Cisco vWAAS appliances deployed in the cloud. It also offers integrated networking services such as the Cisco IOS Zone-based Firewall and HSRP.

Also refer to the presentation, “BRKVIR-2016 - Cisco's Cloud Services Router (CSR): Extending the Enterprise Network to the Cloud,” available from this website:

https://www.ciscolive365.com

6

Application networking solutions and data center security, physical and virtual

Data center design is an act of constant planning. The core infrastructure design, power and cooling, cabling, and location must be carefully analyzed prior to installation and reviewed in an ongoing basis. The services in a data center are often seen as add-ons that are simply connected when needed. The concept of add-on services is a good one and in fact, services should be able to be added as requirements change and the data center grows. But when these devices are in the traffic path for critical applications and data, careful thought must be given to how they connect, where they connect, how they scale, and how they handle failure.

Security is often considered as an after-thought in many architecture designs. In reality, it is easier in the long run if security is considered as part of the core requirements, and not as an add-on. But this depends on several business-related factors and the drivers for security in the data center might change over time due to new application rollouts, compliance requirements, acquisitions, and, not the least of which, security breeches.

This chapter highlights security solutions that are tailored to meet the security requirements of a modern data center environment.

6.1 Network technologies that improve application performance and security

Application performance and security requirements have been essential in the past. With the virtualization of the data center, this has not changed. Instead, the need to have a new security standard seems obvious in a multi-tenant environment.

Typically, data center security contains the following components:

- Threat defense
- Application and content security
- Virtualization security
- Secure access

Apart from the security aspect, the application performance has always been a focus area. One way to improve the performance is to introduce load balancers and WAN optimization to improve the application traffic flow.

Virtualization at a compute layer introduces a new set of challenges when introducing or adding network services:

- Many applications require service chaining (that is, the ability to provide a single policy for a traffic flow as it ingresses to a VM for multiple network services), which further complicates the topic, whether it is a virtualized or a physical host.
- Different choices of x86 hypervisors: ESX, Hyper-v, RHEL, XEN.
- Virtualized hosts are moved around the DC via technologies such as vmotion for various reasons, like maintenance, service recovery/enhancement or migration scenarios. The network service placement/design need to take this mobility into consideration.

6.2 Cisco Application Control Engine (ACE) and SLB: Considerations and solutions

As applications are increasingly deployed to support primary business objectives within multiple organizations, new requirements are emerging to provide scalable application delivery services while reducing overall data center costs. Enterprises and service providers alike require scalable application delivery services that offer application security, maximized application availability, and optimized application performance as seen by users, all within a unified management framework. These services require several key data center technologies, including advanced server load balancing, Secure Sockets Layer (SSL) offloading, multilayer security, and application acceleration.

As virtualized servers can be introduced more easily, the capability to virtualize application delivery services becomes a modern data center requirement.

6.2.1 Benefits of virtualization with Cisco ACE

Virtualization in Cisco ACE delivers many business and technical benefits. Accelerated and lower-cost application rollout and upgrades with Cisco ACE, rolling out a new application or adding application support for another department simply requires the addition of a new ACE virtual partition to create another virtual Cisco ACE device within the existing physical Cisco ACE module, rather than deployment of an additional hardware platform.

The use of Cisco ACE significantly reduces the costs associated with deploying new applications, and it also enables administrators to accelerate application rollouts. Cisco ACE also streamlines application upgrading. With Cisco ACE, the administrator can easily take an application out of service and gracefully add or remove systems in one virtual device without affecting any other applications serviced by the physical Cisco ACE device. Figure 6-1 depicts the logical concept of ACE virtualization.

The end user connects to the applications using a Virtual IP (VIP) address. The ACE itself is subdivided into multiple logical devices, so-called Virtual Contexts (VC). Each VC is a fully feature subsystem with its own Linux kernel and connects to one or multiple VM servers. Adding a new server simply needs to be added to the VC and is brought into service without affecting the other production virtual servers.

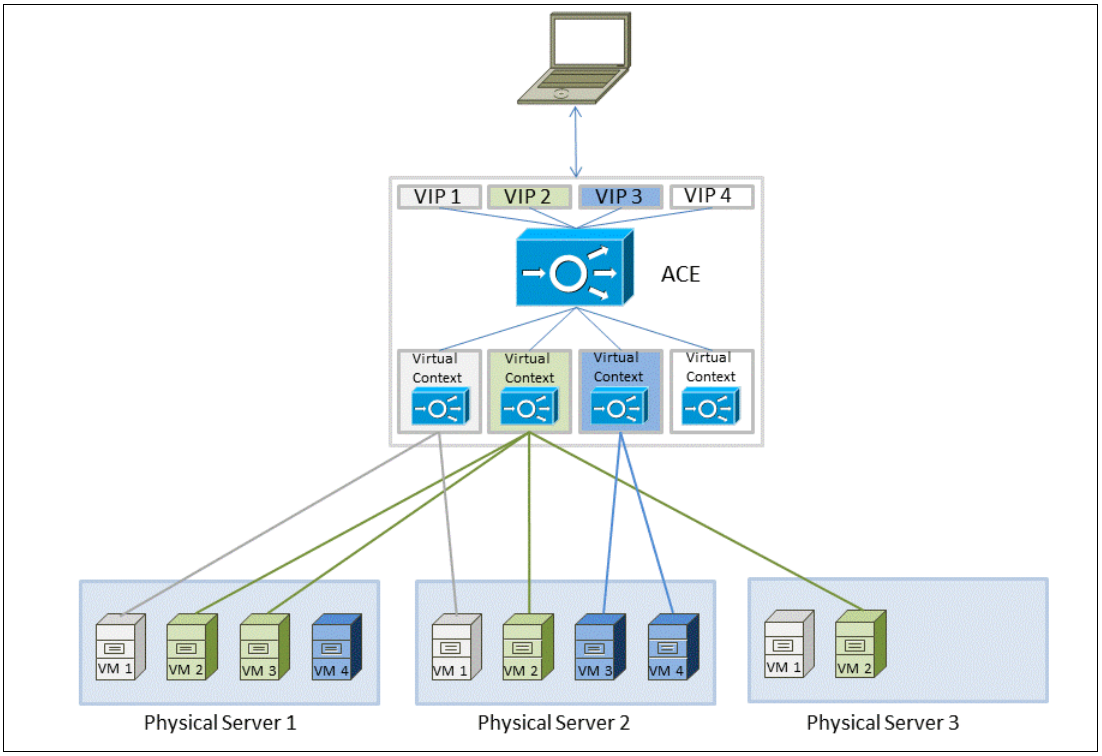


Figure 6-1 ACE Virtualization

There are some limitations with respect to the number of the VCs:

- 250 contexts + Admin context supported on the ACE module for the Catalyst 6500 family
- 20 contexts + Admin context supported on the ACE appliance

Every ACE device contains a special virtual context called “Admin,” which has settings for the ACE device itself. One can configure load balancing within the Admin context but it is recommended that a separate virtual context is created for load balancing. Otherwise, it is possible to allocate 100% of the Admin context resources for load balancing making the administration of the device impossible.

Application availability

To increase application availability, the Cisco ACE product family uses Layer 4 load balancing and Layer 7 content switching algorithms coupled with highly available system software and hardware. Specifically, Cisco ACE provides a failover system with many application health probes that helps ensure that traffic is forwarded to the most available server. To help ensure data center availability on geographically dispersed locations, Cisco ACE is integrated with a Global Site Selector (GSS), as described in 6.3.6, “ACE Global Site Selector appliances” on page 295. Cisco GSS provides connection failover between data centers and helps ensure business continuity.

Application performance

Cisco ACE is designed to accelerate the application experience for all users, whether they are in the office or on the road. To enable optimal application performance for remote and traveling users, Cisco ACE products use a range of acceleration capabilities to improve application response time, reduce bandwidth volume, and improve the efficiency of protocols. These technologies, including hardware-based compression, delta encoding, and Flash Forward, improve performance and reduce response time by minimizing latency and data transfers for any HTTP-based application, for any internal or external end user. When accessing SSL based applications, Cisco ACE can improve the server performance by caching and offloading Secure Sockets Layer (SSL) and TCP.

Security

The Cisco ACE provides protection against application threats and denial-of-service (DoS) attacks with features such as deep packet inspection, network and protocol security, and highly scalable access control capabilities.

The Cisco ACE come in two variations: as a module for the Catalyst 6500 family or as a standalone appliance. Visit Cisco’s web site for further details:

http://www.cisco.com/en/US/products/ps5719/Products_Sub_Category_Home.html

Improved workflow

With traditional application delivery solutions, application deployment often proceeds slowly because of the need for complex workflow coordination. For a new application to be deployed or for an existing application to be tested or upgraded, the application group must work with the network administrator. Coordination is required to make the desired configuration changes on the application delivery device (typically a load balancer), a process particularly problematic for the network administrator, who is responsible for helping ensure that any change to the configuration does not impact existing services.

Virtual partitioning with role-based access control (RBAC) in Cisco ACE mitigates this concern by enabling the network administrator to create an isolated configuration domain for the application group. By assigning configuration privileges within a single isolated virtual device to the application group, the network administrator can stay out of the workflow and eliminate the risk of misconfiguration of existing applications enabled in other virtual devices. This workflow creates a self-service model in which the application group can independently test, upgrade, and deploy applications.

Complete isolation of applications, departments, and customers

With Cisco ACE, administrators have the flexibility to allocate resources to virtual devices. For example, one administrator may want to allocate a virtual device for every application deployed. Another administrator may want to allocate a virtual device for each department's use, even for multiple applications. A service provider administrator may want to allocate a virtual device for each customer. Regardless of how resources are allocated, the Cisco ACE's virtual devices are completely isolated from each other. Configurations in one virtual device do not affect configurations in other virtual devices. As a result, virtual partitioning protects a set of services configured in several virtual devices from accidental mistakes, or from malicious configurations, made in another virtual device. A configuration failure on Cisco ACE is limited to the scope of the virtual device in which it was created. A failure in one virtual device has no effect on other virtual devices in the Cisco ACE. This maximizes the uptime for critical applications, especially when Cisco ACE is deployed in a redundant high-availability configuration.

Optimized capacity utilization

With Cisco ACE, resources can be allocated with a fine granularity to virtual devices. Allocatable resources include, for example bandwidth, connections, connections per second, number of Network Address Translation (NAT) entries, and memory.

Capacity planning practices generally require devices in the data center to be deployed at about 75 percent capacity to allow for periods of peak utilization. However, this approach does not scale with increased application rollouts and growth in business requirements. As the number of devices increases, the amount of unused resources committed for peak usage can exceed the capacity of entire devices.

For example, if four devices are deployed at the recommended 75 percent capacity to allow for scaling, 25 percent of each device is unused. In this simple four-device example, the aggregated unused resources equal the capacity of an entire device (4 times 25 percent equals 100 percent). When extrapolated to real-life data center scenarios in which dozens or hundreds of devices are deployed, the cost associated with such unused reserved resources is significant. Cisco ACE's virtual partitioning capability in combination with the module's scalability, delivering up to 16 Gbps throughput per module, enables multiple virtual devices to share the same physical Cisco ACE device while at the same time allocating resources for future growth.

Reduced data center resource requirements

The virtualization capabilities of Cisco ACE enable customers to reduce both physical and environmental data center resources. As previously discussed, Cisco ACE allows administrators to roll out additional applications simply by configuring additional virtual devices within the same physical Cisco ACE module, rather than deploying additional hardware platforms. As a result, network sprawl is reduced and additional cabling requirements and incremental rack space requirements are eliminated. In addition, because Cisco ACE reduces the number of physical devices in the network, it also significantly reduces power consumption and costs in the data center, even as customers deploy additional applications.

Service continuity through high-availability capabilities

For high availability, Cisco ACE supports stateful redundancy with both active-active and active-standby designs. Stateful redundancy is crucial because it enables user sessions to continue uninterrupted during a failover event. Cisco ACE pairs can be deployed either within the same Cisco Catalyst 6500 Series chassis or, more commonly in redundant data center designs, within two physically separate Cisco Catalyst 6500 Series chassis.

6.3 Application Control Engine (ACE) appliances and modules

At the time of writing, Cisco offers two ACE models:

- A Cisco ACE 4710 appliance, with a one-rack unit (1RU) form factor.
- The Cisco ACE30 (Application Control Engine) Module for Cisco Catalyst 6500 Series switches and Cisco 7600 Series Routers.

Both share the same feature set and can be configured in a similar fashion. There are some performance differences that originate from the different hardware platforms on which they are based, as shown in Table 6-1.

Table 6-1 Performance specifications for the Cisco ACE30 Module and the ACE4710 Appliance

Cisco device	Throughput (Gbps)	Compression (Gbps)	SSL transactions per second (TPS)	Virtual contexts
ACE30-MOD-16-K9	16	6	30,000	250
ACE30-MOD-08-K9	8	6	30,000	250
ACE30-MOD-04-K9	4	4	30,000	250
ACE30-BASE-04-K9	4	1	1,000	5
ACE-4710-04-K9	4	2	7500	20
ACE-4710-02-K9	2	1	7500	20
ACE-4710-01-K9	1	1	7500	20
ACE-4710-0.5-K9	0,5	0,5	7500	20

More details on the hardware limits can be found at:

http://docwiki.cisco.com/wiki/Cisco_Application_Control_Engine_%28ACE%29_Troubleshooting_Guide_--_ACE_Resource_Limits

6.3.1 Bridged mode

When the client-side and server-side VLANs are on the same subnets, the ACE to bridge traffic can be configured on a single subnet mode. No dynamic routing protocols are required.

To enable the bridge group VLANs, one must configure a bridge-group virtual interface (BVI) that is associated with a corresponding bridge group. An IP address on the BVI must be configured. This address is used as a source IP address for traffic from the ACE, for example, ARP requests, or management traffic. The ACE supports up to 4,094 BVIs per system.

The bridged mode has these characteristics:

- Easy migration for servers
- Requires one IP subnet for each server farm
- VIPs can be in the same or different subnet
- Recommend for non-LB traffic

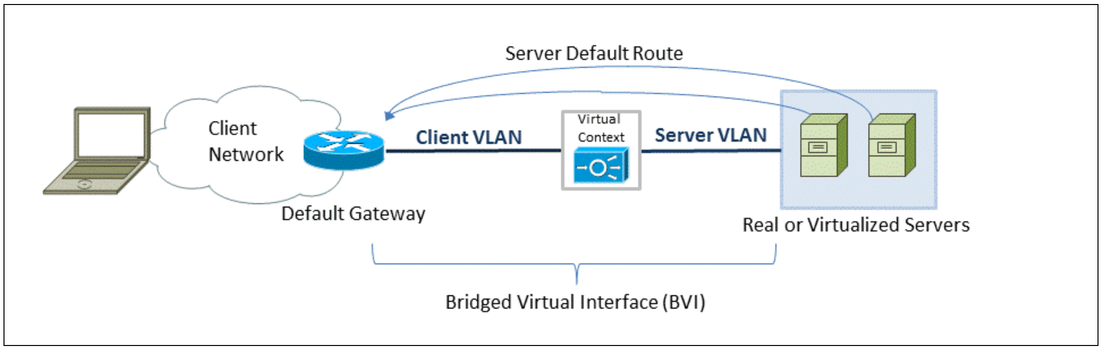


Figure 6-2 ACE bridged mode

6.3.2 Routed mode

When the client-side and server-side VLANs are on different subnets, one can configure the ACE to route the traffic. In routed mode, the ACE is considered a router hop in the network. In the Admin or virtual contexts, the ACE supports static routes only. The ACE supports up to eight equal cost routes for load balancing:

- Easy to deploy with many server IP subnets
- Requires at least two IP subnets
- Servers in dedicated IP subnet
- VIPs usually in different, routable subnet from servers

Figure 6-3 shows ACE routed mode.

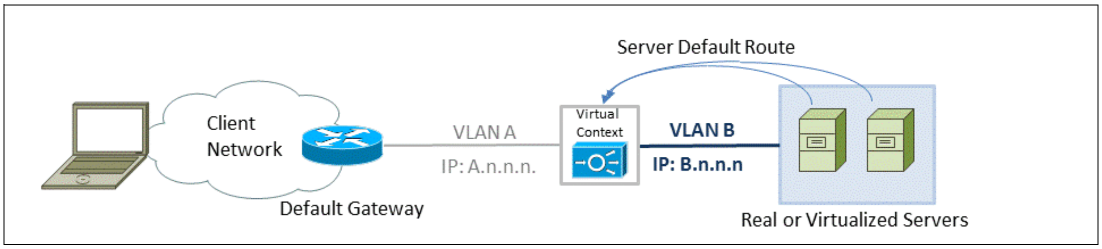


Figure 6-3 ACE routed mode

6.3.3 One-armed mode

In one-armed mode, the ACE is configured with a single VLAN that handles both client requests and server responses. For one-armed mode, the ACE must be configured with client-source network address translation (NAT) or policy-based routing (PBR) to send requests through the same VLAN to the server.

The ACE is not inline with the traffic and receives and sends requests through the default gateway to the servers, see Figure 6-4. The MSFC routes requests to a VIP that is configured on the ACE. When the ACE selects the server for the request based on the configured policy, it rewrites the source IP address with an address in the NAT pool. Then the ACE forwards the request to the server on the same VLAN through the default gateway on the MSFC.

The server sends a response to the default server gateway. The server response contains its source IP address and the NAT address of the ACE as the destination IP address. The gateway forwards the response to the ACE. The ACE receives the response, changes the source IP address to the VIP, and sends it to the MFSC. Then the MFSC forwards the response to the client.

Following are some characteristics of one-armed mode:

- Load balancer not inline
- Allows direct server access
- Requires source NAT

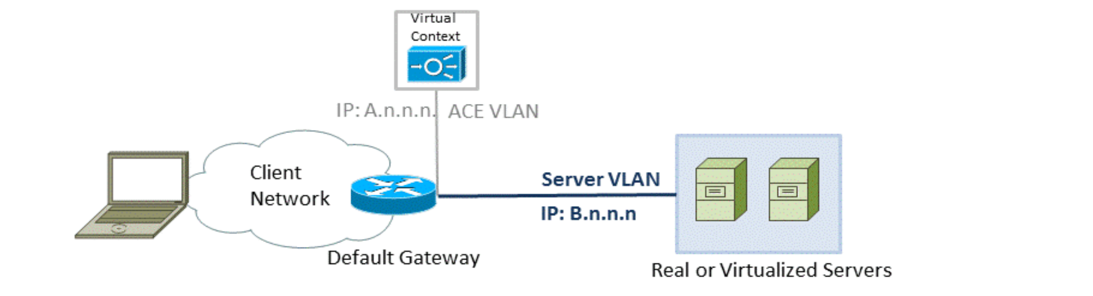


Figure 6-4 ACE one-armed mode

6.3.4 Sample deployments

The traditional solution has been to rely on standalone application delivery appliances, typically connected to Layer 2 to 3 switches in the distribution layer. Multiple physical links are required to protect against link failure, and dedicated links to redundant Layer 2 to 3 switches are often a requirement to account for all possible link and device failures. When additional performance is required, or when specific applications or organizations require dedicated devices, the solution includes additional physical devices connected to the same Layer 2 to 3 switches.

Figure 6-5 shows a data center with two pairs of redundant ACE load balancers connected to the distribution layer and serving requests for one or more tiers of servers. These load balancers can be either the standalone appliance or a separate Catalyst 6500 switches with the ACE30 modules.

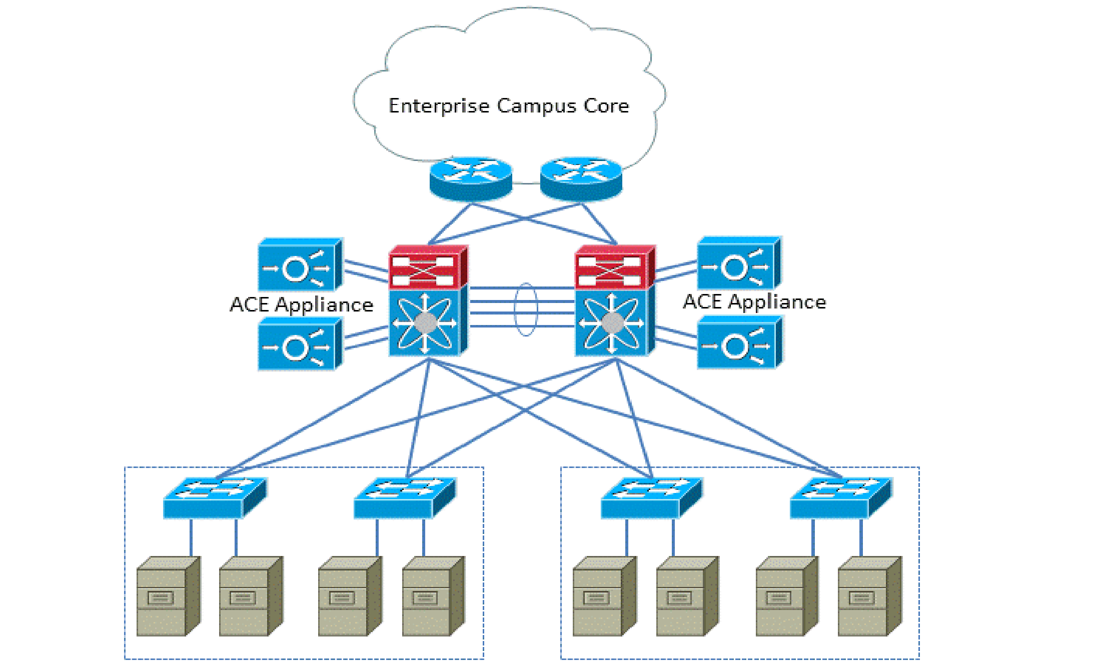


Figure 6-5 Data center with application delivery appliances deployed in the distribution layer

6.3.5 Data center with Cisco ACE module integrated into the distribution layer

A different approach that addresses many of today's data center application infrastructure challenges fully integrates Layer 4 to 7 application delivery services within the routing and switching infrastructure, especially within modular switches. Whether this is accomplished in software alone or with dedicated hardware, the resulting design is much simpler, reducing the need for extra cabling and dedicated power supplies, and much more flexible, allowing rapid changes without the need for any physical modification of the infrastructure. The user retains full control over which flows have to be processed for load balancing or application delivery functions, while all other flows are simply Layer 2 or Layer 3 switched without any affect on performance. In addition, specific integration features can enhance even further the collaboration between the Layer 2 and 3 functions and the application delivery engines and deliver additional security and resiliency.

Cisco delivers such a solution with its integration of the Cisco ACE module into the Cisco Catalyst 6500 Series switches. Figure 6-6 shows the topology when the redundant Cisco Catalyst 6500 Series switches in the distribution layer of a data center are equipped with Cisco ACE modules.

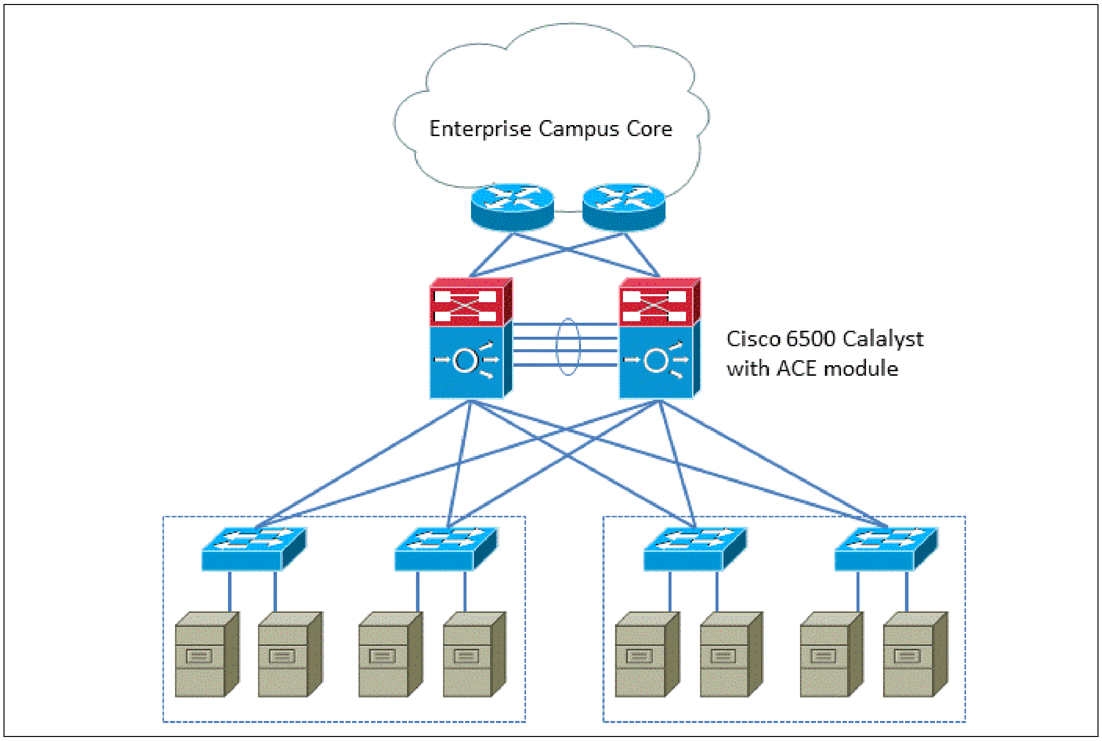


Figure 6-6 Data center with Cisco ACE module integrated into the distribution layer

6.3.6 ACE Global Site Selector appliances

The Cisco Global Site Selector (GSS) is part of the Cisco Application Control Engine (ACE) family. It can be part of any data center or cloud computing architecture that requires an appliance-based, security-focused, global load balancer.

The Cisco GSS allows the end user to access global Internet and intranet applications even if the primary data source is overloaded or not accessible.

The Cisco GSS traffic-management process continuously monitors the load and health of any SNMP-capable device (such as a server load balancing device or real or virtualized server) within each data center. The GSS uses this information in conjunction with customer-controlled load-balancing algorithms to select the best data center, server complex, or server destinations that is available and not overloaded, within user-definable load conditions, in real time.

In this manner, the GSS selects best destination to ensure application availability and performance for any device that uses common DNS request for access. In essence, this means that one can achieve a higher application availability due to provisioning of a failover across distributed data centers and better application performance due to optimized load growth to multiple data centers, servers and switches.

Features and benefits

The Cisco GSS 4492R offers the following benefits:

- Provides a scalable, dedicated hardware platform to help ensure that applications are always available, by detecting site outages or site congestion and performing global traffic management for client requests
- Provides a resilient architecture that is crucial for disaster recovery and for multi-site Web application deployment
- Offers site persistence for e-commerce applications
- Scales to support hundreds of data centers or SLB devices
- Offloads, augments, and optimizes the DNS infrastructure by taking over the domain resolution process for content requests and delivery for all types of static and dynamic Web content, and processes the request-responses at thousands of requests per second
- Offers flexible, heterogeneous support for all Application Networking Services (L4-7) devices and DNS-capable networking products, including third-party load balancers
- Improves the global data center selection process by offering user-selectable global load-balancing algorithms along with universal SNMP load and health probes
- Tightly integrates with Cisco SLB/ADC devices without sacrificing the ability to work in a heterogeneous environment of DNS-capable networking products
- Offers proximity features that use Cisco routers and Layer 4 through 7 content switches to allow the GSS to direct content consumers to the closest data center in real time
- Provides centralized command and control of the DNS resolution process for direct and precise control of the global load-balancing process
- Supports a Web-based GUI and DNS wizard to simplify GSLB command and control
- Supports role-based access control (RBAC) and operation to limit access to GSS functions and features
- Supports configuration using a flat text file, the command-line interface (CLI), and the GUI
- Supported by Cisco Application Networking Manager; see 6.3.7, “Application Networking Manager” on page 298, for unified operations management with Cisco ACE, CSS, and CSM

How the Cisco GSS performs its major tasks

The GSS performs two major functions as part of the global site selection process:

- Takes an active role in the DNS infrastructure to connect the client to the SLB device that supports the requested Website
- Continuously monitors the load and availability of these SLB devices to select the SLB device most capable of supporting the new client

In Figure 6-7, the GSS provides GSLB services for the entire DNS zone (company.com). The Cisco GSS can continuously and simultaneously monitor the load and health of the servers in the data centers. These servers can be located together or at disparate remote or standalone data centers.

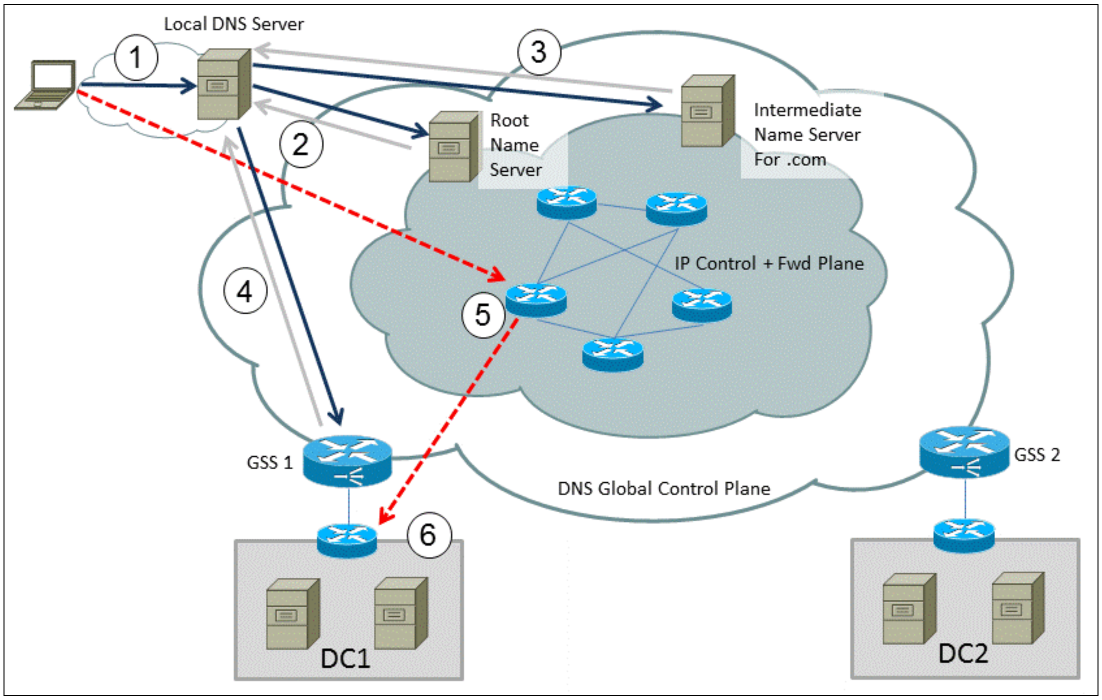


Figure 6-7   Cisco GSS Site Selection Process

How the GSS interacts with the client in the data center selection process is summarized in the following six steps (corresponding to the numbers in Figure 6-7):

Step 1. A client wants to access an application at company.com (Web, e-mail, VPN, and so on). The resolver (client) sends a query for company.com to the local client DNS server (D-proxy).

Step 2. The local D-proxy does not have an IP address for company.com, so it sends a query to a root name server. The root name server can respond to the request in two ways. The most common way is to send the D-proxy directly to the authoritative name server for company.com. In the other method, iterated querying (shown in Figure 2), the root name server sends the D-proxy to an intermediate name server that knows the address of the authoritative name server for company.com.

Step 3. The local D-proxy sends a query to the intermediate name server, which responds, referring the D-proxy to the authoritative name server for company.com.

Step 4. When the local D-proxy sends a query to the authoritative name server for Cisco.com, the name server responds with the IP addresses of the two GSS devices, and tells the D-proxy to ask the GSS for the IP address for company.com or www.company.com.

Step 5. The local D-proxy sends its final request directly to one of the two GSS devices. The GSS is authoritative for the company.com sub domain, so it will be responsible for sending the “A record” response, the IP address, to the D-proxy. The GSS sends the best IP address for that requester at that time - in this case, the IP address that is a Virtual IP on the server load balancer device at data center 1.

In order to send the best IP address to the D-proxy, the GSS manages client traffic flows to each data center, routing users to the closest, least loaded, or otherwise selected “best answer” based on any of the ten global load balancing algorithms that can be applied. The GSS will not send an IP address to the D-proxy if the device is overloaded or not responding.

Step 6. The DNS global load balancing process is complete; the client is directed to the server load balancer device at data center 1 and the client communicates directly with the Virtual IP on the server load balancer at data center 1.

The Cisco GSS can be deployed as a standalone global traffic manager that globally load balances client requests across distributed data centers using network performance metrics such as content use, round-trip time (RTT) between client and the closest data center, routing topology, and any device performance values that are available through SNMP.

6.3.7 Application Networking Manager

Because all virtual devices reside on the same physical Cisco ACE platform, all of the logical devices can be easily and centrally managed. Cisco ACE can be logically partitioned using either the Cisco Application Networking Manager (ANM) GUI, a powerful command-line interface (CLI), or an Extensible Markup Language (XML)-based API. The ANM is an intuitive workstation-based management system capable of simultaneously managing many Cisco ACE modules in multiple Cisco Catalyst 6500 Series chassis, each with many virtual devices.

The ANM can be deployed as a standalone server or a virtual appliance.

Figure 6-8 shows how ANM and the ACE can be integrated with VMware, allowing the creation and management of server farms for application delivery that consist of real servers that are a combination of physical servers and VMware virtual machines (VMs).

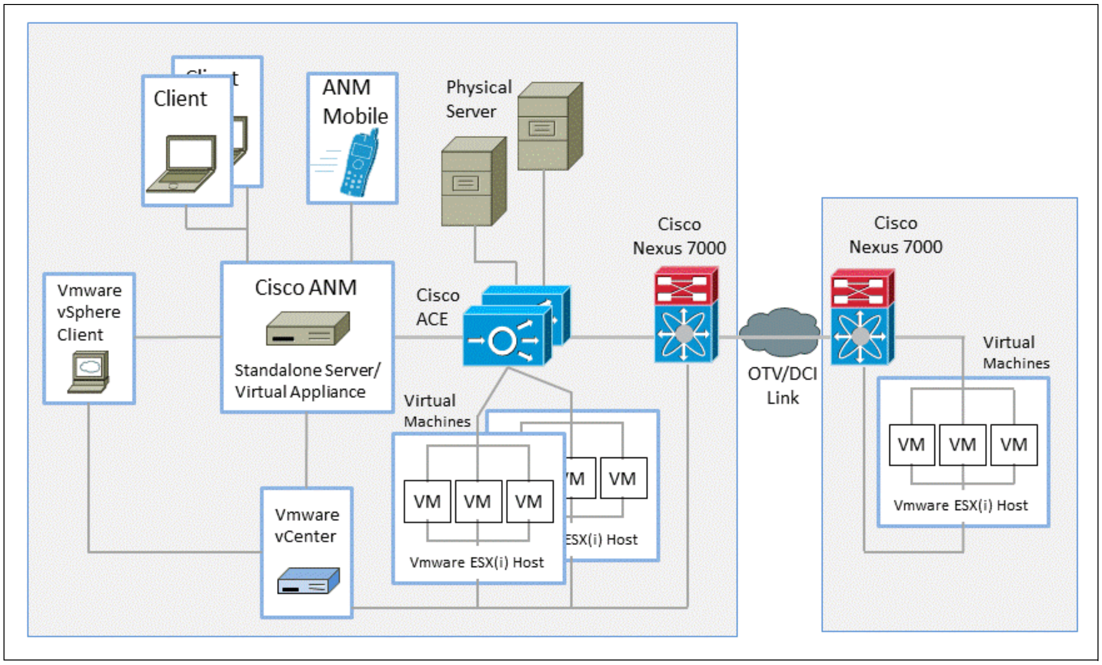


Figure 6-8 The ANM sample network

ANM is a client server application that enables the following functions:

- Configure, monitor, and troubleshoot the functions of supported data center devices.
- Create policies for operations, applications owners, and server administration staff to activate and suspend network-based services without knowledge of, or ability to, change network configuration or topology.
- Manage the following product types:
  - Cisco Application Control Engine (ACE) module or appliance
  - Cisco Global Site Selector (GSS)
  - Cisco Content Services Switch (CSS)
  - Cisco Catalyst 6500 Virtual Switching System (VSS) 1440
  - Cisco Catalyst 6500 series switch
  - Cisco 7600 series router
  - Cisco Content Switching Module (CSM)
  - Cisco Content Switching Module with SSL (CSM-S)
  - VMware vCenter Server

For more information, visit:

http://www.cisco.com/en/US/products/ps6904/tsd_products_support_series_home.html

6.4 Wide Area Application Services (WAAS)

As referenced previously, WAN optimization is another component to enhance application performance across the WAN. The workforce is increasingly global and distributed, and remote workers require high-performance access to application infrastructure to remain productive, achieve revenue goals, and meet customer satisfaction requirements. These requirements have traditionally caused IT departments to distribute application infrastructure like servers, storage, and data protection components to the branch offices to help ensure fast access to information. As storage requirements continue to increase and organizations face ever-evolving regulatory compliance requirements, organizations find themselves in a predicament resolved only by either expanding remote-office infrastructure and its security and data protection posture, or by consolidating that infrastructure from the branch offices into one or more data centers.

The challenge in consolidating infrastructure for immediate cost savings, in hardware, software, data protection, and operating expenses, is addressing the performance implications for the remote users. Access to information over a WAN is an order of magnitude slower than access to information over a LAN, due to limited WAN bandwidth, packet loss, and latency. The first step in controlling costs and enabling the intelligent data center is to reduce the effects of the performance boundary found in the WAN while bridging the gap between infrastructure consolidation and application performance. This challenge can be met through deployment of Cisco Wide Area Application Services (WAAS) Software.

The Cisco WAAS is software deployed on either end of a WAN link in both the remote offices and the data center using router-integrated network modules, virtual appliances or physical appliances. It is coupling intelligent application-layer acceleration with advanced compression and flow optimization capabilities. Cisco WAAS mitigates the negative effects of WAN conditions to provide remote users with LAN-like access to infrastructure in the data center. This LAN-like access allows IT departments to confidently consolidate servers, storage, and data protection infrastructure in the data center without compromising on the performance expectations that remote users have built based on years of having that infrastructure locally.

Cisco WAAS is a software- and hardware-integrated, cloud-ready WAN optimization and application acceleration solution. WAAS appliances offer *outstanding* deployment scalability and design flexibility while WAAS software delivers *best-in-class* application acceleration for the enterprise network. Cisco WAAS 5.0 provides public-cloud-based application acceleration; an elastic, “scale as you grow” enterprise-wide deployment model; and *industry-leading* scalability for secure acceleration of email, file, web, software-as-a-service (SaaS), video, and virtual desktop integration (VDI) applications. Cisco WAAS is able to improve end user experience and reduce bandwidth for key applications including but not limited to; Microsoft Exchange, Citrix XenApp and XenDesktop, SAP, Lotus® Notes®, NetApp SnapMirror, HTTP(S), and file.

Cisco WAAS enables organizations to accomplish crucial business objectives such as these:

- Secure, scalable enterprise-wide Bring Your Own Device (BYOD)
- High performance Virtual Desktop/Experience Infrastructure (VDI/VXI)
- Live and on-demand media applications such as web-casting, e-learning and digital signage
- High-performance public and private cloud services and Software-as-a-Service (SaaS) applications
- Improved application performance and end user experience for applications including web, e-mail, VDI, file, and cloud
- Reduced WAN bandwidth requirements and deferral of expensive bandwidth upgrades
- Reduced branch footprint through server and service consolidation
- Data center consolidation, virtualization, and automation

Cisco WAAS offers the most choices for WAN optimization with the broadest portfolio on the market today:

- Software-based WAN optimization solutions:
  - Cisco WAAS Software on the Cisco Integrated Services Router (ISR) Generation 2 (ISR G2) platform provides router-integrated, on-demand WAN optimization for branch offices. The Cisco Services-Ready Engine (SRE) Modules on the Cisco ISR G2 platform decouples software services from the underlying hardware and can deliver WAN optimization as an on-demand service as required by business objectives and IT budget. This approach makes better use of existing investments while offering business agility.
  - Cisco Virtual WAAS (vWAAS) is a virtual appliance that accelerates business applications delivered from private and virtual private cloud infrastructure, helping ensure an optimal user experience. Cisco vWAAS enables cloud providers to rapidly create WAN optimization services with little network configuration or disruption. vWAAS employs policy-based configuration in Cisco Nexus 1000V Series switches, which allows association with application server virtual machines as they are instantiated or moved. vWAAS also integrates via WCCP as well as AppNav as service insertion techniques.
  - Cisco WAAS Express extends the Cisco WAAS product portfolio with a small-footprint, cost-effective Cisco IOS Software solution integrated into Cisco ISR G2 devices to offer bandwidth optimization capabilities. Cisco WAAS Express increases remote user productivity, reduces WAN bandwidth costs, and offers investment protection by interoperating with existing Cisco WAAS infrastructure.
  - Cisco WAAS Mobile delivers bidirectional compression, application-specific accelerators, and flow optimizers for mobile and remote users in situations in which neither an appliance nor a branch-office router is available or practical, or for public cloud environments that cannot support an appliance.
- Full appliance portfolio:
  - The Cisco Wide Area Virtualization Engine (WAVE) appliances deployed in the branch office enable customers to meet the following business objectives:
    - Improve employee productivity by enhancing the user experience for important business applications delivered over the WAN
    - Reduce the cost of branch-office operations by centralizing IT resources in the data center and lowering the cost of WAN bandwidth as well as hosting Windows-based applications on the WAVE branch-office appliance

- Deliver enterprise-class video while minimizing WAN bandwidth consumption with the Cisco Wide Area Application Services (WAAS) Video Application Optimizer or hosted Cisco Application and Content Networking System (ACNS) Virtual Blade
- Increase IT agility by reducing the time and resources required to deliver new IT services to the branch office
- Simplify branch-office data protection for regulatory compliance purposes
- Flexible deployment models including inline and Web Cache Communication Protocol (WCCP) and APPNav for the highest performance, scalability, and availability.

Cisco WAVE data center appliances provide the high performance, scalable WAN optimization solutions require. Benefits include:

- User-selectable I/O modules with support for 10 Gigabit Ethernet fiber, 1 Gigabit Ethernet copper and 1 Gigabit Ethernet fiber
- Flexible deployment models including inline and Web Cache Communication Protocol (WCCP) for highest performance, scalability, and network availability
- Highest performance for video, Virtual Desktop Infrastructure (VDI), cloud, as well as traditional enterprise applications; in conjunction with Context-Aware Data Redundancy Elimination (DRE), the WAVE appliances can adapt caching behavior based upon the characteristics of individual applications, resulting in higher throughput and lower application latencies

More technical details on WAAS 5.0 can be found at these websites:

http://www.cisco.com/en/US/prod/collateral/contnetw/ps5680/ps6870/white_paper_c11-705319.html

http://www.cisco.com/en/US/products/ps6870/prod_white_papers_list.html

http://www.cisco.com/en/US/prod/collateral/contnetw/ps5680/ps6870/white_paper_c11-676350.html

http://www.cisco.com/en/US/prod/collateral/contnetw/ps5680/ps11231/technical_overview_c17-620098.html

WAAS hardware form factors are discussed here:

http://www.cisco.com/en/US/prod/collateral/contnetw/ps5680/ps6474/data_sheet_c78-685554.html

http://www.cisco.com/en/US/prod/collateral/modules/ps2797/ps9750/data_sheet_c78-605215.html

WAAS software is discussed here:

http://www.cisco.com/en/US/prod/collateral/contnetw/ps5680/ps6870/data_sheet_cisco_wide_area_application_services_mobile.html

http://www.cisco.com/en/US/partner/prod/collateral/contnetw/ps5680/ps11211/datasheet_c78-611644.html

6.4.1 Deployment options

Cisco WAAS provides flexible deployment options leveraging wccp, pbr, inline, vpath or clustering, as summarized in Figure 6-9 and Table 6-2.

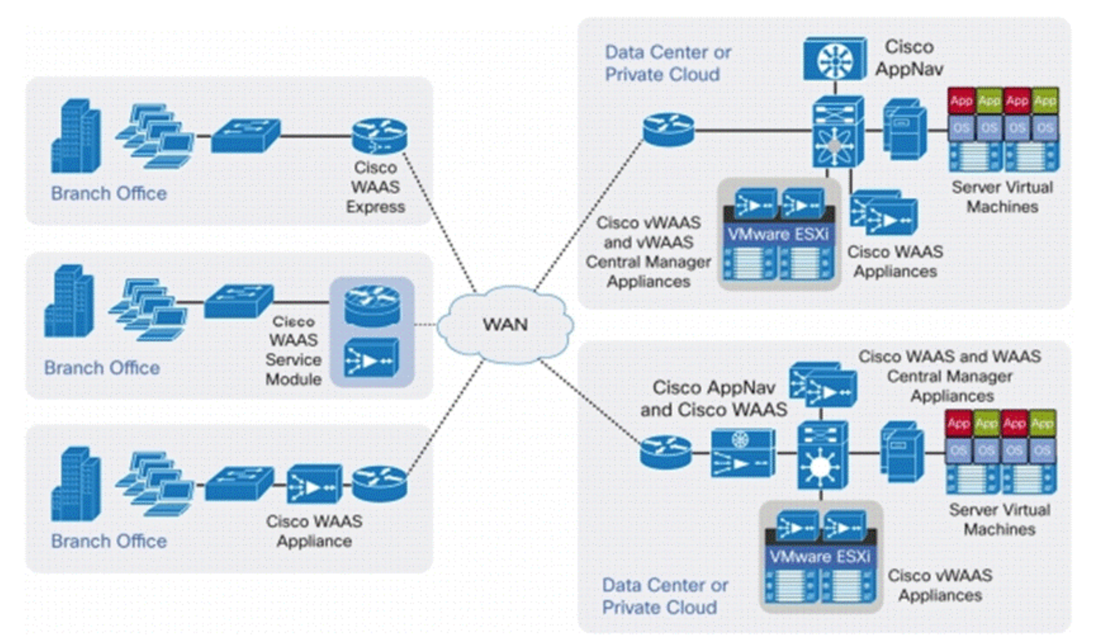


Figure 6-9 WAAS flexible deployment options

Table 6-2 Cisco WAAS product family

Deployment location	Cisco WAAS product family					
	Cisco WAAS Appliances	Cisco vWAAS	Cisco WAAS Modules on ISR and ISR G2	Cisco WAAS Express on ISR G2	Cisco WAAS Mobile Client	Cisco WAAS Mobile Server
Branch office	Yes	Yes	Yes	Yes	-	-
Data center	Yes	Yes	-	-	-	Yes
Private cloud, virtual private cloud, and public cloud	Yes	Yes	-	-	-	Yes
Mobile and home-office PCs	-	-	-	-	Yes	-

WAAS is a symmetric optimization solution; therefore, one requires a WAAS appliance at the branch and the DC to optimize application flows. The WAVEs act as a TCP proxy for both clients and their origin servers within the data center. This is not to be confused with other 3rd party WAN optimization solutions that create optimization tunnels. In 3rd party solutions, the TCP header is modified between the caching appliances. With WAAS, the TCP headers are fully preserved; therefore, network services, such as fws or IPs, have complete visibility of the original client/server connection.

Figure 6-10 shows three connections. TCP connection #2 is the WAAS optimization path between two WAVEs over a WAN connection. Within this path, Cisco WAAS optimizes the transfer of data between the client and server over the WAN connection, minimizing the data the client sends or requests. Traffic in this path includes any of the WAAS optimization mechanisms such as the transport flow optimization (TFO), Data redundancy elimination (DRE), Lempel-Ziv (LZ) compression and application specific optimizations. WAAS design guidance in the branch office is outside the scope of this design guide; however design guidance for the branch office can be found here:

http://www.cisco.com/en/US/docs/solutions/Enterprise/Branch/WAASBr11.html

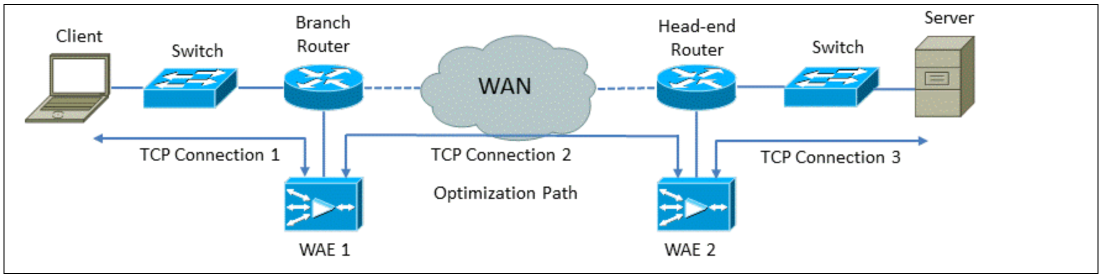


Figure 6-10   TCP connections using WAN

WAAS can be integrated between the client and server in a number of places in the data path. To achieve maximum benefits, optimum placement of the WAVEs between the client (source) and server (destination) is important. Prior to WAAS 5.0, the primary deployment methods were as shown in Table 6-3: WCCP was the most common method of deploying WAAS in the DC. With the release of WAAS 5.0, the focus of this section is on AppNav as the WAAS service insertion technique within the data center as it provides unprecedented scale, policy definition, pooling, and clustering.

Table 6-3   Prior deployment methods

Method	**Comment**
WCCP	Used for transparent interception of application traffic. Used in branch offices and data centers to transparently redirect traffic to the local WAE. The traffic is transparently intercepted and redirected to the local WAE by a WCCP-enabled router or a Layer 3 switch.
PBR	In branch offices, used for wide area application optimization. The branch office router is configured to use PBR to transparently intercept and route both client and server traffic to the WAE that resides in the same branch office. In data centers, used for data center application optimization. The data center router or L3 switch may be configured to use PBR to transparently intercept and route client and server traffic to WAE(s) within the data center. PBR, however, does not support load balancing across multiple WAEs (such as WCCP does).
Inline	Used for transparent interception of application traffic with no traffic redirection configuration requirement on switch/router.
ACE/CSM	Cisco Application Control Engine (ACE) or Catalyst 6500 series Content Switching Module (CSM) installed in the data center for data center application optimization. The ACE or CSM allows for both traffic interception and load balancing across multiple WAE(s) within the data center.

6.5 Cisco APPNAV building blocks

The Cisco AppNav solution is a comprehensive and network-integrated WAN optimization solution that drastically reduces operational complexity in enterprise wide WAN optimization deployments. Cisco AppNav technology enables customers to virtualize WAN optimization resources in the data center by pooling them into one elastic resource in a manner that is policy based and on demand with the best available scalability and performance.

It integrates transparently with Cisco WAAS physical and virtual network infrastructure, (vWAAS) supporting more than a million connections, providing significant investment protection for existing network design objectives as well as the capability to expand the WAN optimization service to meet future demands. Because the Cisco AppNav solution enables organizations to pool elastic resources, it lays the foundation for migration to cloud services as well.

Cisco AppNav technology provides the following benefits:

- Offers efficient and cost-effective expansion of WAN optimization services within a company: The Cisco AppNav solution provides an on-demand pool of Cisco WAAS instances (consisting of appliances or virtual appliance or a mixture of both) that can be used in all data center deployment models.
- Provides flexibility in WAN optimization deployment to address various business needs: The Cisco AppNav solution enables the use of flexible policy definitions that dynamically bind business logical constructs to a set of Cisco WAAS pools, and it makes intelligent flow distribution decisions based on the state of the nodes currently providing services.
- Improves business continuity by providing highly available Cisco WAAS services: The Cisco AppNav clustering mechanism allows redundancy in the pool of Cisco WAAS instances and Cisco AppNav services
- Facilitates migration to the cloud: Cisco AppNav elastic provisioning and deployment of WAN optimization services and the upcoming integration of Cisco AppNav technology with the Cisco Cloud Services Router (CSR) provides a transparent and optimized connection from the branch office to the cloud.
- Addresses challenges posed by today's data center infrastructure: Cisco AppNav native solutions address directional asymmetry in the data center multipath environment while preserving the user's intended network path (network path affinity).

Cisco AppNav technology is deployed on the new-generation Cisco Wide Area Virtualization Engine (WAVE) appliances in the data center or private cloud. The 1-Gbps solution is offered with two types of I/O modules (IOMs) in copper and Small Form-Factor Pluggable (SFP) form factors that can be plugged into Cisco WAVE 8541, 8571, 7541, and 694 devices as shown in Figure 6-11. These solutions can run purely as Cisco AppNav solutions or as Cisco AppNav and Cisco WAAS on the same platform. The 10-Gbps package runs exclusively in Cisco AppNav mode, consists of the chassis as well as AppNav IOM and provides high performance, high availability (dual power supply), and cost effectiveness in a single form factor.

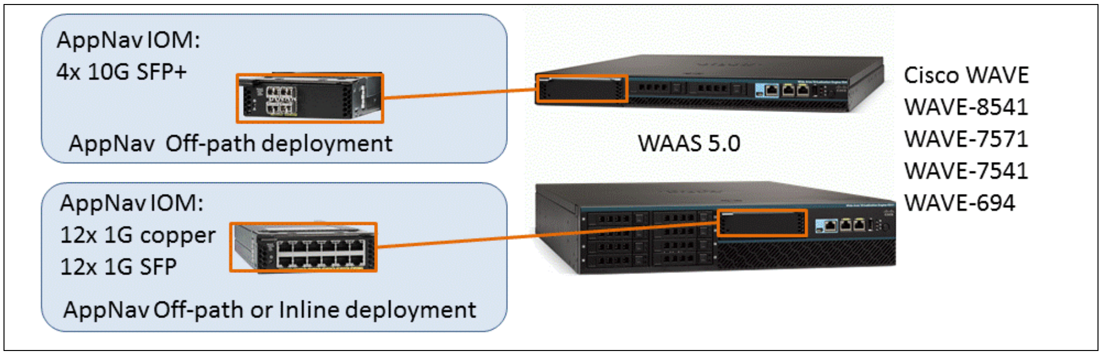


Figure 6-11 Cisco WAVE options

The Cisco AppNav solution can be deployed out of the data path or in the data path, depending on user business and network requirements. Cisco AppNav in-path and off-path deployments provide a comprehensive yet similar level of solution flexibility, scalability, and availability, summarized in Table 6-4:

Table 6-4 Cisco AppNav solution

	Cisco AppNav Solution	
	In-Path	Off-Path
Hardware-based traffic classifier	✓	✓
Hardware-based flow distribution	✓	✓
Scalability up to 8 Cisco AppNav and 32 Cisco WAAS devices (more than 1 million connections)	✓	✓
High availability (AppNav and WAAS)	✓	✓
Asymmetry handling with path preservation	✓	✓
Elastic pooling mix of physical and virtual Cisco WAAS Appliances	✓	✓

The differentiation between in-path and off-path mainly reflects the network characteristics of the deployment type:

- Cisco AppNav in-path: Reduces configuration changes required on adjacent network devices
- Cisco AppNav off-path: Eliminates network throughput bottlenecks that can result from changing interface speeds; less disruptive to deploy and add capacity during the lifecycle of the deployment

Figure 6-12 shows examples of Cisco AppNav in-path and off-path deployments.

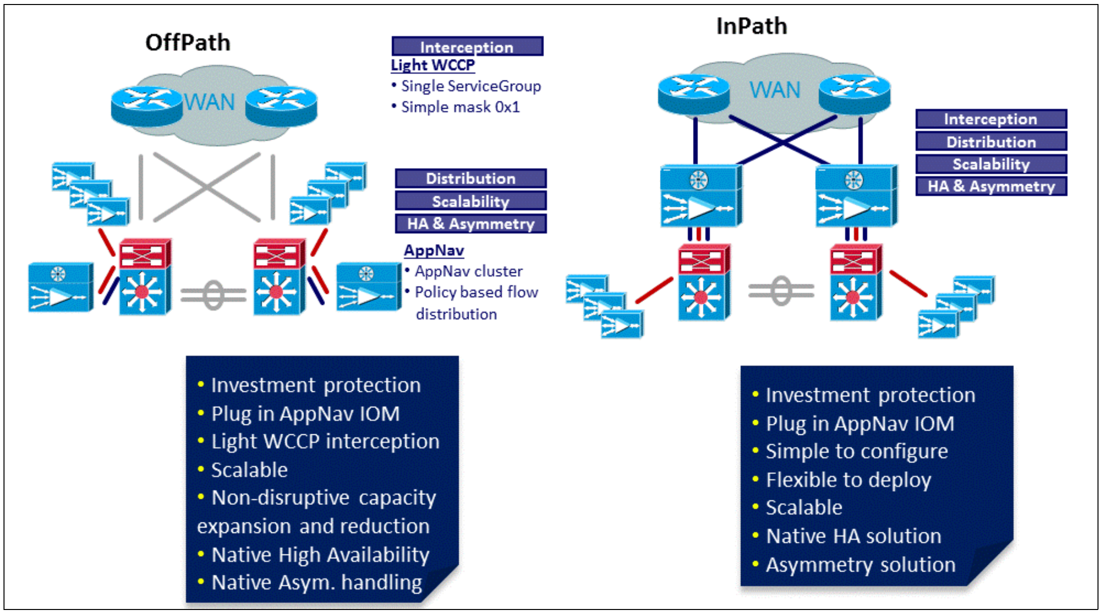


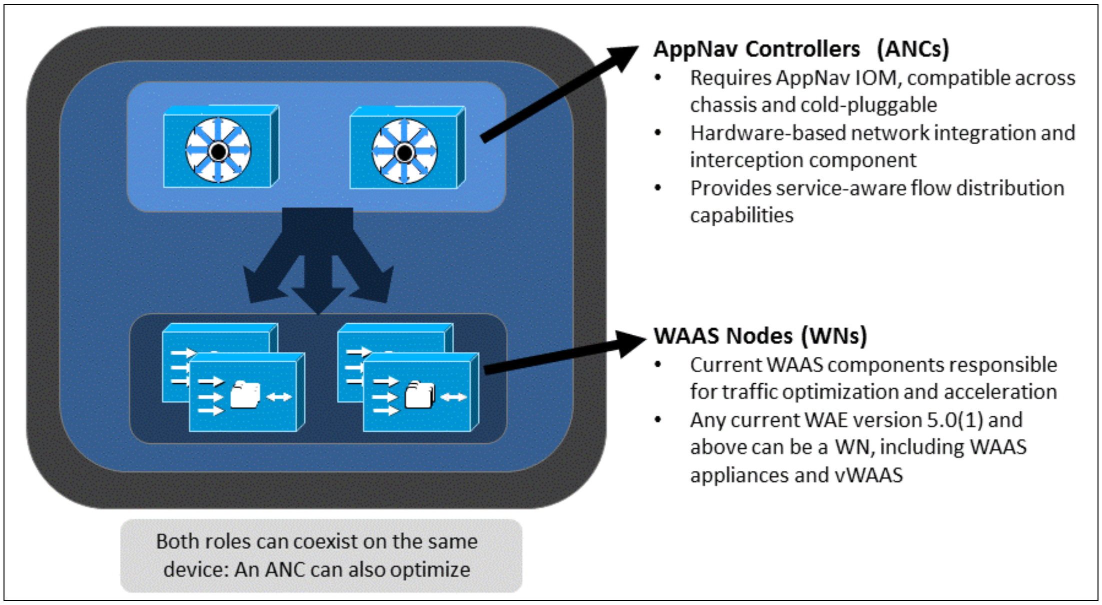
Figure 6-12 In-path and off-path deployment

The following section takes a more detailed look at each one of the AppNav building blocks.

The Cisco AppNav solution can apply flexible and intelligent policies to address the dynamic needs of WAN optimization. These policies can distribute the user traffic flows based on business constructs (such as application or remote location) and deploy new groups of Cisco WAAS devices on demand or based on dynamic load feedback from active Cisco WAAS devices. The Cisco AppNav solution also provides a clustering mechanism that can address the network resiliency that highly available business services need and natively address the challenges posed by today's multipath data centers, which may cause directional asymmetry of user traffic.

The first two building blocks are AppNav Controllers (ANC) and WAAS Nodes (WN). AppNav Controllers are nothing more than existing WAVE devices, with a required hardware in the form of the AppNav I/O module (AppNav IOM). This module provides the necessary data plane resources, so that the WAVE can provide the critical functions of hardware-based interception, redirection, and flow distribution. WAAS Nodes perform the optimization function, and can be pooled together seamlessly. A flexible approach is used, where designers can use physical or virtual appliance form factors and any WAAS model, as long as it runs WAAS Software version 5.0 or above.

It is worth mentioning that both roles, that of the controller and that of the node, can coexist in the same device: An AppNav Controller (ANC) can perform its functions and at the same time optimize like a WN, see Figure 6-13.



AppNav Controllers (ANCs)

- Requires AppNav IOM, compatible across chassis and cold-pluggable
- Hardware-based network integration and interception component
- Provides service-aware flow distribution capabilities

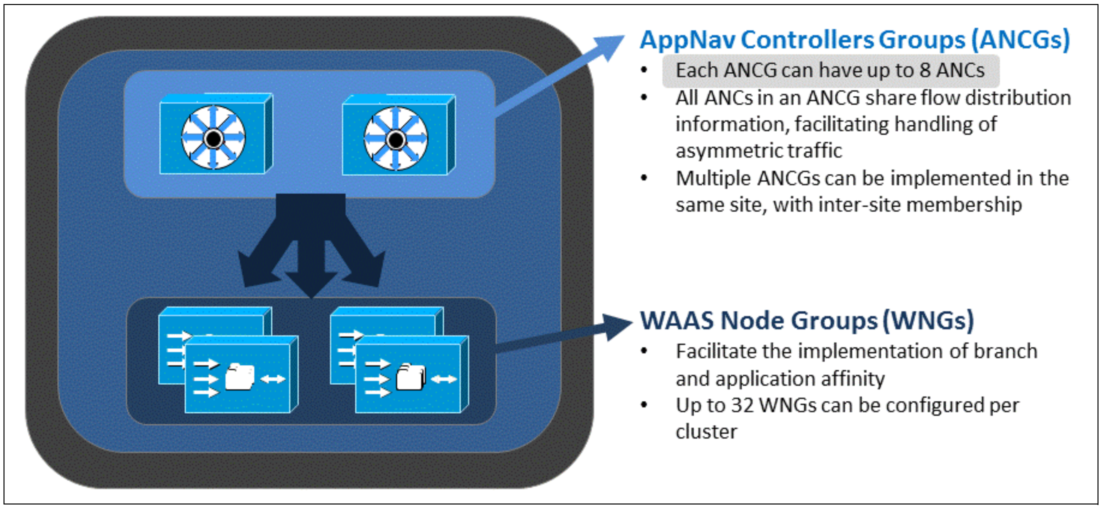
WAAS Nodes (WNs)

- Current WAAS components responsible for traffic optimization and acceleration
- Any current WAE version 5.0(1) and above can be a WN, including WAAS appliances and vWAAS

Both roles can coexist on the same device: An ANC can also optimize

Figure 6-13 AppNav controllers and WAAS nodes

All AppNav building blocks can be grouped logically to accomplish some of the most critical functions. For instance, ANCs can be logically grouped into ANCGs, where each ANCG supports up to eight ANCs as shown in Figure 6-14. This is important for sizing. Remember, all ANCs in the group share flow distribution information, so any one ANC can distribute the flow of traffic independently. WNGs facilitate intelligent flow distribution by pooling optimization resources according to business needs. Up to 32 WWNs can be allocated to a WNG, providing ample scalability based on the specific WAVE model used as a WN.



AppNav Controllers Groups (ANCGs)

- Each ANCG can have up to 8 ANCs
- All ANCs in an ANCG share flow distribution information, facilitating handling of asymmetric traffic
- Multiple ANCGs can be implemented in the same site, with inter-site membership

WAAS Node Groups (WNGs)

- Facilitate the implementation of branch and application affinity
- Up to 32 WNGs can be configured per cluster

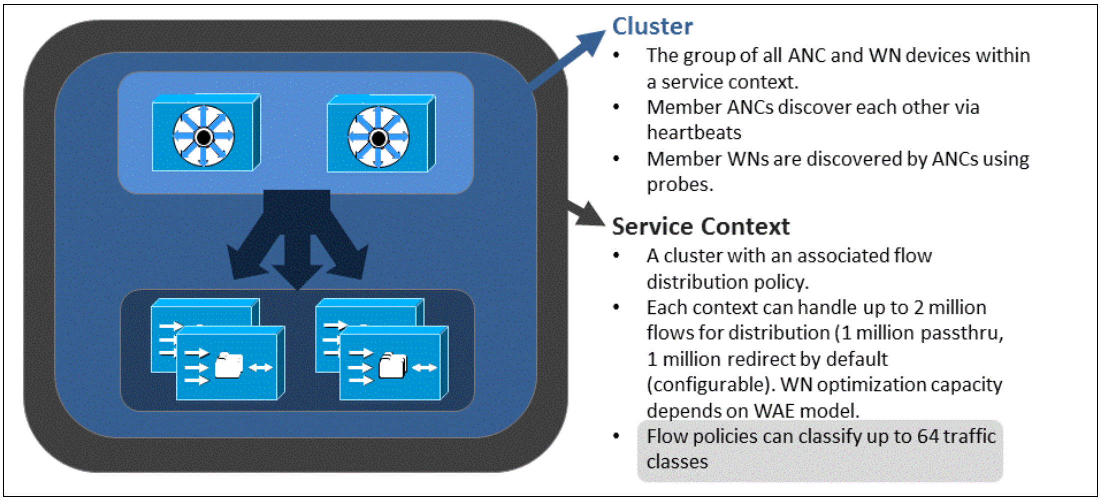
Figure 6-14 ANCGs and WNGs

Clusters are cohesive units that integrate all required functions. WAAS resources and devices acquire different roles based on the functionality they implement. The AppNav Controllers (ANCs) intercept, redirect, and provide load distribution of traffic flows that require WAAS optimization. ANCs are nothing more than WAVEs with the specific hardware that implements the AppNav functions.

The traditional WAN optimization function is the responsibility of WAAS Nodes (WNs). These are the traditional WAVEs implementing WAAS optimization, including appliances and vWAAS. ANCs can be grouped into AppNav Controller Groups (ANCGs). This logical bundle makes ANCs shares resources in such a way that high availability and traffic asymmetry are covered from the perspective of ANC functions. Similarly, WNGs or WAAS Node Groups bundle WNs to create pools of optimization resources that can be made available automatically through elastic provisioning.

It is the combination of ANCGs and WNGs, along with their member devices, that comprise an AppNav Cluster as shown in Figure 6-15. Clusters are cohesive, in that liveliness and resource availability are controlled automatically by the cluster itself. But remember, the main objective here is to intercept and redirect traffic for optimization. All of these functions are triggered by special-purpose flow distribution policies, geared toward implementing business rules.

When such policies are applied to an AppNav Cluster, the end result is a service context, the highest level component of an AppNav solution.



Cluster

- The group of all ANC and WN devices within a service context.
- Member ANCs discover each other via heartbeats
- Member WNs are discovered by ANCs using probes.

Service Context

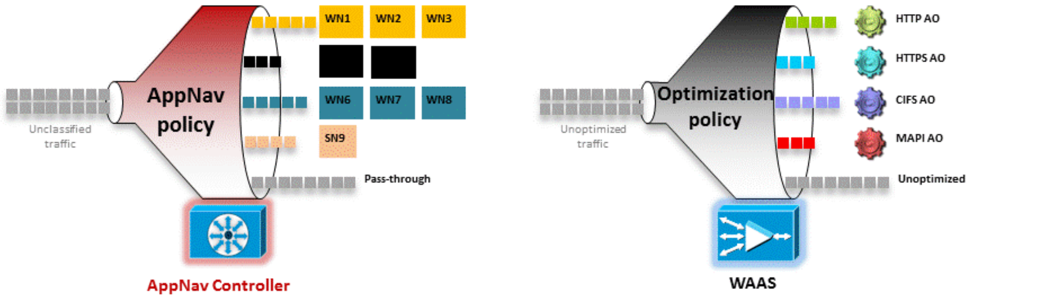
- A cluster with an associated flow distribution policy.
- Each context can handle up to 2 million flows for distribution (1 million passthru, 1 million redirect by default (configurable). WN optimization capacity depends on WAE model.
- Flow policies can classify up to 64 traffic classes

Figure 6-15 AppNav cluster and service context

6.5.1 Intelligent flow distribution

The last AppNav building block is the flow distribution policies themselves. Remember, the objective of WAAS is to optimize network traffic. Traditional policies implement optimization by using class maps to identify the optimization action to apply to each traffic class.

Their actions are related to transport optimization, compression and data reduction, and application layer acceleration. With AppNav, WAAS functionality expands to interception, redirection, and flow distribution. As such, a new type of policy is required, deployed on AppNav controllers. As shown in Figure 6-16, a consistent framework is used to include AppNav class maps that classify traffic to intercept and redirect, based on a rich set of parameters, such as IP addresses and ports, or even the ID of peering WAAS devices. The actions are now related to the specific AppNav such as to redirect traffic to primary or backup pools of WAAS Nodes, or implement complex rules using nested policies.



AppNav policy

Flow distribution policy (AppNav -> WAAS)

- AppNav Classmap traffic classification
  - Per peer OR 3 tuple SRC/DST IP and DST port
  - Mix one peer + SRC/DST IP and DST port
- AppNav Policy
  - Primary (& backup WNGs)
  - Monitoring of Aos
  - Nested policy

Optimization policy

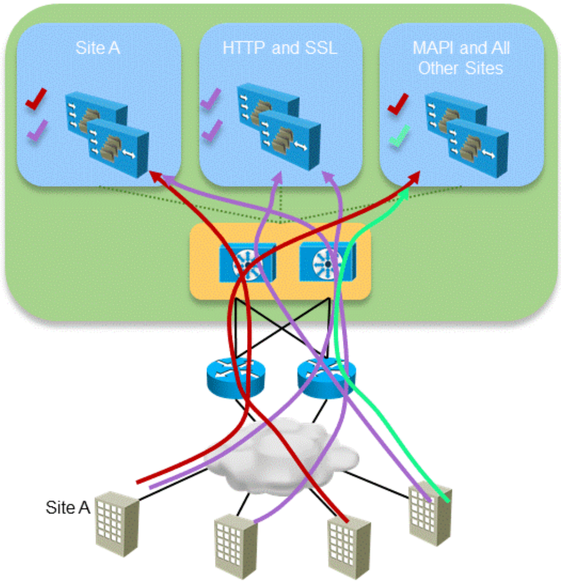
Optimization/accelration associated with given traffic (WAAS)

- Optimization Classmap identifying traffic
- Optimization/acceleration tied to the traffic
  - TFO
  - LZ
  - DRE (Uni, Bi, Adaptive)
  - AO

Figure 6-16 AppNav Policy - optimization policy

By pooling the optimization resources in WAAS Node Groups, those resources can be granularly sized and dedicated to service a variety of traffic flows, clearly classified and defined by business-driven policy rules.

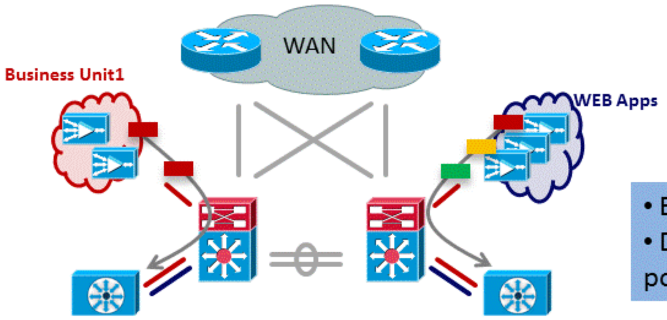
For instance, site or branch affinity can be accomplished simply by matching traffic flows by IP address or branch WAE identifier. In Figure 6-17, the critical Site A can have reserved capacity and have optimal DRE performance by pinning all branch traffic to the same pool of WAAS Nodes. Similarly, service affinity can be accomplished by classifying traffic granularly based on several packet components, such as IP addresses and ports. High-volume applications can have reserved capacity, based on the number of WAAS Nodes in the WNG pool. Application-specific optimization options can be enabled for the specific application. A hybrid approach can also be accomplished by combining site and service affinity in specific WNGs.



- Site affinity:
  - Identified via branch WAE ID or site IP subnet
  - Reserve optimization capacity for critical sites
  - Improves compression performance through DRE
- Application affinity:
  - Identified via source/destination IP addresses and ports
  - Reserve optimization capacity for applications
  - Consolidates application-specific optimization options
- Site + Application (combination)

Figure 6-17 Intelligent flow distribution

Whichever WNG and policy topology is implemented, the end result is a deterministic and highly granular flow distribution approach, one that can be changed quickly by simply changing the business-driven policy rules, as shown in Figure 6-18.



- Business driven bindings
- Dynamic service aware policy

- Logical bindings for WAAS pools
- **Dynamic service aware** flow distribution
  - 🟩 **Green:** WAAS accepts new and existing connections
  - 🟨 **Yellow:** WAAS only accepts existing connections
  - 🟥 **Red:** WAAS can not accept any connections

Figure 6-18 Pool of WAN optimization

6.5.2 APPNav design options

The first design consideration is how to integrate AppNav to the data center topology. This consideration requires knowledge of the roles of interfaces in an AppNav device. Interfaces on the AppNav Controller interface module can have three functions.

Interception is used to receive traffic intercepted from the network and egress traffic to the network as described in Figure 6-19. The interception interface is implied based on the AppNav Controller placement. It does not require explicit configuration for this function.

Distribution interfaces, on the other hand, are used to distribute traffic to the WNs and receive egress traffic from them. The distribution interface is explicitly configured as the cluster interface for intra-cluster traffic, and must be assigned an IP address. A given interface can fulfill the interception and distribution roles at the same time. This role requires explicit configuration of the distribution role.

Finally, management interfaces can be optionally and exclusively designated for management traffic, and isolated from the normal data path. Recommended practices call for the use of one of the appliance's built-in interfaces for management traffic, while reserving the high performance interfaces on the AppNav Controller IO Module for interception and distribution.

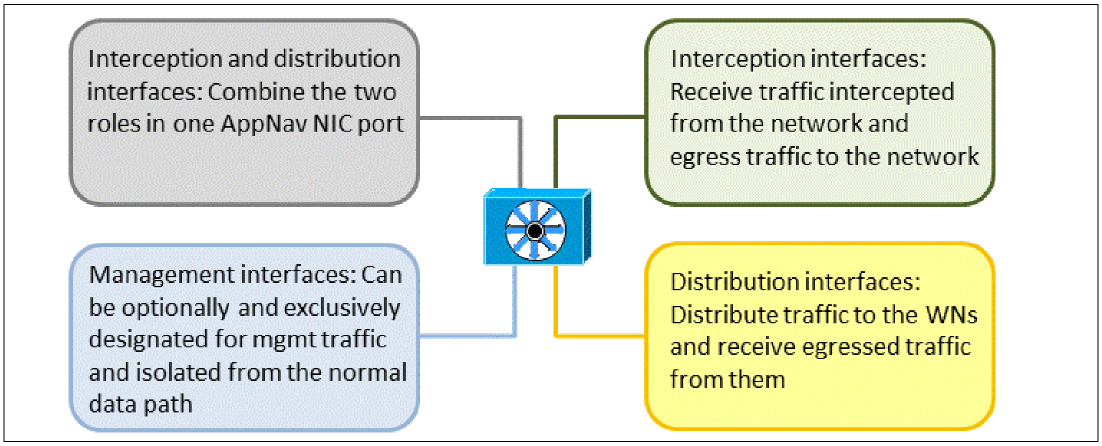

Interception and distribution interfaces: Combine the two roles in one AppNav NIC port
Interception interfaces: Receive traffic intercepted from the network and egress traffic to the network
Management interfaces: Can be optionally and exclusively designated for mgmt traffic and isolated from the normal data path
Distribution interfaces: Distribute traffic to the WNs and receive egressed traffic from them

Figure 6-19 AppNav interface options

AppNav deployments need to accommodate interception, distribution, and, optionally, management interfaces. Figure 6-20 illustrates the topology, requirements, and recommendations for inline deployments. WAAS Nodes are always off-path, regardless of the controller's deployment option. Interception interfaces must behave in bridge mode, which means that at least one inline bridge group, configured on the AppNav Controller IO Module, is required. A bridge group consists of two or more physical or port channel interfaces. These interfaces may or may not be identical. This helps to account for the compression factor and the disparity in throughput on LAN and WAN sides due to it. Functionally, bridge groups are very similar to inline groups in the previous generation of inline WAE devices.

Link state propagation is supported on the bridge group, and is a configurable option. With link state propagation, when one bridge group interface goes down, the other interface on the same bridge group is brought down as well. Distribution interfaces must be different than the interception interfaces when deployed in inline mode.

For simplicity, it is recommended to place all WNGs in the same VLAN, unless traffic segregation is required for the optimization function for the sake of security. Private cloud environments will benefit from multiple VLANs for different WNGs.

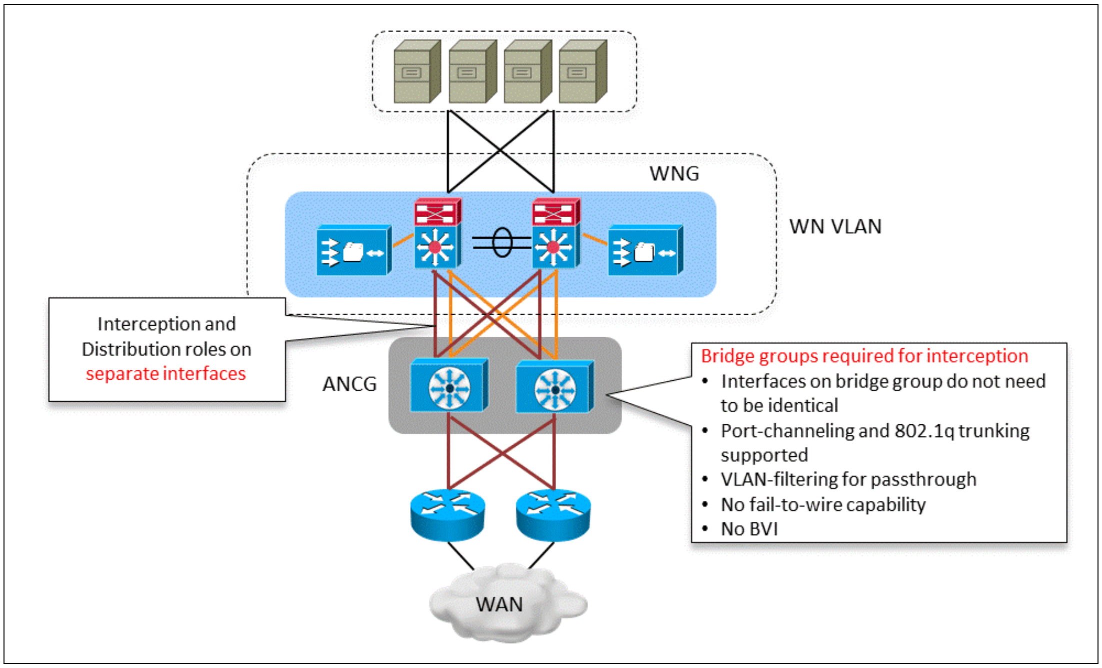
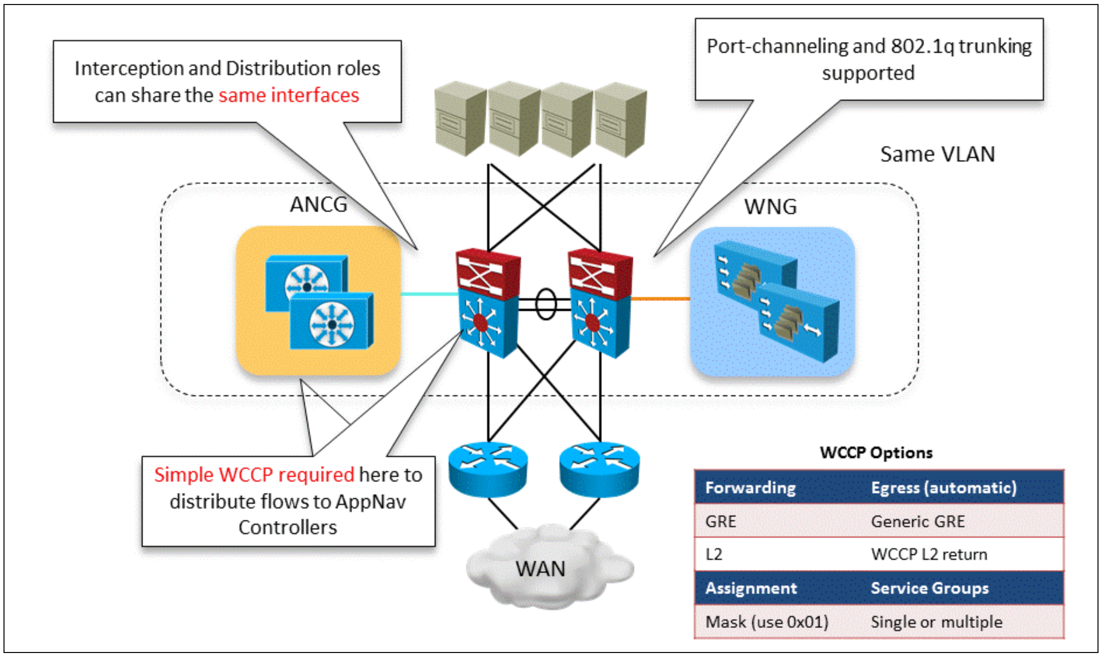

WNG
WN VLAN
Interception and
Distribution roles on
separate interfaces
ANCG
Bridge groups required for interception
Interfaces on bridge group do not need to be identical
Port-channeling and 802.1q trunking supported
VLAN-filtering for passthrough
No fail-to-wire capability
No BVI
WAN

Figure 6-20 AppNav inline deployment

Off-path deployments require different requirements and recommendations. In this environment, AppNav Controllers operate in off-path mode for interception and redirection, and WAAS Nodes operate in off-path mode for optimization. For interception and redirection, this mode requires the use of simple WCCP to redirect traffic from their natural path into the AppNav Controllers. Simple means the use of a single WCCP service group, with a simple destination-based mask, and Layer 2 rewrites as the forwarding and egress method.

One should change to GRE if the AppNav Controllers are not Layer 2 adjacent to the WCCP routers. More flexibility is allowed in off-path mode, as the use of bridged interfaces is not a requirement. Interception and distribution roles can reside on the same set of interfaces. The recommended design for this situation is one physical or logical interface, for simplicity, configured with an IP address. The interface or interfaces fulfill both interception and redirection roles.

WNs typically reside in the same VLAN as ANCs. Port channeling and 802.1q trunking are supported on WAAS Nodes to allow for bandwidth aggregation and scalability. Figure 6-21 illustrates an example with only one WNG, but multiple WNGs can be deployed according to capacity and isolation requirements.


Interception and Distribution roles can share the same interfaces

Port-channeling and 802.1q trunking supported

Simple WCCP required here to distribute flows to AppNav Controllers

WCCP Options

Forwarding	Egress (automatic)
GRE	Generic GRE
L2	WCCP L2 return
Assignment	**Service Groups**
Mask (use 0x01)	Single or multiple

Figure 6-21 Off-path deployment

6.5.3 HA considerations

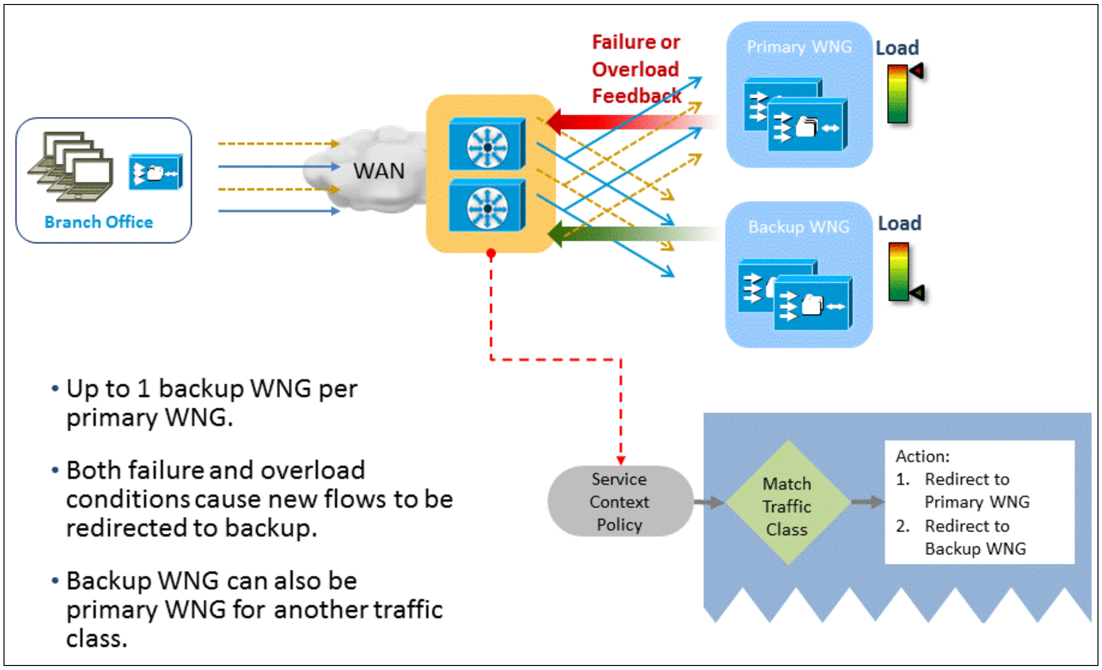
High availability for both AppNav Controllers and WAAS Nodes is an integral part of the architecture of cohesive clusters. Probes and heartbeats allow for seamless operations in the presence of device and group failures.

An optional feature, in the form of backup WAAS Node Groups, is available to deploy an additional high availability mechanism for individual WNGs, or for multiple WNGs at the same time. Using backup WNGs, flow distribution policies are configured to redirect specific traffic classes to a primary WNG, but also to a backup WNG in case of failure or overflow.

Both primary and backup WNGs report their status and, optionally, load. If all WNs in the primary farm report a red or yellow status, indicating they are down or overloaded, new traffic flows will be redirected to the backup WNG until the primary farm is back alive. Up to one backup WNG is supported per primary WNG. Both failure and overload conditions cause new flows to be redirected to backup. Backup WNGs can also be primary WNGs for another traffic class, allowing for N+1 environments where one backup farm can back up multiple primary farms.

Remember, WN monitoring is not enabled by default. Plan to enable Application Accelerator monitoring for traffic classes that will benefit from load-based flow distribution at the Application Accelerator level. Consider that only one Application Accelerator can be enabled for Application Accelerator monitoring per class per policy map. The most critical application for load-based distribution should be selected. Consider ANC and WN high availability as a separate design exercise. ANC scalability relates to interception and redirection, with a maximum threshold of one million redirected flows per cluster. WN scalability relates to the ability to optimize, which is different per platform and model of the selected WAAS Node. It is also subject to per-Application Accelerator considerations.

Consider the implications of oversubscription when defining number of ANCs and WNs, especially for WNs and WNGs. Remember, unless one define backup WNGs, once a farm reaches capacity, all new flows will be passed through. Consider one backup WNG per mission-critical optimization farm. Also consider using one or more WNGs as backups for all farms for N+1 high availability, especially in enterprise data center environments. In private clouds, the recommended strategy is to have backup WNGs per customer, if SLAs call for this type of resiliency. Finally, consider using backup WNGs as a scalability/spillover tool. Remember, the backup farm will be used in the case of primary overload as shown in Figure 6-22, in addition to cases of failure conditions.



- Up to 1 backup WNG per primary WNG.
- Both failure and overload conditions cause new flows to be redirected to backup.
- Backup WNG can also be primary WNG for another traffic class.

Figure 6-22 Backup WNGs

6.5.4 Cluster sizing

Sizing is one of the first considerations in designing AppNav solutions. Figure 6-23 illustrates the sizing numbers for the different building block components of an AppNav Cluster:

- At the service context level, only one flow distribution policy can be active at the same time in an AppNav Controller. Remember, however, that the policy can have multiple actions for different traffic classes.
- The service context includes AppNav Clusters. Up to one million redirected flows are supported per cluster, with no regard to the number of ANCs in the cluster. A maximum limit of another million connections can be passed through.
- Each cluster includes ANCGs and WNGs. For ANCGs, the only limitation at this point is the maximum one ANCG per cluster. This requirement has implications in multi-data center environments, and will be expanded in the asymmetry section of this lesson.
- Each ANCG contains ANCs, up to eight per ANCG. All ANCs in the group share the flow distribution tables, which makes the maximum connections in the ANCG (and the cluster) up to one million optimized, one million passed through. The throughput of each ANC is 10 Gbps and 15 million packets per second.
- Up to 32 WNGs can be configured per cluster, with up to 32 WNs per group. The sizing of each WN depends on the form factor and model of the node.

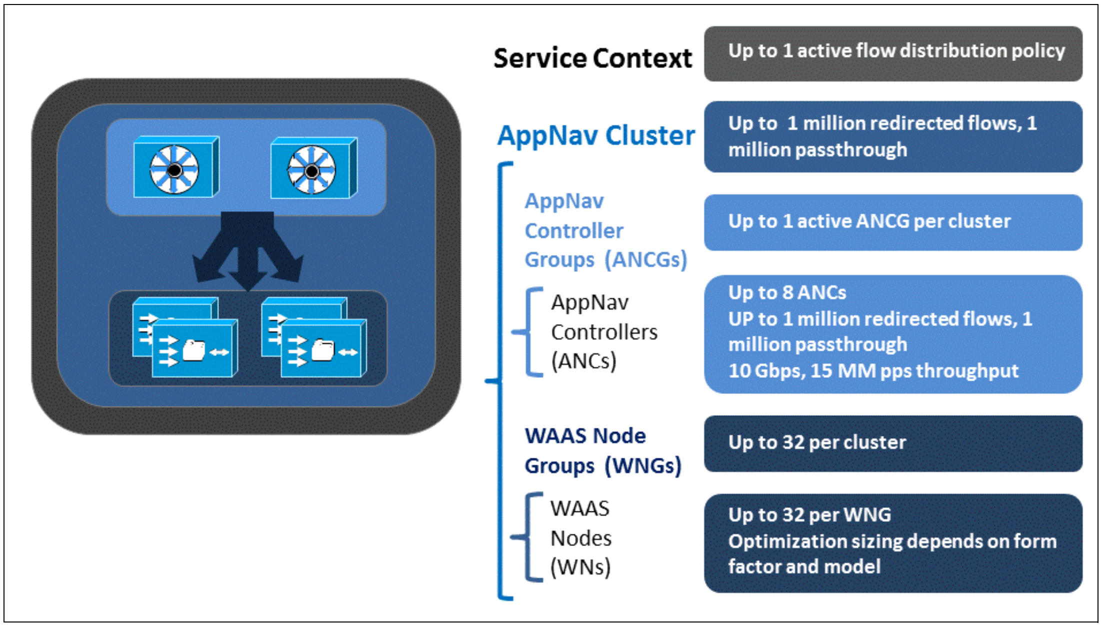

Service Context
Up to 1 active flow distribution policy
AppNav Cluster
Up to 1 million redirected flows, 1 million passthrough
AppNav Controller Groups (ANCGs)
Up to 1 active ANCG per cluster
AppNav Controllers (ANCs)
Up to 8 ANCs
UP to 1 million redirected flows, 1 million passthrough
10 Gbps, 15 MM pps throughput
WAAS Node Groups (WNGs)
Up to 32 per cluster
WAAS Nodes (WNs)
Up to 32 per WNG
Optimization sizing depends on form factor and model

Figure 6-23 Cluster sizing and AppNav scalability

Here are some additional considerations for AppNav scalability:

- AppNav Controller scalability within the cluster is for throughput and resilience. The cluster will support up to 1 million flows with no regards to the number of ANCs (up to 8).
- Create WNGs to warm DRE cache according to affinity (per site in enterprise data centers, per customer in private cloud, and so on).
- Use backup WNGs also for spill-over scalability, but consider cold cache conditions, specially when using one backup WNG for multiple farms.
- Size each WNG according to affinity requirements (branch, application, hybrid), based on traditional optimization sizing (see WAAS sizing tool at this website):

  `http://tools.cisco.com/WAAS/sizing/`
- There is no performance impact when an AppNav Controller is also a WAAS Node.

6.5.5 Flow distribution considerations

Traffic bypass options are one of the first functions to consider, especially in proof of concept scenarios. Bypassed traffic will be passed through, a useful condition when exceptions are created to the optimization policy, or in pilot and test deployments when the focus is on a critical or specific traffic class. Bypass, and the subsequent pass-through condition, can be accomplished at several levels and on several devices of the AppNav architecture.

The traffic bypass options are shown in Figure 6-24. First, WCCP access control lists can be defined to prevent flow redirection from WCCP routers into AppNav Controllers. This option is only useful in off-path deployments. For inline mode, but also for off-path, both VLAN pass-through and intercept ACLs are available. In both cases, the AppNav Controller is configured to consider either certain VLANs, in the case of VLAN pass-through, or certain traffic classes as defined by their IP components, in the case of interception ACLs. IP components include source and destination IP addresses and ports.

For more granularity and individual exceptions per server or client, optimization policies can be modified to include the pass-through action. In this case, traffic is intercepted and redirected, but the WAAS Node does not optimize the traffic flow.

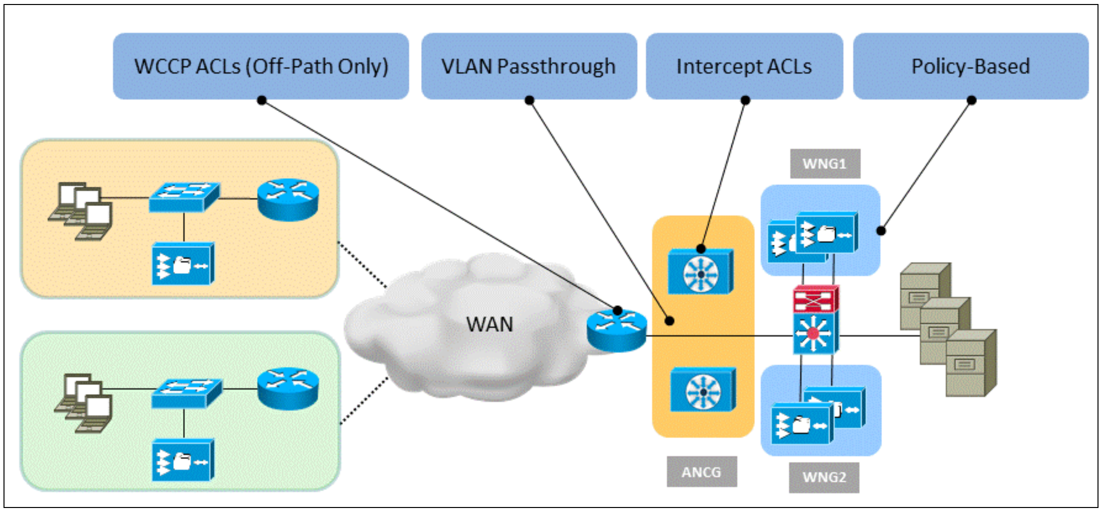


Figure 6-24 Traffic bypass options

Other flow distribution considerations

Here we list some other considerations regarding flow distribution:

- Use interception ACLs when possible as a bypass tool. It is the most granular and efficient way to create pass-through exceptions.
- Use peer ID in site affinity scenarios for simplified processing.
- Use 3-tuple (triple or triplet) of source IP/port and destination IP in branch affinity scenarios for accuracy.
- Used nested policies for modularity to combine site and application affinity.
- Traffic classification: consider per-destination matching for applications that negotiate ports dynamically.
- Consider using different VLANs per WNG in Private Cloud environments to further segregate customer traffic at Layer 2.
- Monitor Application Accelerators for load-based flow distribution. Only TFO (connection load) is monitored by default.

The flow distribution is accomplished with the help of flow distribution policies.

Figure 6-25 illustrates a logical example for an enterprise data center environment, where requirements call for a hybrid approach to traffic classification, pinning critical sites to a WNG, critical applications to another WNG, and all other traffic to a third WNG. The application, FTP in this example, requires special attention. It negotiates ports, and destination IP addressing is necessary for return traffic to match the class, because the FTP server is in active mode. It originates a data connection back to the FTP client using dynamically negotiated destination ports.

In Figure 6-25, a sample policy is shown. The first class matches traffic coming from critical sites using the peer WAE ID for each branch. For this and all other actions in this policy, the same backup WNG will be used for resiliency and overload conditions. The second class map matches ingress FTP traffic by matching FTP connection port 21, redirecting it to the WNG-FTP farm. The next class map completes the configuration related to FTP. It matches egress traffic, going back to FTP clients outside the data center, by matching the FTP client's IP addresses and the FTP server's source address. This is required to optimize traffic on FTP negotiated ports. The rest of the traffic is pinned to a WNG-OTHER farm.

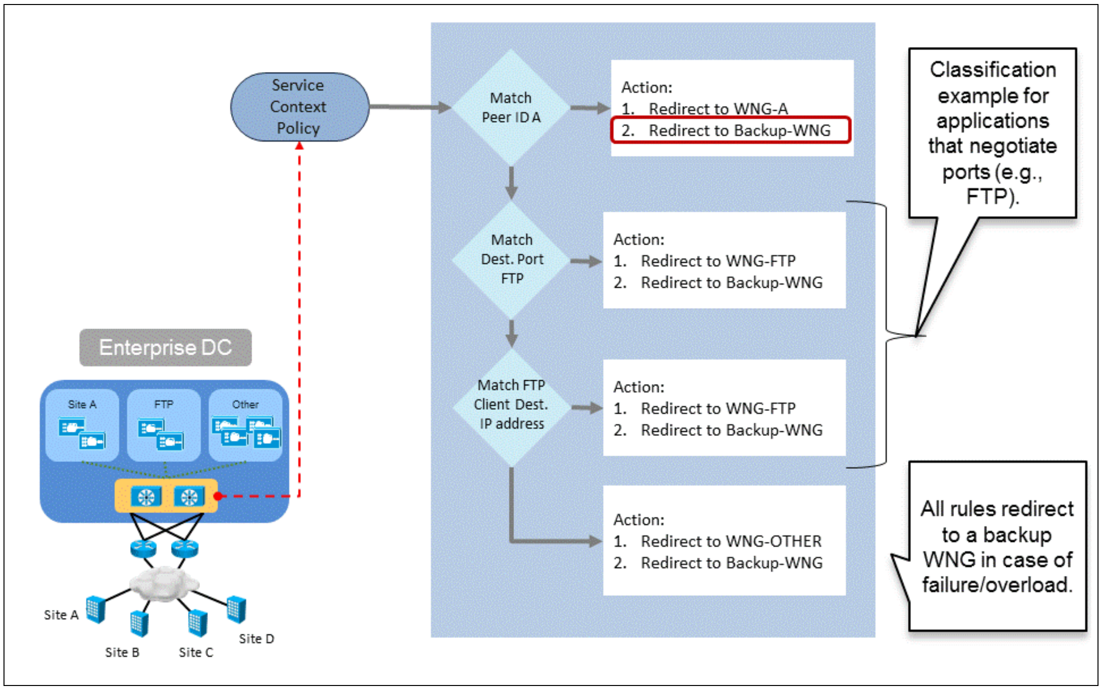

Service Context Policy
Match Peer ID A
Action:
1. Redirect to WNG-A
2. Redirect to Backup-WNG
Match Dest. Port FTP
Action:
1. Redirect to WNG-FTP
2. Redirect to Backup-WNG
Match FTP Client Dest. IP address
Action:
1. Redirect to WNG-FTP
2. Redirect to Backup-WNG
Action:
1. Redirect to WNG-OTHER
2. Redirect to Backup-WNG
Classification example for applications that negotiate ports (e.g., FTP).
All rules redirect to a backup WNG in case of failure/overload.
Enterprise DC
Site A
FTP
Other
Site A
Site B
Site C
Site D

Figure 6-25 Flow distribution example: Enterprise DC

Figure 6-26 shows an example of a private cloud policy.

The policy uses nested policies to implement per-branch, per-application categories, redirecting different application traffic to separate WNGs for each site. A primary policy is used to classify traffic for each site, using the peer WAE ID for simplicity in identifying flows sourced at each site.

Each class map will invoke a nested policy to implement the per application flow distribution strategy for each site. The class maps, or subclass, in this nested policy match applications based on destination ports mostly, and pin each application to a separate WNG. One of the actions per application is to monitor the load of the application on the WNs, allowing for Application Accelerator-aware flow distribution and high availability. The policy for each site will have a similar structure, matching applications, monitoring their load on the Application Accelerator, and pinning them to specific WNGs.

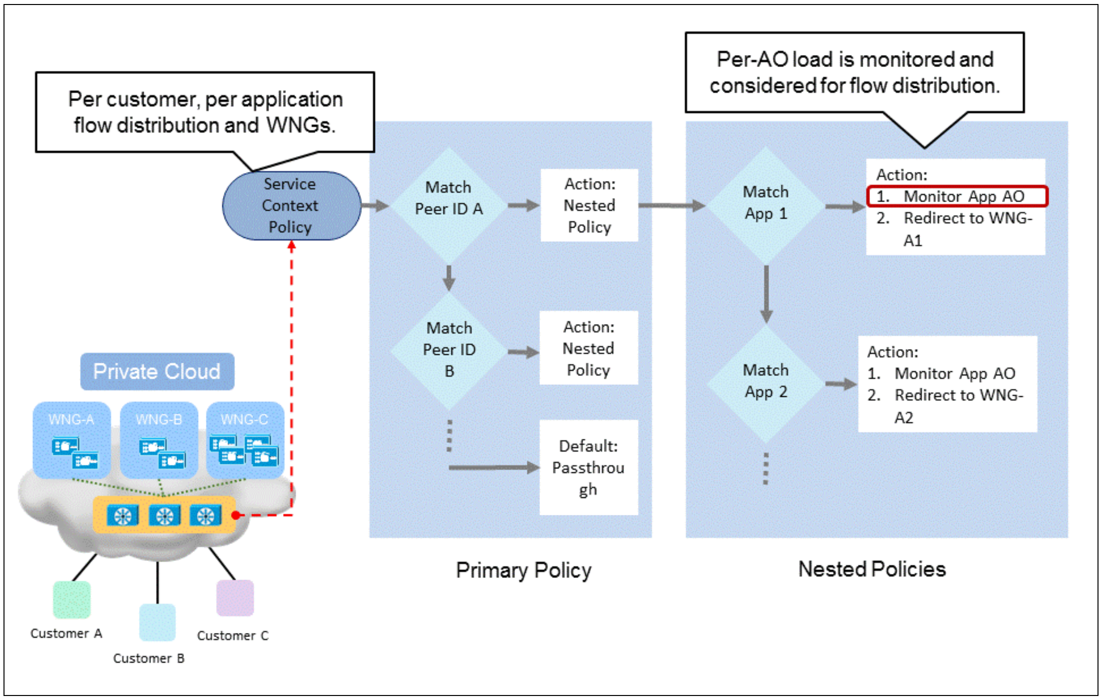


Figure 6-26 Flow distribution example: Private cloud

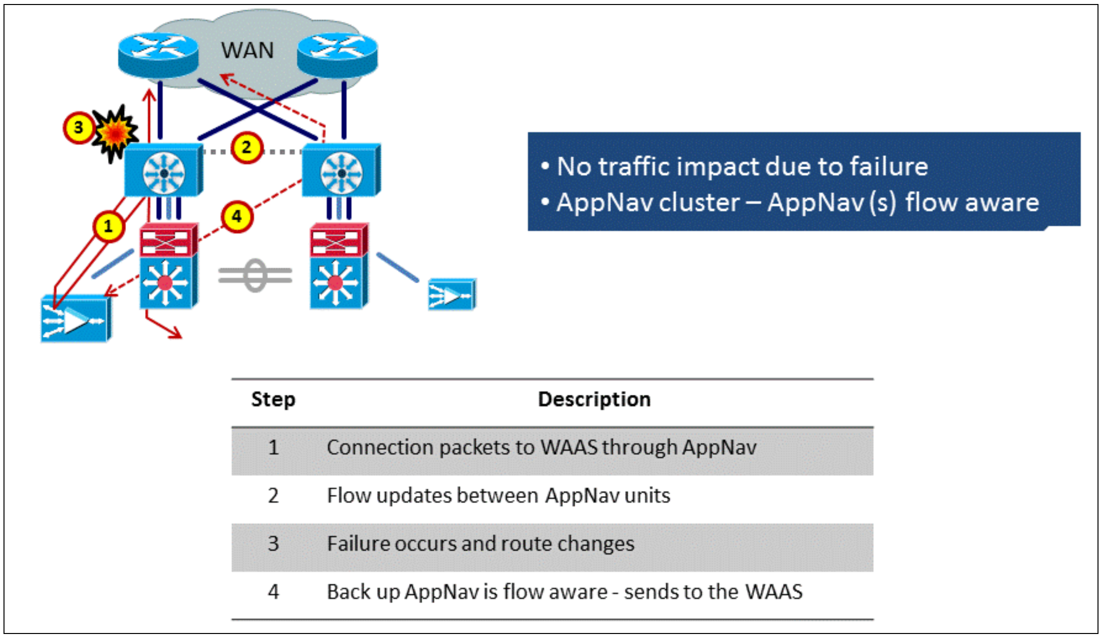
6.5.6 Asymmetric traffic considerations

Automatic asymmetry handling is a built-in feature of AppNav solutions. AppNav Clusters are flow-aware. They maintain the natural traffic path while pinning the asymmetric portions of the same flow to the same optimization node.

Figure 6-27 illustrates an intra-DC scenario for asymmetry handling, considering one physical site and a redundant distribution block with multiple, alternative paths. The table at the bottom describes the path taken by a traffic flow arriving into the data center from the WAN. The AppNav Controller on the left, AppNav1, redirects this flow to the WAAS Node on the left, which then forwards the flow to its intended destination.

At this point, the flow synchronization mechanism, implemented by the cluster membership management process, shares the flow information and the assigned WN with the other ANCs in the cluster.

This makes the ANC on the right, AppNav2, aware of the flow assignment and redirection. This is why the return portion of the flow can be assigned to the same WN, even though asymmetric path conditions cause it to be processed by other ANCs in the cluster.



- No traffic impact due to failure
- AppNav cluster – AppNav (s) flow aware

Step	Description
1	Connection packets to WAAS through AppNav
2	Flow updates between AppNav units
3	Failure occurs and route changes
4	Back up AppNav is flow aware - sends to the WAAS

Figure 6-27 Cisco AppNav high availability

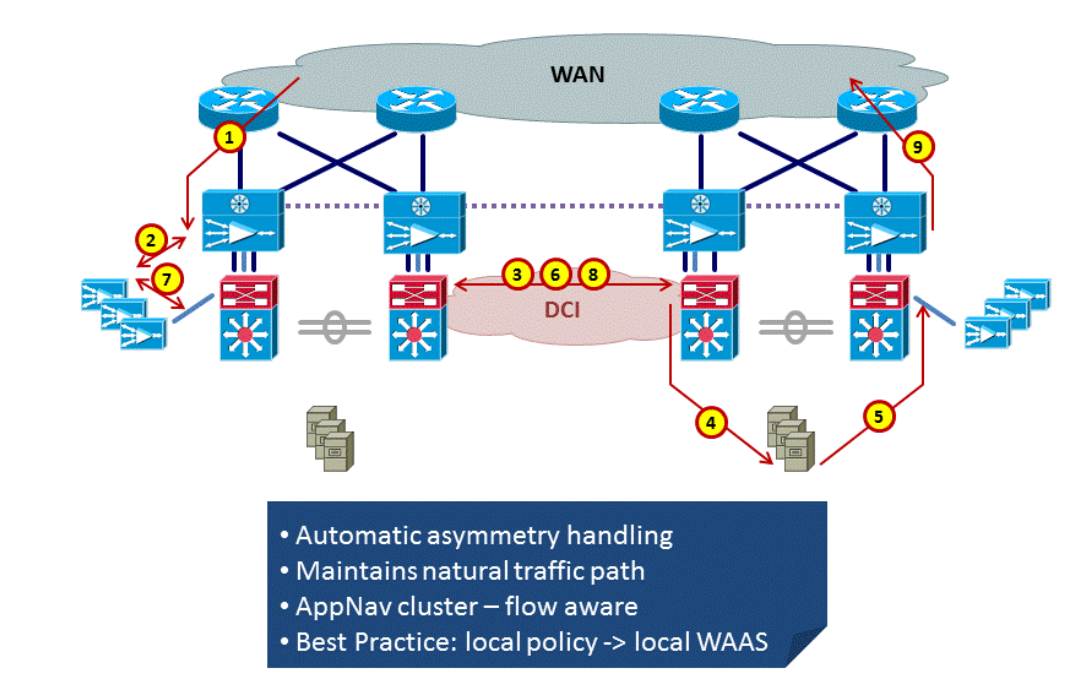
Similarly, asymmetry handling is also automatic and seamless in inter-DC environments, where asymmetric paths are usually a combination of WAN and inter-DC interconnects between two or more data centers. In this scenario, ingress traffic flows may enter from the WAN using one of the data centers, see Figure 6-28.

The local AppNav Controllers intercept and redirect the flow to a local WAAS Node. Remember, this portion of the path can be implemented with a GRE tunnel, in case the WNs are not Layer 2 adjacent to the ANCs. In this example, GRE tunnels are identified with a red segmented arrow. Unoptimized traffic is then forwarded to the intended destination, which in this example is a server on the second data center on the right. The flow traverses the Data Center Interconnect network and arrives at its intended destination.

The return portion of the flow will consistently optimize using the same WN used for the ingress portion. Traffic is intercepted by an AppNav Controller on the destination data center, on the right, and because flow and WN information is synchronized across the ANC group. This ANC knows to redirect the original WN, on the left. This leg of the path is, again, a GRE tunnel. The WN returns the flow through the same GRE tunnel after optimization. At this point, the corresponding ANC forwards the traffic to the original source across the WAN.

All ANCs must belong to the same ANC group in order for inter-DC asymmetry handling to work, even though they reside in different physical locations. This is necessary for flow synchronization, and may require the tuning of dead timers and other settings to allow for heartbeats to be sent across the WAN.

Another consideration relates to the implementation of different policies on the ANCs at each location, to force the selection of local WNs at each location with higher priority.



- Automatic asymmetry handling
- Maintains natural traffic path
- AppNav cluster – flow aware
- Best Practice: local policy -> local WAAS

Figure 6-28 Automatic asymmetry handling - Inter DC

References:

http://www.cisco.com/en/US/prod/collateral/contnetw/ps5680/ps6870/solution_overview_c22-525305.html

http://www.cisco.com/en/US/partner/prod/collateral/contnetw/ps5680/ps11211/datasheet_c78-611644.html

http://www.cisco.com/en/US/prod/collateral/contnetw/ps5680/ps6474/white_paper_c11-705318.html

http://www.cisco.com/en/US/prod/collateral/contnetw/ps5680/ps6474/data_sheet_c78-711502.html

6.6 Virtual WAAS

Cisco virtual WAAS (vWAAS) is a virtual appliance that accelerates business applications delivered from private and virtual private cloud infrastructures, helping ensure an optimal user experience. Cisco vWAAS runs on the VMware ESXi hypervisor providing an agile, elastic, and multi-tenant deployment. vWAAS is the only WAN optimization solution that is deployed in an application-specific, virtualization-aware, and on-demand manner. Using policy-based configuration in the Cisco Nexus 1000V Switch, Cisco vWAAS service is associated with application server virtual machines as they are instantiated or moved. This approach helps enable cloud providers to offer rapid delivery of WAN optimization services with little network configuration or disruption in cloud-based environments.

vWAAS is designed for both enterprises and service providers who plan to offer private and virtual private cloud-based application delivery services over the WAN to their internal or external customers. Cisco vWAAS can be deployed in the physical data center and in private clouds and in virtual private clouds offered by service providers. vWAAS enables migration of business applications to the cloud, reducing the negative effect on performance of cloud-based application delivery to end-users. It enables service providers to offer an excellent application experience over the WAN as a value-added service in their catalogs of cloud services:

- On-demand orchestration: Cisco vWAAS provides the industry's first on-demand orchestrated insertion of WAN optimization in the cloud, using policy-based operations. Using policy-based configuration in the Cisco Nexus 1000V, Cisco vWAAS is associated with application server virtual machines even as they are instantiated or moved.
- High availability: If one Cisco vWAAS virtual machine fails, another Cisco vWAAS virtual machine will become active on the same or a different host using VMware high availability (HA). Cisco vWAAS also supports VMware Distributed Resources Scheduler (DRS) for optimizing and managing resources.
- Fault-tolerant persistent performance: Cisco vWAAS SAN-based storage helps ensure cache preservation and high persistent performance in the event of failure. If a Cisco vWAAS virtual machine fails, a new virtual machine will become active using VMware HA. The new virtual machine will use the same cache storage as the failed virtual machine used, providing the compression benefits of WAN optimization without interruption.

Figure 6-29 shows vWAAS deployment.

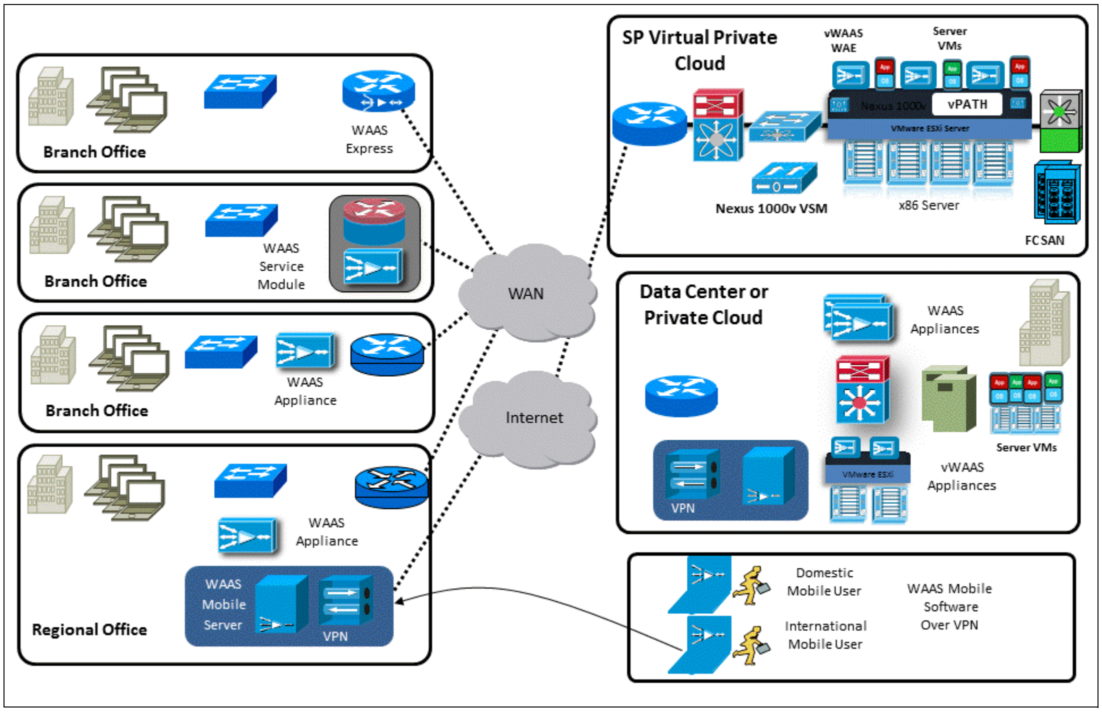


Figure 6-29 vWAAS deployment

Cisco vWAAS supports two deployment options:

1. Traditional WAN edge deployment with WCCP or APPNav: This deployment is applicable in branch offices and data center and private clouds. The WAN edge is the typical place in the network for interception of the WAN-facing traffic to be optimized. Cisco vWAAS provides full support for WCCP Version 2 (WCCPv2) and APPNav. Multiple Cisco vWAAS virtual machines can co-exist together in a single WCCP cluster or APPNav cluster, optimizing all traffic intercepted. These virtual machines can be spread across single or multiple physical servers. Physical and virtual appliances can be mixed in the same WCCP or APPNav cluster. All standard WCCP and APPNav configurations and best practices are applicable to Cisco vWAAS.
2. Deep in the data center with innovative vPATH interception: This deployment is applicable in private clouds and virtual private clouds and helps meet several requirements to help enable WAN optimization as a cloud-ready service, as explained earlier. In this deployment model, Cisco vWAAS virtual machines can be placed next to server virtual machines in the same VMware ESXi host. Using policy-based orchestration of virtual network services provided by the Cisco Nexus 1000V architecture, Cisco vWAAS can be deployed on a per-application or per-server-group basis. vPATH interception in the Cisco Nexus 1000V provides a mechanism for intercepting all traffic to and from these servers and forwarding it to the Cisco vWAAS virtual machine for optimization.

Figure 6-30 shows vWAAS deployment options.

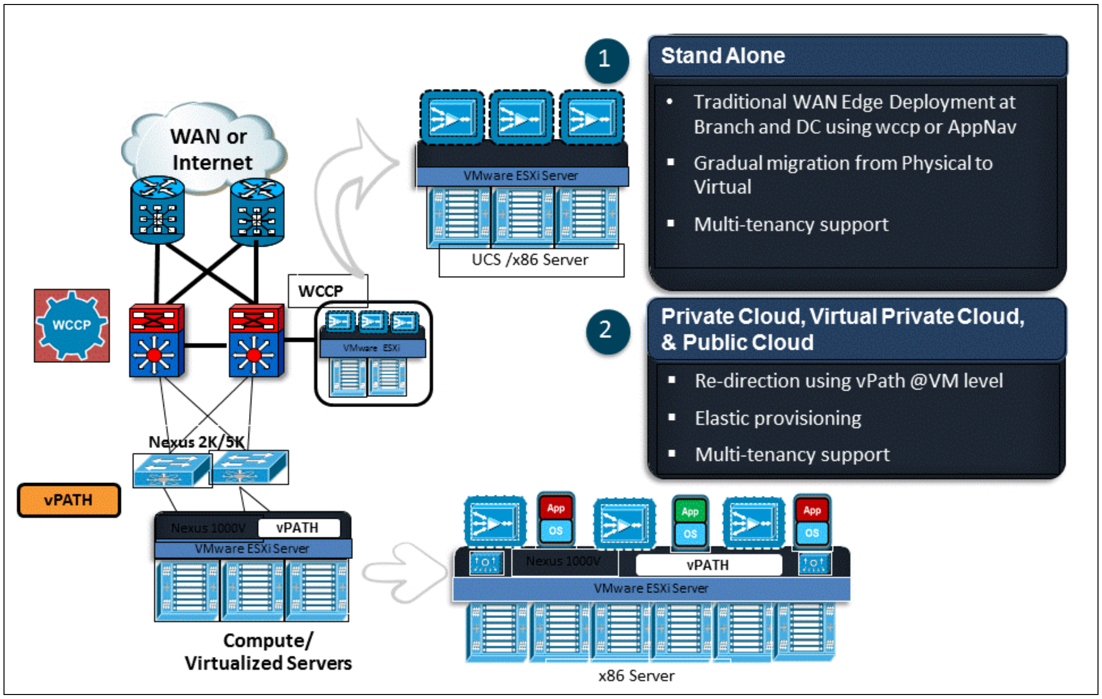


Figure 6-30 Cisco vWAAS deployment options

The focus in this section is on leveraging vPath for inserting vWAAS. The Cisco Nexus 1000V provides a mechanism for attaching Cisco vWAAS to the port profiles of servers that need to be optimized. All traffic to and from these servers will be intercepted by vPATH and forwarded to the Cisco vWAAS virtual machine for optimization. vPATH interception uses Cisco Nexus 1000V port-profile attributes (vn-service) to redirect traffic to Cisco vWAAS. Administrators needs to identify the port profiles of servers to be optimized by Cisco vWAAS. After the port profile is identified, Cisco vWAAS needs to attach to one or multiple port profiles to optimize the traffic. Cisco WAAS autodiscovery helps ensure that a particular TCP connection will be optimized only by the end-point devices (Cisco Wide Area Application Engine [WAE] or Cisco vWAAS).

Cisco vPATH interception, based on the Cisco Nexus 1000V port profiles, provides the following advantages:

- Network attributes move together with virtual machines: This feature provides virtual machine mobility awareness to WAN optimization traffic. Even if server virtual machines move from one VMware ESXi host to another, traffic to and from these servers will continue to be optimized by the same Cisco vWAAS.
- A port profile provides inheritance of network attributes to by any new server virtual machines created with that port profile: Cisco vWAAS starts optimizing traffic immediately for these server virtual machines. This approach provides an on-demand deployment model with minimal network configuration and with elastic scalability.
- Separate port profile can be configured for each tenant in a multi-tenant environment: These port profiles can be attached to separate Cisco vWAAS instances to provide full multi-tenancy.

Figure 6-31 shows service orchestration.

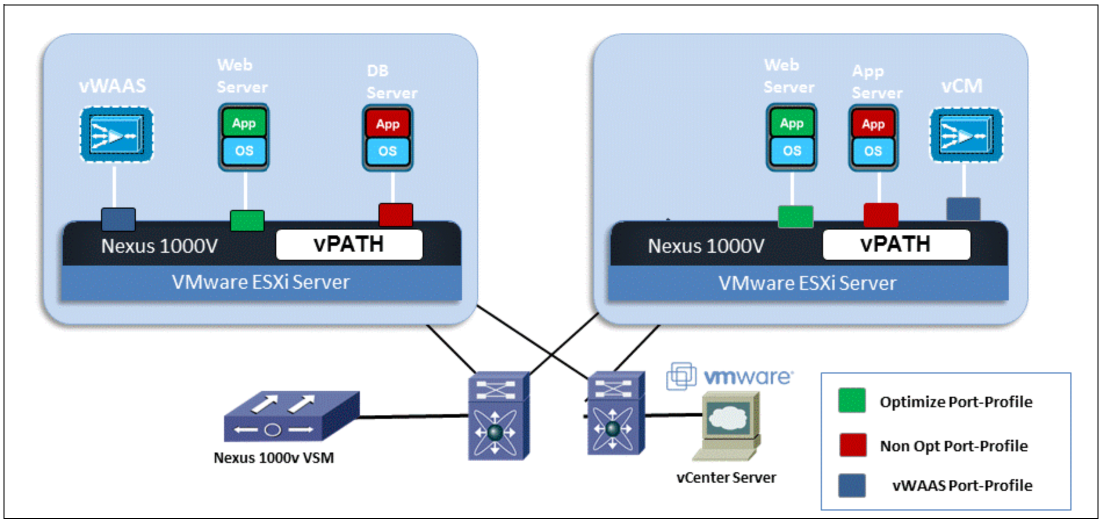


Figure 6-31 On-demand service orchestration in the cloud without network disruption

6.7 ASA series security in the data center

Threats facing today's IT security administrators have grown from the relatively trivial attempts to damage networks, into sophisticated attacks aimed at profit and the theft of sensitive corporate data. Implementation of robust data center security capabilities to protect sensitive mission-critical applications and data is the foundation for the effort to secure enterprise networks.

In addition, attack vectors have moved higher in the TCP/IP stack to subvert network protection and aim directly at applications. HTTP-, XML-, and SQL-based attacks are useful efforts for most attackers because these protocols are usually allowed to flow through the enterprise network and enter the intranet data center.

The data center security challenges does not stop there. New application rollouts, virtualization, and an increasingly transparent perimeter are converging to drive an evolution in the requirements for data center security architectures.

Application rollouts bring there its own set of challenges for securing communications and applying security policy; couple this with a virtualized environment and the challenge of policy enforcement and visibility increases many times over.

Traditionally, the perimeter has been the proverbial shield to stop malicious and unwanted outside traffic from leaking into the enterprise network. Creating a secure perimeter is still valid and essential in defending against attacks and providing traffic filtering. But the amount and type of traffic entering the enterprise network has increased and continues to do so. Extranet connections for business partners, vendor connections, supply chain transactions, and digital communications all required more openings to be created at the perimeter to allow communication. Permitting these business-driven openings creates greater opportunities for attack and elevates the risk to a network.

Virtualization is driving change in the way data centers are being developed. Server virtualization is becoming a prevalent tool for consolidation, power savings, and cost reduction. It is also creating new challenges for infrastructure and security teams to be able to provide consistent levels of isolation, monitoring, and policy enforcement-similar to what is available for physical servers and systems today.

The Cisco ASA firewalls enables businesses to segment and secure multi-tenant data center environments by integrating firewall security directly into the network infrastructure. This chapter explains the different hardware options and the operational modes that they offer to secure modern data centers.

6.7.1 Cisco Catalyst 6500 Series ASA Services Module

In a Cisco Nexus 7000 Series deployment, the Cisco ASA 5585-X is a perfect fit. In case the Cisco Catalyst 6500 Series switch is being used, the ASA Services Module is more appropriate. The software code remains the same. If intrusion prevention services (IPS) functionality is required, the standalone ASA 5585-X appliance is the better choice.

The ASA Firewall Module is designed for data centers that require an integrated security solution that combines full-featured switching with security functionalities. The Cisco ASA Services Module integrates seamlessly with Cisco Catalyst 6500 Series switches.

Figure 6-32 shows the Catalyst 6500 ASA Services Module.

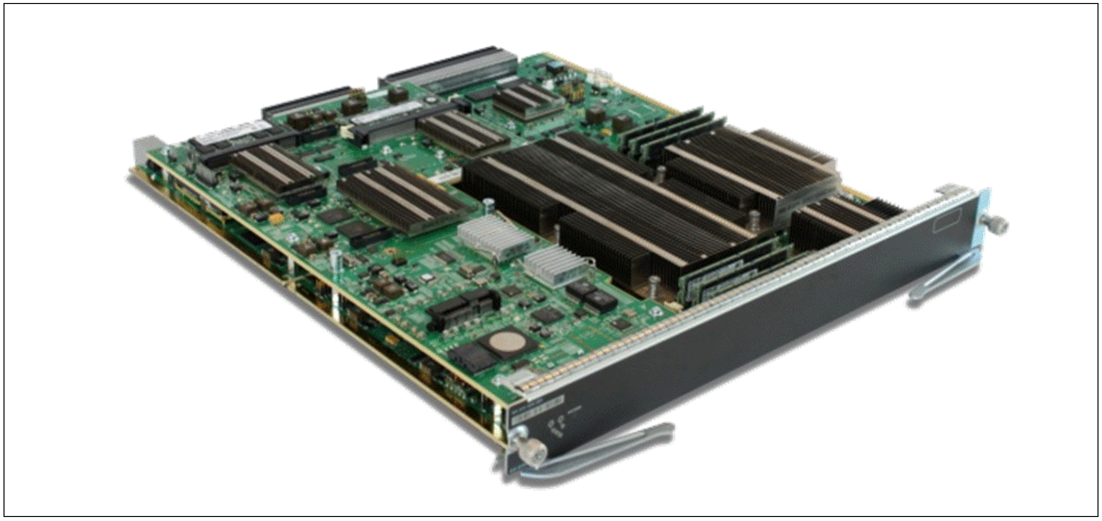


Figure 6-32 ASA Firewall Module

The ASA Services Module will add full firewall capabilities to an existing infrastructure by sliding a blade into an empty slot in an existing Catalyst 6500 Series switch; no additional rack space, cabling, power, or physical interface is required. It also works in tandem with other modules in the chassis to deliver security feature throughout the entire chassis, effectively making every port a security port.

The module delivers 20 Gbps of maximum firewall throughput and 16 Gbps of multi-protocol firewall throughput. Up to four blades can be installed in a single chassis, providing scalability to 64 Gbps per chassis.

The performance of the ASA Firewall Module is comparable to the stand-alone ASA 5585-x as shown in Table 6-5.

Table 6-5 Firewall performance of the ASA Services Module compared with ASA 5585-X

	ASA Services Module	**ASA 5585-X w/SSP-60**
Maximum firewall throughput	20 Gbps	40 Gbps
Multiprotocol firewall throughput	16Gbps	20 Gbps
Concurrent connections	10,000,000	10,000,000
Connections per second	300,000	350,000
Security contexts	250	250
VLANs	1000	1000

Features and benefits

The ASA Services Module helps data centers increase effectiveness and efficiency in protecting their networks and applications. The module delivers exceptional protection of a Cisco Catalyst 6500 Series investment and helps to reduce the total cost of network ownership - all while lowering operating costs and addressing intangible opportunity costs.

This is accomplished through the following elements:

- Seamless integration. The ASA Services Module seamlessly integrates with Cisco Catalyst 6500 Series switches. Full firewall capabilities are added by simply sliding the ASA Services Module into an empty slot in the existing Catalyst 6500 Series switch. No rack space is required, since it populates an empty slot within the existing switch; all interfaces are virtual, eliminating the need to manage physical interfaces; and the module uses the existing switch connections, so no re-cabling is required. As a result, the time required for installation and configuration is reduced.
- Simplified maintenance and management. The ASA Services Module integrates with the Catalyst 6500 Series chassis, using the same connections and management software as the rest of the switch. In effect, the module becomes part of the switch, with almost no increase in the time, effort, and cost of managing and maintaining the switch.
- Minimal environmental costs. As a fully integrated component of the Cisco Catalyst 6500 Series switch, the ASA Services Module utilizes the power and cooling from the switch.

An overview of the ASA features can be found here:

http://www.cisco.com/en/US/prod/collateral/vpndevc/ps6032/ps6094/ps6120/prod_brochure0900aecd80285492.pdf

For more security features, visit this website:

http://www.cisco.com/en/US/partner/docs/security/asa/roadmap/asa_new_features.html

6.7.2 Cisco ASA 5585-X Adaptive Security Appliance

The Cisco ASA 5585-X is a high-performance, 2-slot chassis, with the firewall/VPN Security Services Processor (SSP) occupying the bottom slot, and the IPS Security Services Processor (IPS SSP) in the top slot of the chassis. The firewall/VPN SSP is required to run the IPS SSP on the Cisco ASA 5585-X.A Security Plus license is needed to enable 10 Gigabit Ethernet Enhanced Small Form-Factor Pluggable (SFP+) ports on the SSP-10, SSP-20, IPS SSP-10, and IPS SSP-20.

For higher interface density requirements, Cisco ASA 5585-X I/O Modules provide high interface density to mission-critical data centers that require exceptional flexibility and security. With two 8-port 10 Gigabit Ethernet modules and one SSP-40 or SSP-60 module, Cisco ASA 5585-X can support up to twenty 10 Gigabit Ethernet ports in a 2RU chassis. With two 20-port 1 Gigabit Ethernet modules and one SSP module, Cisco ASA 5585-X can support up to fifty 1Gigabit Ethernet ports in a 2RU chassis.

Figure 6-33 shows the Cisco ASA 5585-X chassis with two I/O modules.

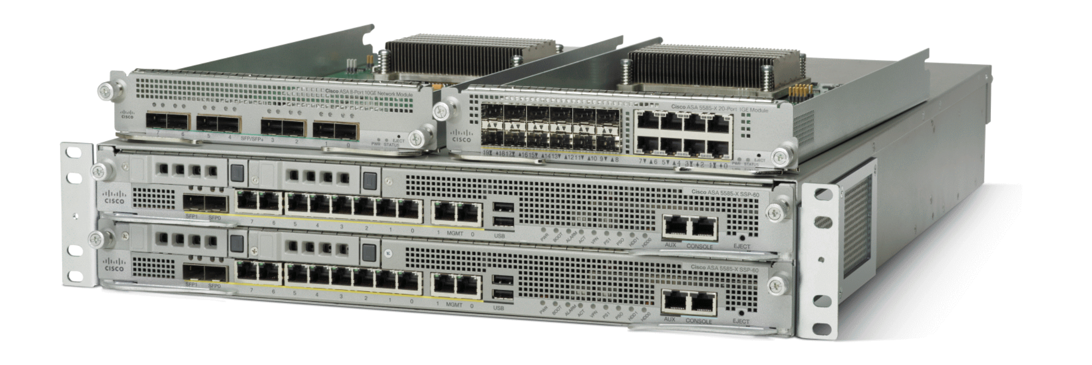


Figure 6-33 Cisco ASA 5585-X Adaptive Security Appliance

6.7.3 Firewall modes of operation

The Cisco ASA security appliances can be operated in a high availability mode. It offers features such as sub-second failover and interface redundancy, including full-mesh Active/Standby and Active/Active failover configurations.

With Active/Active failover, both units can pass network traffic. This also can be used to configure traffic sharing on the network. Active/Active failover is available only on units running in “multiple” context mode. With Active/Standby failover, a single unit passes traffic while the other unit waits in a standby state. Active/Standby failover is available on units running in either “single” or “multiple” context mode. Both failover configurations support stateful or stateless failover.

The unit can fail if one of these events occurs:

- The unit has a hardware failure or a power failure.
- The unit has a software failure.
- Too many monitored interfaces fail.
- The administrator has triggered a manual failure by using the CLI command: `no failure active`

Even with stateful failover enabled, device-to-device failover may cause some service interruptions. Here are some examples:

- Incomplete TCP 3-way handshakes must be re-initiated.
- In Cisco ASA Software Release 8.3 and earlier, Open Shortest Path First (OSPF) routes are not replicated from the active to standby unit. Upon failover, OSPF adjacencies have to be reestablished and routes re-learnt.
- Most inspection engines' states are not synchronized to the failover peer unit. Failover to the peer device loses the inspection engines' states.

Active/Standby failover

The Active/Standby failover lets one use a standby security appliance to take over the functions of a failed unit. When the active unit fails, it changes to the standby state while the standby unit changes to the active state. The unit that becomes active assumes the IP addresses (or, for transparent firewall, the management IP address) and MAC addresses of the failed unit and begins passing traffic. The unit that is now in standby state takes over the standby IP addresses and MAC addresses. Because network devices see no change in the MAC to IP address pairing, no Address Resolution Protocol (ARP) entries change or time out anywhere on the network.

In an Active/Standby failover, a failover occurs on a physical unit basis and not on a context basis in multiple context mode. An active/Standby failover is the most commonly deployed method of high availability on the ASA platform.

Active/Active failover

The Active/Active failover is available to security appliances in “multiple” context mode. Both security appliances can pass network traffic at the same time, and can be deployed in a way that they can handle asymmetric data flows. The security contexts can be divided on the security appliance into failover groups. A failover group is simply a logical group of one or more security contexts. A maximum of two failover groups on the security appliance can be created.

The failover group forms the base unit for failover in Active/Active failover. Interface failure monitoring, failover, and active/standby status are all attributes of a failover group rather than the physical unit. When an active failover group fails, it changes to the standby state while the standby failover group becomes active. The interfaces in the failover group that becomes active assume the MAC and IP addresses of the interfaces in the failover group that failed. The interfaces in the failover group that is now in the standby state take over the standby MAC and IP addresses. This is similar to the behavior that is seen in physical Active/Standby failover.

Routed mode

The routed mode is the traditional mode of the firewall operation. Two or more interfaces separate L3 domains.

In routed mode, the firewall is considered to be a router hop in the network. It can use OSPF or RIP (in single context mode). Routed mode supports many interfaces. Each interface is on a different subnet. Interfaces can be shared between contexts.

The firewall acts as a router between connected networks, and each interface requires an IP address on a different subnet. In single context mode, the routed firewall supports OSPF, EIGRP, and RIP. Multiple context mode supports static routes only. It is recommended to use the routing capabilities of the upstream and downstream routers instead of relying on the ASA firewall for extensive routing needs. These are the routed mode characteristics:

- Traditional mode of the firewall (Layer 3 hop)
- Separates two L3 domains, as depicted in Figure 6-34
- Often represents a NAT boundary
- Policy is applied to flows as they transit the firewall

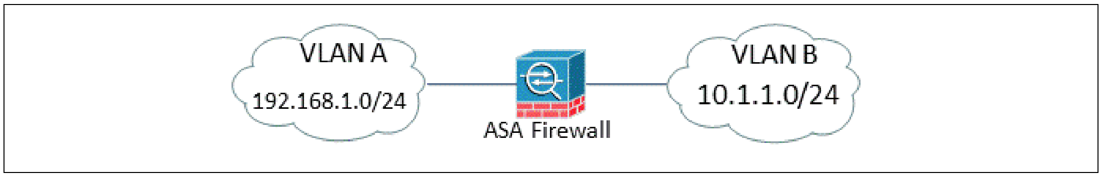


Figure 6-34 ASA routed mode

Transparent mode

The transparent mode is where the firewall acts as a bridge functioning mostly at Layer 2.

A transparent firewall is a Layer 2 firewall that acts like a “bump in the wire,” or a “stealth firewall,” and is not seen as a router hop to connected devices. The ASA connects the same network between its interfaces. Because the firewall is not a routed hop, it is easy to introduce a transparent firewall into an existing network. These are the transparent mode characteristics:

- Operates at Layer 2, transparent to the network
- Drops into existing networks without re-addressing or redesign as IP address schemes must not be modified
- Simplifies internal firewalling and network segmentation

Figure 6-35 shows how the ASA firewall is connected in the 10.1.1.0/24 network.

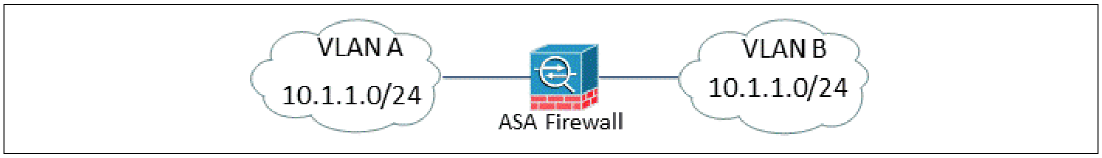


Figure 6-35 ASA transparent mode

There are some advantages when using the transparent mode:

- Routing protocols can establish adjacencies through the firewall.
- Protocols such as HSRP, VRRP, GLBP and multicast streams can cross the firewall.
- Non-IP traffic can be allowed (IPX, MPLS, BPDUs).
- There is no dynamic routing protocol support (on the FW itself) or VPN support.
- There is no QoS or DHCP Relay support.

Multi-context mode

Multi-context mode involves the use of virtual firewalls, which can be either routed or transparent mode. A context is a virtual firewall as shown in Figure 6-36. The firewall is divided in multiple logical firewalls. The system context assigns the physical ports to the logically separated, virtual firewalls. The admin context provides remote root access and connects to all virtual firewalls and must be defined in addition to the virtual contexts. There is a limit of 250 virtual firewall on the ASA platform.

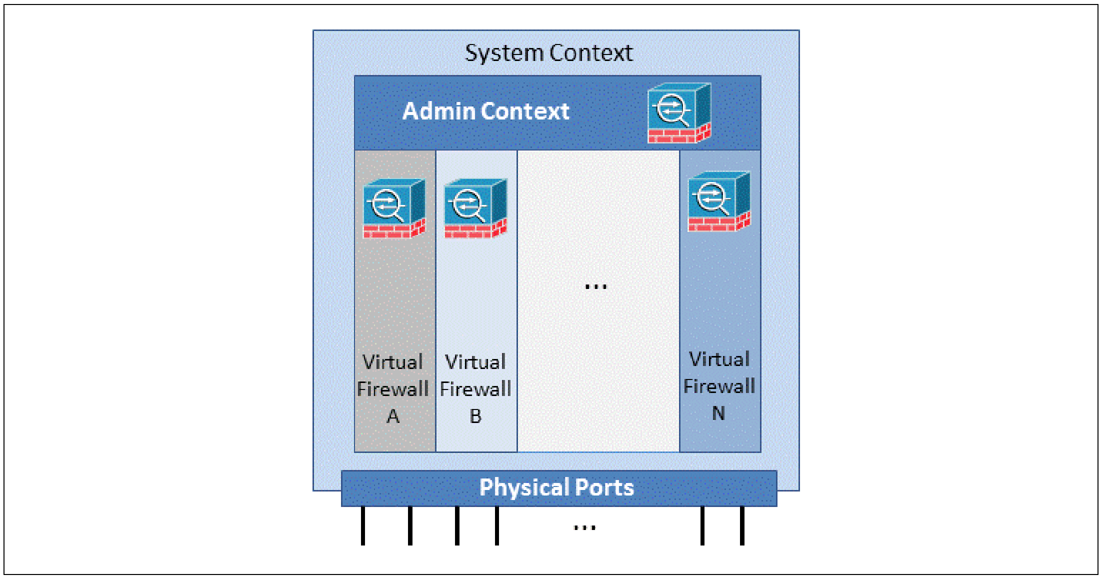


Figure 6-36 ASA Virtual Firewall Concept

There are several unsupported features in ASA Multi-Context mode:

- Dynamic routing protocols (EIGRP, OSPF, RIP)
- Multicast routing (multicast bridging is supported)
- MAC addresses for virtual interfaces are automatically set to physical interface MAC
- VPN services
- Route tracking is not available (routed mode only)

The admin context is just like any other context, except that when a user logs in to the admin context, then that user has system administrator rights and can access the system and all other contexts. The admin context is not restricted in any way, and can be used as a regular context. However, because logging into the admin context grants administrator privileges over all contexts, one might need to restrict access to the admin context to appropriate users.

Mixed mode

Mixed mode is the concept of using virtualization to combine routed and transparent mode virtual firewalls and was introduced with software version 8.4. Earlier versions only allowed either a transparent or routed mode on the firewall. There are some restrictions with the use of mixed mode:

- It is only supported on the ASA-SM today with code level 8.5 and above
- Up to 8 pairs of interfaces are supported per context

There are some caveats and dependencies; see the Release Notes:

http://www.cisco.com/en/US/partner/docs/security/asa/asa84/release/notes/asarn84.html

More information about the configuration of various operational modes of the ASA firewalls can be found at this website:

http://www.cisco.com/en/US/partner/products/ps6120/tsd_products_support_configure.html

6.7.4 Security layers in data centers

Device virtualization provides new design opportunities and options for creating flexible data center architectures. Features that provide control plane and data plane isolation are offering a multitude of design options for device placement, Layer 2 and Layer 3 designs, and service integration. This document focuses on three areas of data center security: isolation, policy enforcement, and visibility. These can be defined as follows:

- Isolation: Isolation can provide the first layer of security for the data center and server farm. Depending on the goals of the design it can be achieved through the use of firewalls, access lists, Virtual Local Area Networks (VLAN), virtualization, and/or physical separation. A combination of these can provide the appropriate level of security enforcement to the server farm applications and services.
- Policy Enforcement: There is no shortage on the variety of traffic flows, protocols, and ports required to operate within the data center. Traffic flows can be sourced from a variety of locations, including client to server requests, server responses to these requests, server originated traffic, and server-to-server traffic. Because of the amount of traffic flows and the variety of sources, policy enforcement in the data center requires a considerable amount of up-front planning; couple this with a virtualized environment and the challenge of both policy enforcement and visibility becomes greater.
- Visibility: Data centers are becoming very fluid in the way they scale to accommodate new virtual machines and services. Server virtualization and technologies, such as VMotion, allow new servers to be deployed and to move from one physical location to another with little requirement for manual intervention. When these machines move and traffic patterns change, a security administration challenge can be created in terms of maintaining visibility and ensuring security policy enforcement.

The security services and examples in this document have been integrated into the example architecture defined here with these three considerations in mind. Because security models can differ depending on the business goals of the organization, compliance requirements, the server farm design, and the use of specific features (such as device virtualization), there is no magic blueprint that covers all scenarios. However, the basic principles introduced in this document for adding security to the data center architecture can apply to other environments.

Figure 6-37 displays the various data center layers with their security elements.

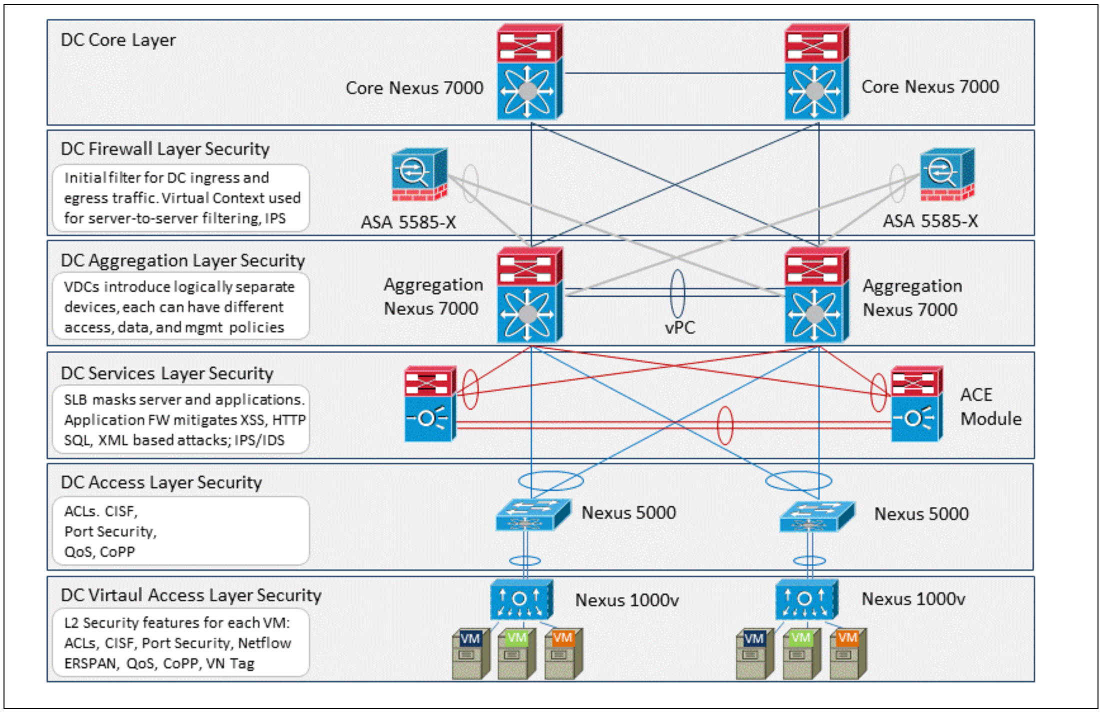

DC Core Layer
Core Nexus 7000
Core Nexus 7000
DC Firewall Layer Security
Initial filter for DC ingress and egress traffic. Virtual Context used for server-to-server filtering, IPS
ASA 5585-X
ASA 5585-X
DC Aggregation Layer Security
VDCs introduce logically separate devices, each can have different access, data, and mgmt policies
Aggregation Nexus 7000
vPC
Aggregation Nexus 7000
DC Services Layer Security
SLB masks server and applications. Application FW mitigates XSS, HTTP SQL, XML based attacks; IPS/IDS
ACE Module
DC Access Layer Security
ACLs. CISF, Port Security, QoS, CoPP
Nexus 5000
Nexus 5000
DC Virtaul Access Layer Security
L2 Security features for each VM: ACLs, CISF, Port Security, Netflow ERSPAN, QoS, CoPP, VN Tag
Nexus 1000v
Nexus 1000v
VM
VM
VM
VM
VM
VM

Figure 6-37 Security Services in a virtualized data center

Using a multi-layered security model to provide protection for the enterprise data center from internal or external threats is a best practice for creating a multi-layered security model.

The architectures discussed in this document are based on the Cisco data center design best practice principles. This multi-layered data center architecture is comprised of the following key components: core, firewall, aggregation, services, and access. This architecture allows for data center modules to be added as the demand and load increases. The data center core provides a Layer 3 routing module for all traffic in and out of the data center.

The aggregation layer serves as the Layer 3 and Layer 2 boundary for the data center infrastructure. In these design, the aggregation layer also serves as the connection point for the primary data center firewalls. Services such as server load balancers, intrusion prevention systems, application-based firewalls, network analysis modules, and additional firewall services are deployed at the services layer. The data center access layer serves as a connection point for the server farm. The virtual-access layer refers to the virtual network that resides in the physical servers when configured for virtualization.

Data center core layer

The data center core module provides Layer 3 connectivity between the data center and the campus network. The core is a centralized Layer 3 routing module in which one or more data center aggregation layers connect. This usually serves as the initial entry point from the campus network into the data center infrastructure.

Firewall layer

The aggregation layer provides an excellent filtering point and first layer of protection for the data center. This layer provides a building block for deploying firewall services for ingress and egress filtering. The Layer 2 and Layer 3 recommendations for the aggregation layer also provide symmetric traffic patterns to support stateful packet filtering.

Depending on the requirements, one can select the Cisco ASA5585-X with its modular capabilities including IPS or in case an IPS is not required, the ASA security module for the Catalyst 6500 family may be chosen. Both models meet the high performance data center requirements.

The Cisco ASA firewalls can be configured in various modes, see 6.7.3, “Firewall modes of operation” on page 328. The firewalls can be configured in an active-active design. This design allows load sharing across the infrastructure based on the active Layer 2 and Layer 3 traffic paths. Each firewall can be configured for two virtual contexts or more. Virtual context 1 is active on the ASA 1 and virtual context 2 is active on ASA 2. The backup firewalls are configured the other way: backup VC1 on ASA2 and backup VC2 on ASA1.

> **Note:** The modular aspect of this design allows additional firewalls to be deployed at the aggregation layer as the server farm grows and the performance requirements increase.

Deployment recommendations

The firewall policy will differ based on the organizational security policy and the types of applications deployed. In most cases a minimum of the following protocols will be allowed:

- Domain Name System (DNS),
- Hypertext Transfer Protocol (HTTP),
- Hypertext Transfer Protocol over Secure Sockets Layer (HTTPS),
- Simple Mail Transfer Protocol (SMTP),
- File Transfer Protocol (FTP),
- Routing protocols,
- Unified communications: voice over IP (VoIP) protocols, video protocols,
- Multicast,
- Terminal services,
- Internet Control Message Protocol (ICMP) to some extent

Regardless of the number of ports and protocols being allowed either to and from the data center or from server-to-server, there are some baseline recommendations that will serve as a starting point for most deployments. The firewalls should be hardened in a similar fashion to the infrastructure devices. The following configuration notes apply:

- Use HTTPS for device access. Disable HTTP access.
- Configure Authentication, Authorization, and Accounting (AAA) for role-based access control and logging. Use a local fallback account in case the AAA server is unreachable.
- Use out-of-band management and limit the types of traffic allowed over the management interface(s).
- Use Secure Shell (SSH). Disable Telnet.
- Use Network Time Protocol (NTP) servers.

Aggregation layer

The aggregation layer provides an excellent filtering point and first layer of protection for the data center. This layer provides a building block for deploying firewall services for ingress and egress filtering. The Layer 2 and Layer 3 recommendations for the aggregation layer also provide symmetric traffic patterns to support stateful packet filtering.

The aggregation layer provides an excellent filtering point and first layer of protection for the data center. This layer provides a building block for deploying firewall services for ingress and egress filtering. The Layer 2 and Layer 3 recommendations for the aggregation layer also provide symmetric traffic patterns to support stateful packet filtering.

The aggregation switches used in this design are a pair of Cisco Nexus 7000 Series switches. They serve as a high performance 10-Gigabit aggregation point for data center traffic and services.

The Cisco Nexus 7000 introduces the concept of Virtual Device Context (VDC). The VDC feature allows for the virtualization of the control plane, data plane, and management plane of the Cisco Nexus 7000 as described in Chapter 9, “NX-OS network operating system” on page 481.

From a security standpoint this virtualization capability can provide an enhanced security model. Because the VDCs are logically separate devices, each can have different access, data, and management policies defined.

Services layer

Data center security services can be deployed in a variety of combinations. The type and the combination of security deployed depend largely on the business model of the organization. The services deployed in this layer are used in a combination to provide isolation, application protection, and visibility into data center traffic flows. From a larger viewpoint, it is also important to consider the scalability of security services in the data center. The goal of these designs is to provide a modular approach to deploying security by allowing additional capacity to easily be added for each service. Additional web application firewalls, IDS, IPS and monitoring services can all be scaled without requiring a re-architecture of the overall data center design.

In Figure 6-37 on page 332 the Cisco Application Control Engine (ACE) service module for the Cisco Catalyst 6500 is shown. The Cisco ACE is designed as an application and server scaling tool, but it has security benefits as well. The Cisco ACE can mask the servers real IP address and provide a single IP for clients to connect over a single or multiple protocols such as HTTP, HTTPS, FTP, an so on.

The Cisco ACE devices can also be used to scale the web application firewall appliances. The web application firewalls are configured as a server farm and the Cisco ACE is distributing connections to the web application firewall pool.

To offload servers from the Secure Socket Layer (SSL) processing, the Cisco ACE can store server certificates locally. This allows the Cisco ACE to proxy SSL connections for client requests and forward the client request in clear text to the server. The following configuration fragment shows the SSL proxy service configuration on the Cisco ACE module.

Access layer

The data center access layer provides Layer 2 connectivity for the server farm. In most cases the primary role of the access layer is to provide port density for scaling the server farm.

Security at the access layer is primarily focused on securing Layer 2 flows. Using VLANs to segment server traffic and associating access control lists (ACLs) to prevent any undesired communication are best practice recommendations. Additional security mechanisms that can be deployed at the access layer include private VLANs (PVLANs), the Catalyst Integrated Security Features (CISF), which include these:

- Dynamic Address Resolution Protocol (ARP) inspection
- Dynamic Host Configuration Protocol (DHCP) snooping
- IP Source Guard.

Port security can also be used to lock down a critical server to a specific port.

Control Plane Policing (CoPP) provides QoS-based prioritization and protection of control plane traffic that arrives at the switch in the data plane, which ensures network stability, reachability, and packet delivery.

Virtual access layer

Virtualization is changing the way data centers are designed. Server virtualization is creating new challenges for security deployments. Visibility into virtual machine activity and isolation of server traffic becomes more difficult when virtual machine-sourced traffic can reach other virtual machines within the same server without being sent outside the physical server.

In the traditional access model, each physical server is connected to an access port. Any communication to and from a particular server or between servers goes through a physical access switch and any associated services such as a firewall or a load balancer.

In case applications reside on virtual machines and multiple virtual machines reside within the same physical server, it might not be necessary for traffic to leave the physical server and pass through a physical access switch for one virtual machine to communicate with another. Enforcing network policies in this type of environment can be a significant challenge. The goal remains to provide many of the same security services and features used in the traditional access layer in this new virtual access layer.

The virtual access layer resides in and across the physical servers running virtualization software. Virtual networking occurs within these servers to map virtual machine connectivity to that of the physical server. A virtual switch is configured within the server to provide virtual machine ports connectivity. The way in which each virtual machine connects, and to which physical server port it is mapped, is configured on this virtual switching component. While this new access layer resides within the server, it is really the same concept as the traditional physical access layer. It is just participating in a virtualized environment.

More details can be found here:

http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DC_3_0/dc_sec_design.pdf

6.7.5 Traffic flows

Traditionally, a firewall is a routed hop and acts as a default gateway for hosts that connect to one of its screened subnets. A transparent firewall, on the other hand, is a Layer 2 firewall that acts like a “bump in the wire,” or a “stealth firewall,” and is not seen as a router hop to connected devices. This chapter briefly describes the two modes of operation.

Transparent mode traffic flow

These steps describe how data moves through the ASA as shown in Figure 6-38.
A user on the outside network requests data from the inside virtual server:

1. The ASA receives the packet and adds the source MAC address to the MAC address table, if required.
2. In the multiple context mode, the ASA first classifies the packet in accordance with a unique interface. Because it is a new session, it verifies that the packet is allowed in accordance with the terms of the security policy (access lists, filters, or AAA). The ASA records that a session is established only if the outside user has the valid access to the internal web server. The access list must be configured to allow the outside user to get the access for the web server.
3. If the destination MAC address is in its table, the ASA forwards the packet out of the inside interface. If the destination MAC address is not in the ASA table, the ASA attempts to discover the MAC address when it sends an ARP request and a ping. The first packet is dropped.

The web server responds to the request. Because the session is already established, the packet bypasses the many lookups associated with a new connection. The ASA forwards the packet to the outside user. There is no NAT required as the ASA acts as L2 bridge. The Switched Virtual Interface (SVI) in the Nexus 7000 will route the data to the adjacent L3 routing instance on the same device.

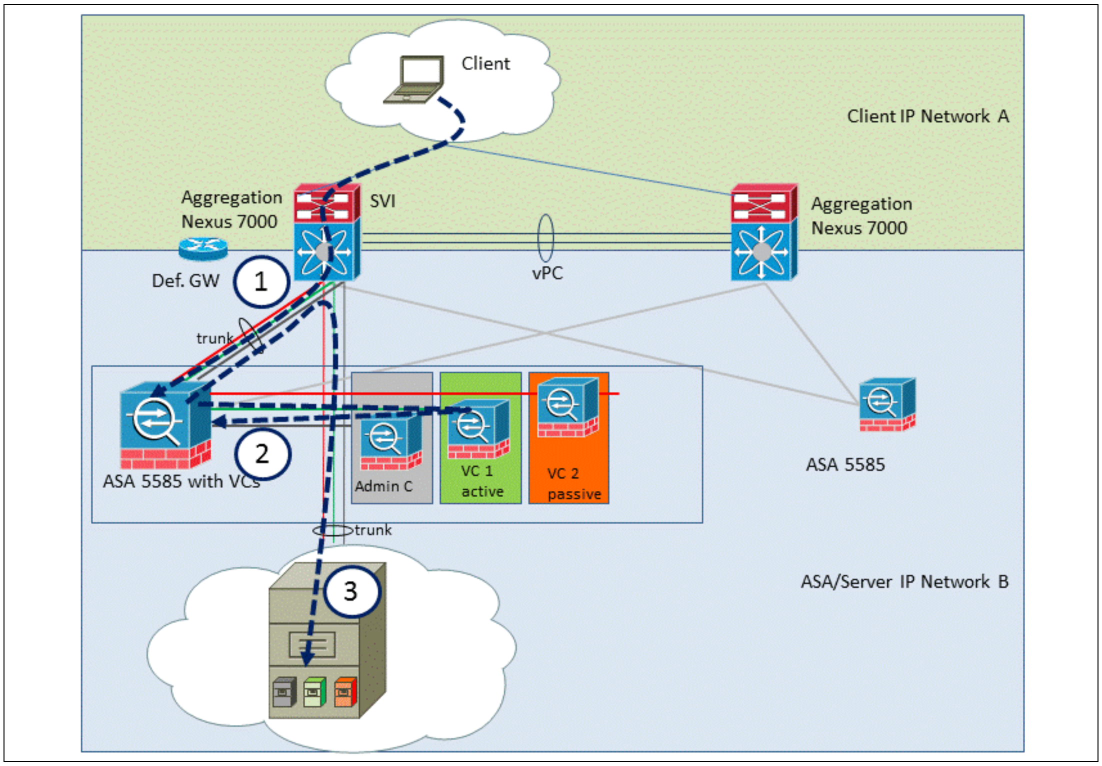


Figure 6-38 Transparent mode ASA flow through

Routed mode traffic flow

The ASA acts as a router between connected networks, and each interface requires an IP address on a different subnet. In single context mode, the routed firewall supports OSPF and RIP (in passive mode). Multiple context mode supports static routes only. We recommend using the advanced routing capabilities of the upstream and downstream routers instead of relying on the ASA for extensive routing needs.

As shown in Figure 6-39, the user on the outside network requests data from a virtual server:

1. The ASA receives the packet and because it is a new session, the ASA verifies that the packet is allowed according to the terms of the security policy (access lists, filters, AAA).
2. For multiple context mode, the ASA first classifies the packet according to either a unique interface or a unique destination address associated with a context; the destination address is associated by matching an address translation in a context. In this case, the classifier "knows" that the server address belongs to a certain context because of the server address translation. The ASA translates the destination address to the real address IP address.
3. The ASA then adds a session entry to the fast path and forwards the packet from the server side interface.

When the server responds to the request, the packet goes through the ASA and because the session is already established, the packet bypasses the many lookups associated with a new connection. The ASA performs NAT by translating the IP address and forwards the packet to the outside user.

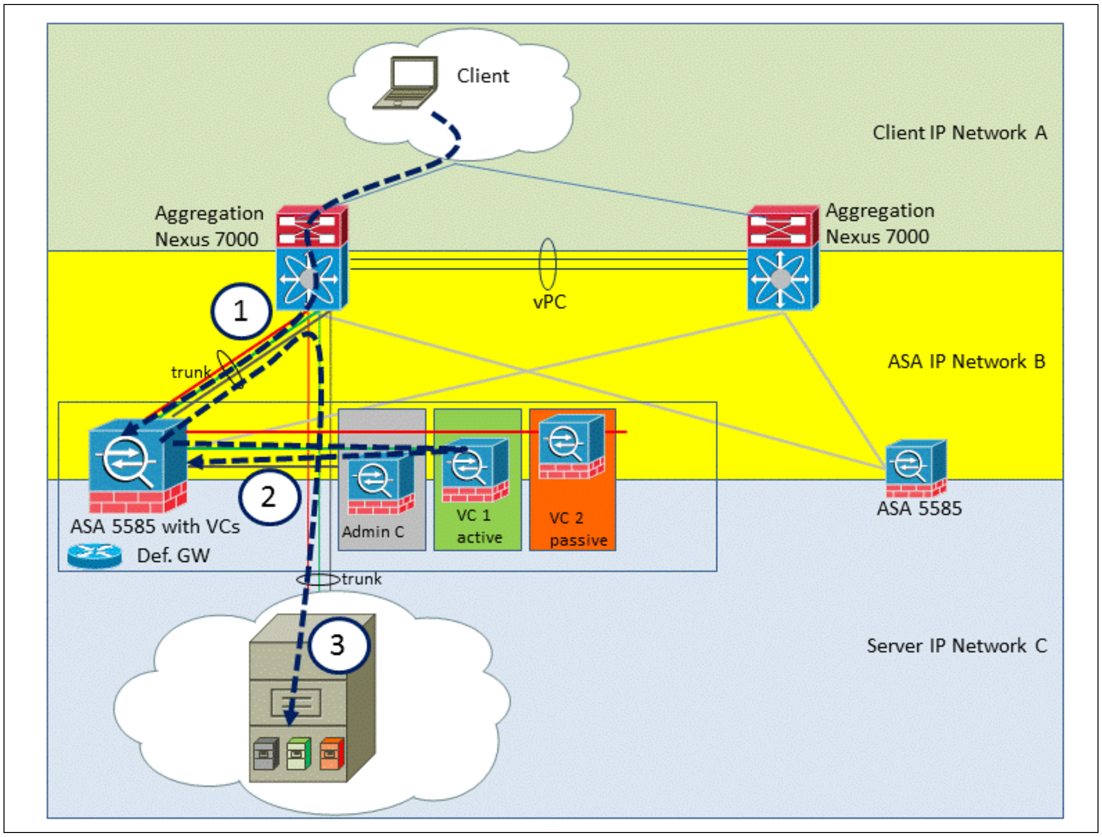


Figure 6-39 Routed mode ASA flow through

Server to server in routed mode traffic flow

Server A requests a data from server B as shown in Figure 6-40

1. The ASA receives the packet and because it is a new session, the ASA verifies that the packet is allowed according to the terms of the security policy (access lists, filters, AAA).
2. For multiple context mode, the ASA first classifies the packet according to either a unique interface or a unique destination address associated with a context; the destination address is associated by matching an address translation in a context. In this case, the interface is unique; the server IP address does not have a current address translation.

3. The ASA then records that a session is established and forwards the packet out of the interface of the cross-connected Nexus 7000 switch.
4. When server B responds to the request, the packet goes through the fast path, which lets the packet bypass the many lookups associated with a new connection. The ASA forwards the packet to server A.

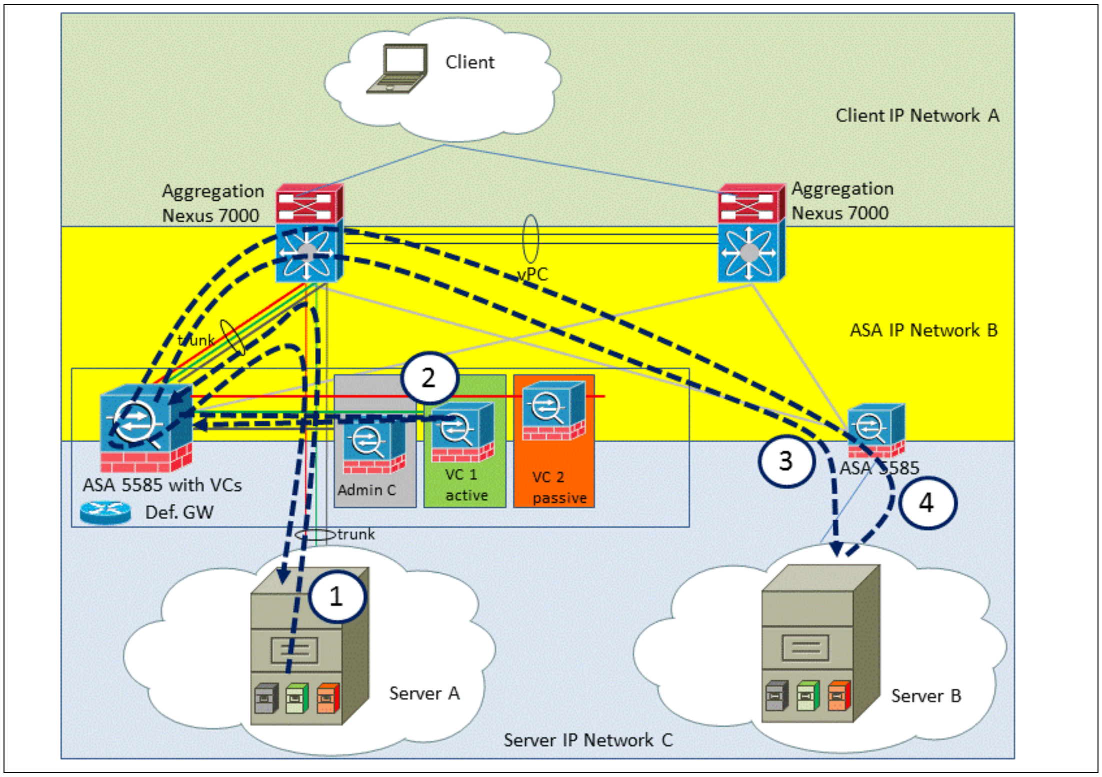


Figure 6-40 Routed mode ASA server to server flow through

6.8 Cisco Security Manager (CSM)

Cisco Security Manager (CSM) is a security management application that provides insight into and control of Cisco security and network devices. The CSM can manage Cisco security appliances, including Cisco ASA 5500 Series Adaptive Security Appliances, Cisco IPS 4200 Series Sensor Appliances, Cisco Integrated Services Routers (ISRs), Cisco ASA Services Modules, and Cisco Catalyst 6500 Series switches.

The CSM integrates a variety of capabilities, including policy and object management, event management, reporting, and troubleshooting.

6.8.1 Security policy management

Security administrators can create reusable network objects such as network addresses and services, which are defined once and used any number of times. Cisco Security Manager also allows policies to be defined once and shared across devices. This minimizes errors associated with manual entry of data, and makes the management of security objects and policies more efficient. Administrators can implement both on-demand and scheduled security deployments, and can roll back to a previous configuration if required. Role-based access control and deployment workflows help ensure that compliance processes are followed.

Cisco Security Manager also enables flexible provisioning of IPS signature updates, providing administrators with the ability to incrementally provision new and updated signatures, create IPS policies for those signatures, and then share the policies across devices. Additionally, insight into Cisco Security Intelligence Operations (SIO) recommendations allows administrators to fine-tune their environment prior to deploying signature updates.

Event management and troubleshooting

Integrated event management provides administrators with real-time incident analysis and rapid troubleshooting, while advanced filtering and search capabilities enable them to quickly identify and isolate interesting events. Cross-linkages between the event manager and configuration manager reduce troubleshooting time for firewall rules and IPS signatures.

Reporting

The CSM has a report manager, which generates system reports from events that it receives and displays them as either charts or data grids. It also allows administrators to define and save custom reports using report criteria to meet their specific reporting needs.

6.8.2 Cisco security manager primary use cases

Within data centers, the following two management capabilities will primarily be used:

Firewall management

This capability includes the following activities:

- Object and rule sharing capabilities enable administrators to efficiently and consistently maintain their firewall environment.
- Policy query feature displays which rules match a specific source, destination, and service flow, including wildcards. This feature allows the administrator to define policies more efficiently.
- To ease configuration, device information can be imported from a device repository or configuration file, or added in the software; additionally, firewall policies can be discovered from the device itself - this feature simplifies initial security management setup.
- Consumption of ASA/ASASM/FWSM syslog events.
- System-generated and customized reports, including firewall traffic and botnet reports.

IPS management

This capability includes the following activities:

- Incremental provisioning of new and updated signatures.
- Insight into Cisco SIO recommended defaults allows customers to tune their environment before distributing the signature update.
- IPS signature policies and event action filters can be inherited and assigned to any device. All other IPS polices can be assigned to and shared with other IPS devices; IPS management also includes policy rollback, a configuration archive, and cloning or creation of signatures.
- IPS update administration and IPS subscription licensing updates allow users to manage IPS software, signature updates, and licensing based on local and shared polices.
- Consumption and viewing of IPS Security Device Event Exchange (SDEE) events.
- System and custom IPS reports, including top attackers, top signatures.

The features of the current version of Cisco's CSM can be found here:

http://www.cisco.com/en/US/prod/collateral/vpndevc/ps5739/ps6498/data_sheet_c78-711505.html

6.9 Cisco TrustSec

The explosion in consumer IT devices, the need for global collaboration, and the move to cloud-based services and virtualization means that the traditional corporate network boundary is a thing of the past. Mobile workforce, collaboration, and productivity requirements have never been greater.

The Cisco TrustSec architecture recognizes this fundamental change by telling you who and what is connecting to your wired or wireless network, and providing an innovative mechanism to control of what they can do and where they can go while they are there.

The Cisco TrustSec Solution includes these benefits:

- Enables business productivity: Gives increasingly mobile and complex workforces access to the right resources with assured security.
- Delivers security and compliance risk mitigation: Provides visibility into who and what is connecting to the network, and control over what they can access and where they can go.
- Improves IT operational efficiency: Reduces IT overhead through centralized security services, integrated policy management, and scalable enforcement mechanism.

Cisco TrustSec uses your networking infrastructure to deliver the following technical capabilities in a scalable manner:

- Identity-enabled networking: Flexible authentication (FlexAuth) methods, including IEEE 802.1X, WebAuth, and MAC Address Authentication, using the network to ascertain the identity of users and devices on your network.
- Context awareness: The distributed, granular scanning and endpoint inspection elements in Cisco TrustSec provide contextualized visibility into the "who, how, what, and when" for the identities of users and devices accessing the network.
- Highest precision device profiling for any endpoint: Automatically identifies and classifies devices by collecting endpoint data through the built-in ISE probes or network-based device sensors on the Cisco Catalyst switching and wireless infrastructure, and further refines the classification with directed policy-based active endpoint scanning technology.
- Guest user access and lifecycle management: Sponsored guests receive restricted access to specific resources through a customized web portal. Internal network access is blocked, and activity is tracked and reported.
- Centralized policy and enforcement: A centralized policy platform enables coordinated policy creation and consistent, context-based policy enforcement across the entire corporate infrastructure, spanning the head office, data center, branch office, and remote users. Non-compliant devices can be quarantined, remediated, or given restricted access.
- Topology-independent access control: Provides a scalable and flexible way to assign roles via network "tags" to authorize users and devices, and to enable any network to enforce policies based on these tags. This offers a unique, scalable architecture for network enforcement without network redesign using VLANs or having to manage a multitude of lengthy ACLs.
- Data integrity and confidentiality: Hop-by-hop standards-based MACsec encryption provides data confidentiality with visibility in the flows for security-based access policy enforcement.

- Monitoring, management, and troubleshooting: Centralized, policy-based corporate governance and compliance includes centralized monitoring and tracking of users and devices to maintain policy compliance. Provides sophisticated troubleshooting, detailed auditing, and historical and real-time reporting.
- Integration with Cisco Prime Network Control System: Provides a unified view of all network functions to streamline your network management efforts.

Figure 6-41 illustrates the overall scope of the TrustSec solution: to secure customer networks by building and enforcing identity and context-based access policies for users and devices while protecting critical data throughout the network.

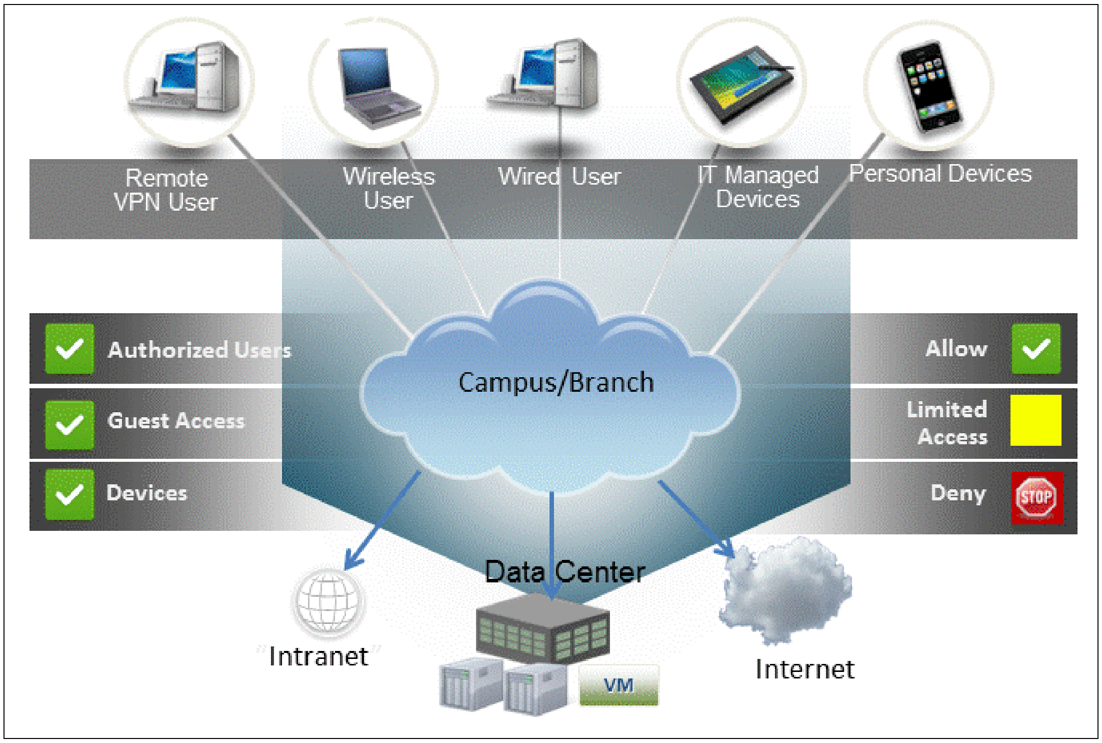


Figure 6-41 Scope of the TrustSec solution

Cisco Catalyst 2960, 3560, 3750, 4500, and 6500 Series switches, Cisco Nexus 7000, 5000, 2000, and 1000v Series switches, Cisco Wireless LAN Controllers, Cisco Integrated Services Router Generation 2 (ISR-G2) platforms, and Cisco ASR 1000 Series Aggregation Services Routers interact with network users for authentication and authorization. Access to the wired or wireless network is dictated by policy, user identity, and other attributes.

Flexible authentication methods include 802.1X, web authentication, and MAC authentication bypass, all controlled in a single configuration for each switch port. Device sensors in the wired and wireless infrastructure automatically detect and help to classify devices attached to the network with minimal effort. Cisco switches that use Security Group Access technology can tag each data packet with user identity information so that further controls can be deployed anywhere in the network. In addition, Cisco Nexus switches support MACsec (IEEE 802.1AE standard encryption) today for data-in-motion confidentiality and integrity protection.

Refer to the TrustSec page for a more detailed overview of the TrustSec architecture and solution components, as well as the latest matrix of the TrustSec features available on different Cisco platforms at this website:

http://www.cisco.com/go/trustsec

6.9.1 Trustsec in the data center

Although TrustSec technologies do play a key role in enforcing security policies in the access layer of a campus network, it also includes capabilities that provide a powerful tool for access control within the data center. Here are a few of the key business drivers for Trustsec in the data center:

- Compliance:

  Enterprises must comprise to regulations, and show how they are compliant. TrustSec segmentation and encryption technologies reduce the scope of compliance.

- Bring Your Own Device (BYOD):

  Enterprises require differentiated access to data center assets based on users/devices. TrustSec provides a scalable way to enforce access to the valuable assets in the data center, regardless of what device type the user is using to access the resources.

- Virtual Desktop Infrastructure:

  Security for the VM user is a big concern for enterprises. TrustSec provides enforcement for the VM user to access the appropriate resources in the data center.

There are two main technologies that we explore in regards to Trustsec in the data center:

- Security Group Access
- MACSec Link Encryption

6.9.2 Security Group Access

Security Group Access (SGA) is a next-generation of Access Control Enforcement that was created to address the growing operational expenses with maintaining firewall rules and access-lists. SGA is a complimentary enforcement technology that removes the concerns of overrunning switch TCAM memory and difficult to manage Access Control Lists.

The ultimate goal of SGA is to assign a TAG (known as a Security Group Tag, or SGT) to the user/device's traffic at ingress (inbound into the network), and then enforce the access elsewhere in the infrastructure (in the data center, for example). So, SGA assigns a TAG at login, and enforces that TAG elsewhere in the network (egress enforcement).

There are three foundational pillars of Security Group Access: Classification, Transport, and Enforcement. Where classification is the ability to accept the tag for a particular network authentication session; transport is the ability to send that assigned tag to upstream neighbors either via native tagging, or using Security group eXchange Protocol (SXP). Enforcement may be on switches using Security Group ACLs (SGACLs) or on a Security Group Firewall (SG-FW).

Security Group Tag

Security Group Tag (SGT) is a 16-bit value that is assigned (usually by the Cisco Identity Solutions Engine) to the user or endpoint's session upon login. The Network infrastructure views the SGT as another attribute to assign to the session, and will insert the Layer 2 tag to all traffic from that session. The SGT can represent the context of the user and device.

In most cases, 802.1X will not be used in the data center. Servers are not usually required to authenticate themselves to the data center switch, as the DC is normally considered physically secure and there is no network access control applied there. However, the servers themselves will be the destination of traffic coming from the campus and from within the data center itself.

Since 802.1X is not typically used in the data center, we need a manual way to apply the SGT. This is configured at the interface level of the Nexus configuration and is manually applied to the port itself. The Nexus 7000, 5000, 2000, and 1000v switches support this classification capability.

Figure 6-42 illustrates a Contractor IT Administrator accessing the network. The user's traffic is tagged with an SGT = 100 as it enters the network. The traffic traverses the network with the associated tag, and ultimately enforced in the data center. In the data center, traffic from the IT Server is tagged with SGT = 10, while traffic from the database server is tagged with SGT = 4. As the user attempts to access the resources in the data center, he or she is allowed to access the IT Server but denied access to the Database server. These security policies are enforced purely on the Security Group Tags rather than IP addresses, port numbers, or other traditional network information.

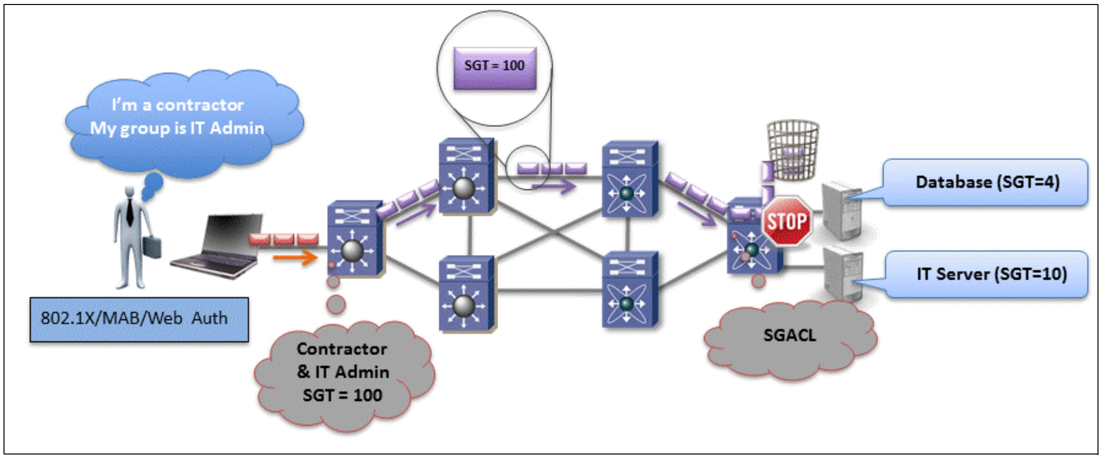


Figure 6-42 Contractor IT Administrator access

Security Group eXchange Protocol

Ideally, all of your Access-Layer devices will support tagging the users traffic natively. Not all devices will support native tagging. However, all supported Cisco Switches will support the assignment of the SGT to the authentication session (known as classification).

Cisco has developed a peering protocol (like BGP) to allow devices to communicate their database of IP address to SGT mappings to one another. This peering protocol is called Security Group eXchange Protocol (SXP). As this is a peering protocol, it is possible to be very specific and deterministic as to which devices send updates (speaker) and which ones receive updates (receiver). A network device may peer via SXP directly to the enforcement device (the data center switch or firewall).

The Nexus 7000, 5000, 2000, and 1000v switches support SXP.

Enforcement

Now that we have Security Groups being assigned (Classification) and they are being transmitted across the network (Transportation), it is time to focus on the third pillar of Security Group Access: Enforcement. Enforcing a security policy based on Security Group Tags is far simpler than the traditional network-based ACLs that are traditionally used in networks.

There are multiple ways to enforce traffic based on the tag, but ultimately, we can summarize them into two major types:

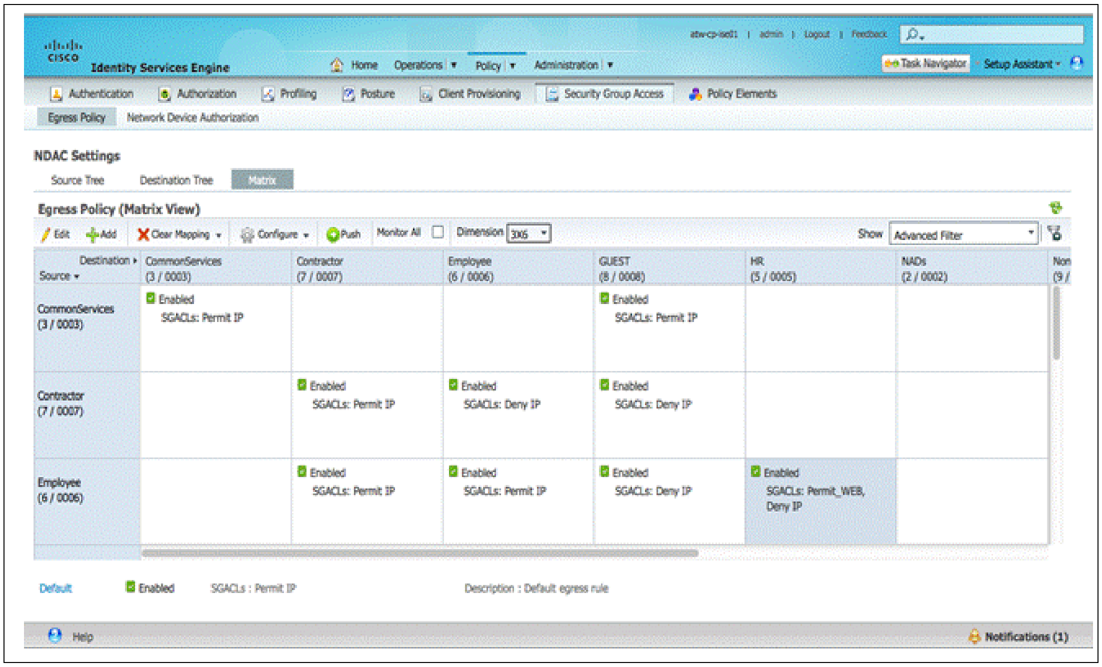
- Enforcement on a Switch (SGACL)
- Enforcement on a Firewall (SG-FW)

SGACL (Security Group Access Control List)

Historically, enforcement with SGACL was the only option available. It started with the Nexus 7000 Series and has expanded to the Nexus 5000 Series, Catalyst 6500 (Supervisor 2T), and the 3000X Series switches. A major benefit to SGACL usage is the consolidation of Access Control Entries (ACEs) and the operational savings involved with maintenance of those traditional Access Lists.

An SGACL can be visualized in a format very similar to a spreadsheet. It is always based on a Source Tag to a Destination Tag.

Figure 6-43 shows an example SGACL policy on the Cisco Identity Solutions Engine, which represents the SGACLs in a columns and rows presentation. The box that is highlighted in gray shows that when traffic with a source of the Employee SGT (6) attempts to reach a destination with the HR SGT (5), an SGACL named “Permit_WEB” will be applied and a catch-all of “Deny IP”.



Destination ▸ Source ▾	CommonServices (3 / 0003)	Contractor (7 / 0007)	Employee (6 / 0006)	GUEST (8 / 0008)	HR (5 / 0005)	NADs (2 / 0002)	Non (9 /
CommonServices (3 / 0003)	☑ Enabled SGACLs: Permit IP			☑ Enabled SGACLs: Permit IP			
Contractor (7 / 0007)		☑ Enabled SGACLs: Permit IP	☑ Enabled SGACLs: Deny IP	☑ Enabled SGACLs: Deny IP			
Employee (6 / 0006)		☑ Enabled SGACLs: Permit IP	☑ Enabled SGACLs: Permit IP	☑ Enabled SGACLs: Deny IP	☑ Enabled SGACLs: Permit_WEB, Deny IP		

Default ☑ Enabled SGACLs : Permit IP Description : Default egress rule

Figure 6-43 SGACL policy example

There are two main ways to deploy SGACLs: North-South and East-West. North-South refers to the use-case of a user or device being classified at the Access-Layer, but enforcement with the SGACL occurring at the data center. For example, a guest entering the access-layer is assigned a GUEST Security Group Tag. Traffic with a GUEST SGT will be dropped if it tries to reach a server with financial data.

East-West refers to the use-case of an SGACL protecting resources that exist on the same switch. For example, a development server and a production server on the same Nexus 5000 Series switch in the data center, an SGACL may be deployed to prevent the development server from ever communicating with the production server.

The Nexus 7000, 5000, 2000, and 1000v switches support SGACLs.

Security Group Firewalls (SG-FW)

Although traffic enforcement can be done on the switching infrastructure with SGACLs, it can also be done on a device that was purpose-built to do traffic filtering, a firewall. Cisco has added the ability to enforce traffic on firewalls by implementing the Security Group Firewall (SG-FW). There are two different types of Security Group Firewalls that exist; there is the ASA-based SG-FW and the router-based SG-FW. The ASA is the typical SG-FW device to be implemented in the data center.

Beginning with ASA version 9.0, the ASA firewall gains SG-FW functionality. ASDM supports the full configuration, and therefore the ASA is the only SG-FW that has a GUI (as of the writing of this document). The SG-FW in the ASA is a very simple concept. The powerful firewall policy in the firewall has been expanded to include Source and Destination Security Groups into the decision.

Figure 6-44 illustrates the use SGT, SXP, and SGACLs together in a network. SXP is used to propagate the Security Group Tags across a campus network in which the devices along the path may or may not support SGT/SXP capability. As long as the sending and receiving network device supports SXP, the assigned Secure Group Tags can be distributed across the network. In this example, SXP is used to transmit the SGT information to the data center edge, and ultimately the SGT enforcement is done by the Nexus 5000 switch in the data center.

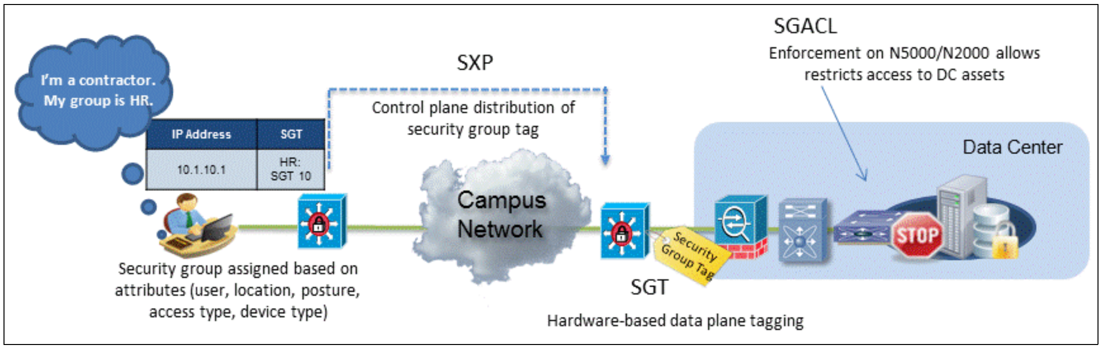

I'm a contractor.
My group is HR.
IP Address
SGT
10.1.10.1
HR:
SGT 10
SXP
Control plane distribution of
security group tag
Campus
Network
SGACL
Enforcement on N5000/N2000 allows
restricts access to DC assets
Data Center
STOP
Security
Group Tag
SGT
Security group assigned based on
attributes (user, location, posture,
access type, device type)
Hardware-based data plane tagging

Figure 6-44 SGT, SXP, and SGACLs

6.9.3 MACSEC link encryption

Enterprises and service providers store data in data centers that contain information that needs to be protected as intellectual property. Examples of such information are patient data, financial data, and regulatory data. Depending on the type of corporation, there might be regulatory requirements that have to be followed by the respective organization. Examples of data protection requirements are HIPAA, PCI, Sarbanes-Oxley (SOX), Basel II, and FDA.

In order to properly protect sensitive data, it is sometimes desirable to encrypt data that is passed between data centers, between DC rooms, or in some cases, between switches. This requirement holds true for data base synchronization traffic, applications exchanging state information as well as users accessing applications. Encrypting data in this manner not only assures data privacy but also data integrity.

In 2006 the IEEE ratified the 802.1AE standard, also known as MAC security standard (MACsec). MACsec encrypts all Ethernet frames, irrespective of the upper layer protocol. With MACsec, not only routed IP packets but also IP packets where the source and destination is in the same subnet or even non-IP traffic are encrypted.

As part of the Cisco TrustSec infrastructure, the Nexus 7000 supports IEEE 802.1AE standards based encryption (also known as MACsec) on all M-Series module interfaces. This hardware level encryption is where each port has its dedicated crypto engine that can be turned on in NX-OS configuration. Since each port has dedicated resources, there is no tax on the supervisor or line card CPU. All data encryption and decryption is performed at the port level. Apart from the additional MACsec header, there is no overhead or performance impact when turning on port level encryption. The Nexus 6000, 5000, and 1000v switches do not currently support MACsec functionality.

From a deployment standpoint, the main difference between MACsec and IPSec is that MACsec is performed at a link layer (meaning hop-by-hop basis), while as IPSec is fundamentally building a tunnel that goes from one IP address to another over potentially many hops.

802.1AE not only protects data from being read by others sniffing the link, it assures message integrity. Data tampering is prevented by authenticating the relevant portions of the frame. Figure 6-45 illustrates how a regular Layer 2 frame is encrypted as it traverses the Nexus 7000 platform.

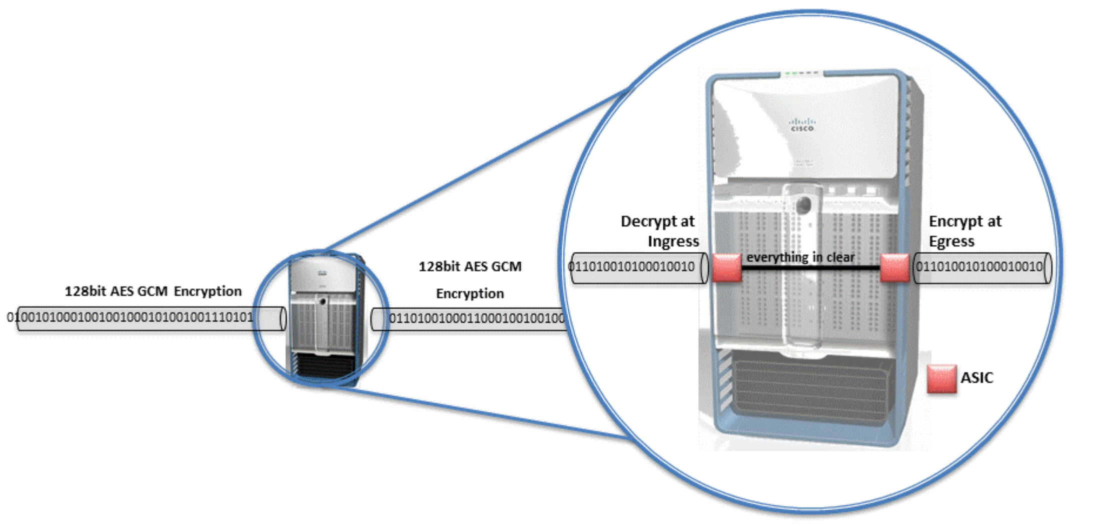


Figure 6-45 Layer 2 frame encryption

6.10 Virtualized Network Services

Network services are frequently placed in-line or in the flow of traffic forming a line of Layer 4 - Layer 7 services. As applications are virtualized, their movement may take them out of the traffic path, creating challenges maintaining network services to VMs and their applications. In most data centers a mix of physical and virtual network services is emerging as well as a mix of virtual servers and physical servers. Network services can be offered to a VM and its associated traffic independent of its form factor, whether it is a physical appliance or a dedicated module of virtualized network services, as long as the VM and virtual distributed switch send traffic to the appropriate services as the application VM moves around the data center.

This is important as traffic pattens have shifted from primarily north to south to a mix of east to west and north to south, resulting in the need for network services to offer much greater flexibility in their service to VMs and the applications they contain. In this approach, network services are logically wrapped around a VM through a policy and they move together. This addresses the challenge of intensive change management in a virtualized environment. In addition to making network services accessible independent of their location, there is the added benefit that virtual network services decrease the number of hardware appliances in the data center, which reduces complexity, total cost of ownership, and energy consumption.

6.10.1 Virtual Services Appliances

Cisco UNS is not just s suite of Layer 4-Layer 7 network services offerings, but a framework for transparently inserting network services into a virtualized environment. The services are not limited to just the products themselves, but also include the key enabling technologies to redirect VM traffic to and from the virtual appliance through the Cisco Nexus 1000V and vPath. Today, DCs have the option of choosing a physical appliance, which can be virtualized using virtual device contexts, or a virtual appliance to meet the network service requirements as shown in Figure 6-46 and Figure 6-47 next. The decision factors that go into the decision of physical or virtual are scale, performance, multi-tenancy capabilities, application mobility, service-chaining, and configuration and device management as it relates to vm-attributes. The remaining focus will be on virtual network services.

The following technologies are covered in this section:

- vPath
- Virtual Security Gateway (VSG)
- Virtual ASA (vASA)
- Virtual WAAS (vWAAS)
- Cisco Prime Network Analysis Module (vNAM)

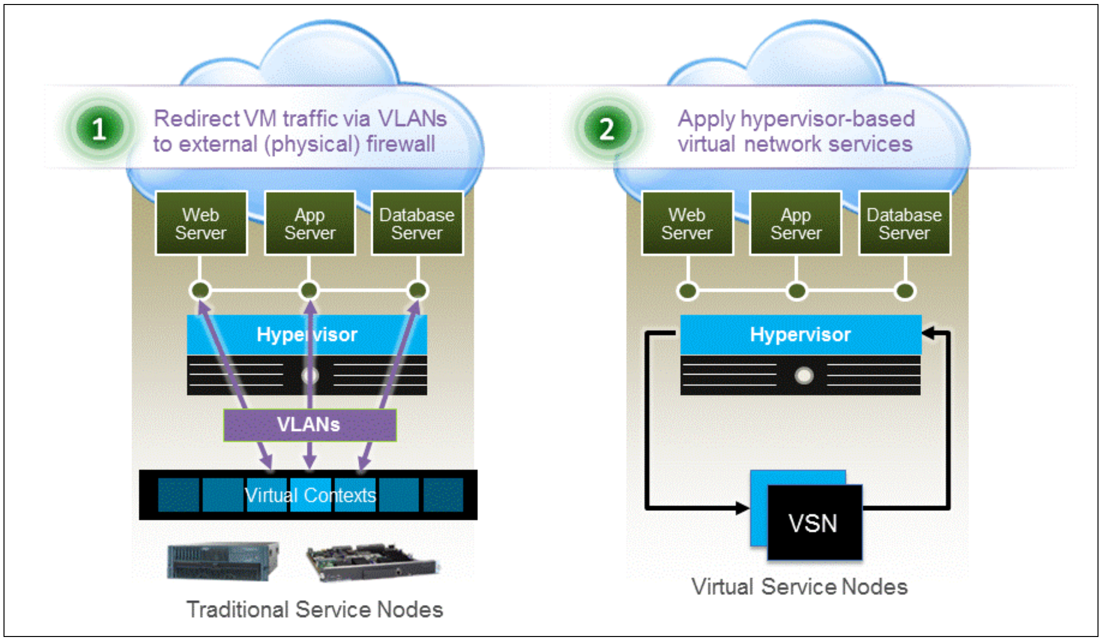


Figure 6-46 Deployment options in virtualized or cloud DC

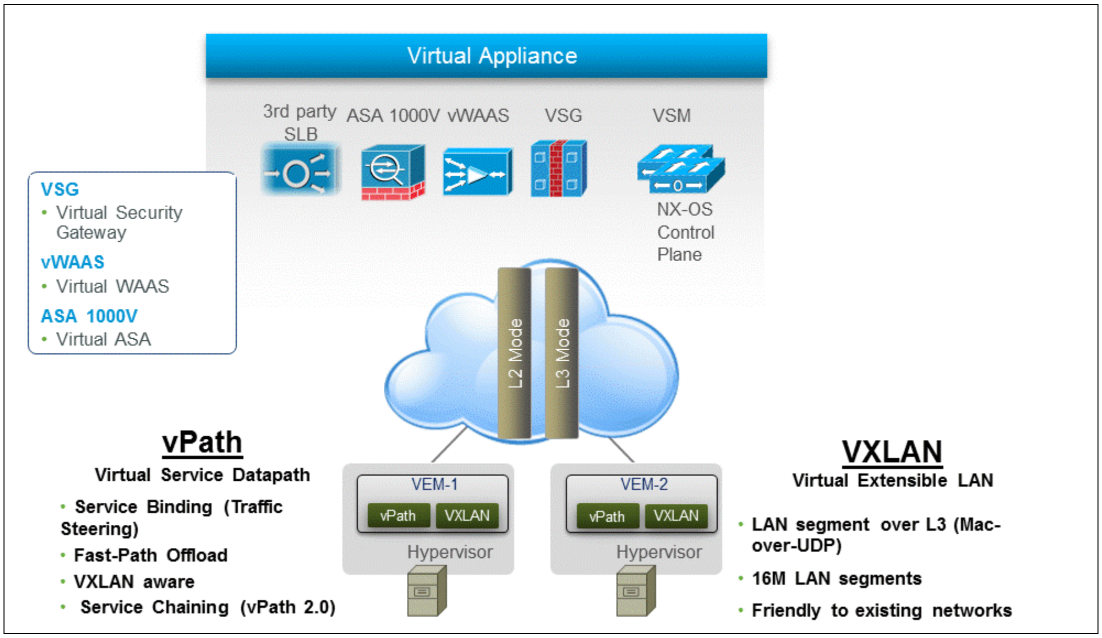


Figure 6-47 Embedding Intelligence for Virtual Services: VXLAN

6.10.2 Virtual network service data path (vPath)

A challenge with deploying any network service is how to insert or deploy the service with minimal complexity while also allowing for future service-chaining requirements. Cisco vPath is a transparent service insertion framework to insert services to ease the burden of inserting network services without the need for VLAN chaining, pbr, wccp, and so on.

Cisco vPath provides embedded intelligence within Cisco Nexus 1000V Series virtual Ethernet modules (VEMs). The Cisco vPath architecture provides a forwarding-plane abstraction and a programmable framework for inserting network services such as firewalls, load balancers, and WAN optimization at the virtual access layer. Cisco vPath is implemented in the switch data path with a set of configuration tables, flow tables, and well-defined interfaces for adding and removing flow entries. Each entry in the flow table associates a data path action (for example, a redirect, permit, or drop action). Cisco vPath enables separation of the control plane and data plane with well-defined APIs.

Because Cisco vPath resides in the switch data path, if configured, it can intercept the incoming traffic and the virtual machine-to-virtual machine traffic and redirect it to appropriate network services for processing. The network service nodes can be deployed as either Layer 2 or Layer 3 adjacent to Cisco vPath. Cisco vPath uses overlay tunnels for redirecting packets to network service nodes. Network services can optionally offload packet processing from network service nodes by programming appropriate flow entries in the Cisco vPath flow table, and also can insert flow entries to intercept traffic on demand. Cisco vPath also provides data-path orchestration to chain multiple services as well as service clustering to scale services.

vPath provides the forwarding plane abstraction and programmability required to implement the Layer 2 to Layer 7 network services such as segmentation firewalls, edge firewalls, load balancers, WAN optimization, and others. It is embedded in the Cisco Nexus 1000V Series switch Virtual Ethernet Module (VEM). It intercepts the traffic whether external to the virtual machine or traffic from virtual machine to virtual machine and then redirects the traffic to the appropriate virtual service node (VSN) such as Cisco Virtual Security Gateway (VSG), Cisco ASA 1000V, Cisco Virtual Wide Area Application Services (vWAAS) for processing. vPath uses overlay tunnels to steer the traffic to the virtual service node and the virtual service node can be either Layer 2 or Layer 3 adjacent.

The basic functions of vPath includes traffic redirection to a virtual service node (VSN) and service chaining. Apart from the basic functions, vPath also includes advanced functions such as traffic off load, acceleration and others.

vPath steers traffic, whether external to the virtual machine or from a virtual machine to a virtual machine, to the virtual service node. Initial packet processing occurs in the VSN for policy evaluation and enforcement. Once the policy decision is made, the virtual service node may off-load the policy enforcement of remaining packets to vPath.

Cisco vPath for Cisco Nexus 1000V Series switches provides the following benefits:

- Enables granular enforcement of network policies
- Eliminates the need for placing network services inline; using the overlay model, Cisco vPath can tunnel traffic to network services placed anywhere in the network
- Provides virtual machine mobility awareness
- Provides multi-tenancy awareness
- Enables data-path acceleration by offloading policy from network services to Cisco vPath
- Provides a scale-out model for horizontal scaling of network infrastructure
- Enables chaining of multiple network services

The other components of the virtual service architecture includes as shown in Figure 6-48:

- The Virtual Network Management Center (VNMC), a multi tenant policy manager responsible for device and policy management and integration with VMware vCenter. VNMC is the overall management and orchestration component of the virtual service architecture.
- The Cisco Nexus 1000V Series switch VSM, responsible for all the interactions with vPath and with VNMC. The Virtual Service Agent on the Cisco Nexus 1000V Series switch is responsible for all the control aspects of vPath such as traffic classification, traffic redirection, service chaining, traffic off loading and acceleration.
- Virtual Service Node (VSN), responsible for the service processing. The various virtual services supported include VSG, vWAAS, ASA for 1000V and others. The VSNs can include many instances of the same virtual service or different virtual service types.

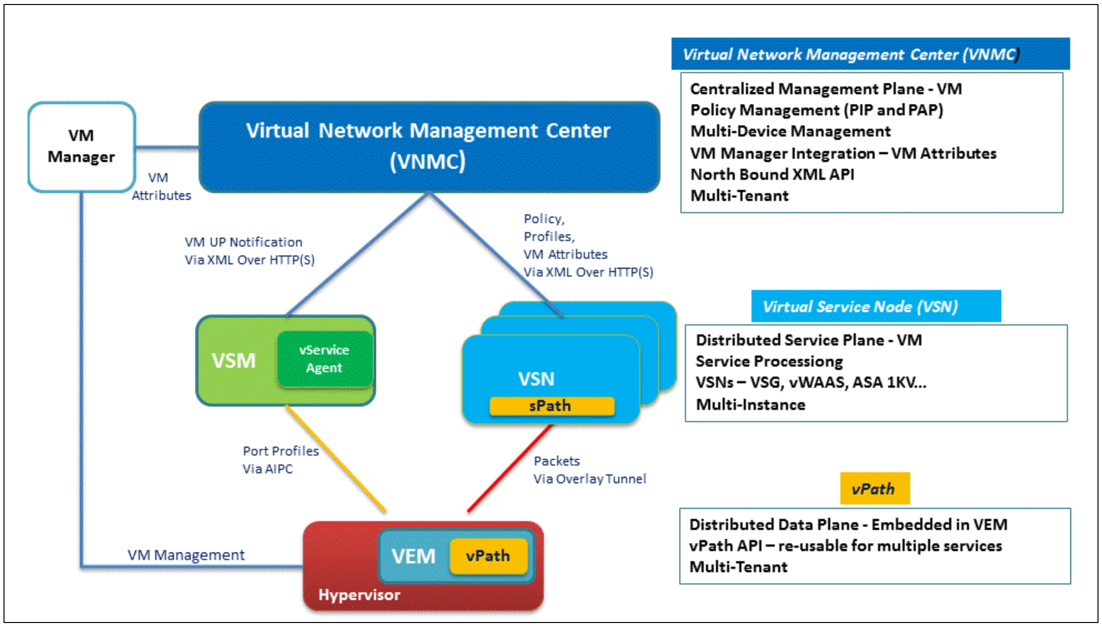

Virtual Network Management Center (VNMC)
Centralized Management Plane - VM
Policy Management (PIP and PAP)
Multi-Device Management
VM Manager Integration – VM Attributes
North Bound XML API
Multi-Tenant
VM Manager
VM Attributes
Virtual Network Management Center (VNMC)
VM UP Notification Via XML Over HTTP(S)
Policy, Profiles, VM Attributes Via XML Over HTTP(S)
Virtual Service Node (VSN)
Distributed Service Plane - VM
Service Processiong
VSNs – VSG, vWAAS, ASA 1KV...
Multi-Instance
VSM
vService Agent
VSN
sPath
Port Profiles Via AIPC
Packets Via Overlay Tunnel
vPath
Distributed Data Plane - Embedded in VEM
vPath API – re-usable for multiple services
Multi-Tenant
VM Management
VEM
vPath
Hypervisor

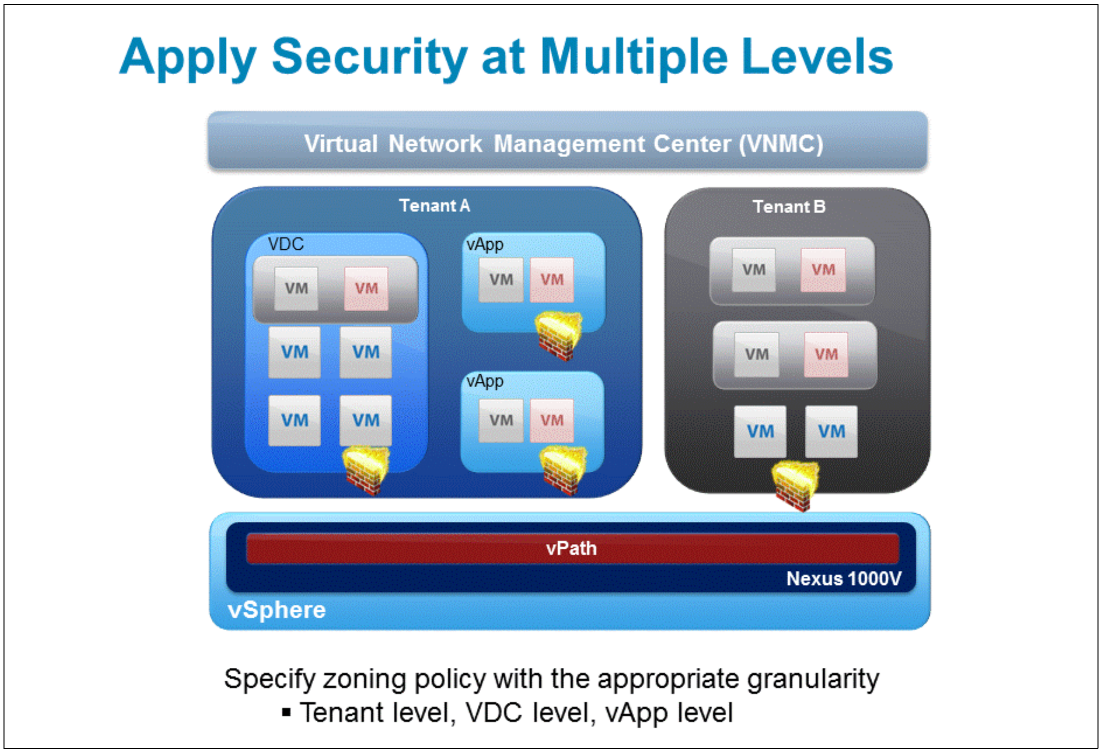
Figure 6-48 Virtual services architecture

Dynamic service provisioning

vPath supports dynamic provisioning of virtual machines via service profiles and ensures that the service profiles follow vMotion events in ESX environments. The service parameters are configured in a service profile and then attached to a port profile. When the virtual machines get instantiated and attached to a port profile, the service profile also gets dynamically attached to the virtual machine. Once associated, all the policies are dynamically provisioned to a virtual machine as the virtual machine comes up or moves from one server to another.

The service parameters are configured in a service profile and then attached to a port profile. When the virtual machines get instantiated and attached to a port profile, the service profile also gets dynamically attached to the virtual machine. Once associated all the policies are dynamically provisioned to a virtual machine as the virtual machine comes up or moves from one server to another.

The virtual services architecture supports a collaborative management model where the roles and responsibilities of network administrator, server administrator and service administrator are clearly defined as shown in Figure 6-49.



Apply Security at Multiple Levels

Specify zoning policy with the appropriate granularity

- Tenant level, VDC level, vApp level

Figure 6-49 Apply security at multiple levels

The virtual services architecture enables the mobility of the virtual machine as well as the virtual service node. Dynamic service provisioning ensures that the virtual machine traffic flow continues to be handled by the appropriate virtual service node. This is possible since the service profile remains the same in the port profile and the port profile moves along with the virtual machine. As a result the virtual machine in the new host will continue to use the same virtual service node for service processing.

Virtual network service chaining

The Cisco Nexus 1000V vPath service-chaining feature provides the capability to define multiple network services in a specific order that network traffic flow must follow to reach the end application. Network administrators can define the network services in the Cisco Nexus 1000V Series port profile, and they can also define the specific order that they want the traffic flow to follow for the defined network services. This Cisco Nexus 1000V vPath capability significantly simplifies the deployment of virtual network services. Because network services can be defined in a centralized place, it also makes insertion and removal of network services much simpler.

The following virtual services are supported by Cisco Nexus 1000V Series switch using the vPath:

- Cisco Virtual Security Gateway (VSG): provides trusted multi-tenant access with granular zone-based security policies for VMs. Cisco VSG delivers security policies across multiple servers. It supports VM mobility across physical servers for workload balancing, availability, or scale.
- Cisco Virtual Wide Area Network Application Services (vWAAS): a WAN optimization solution, helps deliver assured application performance acceleration to IT users connected to enterprise data centers and enterprise private clouds.
- Cisco ASA for 1000V: provides trusted security to multi-tenant virtual and cloud infrastructures at the edge. When implemented with the Cisco Nexus 1000V Switch, it provides consistent security across physical, virtual, and cloud infrastructures as shown in Figure 6-50.

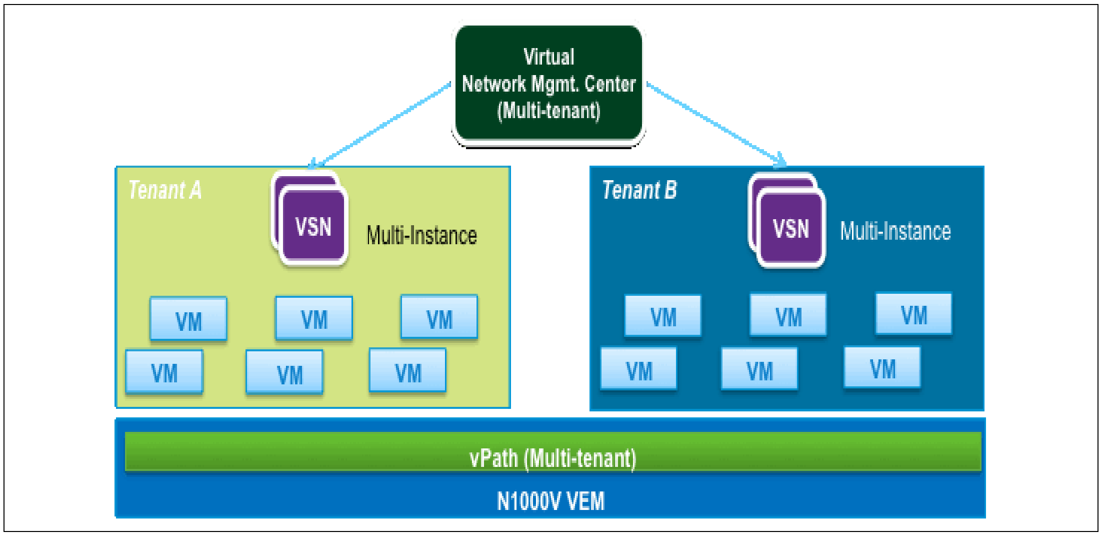


Figure 6-50 VNM

The Cisco vPath service-chaining capability is available with Cisco Nexus 1000V Series switches Release 1.5(2) and later. The first release of Cisco vPath service-chaining supports two network services, Cisco ASA 1000V and Cisco VSG. Service-Chaining with vWAAS will be support in CY2013.

Figure 6-51 illustrates the virtual multi-tenant security use case using Cisco Nexus 1000V Series service-chaining capabilities with Cisco VSG and Cisco ASA 1000V. Securing a virtual multi-tenant data center requires tenant-edge firewalls that can protect northbound and southbound traffic and also firewalls within the tenant that can protect eastbound and westbound traffic. Cisco ASA 1000V and Cisco VSG provide a comprehensive security solution for both tenant-edge and intra-tenant security needs using the Cisco Nexus 1000V vPath service-chaining feature.

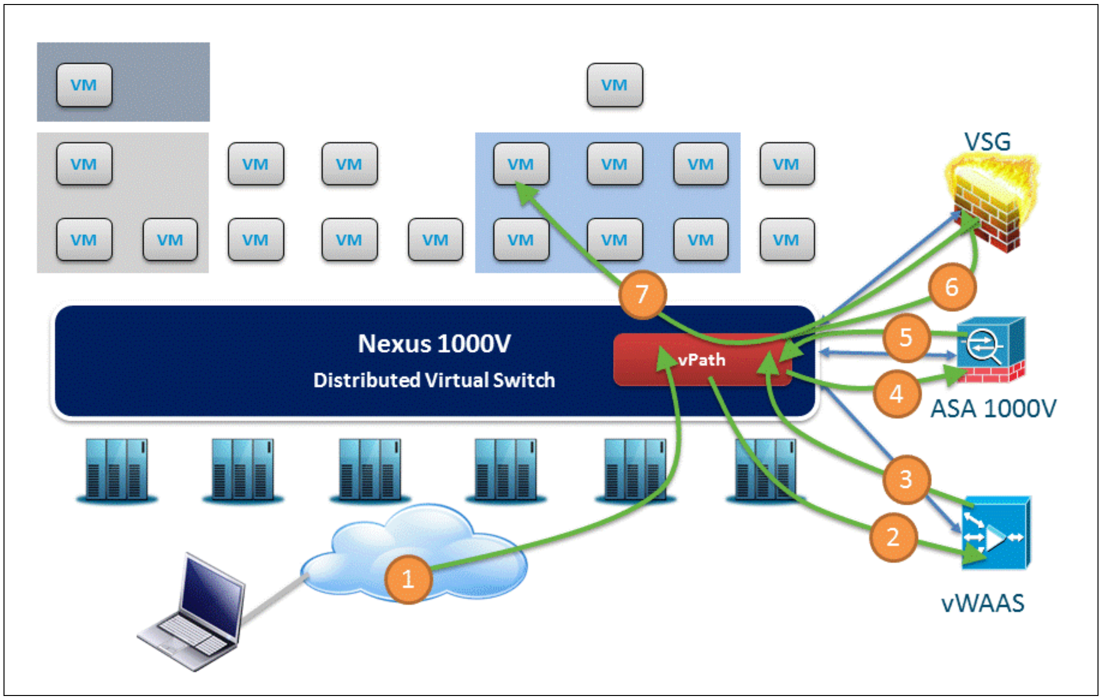


Figure 6-51 Intelligent traffic steering through multiple network services

The Cisco Nexus 1000V vPath is flow aware. It maintains flow tables to track the state of each flow. For example, a given flow may include the following events:

1. Cisco vPath intercepts packets for a new flow destined for a virtual machine. It looks up the flow table and determines that the new flow does not exist.
2. On the basis of the user configuration, Cisco vPath forwards packets to the appropriate network services (firewall, load balancer, and so on) by encapsulating the original packet using Cisco vPath encapsulation. Cisco vPath encapsulation also includes the policy information that network services need to apply to the traffic flow.
3. Network services evaluate the policy and send the policy action back to Cisco vPath. In addition, network services can optionally instruct Cisco vPath to offload further processing of the flow to vPath, providing performance acceleration for the network services.
4. On the basis of the policy result, Cisco vPath will add the flow to its flow table and forward the original packet to the eventual destination or to any additional network services.

Multiple network services can be chained using Cisco vPath with centralized configuration.

See the *vPath Reference Guide* at the following website:

`http://www.cisco.com/en/US/docs/switches/datacenter/vsg/sw/4_2_1_VSG_1_4_1/vpath_vservices/reference/guide/vPath_vService_Reference.html`

6.10.3 Cisco Virtual Security Gateway

Cisco Virtual Security Gateway (VSG) for the Cisco Nexus 1000V Series switch is a virtual firewall appliance that provides trusted access to virtual data center and cloud environments with dynamic policy-driven operation, mobility-transparent enforcement, and scale-out deployment for dense multi-tenancy. The Cisco VSG enables a broad set of multi-tenant workloads that have varied security profiles to share a common compute infrastructure. By associating one or more VMs into distinct trust zones, the VSG ensures that access to trust zones is controlled and monitored through established security policies.

Together, the Cisco VSG and Cisco Nexus 1000V Virtual Ethernet Module provide the following benefits:

- Efficient deployment: Each Cisco VSG can protect VMs across multiple physical servers, which eliminates the need to deploy one virtual appliance per physical server.
- Performance optimization: By offloading Fast-Path to one or more Cisco Nexus 1000V VEM vPath modules, the Cisco VSG boosts its performance through distributed vPath-based enforcement.
- Operational simplicity: A Cisco VSG can be deployed in one-arm mode without creating multiple switches or temporarily migrating VMs to different switches or servers. Zone scaling is based on security profile, not on vNICs that are limited for virtual appliances.
- High availability: For each tenant, one can deploy a Cisco VSG in an active-standby mode to ensure a highly-available operating environment with vPath redirecting packets to the standby Cisco VSG when the primary Cisco VSG is unavailable.
- Independent capacity planning: One can place a Cisco VSG on a dedicated server controlled by the security operations team so that maximum compute capacity can be allocated to application workloads. Capacity planning can occur independently across server and security teams and one can maintain operational segregation across security, network, and server teams.

Cisco VSG is a virtual security appliance that is tightly integrated with the Cisco Nexus 1000V distributed virtual switch. Cisco VSG uses the virtual network service path (vPath) technology embedded within the Cisco Nexus 1000V Virtual Ethernet Module (VEM). The vPath capability within the Cisco Nexus 1000V offloads the switching logic directly to the host, providing high performance, seamless interaction with other virtual appliances, and resiliency in case of appliance failure (see Figure 6-52). In addition, the Cisco Nexus 1000V vPath is tenant-aware, which allows for the implementation of security policies within and across multiple tenants.

The initial packet traffic flow through the Cisco Nexus 1000V and VSG is as follows:

1. The first packet of a flow enters the Cisco Nexus 1000V (this can be from an external user or from another VM) and is intercepted (per policy) by vPath (as shown in Figure 6-52).
2. vPath recognizes that this is a new flow that needs to be inspected by the designated VSG and sends it to the proper VSG (among multiple VSGs across multiple tenants).
3. VSG is a policy decision point. It applies the policy and then caches the decision in the Cisco Nexus 1000V.
4. If the decision is permit, the packet is sent to the destination VM, in this case in the purple zone. Otherwise the packet is dropped.
5. After the initial decision, ACL enforcement is off-loaded to vPath on the Cisco Nexus 1000V. Subsequent packets are processed in the hypervisor switch. There is a significant performance improvement, since most of the packets are processed by the fast path (as shown in Figure 6-53.

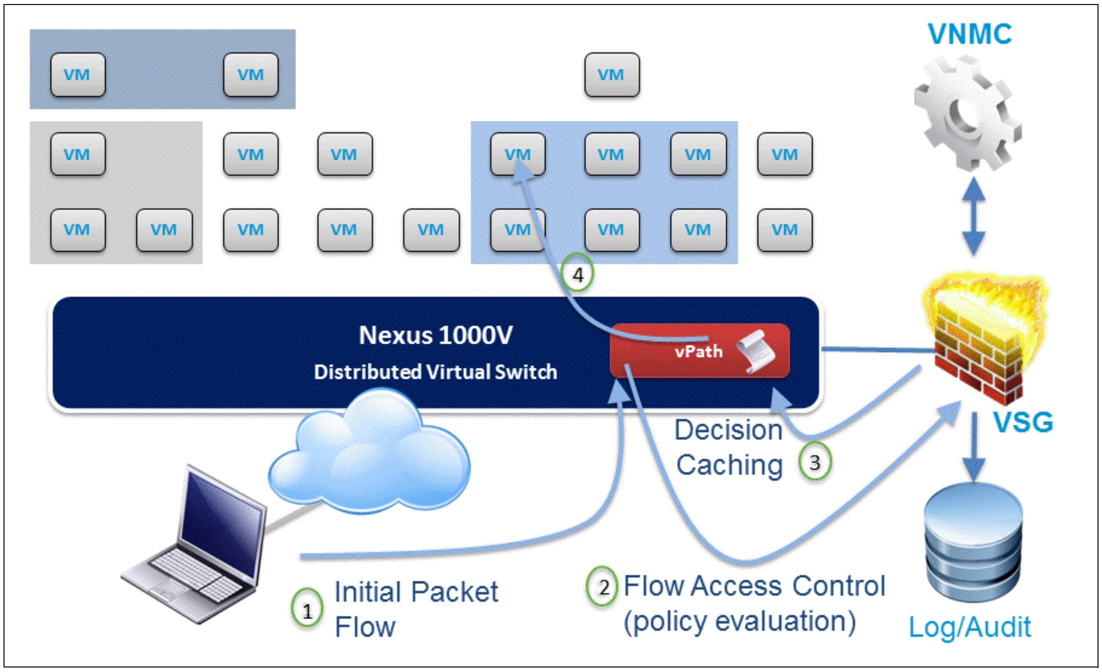


Figure 6-52 Intelligent traffic steering with vPath

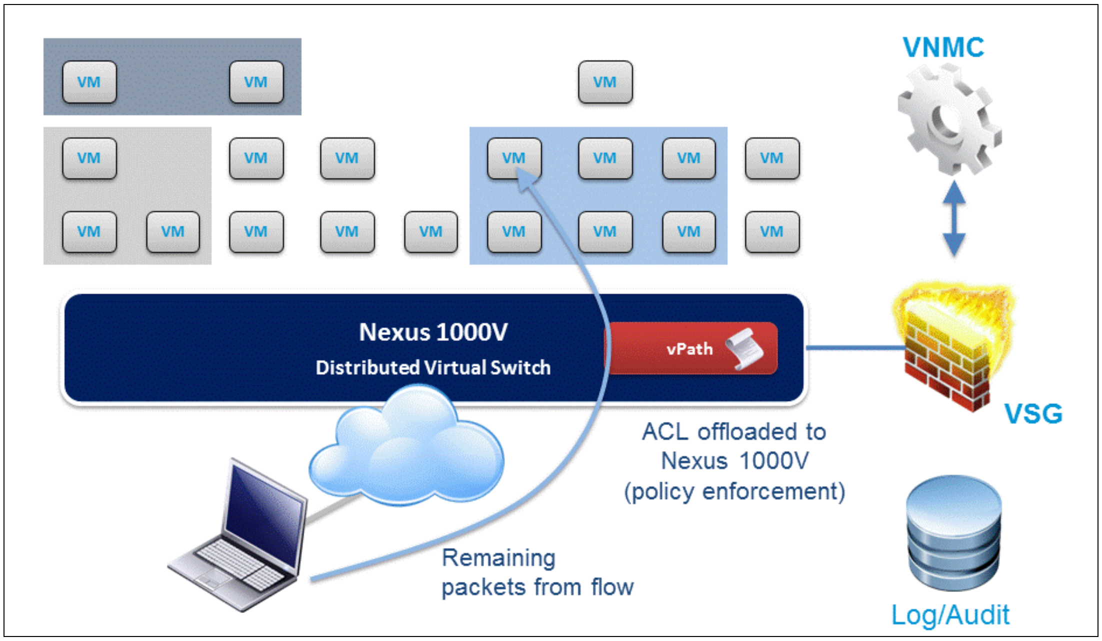


Figure 6-53 Performance acceleration with vPath

VSG also leverages the port-profile capability of the Cisco Nexus 1000V. Security policies defined on the VSG are bound to a particular port-profile. The Cisco Nexus 1000V manages and enforces security and port policies within each port-profile. A VM can easily inherit the attributes of port profile to which it has been assigned. Similarly, moving VMs to different hosts (by using VMware's vMotion capability) is seamless and all security attributes are preserved during this vMotion operation.

In addition to leveraging capabilities of the Cisco Nexus 1000V, the Cisco VSG delivers a rich array of features that makes it a perfect match for a multi-tenant, multi-segmented organization. Here are some of the VSG features:

- VM-level granularity and context-aware rules
- Ability to assign separate instances of VSG per tenant
- Capability of mapping rules and policy engines to an organization with a complex multi-segmented network and logical infrastructure
- Rich array of VM attributes that are leveraged in the VSG's policy engine, which enable the implementation of complex policies with fewer firewall rule-sets
- Bounded failure domains and a resilient architecture

The Cisco VSG uses VNMC for centralized management and for centralized policy enforcement. VNMC also interacts with the Cisco Nexus 1000V VSM and VMware vCenter as shown in Figure 6-54.

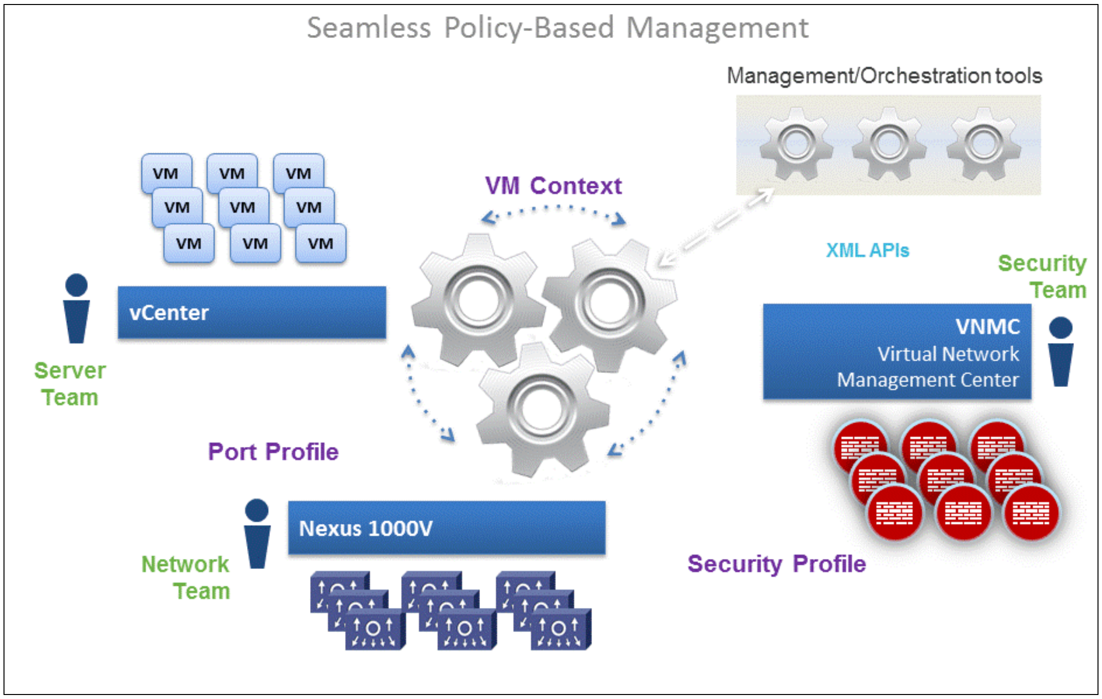


Figure 6-54 Virtual Network Management Center (VNMC)

The following summary describes the interactions between VNMC, VSG, Cisco Nexus 1000V, and VMware vCenter:

- VNMC communicates with VMware vCenter and accesses VMware vCenter VM attributes to use in defining security policies.
- VSG and VNMC communicate over secure Layer 3 (SSL) and with a pre-shared key. Using this secure channel VNMC publishes device and security policies to the VSG.
- VSM and VNMC also communicate over SSL with a pre-shared key. VSM provides VM-to-IP mapping information to VNMC.
- VEM communicates with VSG over a Layer 2 service VLAN. vPath redirects the data traffic over Service VLAN and the policy results from VSG are sent to vPath (VEM) by VSG.

VSG high availability

High availability and resilience against failures is of paramount importance in a data center. High availability of security-related appliances and services are no exception. VSG functional components utilize a redundant and resilient architecture. One can deploy redundant VSGs that synchronize their configurations and operate in active/standby mode. During a failure of a VSG, the standby VSG becomes active and operational within six seconds.

Similar functionality also exists in the Cisco Nexus 1000V Virtual Supervisor Module (VSM), where it also operates in an active/standby configuration. The VSG Virtual Network Management Center (VNMC) leverages the VMware high availability functionality to provide redundancy in case of hardware failures. VNMC deployment via the Cisco Nexus 1010 appliance offers an additional level of resiliency in the near future by decoupling the dependency on vSphere ESXi servers. It is also important to note that vPath data flow is not interrupted in case of failures as shown in Figure 6-55.

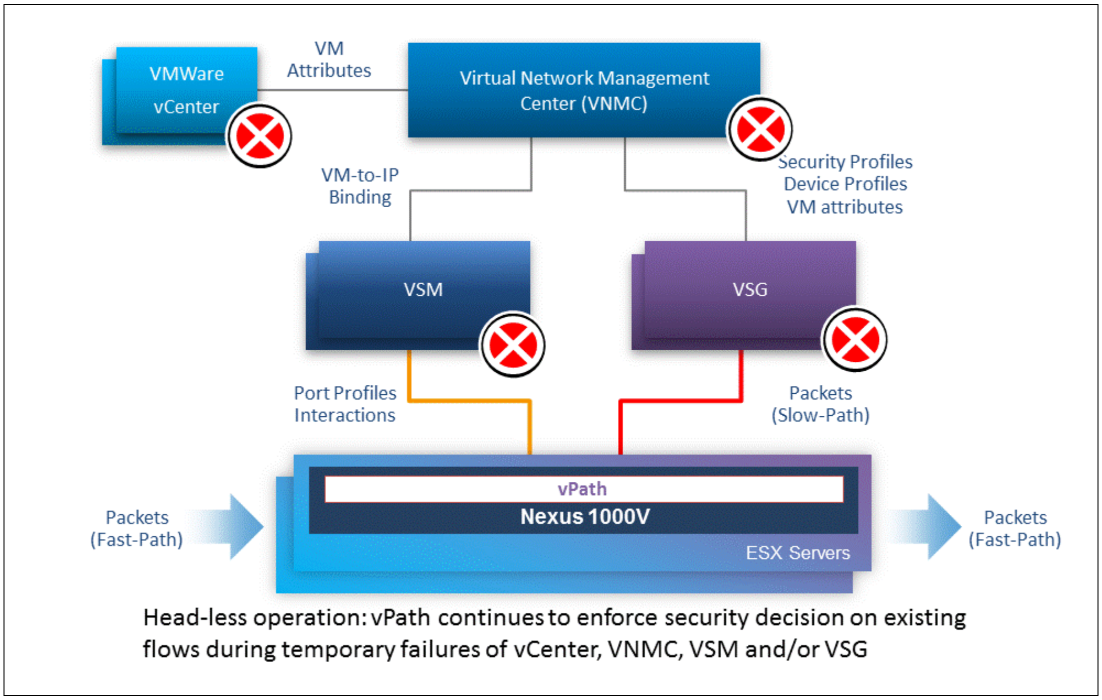

VMWare vCenter
VM Attributes
Virtual Network Management Center (VNMC)
VM-to-IP Binding
Security Profiles
Device Profiles
VM attributes
VSM
VSG
Port Profiles Interactions
Packets (Slow-Path)
Packets (Fast-Path)
vPath
Nexus 1000V
ESX Servers
Packets (Fast-Path)
Head-less operation: vPath continues to enforce security decision on existing flows during temporary failures of vCenter, VNMC, VSM and/or VSG

Figure 6-55 vPath data flow in case of failures

Security policy foundational components

The following discussion summarizes the building blocks that together define a security policy enforced by the VSG policy engine:

- Rules: A particular firewall policy instance is defined here. For example, a specific rule may only allow access to a particular VM from a specific subnet. These firewall rules can be based on various attributes, such as IP address/subnet, VM name, VM guest operating systems, zone name, cluster names, and so on.
- Policies: A policy aggregates more than one rule within its construct. A policy could contain a few different rules that apply to different VMs or subnets. Policies are set at different points within the tree and different policies can re-use the same rule.
- Policy set: A policy set contains one or more policies and can be defined at different points within the tenant tree.

- Security profile: A policy set is then mapped to a security profile. That security profile contains all the security definitions and rules that are contained within the policy set. As outlined below, that security profile then can be mapped to port profile within the Cisco Nexus 1000V.

Policy set containing a single policy

In a simple case, one can have a single policy within a policy set. This scenario is shown in Figure 6-56:

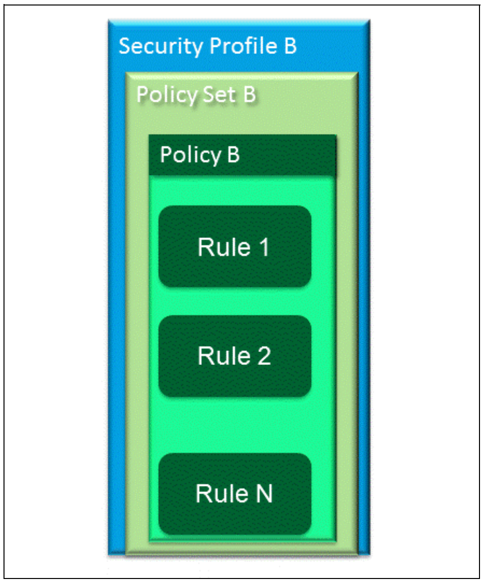


Figure 6-56 Security policy building block

In this example, a policy is defined at point B in the tenant structure that contains a set of rules. The policy set and security profiles contain that policy and every tenant-substructure at and below that point is bound to that policy. That security profile becomes operational once it is statically bound to one or more port profiles. The firewall rules that are defined within the security profile are subsequently applied to all VMs that belong to those port-profiles. VNMC's administration features can be configured to only allow specific users to only have access to change attributes within security policy B and nothing else. This allows the cloud administrator to decentralize access boundaries to tenants and subtenant personnel without compromising security of the whole infrastructure.

Policy set containing multiple policies

In some instances, it may be desirable to have a hierarchy of policies implemented. For example, one may have a general policy applied at point A within the tenant tree (which effects all VMs under it) and more specific policies defined at points C and D, which are both located lower than point A within the tenant structure. Figure 6-57 shows how this scenario may be implemented using the security building blocks explained before.

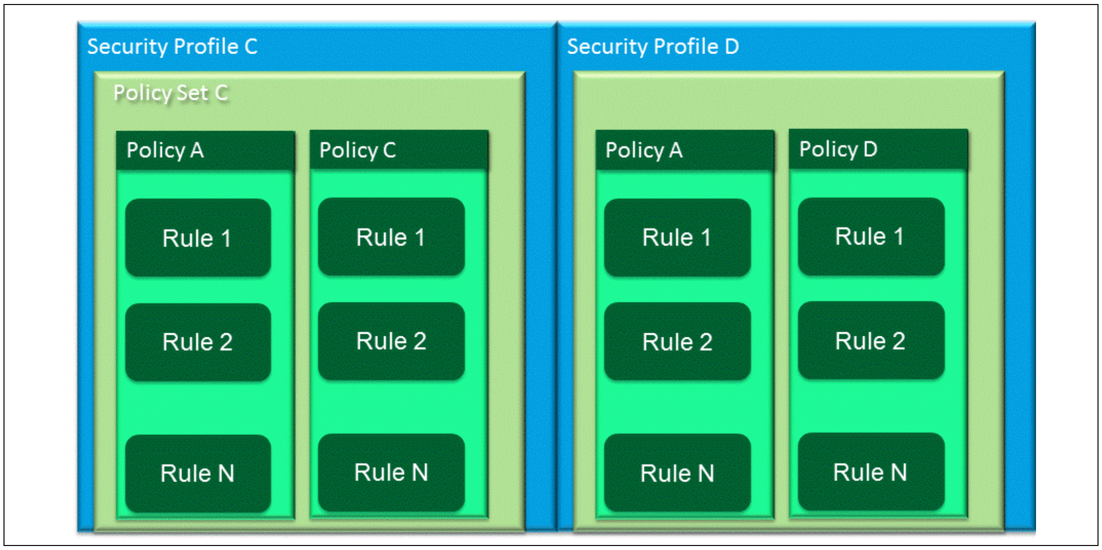


Figure 6-57 Security Multiple Policy Building Block

As shown, the security profiles defined at points C and D include the more general policy defined at point A as well as the policy defined at points C and D. Any packet entering VSG is first subjected to rules defined at policy-A, then it is subjected to the more specific rules defined at points C or D.

In summary, one can define one or more security profiles within a single VSG instance at different points within the tenant tree structure. In case of a hierarchical policy structure, a specific security profile may include several policies which are defined at different points within the tenant tree. It is recommended that more general policies be defined at points that are closer to the root. Each security profile can then be mapped to VMware's port group and Cisco Nexus 1000V port-profile.

Tenant hierarchy within VSG

The tenant construct within the VSG policy engine provides a powerful tool to overlay complex rules within a virtualized and a multi-tenant environment. Each tenant can be subdivided into three different sub-levels, which are commonly referred to as vDC, vAPP, and tier levels in the VSG documentation. Security rules and policy definitions can be set at any point in the organization. These rules apply to all VMs that reside on the “leaves” at or below the enforcement point. It is recommended that inter-tenant communication be established through routing at the aggregation layer. To provide proper tenant separation and policy control, a unique instance of VSG must be deployed for each tenant.

Figure 6-58 is an example of how VSG can be deployed within this construct. In this example, an instance of VSG is placed at the tenant level (Tenant A), but the policies are applied at two different levels within the tenant. Policy P1 is applied at the data center level, which means that the entire data center DC2 and all the sublevels within DC2 are subjected to P1 policy evaluation. Policy P2 is specific to App2 only and is placed at that organizational level. The general guideline is to have more generic policies higher in the organizational structure while more specific policies are placed closer to the organization where they are more meaningful.

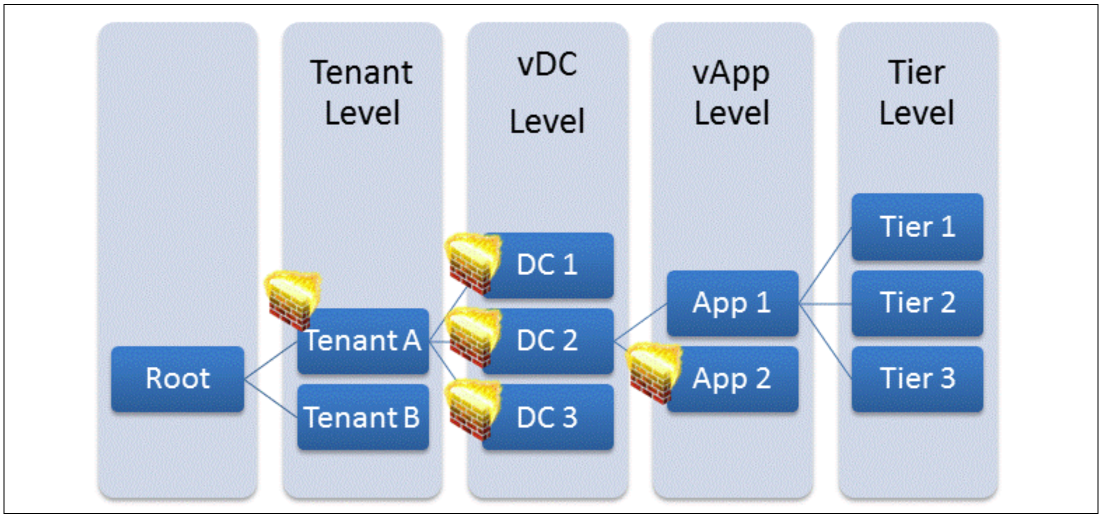


Figure 6-58 Multi-tenant structure

It is important to note that the classification system used in this example is not mandatory or predefined within VSG. The categorization of the three zone levels below the tenant definition within VSG policy framework is user-defined and not bound to these descriptions. This is important when mapping VSG into VMDC because tenancy is defined within the framework of the VMDC tenant container and not by the VSG.

The tenant as defined within the context of the VSG is equivalent to the tenant container in the VMDC Reference Architecture. Tenant containers in VMDC are typically consumed by business units or individual applications within the enterprise.

Tenant segments are flexible constructs (that is, business units, applications, or application tiers); VSG can apply policies within any of the segments and at multiple levels, that is, VSG zones. In Figure 16, we can see the tenant container model on the left and the VSG organizational hierarchy on the right. At the first level of the VSG organizational structure we have the single tenant definition for tenant 1. The first tier of the zone tree maps to the tenant segments for both shared and non-shared resources. The next two levels of the tree map to tenant sub-segments. Individual security policies are applied at any of the three zone levels and are then combined into a policy set.

VSG tenant-based role-based access

VSG administrative features allow the setting of user access based on the position in the tenant tree. For example in Figure 18, the administrator for the VDI environment can be granted access to change policies for users below the VDI position in the tenant tree. Similarly, one can set localized policy access for Tenant1-resources, Office-Users, or Lab-Users administrators that allows them to set policies for all users within their domain in the tenant-tree structure. In addition a global access can be granted to the tenant-administrator for the whole tenant at the root of the tenant tree.

Use cases

As it can be seen from the design framework discussed previously, VSG can be incorporated to satisfy a wide variety of use cases, such as these:

- Logical separation and creation of security zones based on applications: In such an environment, one can use VSG to apply security policies that isolate or restrict certain applications from a predefined set of users or network subnets. VSG's versatile rule-definition capabilities can use VM attributes, such as VM name, guest operating system, and so on, to set security rules within its policy engine.

- Logical separation and creation of security zones based on organizational internal structure: In such a use case, the VSG can be used to set security policies that incorporate and overlay on top of the organizational structure of the enterprise. As shown previously, the VSG's multi-tenant capabilities can be leveraged to provide decentralized policy control and a hierarchical security structure.
- Implementation of a multi-tenant data center: Similar to the use case before, VSG can be used to facilitate the adoption of a multi-tenant environment where a shared integrated compute stack can be logically divided to support multiple class of users or applications.
- Securing a three-tier application within a data center: VSG can be used to secure typical three-tier applications. VSG can be used to implement Web, APP, and Database zones and set policies between them.
- Secure implementation of specialized compute and application environments: VSG can be used to provide customized security policies for specialized applications, such as VDI.
- Consistent security policies across different hosts and clusters, supporting vMotion and redundant data centers: The VSG/Nexus 1000V combination provides consistent security policies when users are moved to different hosts.

For more information on deployment options and configuration details for the Cisco Nexus 1000V, Cisco Nexus 1010, and VSG, see the following documents:

- Cisco Nexus 1000V Deployment Guide:

  `http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9902/guide_c07-556626.html`

- Cisco Nexus 1010 Deployment Guide:

  `http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9902/white_paper_c07-603623.html`

- VSG Deployment Guide:

  `http://www.cisco.com/en/US/prod/collateral/modules/ps2706/ps11208/deployment_guide_c07-647435_ps9902_Products_White_Paper.html`

6.10.4 ASA 1000V Cloud Firewall

As the VSG is an intra-tenant firewall, an inter-tenant firewall is also at times required. The Cisco ASA 1000V Cloud Firewall provides consistent, enterprise-class security for private and public clouds. By employing mainstream, proven ASA security technology, the ASA 1000V secures the tenant edge in multi-tenant private and public cloud deployments. It acts as a default gateway, provides edge firewall functionalities, and secures against network based attacks. The ASA 1000V complements the Cisco Virtual Security Gateway (VSG), which provides granular inter-VM security within a tenant.

The ASA 1000V integrates with the Cisco Nexus 1000V Series Switch for unparalleled flexibility and operational efficiency:

- Efficient deployment: A single ASA 1000V instance can span and secure multiple physical servers for enhanced deployment flexibility, simplified management and optimized resource utilization. This architecture eliminates the need to deploy one virtual appliance per physical server.
- Independent capacity planning: Cisco ASA 1000V can be placed on a dedicated server controlled by the security operations team so that appropriate computing capacity can be allocated to application workloads, capacity planning can occur independently across server and security teams, and operational segregation can be maintained across security, network, and server teams.

- Scalable cloud networking: Cisco ASA 1000V acts as a VXLAN gateway to send traffic to and from the VXLAN to a traditional VLAN.
- Service chaining: vPath supports service chaining so that multiple virtual network services can be used as part of a single traffic flow. For example, by merely specifying the network policy, vPath can direct the traffic to first go through the ASA 1000V Cloud Firewall, providing tenant edge security, and then go through the Virtual Security Gateway, providing zone firewall capabilities.
- Multi-hypervisor capability: Cisco ASA 1000V does not need to be custom built for each hypervisor. Nexus 1000V Switch provides with hooks into the hypervisor. Therefore, integration with the Nexus 1000V switch enables the ASA 1000V to secure heterogeneous hypervisor environments with ease, as Nexus 1000V spans across these heterogeneous infrastructures.

The ASA 1000V also employs the Cisco Virtual Network Management Center (VNMC) for centralized, multi-device and multi-tenant policy management (Figure 6-59). VNMC provides the following key benefits:

- Rapid and scalable deployment through dynamic, template-driven policy management based on security profiles
- Easy operational management through XML APIs to help enable integration with third-party management and orchestration tools
- A non-disruptive administration model that enhances collaboration across security and server teams while maintaining administrative separation and reducing administrative errors
- Role-based access privileges for more precise control

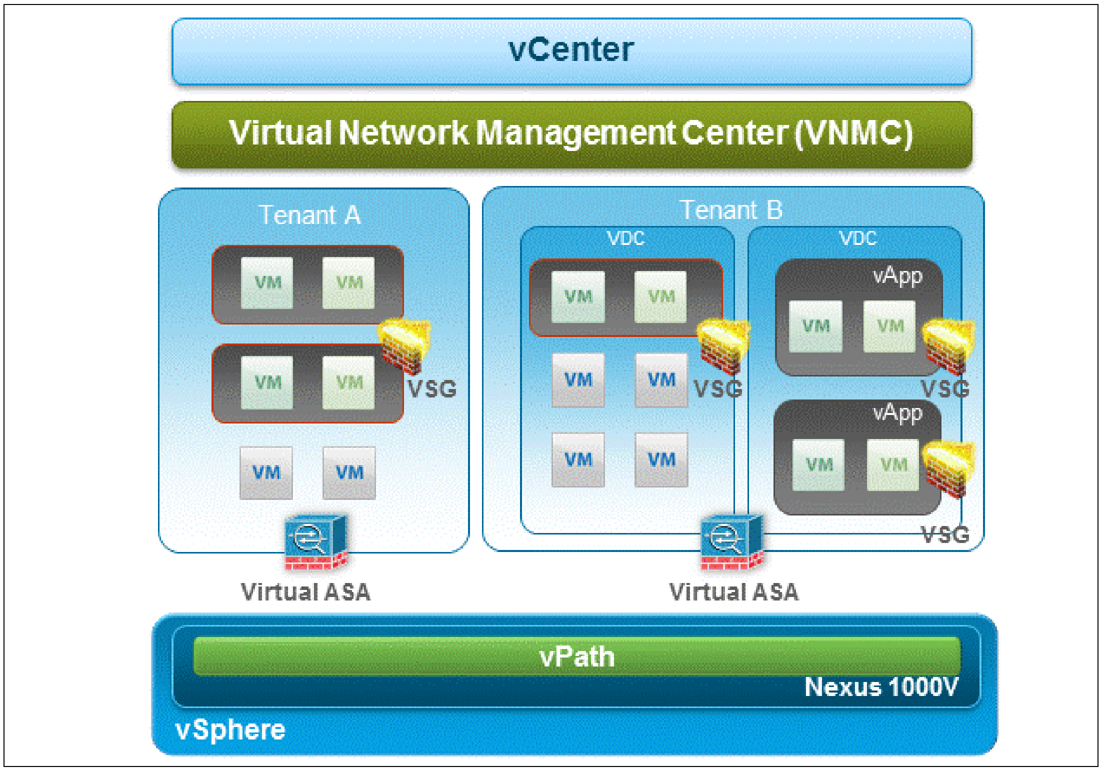


Figure 6-59 Cisco ASA 1000V Cloud Firewall Tenant Edge Security

The Cisco ASA 1000V Cloud Firewall provides strong tenant edge security as shown in Figure 6-59 on page 362, while the Virtual Security Gateway (VSG) secures traffic within a tenant (intra-tenant security for traffic between virtual machines [VMs]).

Table 6-6 shows a comparison of ASA 1000V and VSG.

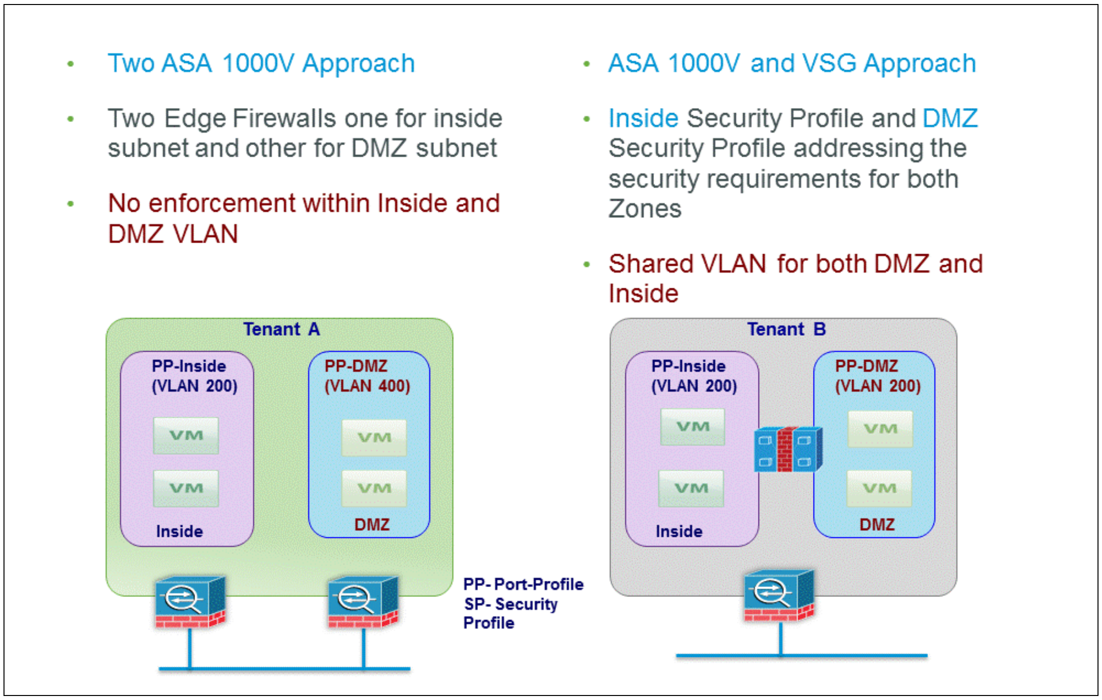
Table 6-6 Comparison of ASA 1000V and VSG

ASA 1000V Cloud Firewall (vASA)	Virtual Security Gateway (VSG)
Secures the tenant edge	Secures intra-tenant (inter-VM) traffic
Default gateway (All packets pass through the Cisco ASA 1000V)	VSG offloads traffic to the vPath offering performance acceleration; only the first packet goes through the VSG
Layer 3 firewall (for north-to-south traffic)	Layer 2 firewall (for east-to-west traffic
Edge firewall features and capabilities including network-attribute-based access control lists (ACLs), site-to-site VPN, Network Address Translation (NAT), Dynamic Host Configuration Protocol (DHCP), stateful inspections, IP audit, TCP intercept, authentication, authorization, and accounting (AAA)	Network attribute and VM-attribute based policies

While Cisco ASA 1000V and Cisco VSG provide complementary functionality, they are separate products that can be run independently from one another. Based on the customer environment, the ASA 1000V and VSG can coexist to provide end-to-end (intra-tenant and tenant edge) security for private and public clouds. However, Cisco VSG is not a required component of the Cisco ASA 1000V solution.

A use case worth noting is using the ASA and VSG in a virtualized DMZ.

In Figure 6-60 there are two different design scenarios. Option 1 uses separate vASA and VLANs to provide segmentation. Option 2 uses a single vASA for perimeter security and single VLAN and VSG to provide secure L2 isolation within the VLAN.



- Two ASA 1000V Approach
- Two Edge Firewalls one for inside subnet and other for DMZ subnet
- No enforcement within Inside and DMZ VLAN

- ASA 1000V and VSG Approach
- Inside Security Profile and DMZ Security Profile addressing the security requirements for both Zones
- Shared VLAN for both DMZ and Inside

Figure 6-60 ASA 1000V DMZ use case

Virtual Network Management Center (VNMC)

Cisco Virtual Network Management Center (VNMC) is a multi-tenant-capable, multi-device, policy-driven management solution for Cisco virtual security services (such as Cisco ASA 1000V Cloud Firewall and Cisco Virtual Security Gateway [VSG]) to provide end-to-end security of virtual and cloud infrastructures. It has been custom-created for the virtualization-specific workflows. VNMC delivers the following benefits:

- Helps enable rapid and scalable deployment through dynamic, template-driven policy management based on security profiles
- Integrates with VMware vCenter to use VM attributes (required for policies based on VM attributes)
- Enhances management flexibility through an XML API that helps enable programmatic integration with third-party management and orchestration tools
- Helps ensure collaborative governance with role-relevant management interfaces for network, server, and security administrators

Figure 6-61 shows the VC virtual services architecture.

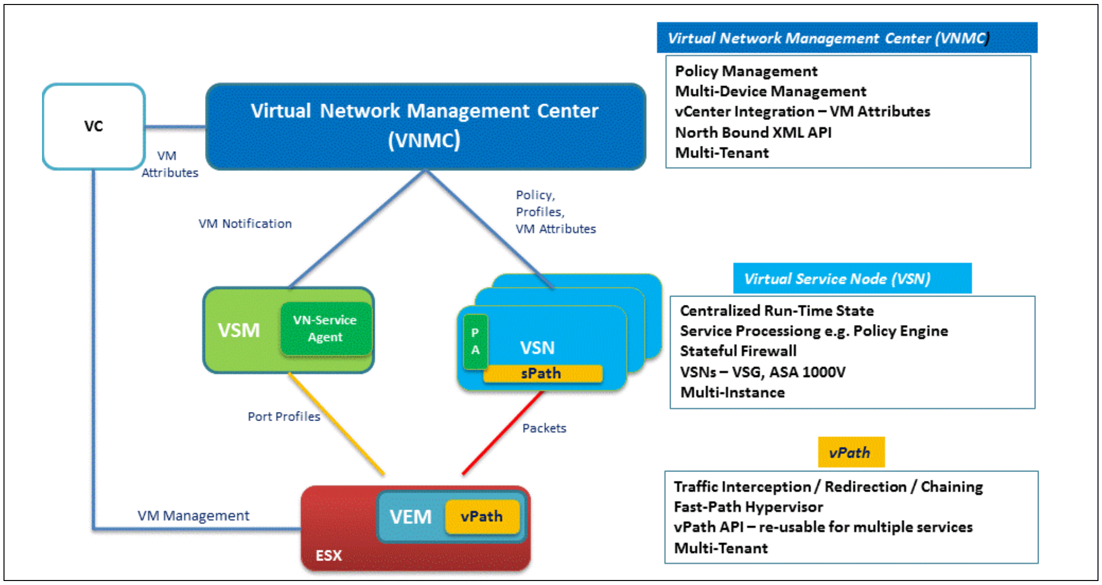

VC
Virtual Network Management Center (VNMC)
VM Attributes
VM Notification
Policy, Profiles, VM Attributes
VSM
VN-Service Agent
P A
VSN
sPath
Port Profiles
Packets
VM Management
VEM
vPath
ESX
Virtual Network Management Center (VNMC)
Policy Management
Multi-Device Management
vCenter Integration – VM Attributes
North Bound XML API
Multi-Tenant
Virtual Service Node (VSN)
Centralized Run-Time State
Service Processing e.g. Policy Engine
Stateful Firewall
VSNs – VSG, ASA 1000V
Multi-Instance
vPath
Traffic Interception / Redirection / Chaining
Fast-Path Hypervisor
vPath API – re-usable for multiple services
Multi-Tenant

Figure 6-61 VC Virtual Services Architecture

Consistent, efficient execution of security policies

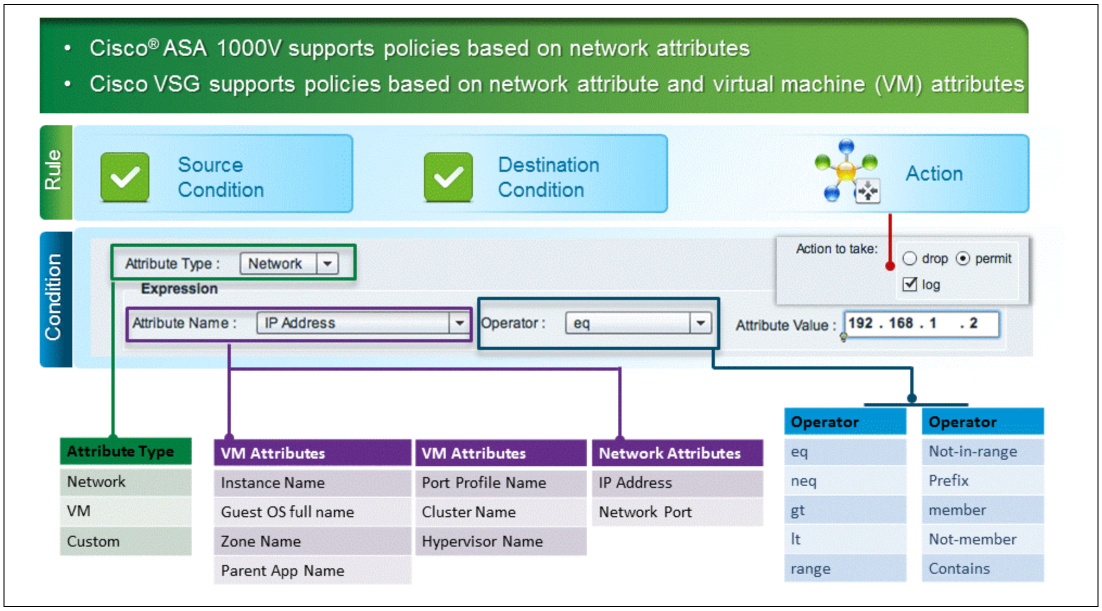
Cisco VNMC uses security profiles for template-based configuration of security policies. A security profile is a collection of security policies that can be predefined and applied on demand at the time of virtual machine instantiation. This profile-based approach significantly simplifies authoring, deployment, and management of security policies, including in a dense multi-tenant environment, while enhancing deployment agility and scaling. Security profiles also help reduce administrative errors and simplify audits.

The XML API for Cisco VNMC facilitates integration with northbound network provisioning tools for programmatic network provisioning and management of Cisco VSG and Cisco ASA 1000V. The option of programmatic control of Cisco VSG and Cisco ASA1000V can greatly simplify operation processes and reduce infrastructure management costs.

Non-disruptive administration model

By providing visual and programmatic controls, Cisco VNMC can enable the security operations team to author and manage security policies for virtualized infrastructure and enhance collaboration with the server and network operations teams. This non-disruptive administration model helps ensure administrative segregation of duties to reduce administrative errors and simplify regulatory compliance and auditing. Cisco VNMC operates in conjunction with the Cisco Nexus 1000V Series Virtual Supervisor Module (VSM) to achieve the following workflow as shown in Figure 6-62:

- Security administrators can author and manage security profiles and manage VSG and ASA 1000V instances. Security profiles are referenced in Cisco Nexus 1000V Series port profiles.
- Network administrators can author and manage port profiles, as well as manage Cisco Nexus 1000V Series distributed virtual switches. Port profiles with referenced security profiles are available in VMware vCenter through the Cisco Nexus 1000V Series VSM's programmatic interface with VMware vCenter.
- Server administrators can select an appropriate port profile in VMware vCenter when instantiating a virtual machine.



- Cisco® ASA 1000V supports policies based on network attributes
- Cisco VSG supports policies based on network attribute and virtual machine (VM) attributes

Attribute Type : Network

Attribute Name : IP Address

Operator : eq

Attribute Value : 192 . 168 . 1 . 2

Action to take: ○ drop ● permit ☑ log

Attribute Type
Network
VM
Custom

VM Attributes	VM Attributes	Network Attributes
Instance Name	Port Profile Name	IP Address
Guest OS full name	Cluster Name	Network Port
Zone Name	Hypervisor Name	
Parent App Name		

Operator	Operator
eq	Not-in-range
neq	Prefix
gt	member
lt	Not-member
range	Contains

Figure 6-62 Cisco 1000V: Logical way of authoring policies

References

Data center management:

https://supportforums.cisco.com/community/netpro/data-center/data-center-management

Cisco ASA 1000V: A Firewall for the Cloud:

http://www.youtube.com/watch?v=hN9tqvaxPkM

Cisco ASA 1000V: Cloud Firewall and Virtual Security:

http://www.youtube.com/watch?v=AIDOIu5mH7A

Cisco ASA 1000V: Configuration Tips and Tricks:

http://www.youtube.com/watch?v=5Vwo6n5tXao

http://www.cisco.com/en/US/prod/collateral/vpndevc/ps6032/ps6094/ps12233/data_sheet_c78-687960.html

6.10.5 Virtual NAM: Network Analysis Module

From a support perspective, it is critical to have the ability to monitor traffic flows for security, analytical or billing purposes within the data center. The NAM and now vNAM, provides this functionality for both physical and virtual hosts.

A new version of the Cisco Network Analysis Module (NAM), Cisco Prime NAM, is offered in a virtual appliance form factor. Cisco Prime NAM offers traffic and performance analysis capabilities that empower network administrators to quickly troubleshoot performance issues and ensure optimal use of network resources in the VM network. Integrated with the Cisco Nexus 1010 Virtual Services Appliance, Cisco vNAM reduces network footprint, eases deployment, and lowers administrative cost.

With Cisco Prime NAM, the following actions can be performed:

- Analyze network usage behavior by application, host/VM, and conversation to identify bottlenecks that may impact performance and availability
- Troubleshoot performance issues with extended visibility into VM-to-VM traffic, virtual interface statistics, and application response times
- Improve the efficiency of the virtual infrastructure with deeper operational insight

Key features of Cisco Prime NAM include these:

- Workflow-oriented user experience improves operational efficiency and user productivity.
- Historical analysis helps tackle unanticipated performance issues.
- Traffic analysis provides multifaceted insight into network behavior.
- Packet Capture, Decode, Filters, and Error scan expedite root-cause analysis.
- Standards-based API preserves investments in existing management assets.

In conjunction with these virtual services offerings, a dedicated platform has been introduced to support the deployment of the service virtual appliances on a VMware vSphere ESXi host. The Cisco Nexus 1010 Virtual Services Appliance offers a dedicated hardware platform for the deployment of services critical to virtualization infrastructure.

Here are some benefits of such a platform:

- Placing management and control path elements, such as the Cisco Nexus 1000V Virtual Supervisor Module (VSM), on the Cisco Nexus 1010 allows managing policies separate from VMware virtualization administrators, which makes it easier to attain compliance and audit requirements and reduces administrative errors.
- Offloading VSM to a dedicated appliance delivers scalability and performance improvements to the virtualized data center infrastructure.
- A virtual services platform close to, but not resident within, the virtualization infrastructure permits a VM-aware solution, such as the Cisco Network Analysis Module, to gain accurate network statistics directly from data sources, including virtual ports.

For more information on deployment options and configuration details for the Cisco virtual NAM, refer to the following document:

http://www.cisco.com/en/US/products/ps10846/index.html

Other references:

http://lippisreport.com/2011/01/lippis-report-164-cisco-builds-a-modern-network-service-layer-for-virtualized-and-cloud-infrastructure/

http://www.cisco.com/en/US/prod/collateral/modules/ps2706/ps11208/data_sheet_c78-618245.html

http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9902/white_paper_c11-713736.html

7

Convergence of LAN and SAN: Fibre Channel over Ethernet

In this chapter, we describe the possibilities and benefits of converged networking. The LAN and SAN environments have existed independently for a long time, but in recent years we have seen significant cost and performance benefits by mixing these two types of networks together on a single infrastructure. We describe the architecture, functionality, and benefits of using Fibre Channel over Ethernet (FCoE) to converge the LAN and SAN together.

It is beyond the scope of this chapter to cover attaching BladeCenters or IBM PureFlex and IBM Flex systems to the network. See Chapter 2, “IBM and Cisco building blocks for the data center” on page 29 and Chapter 4, “Data center physical access layer evolution” on page 191 for reference information on BladeCenter or Pure Flex, or consult the website:

http://www.ibm.com

We explore topics such as these:

- Converged: What does that really mean?
- How is FCoE different from FC and iSCSI?
- Data Center Bridging
- Business drivers for FCoE

7.1 Converged networking in data centers

Data centers typically run multiple separate networks, including an Ethernet network (LAN) for client to server and server to server communications and a Fibre Channel Storage Area Network (SAN), which is a fiber connection from server to storage and storage to storage. To support various types of networks, data centers use separate redundancy mechanisms for each network: Ethernet network interface cards (NICs) and Fibre Channel interfaces (HBAs) in the servers, and redundant pairs of switches at each layer in the network architecture. Use of parallel infrastructures increases capital costs, makes data center management more difficult, and diminishes business flexibility. However converged networking enables us to merge these two traffic streams using a single physical connection.

Typical servers in data centers have multiple adapters to control their I/O traffic. By selecting a unified I/O adaptor that supports all the requirements of a data center, applications can reduce the number of network/storage devices and cables used to interconnect them. Unified I/O enables significant reductions in data center power requirements and cabling. The name FCoE is an acronym for Fibre Channel over Ethernet. FCoE allows the Fibre Channel frame to be transmitted using enhanced Ethernet protocols.

To support the convergence of FC traffic onto Ethernet, a series of new standards were created to “enhance” the current Ethernet so it becomes usable for FC frames to be transported via LAN infrastructure.

FCoE is defined in the Fibre Channel Backbone Generation 5 (FC-BB-5) standard, completed by the INCITS T11 Technical Committee in June 2009 and published by ANSI in May 2010. FC-BB-5 is part of the Fibre Channel family of standards and specifies how to transport Fibre Channel over other network technologies. FC-BB-5 defines the FCoE frame format, addressing structures, forwarding behavior, discovery protocols, and requirements for deployment.

Figure 7-1 shows what standards are combined under FCoE.

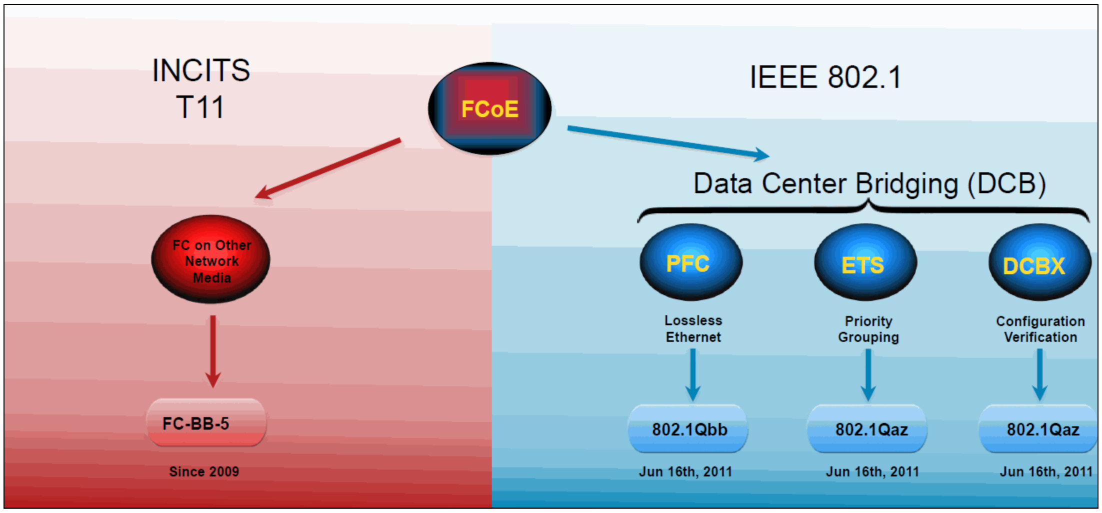


Figure 7-1 FCoE standards

There are several ways for us to connect our storage devices, and the most common are Fibre Channel (FC), Internet Small Computer System Interface (iSCSI) and Network Attached Storage (NAS). Figure 7-2 gives an overall picture of the most commonly used storage connections using advanced and normal Ethernet connection to communicate.

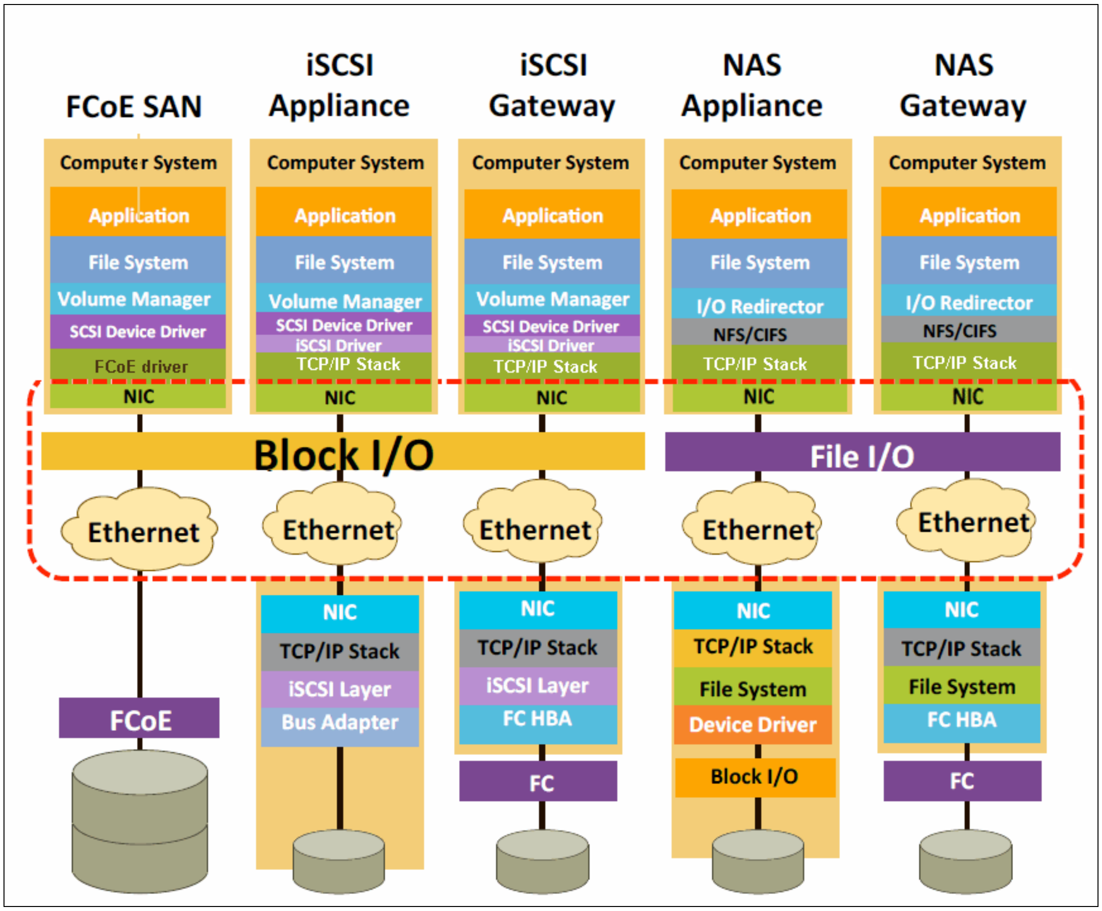


Figure 7-2 Most common methods of connecting to storage using enhanced or normal Ethernet

7.2 Business benefits and customer drivers for Unified Fabric

Traditionally organizations have been running separate, parallel networks in their data center: an Ethernet-based LAN to connect servers, clients, and the broader Internet, and a storage area network (SAN) based on Fibre Channel to connect servers to the storage pool (disks and tapes), as illustrated in Figure 7-3.

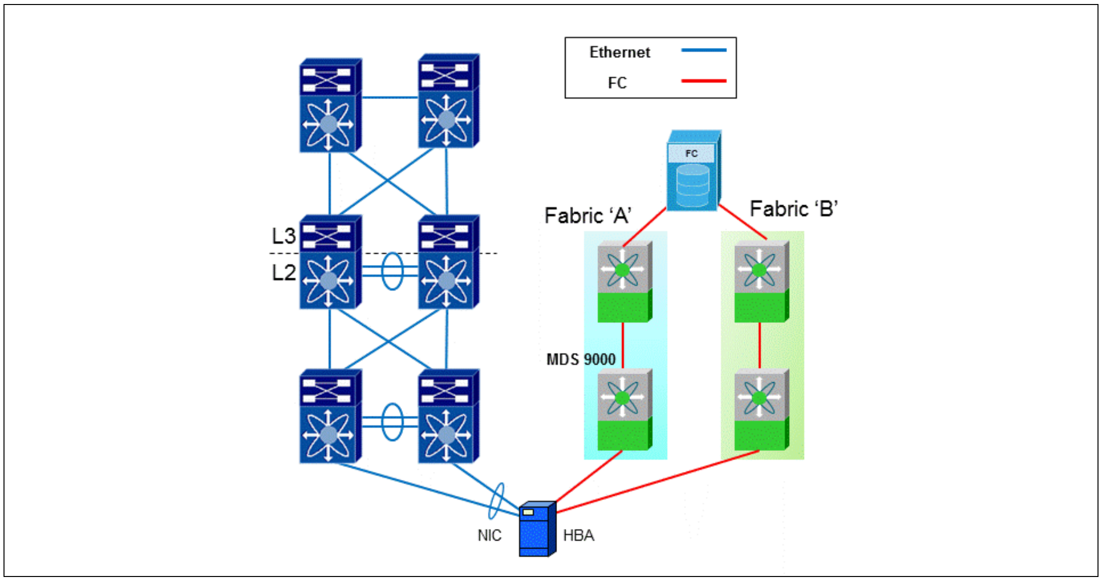


Figure 7-3 Traditional data center design

Storage and LAN networking are both critical networks of any data center. Keeping them up and running, eliminating performance barriers, and enabling smooth expansions and refreshes are top priorities for all corporations. To this end, there must be significant benefit to any major changes that outweigh the potential risks and disruption.

This section covers the benefits of network convergence using Fibre Channel over Ethernet (FCoE). With FCoE, companies can collapse the LAN and SAN into a single common network / infrastructure and provides the following benefits:

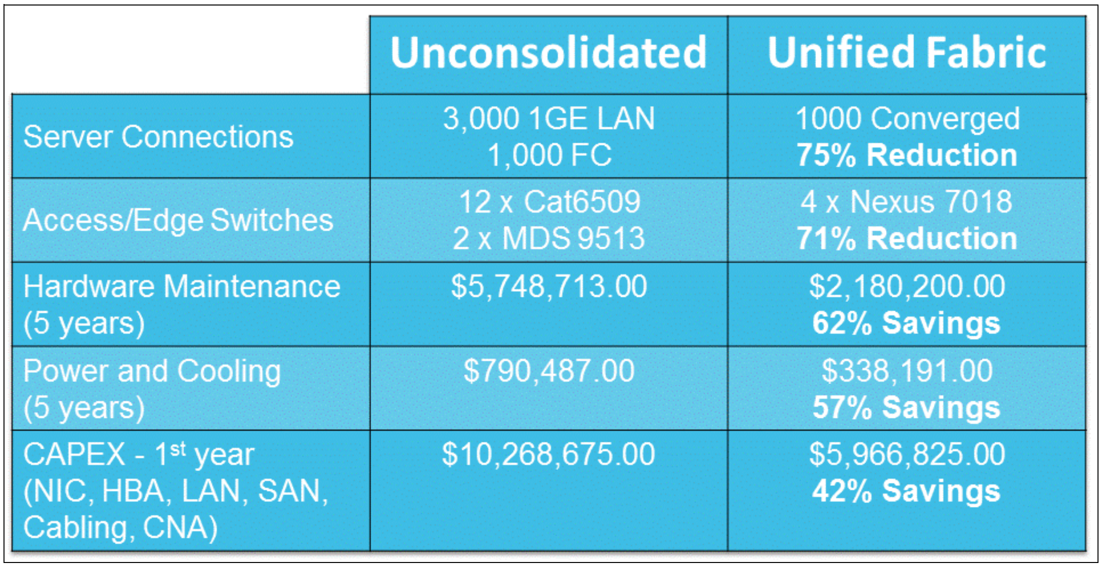
- Capital expenditures (CAPEX) savings by eliminating redundant switches, cables, networking cards and adapters.
- Operating costs (OPEX) reduction by simplifying administration and reducing the number of managed devices. Also reduction on hardware maintenance cost, power and cooling requirements.
- Improved Performance by using higher performance, higher density Ethernet switches. While FCoE is supported over 10 Gbps Ethernet and 40 Gbps Ethernet, Fibre Channel currently supports up to 16Gbps with the vast majority of implementations still based on 4 and 8 Gbps links.
- Investment Protection by integrating with existing Fibre Channel SANs, as demonstrated on the design section.

7.2.1 Real-world comparison

The following section presents a real-world example that shows the potential benefits of using FCoE. This example is based in a data center with:

- 500 Servers
- Power Usage Effectiveness (PUE) of 2.5
- Cost of power per kWhr $0.10
- 5-year period
- Annual growth rate of servers 10%

As seen in Figure 7-4, which illustrates cost comparison between using separated, isolated networks as opposed to a converged fabric.



	Unconsolidated	Unified Fabric
Server Connections	3,000 1GE LAN 1,000 FC	1000 Converged **75% Reduction**
Access/Edge Switches	12 x Cat6509 2 x MDS 9513	4 x Nexus 7018 **71% Reduction**
Hardware Maintenance (5 years)	$5,748,713.00	$2,180,200.00 **62% Savings**
Power and Cooling (5 years)	$790,487.00	$338,191.00 **57% Savings**
CAPEX - 1st year (NIC, HBA, LAN, SAN, Cabling, CNA)	$10,268,675.00	$5,966,825.00 **42% Savings**

Figure 7-4 Cost comparison

In an unconsolidated solution each server needs to be equipped with six 1 GE connections, including interfaces for production traffic, installation, administration and backup/recovery. In addition to that each server requires two Fibre Channel Host Bus Adapters (HBA's), one connected to Fibre Channel Fabric A and the other to Fabric B. When compared with the converged approach, where each server has only two 10 Gbps connections that transport Ethernet and Fibre Channel over Ethernet (FCoE) traffic, this represents a 75% reduction on the number of server connections required on an unconsolidated environment.

The number of access switches required to build this topology is also reduced by utilizing the Nexus 7018 which supports a much higher number of 10 GE ports when compared with the Cisco Catalyst 6509.

The reduction on the number of devices also reduces the cost of the services and support contracts required for the hardware. It also reduces the power and cooling requirements significantly.

The initial hardware cost (CAPEX) is reduced as a result of the decrease in the number of network interface cards and HBA's required on the servers. Savings also comes from the reduced number of Ethernet (LAN) ports, Fibre Channel (SAN) ports, and switches required.

The table in Figure 7-4 provides the details of the savings achieved by utilizing a converged architecture.

The following white papers discuss in more detail the business benefits of using Fibre Channel of Ethernet (FCoE) to converge LAN and SAN:

- Forrester Research: Exploring the Potential Benefits of End-to-End Convergence of Data Center Networks:

  http://www.cisco.com/en/US/prod/collateral/modules/ps5991/white_paper_Forrester_Exploring_the_Potential_of_DC_Networks.pdf

- DC: The ROI of Converged Networking Using Unified Fabric:

  http://www.cisco.com/en/US/solutions/collateral/ns340/ns517/ns224/ns945/ns1060/white_paper_ROI_Converged_NW.pdf

- Storage I/O: Business and Technology Benefits of Converged I/O Networking Infrastructures:

  http://www.cisco.com/en/US/solutions/collateral/ns340/ns517/ns224/ns945/ns1060/SIO_IndustryPerspective_Converged_Mar08_2011.pdf

- Energy-Efficient Unified Fabrics: Transform the Data Center Infrastructure with Cisco Nexus Series:

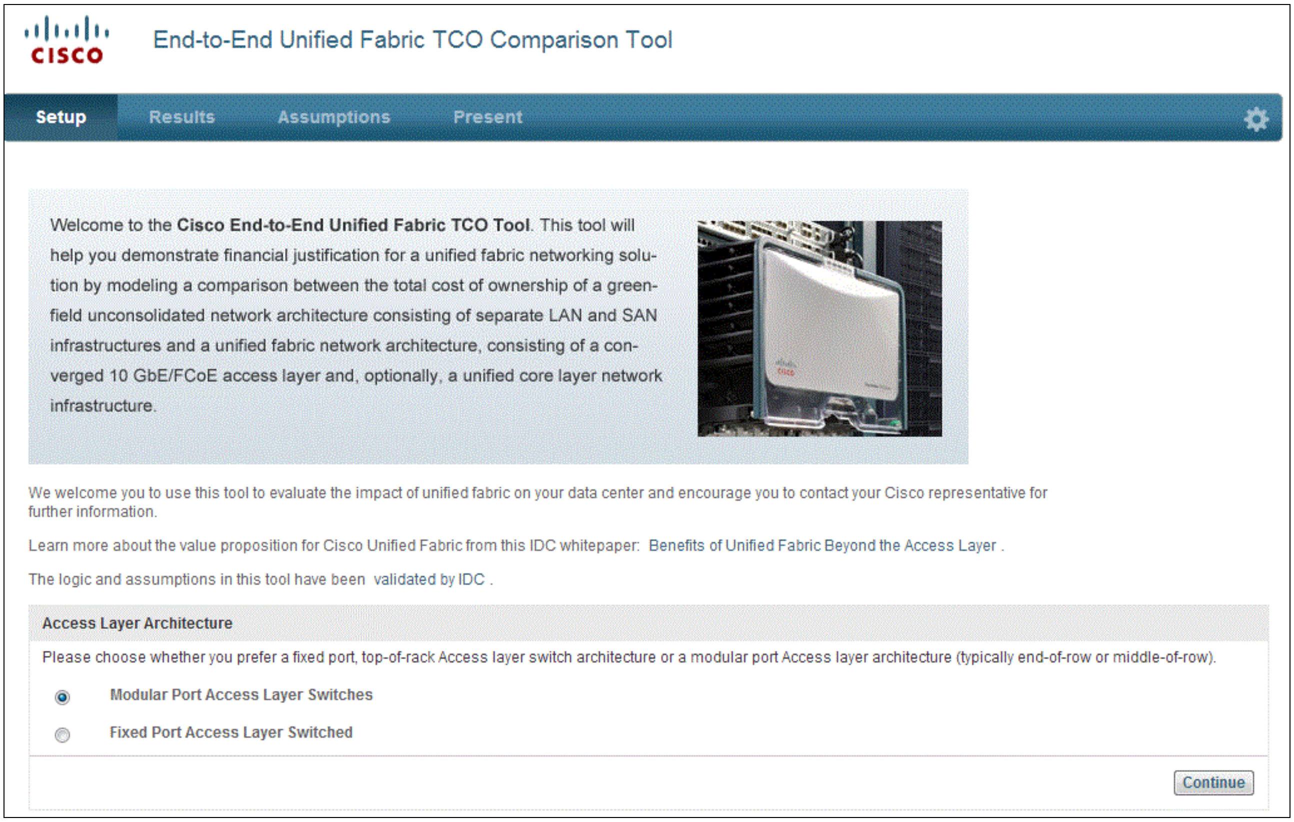
  http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9670/white_paper_c11-508518_ps9670_Products_White_Paper.html

7.2.2 End-to-End Unified Fabric TCO Calculator

The Cisco End-to-End Unified Fabric TCO Comparison Tool in Figure 7-5 provides an easy way to evaluate the impact of unified fabric on a data center.

The tool compares the total cost of ownership (TCO) of two network architectures: a traditional network architecture with separate LAN and SAN infrastructure at the access and core layers, and a unified fabric architecture with a converged access and core layer.

By entering the number of servers, annual growth rate of servers, number of interfaces per server, and oversubscription ratio, the tool will provide a comparison between both architectures.



Welcome to the **Cisco End-to-End Unified Fabric TCO Tool**. This tool will help you demonstrate financial justification for a unified fabric networking solution by modeling a comparison between the total cost of ownership of a greenfield unconsolidated network architecture consisting of separate LAN and SAN infrastructures and a unified fabric network architecture, consisting of a converged 10 GbE/FCoE access layer and, optionally, a unified core layer network infrastructure.

We welcome you to use this tool to evaluate the impact of unified fabric on your data center and encourage you to contact your Cisco representative for further information.

Learn more about the value proposition for Cisco Unified Fabric from this IDC whitepaper: Benefits of Unified Fabric Beyond the Access Layer .

The logic and assumptions in this tool have been validated by IDC .

Access Layer Architecture

Please choose whether you prefer a fixed port, top-of-rack Access layer switch architecture or a modular port Access layer architecture (typically end-of-row or middle-of-row).

● Modular Port Access Layer Switches

○ Fixed Port Access Layer Switched

Figure 7-5 Cisco End-to-End Unified Fabric TCO Calculator

The End-to-End Unified Fabric TCO Calculator Tool can be accessed at this website:

`https://express.salire.com/go/cisco/end-to-end-unified-fabric-tco-comparison-tool.aspx`

7.3 But what is FCoE all about, is it SAN or LAN?

SAN architectures use the SCSI protocol for sending and receiving storage data over their own network, often called the Fibre Channel network, FCP SAN, or just the SAN. What the SAN usually does is to implement the SCSI protocol within the FC frames on a Fibre Channel network. iSCSI SANs implement the same SCSI protocol within TCP/IP packets in an Ethernet network.

FCoE differs from iSCSI in that instead of using TCP/IP, it simply encapsulates the native FC protocol within Layer 2 Ethernet packets, and Data Center Bridging (DCB) is used on the underlying Ethernet network to provide a lossless fabric for FCoE traffic to co-exist with the normal network traffic, even under congestion situations in the network.

Figure 7-6 shows the storage transport layers utilized by each storage protocol. It also shows the same storage layers as in Figure 7-2 on page 371, in a more simplified way, and adds InfiniBand as well.

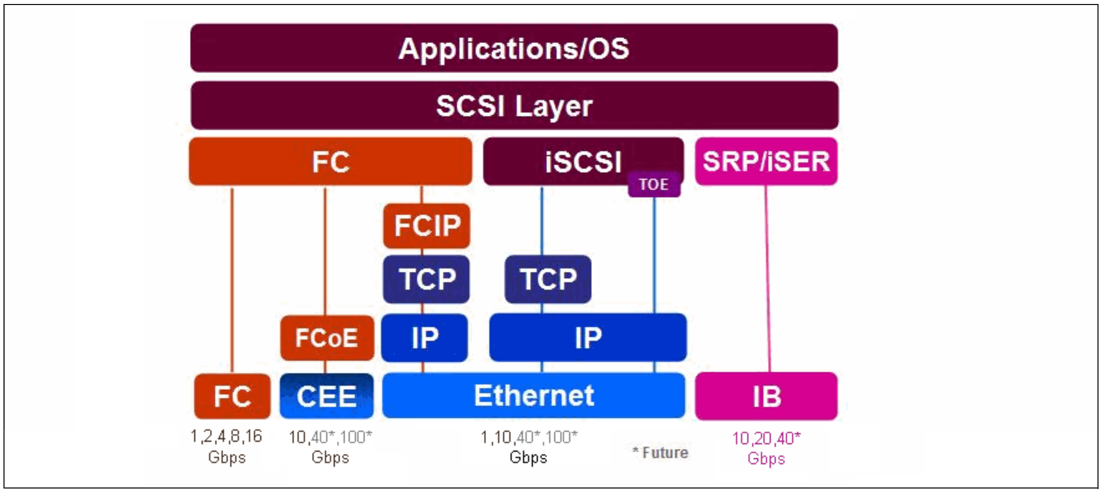


Figure 7-6 Storage transport layers

And if we take a look at the OSI stack, we can clearly see where we mix the FC stack and the OSI stack to create room for FCoE to work as shown in Figure 7-7.

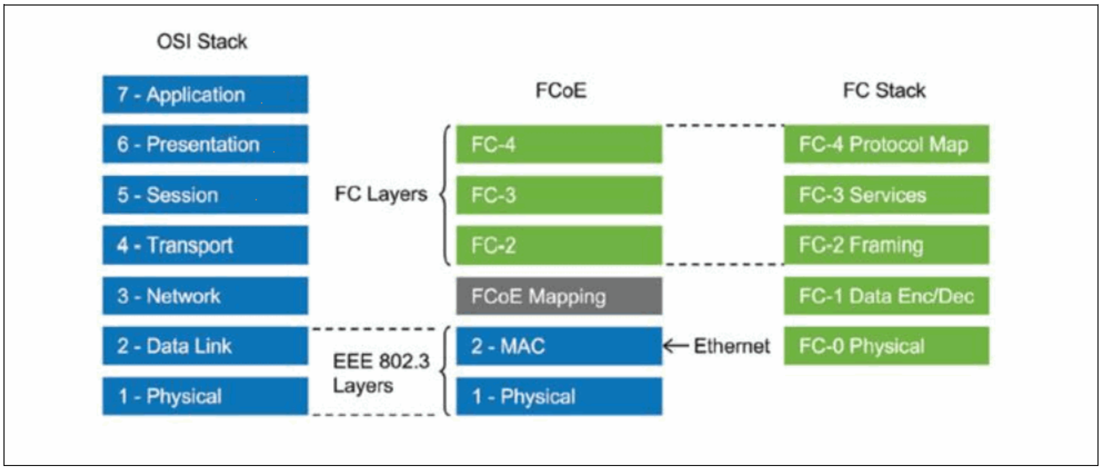


Figure 7-7 OSI stack

The area where FCoE differs from iSCSI is where FCoE replaces the TCP/IP layer used in iSCSI, and instead relies on the enhancements in the Ethernet layer as we describe in 7.4, “Fibre Channel over Ethernet architecture” on page 378.

Before we deep dive into FCoE, we provide a high level view of how FCoE works using an analogy. We hope this will make it easier to understand what we discuss later in this chapter where we go into deeper details of the functionality.

7.3.1 FCoE high level overview

If we imagine we have a railroad company with multiple divisions, and one of them is called *Division A* and they only deliver a *one train-carriage* called *FCP* at any given time. This carriage is filled with an FC payload and its destination address is known before the journey starts. The FCP carriage can only run on special rails (FC) designed for FCP carriages and these rails are quite expensive and mainly used by medium and large size companies (our SAN).

This division has been doing this for many years and are in control of their work process. They use processes, such as flow control and class of service, to ensure timely delivery of all their packages, and they have been awarded the lossless (never to lose a package) award many years in a row.

Division A is our FC standard.

There is also another division called *Division B* in the same railroad company that has been around for many years. They can carry *multiple carriages*, called *IP*. Their carriages use well known rails (LAN) and are used by every household, and company, worldwide. They have a reputation for losing packages, but their re-delivery (retransmission) is very fast. They have recently *enhanced* their delivery system and can now offer lossless transport using carriages they call *Jumbo* in combination with some newly implemented enhanced standards.

Division B is our Ethernet standard.

Since Division B is doing so well, we want to use their IP carriages to deliver our FC payload and FCP carriages. Although the rails are not natively suited for FCP, they have come up with a solution to support it. We simply put the FCP carriage into the belly of the Jumbo. Remember, Jumbo uses IP rails. It is even possible for the Jumbo to take several FCP carriages using the same ride.

This simple analogy demonstrates how the FC frames can fit into the Jumbo Ethernet frame.

Now we need to know how the IP division knows to what address the FC payload is to be delivered. This is done by using the Data Center Bridging Xchange (DCBX) standard.

It is clear that normally when we use FCP, we communicate the source and destination address to FC using normal FC standards, but in FCoE, we are actually using the Ethernet addressing scheme and therefore IP will deliver our FC frame using Ethernet addressing. The FC frame is totally encapsulated, address and all, in the IP carriage, and therefore the FC address cannot be used for this transport.

However, our devices are now introduced to the FCoE world using a new initialization mode. It is similar to the addressing scheme used by FC. With FC we are used to a process where we receive an FC-ID, but when using FCoE, we receive something called an FC-MAC address, which is a combination of FC-ID and an FC-MAP address. When the IP side has received the MAC address, using DCBX, they can deliver the FC carriage using the same fashion as they would normally do with all their carriages.

But what happens if the IP side gets stuck in traffic or arrives late; what happens then?

It is a fact that the FCP standard is more effective than Ethernet when it comes to lossless delivery and by using FCoE we are putting all our faith in Ethernet delivering our FC package without losing it. To help this, we have utilized some of the working methods and processes that FC uses, and created an *enhanced* Ethernet. DCBX is actually designed to make Ethernet operate more like Fibre Channel by creating a special purpose traffic class that emulates the behavior of a SAN fabric. By incorporating the enhanced Ethernet standards (PFC, ETS, DCBX), we have our FC payload reliably delivered to our destination.

The fact that the Fibre Channel payload remains intact throughout its FCoE journey is because FCoE preserves the Fibre Channel constructs and services, and enables FCoE solutions to utilize existing management applications. As a result, FCoE solutions are designed to integrate seamlessly into existing Fibre Channel environments without introducing incompatibilities or disrupting existing infrastructures.

We expand on this topic in 7.4, “Fibre Channel over Ethernet architecture”.

7.4 Fibre Channel over Ethernet architecture

Given the significant number of protocol options and the possibilities that FCoE introduces to both new and older designs, it is quite easy to understand that architects are looking carefully at this in their designs. One of the big advantages in moving to converged networking is that it does not require an “all or nothing” decision. This allows for users to try small implementations before selecting one technology over the other. Keep in mind that when we look at the basic functionality of FC, (to move SCSI data), application and or performance requirements must also be considered. SCSI is very sensitive if frames are dropped or if latency is high or variable, so lossless delivery and proper storage network design are critical in any storage architecture, regardless of which technology is utilized.

Figure 7-8 shows the encapsulation of FC frame.

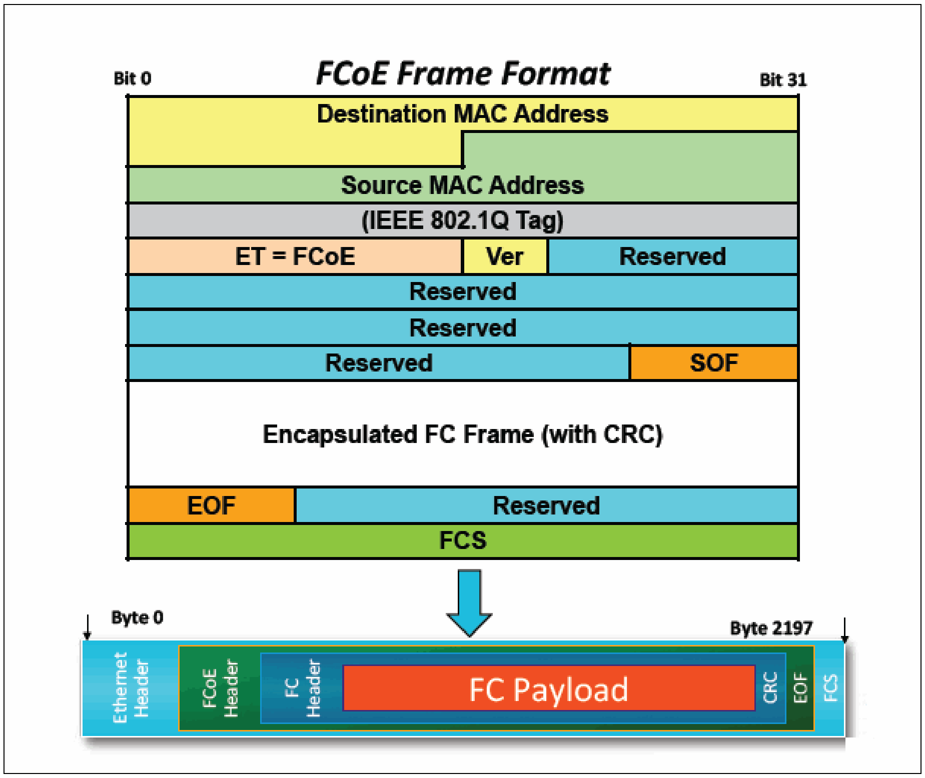


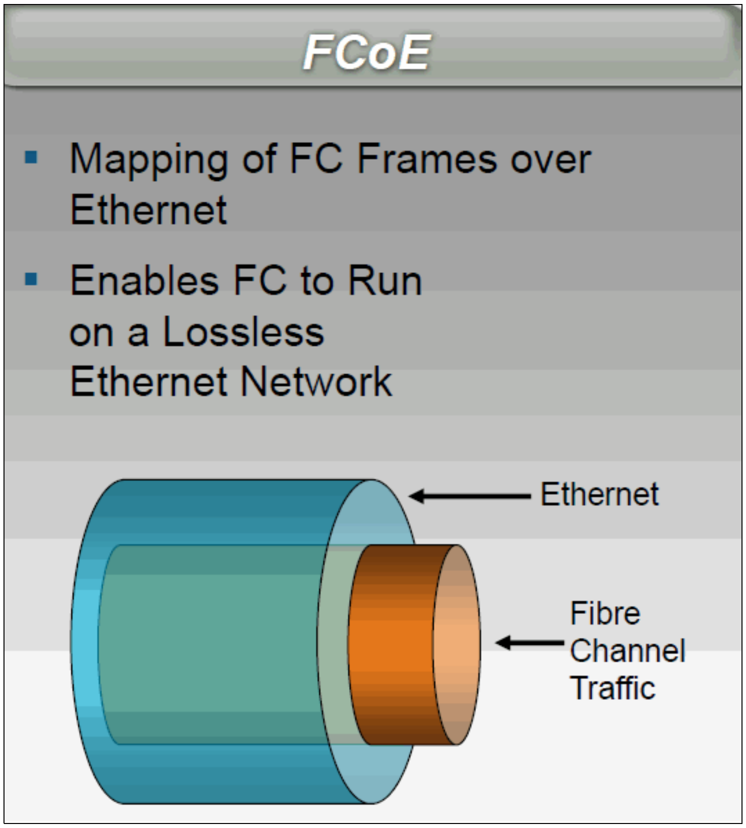
Figure 7-8 FCoE frame

FCoE is encapsulated over Ethernet with the use of a dedicated Ethertype, 0x8906. A single 4-bit field (ver) satisfies the IEEE sub-type requirements. The SOF (start of frame) and EOF (end of frame) are encoded as specified in the FCoE standard. Reserved bits are present to guarantee that the FCoE frame meets the minimum length requirement of Ethernet. Inside the encapsulated Fibre Channel frame, the frame header is retained so as to allow connecting to a storage network by passing on the Fibre Channel frame directly after de-encapsulation. FCoE also requires a new larger Ethernet frame size called “baby jumbo frame” that accommodates the Fibre Channel frame and the associated Ethernet and FCoE headers. These standards can be seen in Figure 7-1 on page 370.

When looking at the FCoE protocol stack, then it is simply FC upper services (Layers FC 2, FC 3, and FC 4) placed on top of Ethernet physical and Data Link layers (Layer 2). The Data Center Bridging (DCB) layer is the Ethernet Link Layer (Layer 2) where DCB enhancements makes Ethernet capable of transporting FC frames. Between the FC and Ethernet layers is the FCoE layer.

FCoE maps Fibre Channel directly over Ethernet while being independent of the Ethernet forwarding scheme. The FCoE protocol specification replaces the FC0 and FC1 layers of the Fibre Channel stack with Ethernet. By retaining the native Fibre Channel constructs, FCoE was meant to integrate with existing Fibre Channel networks and management software.

Figure 7-9 illustrates how Fibre Channel traffic and Ethernet traffic run on the same physical cable.



FCoE

- Mapping of FC Frames over Ethernet
- Enables FC to Run on a Lossless Ethernet Network

Figure 7-9 Ethernet and Fibre Channel in same cable

7.4.1 Key terminology

T11 FC-BB-5 FCoE defines two types of endpoints for the Ethernet encapsulation of Fibre Channel frames: FCoE nodes (ENodes) and FCoE forwarders (FCFs). ENodes are the combination of FCoE termination functions and Fibre Channel stack on the CNAs, and in that sense they are equivalent to host bus adapters (HBAs) in native Fibre Channel networks. FCFs (Fibre Channel Forwarders) are the combination of FCoE termination functions and Fibre Channel stack on Ethernet switches (dual-stack switches) and are therefore equivalent to Fibre Channel switches in native Fibre Channel networks.

ENodes present virtual FC interfaces in the form of VN_Ports, which can establish FCoE virtual links with FCFs' VF_Ports. FCFs present virtual FC interfaces in the forms of VF_Ports or VE_Ports; a VF_Port establishes FCoE virtual links with a CNA's VN_Port, and VE_Ports enable FCFs to establish FCoE virtual links with one other. These interface types have their equivalents in native Fibre Channel's N_Ports, F_Ports, and E_Ports.

7.4.2 Fibre Channel

In a Fibre Channel network, the flow control is a mechanism called buffer to buffer credits, or sometimes as B2B (not to be confused with business-to-business) that is a credit-based flow-control used to control the flow and provide lossless behavior on the Fibre Channel link.

At a high level, each FC device knows the amount of buffer spaces available on the next hop device based on the agreed upon frame size (typically 2148 bytes.) The receiver gives credit to the transmitter, specifying the amount of buffering available for reception of Fibre Channel frames. The transmitter is therefore bound by the number of credits available to it when transmitting frames. Credit-based flow control ensures that the receiver always has enough buffers to process the incoming frames and avoid frame loss in a SAN. This means that a device will never send a frame to a next hop device that does not have a buffer available.

Fibre Channel requires three primary extensions to deliver the capabilities of Fibre Channel over Ethernet networks:

- Encapsulation of native Fibre Channel frames into Ethernet Frames.
- Extensions to the Ethernet protocol itself to enable an Ethernet fabric in which frames are not routinely lost during periods of congestion.
- Mapping between Fibre Channel N_port IDs and Ethernet MAC addresses.

7.4.3 Data Center Bridging (DCB)

Data Center Bridging (DCB) defines the standards development efforts underway by the IEEE 802.1 work group. The work is aimed at adding new extensions to bridging and Ethernet to turn it into a lossless transport suitable for transporting storage traffic without losing data. The ability to move various types of data flows over a single shared lossless Ethernet link creates a convergence-friendly transport and opens the door for supporting requirements of applications from different purpose-built networks.

Data Center Bridging Capabilities Exchange (DCBX) protocol is a capabilities-exchange protocol that is used by DCB-capable switches to identify and communicate with each other. Neighboring devices use DCBX to exchange and negotiate configuration information and detect misconfigurations. Among other tasks, devices use DCBX to exchange information about FCoE, information about whether priority flow control (PFC) is enabled on the port, and ETS priority group information, including 802.1p priority values and bandwidth allocation percentages.

DCBX distributes the local configuration and detects the misconfiguration of ETS and PFC between peers. It also provides the capability for configuring a device with PFC, ETS, and application parameters. The application parameter configures the device on what priority to use for different applications such as FCoE and IPC.

By definition: The DCBX allows the mutual discovery of DCB-capable hosts and switches, allowing DCB-specific parameters, such as those of ETS and PFC, to be exchanged before Ethernet links are shared.

7.4.4 Enhanced Transmission Selection (ETS)

As previously discussed, PFC creates up to eight virtual lanes on a single Ethernet link. ETS augments PFC by enabling optimal bandwidth management of the virtual lanes. ETS provides prioritized processing of traffic based on the bandwidth allocation assigned to each traffic class, thereby providing guaranteed bandwidth for specific traffic classes.

ETS is a dynamic approach, as if a certain group or class does not use the allocated bandwidth, all other lanes for that port can use the available bandwidth if needed. The value selected for your lane (group) manages the bandwidth on the Ethernet link. This means that you can have multiple groups, each with a different priority using the available bandwidth. For example, you can assign a higher priority and a certain percentage of guaranteed bandwidth to storage traffic. The assigned percentage indicates the maximum bandwidth a priority group or traffic class group can use. Bandwidth allocation delivers quality of service to applications and data flows that need it, such as storage or control traffic.

Figure 7-10 shows lanes with different flow control mechanisms.

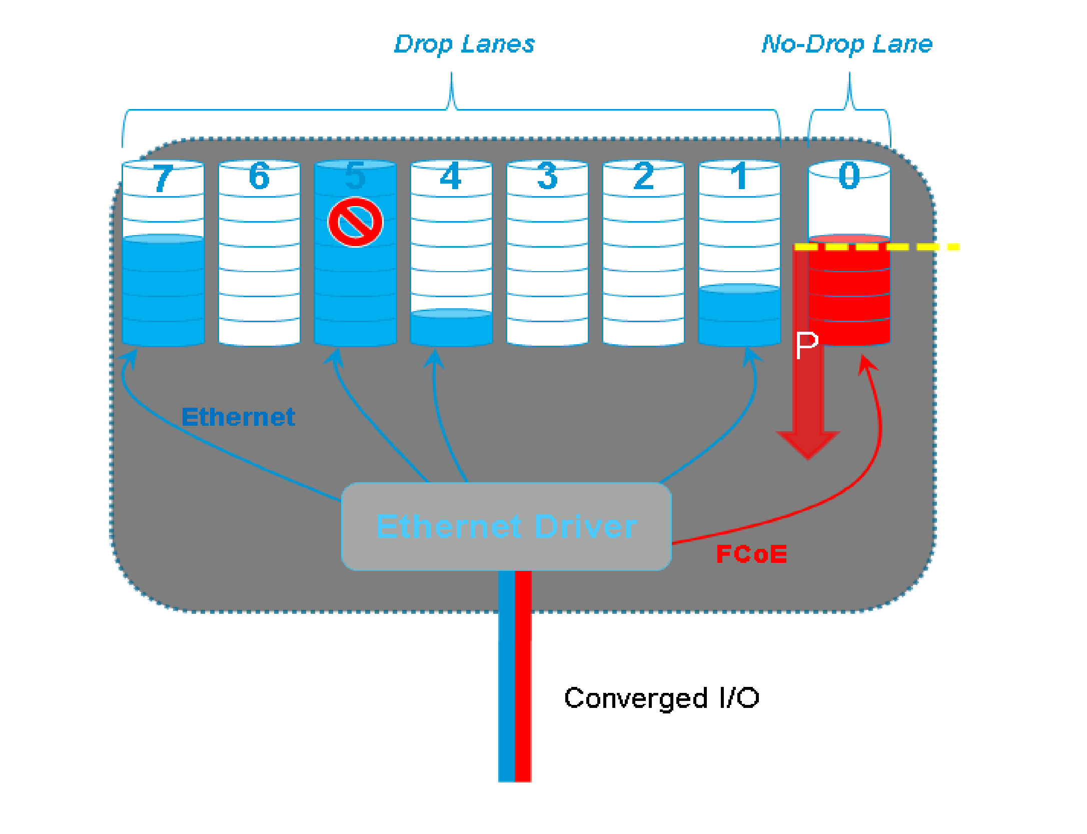


Figure 7-10 flow control lanes

In Figure 7-11, we demonstrate how the bandwidth adjusts dynamically to the bandwidth needed by the different types of traffic.

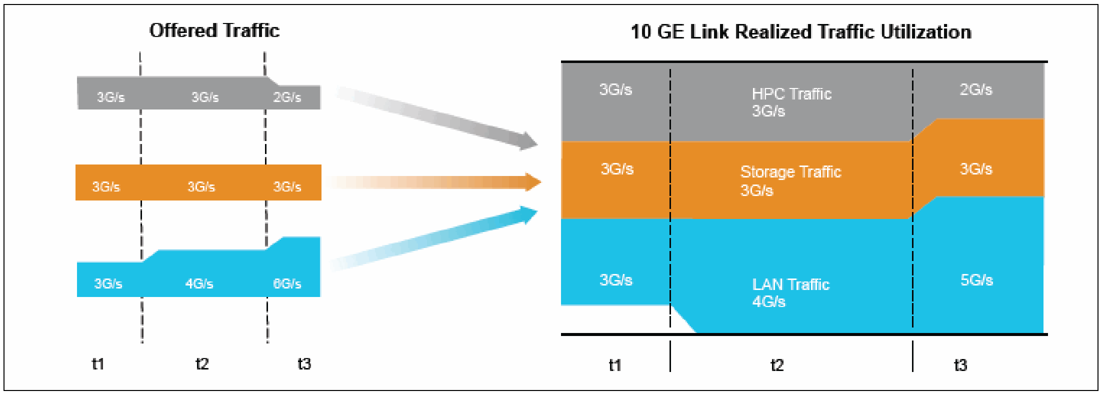


Figure 7-11 Dynamic bandwidth adjustments

As shown in Table 7-1 we can have three priority groups: priority 1 (IPC), priority 2 (FCoE) and priority 3 (LAN). All IPC traffic that is transmitted has the highest priority of the available bandwidth, and then the FCoE and the remaining bandwidth is then used by the LAN. As a result, the IPC applications, which might be sensitive to latency, are never starved for bandwidth by the other applications.

Table 7-1 shows priority groups

Priority group	Bandwidth	Type of traffic
1	-	IPC
2	50%	FCoE
3	50%	LAN

And keep in mind that these new components of enhanced Ethernet can affect existing layers in the OSI standard, for example with better flow control and class of service.

ETS incorporates a mechanism called Data Center Bridging Exchange (DCBX). DCBX controls the discovery and initialization protocol and collects MAC addresses from devices attached to the DCB network. The DCBX process uses the capabilities of IEEE 802.1AB Link Layer Discovery Protocol (LLDP) which is an open link layer protocol used by network devices to report their identity, capabilities and neighbors on the LAN. The Ethernet frame used in LLDP stores the destination address.

7.4.5 Priority Flow Control (PFC)

Functionally, Priority Flow Control or PFC is very similar to the Fibre Channel buffer to buffer credit system, described in 7.4.2, “Fibre Channel” on page 380, where the sender keeps track of buffer availability at the receiver's end and will transmit packets only if it knows that buffers are available. A traditionally lossy Ethernet network can be made lossless through the use of a special mechanism usually referred to as the *pause mechanism*.

PFC itself is standardized as part of IEEE 802.1Qbb. It enables the Ethernet pause capability based on priority or quality of service. This mechanism provides detailed congestion control on the link, allowing administrators to provide no-drop (lossless) behavior for FCoE traffic that shares the link with Ethernet traffic where frame drops can be gracefully handled by upper-level applications. PFC allows lossless behavior to be assigned to certain priority groups, according to the requirements of the mixed traffic types on a link-by-link basis. This feature essentially enables I/O consolidation on the same infrastructure carrying a mix of traffic with different requirements. A total of eight such priority groups, representing virtual lanes of traffic, can be available in an enhanced Ethernet network.

In short: The pause mechanism is when an Ethernet frame is sent by the *receiver* to indicate exhaustion of resources. Upon receipt of a pause frame, the *sender* will withhold transmission until either a timer has expired (a time-out defined by the receiver) or the receiver allows transmission to proceed.

Use of a pause mechanism enables flow control on the physical link and simulates a lossless network.

Figure 7-12 shows the pause mechanism on lane 3.

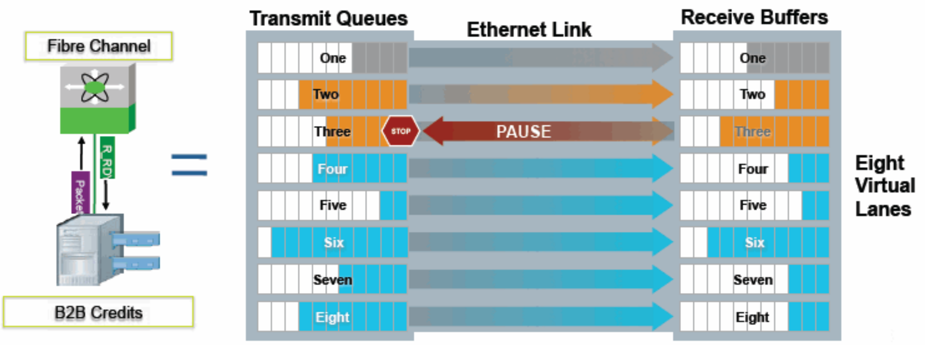


Figure 7-12 Pause mechanism

As stated previously, PFC provides a link level flow control mechanism that can be controlled independently for each priority. The goal of this mechanism is to ensure zero loss due to congestion in DCB networks. For FCoE, this is done by creating and defining the following parameters:

- Eight priority groups based on CoS marking
- FCoE traffic marked to be *No-Drop*
- PFC sent for No-Drop traffic when buffer threshold is reached

This is actually an enhancement to the Ethernet control mechanism defined in IEEE 802.3X often referred to as the “PAUSE” effect. It basically sends messages to the sender to stop sending frames due to the “PAUSE” effect, therefore preventing loss of FCoE frames in the No-Drop traffic class.

7.4.6 Quantized Congestion Notification (CN or QCN)

QCN is a part of the DCB standard and provides end-to-end congestion management for Layer 2 networks that do not already have congestion control mechanisms built in, and needs to be enabled on every device data path throughout the network.

In essence, this is a throttling mechanism that controls end-to-end congestion management which is defined in IEEE 802.1.Qau. The purpose of end-to-end congestion management is to ensure that congestion is controlled from the sending device to the receiving device in a dynamic fashion that can deal with changing bottlenecks.

> **Note:** PFC and ETS are independent protocols that do not require QCN to be implemented.

Vendors have made different statements about to whether CN is required for FCoE deployments. However, IBM and many other vendors have taken the position that CN is a “nice to have” feature but is optional, at least for initial FCoE deployments.

The applicability of CN is generally considered to be related to the size of the switched LAN over which FCoE transmission will occur. In an FCoE deployment, with a single or small number of intermediate DCB switches between source server and target FCoE or FCoE gateway device, PFC is most likely sufficient to ensure lossless behavior. In a larger network, with multiple DCB switches and multiple possible paths between source and destination of the FCoE flow, intermediate congestion might be possible despite PFC, increasing the need for CN.

7.4.7 Fibre Channel forwarder

The technology that makes a switch able to forward the FC frame is called Fibre Channel Forwarder (FCF). As shown in Figure 7-15 on page 386, the FCF is an element within the switch that forwards frames based on FC information. Each FCF has its own Domain ID just like a normal FC switch, and when two FCFs are connected, their connection ports are called VE-ports.

During the initialization of an FCoE Enode connection, the FLOGs are done in the FCF. The FCF assigns the Enode MAC address to be used for FCoE forwarding called Fabric Provided MAC Address (FPMA).

> **FCF**: A combination of FCoE termination functions and Fibre Channel stack on Ethernet switches (dual-stack switches) and are therefore equivalent to Fibre Channel switches in native Fibre Channel networks.

> **ENodes**: A combination of FCoE termination functions and Fibre Channel stack on the CNAs, and in that sense they are equivalent to host bus adapters (HBA's) in native Fibre Channel networks.

Each FCID is assigned a unique Enode FCoE MAC address.

- An FCoE Mapped MAC Address is assigned by the VN_Port “owns” the WWPN and WWNN.
- Fabric basically owns the FCID and FCoE mapped MAC address or the Layer 2 addresses.

- FC stack is migrated to new platform independent of the Ethernet (TCP/IP) stack and will get a new FCoE MAC address assigned.

Figure 7-13 shows FCF connecting devices.

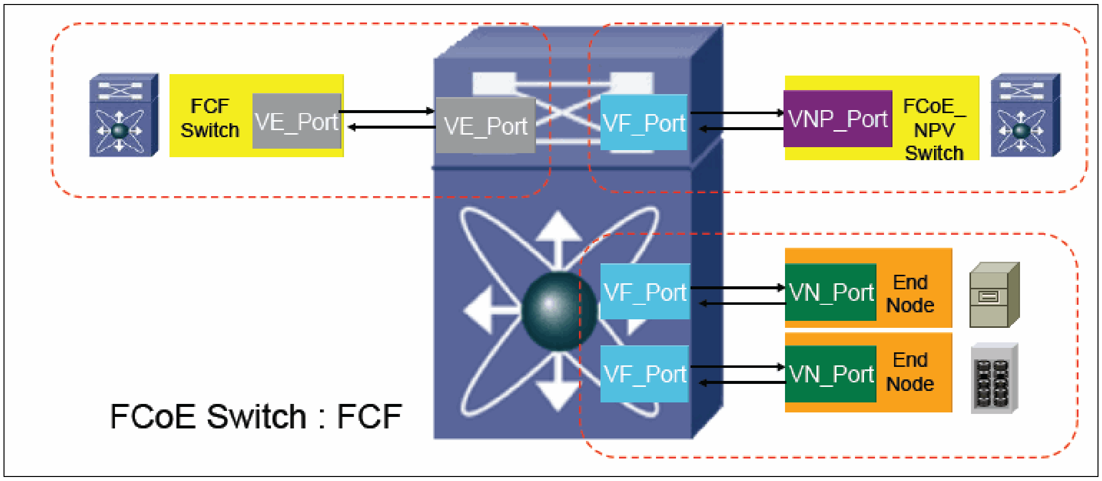


Figure 7-13 FCF and ports names

To explain the different port types, Figure 7-14 shows where ENodes are connected to FCF, but we also see that normal FC switches connect using normal FC ports.

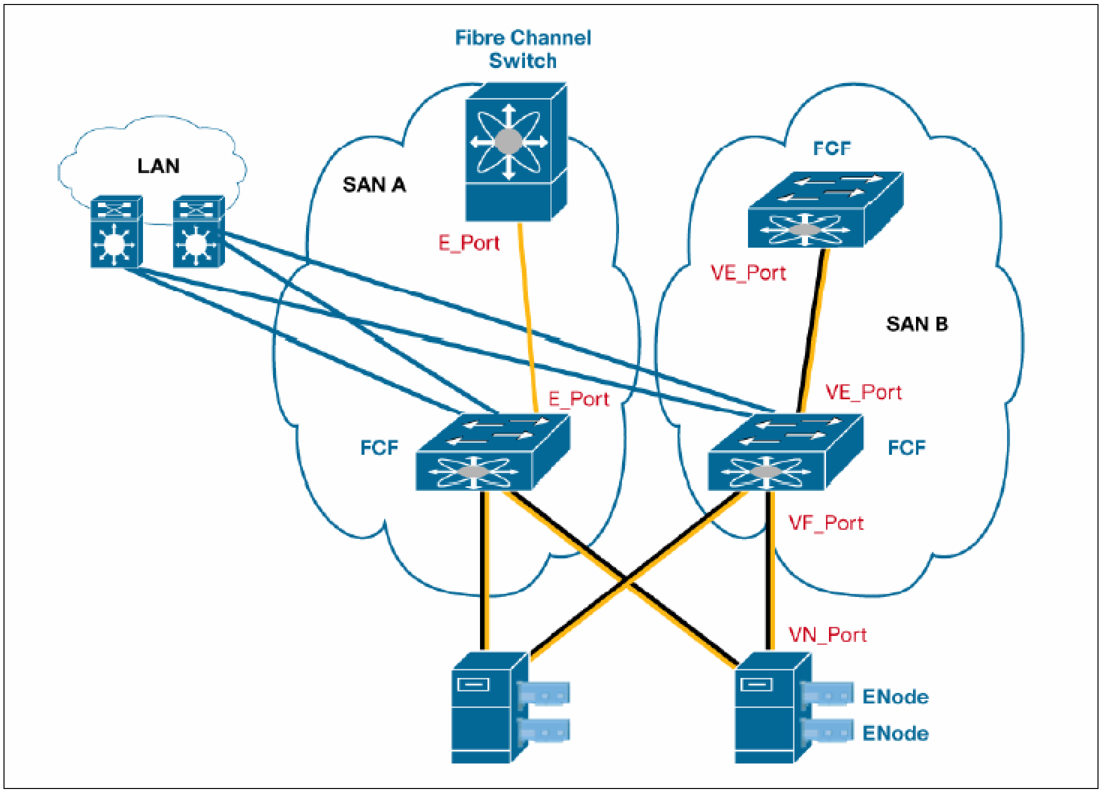


Figure 7-14 ENodes and FCFs

DCB switches typically have an “embedded” FCF as shown in Figure 7-15.

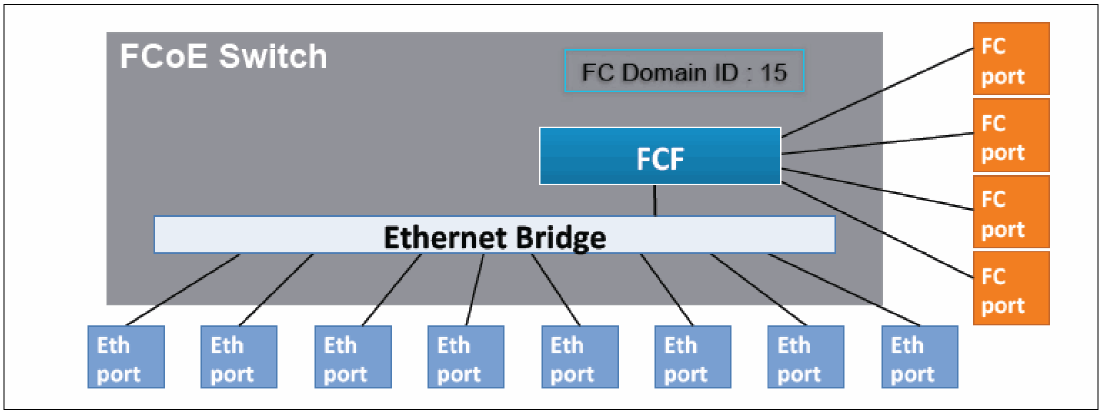


Figure 7-15 shows an FCF in an enhanced Ethernet switch

7.4.8 FCoE Initialization Protocol

FCoE Initialization Protocol (FIP) is the control plane protocol that is used in the discovery and initialization stages of establishing links among the elements of the fabric. Once discovery and initialization are established and the nodes have successfully performed fabric login, the FCoE data plane frames are used for data transmission operations. Unlike FCoE frames, FIP frames do not transport Fibre Channel data, but contain discovery and Login/Logout parameters. FIP frames are assigned a unique EtherType code to distinguish them from FCoE frames with Fibre Channel storage data.

In FCoE, ENodes discover FCFs and initialize the FCoE connection through the FCoE Initialization Protocol (FIP). The FIP has a separate EtherType from FCoE.

Table 7-2 shows the port definitions.

Table 7-2 Port definitions

Name	Description
FC Node	The end-point entity in a host, providing N_Ports
FC Switch	The FC frames switching entity, providing F_Ports and E_Ports
N_Port to F_Port	A link between a FC Node and a FC Switch
E_Port to E_Port	A link between two FC Switches
ENode	An FC Node supporting FCoE on at least one Ethernet interface
FCF	A switch supporting FCoE on at least one Ethernet interface
VN_Port to VF_Port	A Virtual Link between an ENode and an FCF
VE_Port to VE_Port	A Virtual Link between two FCFs

Rather than just providing an FCID, FIP will provide an FPMA, which is a MAC address composed of two parts: FC-MAP and FCID and FIP uses a separate Ethertype from FCoE and its frames are standard Ethernet size (1518 Byte 802.1q frame) whereas FCoE frames are 2242 Byte Jumbo Frames.

FPMAs use the 24-bit-wide Fibre Channel ID (FC_ID) assigned to the CNA during the FIP FLOGI and FDISC exchange, and therefore they cannot be available to the CNA before the fabric login has occurred. The FPMA is built by appending the FC_ID to a 24-bit quantity called the FCoE MAC address prefix (FC-MAP) as shown in Figure 7-16.

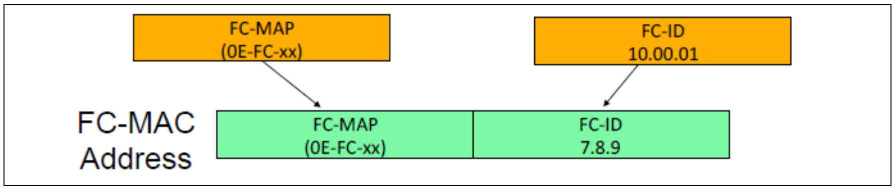


Figure 7-16 FC-MAC address

FC-BB-5 defined a range of 256 FC-MAPs to facilitate FCoE deployments. Cisco has established best practices that make the manipulation of FC-MAPs unnecessary, and most users should find the default FC-MAP value 0E-FC-00 sufficient. The 256 different FC-MAPs make available to users up to 256 pools of locally unique MAC addresses. The pools are useful when the FC_IDs are not unique on an Ethernet VLAN; for instance, when different Fibre Channel fabrics or different VSANs are encapsulated in the same Ethernet VLAN, the ranges of FC_IDs assigned in each Fibre Channel fabric may overlap. Cisco strongly recommends that you never attempt to map multiple Fibre Channel fabrics onto the same Ethernet VLAN. Most users will not ever need to map multiple Fibre Channel fabrics onto the same physical Ethernet network, but if such a need arises, each Fibre Channel fabric should be encapsulated in a separate VLAN.

Figure 7-17 shows initial FIP transactions.

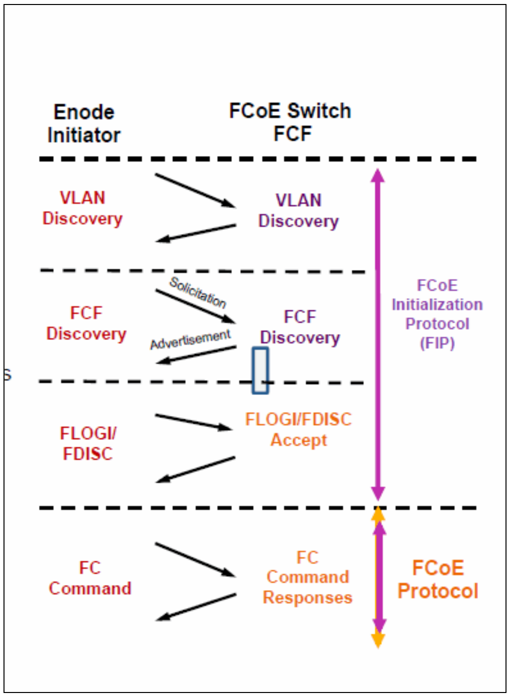


Figure 7-17 FIP transactions

7.4.9 FIP VLAN discovery

FIP VLAN discovery discovers the FCoE VLAN that will be used by all other FIP protocols as well as by the FCoE encapsulation for Fibre Channel payloads on the established virtual link. One of the goals of FC-BB-5 was to be as non-intrusive as possible on initiators and targets, and therefore FIP VLAN discovery occurs in the native VLAN used by the initiator or target to exchange Ethernet traffic. The FIP VLAN discovery protocol is the only FIP protocol running on the native VLAN; all other FIP protocols run on the discovered FCoE VLANs.

The ENode sends a FIP VLAN discovery request to a multicast MAC address called All-FCF-MACs, which is a multicast MAC address to which all FCFs listen. All FCFs that can be reached in the native VLAN of the ENode are expected to respond on the same VLAN with a response that lists one or more FCoE VLANs that are available for the ENode’s VN_Port login. This protocol has the sole purpose of allowing the ENode to discover all the available FCoE VLANs, and it does not cause the ENode to select an FCF.

FIP VLAN discovery is an optional protocol in FC-BB-5. An ENode implementation can choose to offer only manual configuration for FCoE VLANs, and therefore choose not to perform FIP VLAN discovery. It is commonly assumed that such implementation will default to VLAN 1002 for its FCoE VLAN.

7.4.10 FIP FCF discovery

FIP FCF discovery is the protocol used by ENodes to discover FCFs that can accept logins. FCFs periodically send FIP FCF discovery advertisement messages on each configured FCoE VLAN; these messages are destined for the multicast MAC address All-ENode-MACs, a multicast MAC address to which all ENodes listen. The FIP FCF discovery advertisement is used by the FCF to inform any potential ENode in the VLAN that FCF VF_Ports are available for virtual link establishment with ENodes' VN_Ports. The advertisement includes the MAC address of the FCF as well as other parameters useful for tuning the characteristics of the virtual link (FIP time-out values, FCF priority, and so on).

Given the periodic nature of the advertisements, new ENodes joining the network will typically not want to wait to collect multicast FIP FCF discovery advertisements from all FCFs, and therefore FC-BB-5 allows ENodes to solicit unicast advertisements by sending a FIP FCF discovery solicitation to the All-FCF-MACs multicast MAC address. FCFs receiving the solicitation can generate a unicast FIP FCF discovery advertisement addressed to the requesting ENode. Upon collection of these advertisements, the ENode can make the final decision as to which FCF to contact for the establishment of a virtual link with its VN_Port.

7.4.11 FIP FLOGI and FDISC

After the ENode has discovered all FCFs and selected one for login, the last step is to inform the selected FCF of the intention to create a virtual link with its VF_Port. After this step, Fibre Channel payloads (encapsulated in FCoE frames) can start being exchanged on the new virtual link just established. On any native Fibre Channel link between an N_Port and an F_Port, the first protocol exchange performed as part of activating the data-link layer is the fabric login, or FLOGI, which results in the assignment of an FC_ID to the N_Port. In designing FIP, the T11 committee decided to merge the logical step of FCF selection by an ENode in FIP with the native Fibre Channel fabric login exchange. The result of this optimization is a single FIP exchange that serves both purposes of FCF selection, as well as fabric login and FC_ID allocation. This optimization is not only convenient; it is a requirement for obtaining an appropriate FPMA for the ENode to use in the subsequent FCoE encapsulated frames.

FIP FLOGI and FDISC are unicast frames almost identical to the native Fibre Channel FLOGI and FDISC frames they replace. The VN_Port sends an FLOGI or an FDISC request, followed by the corresponding FLOGI or FDISC accept payload from the FCF. Completion of this exchange terminates the FIP virtual link establishment phase.

So if we look at the LIP process shown in Figure 7-18, then we could divide FIP into a discovery process and login process similar to this:

Step 1: FIP discovery process:

1. FCoE VLAN discovery
2. FCF discovery
3. Verifies that lossless Ethernet is capable of FCoE transmission

Step 2: FIP login process:

1. Similar to existing Fibre Channel Login
2. Process: Sends FLOGI to upstream FCF
3. FCF assigns the host an enode MAC address to be used for FCoE forwarding: Fabric Provided MAC Address (FPMA)

Figure 7-18 shows the FIP discovery and port naming.

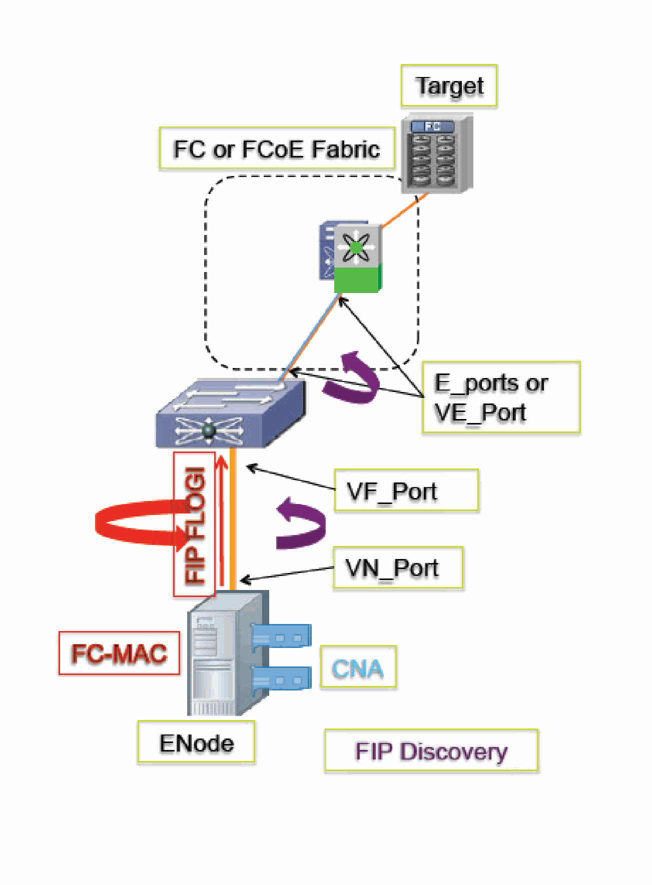


Figure 7-18 FIP discovery

7.4.12 FIP snooping

The FIP snooping firewall filters deny any FCoE traffic on the VLAN except for traffic originating from ENodes that have already logged in to the FCF. When FIP snooping is enabled, the FCoE transit switch inspects both FIP frames and FCoE frames.
So FIP snooping is only needed on forwarders that do not encapsulate the FC frames.

In the industry, FCoE pass-thru switches that implement either Recommendation D.4.1 or D.4.2 are commonly referred to as FIP snooping switches, to distinguish them from FCoE pass-thru switches that simply implement IEEE DCB and do not offer any additional FCoE security capability on top of that.

A definition of D.4.1 and D.4.2 can be found here:

http://www.fcoe.com/09-056v5.pdf

7.5 Multi-hop Fibre Channel over Ethernet

In Fibre Channel over Ethernet (FCoE), a “hop” is when the frame needs to travel from one FCF to another FCF through VE-ports, just as in the FCP environment when an FC frame goes from one domain ID to another domain ID.

But we need to be careful when we talk about multi-hop since we are allowed by the standard to insert DCB lossless bridge between our VN port and our VF port, so we cannot always trust just counting switches in our topology.

This method to insert a “tunnel” between VN and VF ports complicates troubleshooting and minimizes the security of the fabric since this “tunnel” is not visible to the storage administrator. But by adding FIP snooping we can ensure authentication of our frames. The FIP snooping bridge uses dynamic ACLs to enforce the FC rules within the DCB network.

FCoE is forwarded on a hop by hop basis where the switches do not have knowledge of the MAC address beyond the next hop, once it receives the FC frame and claims ownership.

Hop behavior: Because FCoE is Fibre Channel, it behaves according to Fibre Channel rules as to what a “hop” is.

In a multi-hop topology, the design becomes more interesting as our Fibre Channel network is not connected to the first hop switch, and we now have multiple Ethernet paths through the network. Spanning Tree Protocol (STP) will block links in our network to provide a loop free topology which won't meet Fibre Channel's needs. Cisco Fabric Path and Transparent Interconnections of Lots of Links (TRILL) help to overcome this.

Cisco FabricPath, as its name implies, is a Cisco proprietary implementation of MAC routing while Transparent Interconnections of Lots of Links (TRILL) is an Internet Engineering Task Force (IETF) project. With either Cisco FabricPath or TRILL, we can provide topologies with mammoth amounts of parallel links and lots of loops that are forwarding all the time due to the absence of STP. When Fabricpath/TRILL is coupled with DCB, it provides a network that SAN administrators can place their lossless traffic on without worry of dropped frames and gain all of the benefits of network multi-pathing.

As discussed previously, the Data Center Bridging capabilities in PFC, ETS, DCBX and QCN have enhanced Ethernet to make it reliable transport for FCoE frames. DCB will also enable FCoE to be used reliably as a multi-hop protocol.

7.6 iSCSI technology

When we talk about networking storage with FCoE or FC, then iSCSI is always part of that discussion. In essence, iSCSI is the encapsulation of the standard SCSI protocol within TCP/IP packets. Now we might think that iSCSI and FC0E are the same thing (block level storage moved using Ethernet), but the fundamental difference of these two protocols is that iSCSI adds an additional name space and utilizes TCP/IP, while FCoE leverages the existing FC protocol semantics and utilizes Ethernet Transport with TCP/IP overhead. Due to the absence of TCP/IP and additional naming conventions, FCoE (like FC) is better suited for low latency, high performance applications. iSCSI is better suited for higher latency environments, low to midrange performance applications.

With an iSCSI network, you can use an existing traditional Ethernet networking infrastructure, reducing the costs for specialized storage area networking devices, software, and licenses. iSCSI has a head start for many businesses because it already has a stable network infrastructure in place. iSCSI uses the reliable TCP protocol to transport SCSI I/O commands over a network, providing block-level data access without needing specialized hardware requirements. It can also operate with various peripheral devices.

The iSCSI protocol allows for longer distances between a server and its storage when compared to the traditionally restrictive parallel SCSI solutions or the newer serial-attached SCSI (SAS). iSCSI technology can use a hardware initiator, a host bus adapter (HBA), or a software initiator to issue requests to target devices. Within iSCSI storage terminology, the initiator is typically known as the *client*, and the target is known as the *storage device*. The iSCSI protocol encapsulates SCSI commands into protocol data units (PDUs) within the TCP/IP protocol, and then transports them over the network to the target device.

The iSCSI protocol is a transport for SCSI over TCP/IP. Figure 7-2 on page 371 illustrates a protocol stack comparison between FC and iSCSI. iSCSI provides block-level access to storage, as does FC, but uses TCP/IP over Ethernet instead of the FC protocol.

Because iSCSI uses Ethernet-based TCP/IP rather than a dedicated (and different) SAN technology, it is attractive for its relative simplicity and uses widely available Ethernet skills. Its limitations are mostly the relatively lower speeds of Ethernet compared to FC and the extra TCP/IP encapsulation that is required. iSCSI transactions occur between an iSCSI initiator (hardware or software) that transmits a request (such as read/write) and an iSCSI target. This iSCSI target processes the request and responds with the appropriate information, such as data and sense. iSCSI initiators are typically application servers or users, where iSCSI targets are typically SAN access points or actual storage controllers. Because an iSCSI request is an encapsulation of a SCSI request, the SCSI concept of command descriptor blocks (CDB) is applicable to iSCSI. CDBs define the type of SCSI operation, the logical block address where to start, the length of data involved, and other control parameters.

Figure 7-19 shows where iSCSI sits in the stack.

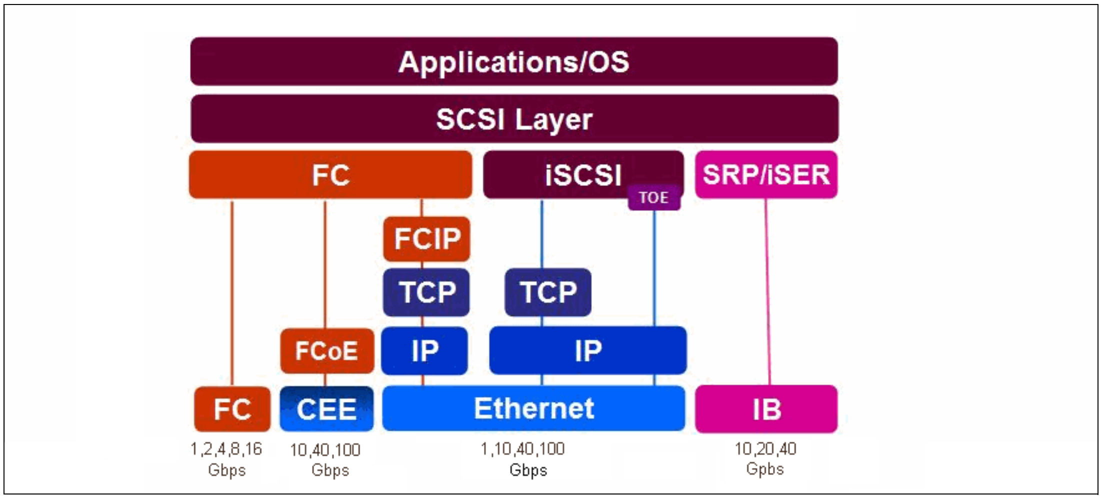


Figure 7-19 The flow stack

7.7 Cisco and IBM solutions for FCoE

This section focus on reviewing the Cisco switches that support FCoE and then shows the multiple topologies for connecting IBM Servers (System x, System p and BladeCenter) to Cisco switches.

Some of the topologies shown on this section are fully certified and tested by IBM and Cisco, while other topologies may not have been fully certified and tested, yet they are technically possible.

All technically possible topologies are shown given that the number topologies fully tested certified will continue to increase over time. For any solution that is not officially supported and certified, an IBM SCORE/RPQ may be submitted.

7.7.1 Cisco FCoE portfolio

Cisco supports FCoE on the following products:

- Nexus 2232PP: The Nexus 2232PP is a Fabric Extender (FEX) that supports FCoE when it is connected to the Nexus 5x00/6000 switches. It has 32 1/10 Gigabit Ethernet host interfaces that support Fibre Channel over Ethernet (FCoE) and 8 10 Gigabit Ethernet for uplinks to the Nexus 5x00 that are the fabric interfaces and transport FCoE between the Nexus 2232PP and the Nexus 5x00 that it is connected to. The maximum distance between the Nexus 2232PP and the parent Nexus 5x00 when using FCoE is 300 meters.
- Nexus 2248PQ: The Nexus 2248PQ is a Fabric Extender (FEX) that supports FCoE when it's connected to the Nexus 5x00/6000 switches. It has 48 1/10 Gigabit Ethernet host interfaces that support Fibre Channel over Ethernet (FCoE) and 4 x 40 Gigabit Ethernet uplinks (which can also be used as 16 x 10 GE uplinks) to the parent Nexus 5x00/6000 switch. The uplinks are the fabric interfaces to transport FCoE between the Nexus 2248PQ and the Nexus 5x00/6000 that it's connected to. The maximum distance between the Nexus 2248PQ and the parent Nexus 5x00/6000 switch when using FCoE is 300 meters.

- Nexus 4001I: The Nexus 4001I Switch Module for IBM BladeCenter is a blade switch solution for IBM BladeCenter H and HT chassis. It is a line-rate, extremely low-latency, non-blocking, Layer 2, 10 Gigabit Ethernet blade switch that supports Fibre Channel over Ethernet (FCoE) and IEEE 802.1 Data Center Bridging (DCB) standards. The Nexus 4001I has fourteen fixed 10 Gigabit Ethernet server-facing downlinks (auto-sensing ports; can also operate in Gigabit Ethernet mode) and six fixed 10 Gigabit Ethernet uplinks (auto-sensing ports; can also operate in Gigabit Ethernet mode). The Cisco Nexus 4001I inserts into the high-speed slot (HSS) of the IBM BladeCenter H or HT chassis. The Cisco BladeCenter H and HT chassis are designed to support up to four Cisco Nexus 4001I switches per chassis.
- Nexus 5010: The Cisco Nexus 5010 is a 1RU, 10 Gigabit Ethernet, FCoE, and Fibre Channel switch. It has 20 fixed 10 Gigabit Ethernet and FCoE SFP+ ports. One expansion module slot can be configured to support up to 6 additional 10 Gigabit Ethernet and FCoE SFP+ ports, up to 8 4Gbps SFP Fibre Channel switch ports, up to 6 8Gbps SFP+ Fibre Channel switch ports or a combination of 4 additional 10 Gigabit Ethernet and FCoE SFP+ ports with 4 additional 4/2/1-Gbps Fibre Channel switch ports.
- Nexus 5020: The Cisco Nexus 5020 is a two-rack-unit (2RU), 10 Gigabit Ethernet, FCoE, and Fibre Channel switch. It has 40 fixed 10 Gigabit Ethernet and FCoE Small Form-Factor Pluggable Plus (SFP+) ports. Two expansion module slots can be configured to support up to 12 additional 10 Gigabit Ethernet and FCoE SFP+ ports, up to 16 Fibre Channel switch ports, or a combination of both.
- Nexus 5548P: The Cisco Nexus 5548P is a one-rack-unit (1RU) 10 Gigabit Ethernet and FCoE switch offering up to 960-Gbps throughput and up to 48 ports. The switch has 32 1/10-Gbps fixed SFP+ Ethernet and FCoE ports and one expansion slot.
- Nexus 5548UP: The Cisco Nexus 5548UP is a 1RU 10 Gigabit Ethernet, Fibre Channel, and FCoE switch offering up to 960 Gbps of throughput and up to 48 ports. The switch has 32 unified ports and one expansion slot.
- Nexus 5596UP: The Cisco Nexus 5596UP Switch is a 2RU 10 Gigabit Ethernet, Fibre Channel, and FCoE switch offering up to 1920 Gbps of throughput and up to 96 ports. The switch has 48 unified ports and three expansion slots.
- Nexus 5596T: The Cisco Nexus 5596T Switch is a 2RU form factor, with 32 fixed ports of 10G BASE-T and 16 fixed ports of SFP+. The switch also supports up to three expansion slots. The switch supports 10 Gigabit Ethernet (fiber and copper), Fibre Channel, and FCoE, offering up to 1920 Gbps of throughput and up to 96 ports. The switch supports unified ports on all SFP+ ports. The hardware for the 10G BASE-T ports is capable of supporting FCoE.
- Expansion Module Options for the Cisco Nexus 5548P, 5548UP, 5596UP, and 5596T Switches: The Cisco Nexus 5500 platform is equipped with expansion modules that can be used to increase the number of 10 Gigabit Ethernet and FCoE ports or to connect to Fibre Channel SANs with 8/4/2/1-Gbps Fibre Channel switch ports, or both.
- Nexus 6004: The Cisco Nexus 6004 Switch is a 4RU 10/40 Gigabit Ethernet and FCoE switch offering up to 96 40 GE ports (or 384 10 GE ports). The switch has 48 40G ports and four expansion slots, each capable of housing an expansion module of 12 x 40 GE ports. Each 40G port in the Nexus 6004 can be configured as either a single 40 GE port or 4 x 10 GE ports.
- Nexus 6001: The Cisco Nexus 6001 Switch is a 1RU 10/40 GE Gigabit Ethernet and FCoE switch. The Nexus 6001 has 48 1/10 GE ports and 4 x 40 GE uplink ports, each of which can be configured as either a single 40 GE port or 4 x 10 GE ports. All ports in the Nexus 6001 support both native 10/40 GE and FCoE.

- Nexus 7000: FCoE is supported on the Nexus 7004, 7009, 7010 and 7018, which have 2, 7, 8 and 16 linecard slots respectively. FCoE on the Nexus 7000 is supported on the F1 (N7K-F132XP-15), F2 (N7K-F248XP-25), and F2E (N7K-F248XP-25E) linecards. The Nexus 7004 does not support the F1 linecard. The F2 and F2E linecards requires Supervisor 2 or Supervisor 2E for FCoE support.
- MDS 9500/9700 Series Multilayer Storage Directors: The MDS 9500 and 9700 are director-class SAN switches. FCoE is supported on the MDS 9506, 9509, 9513 and 9710. Those MDS switches have 4, 7, 11 and 8 linecard slots respectively. FCoE on the MDS 9500 switches is supported on the 10-Gbps 8-Port FCoE Module (DS-X9708-K9). This module provides the scalability of 88 line-rate ports per chassis (MDS 9513), without oversubscription, and PortChannels with up to 16 physical links, creating 160-Gbps logical links. The 10-Gbps 8-Port FCoE Module for the MDS 9500 requires Supervisor-2A Module to be installed on the chassis. Multihop FCoE supported on the MDS 9500 is fully compatible with FCoE connectivity to the Nexus 5000 and the Nexus 7000. The 9710 will support a 48 port 10G FCoE module in a future release.
- MDS 9250i Multiservice Fabric Switch: The Cisco MDS 9250i offers up to forty 16-Gbps Fibre Channel ports, two 1/10 Gigabit Ethernet IP storage services ports, and eight 10 Gigabit Ethernet Fibre Channel over Ethernet (FCoE) ports in a fixed two-rack-unit (2RU) form factor. Using the eight 10 Gigabit Ethernet FCoE ports, the Cisco MDS 9250i platform attaches to directly connected FCoE and Fibre Channel storage devices and supports multitiered unified network fabric connectivity directly over FCoE.

More details about the MDS 9500 Series Multilayer Directors can be found at:

http://www.cisco.com/en/US/products/ps5990/index.html

More details about the MDS 9710 Series Multilayer Directors can be found at:

http://www.cisco.com/en/US/prod/collateral/ps4159/ps6409/ps12970/data_sheet_c78-727769.html

More details about the MDS 9250i Multi Service Fabric Switch can be found at:

http://www.cisco.com/en/US/prod/collateral/ps4159/ps6409/ps5988/ps13021/data_sheet_c78-727493.html

For details on the Cisco Nexus switches mentioned, see Chapter 2, “IBM and Cisco building blocks for the data center” on page 29.

7.7.2 Topologies for connecting IBM servers to Cisco FCoE switches

The following sections describe the multiple topologies for connecting IBM servers (System x, System p, and BladeCenter) to Cisco switches that support FCoE.

7.7.3 Topology 1: FCoE at the access layer with Nexus 5x00 and FC uplink to the SAN

Introducing FCoE at the edge (access layer) of the network is the simplest way to converge the LAN and SAN using FCoE. This topology (illustrated on Figure 7-20) does not require any change on the Data Center LAN Aggregation Layer or the Fibre Channel SAN Core Layer.

The consolidation happens on the access layer and FCoE only runs between the Server CNA (Converged Network Adapter) and the Nexus 5x00 switches which are the converged access layer switches.

On this topology the link between the server and the Nexus 5x00 is a converged link transporting Ethernet and FCoE traffic. When the FCoE traffic arrives on the Nexus 5x00, it removes the FCoE Encapsulation and the original Fibre Channel frame is sent on the native Fibre Channel uplink towards the FC SAN Core. The Ethernet packets when they arrive on the Nexus 5x00 are forwarded on the LAN uplinks towards the LAN aggregation layer.

The FCoE traffic is carried on a VLAN dedicated to FCoE traffic, while the Ethernet traffic is carried on separate VLANs.

This topology, like any other FCoE topology discussed in this section, preserves the Fabric A / Fabric B separation commonly found in Fibre Channel-based Storage Area Networks (SANs).

In this topology, the Nexus 5x00 has a native Fibre Channel uplink to the SAN core and it can run in FC Switch Mode or in NPV Mode:

- In FC Switch mode the ports between the FC director (Core) and the Nexus 5x00 are expansion ports (E ports) and the Nexus 5x00 is assigned a domain ID. In switch mode the Nexus 5x00 actively participates on the FC fabric as any other FC switch does.
- N-Port Virtualizer (NPV) mode is a Cisco feature that takes advantage of N-Port ID virtualization (NPIV) defined in the FC standard. When the Nexus 5x00 is running in NPV mode, it is not an FC switch but instead it passes to the upstream director all the fabric login requests received from the hosts. When operating in NPV mode the Nexus 5x00 function as an “aggregator” of physical connections for the servers and then passes all the traffic to the upstream director. In NPV mode the Nexus 5x00 is a very simple port aggregator and thus eliminates the concerns related with domain ID scalability because it does not consume a domain ID in the fabric and also interoperability concerns because the Nexus 5x00 is not functioning as a FC switch. When the Nexus 5x00 is operating in NPV mode the Core Director needs to be configured to support N-Port ID virtualization (NPIV) which is a Fibre Channel standard and is supported by Cisco and Brocade.

More details about operating the Nexus 5x00 in NPV or in switch mode can be found at:

http://www.cisco.com/en/US/partner/docs/solutions/Enterprise/Data_Center/UF_FCoE_final.html#wp1146167

Cisco recommends that the Nexus 5x00 is configured in NPV mode in this topology. This keeps the configuration and interaction between the Nexus 5x00 and the Core FC switches very simple. With this approach the FC SAN core can be based on Cisco MDS or on Brocade FC switches.

Figure 7-20 shows FCoE at the access layer with Nexus 5x00 and FC uplink to the SAN. The servers connect to the Nexus 5x00 using their CNA's. If we consider Server 1, the FCoE traffic from the port of the CNA that is going to Fabric A (left link from the server) is transported on VLAN 20, marked in Red. The FCoE traffic from the port of the CNA that is going to Fabric B (right link from the server) is transported on VLAN 30, marked in Red as well. VLAN 20 is only present on the left link between the server 1 and the Nexus 5x00 on the left. VLAN 30 is only present on the right link between server 1 and the Nexus 5x00 on the right. This way Fabric A and Fabric B are completely separated. The Ethernet traffic is transported on VLAN 10 in this example, and is present on both links from the server to the Nexus 5x00.

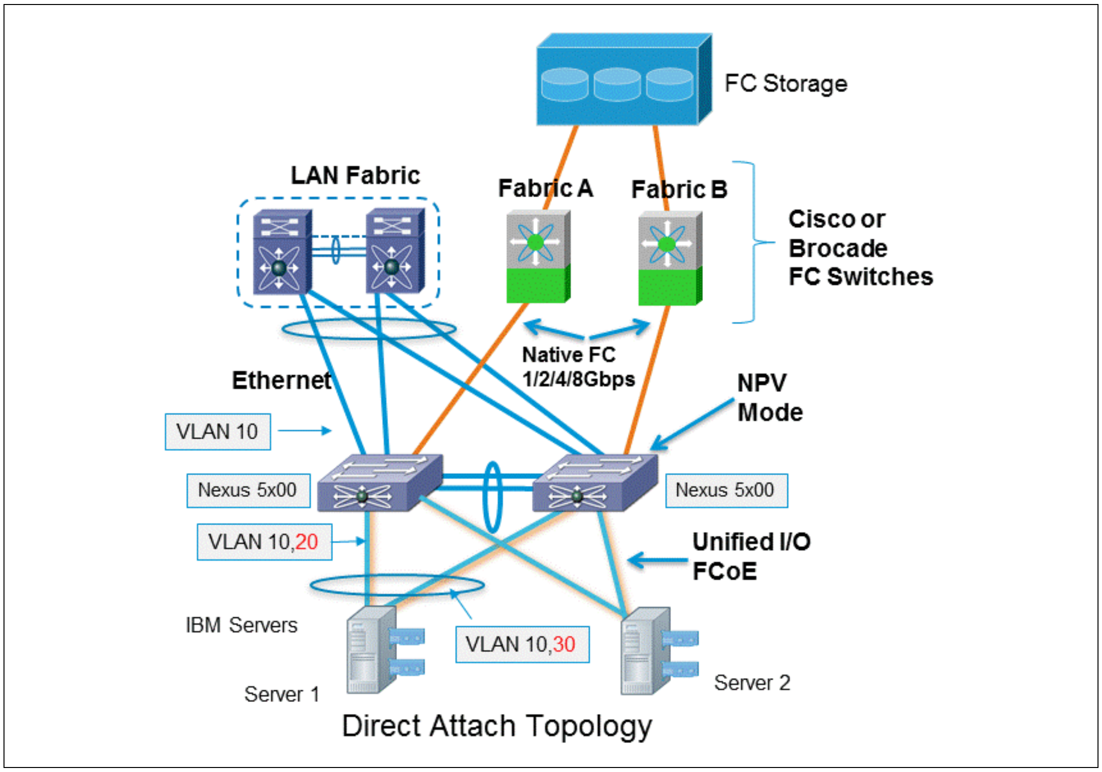


Figure 7-20 FCoE at the access layer with Nexus 5x00 and FC uplink to the SAN

The servers can connect to the Nexus 5x00 using vPC (Virtual Port-Channel), In this case the Ethernet traffic (Server 1, VLAN 10 in the example) is load balanced between both 10G links, or the servers can connect to the Nexus 5x00 using NIC Teaming (like Server 2 in the diagram).

The Nexus 5x00s are configured in NPV mode and have native Fibre-Channel (1/2/4/8Gbps) uplinks to the FC SAN Core. Because the Nexus 5x00 is running in NPV mode the FC core switches can be Cisco MDS or Brocade as long as they are configured with NPIV.

The connection between the Nexus 5x00 and LAN aggregation layer is a regular Ethernet connection and can utilize vPC as shown on the diagram or it can be based on spanning-tree. The Data Center LAN aggregation layer can be based on Cisco Nexus or any other Ethernet switch.

7.7.4 Topology 2: FCoE at the access layer with Nexus 5x00/6000 and FCoE uplink to the SAN

Note : This design option is described using the Nexus 5000, but the Nexus 6000 can also be used in place of the Nexus 5000 in this design. The same design concepts apply with the Nexus 6000, as it supports FCoE on all of its 10G/40G Ethernet ports.

This topology is very similar to Topology 1; the only difference is that the uplink between the Nexus 5x00 and the SAN core switches utilizes Fibre Channel over Ethernet (FCoE) instead of native Fibre Channel. Because there is no interoperability between Cisco and Brocade FCoE switches, the SAN core switches must be Cisco MDS 9500s, MDS 9710 or 9250i in this topology.

The Cisco MDS 9500 will use the MDS 10-Gbps 8-Port FCoE Module (DS-X9708-K9) to support the connection between the Nexus 5x00 and the MDS. The 9250i natively supports 8 ports of 10G FCoE and the 9710 will support a 48 port 10G FCoE card in the future.

When the Fibre Channel over Ethernet (FCoE) frame arrives on the Nexus 5x00, it is forwarded with the FCoE encapsulation to the MDS 9500 via the FCoE uplinks. The MDS will then remove the FCoE encapsulation before sending the native Fibre Channel frame to the attached FC storage systems. This FCoE connection between the Nexus 5x00 and the MDS only transport FCoE frames; it does not transport IP Ethernet or any non-FCoE frames.

On this topology the Nexus 5500 can be configured in FCoE NPV mode or in FCoE Switch mode. FCoE NPV mode is similar to FC NPV.

More details about using FCoE NPV can be found at this website:

http://www.cisco.com/en/US/partner/docs/switches/datacenter/nexus5000/sw/mkt_ops_guides/513_n1_1/n5k_ops_fcoe.html

Figure 7-21 shows FCoE at the access layer with Nexus 5x00 and FCoE uplink to the SAN. The only difference between this topology and topology 1 is that when the FCoE frames arrive on the Nexus 5x00, they are forwarded with the FCoE encapsulation to the MDS core switches. The MDS core switches then remove the FCoE encapsulation before sending the native Fibre-Channel frames to the FC Storage system. The only VLANs that exist on the link between the Nexus 5x00 and the MDS switches are the VLANs dedicated for FCoE traffic. For example, on the link between the Nexus 5x00 on the left and the MDS on the left (Fabric A) only VLAN 20 would be allowed on this connection. On the link between the Nexus 5x00 on the right and the MDS on the right (Fabric B) then only VLAN 30 would be allowed.

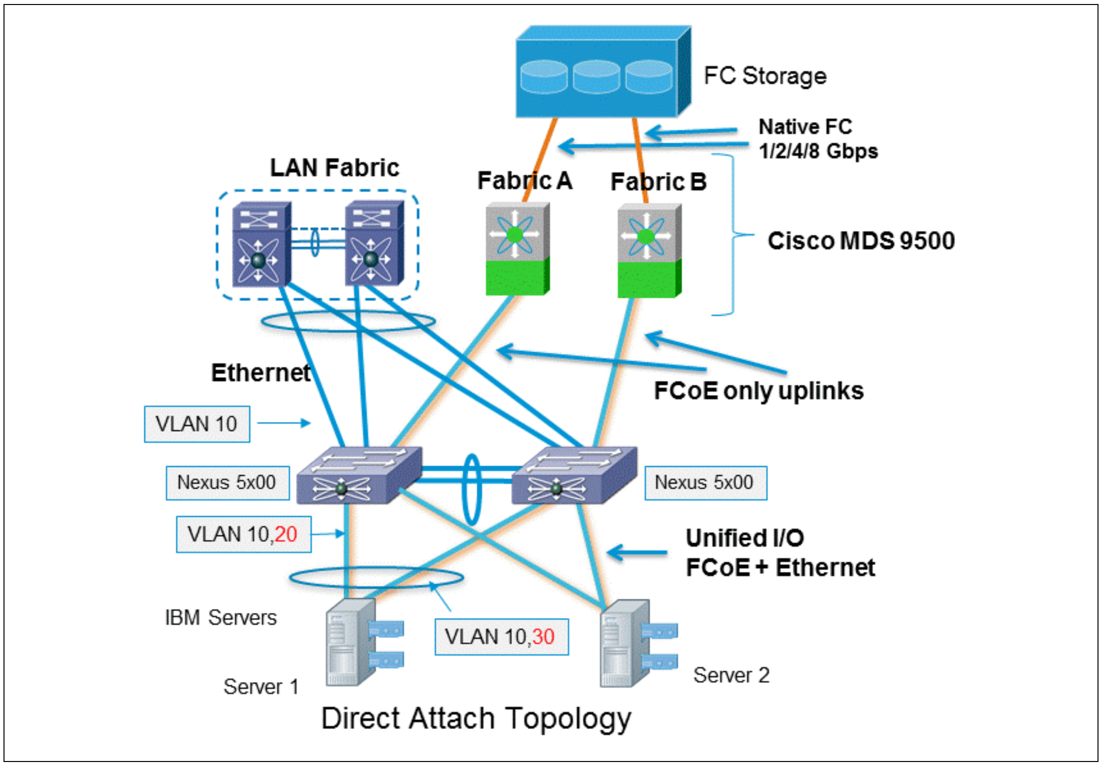


Figure 7-21 FCoE at the access layer with Nexus 5x00 and FCoE uplink to the SAN

7.7.5 Topology 3: FCoE at the access layer with Nexus 7000 and FCoE uplink to the SAN

This topology is almost the same as topology 2, except that the Nexus 7000 is used as the converged access layer switch.

This solution is applicable for customers who require director-class redundancy at the access (edge) layer. The use of the Nexus 7000 represents an evolution for customers who currently utilize Director-Class Fibre Channel switches on the edge of their networks as the Nexus 7000 is a Director-Class FCoE switch. More details about the Director-Class capabilities of the Nexus 7000 that are relevant for a converged network are available at:

http://www.cisco.com/en/US/solutions/collateral/ns340/ns517/ns224/ns945/ns1060/white_paper_c11-644817.html

The Nexus 7000 does not have native Fibre-Channel interfaces so the uplink between the Nexus 7000 and the SAN Core switches utilizes Fibre Channel over Ethernet (FCoE). Because there is no interoperability between Cisco and Brocade FCoE switches, then the SAN Core switches must be Cisco MDS 9500s in this topology.

The Cisco MDS 9500 will use the MDS 10-Gbps 8-Port FCoE Module (DS-X9708-K9) to support the connection between the Nexus 7000 and the MDS. The Nexus 7000 requires the F1 (N7K-F132XP-15), F2 (N7K-F248XP-25), or F2E (N7K-F248XP-25E) linecard on the Nexus 7000 side.

Figure 7-22 shows FCoE at the access layer with Nexus 7000 and FCoE uplinks to the SAN that is based on Cisco MDS 9500.

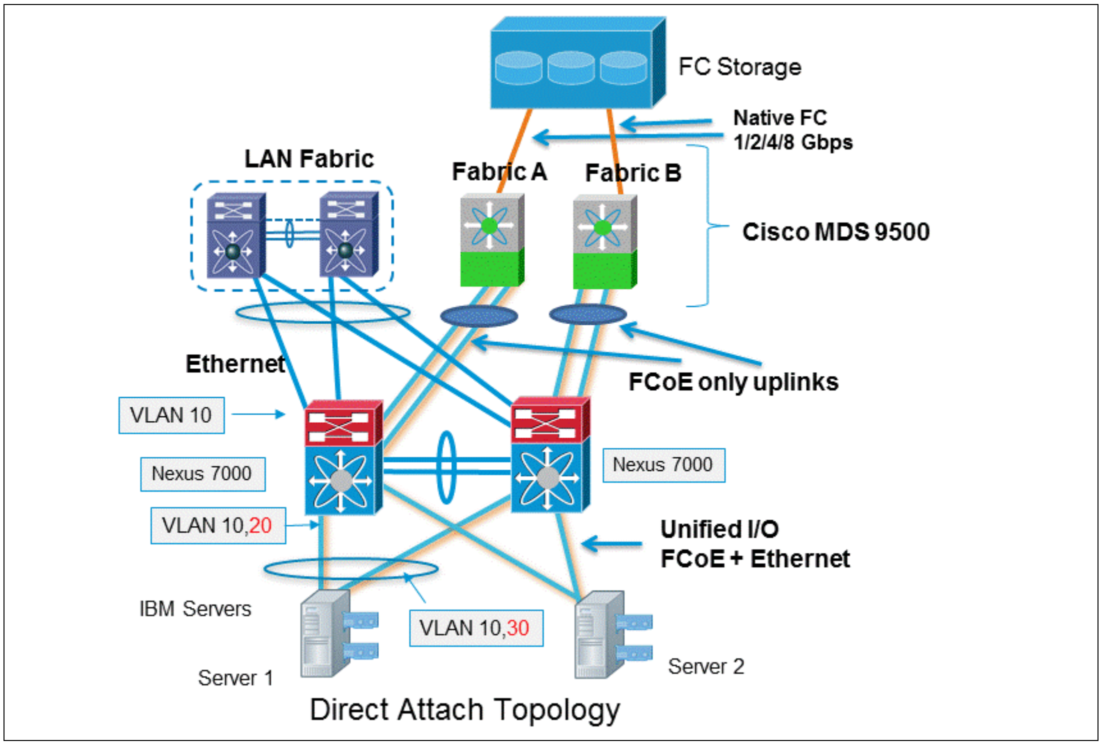


Figure 7-22 FCoE at the access layer with Nexus 7000 and FCoE uplink to the SAN

When the Fibre Channel over Ethernet (FCoE) frame arrives on the Nexus 7000, the FCoE frames are processed within the Storage VDC, as described in 2.14.2, “Nexus 7000” on page 95. They are then forwarded with the FCoE encapsulation to the MDS via the FCoE uplinks. The MDS will then remove the FCoE encapsulation before sending the native Fibre Channel frame to the attached FC storage systems. This FCoE connection between the Nexus 7000 and the MDS only transports FCoE frames; it does not transport IP Ethernet or any non-FCoE frames. Up to 16 FCoE links can be placed in a PortChannel to increase the aggregate bandwidth available between the Nexus 7000 and the MDS switches.

7.7.6 Topology 4: FCoE at the access layer with Nexus 5x00 + single-attached Nexus 2232PP/2248PQ and FC uplink to the SAN

Note : This design option will be described using the Nexus 2232PP as the FEX, but the Nexus 2248PQ can also be used in place of the Nexus 2232PP in this design. The same design concepts apply with the Nexus 2248PQ. The 2248PQ does provide additional 10 GE FCoE port density (48 host ports) over the 2232PP (32 host ports). In addition, the 2248PQ has twice the uplink bandwidth (160G) of the 2232PP (80G).

This topology is almost the same as topology 1: FCoE at the access layer with Nexus 5x00 and FC uplink to the SAN, the difference is the introduction of the Nexus 2232PP between the Servers and the Nexus 5x00.

In this topology, the Nexus 2232PP is single-attached to one Nexus 5x00, as shown in Figure 7-23, with the Nexus 2232PP on the left connected to the Nexus 5x00 on the left, and the Nexus 2232PP on the right connected to the Nexus 5x00 on the right.

As described in Chapter 2, “IBM and Cisco building blocks for the data center” on page 29**,** the Nexus 2232PP is a Fabric Extender and not an independent switch so it is not managed separately. Therefore, the introduction of the Nexus 2232PP in this topology does not add any extra device that needs to be managed.

When connecting the servers to the Nexus 2232PP, one area that requires careful consideration is over-subscription. The Nexus 2232PP has 32 1/10G ports that can be used to connect the servers and 8 x 10 GE ports for uplinks to the Nexus 5x000. If all 32 host ports are used and all 8 uplink ports are connected to the upstream Nexus 5x00 then there is a 4:1 oversubscription ratio between the Nexus 2232 and the Nexus 5x00 (32x 10G = 320 Gbps for host connectivity divided by 8x 10G = 80 Gbps for uplinks equals a 4:1 oversubscription rate).

Also, considering that up to twenty-four (24) Nexus 2232PP can be connected to a single Nexus 5500, careful consideration must also be given regarding the oversubscription rate on the uplinks between the Nexus 5500 and the Fibre Channel (SAN) Core. As more Nexus 2232PP are added / connected to a Nexus 5x00 then more native Fibre Channel uplinks may be required between the Nexus 5x00 and the Fibre Channel (SAN) core switches to keep an acceptable level of oversubscription.

The maximum distance between the Nexus 2232PP and the Nexus 5x00 when FCoE is used on the servers connected to the Nexus 2232PP is 300 meters.

See Figure 7-23 for a diagram of this topology.

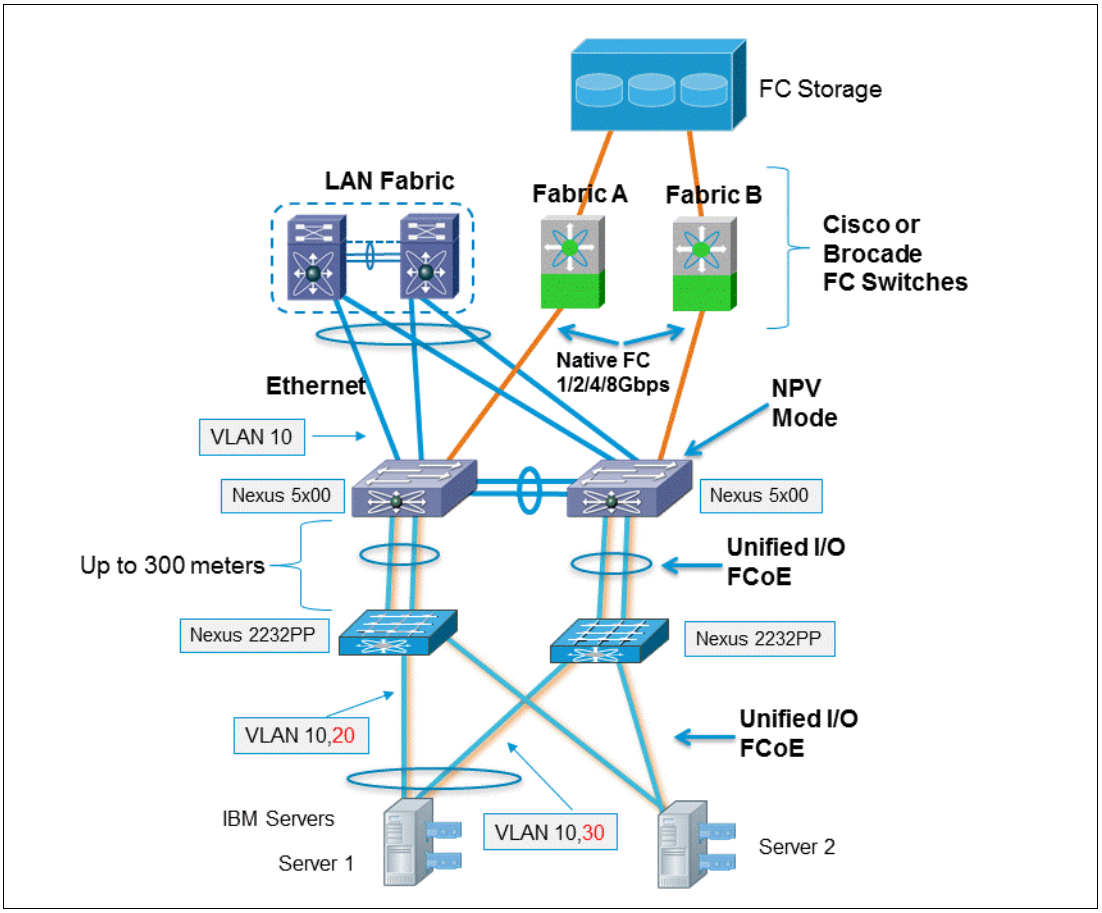


Figure 7-23 FCoE at the access layer with Nexus 5x00 + Single-Attached Nexus 2232PP and FC uplink to the SAN

Figure 7-23 shows FCoE at the access layer with Nexus 5x00 plus Single-Attached Nexus 2232PP and FC uplink to the SAN. The servers connect to the Nexus 2232PP using their CNA's, the Nexus 2232PP on the left is only connected to the Nexus 5x00 on the left (single-attached), similarly Nexus 2232PP on the right connects only to the Nexus 5x00 on the right. The links, up to 8, between the Nexus 2232PP and Nexus 5x00 are converged links transporting Ethernet and FCoE frames.

If we consider Server 1, the FCoE traffic from the port of the CNA that is going to Fabric A (left link from the server) is transported on VLAN 20, marked in Red. The FCoE traffic from the port of the CNA that is going to Fabric B (right link from the server) is transported on VLAN 30, marked in Red as well. VLAN 20 is only present on the left link between the server 1, the Nexus 2232PP and the Nexus 5x00 on the left. VLAN 30 is only present on the right link between server 1, the Nexus 2232PP and the Nexus 5x00 on the right. This way Fabric A and Fabric B are completely separated. The Ethernet traffic is transported on VLAN 10 on this example and is then present on both links from the server to the Nexus 2232PP and Nexus 5x00s.

The servers can connect to the Nexus 2232PP using vPC (Virtual Port-Channel), in this case the Ethernet traffic (Server 1, VLAN 10 on the example) is load balanced between both 10G links, or the servers can connect to the Nexus 2232PP using NIC Teaming, Server 2 in the diagram.

The Nexus 5x00 are configured in NPV mode and have native Fibre-Channel (1/2/4/8Gbps) uplinks to the FC SAN Core. Because the Nexus 5x00 is running in NPV mode the FC core switches can be Cisco MDS or Brocade as long as they are configured with NPIV.

The connection between the Nexus 5x00 and LAN aggregation layer is a regular Ethernet connection and can utilize vPC as shown on the diagram or it can be based on spanning-tree. The Data Center LAN aggregation layer can be based on Cisco Nexus or any other Ethernet switch.

7.7.7 Topology 5: FCoE at the access layer with Nexus 5x00/6000 + single-attached Nexus 2232PP/2248PQ and FCoE uplink to the SAN

Note : This design option will be described using the Nexus 5000, but the Nexus 6000 can also be used in place of the Nexus 5000 in this design. The same design concepts apply with the Nexus 6000, as it supports FCoE on all of its 10G/40G Ethernet ports. The FEX in this design option will be described using the Nexus 2232PP, but the Nexus 2248PQ can also be used in place of the Nexus 2232PP in this design. The same design concepts apply with the Nexus 2248PQ. The 2248PQ does provide additional 10 GE FCoE port density (48 host ports) over the 2232PP (32 host ports). In addition, the 2248PQ has twice the uplink bandwidth (160G) of the 2232PP (80G).

This topology is very similar to Topology 4, with the only difference being that the uplink between the Nexus 5x00 and the SAN Core switches utilizes Fibre Channel over Ethernet (FCoE) instead of native Fibre Channel. Because there is no interoperability between Cisco and Brocade FCoE switches, the SAN Core switches must be Cisco MDS 9500s, 9250is or 9710s when the 48 port 10G FCoE card is supported.

The description under 7.7.4, “Topology 2: FCoE at the access layer with Nexus 5x00/6000 and FCoE uplink to the SAN” on page 397 where it is detailed what is required to connect the Nexus 5x00 to the MDS using FCoE is equally applicable here.

Figure 7-24 shows FCoE at the access layer with Nexus 5x00 + Single-Attached Nexus 2232PP and FCoE uplink to the SAN.

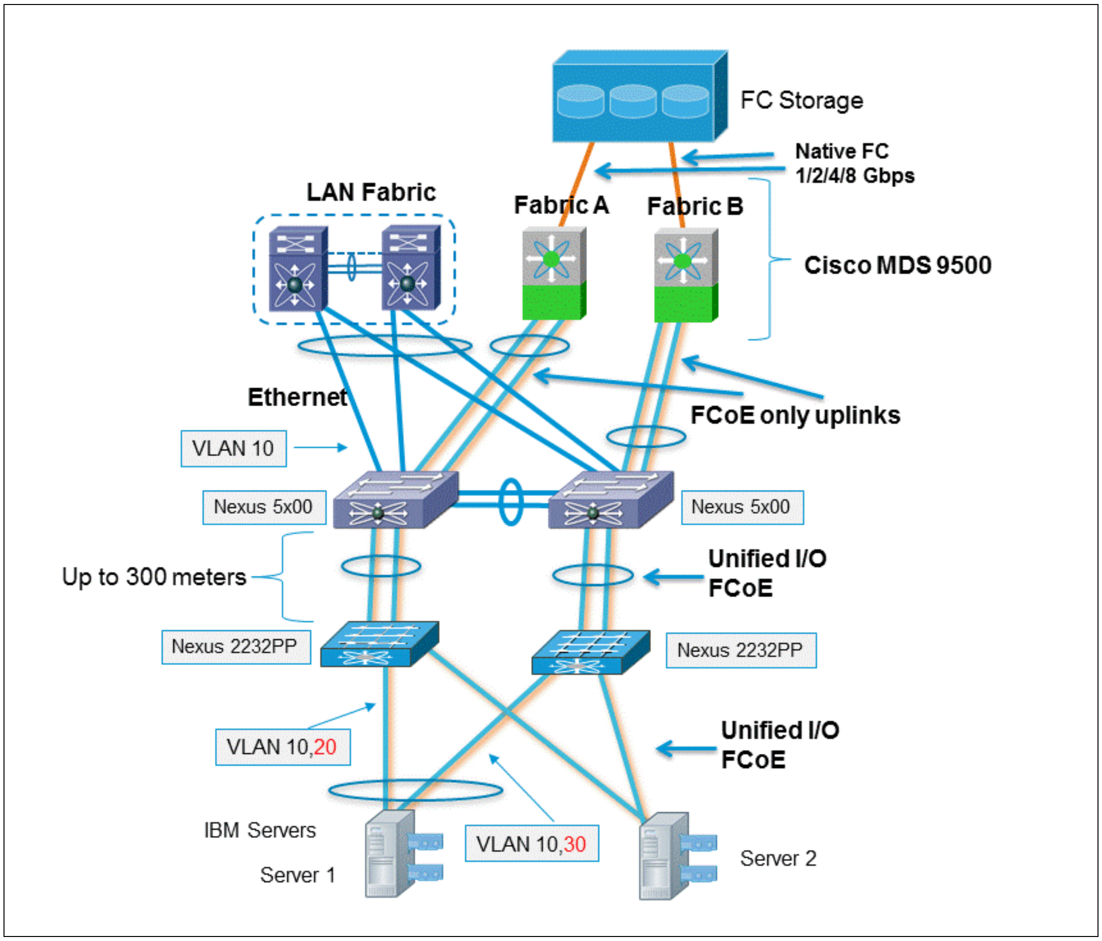


Figure 7-24 FCoE at the access layer with Nexus 5x00 + Single-Attached Nexus 2232PP and FCoE uplink to the SAN

7.7.8 Topology 6: FCoE at the access layer with Nexus 5x00 + dual-homed Nexus 2232PP/2248PQ and FC uplink to the SAN

Note : This design option will be described using the Nexus 2232PP as the FEX, but the Nexus 2248PQ can also be used in place of the Nexus 2232PP in this design. The same design concepts apply with the Nexus 2248PQ. The 2248PQ does provide additional 10 GE FCoE port density (48 host ports) over the 2232PP (32 host ports). In addition, the 2248PQ has twice the uplink bandwidth (160G) of the 2232PP (80G).

This topology is similar to Topology 5, however starting from Cisco NX-OS 5.1(3)N1(1) for the Nexus 5x00, you can use FCoE in conjunction with dual-homed fabric extenders.

To preserve the Fabric A and Fabric B separation required for Fibre Channel, each fabric extender is configured as belonging to either Fabric A or Fabric B.

Figure 7-25 illustrates how FCoE works in conjunction with a fabric extender Dual-homed (active-active) to both Nexus 5x00. Each CNA port connects to a fabric extender that belongs to a different fabric, A or B – see the configuration example below to understand how you assign a fabric extender to Fabric A or B. Even if the fabric extender is dual-homed, the FCoE traffic from a given fabric extender goes only to the Nexus 5500 platform associated with the SAN fabric to which the fabric extender belongs.

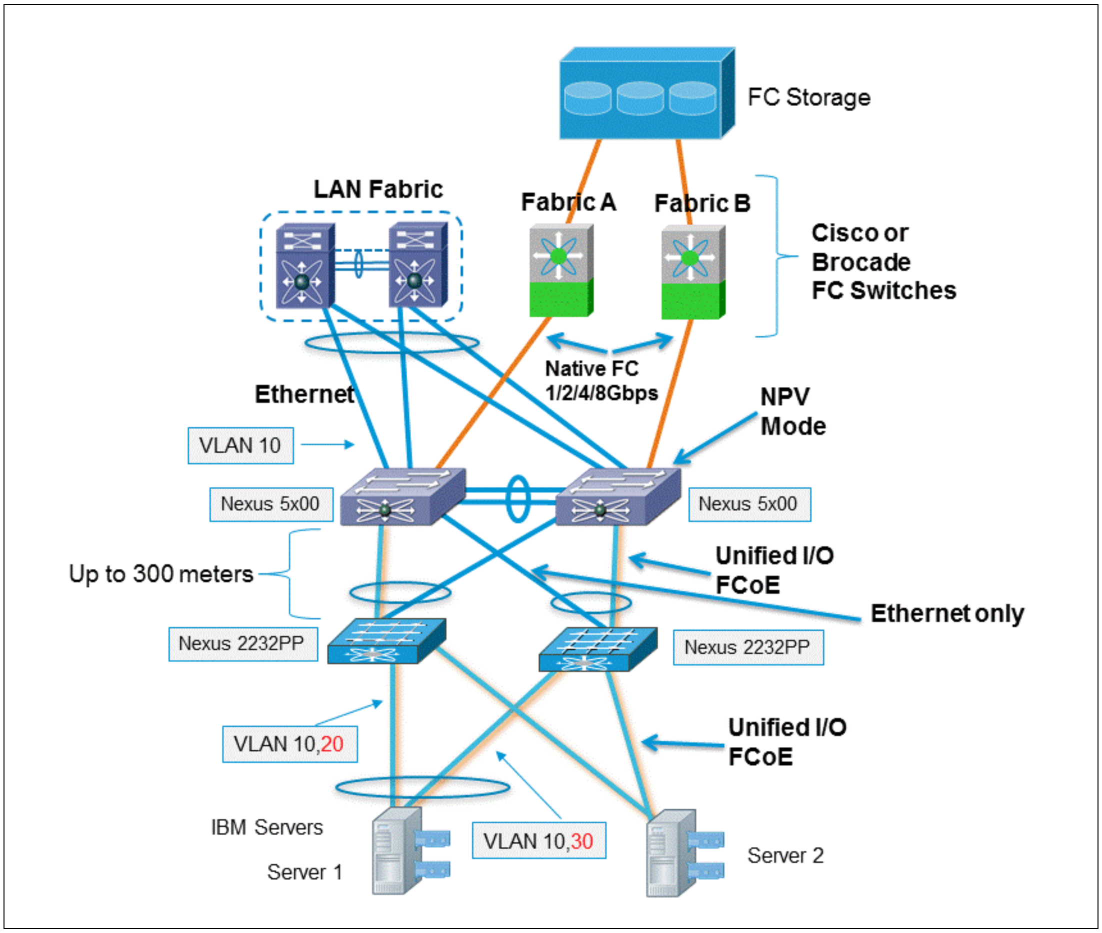


Figure 7-25 FCoE at the access layer with Nexus 5x00 + Dual-Homed Nexus 2232PP and FC uplink to the SAN

For the fabric extender to know the fabric to which it belongs, the network administrator must configure the keyword **fcoe** under the fabric extender configuration of the Cisco Nexus 5500 platform that is associated with the desired SAN fabric.

For instance, if the left fabric extender in Figure 7-25 on page 404 is FEX 101, the configuration of FEX 101 on the left Cisco Nexus 5500 platform will be as follows:

```
fex 101
fcoe
```

The FEX 101 configuration on the Cisco Nexus 5500 platform on the right will not include the fcoe keyword.

Because of the configuration above, then FCoE traffic received on the Nexus 2232PP on the left (FEX 101 on this example) will only be sent on the uplinks connected to the Nexus 5500 on the left. The links between the Nexus 2232PP on the left and the Nexus 5500 on the right are only used for non-FCoE traffic.

One of the drawbacks of this topology is that the fabric links connecting from the fabric extenders to the two upstream Nexus 5500 switches may have different traffic loads since each fabric extender will use only half the links for FCoE, but all the links for regular Ethernet traffic. So in the topology above the link between the Nexus 2232 on the left and the Nexus 5500 on the left will be used for FCoE and non-FCoE traffic while the uplink from the Nexus 2232 on the left and the Nexus 5500 on the right will be used for non-FCoE traffic only. FCoE will never go between Nexus 2232 on the left and Nexus 5500 on the right, thus preserving Fabric A and Fabric B separation.

7.7.9 Topology 7: FCoE at the access layer with Nexus 5x00/6000 + dual-homed Nexus 2232PP/2248PQ and FCoE uplink to the SAN

Note : This design option will be described using the Nexus 5000, but the Nexus 6000 can also be used in place of the Nexus 5000 in this design. The same design concepts apply with the Nexus 6000, as it supports FCoE on all of its 10G/40G Ethernet ports. The FEX in this design option will be described using the Nexus 2232PP, but the Nexus 2248PQ can also be used in place of the Nexus 2232PP in this design. The same design concepts apply with the Nexus 2248PQ. The 2248PQ does provide additional 10 GE FCoE port density (48 host ports) over the 2232PP (32 host ports). In addition, the 2248PQ has twice the uplink bandwidth (160G) of the 2232PP (80G).

This topology is very similar to Topology 6, with the only difference being that the uplink between the Nexus 5x00 and the SAN Core switches utilizes Fibre Channel over Ethernet (FCoE) instead of native Fibre Channel. Because there is no interoperability between Cisco and Brocade FCoE switches, the SAN Core switches must be Cisco MDS 9500s, 9250is or 9710s when the 48 port 10G FCoE card is supported.

The description under 7.7.4, “Topology 2: FCoE at the access layer with Nexus 5x00/6000 and FCoE uplink to the SAN” on page 397 where it is detailed what is required to connect to the Nexus 5x00 to the MDS using FCoE is equally applicable here.

Figure 7-26 shows FCoE at the access layer with Nexus 5x00 + Dual-Homed Nexus 2232PP and FCoE uplink to the SAN.

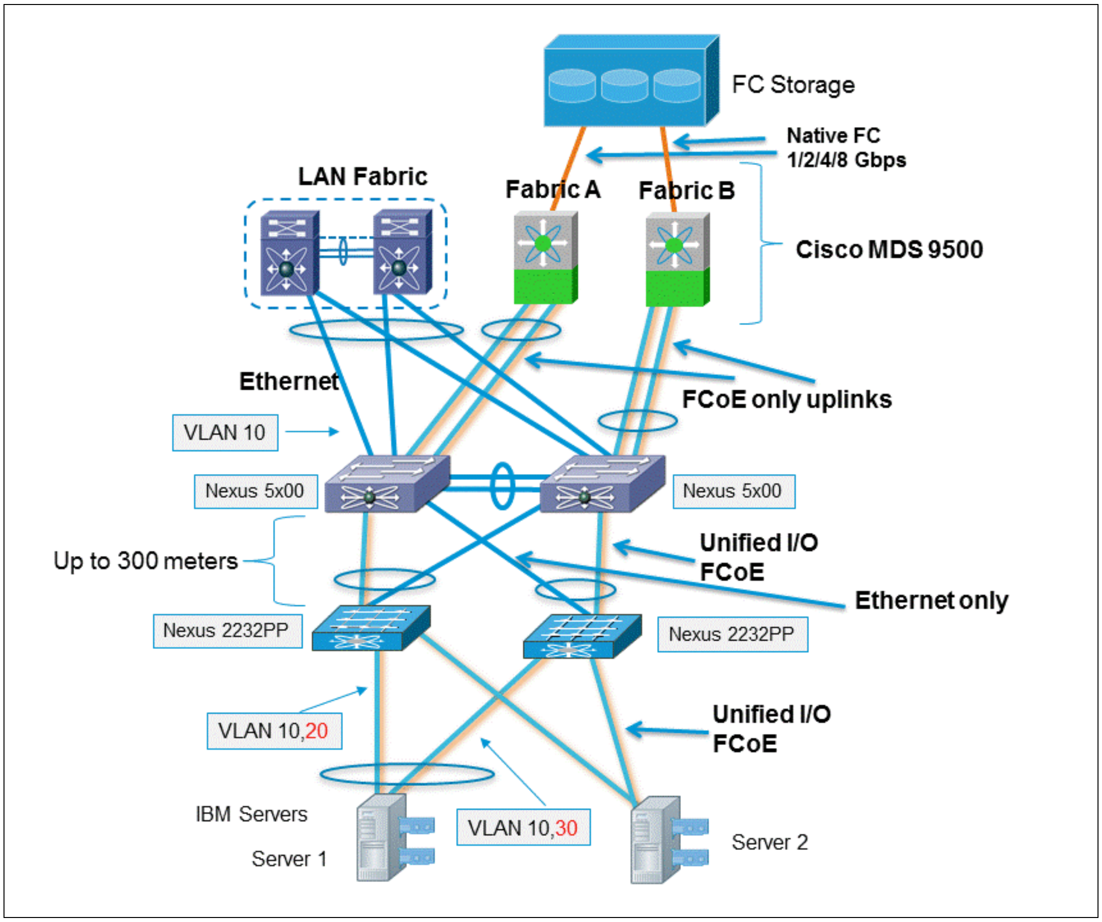


Figure 7-26 FCoE at the access layer with Nexus 5x00 + Dual-Homed Nexus 2232PP and FCoE uplink to the SAN

7.7.10 Topology 8: FCoE for the IBM BladeCenter with Nexus 4000 + Nexus 5x00 and FC uplink to the SAN

The Nexus 4000 is a lossless Ethernet switch that is capable of sending FCoE traffic. It supports FCoE on its interfaces facing the IBM blade servers (internal interfaces) as well as on its uplink interfaces (external interfaces). The Nexus 4000 Series blade switch is supported by IBM BladeCenter models BCH and BCH-T.

Since the Nexus 4000 is not an FCF, the CNA's in the blade servers will not do a fabric login on the Nexus 4000. This will require forwarding the login process for those CNA's to an actual FCF. The Nexus 5x00 function as the FCFs on this topology.

This topology shows that the IBM BladeServers can utilize FCoE between the servers' CNAs and the Nexus 4000 server facing ports. FCoE frames are sent on the uplink interfaces of the Nexus 4000 to the Nexus 5x00. The Nexus 5x00 receives the FCoE frame, removes the FCoE encapsulation and forwards the native Fibre Channel frame on the Fibre Channel uplinks towards the SAN core.

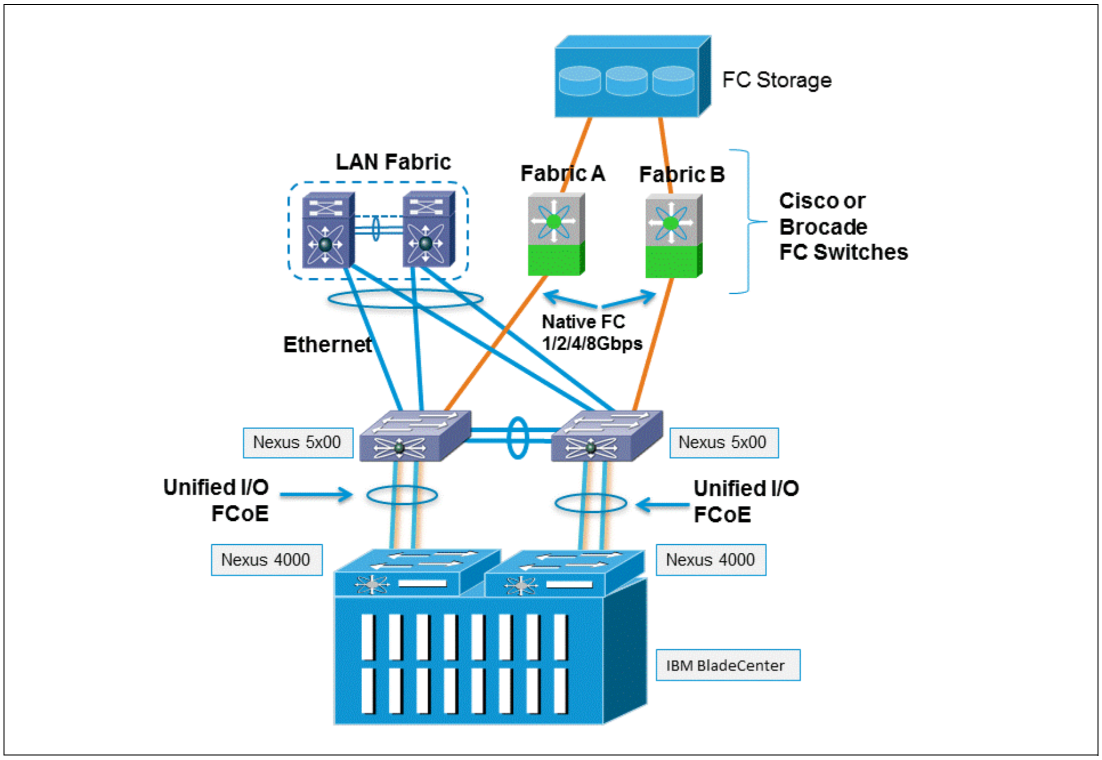


Figure 7-27 FCoE for the IBM BladeCenter with Nexus 4000 + Nexus 5x00 and FC uplink to the SAN

For FCoE to maintain the dual storage area network (SAN) fabrics on the topology above, the Nexus 4000 uplinks are linked to a single Nexus 5x00. This type of topology will allow both FCoE and Ethernet traffic to flow through the same physical links. Figure 7-27 depicts the supported FCoE topology. The Ethernet links between the Nexus 5x00 switches do not allow FCoE traffic. This ensures the separation of the Fabric A and Fabric B fabrics.

For more information on this topology, see the Cisco Nexus 4000 Series Design Guide at:

http://www.cisco.com/en/US/partner/prod/collateral/switches/ps9441/ps10596/deployment_guide_c07-574724.html

7.7.11 Topology 9: FCoE for the IBM BladeCenter with Nexus 4000 + Nexus 5x00/6000 and FCoE uplink to the SAN

Note : This design option will be described using the Nexus 5000, but the Nexus 6000 can also be used in place of the Nexus 5000 in this design. The same design concepts apply with the Nexus 6000, as it supports FCoE on all of its 10G/40G Ethernet ports.

This topology is very similar to Topology 8, with the only difference being that the uplink between the Nexus 5x00 and the SAN Core switches utilizes Fibre Channel over Ethernet (FCoE) instead of native Fibre Channel. Because there is no interoperability between Cisco and Brocade FCoE switches, the SAN Core switches must be Cisco MDS 9500s, 9250is or 9710s when the 48 port 10G FCoE card is supported. The description under 7.7.4, “Topology 2: FCoE at the access layer with Nexus 5x00/6000 and FCoE uplink to the SAN” on page 397 where it is detailed what is required to connect the Nexus 5x00 to the MDS using FCoE and is equally applicable here.

Figure 7-28 shows FCoE for the IBM BladeCenter with Nexus 4000 + Nexus 5x00 and FCoE uplink to the SAN.

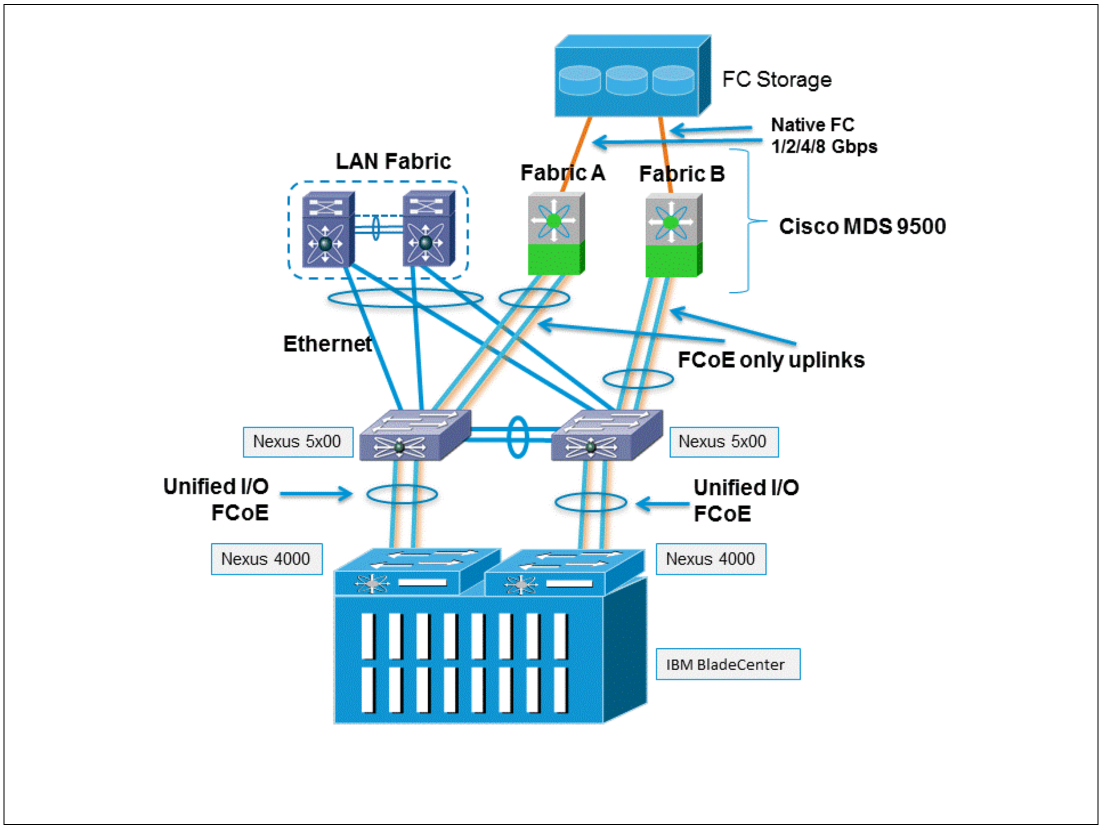


Figure 7-28 FCoE for the IBM BladeCenter with Nexus 4000 + Nexus 5x00 and FCoE uplink to the SAN

7.7.12 Topology 10: FCoE for the IBM BladeCenter with Nexus 4000 + Nexus 7000 and FCoE uplink to the SAN

This topology is very similar to Topology 9, except that the Nexus 4000 is connected to a Nexus 7000 instead of the Nexus 5x00.

This solution is applicable for customers who require director-class redundancy on the device concentrating the FCoE uplinks coming from the Nexus 4000 installed on the IBM BladeCenters. This could be desirable, particularly if multiple BladeCenter chassis connect to a single Nexus 7000 for accessing Fabric A and then to another Nexus 7000 for accessing Fabric B.

More details about the Director-Class capabilities of the Nexus 7000 that are relevant for a converged network are available at:

http://www.cisco.com/en/US/solutions/collateral/ns340/ns517/ns224/ns945/ns1060/white_paper_c11-644817.html

The Nexus 7000 does not have native Fibre Channel interfaces so the uplink between the Nexus 7000 and the SAN Core switches utilizes Fibre Channel over Ethernet (FCoE). Because there is no interoperability between Cisco and Brocade FCoE switches, the SAN Core switches must be Cisco MDS 9500s, 9250is or 9710s when the 48 port 10G FCoE card is supported.

Figure 7-29 shows FCoE for the IBM BladeCenter with Nexus 4000 + Nexus 7000 and FCoE uplink to the SAN.

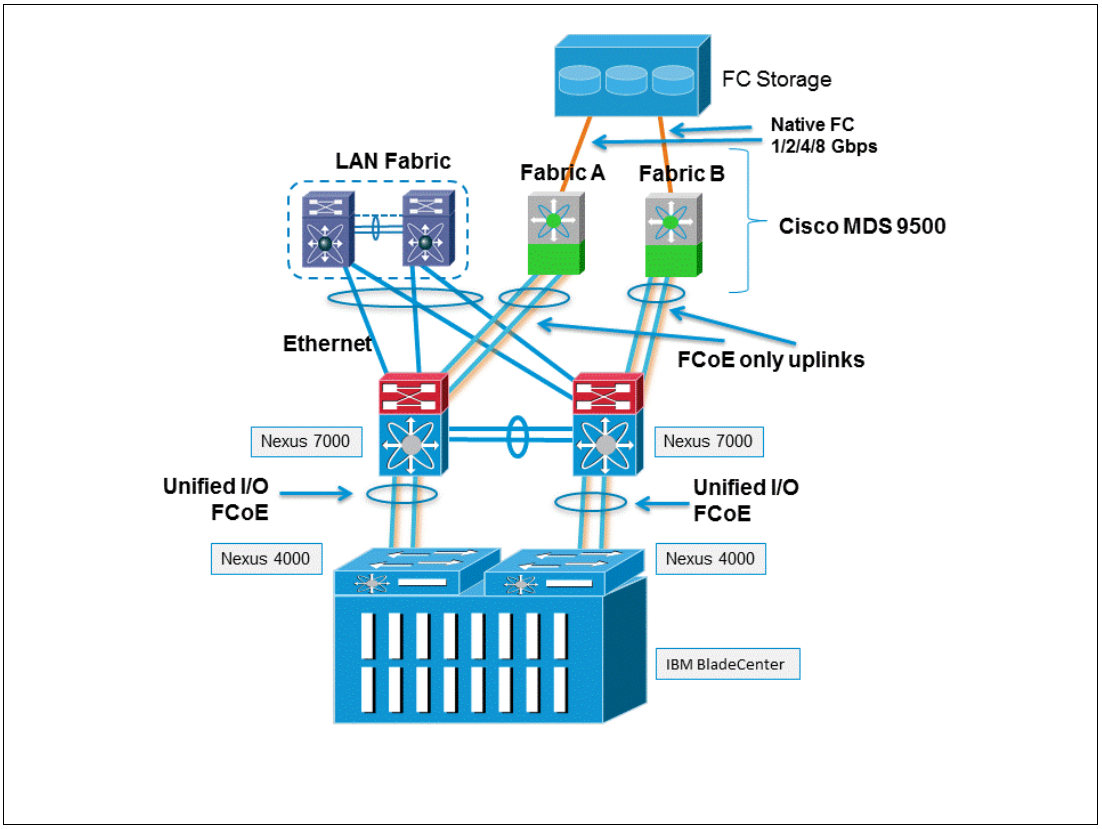


Figure 7-29 FCoE for the IBM BladeCenter with Nexus 4000 + Nexus 7000 and FCoE uplink to the SAN

7.7.13 Topology 11: FCoE beyond the access layer with multi-hop FCoE

With the introduction of multi-hop Fibre Channel over Ethernet (FCoE), network convergence is no longer restricted to the first-hop access layer. Multi-hop FCoE helps extend the flexibility and scalability of convergence further into the data center while preserving investments in Fibre Channel SANs.

A variety of multi-hop FCoE deployment options (topologies) are available. Two of the most common topologies that extend convergence beyond the access layer are shown in Figure 7-30 on page 410:

- On the left, the Nexus 7000 is used on the Access (edge) layer; Nexus 7000 is also used on the Aggregation Layer and MDS 9250i, 9500 or 9700 is used in the SAN Core;
- On the right, the Nexus 5x00 is used on the Access (edge) layer; Nexus 7000 is used on the Aggregation Layer and MDS 9250i, 9500 or 9700 is used in the SAN Core;

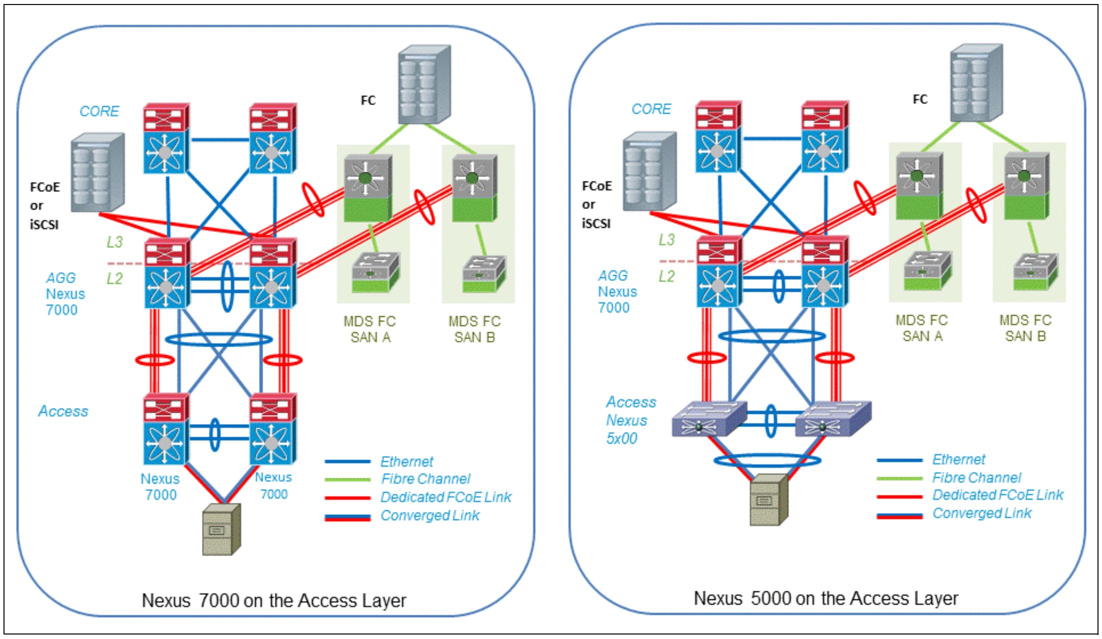


Figure 7-30 Common Multi-hop FCoE Deployment Scenarios.

At the time of this writing, the Nexus 6000 platform was just being introduced. Since the Nexus 6000 supports both 10G FCoCe and 40G FCoE, it has a variety of potential use cases for converged networks. As mentioned throughout this chapter, the Nexus 6000 can be inserted in place of the Nexus 5000 platform as an FCoE access layer switch, providing the ISLs to the SAN are 10G (or 40G) FCoE. This applies to both Single-hop and Multi-hop FCoE deployments. In addition to the FCoE access layer use cases, the Nexus 6000 can also be used as an FCoE aggregation layer platform (similar to the Nexus 7000) in Multi-hop FCoE deployments, providing the ISLs to the SAN are 10G or 40G FCoE. In the future, native FC interface support will be added to the Nexus 6004 platform via an expansion module.

Some of the benefits of converging beyond the access layer include these:

- Higher overall throughput and port density compared to native Fibre Channel switches.
- Flexibility to support file (network-attached storage [NAS]) and block (Small Computer System Interface over IP [iSCSI], FCoE and FC) storage traffic on the same infrastructure.
- Evolution to 40 and 100 Gigabit Ethernet that are capable of transporting NAS, iSCSI and FCoE at much higher rates than native FC SANs.

For more details on the multi-hop FCoE Designs, see these websites:

- Multi-hop Fibre Channel over Ethernet Designs:

  http://www.cisco.com/en/US/prod/collateral/modules/ps5991/ps11581/at_a_glance_c45-648467.pdf

- Director-Class FCoE: Converge the Network with Cisco Nexus 7000 Series switches and Cisco MDS 9500 Series Multilayer Directors:

  http://www.cisco.com/en/US/partner/solutions/collateral/ns340/ns517/ns224/ns945/ns1060/guide_c07-648629_ps5991_Products_White_Paper.html

- Initial Findings: Evaluating Multi-hop FCoE:

  http://www.cisco.com/en/US/partner/docs/solutions/Enterprise/Data_Center/VMDC/tech_eval/mFCoEwp.html

7.7.14 FCoE maximum supported distances

There are certain maximum supported distances between two switches transporting FCoE. Those distances depend on the transceiver being used as well as on the amount of buffer available on the switch port.

Figure 7-31 shows the most common connections and their maximum distance.

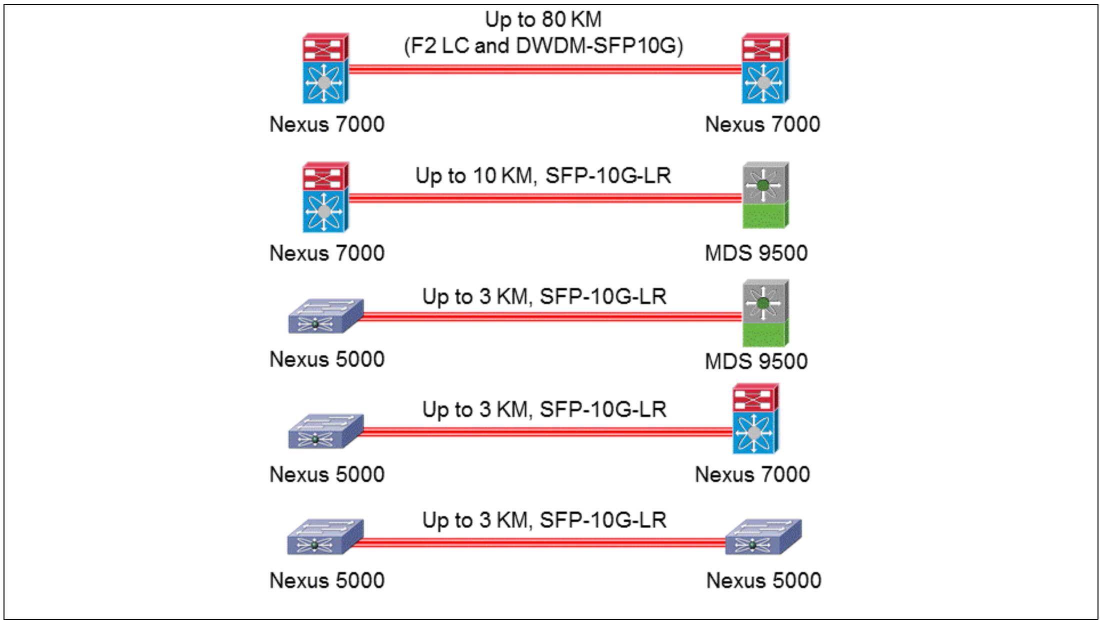


Figure 7-31 The most common connections and their maximum distance.

The maximum distance of the connection between two switches is based on the switch that supports the shortest distance:

- The Nexus 5000 maximum supported distance for FCoE is 3KM.
- The maximum distance supported by the Nexus 7000 is 80KM.
- The maximum distance supported by the MDS 9500 is 10KM.

Based on these, for example, when a Nexus 5000 connects to a Nexus 7000 the maximum supported distance is 3KM because the Nexus 5000 maximum distance of 3KM becomes the limiting factor.

7.7.15 Converged Network Adapters

Converged Network Adapters (CNAs) provide an opportunity to reduce data center costs by converging data and storage networking. Standard TCP/IP and Fibre Channel (FC) traffic can both run on the same high-speed 10 Gbps Ethernet wire, resulting in cost savings through reduced adapter, switch, cabling, power, cooling and management requirements.

Figure 7-32 shows normal and converged adapters.

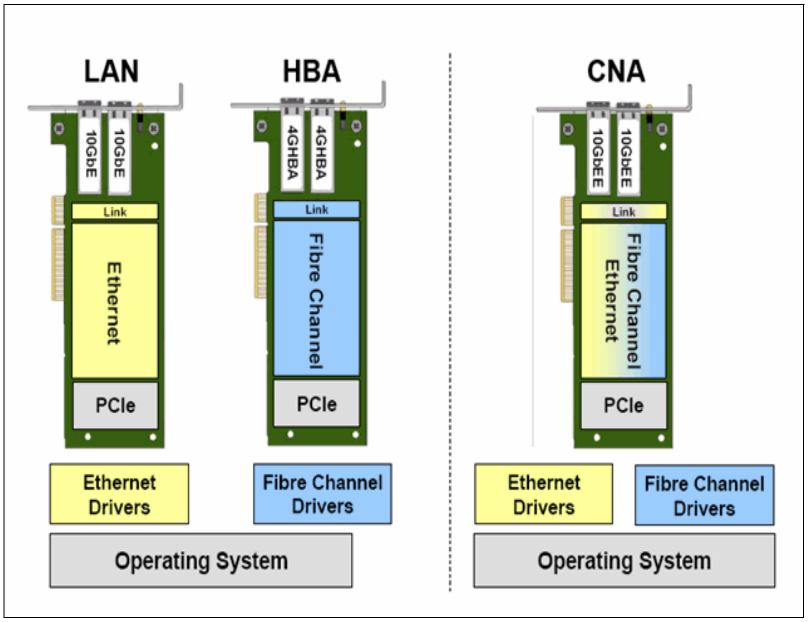


Figure 7-32 Adapters

Today, multiple CNAs are available, with more becoming available over time. IBM currently supports Brocade, Emulex, and QLogic CNAs. IBM also supports Broadcom, Intel, and Mellanox 10 Gbps Ethernet adapters that can be used for Internet Small Computer Systems (iSCSI). For a full list of supported CNAs and other network adapters, see the following references:

- IBM System x Configuration and Options Guide

  http://www.ibm.com/systems/xbc/cog/

- *IBM BladeCenter Interoperability Guide*

  http://www.ibm.com/support/entry/portal/docdisplay?brand=5000008&lndocid=MIGR-5073016

- IBM Server Proven: Compatibility for hardware, applications, and middleware

  http://www.ibm.com/systems/info/x86servers/serverproven/compat/us/

Figure 7-33 shows FC and FCoE port differences.

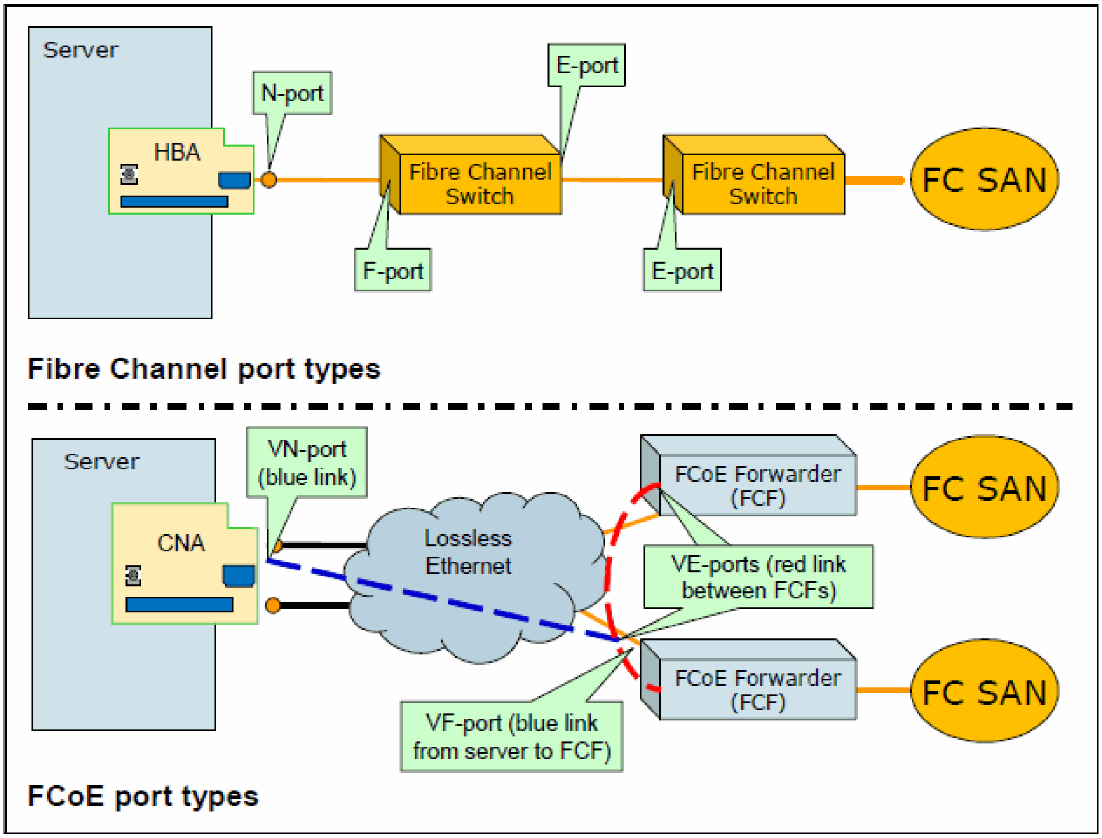


Figure 7-33 Difference between FC port types and FCoE port types

7.7.16 Emulex 10GbE Virtual Fabric Adapter and Emulex 10GbE Virtual Fabric Adapter II for IBM BladeCenter

The existing IBM BladeCenter Virtual Fabric portfolio includes support for the following adapters:

- Emulex 10GbE Virtual Fabric Adapter
- Emulex 10GbE Virtual Fabric Adapter Advanced
- Emulex 10GbE Virtual Fabric Adapter II
- Emulex 10GbE Virtual Fabric Adapter Advanced II

By using these adapters, IBM clients can simplify their I/O infrastructure by reducing the number of switches needed inside the chassis. The Emulex 10GbE Virtual Fabric Adapter II is a dual-port 10 Gbps Ethernet card that supports 1-Gbps or 10-Gbps traffic, or up to eight virtual network interface card (NIC) devices.

Figure 7-34 shows an Emulex 10GbE Virtual Fabric Adapter.

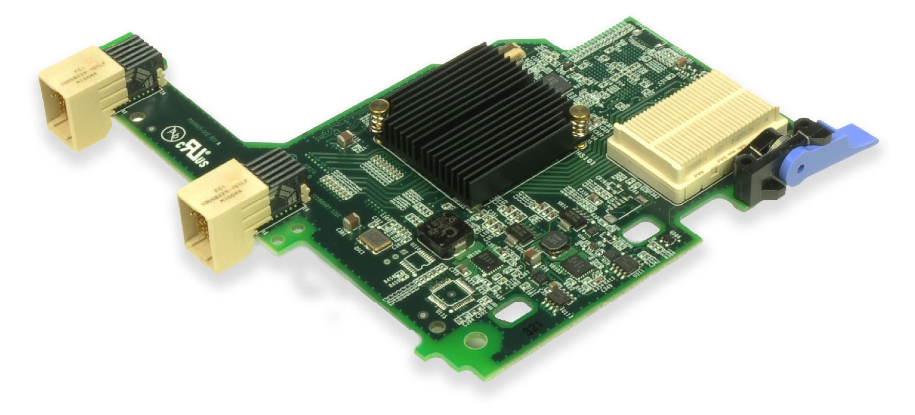


Figure 7-34 Emulex 10GbE Virtual Fabric Adapter II for IBM BladeCenter

The Emulex 10GbE Virtual Fabric Adapter and Emulex 10GbE Virtual Fabric Adapter II provide two physical Ethernet ports, supporting 1 Gbps and 10 Gbps traffic. Each physical 10-Gbps port can be divided into four virtual ports with bandwidth allocation in 100 Mbps increments to a maximum of 10 Gbps per physical port.

The Emulex 10GbE Virtual Fabric Adapter Advanced and Emulex 10GbE Virtual Fabric Adapter Advanced II offer network functions, pNIC and vNIC functions, and iSCSI or FCoE. They can also be used as a simple NIC card. The adapters must be configured for iSCSI or FCoE. You cannot use both iSCSI and FCoE at the same time unless software iSCSI is involved. The Emulex 10GbE Virtual Fabric Adapter provides an iSCSI hardware initiator.

The Emulex 10GbE Virtual Fabric Adapter and Emulex 10GbE Virtual Fabric Adapter II can also be updated by applying a license.

For more information, see the following IBM Redbooks Product Guides:

- *Emulex 10GbE Virtual Fabric Adapter and Virtual Fabric Adapter Advanced for IBM BladeCenter*, TIPS0748
- *Emulex 10GbE Virtual Fabric Adapter II for IBM BladeCenter*, TIPS0828

7.7.17 QLogic 2-port 10Gb Converged Network Adapter (CFFh) for IBM BladeCenter

The QLogic 2-port 10Gb Converged Network Adapter (CFFh) for IBM BladeCenter offers robust 8-Gbps Fibre Channel (FC) storage connectivity and 10-Gbps networking over a single Converged Enhanced Ethernet (CEE) link. This adapter combines the functions of a NIC and a host bus adapter (HBA) on a single converged adapter. Therefore, clients can realize potential benefits in cost, power and cooling, and data center footprint by deploying less hardware.

Figure 7-35 shows the QLogic 2-port 10Gb Converged Network Adapter (CFFh). The card offers two network ports and two FCoE ports. To perform iSCSI, you must use a software initiator.

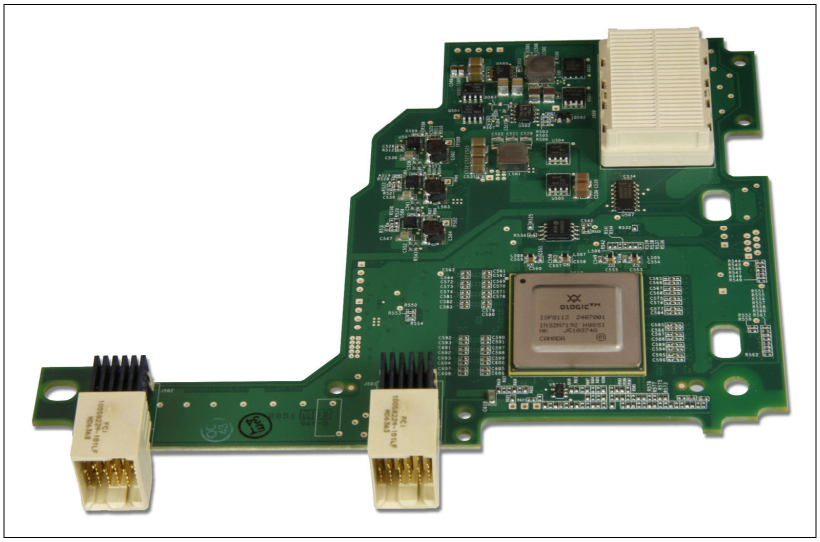


Figure 7-35 The QLogic 2-port 10 Gb Converged Network Adapter (CFFh)

For more information, see the IBM Redbooks publication Product Guide, *QLogic 2-port 10Gb Converged Network Adapter (CFFh) for IBM BladeCenter*, TIPS0716.

7.7.18 Brocade 2-port 10GbE Converged Network Adapter for IBM BladeCenter

The Brocade 2-port 10GbE Converged Network Adapter with the Brocade Converged 10GbE Switch Module is part of a converged Ethernet solution for IBM BladeCenter. It offers native Ethernet and FC connectivity, maximum bandwidth and performance, and simplicity in a converged environment. This card combines the functions of an HBA and NIC on one PCIe 2.0 x8 card. This expansion card supports full FCoE protocol offload and allows FC traffic to converge onto 10 Gbps CEE networks. FCoE and 10-Gbps CEE operations run simultaneously.

Figure 7-36 shows the Brocade 2-port 10GbE Converged Network Adapter. The card offers two network ports and two FCOE ports. To perform iSCSI, you must use a software initiator.

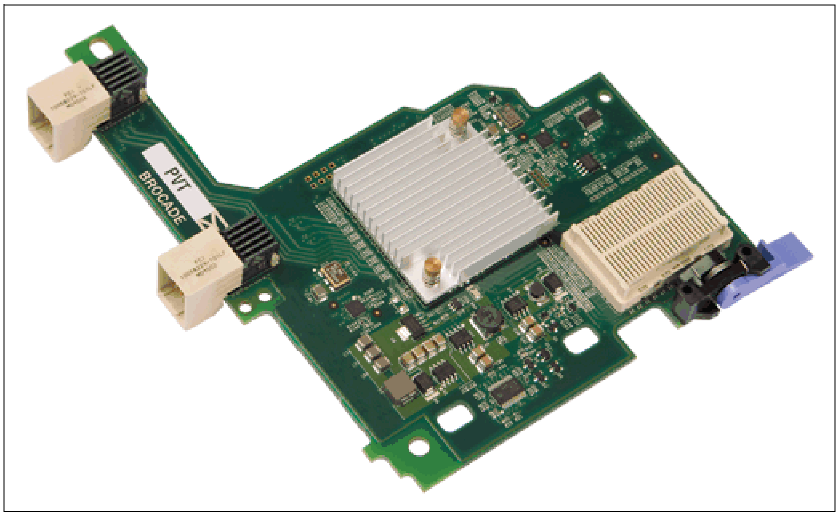


Figure 7-36 Brocade 2-port 10GbE Converged Network Adapter

For more information, see the IBM Redbooks publication Product Guide, *Brocade 2-port 10GbE Converged Network Adapter for IBM BladeCenter*, TIPS0790.

7.7.19 Cisco FCoE networking products

This section describes several FCoE networking products.

Nexus 5500

The Nexus 5548UP switch provides 32 unified ports (support 1GE, 10GE, 10G FCoE, and 1/2/4/8G FC), with expansion to 48 total unified ports through the addition of a 16-port expansion module.

The Nexus 5596UP switch provides 48 unified ports (support 1GE, 10GE, 10G FCoE, and 1/2/4/8G FC), with expansion to 96 total unified ports through the addition of up to three 16-port expansion modules.

Additional information on the Nexus 5500 family switches can be found in Chapter 2, or at the following website:

http://www.cisco.com/go/nexus5000

Nexus 6000

The Nexus 6004 switch provides 48 40 GE ports (each supports 1x40 GE or 4x10 GE with or without FCoE), with expansion to 96 total 40 GE ports (or up to 384 10 GE ports) through the addition of up to four 12-port 40 GE expansion modules.

The Nexus 6001 switch provides 48 1/10 GE ports and 4 x 40 GE uplink ports, each of which can be configured as either a single 40 GE port or 4 x 10 GE ports. All ports in the Nexus 6001 support both native 10/40 GE and FCoE.

Additional information on the Nexus 6000 family switches can be found in Chapter 2, or at the following website:

http://www.cisco.com/go/nexus6000

Nexus 7000

The Nexus 7000 Switch family provides a variety of high density FCoE deployment options in a Director-Class switching platform. FCoE deployment density ranges from 96 10 GE FCoE ports in the Nexus 7004 all the way up to 768 10 GE FCoE ports in the Nexus 7018. Medium density options also exist with the Nexus 7009 and Nexus 7010 switching platforms. The modules in the Nexus 7000 that support FCoE include the F1 (N7K-F132XP-15), F2 (N7K-F248XP-25), and the F2E (N7K-F248XP-25E).

Additional information on the Nexus 7000 family switches can be found in Chapter 2, or at the following website:

http://www.cisco.com/go/nexus7000

Nexus 2000

The Nexus 2000 Series Fabric Extenders (FEX) provide an efficient method of extending the Unified Fabric to the server rack while simplifying cabling and management. The Nexus 2000 Fabric Extenders can be used in conjunction with the Nexus 5500, Nexus 6000, and Nexus 7000 switches for FCoE deployments.

Additional information on the Nexus 2000 family switches can be found in Chapter 2, or at the following website:

http://www.cisco.com/go/nexus2000

Nexus 4000

The Nexus 4001I switch provides a 10 GE FCoE solution for IBM BladeCenter H and BladeCenter HT. The 40001I supports 10 GE connectivity to each of the fourteen blade servers in the BladeCenter chassis, and up to 6 10 GE uplinks to connect to the external network. The Nexus 4001I supports FIP Snooping functionality (FCoE pass-thru module) in conjunction with the Nexus 5500, 6000, or 7000 switches (providing FCF) to provide an FCoE solution to the blade servers.

Additional information on the Nexus 4001I switch can be found in Chapter 2, or at the following website:

http://www.cisco.com/go/nexus4000

7.8 Operational models enabling FCoE

Converging SAN and LAN traffic onto a common infrastructure enables customers to realize significant cost efficiencies through reducing power consumption, cooling costs, adapters, cables, and switches. Most customers would like to derive these benefits and adopt a converged network, but in some cases the shared operational model driven by network convergence can be challenging to adopt and implement successfully. We explore the tools used to create a viable operational model for FCoE deployments below.

7.8.1 Role Based Access Control

Cisco has a very useful tool, Role Based Access Control (RBAC), in the Nexus platforms, which allows multiple administrative domains to be created in a single switch. With the use of RBAC, separate user roles for LAN network team and SAN team can be defined that allows both teams to administer only the portion of the configuration that is relevant to their respective domain. For example, a LAN Admin role can be created that only allows configuration access to non-FCoE VLANs, non-FCoE ports, access-lists among others. While a SAN admin role can be created that only allows configuration access to VSANs, zoning, FC ports.

Using RBAC to provide distinct administrative access policies for each team, although they are sharing common network devices, allows for a very similar administrative experience to having completely separate LAN and SAN infrastructure. However, even with the effectiveness of RBAC, communication and teamwork is required and expected between the LAN and SAN teams for a successful FCoE deployment, as there are several joint administrative tasks to be accomplished. These joint administrative tasks include selecting VLANs to be used for FCoE, defining what ports will run FCoE, defining uplinks (FC or FCoE) to the SAN Core, ongoing firmware upgrades.

RBAC is supported on all Nexus switching platforms. See the following website for more information on NX-OS Role Based Access Control:

http://www.cisco.com/en/US/docs/switches/datacenter/nexus5000/sw/configuration/guide/cli/sec_rbac.html

7.8.2 Nexus 7000 Storage Virtual Device Context

In additional to RBAC support, the Nexus 7000 supports an additional level of administrative isolation for FCoE deployments using its Virtual Device Context (VDC) capability. VDCs (covered more thoroughly in Chapter 9 of this book) provide a device level virtualization of the Nexus 7000 switching platform. Each VDC has its own configuration file, a separate set of access control policies (including separate RBAC for each VDC), and provides hardware level traffic isolation within the switch.

FCoE deployments using the Nexus 7000 platform require the use of a special VDC, called the Storage VDC. The Storage VDC provides a “virtual Director” switch within the Nexus 7000 for the SAN Administrator to own and manage. This allows the SAN Administrator to manage this virtual Director platform in the same manner as they manage an FC Director switch, having complete control of all FC configuration elements (vsans, zoning, FC domain services, FC ports, etc…) from within the Storage VDC.

See the following for more detail on Nexus 7000 VDCs:

http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9402/ps9512/White_Paper_Tech_Overview_Virtual_Device_Contexts.html

Storage VDC on the Nexus 7000 platform builds on the access control isolation provided by NX-OS RBAC, allowing for an even greater level of autonomy for LAN and SAN teams sharing network devices in a converged network fabric.

7.8.3 Licensing requirements for FCoE

On Cisco Nexus 5500 Series switches, FCoE capability is licensed in groups of 8 ports. If a port is to run FCoE or native FC, it must be licensed using the Nexus 5500 8-Port Storage Protocol Services License (N55-8P-SSK9=).

Similar to the Nexus 5500, the Nexus 6000 Series switches also license FCoE in groups of ports, but in groups of 16x10G (or 4x40G)ports rather than groups of 8 ports. The part number for the Storage Protocol license for 16x10G / 4x40G ports is (N6004-4Q-SSK9).

On Cisco Nexus 7000 Series switches, FCoE capability is licensed on a per module basis. For the modules that support FCoE (F1, F2, F2E) one of the following licenses must be ordered if you are running FCoE on ports from that module:

- N7K-FCOEF132XP- FCoE License for Nexus 7000 32-port 10G SFP+ (F1)
- N7K-FCOEF248XP- FCoE License for Nexus 7000 48 port 10G SFP+ (F2/F2E)

No special license is required for FCoE on the MDS 9000 switches.

7.9 Acronyms

Table 7-3 provides a list of acronyms used throughout this chapter.

Table 7-3 Acronyms

DCB	Data Center Bridging	A collection of standards that extend classical Ethernet protocols for use in the data center
DCBX	Data Center Bridging Exchange	Capabilities through which peers can exchange DCB parameters to help ensure lossless Ethernet behavior
FC	Native Fibre Channel Protocol	The dominant protocol for use in the SAN; its deterministic and lossless reliability make it well suited for storage traffic
FCF	Fibre Channel Forwarder	The function in an FCoE-capable device that allows FCoE frames to be switched across multiple hops
ETS	Enhanced transmission selection	A feature defined in the IEEE 802.1Qaz standard; enables a percentage of available bandwidth on a link to be divided among specified priority groups
PFC	Priority Flow Control	A feature defined in the IEEE 802.1Qbb; Priority-based flow control is intended to eliminate frame loss due to congestion
FCID	Fibre Channel ID (24 Bit routable address)	
FCoE	Fibre Channel over Ethernet	
FC-MAP	24 Bit value identifying an individual fabric	
FIP	FCoE initialization protocol	An initialization protocol that enables establishment of point-to-point virtual Fibre Channel links over a multi-access network such as Ethernet
FLOGI	FC fabric login	
FPMA	Fabric provided MAC address	
PLOGI	FC port login	
PRLI	process login	
SAN	Storage area network	A network designated for storage traffic
SCSI	Small Computer Systems Interface	

8

Overcoming the limitations of traditional data center networks

This chapter introduces the main challenges that network architects have to face when trying to design and implement networks in a data center environment. We elaborate on the Cisco solutions that can be leveraged to overcome these shortcomings:

- First we introduce the traditional three tiered data center network model and explain why traditional Ethernet networks cannot meet the requirements that modern data centers bring in terms of efficiency, flexibility, scalability, and manageability.
- After explaining the reasons why Ethernet networks have to evolve and where the drivers for change come from, we explore the available Cisco solutions that can be leveraged to overcome the shortcomings we have introduced at the beginning of the chapter.
- These new Cisco data center networking solutions possess different degrees of maturity and are positioned differently in terms of use cases and scope. We explore these solutions in detail in this chapter, also highlighting what type of standard based solutions can be used and elaborating on the key differences. The solutions are divided as follows:
  - Multi-chassis EtherChannel techniques: Port Channels, virtual Port Channels, Enhanced virtual Port Channels, virtual Port Channel +.
  - Data Center Network Fabrics: Cisco FabricPath and TRILL
  - Data Center Network overlays techniques: OTV, VPLS and LISP.

8.1 Traditional data center network model and its limitations

This section first introduces the three tiered data center network model and then will explore the most commonly agreed upon limitations of such networks in order to then elaborate on how Cisco data center networking solutions can be leveraged to address and overcome those limitations.

8.1.1 Traditional three tiered data center network model

As introduced in 1.1, “Data center drivers and trends” on page 2 and 1.2, “Architecture evolution” on page 20, modern data center networks need to address requirements that are challenging the three tiered model, forcing the network to evolve beyond the traditional approaches and best practices. For example, usually the availability requirements have been met by keeping the environment static and limiting the amount of changes that can be applied on the data center network infrastructure. At the light of IT trends such as virtualization and cloud, the network must evolve and become flexible, responsive and tightly integrated and aligned with IT infrastructure management and IT Service Management disciplines.

Figure 8-1 shows a high level diagram of a three tiered data center network: it consists of core, aggregation, and access layers. This approach has many benefits, among which are simplifying the network design through the use of common building blocks, ease of growth and scalability from medium-sized to very large data centers. Usually Layer 2 network adjacency is needed at the access layer while the Core is implemented based on Layer 3. The Distribution layer typically acts as a Layer 2/3 boundary with this model.

The hierarchical network design model gains much of its stability and high availability characteristics by splitting out switching nodes based on their function, and providing redundant switching units for each functional layer required. Each one of the three traditional data center network layers is briefly described below.

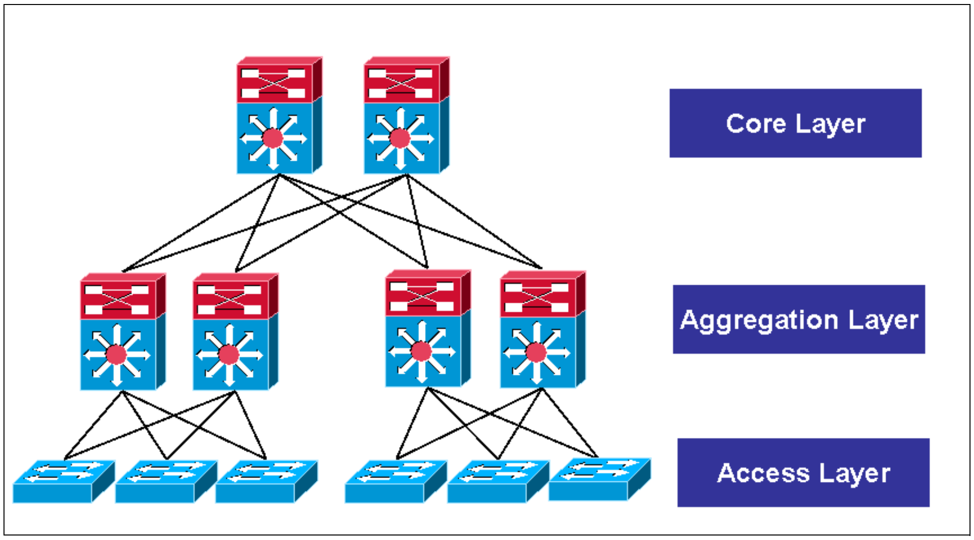


Figure 8-1 Three tiered data center network

Core layer

The core of a data center network is typically broken out into a pair of high performance, highly available, chassis-based switches. In larger or geographically dispersed network environments, the core is sometimes extended to contain additional switches, and the recommended approach is to scale the network core continuing to use switches in redundant pairs. The primary function of the data center network core is to provide highly available, high performance switching for IP traffic. By configuring all links connecting to the network core as point-to-point Layer-3 connections, rapid convergence around any link failure is provided, and the control plane of the core switches is not exposed to broadcast traffic from end node devices or required to participate in STP for Layer 2 network loop prevention.

Aggregation layer

The aggregation layer of the data center provides a consolidation point where access layer switches are connected providing connectivity between servers for multi-tier applications, as well as connectivity across the core of the network to clients residing within the campus, WAN, or Internet. The aggregation layer typically provides the boundary between Layer-3 routed links and Layer 2 Ethernet broadcast domains in the data center. The access switches are connected to the aggregation layer using 802.1Q VLAN trunks to provide the capability of connecting servers belonging to different VLANs and IP subnets to the same physical access switch. An alternate aggregation layer option is to use Layer-3 routed point-to-point links also between the aggregation layer and access layers, referred to as a routed-access design.

The routed-access design provides deterministic failover characteristics; it features very small broadcast and failure domains contained to a single switch. However, because the routed access model limits a given VLAN or IP subnet to a single access switch, it also severely limits the mobility of servers without requiring IP address changes and can complicate designs, such as Network Interface Card (NIC) teaming approaches where multiple interfaces from a server carry the same IP address. Usually then access switches and aggregation switches are connected via Layer 2 using the spanning tree protocol, which blocks half of the inter-layer links. Often additional aggregation switches are added to have a physical separation between various services or business units in order to simplify the day by day management; that is, separate maintenance windows per aggregation block. Moreover the creation of additional aggregation switches blocks is also required when the maximum number of ports are exhausted on the existing aggregation switches.

Access layer

The access layer of the network provides connectivity for server end nodes residing in the data center. Design of the access layer is dictated by server density, form factor, server virtualization, and connectivity (GigE, 10GbE, FC, and so on).

Note that Layer 2 adjacency requirement is needed for HA protocols and some applications that require NIC cards to be in the same Layer 2 broadcast domain. One of the main requirements of the access layer is to provide the maximum number of access ports at the lowest port cost. Although most access layer implementations use Layer 2 switches, it can also be implemented using Layer 3 switches for those cases where Layer 2 adjacency beyond a switch or a pair is not needed.

There are several advantages for implementing a Layer 3 access, such as convergence time of Layer 3 protocols are usually better than spanning tree protocols and all uplinks are active (no spanning tree blocking). However this approach can be more costly and also limits the scope of a given VLAN to just one or two access switches, thus limiting overall flexibility.

8.1.2 Shortcomings of the tiered data center network

For more than a decade, data center networks have been designed the same way campus networks were designed: using a hierarchical approach with a number of layers (typically three to five). This dates back from an era of limited switch performance, where core devices had to be limited to simple functionality to maximize performance. Aggregation layers were therefore used to distribute sophisticated functions, such as access control lists. Also the ports in the core devices were usually rare and expensive, so often the aggregation layer was required to increase the number of ports. Basically data centers networks have been designed very similarly to campus networks where servers would be the endpoints instead of workstations. Today this model is not up to the task of satisfying network requirements in the data center and data center network design has become its own discipline with a specific solution space that is very different from the campus networking one.

LAN switching performance has outstripped demand and switches are now available with more than a terabit of internal switching capacity and port densities that were not imaginable just a few years ago. They can also deliver this level of performance when performing sophisticated networking functions. With such switches, it is possible to redesign data center LANs to collapse the number of layers, resulting in fewer switches and correspondingly fewer inter-switch trunks, reducing power space and cooling requirements. This simplified topology means lower latency and it is easier to manage, provision and troubleshoot.

From a high level point of view, the evolutionary journey in the data center network can be represented as shown in Figure 8-2. In fact the data plane / control plane of the data center network is evolving into a single logical model so that an increase in performance and stability does not come to the expense of overall manageability and so that the network can offer a simplified view of itself to the applications and IT infrastructure teams.

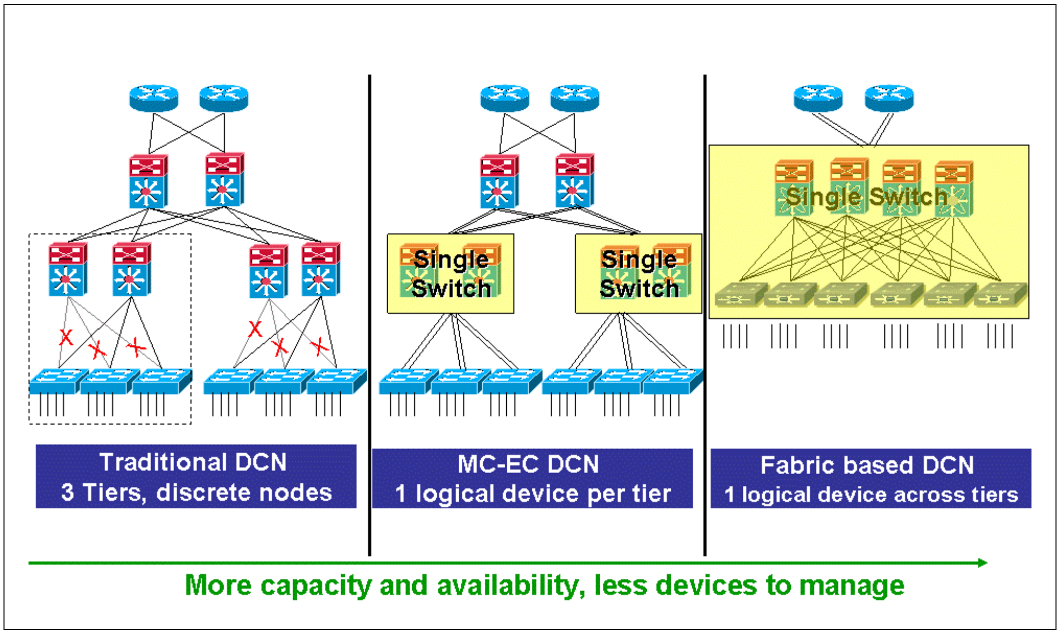

Single Switch
Single Switch
Single Switch
Traditional DCN
3 Tiers, discrete nodes
MC-EC DCN
1 logical device per tier
Fabric based DCN
1 logical device across tiers
More capacity and availability, less devices to manage

Figure 8-2 Data center network data plane / control plane evolution

- A first evolutionary step (shown in the middle of Figure 8-2) is logically aggregating network nodes belonging to the same hierarchical layer, thus obtaining a single switch view at the aggregation layer, for example. These techniques are commonly referred to as Multi-chassis EtherChannel techniques (MC-EC), as discussed in 8.2, “Multi-chassis EtherChannel techniques” on page 428.
- The end goal of this evolution (shown on the right hand side in Figure 8-2) is extending this paradigm to the whole data center, enabling a single logical view for all the network nodes in the data center. From a data plane point of view Cisco FabricPath can be leverage to achieve this end state. We discuss FabricPath in 8.3, “Data center network fabrics” on page 442.

Also, enterprises are planning to enable a single logical data center view across multiple physical data centers, enabling active-active production sites so that client requests can be load balanced, for example, or demand fluctuations and failures can be handled seamlessly. In order to do this properly, it is necessary to provide data center to data center connectivity across different layers of the OSI stack:

- Layer 1 Extension: Usage of Optical links, or DWDM, for protocol-agnostic transport: Ethernet, Fibre Channel, InfiniBand and other traffic types
- Layer 2 Extension: The widespread adoption of server virtualization technologies drives a significant expansion of the Layer 2 domain, and also brings the need to extend Layer 2 domains across physically separated data centers in order to stretch VLANs to enable VM mobility. These are very challenging requirements to satisfy in order to achieve the service delivery objectives since they directly impact the scalability (in terms of new MAC addresses that can be included in a Layer 2 domain) and flexibility (in terms of the time and effort needed to deploy new services) of the data center network.

  Common drivers for Layer 2 extension are as follow:

  - Applications exist that use “hard-coded” IP parameters that cannot be easily modified to be available from different locations. That is, the same subnet needs to be available in different remote locations.
  - Data center consolidation initiatives or outsourcing transitions can require that a portion or the whole server farm needs to be moved to another location with minimal effort and disruption.
  - Layer 2 heartbeat mechanisms are used for highly available server clusters.
  - Virtual machine mobility (for example, VMware VMotion and POWER Live Partition Mobility)
  - There are network, server load balancing, and security devices that require Layer 2 adjacency for high availability, as well as heartbeat.
  - Network Attached Storage subnet design requirements exist.
- Layer 3 Extension: This is the traditional way of providing data center to data center network connectivity using private or service provider owned WAN infrastructures and leveraging proven technologies such as MPLS.

We now explore in detail the most important limitations of traditional data center networks so that in the next section we can elaborate on how Cisco data center networking solutions can overcome these limitations and enable a truly modern data center that can satisfy the business and IT requirements we introduced in Chapter 1, “The modern data center” on page 1 and Chapter 3, “Designing networks for the modern data center” on page 131.

Figure 8-3 captures the most common limitations of traditional data center networks; we now explore each of these in detail.

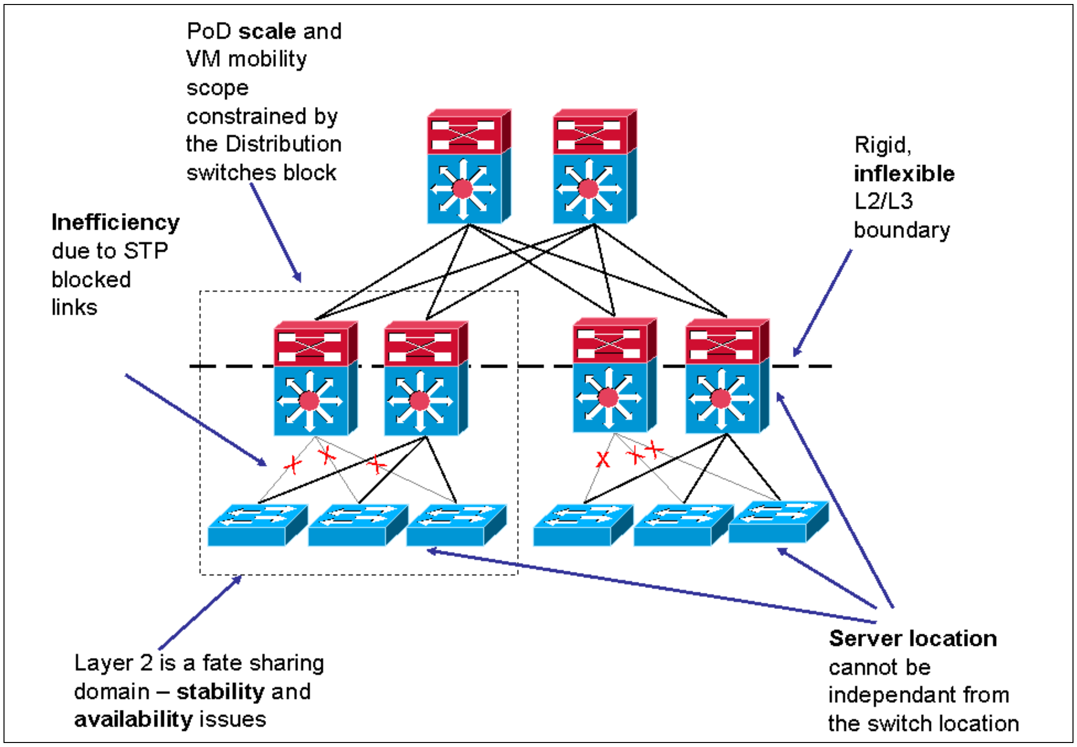

PoD scale and VM mobility scope constrained by the Distribution switches block
Rigid, inflexible L2/L3 boundary
Inefficiency due to STP blocked links
Layer 2 is a fate sharing domain – stability and availability issues
Server location cannot be independant from the switch location

Figure 8-3 Limitations of traditional data center networks

Network resource utilization inefficiency

The usage of spanning tree protocol or one of its variants as the control plane protocol comes at the expense of inefficient data center network resources utilization. In fact, in order to avoid loops in a Layer 2 network, a tree topology must be enforced by disabling a subset of the available links. So usually half of the available links are idle and the usage of the available capacity is suboptimal at best. This can be mitigated by balancing the available VLANs across different STP instances, but obviously, being able to exploit 100% of the available capacity would be a huge improvement from a data center network efficiency point of view.

Also, from a path optimization point of view, a static tree topology has a drawback in terms of the ability of leveraging the most direct path between two hosts in the data center. In fact the chosen path is always the most optimal only from the elected root bridge perspective.

We introduce virtual port channels as a way to improve link utilization in 8.2, “Multi-chassis EtherChannel techniques” on page 428 and FabricPath as a way of enabling multipathing at Layer 2 in 8.3.1, “FabricPath” on page 442

Network inflexibility and limited range of virtual machine mobility

The Virtual Machine (VM) is the new atomic building block in the data center and the importance of physical server NICs for the network architecture fades when compared to the virtual networking realm inside the server platforms. The network in this sense must become VM-aware and support the live migration of Virtual Machines at different levels:

- Between physical servers located in the same PoD: This can be achieved very easily because usually each PoD is a Layer 2 domain, hierarchically below the L2/L3 boundary which usually sits at the aggregation layer. A limiting factor in this scenario is that usually with a three tiered design, the logical PoD size is relatively small.

- Between physical servers located in different PoDs within the same data center: This can be difficult to achieve in large data centers as PoDs may be very distant from each other from a physical placement point of view and as PoDs are usually interconnected via Layer 3. In fact as illustrated on Figure 8-3 on page 426 there is Layer 3 core link between aggregation layers, which limits VM mobility and the ability to add more servers on the same VLAN. This also has an impact on resource utilization as the available rack space may not be fully utilized.
- Between physical servers located in different data centers: This can be difficult to achieve as a Layer 2 interconnection between data centers is not a common best practice and it may impact the stability of the stretched Layer 2 domains.

We introduce OTV as a way to stretch VLANs across data centers in 8.4.1, “Overlay Transport Virtualization” on page 458 and FabricPath as a way of creating stable Layer 2 domains in large data centers in 8.3.1, “FabricPath” on page 442.

Layer 2 scalability and manageability limitations

The traditional three tier data center network can become a bottleneck in terms of these capabilities:

- The ability of the single PoD to scale, which means accommodating additional server and storage endpoints
- The ability of the single data center to scale, which means accommodating additional PoDs.

This situation can be caused by several factors depending on the specific data center deployment. Here are some common examples:

- Maximum number of MAC addresses the aggregation switches can handle. In fact in traditional Ethernet networks the aggregation switches must learn all the MAC addresses of the hosts that are attached to the downstream access switches
- Maximum number of ARP requests that can be handled by the aggregation switches
- Maximum number of ports that can be provisioned at the aggregation layer which limits the number of access switches that can be connected and the amount of bandwidth that can provisioned between access and aggregation layers.

We introduce FabricPath as a way of overcoming these scalability limits in 8.3.1, “FabricPath” on page 442

Layer 2 domain stability and availability

The stability of the control plane in a typical Layer 2 DCN is typically controlled by the spanning tree protocol (STP) or one of its many variants.

This approach, however, does not have the robustness, flexibility and efficiency that is required to assure highly available service delivery. For example, the typical convergence time required to recover from a link failure (order of magnitude of seconds even with RSTP) is not in synch with the needs of today’s network-dependent and high-paced business environment. This limitation has a huge impact on the availability and reliability of the data center network.

Also Layer 2 domains are fate sharing domains in the sense that a fault affecting one of the network nodes part of that domain (for example accidentally creating a loop in the topology) can cause a ripple effect and bring down the whole Layer 2 domain: local issues can trigger data center network wide outages.

We introduce FabricPath as a way of having more stability and availability in Layer 2 networks in 8.3.1, “FabricPath” on page 442. For example Fabricpath has improvements such as an added Time to Live field to avoid Layer 2 packets looping indefinitely.

Server location independence

Another critical aspect to consider in the context of inter data center VM mobility is asymmetrical client to server or server to server traffic patterns: these could occur because after a VM moves it keeps its IP address (identity) but clients may still try to access it at the old location, instead of the new one. This is caused by a structural flaw of the Internet routing architecture and its namespace: an IP address consolidates in fact two very distinct data points:

- Identity information: Who the host is
- Location information: Where the host is

In fact, a host IP address changes when the host changes location, even if its identity (function) has not changed at all. This happens because as mentioned above today the IP Address is used as the host identity. In the same way if a new host is attached to a port that was previously assigned to another host it will need a new IP address on the same subnet as the previous host. In other words changing the host location will trigger different IP addresses being assigned to the host.

Effectively, then, the amount of freedom a VM has in moving from physical server to physical server is constrained as identity and location data are one and the same, while ideally the VM would need to keep its identity regardless of location to simplify mobility and avoid asymmetrical traffic patterns.

We introduce LISP as a way of solving this problem by separating host identity and host location information in 8.4.5, “LISP” on page 470.

8.2 Multi-chassis EtherChannel techniques

In this section, we discuss port-channel and multi-chassis EtherChannel techniques.

8.2.1 Port channeling

A port channel bundles several individual interfaces from a single device into a a single logical interface to provide increased bandwidth and redundancy in the connectivity between switches, routers and servers. Port channeling load balances traffic, by flow, across these physical interfaces. This means that all the traffic belonging to a particular traffic flow will always use the same physical interface. The port channel can stay operational as long as at least one physical interface within the port channel is operational.

You create a port channel by bundling compatible interfaces. You can configure and run either static port channels or ports channels running the Link Aggregation Control Protocol (LACP) which is defined in IEEE 802.3ad. Any configuration changes that you apply to the port channel are applied to each member interface of that port channel. For example, if you prune a VLAN on the logical port channel interface, that VLAN is pruned on each interface in the port channel.

Port channels can be configured as either Layer 2 or layer-3 interfaces. As Layer 2 interfaces they can be configured as access ports, carrying a single VLAN, or 802.1Q trunk ports.

Individual interface compatibility requirements

When you add an individual interface to a channel group to form a port-channel, the Cisco NX-OS checks certain interface attributes to ensure that the interface is compatible with the channel group. Making sure that physical attributes like speed and QoS capabilities match ensures that traffic hitting any individual interface in the logical port channel is handled equally.

The Cisco NX-OS also checks a number of operational attributes for an interface before allowing that interface to participate in the port-channel aggregation.

The compatibility check includes the following operational attributes:

- Port channel mode (on, off or active)
- Access VLAN
- Trunk native VLAN
- Allowed VLAN list
- Speed
- 802.3x flow control setting
- MTU
- Broadcast/Unicast/Multicast Storm Control setting
- Priority-Flow-Control
- Untagged CoS
- SPAN-cannot be a SPAN source or a destination port
- Layer 3 ports-cannot have sub-interfaces
- Storm control
- Flow-control capability
- Flow-control configuration

In some cases, an individual interface that does not pass the matching attribute check can still be forced to become a part of a port channel. This is not recommended as it can cause unexpected problems.

Load balancing using port channels

The Cisco NX-OS load balances traffic across all operational interfaces in a port channel by reducing part of the binary pattern formed from the addresses in the frame to a numerical value that selects one of the links in the channel. Port channels provide load balancing by default and the basic configuration uses the following criteria to select the link:

- For a Layer 2 frame, it uses the source and destination MAC addresses.
- For a Layer 3 frame, it uses the source and destination MAC addresses and the source and destination IP addresses.
- For a Layer 4 frame, it uses the source and destination MAC addresses, the source and destination IP addresses, and the source and destination port number.

You can configure the switch to use one of the following methods to load balance across the port channel:

- Destination MAC address
- Source MAC address
- Source and destination MAC address
- Destination IP address
- Source IP address
- Source and destination IP address
- Destination TCP/UDP port number
- Source TCP/UDP port number
- Source and destination TCP/UDP port number

8.2.2 Link Aggregation Control Protocol (LACP)

The Link Aggregation Control Protocol (LACP) is described within the IEEE 802.3AD specification. It provides a method to control the bundling of several physical ports together to form a single logical channel. LACP is defined to allow a network device to negotiate an automatic bundling of links by sending LACP Protocol Data Units (LACP PDUs) to its directly connected peer.

In static Link Aggregation, each port-channel is created with no awareness of the port channel on it peer switch. This can cause packets to be dropped if the two switches are not configured exactly the same way. Using the Link Aggregation Control Protocol, dynamic Link Aggregation Groups (LAGs) are aware of the state of the port channel on the partner switch. LACP enables you to form a single Layer 2 or Layer-3 link automatically from two or more Ethernet links. This protocol ensures that both ends of the Ethernet link are functional and agree to be members of the aggregation group.

LACP performs the following functions in the system:

- Maintains configuration information to control aggregation.
- Exchanges configuration information with other peer devices.
- Attaches or detaches ports from the LAG based on the exchanged configuration information.

In real world terms this means that if you accidentally have physical interfaces from a single port channel on switch A, connecting to physical interfaces that are part of two separate port channels on switch B, if you have LACP enabled your incorrectly cabled ports will not come up. If LACP is disabled, by having the port channel mode on the interfaces set to on, the interfaces would come up and possibly cause a problem in your network.

8.2.3 Port channel benefits

From a routing protocol or spanning-tree perspective a port channel is seen as a single logical interface, even if it is made up of multiple physical interfaces. Because of this only a single neighbor relationship is established over the entire port channel bundle. This can save a fair amount of CPU cycles over non bundled interfaces. For example if you have four physical Layer 2 links between two switches, and they are not part of a port channel, you will have to exchange BPDUs across all four links, and create a spanning tree relationship across all of them. Also, depending on how the Layer 2 links are configured, you may end up blocking on some of the four ports, maybe even 3 of them, to avoid creating a loop. Configuring all four interfaces as a port channel between the two devices allows them to only need to form a single spanning tree neighbor relationship, to send BPDUs on only one of the physical interfaces for the entire logical bundle, and to forward across all four physical ports without worry of a spanning tree loop.

The same is true if the four interfaces are configured as layer-3, and you are running a routing protocol like OSPF. If you want to forward traffic across all four interfaces, you will have to put an IP address on each one of them, and if configured correctly, there will be four OSPF neighbor-ships being maintained between the two physical devices. On the other hand, if we use a port channel, we only need a single IP address on each device, configured on the logical port channel interface, and a single OSPF neighbor-ship will be formed.

Since the port channel remains up as long as at least one physical interface is active between the peers, this can also help keep your network more stable since a single, or even multiple interface, failure will not cause a protocol re-convergence. This allows for very fast, generally sub second, traffic forwarding re-convergence, when an interface that is part of port channel goes down. The forwarding just fails over to another member in the port channel bundle, as long as one is available. If you configure your port channels using multiple physical interfaces on separate line cards, your network would also be able to survive a line card failure without having to re-converge. This can help network stability because when a neighbor relationship is torn down, it does not only cause a recalculation of the network for the two neighbors with the failed interface, but very often for all of the neighbors connected to them, and sometimes for the entire system, be it the autonomous system of a routing protocol like OSPF or the spanning tree domain.

Port channels allow you to easily aggregate bandwidth. With a few commands you can bundle sixteen 10 gigabit interfaces into a single 160 gigabit logical interface. The logical port channel interface keeps interface statistics and counters for the entire bundle. When you look at the traffic statistics of the port channel interface you see the combined statistics for all of the physical interfaces in the bundle. You can also look at statistics for each individual physical interface for a more granular view.

8.2.4 Solving regular port channel limitations

A limitation of regular port channels is that all the physical ports in the aggregation group must reside on the same physical device. Therefore it can still be a single point of failure in your network design if you have a chassis failure. Cisco solves this limitation by using a concept called Multi-Chassis EtherChannel (MEC) allowing ports to be aggregated on two different physical chassis that form a single “virtual switch.” This concept is implemented in various ways depending on the type of device we are discussing. It is called:

- Multi-chassis LACP in IOS
- Virtual Switching System (VSS) on the Catalyst 6500
- Virtual Port Channel (vPC) in NX-OS.
- vLAG (virtual link aggregation) is an IBM Networking OS feature that is equivalent to vPC and is also proven to be interoperable with Cisco upstream switches in vPC mode.

Since this book is concentrating most heavily on Cisco devices that run NX-OS, we explore the vPC feature in this chapter.

If you would like to learn more about Multi-chassis LACP, visit this website:

http://www.cisco.com/en/US/docs/ios-xml/ios/cether/configuration/15-1s/ce-multichass-lacp.html

If you would like to learn more about VSS, visit this website:

http://www.cisco.com/en/US/docs/switches/lan/catalyst6500/ios/12.2SX/configuration/guide/vss.html

Figure 8-4 shows the physical difference between a regular port channel and a Multi-Chassis EtherChannel.

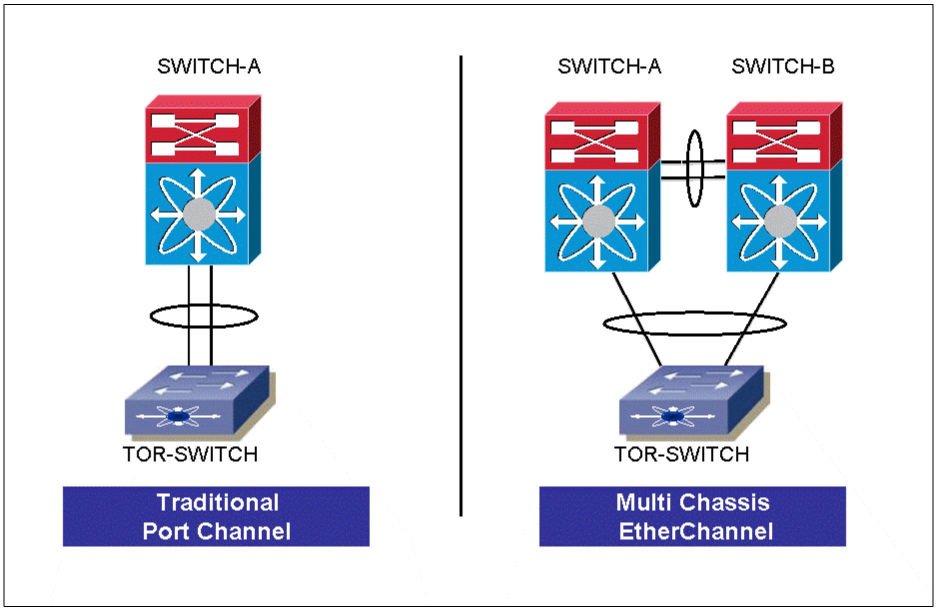


Figure 8-4 Difference between a regular port channel and a Multi-Chassis EtherChannel

With a traditional port channel, the failure of Switch-A in Figure 8-4 causes the TOR-Switch to lose connectivity. While with a Multi-Chassis EtherChannel even if we lose Switch-A the TOR-Switch stays connected and fails over very quickly, without having to recalculate any routing protocol or spanning tree topology.

8.2.5 vPC: Virtual Port Channel

The Virtual Port Channel (vPC) is Cisco's implementation of Multi-Chassis EtherChannel in NX-OS. vPC is supported on the Nexus 7000, 6000, 5000 and 3000 switches. vPC is a virtualization technology that presents two Cisco Nexus switches as a unique Layer 2 logical node to access layer devices or endpoints. In other words a virtual port channel (vPC) allows links that are physically connected to two different Nexus switches to appear as a single port channel to a third device as shown in Figure 8-5.

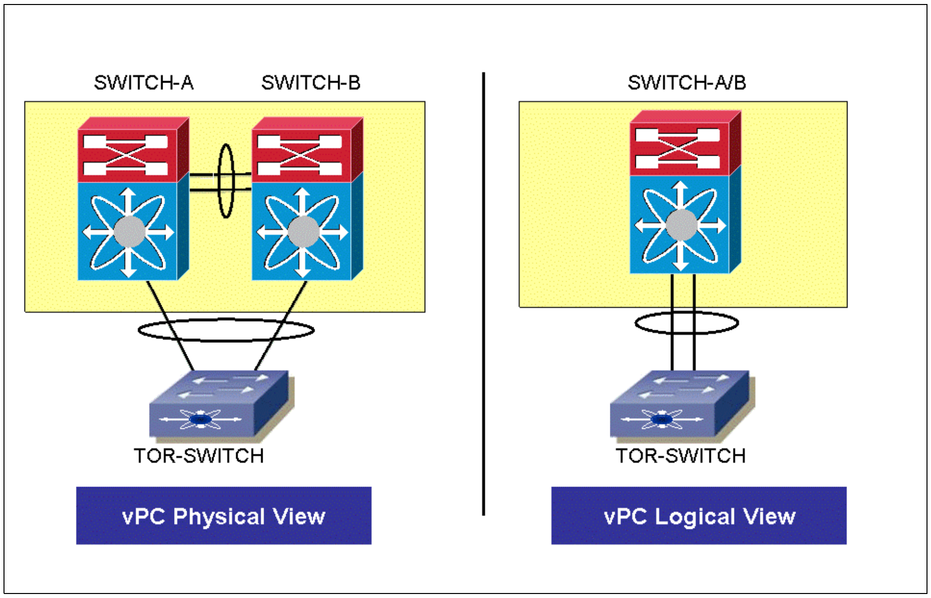


Figure 8-5 Depicts the difference between the physical view and the logical view

vPC provides the following technical benefits:

- Eliminates spanning tree protocol (STP) blocked ports
- Uses all available uplink bandwidth
- Allows dual-homed servers to operate in active-active mode
- Provides fast convergence upon link or device failure
- Offers dual active/active default gateways for servers
- Simplifies network design
- Build highly resilient and robust Layer 2 network
- Enables seamless virtual machine mobility and server high-availability clusters
- Scales available Layer 2 bandwidth, increasing bisectional bandwidth
- Grows the size of the Layer 2 network

The vPC feature leverages both hardware and software redundancy aspects to build a highly scalable and available network. It uses all port-channel member links available so that in case an individual link fails, hashing algorithm will redirect all flows to the remaining links.

A vPC domain is composed of two peer devices. Each peer device processes half of the traffic coming from the access layer, or a directly connected host. In case a peer device fails, the other peer device will absorb all the traffic with minimal convergence time impact. Each peer device in the vPC domain runs its own control plane, and both devices work independently. Any potential control plane issues stay local to the peer device and does not propagate or impact the other peer device.

From a spanning tree protocol (STP) standpoint, vPC eliminates STP blocked ports and uses all available uplink bandwidth. Spanning tree is used as a fail safe mechanism and does not dictate L2 path for vPC-attached devices.

Within a vPC domain, customers can connect access devices in multiple ways:

- vPC-attached connections leveraging active/active behavior with port-channel, this is the recommended way.
- Active/standby connectivity using spanning-tree
- Single attachment without spanning-tree running on the access device.

Note: The vPC feature is included in the base NX-OS Software license, and as such does not require any additional licensing.

8.2.6 vPC components and terminology

The components of a vPC domain are shown in Figure 8-6:

- vPC peer: A vPC switch, one of a pair
- vPC member port: One of a set of ports (Port Channels) that form a vPC
- vPC: The combined Port Channel between the vPC peers and the downstream device
- vPC peer-link: A Layer 2 link used to synchronize state between vPC peer devices, it must be 10GbE and it should be a port channel
- vPC peer-keepalive link: The keepalive link between vPC peer devices, this link is used to monitor the health of the peer switch.
- vPC VLAN: One of the VLANs carried over the peer-link and used to communicate via vPC with a peer device.
- non-vPC VLAN: One of the STP VLANs not carried over the peer-link
- CFS: Cisco Fabric Services protocol, used for state synchronization and configuration validation between vPC peer devices

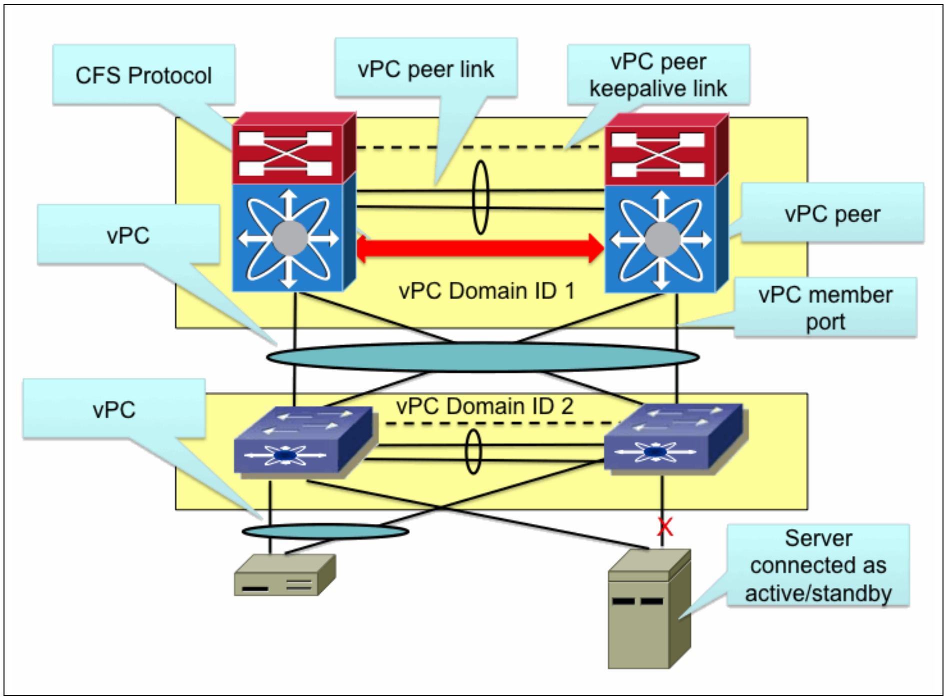


Figure 8-6 vPC components and terminology

Building a vPC

To build a vPC, you must first enable the vPC feature in your NX-OS switch. Once that is done, you can build your vPC by doing the following steps:

1. Define a vPC Domain ID
2. Define the systems vPC role
3. Create a vPC Keepalive Link
4. Create a vPC Peer-link
5. Create the vPC

These steps are described in the next few sections.

The vPC Domain ID

The vPC domain defines the grouping of switches participating in the vPC. In other words the two vPC peer switches must use the same domain ID. The valid vPC domain IDs are from 1 to 1000. The domain ID must be unique when connecting two sets of vPC peers as shown in Figure 8-6 between domain ID 1 and domain ID 2. In that figure the two top Nexus 7000 switches (in domain ID 1) form one vPC domain and the two bottom Nexus 5500 switches (in domain ID 2) form a second one. In this scenario, the vPC domain identifiers must be different in both layers because this information is used as part of the LACP protocol. Using the same vPC domain identifiers would generate continuos flaps on the vPC interconnecting the Nexus 5500s to Nexus 7000s.

Once configured, both peer devices use the vPC domain ID to automatically assign a unique vPC system MAC address, as follows:

```
vPC system-mac = 00:23:04:ee:be:<vpc domain-id in hexadecimal>
```

For instance, vPC domain 10 will result in a vPC system-mac of 00:23:04:ee:be:0a.

The vPC system MAC is then identical on both peer devices. This is the foundation for the L2 virtualization technique with vPC. When a pair of vPC peer switches need to present themselves as a unique logical device they will use this unique and shared system-MAC across both peer devices. This will make both peers look as a single switch from a spanning tree protocol perspective.

> **Note:** It is possible to manually configure the vPC system-MAC value with the command "system-MAC" inside the vPC domain configuration context.

A vPC local system MAC is also owned by each peer devices and is unique per device. The vPC local system MAC is derived from the system or VDC MAC address. The vPC local system MAC is used whenever vPC systems do not need to present themselves as a single logical device. The local system MAC is used for regular STP and is used in for non vPC connected devices such as switches connected to both peers without using a port-channel.

The vPC system role

There are two defined vPC roles: primary and secondary.

The vPC role defines which of the two vPC peer devices processes Bridge Protocol Data Units (BPDUs) and responds to Address Resolution Protocol (ARP). By default only the primary system will do these roles. If the secondary system receives a BPDU or an ARP request it forwards them to the primary switch over the vPC peer link (to be described later in this chapter,) and the primary switch sends a response as required.

You can use the "role priority <value> command" (under vPC domain configuration context) to define which switch will be the vPC primary peer device:

<value> ranges from 1 to 65535 and the lowest value will dictate the primary peer device.

In case of a tie (same role priority value defined on both peer devices), the lowest system MAC will dictate which switch will become the primary peer device. This value must be defined before the peer relationship is formed. Once a peer relationship forms a change to the configuration will not have any effect until the peer relationship re-forms. In other words the vPC peer relationship will have to go down due to a peer-link going down (or being shut down) or a switch reload, both of which can be disruptive events if any vPCs are active.

vPC role (primary or secondary) matters due to vPC peer behavior in a peer-link failure. When a peer-link failure occurs the secondary peer device sets all of its vPC member ports to down state and shuts down all its vPC VLAN interfaces. This means that any SVI (Switch Virtual Interfaces) that belong to VLANs that are part of a vPC are shut down on the secondary device. This is done to avoid creating a split in the network.

The vPC peer-keepalive link

The vPC peer-keepalive link is a Layer 3 link that connects the vPC peer devices to each other. The vPC peer-keepalive link carries periodic heartbeats between the vPC peers. The keepalive link is physically *separate* from the peer-link.

It is used at the boot up of the vPC systems to guarantee both peer devices are up before forming a vPC domain, and also when vPC peer-link fails. During a peer-link failure the vPC peer-keepalive link is leveraged to avoid split brain scenario (both vPC peer devices becoming active-active). The assumption is that if the peer-link is down, *and* there are no messages received on the keepalive link, then the other device is actually down.

When the vPC peer-link is down, there is no more real time synchronization between the two peer devices, this could cause a problem. If the primary and secondary devices are still up, because they are both receiving keepalive messages, the secondary switch will shut down its vPC member ports and SVIs as mentioned in the vPC System Role section of this chapter. If the primary switch looses its peer link and does not see messages from the secondary on the keepalive link, it declares the secondary switch down and continues to forward. On the other hand if the secondary switch looses its peer-link and does not receive any keepalive messages from the primary switch, it declares the primary switch down, becomes primary and continues to forward traffic.

The keepalive timers (hello interval, hold time-out and keepalive time-out) are configurable.

In terms of data structure, a vPC peer-keepalive message is a User Datagram Protocol (UDP) message on port 3200 that is 96 bytes long, with a 32-byte payload.

The vPC peer-link

The vPC peer-link is a 802.1Q trunk between the vPC domain peers. It carries all of the Cisco Fabric Services (CFS) Protocol information between the peers. The vPC peer-link is the most important connectivity element of the vPC system. This link is used to create the illusion of a single control plane by forwarding BPDUs, LACP or ARP packets to the primary vPC switch from the secondary vPC switch. It is also used for forwarding traffic that originates at or is destined for non-vPC connected devices (orphan ports). It carries flooded frames, and in the case of the Cisco Nexus 7000 Series, Hot Standby Router Protocol (HSRP) frames. It is also used to synchronize the MAC address tables between the two peers.

This process is shown in Figure 8-7.

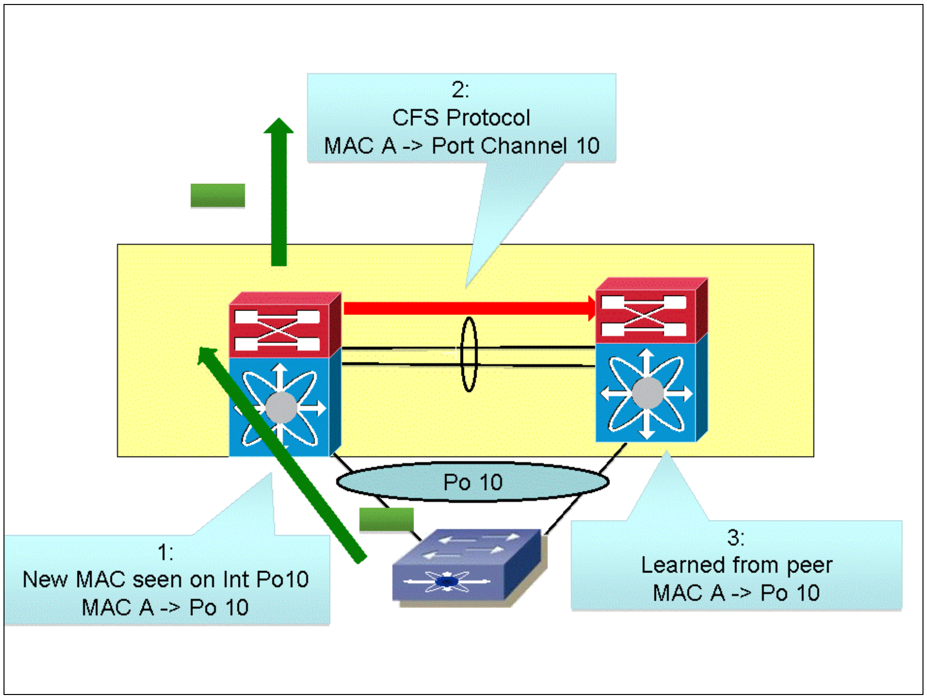


Figure 8-7 MAC address learning with vPCs

As seen in Figure 8-7, when a packet with a new MAC address arrives on a vPC member port, the switch adds it into its MAC address table, which is normal MAC address learning behavior for a switch. The difference in a vPC system is that it also send the learned MAC address to its peer device using the CFS Protocol. The peer switch then adds the new MAC address into its forwarding table pointing at the local port channel. This ensures that the peers can forward locally since both of them have a physical interface on the port channel.

Because of the importance of the vPC peer-link, it is strongly recommended that it is created as a port channel. In the case of a Nexus 7000, if possible, it is recommended that it is created as a port channel across two separate line cards, to provide extra reliability.

The vPC peer-link must use a 10 Gbps interface. It is mandatory that both sides of vPC peer-link are strictly identical. As a consequence, it is not possible to mix F series linecard port on one side and M series linecard port on the other side. The vPC peer-links must be directly connected to each other.

> **Note:** The M132XP and M108X2 (that is, a 32 10-Gigabit Ethernet line card and 8 10-Gigabit Ethernet line card, respectively) can form a port-channel together. However, be aware of the requirement that ports on M132XP can form a port-channel with ports on M108X2 only if the port is configured in dedicated mode.
>
> This is consistent with vPC peer-link recommendations: vPC peer-link member ports should have full 10-Gigabit Ethernet capability - in other words, no oversubscription

When deciding how much bandwidth you need for your vPC peer-link you should understand that for unicast traffic, vPC peer devices always use local forwarding using the vPC member ports. Therefore the vPC peer-link is usually not loaded with unicast traffic unless all local vPC member ports to a device fail on one side of vPC domain. For multicast traffic, a copy of the stream is replicated on the vPC peer-link (except when the vPC peer-link is built with F2 ports as the line card do not support dual DR [Designated Router] in multicast environment). Multicast traffic needs to be taken into account when dimensioning the vPC peer-link.

> **Note:** Note that any VLANs that are allowed on a vPC member port must be allowed on the vPC peer-link.

Cisco Fabric Services (CFS) protocol

The Cisco Fabric Services (CFS) protocol provides reliable synchronization and consistency check mechanisms between the 2 peer devices and runs on across the vPC peer-link. The protocol was first implemented on MDS products and then ported to Nexus switches.

The Cisco Fabric Services (CFS) protocol performs the following functions:

- Configuration validation and comparison (consistency check)
- Synchronization of MAC addresses for vPC member ports
- vPC member port status advertisement
- Spanning tree protocol management
- Synchronization of HSRP and IGMP snooping
- Cisco Fabric Services, which is enabled by default when vPC feature is turned on

8.2.7 vPC data plane loop avoidance

vPC performs loop avoidance at the data-plane layer instead of using the control plane layer. All the loop avoidance logic is implemented directly in hardware on the vPC peer-link ports, avoiding any dependency on using the CPU.

vPC peer devices always forward traffic locally whenever possible. This is possible because both vPC peers have a physical port connected on each vPC. For example if the L2 forwarding table on a peer says that a packet is supposed to go to port channel 7, a physical interface exists to that vPC on both switches and they can forward locally without having to go over the peer-link.

As mentioned earlier, the vPC peer-link does not typically forward data packets and it is usually considered a control plane extension in a steady state network. During steady state operation, the only non control-plane traffic normally forwarded on the vPC peer-link is broadcast, multicast or unknown unicast (anything that does not have a matching MAC address in the forwarding table) traffic. The vPC loop avoidance rule states that traffic coming from a vPC member port, then crossing the vPC peer-link is NOT allowed to egress ANY vPC member port; however it can egress any other type of port such as an L3 port or an orphan port. This is depicted in Figure 8-8.

The only exception to this rule occurs when a vPC member port goes down. At that point the vPC peer devices exchange member port states and reprogram the hardware forwarding for the vPC loop avoidance logic for that particular vPC. The peer-link is then used as backup path for optimal resiliency.

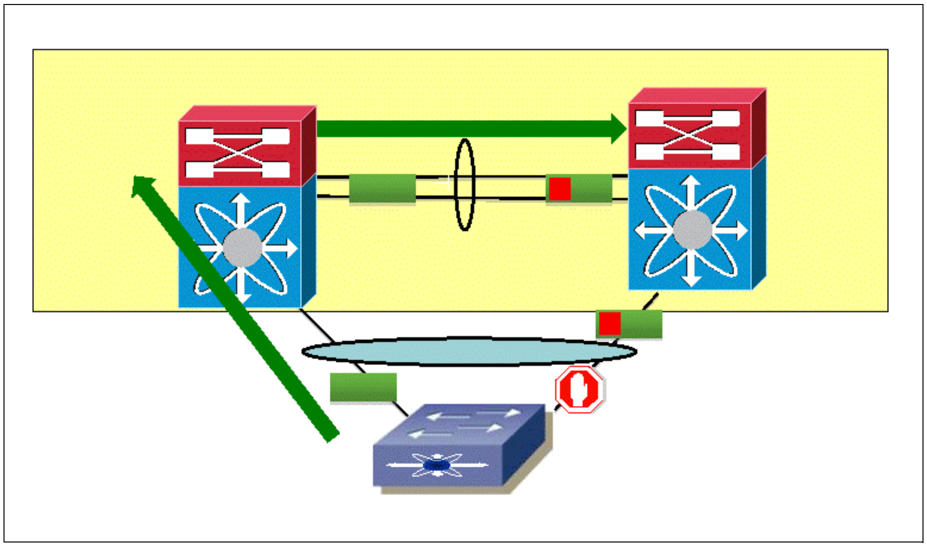


Figure 8-8 The vPC loop avoidance mechanism.

vPC and First Hop Redundancy Protocols

With vPC, First Hop Redundancy Protocols (FHRP) like Hot Standby Router Protocol (HSRP) and Virtual Router Redundancy Protocol (VRRP) are supported in an active/active configuration. In other words both routers forward for the FHRP MAC address. This is an advantage from a data center design perspective because:

- No additional configuration is required
- The Standby device communicates with the vPC manager process to determine if its vPC peer is the “Active” HSRP/VRRP peer
- The “Standby” peer turns on the “gateway capability” and forwards any traffic it receives destined to the FHRP active MAC address
- General HSRP best practices still apply. Best practice states that you should configure HSRP on the primary vPC peer device as active and on the vPC secondary device as standby.
- When running active/active aggressive timers can be relaxed since both devices are forwarding

vPC orphan ports

As described earlier, in the rare case that a vPC peer-link goes down while both peer-switches are still alive all the vPCs on the secondary vPC peer will be suspended. This can cause any devices that are single attached or have their active connection attached to the secondary vPC peer to loose connectivity to the rest of the network. Those hosts are known as orphan devices. The ports they are connected to are also known as orphan ports.

In case of a peer-link shut or restoration, an orphan port's connectivity may be bound to the vPC failure or restoration process. If the device is single attached (connected to only one of the vPC peers) it is recommended that it be connected to active device, since it will keep its vPCs active in the case of a vPC peer-link failure. In the case that a device is connected to both vPC peer switches and is using regular STP, NIC teaming or another active/standby redundancy feature to forward on only one of the interfaces, then the “orphan port suspend” feature should be used.

The vPC orphan ports suspend feature was developed for devices that are dual-attached to the vPC domain in an active/standby mode (server, firewalls or load-balancers for instance). As described earlier in this chapter, when the vPC peer-link goes down, the vPC secondary peer device shuts down all its vPC member ports, as well as its vPC configured orphan ports, causing the dual attached device to failover. When the vPC peer-link is restored configured vPC orphan ports on the secondary vPC peer device are brought back up along with the vPC member ports.

8.2.8 Enhanced vPC

Enhanced vPC enables you to support server connectivity with one topology and address requirement for both high availability and high bandwidth. Enhanced vPC is a technology that supports the topology shown in Figure 8-9, where a Cisco Nexus 2000 Fabric Extender (FEX) is dual-homed to a pair of Cisco Nexus 5500 or Nexus 6000 Series devices while the hosts are also dual-homed to a pair of FEXs using a vPC.

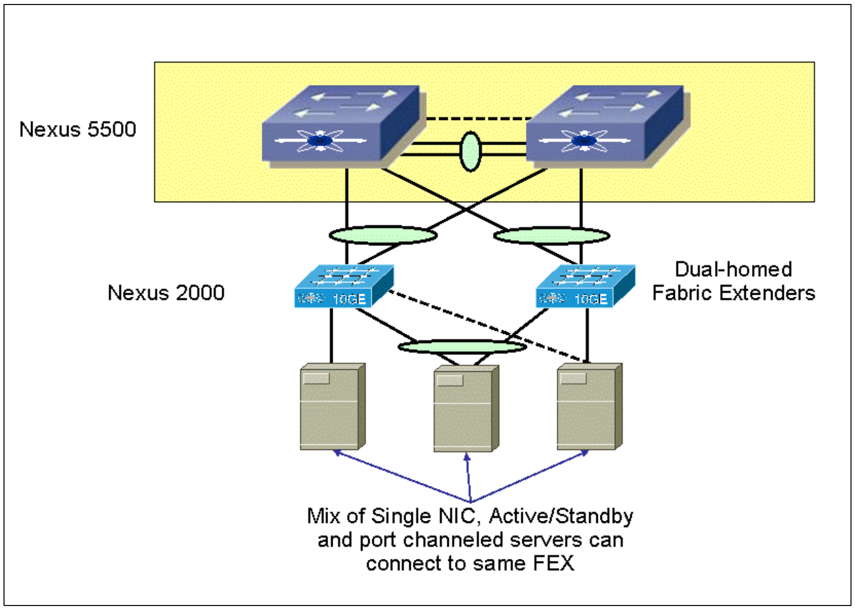


Figure 8-9 An example of Enhanced vPC

Enhanced vPC allows you to have dual tiered vPCs, where a server can be connected over a vPC to two FEXs that are connected using a vPC to two separate Nexus 5500s or Nexus 6000. It adds an extra level of availability and capacity for end hosts.

To learn more about the Enhanced vPC feature, visit this website:

http://www.cisco.com/en/US/docs/switches/datacenter/nexus5000/sw/mkt_ops_guides/513_n1_1/n5k_enhanced_vpc.pdf

8.2.9 Other vPC enhancements

Many vPC enhancements have been added into NX-OS since the feature first came out. Here are some of them:

- vPC Peer-Gateway
- vPC Peer-Gateway Exclude-Vlan
- vPC ARP Sync
- vPC Delay Restore
- vPC Graceful Type-1 Checks
- vPC Auto-Recovery

Details are given at the following website:

http://www.cisco.com/en/US/docs/switches/datacenter/sw/design/vpc_design/vpc_best_practices_design_guide.pdf

Note: Although all the examples and diagrams in this vPC section show the Nexus 5500 and/or Nexus 7000, the Nexus 6000 that has been recently announced by Cisco also supports vPC and Enhanced vPC.

8.3 Data center network fabrics

This section introduces the proprietary Cisco Data Center Fabric solution: FabricPath. Because of the limitations of Layer 2 domains, introduced in 8.1.2, “Shortcomings of the tiered data center network” on page 423, Ethernet has to evolve in the data center to achieve the following objectives:

- Faster convergence time than spanning tree based networks
- Multipathing at Layer 2 so all the available links can be used for active data traffic
- Improved scalability of the Layer 2 domain

We now introduce Cisco FabricPath, explaining how it addresses these three key requirements for next generation Layer 2 networks in the data center.

> **FabricPath and TRILL:** FabricPath is a Cisco proprietary pre-standard implementation of the IETF Transparent Interconnection of Lots of Links (TRILL). The ultimate aim of TRILL is to provide an IETF standards based approach to support a multi-path Ethernet network and provide interoperability between different vendors networks.
>
> Cisco supports FabricPath and will support TRILL in the near future, giving customers a choice which one to deploy. Customers implementing FabricPath today and wanting to switch over to TRILL will be able to do so with a simple software upgrade. For the source, see this website:
>
> http://blogs.cisco.com/datacenter/cisco-fabricpath-%E2%80%93-scaling-your-data-center-network/

8.3.1 FabricPath

In this section, we discuss the Cisco FabricPath feature. Its goal is to provide the robustness and scalability of Layer 3 (routing) while keeping the simplicity and transparency of Layer 2 (switching), as shown in Figure 8-10.

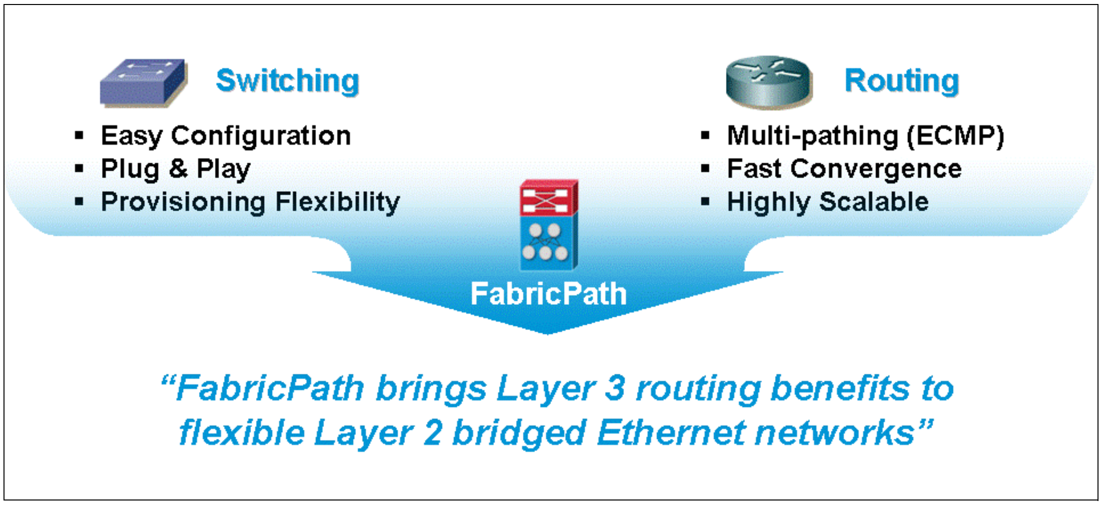

Switching
▪ Easy Configuration
▪ Plug & Play
▪ Provisioning Flexibility
Routing
▪ Multi-pathing (ECMP)
▪ Fast Convergence
▪ Highly Scalable
FabricPath
“FabricPath brings Layer 3 routing benefits to flexible Layer 2 bridged Ethernet networks”

Figure 8-10 FabricPath goal

Overview

A high level FabricPath architecture is shown in Figure 8-11. The resulting data center network is split in two domains:

- A FabricPath (FP) domain, shown in blue in Figure 8-11:

  An FP domain is made up by a Spine switch layer and a Leaf switch layer: both spine switches and leaf switches are assigned Switch IDs. Usually a given leaf switch is connected to all the spine switches, and a given spine switch is connected to all the leaf switches (3 stages Clos topology) but this is not required: Fabricpath does not impose a specific physical topology. Ports participating in an FP domain are labeled FP core ports. These ports form FabricPath adjacencies (using a modified version of IS-IS), do not run STP nor send BPDUs, and send/receive only FabricPath frames.

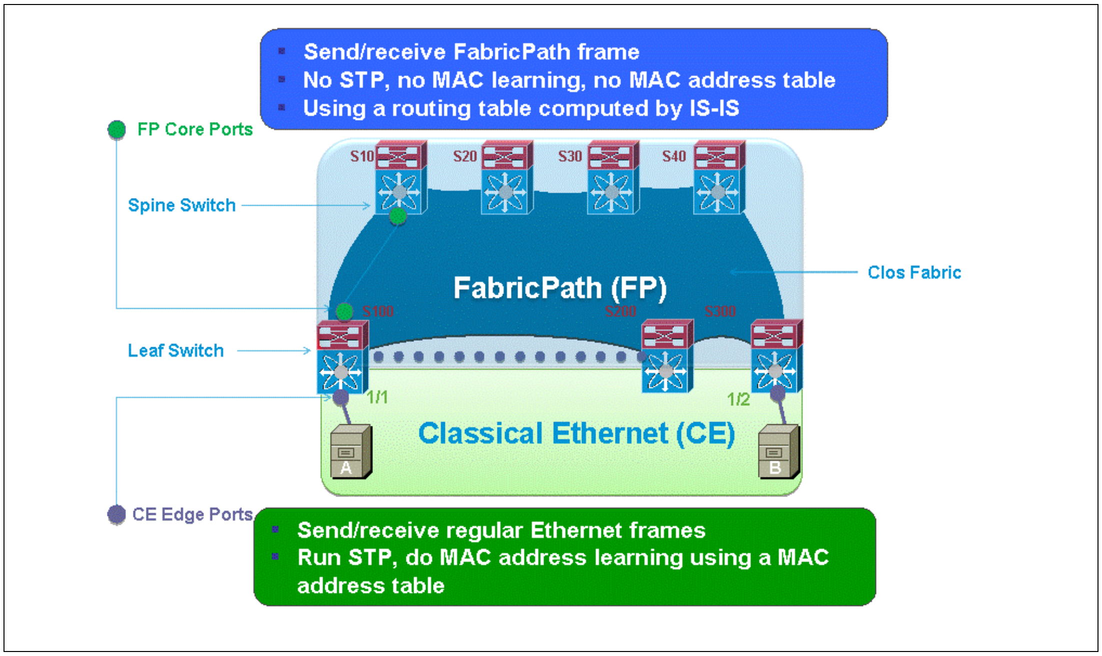

Send/receive FabricPath frame
No STP, no MAC learning, no MAC address table
Using a routing table computed by IS-IS
FP Core Ports
S10
S20
S30
S40
Spine Switch
FabricPath (FP)
Clos Fabric
S100
S200
S300
Leaf Switch
1/1
1/2
Classical Ethernet (CE)
A
B
CE Edge Ports
Send/receive regular Ethernet frames
Run STP, do MAC address learning using a MAC address table

Figure 8-11 FabricPath architecture overview diagram

- A Classical Ethernet (CE) domain - shown in green in Figure 8-11:

  Hosts connect to CE edge ports so the FP domain is transparent for the endpoints.

As shown in Figure 8-11, the Classical Ethernet domain ports still function the same way traditional Ethernet ports and switches work: they receive regular Ethernet frames, learn MAC addresses, populate local MAC address tables, and run spanning tree protocols and their variants.

On the other hand, the FabricPath domain obeys different rules: they handle special frames (FabricPath frames) and do not learn MAC addresses, but instead they rely on a MAC routing table, computed using the IS-IS routing protocol. Also FabricPath uses a new special VLAN type called FabricPath, used within the Fabric and on the CE interfaces on the edge. This VLAN is allowed to cross a FabricPath domain. CE VLANs are not allowed to cross a Cisco FabricPath domain.

As indicated in Figure 8-11 on page 443, FabricPath is usually an implementation of a three stage Clos network, which can be non-blocking if the stages are properly designed (number of Spine switches and number of links between Spine switches and Leaf switches), as Charles Clos demonstrated in 1953[1].

From a high level point of view, then FabricPath enhances the traditional Ethernet functionalities from three main standpoints:

- Hierarchical network addressing scheme: Instead of the flat MAC addressing scheme traditional Ethernet networks use. This is accomplished using encapsulation by adding a new FabricPath header in the Ethernet frame which contains:
  - Switch ID: Unique number assigned to help identify each device that is part of L2 Fabric
  - Port ID: Used to provide information about MAC-to-Interface association at L2 Fabric boundary
  - Tree ID: Unique number assigned to help identify each distribution “Tree”
  - TTL: Decrement at each hop to prevent frames looping infinitely in the fabric in case of unexpected failure
- Intelligent MAC address learning: Instead of traditional data plane learning it utilizes conversational learning - eliminating the need for aggregation switches to learn all the MAC addresses of the hosts attached to the downstream access ports
- Link-state based control plane: FabricPath uses trees, only for broadcast, multicast and unknown unicast but forwards unicast frames following the shortest path, this is possible because it uses a link-state control plane based on ISIS rather than distance vector like spanning tree. This allows FabricPath ports to be in a forwarding state on all paths (maximum 16) without any blocked links for unicast traffic. It also allows faster convergences in case of failure similar to routing protocol.

Figure 8-12 compares the traditional data center network architecture with a FabricPath data center network. The latter provides three main high level advantages:

- Easier application deployment: Any VLAN can be defined on any of the access switches without requiring special configurations or protocols so network constraints are eliminated from an application and server placement standpoint.
- Extended VM mobility range: FP access switches can reach each other at Layer 2, enabling simple administration and movement of virtual machines. A server does not have to be physically located in a particular PoD, making provisioning easy and dynamic. This is captured by the red VM mobility range box in Figure 8-12, which is no longer constrained by the aggregation switches block but can become data center wide.
- Increased capacity: Usage of up to 256 x 10 GE links and Layer 2 multipathing (so all links are active) increasing link utilization and total capacity
- Simplicity of configuration: Compared to STP and vPC based networks.
- Improved reliability: FabricPath provides N+1 redundancy so for example with 4 spines, the failure of one node will reduce the bandwidth by just 25% while with STP and vPC this will be reduced by 50%.

[1] *Clos, Charles (Mar 1953). “A study of non-blocking switching networks,” Bell System Technical Journal 32 (2): 406–424. ISSN 00058580*

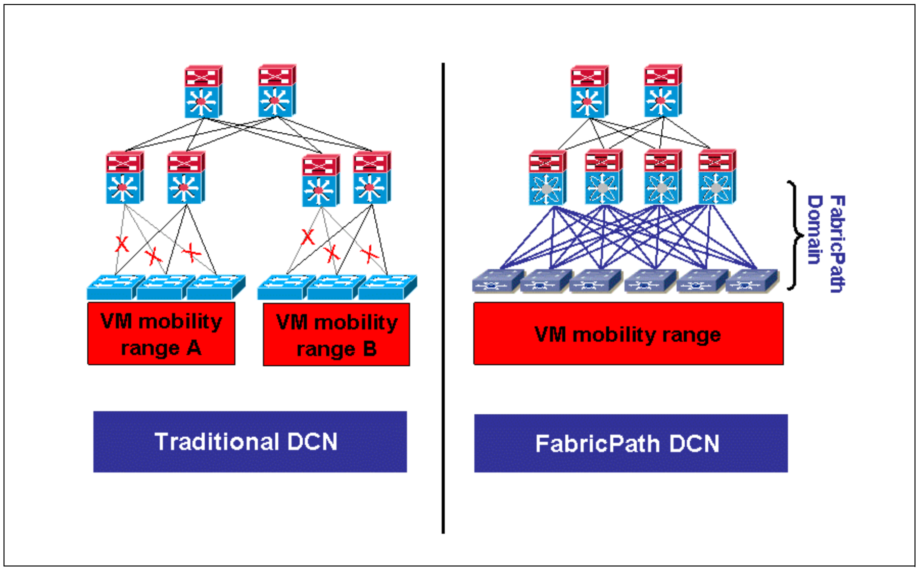


Figure 8-12 Advantages of a FabricPath enabled data center network architecture

These benefits are accomplished by leveraging the following functionalities:

- Conversational learning to remove MAC address scalability limitations: Spine switches do not learn any MAC addresses in a pure Layer 2 implementation. Spine switches forward frames based on destination Switch-ID.
- Spanning tree protocol independence: No reliance on Spanning Tree anymore. Each switch has a complete view of the Layer 2 topology, and it calculates the Layer 2 forwarding table based a shortest-path-first calculation.
- Traffic distribution for unicast: Unicast Layer 2 traffic can take multiple equal-cost Layer 2 paths.
- Traffic distribution for multicast on multiple distribution trees: Multicast traffic can be distributed along two multi-destination trees.
- More direct communication paths: Any topology is possible, so cabling two access or edge switches directly to each other creates a direct communication path, unlike what happens with Spanning Tree.
- Loop mitigation with TTL in the frame field: Layer 2 loops, as they are known of today in Spanning-Tree-Protocol-based Layer 2 networks, are mitigated by dropping frames that have been propagated across too many hops. The Layer 2 FabricPath frames include a Time to Live (TTL) field that is decremented at each hop. The value of the TTL field is 32.

Cisco FabricPath hardware and software requirements

Cisco FabricPath can be implemented on the following Cisco hardware platforms:

- Cisco Nexus 5500 products can act both as Leaf or Spine switches in a FabricPath topology with NX-OS Software Release 5.1(3)N1(1) or later.
  - N5K-C5548P-FA - 32 Fixed Ports 1/10GB
  - N5K-C5548UP-FA - 32 Fixed Ports 1/10GB
  - N5K-C5596UP-FA - 48 Fixed Ports 1/10GB
  - N5K-C5596T-FA - 32 10G BASE-T fixed ports and 16 1/10G SFP+ fixed ports
  - Additional license needed to run FabricPath (FabricPath - N5548-EL2-SSK9/N5596-EL2-SSK9)

- Cisco Nexus 6004 and Nexus 6001can act both as Leaf or Spine switches in a FabricPath topology with NX-OS Software Release 6.0 or later.
- Cisco Nexus 7000 Series products can act both as Leaf or Spine switches in a FabricPath topology:
  - F1 line cards using NX-OS 5.1 or above
  - F2 line cards using NX-OS 6.0 or above
- Proper licensing is needed to run FabricPath:
  You need the ENHANCED_LAYER2_PKG license installed. For details, visit:

http://www.cisco.com/en/US/docs/switches/datacenter/sw/nx-os/licensing/guide/Cisco_NX-OS_Licensing_Guide_chapter1.html.

Notes:

- The F1 cards are complemented by M1 or M2 cards for routing purposes. When using M1 or M2 cards in the same virtual device context (VDC) as the F1 card, routing is offloaded to the M1/M2 cards, and more routing capacity is added to the F1 card by putting more M1 or M2 cards into the same VDC as the F1 card.
- F2 cards cannot belong to the same VDC as F1, M1, or M2 cards. It is worth noticing that the F2E (Enhanced) line card can exist in the same VDC as M1 or M2 cards.
- For use for FabricPath, Nexus 2000 Fabric Extenders can be attached only to F2/F2E cards.
- Although all the examples and diagrams on the FabricPath section show the Nexus 5500 and/or Nexus 7000, the Nexus 6000 that has been recently announced by Cisco also supports FabricPath, including EvPC and vPC+ with FabricPath.

For an overview of Virtual Device Contexts (VDCs), see 9.2, “Virtual device contexts overview” on page 486.

8.3.2 FabricPath use cases

Cisco FabricPath technology can be leveraged whenever robust and stable Layer 2 domains need to be created. This is a fairly common requirement in modern data centers because of server virtualization requirements or server clustering requirements, for example. Because of this FabricPath may be a very attractive solution for very different type of environments.

In this section, we explore three main use cases in terms of positioning FabricPath in modern data centers:

1. High performance computing for low latency, high density Layer 2 large clusters
2. Internet Exchange Point for Layer 2 peering
3. Enterprise data center for data center wide workload flexibility

(1) High performance computing

High performance computing clusters aggregate a very large number of computing nodes to form a single high performing system for a variety of computing tasks in academia and industry. Usually these systems require a very high density of compute nodes and traffic patterns are typically kept within the system itself, with the vast majority of traffic being east to west (or server to server). Latency has to be kept as low as possible in order to avoid the data center network being a bottleneck for the HPC system. When possible, then, it would be preferable to have a non-blocking fabric to handle this type of traffic. Also being a large cluster, Layer 2 connectivity is usually a requirement. For more information on HPC data center requirements, see Chapter 1, “The modern data center” on page 1.

Figure 8-13 shows how FabricPath can be leveraged in an HPC environment to minimize latency and oversubscription inside the fabric.

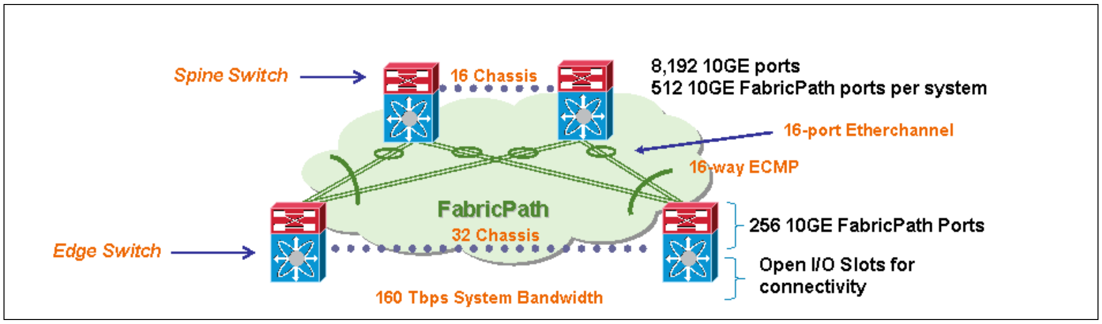

Spine Switch
16 Chassis
8,192 10GE ports
512 10GE FabricPath ports per system
16-port Etherchannel
16-way ECMP
FabricPath
32 Chassis
256 10GE FabricPath Ports
Open I/O Slots for connectivity
Edge Switch
160 Tbps System Bandwidth

Figure 8-13 FabricPath in a high performance computing data center

In this context, FabricPath enables building a high-density fat-tree network using Cisco Nexus 7018 switches for both the Spine and the Leaf layers. At the leaf layer there are 32 chassis, each using 256 10 GE ports (16 x 16-port EtherChannel links) to 16 spine switches with 256 ports for host connectivity (using 32 ports line cards). This creates a fully non-blocking Fabric with Layer 2 Equal Cost multipathing and all links can be used at the same time so latency is minimized and this configuration can accommodate up to 8,192 10 GE ports at the server access layer. This is just one example and with the F2 line card (which has 48 x 10 GE ports) this Fabric can accommodate more ports.

(2) Internet Exchange Point (IXP)

Another example of where FabricPath technology can be leveraged is the Layer 2 Internet Exchange Point, as shown in Figure 8-14.

Layer 2 peering enables multiple providers to peer their internet routers with one another leveraging a 10 GE non-blocking fabric that can scale to thousands of ports.

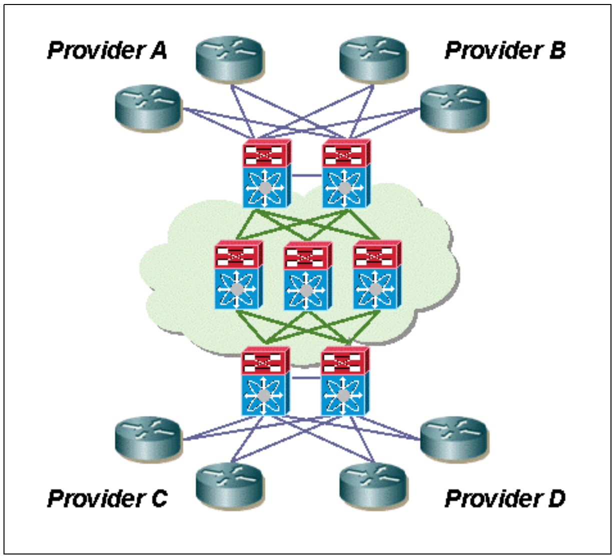


Figure 8-14 FabricPath in an Internet Exchange Point

For the purpose of this book, this use case will not be explored in detail. Instead, we focus on the next use case, which is very attractive to enterprises that are thinking about modernizing their data center network while addressing present and future requirements.

(3) Workload flexibility for enterprise data centers

The most interesting use case for the purpose of this publication is for enterprise data centers requiring data center network flexibility in terms of virtual servers placement and mobility. Figure 8-15 shows how FabricPath enables this flexibility by allowing you to decouple the logical network placement and the physical location of the virtual server itself.

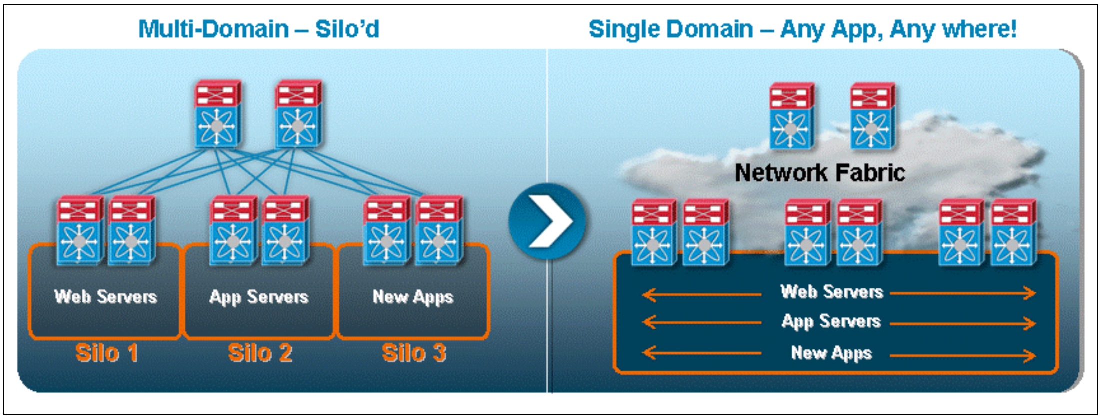


Figure 8-15 FabricPath for workload flexibility in an enterprise data center

These are the main benefits FabricPath brings in this context:

- Instead of having different siloed physical network environments for application, web, and database servers, with FabricPath, any VLAN can be defined anywhere in the fabric, allowing the utmost flexibility in terms of application deployment.
- Instead of being constrained by traditional Ethernet boundaries, with FabricPath, very large yet stable Layer 2 domains can be created, enabling data center wide VM mobility with very high performance in terms of available capacity because all the available links are used together.

Depending on the scope of interest for FabricPath, two main use cases can be further defined for the enterprise workload mobility scenario, as depicted in Figure 8-16:

- Intra-PoD Fabric: This scenario allows for a simplified adoption of FabricPath in terms of impact on the existing environment. In fact a FabricPath PoD can be created for a specific application environment that benefits from the increased capacity, stability and flexibility of a FabricPath enabled Layer 2 domain. With this scenario, shown at the left hand side in Figure 8-16, seamless integration with Layer 3 can be achieved also allowing to scale to a very high number of MAC addresses with virtually unlimited bandwidth.
- Inter-PoD Fabric: This scenario, shown at the right hand side in Figure 8-16, the FabricPath scope is data center wide. VLANs can be either terminated at the data center network aggregation layer (FabricPath leaf layer) or be stretched between PoDs. If integration with STP is needed to accommodate legacy access layers, STP will not be extended between PoDs. Using this approach some VLANs can be just confined in a single PoD, while others can be data center wide. This scenario can be further broken down into two additional options:
  - Converged Layer 2 and Layer 3 core: The FabricPath domain provides both Layer 2 and Layer 3 functionalities
  - Parallel Layer 2 core: The FabricPath domain provides pure Layer 2 connectivity across PoDs and it is deployed in parallel with the existing Layer 3 core. These can also be deployed on the same Nexus 7000 chassis as separate VDCs to decrease the physical footprint of the overall solution.

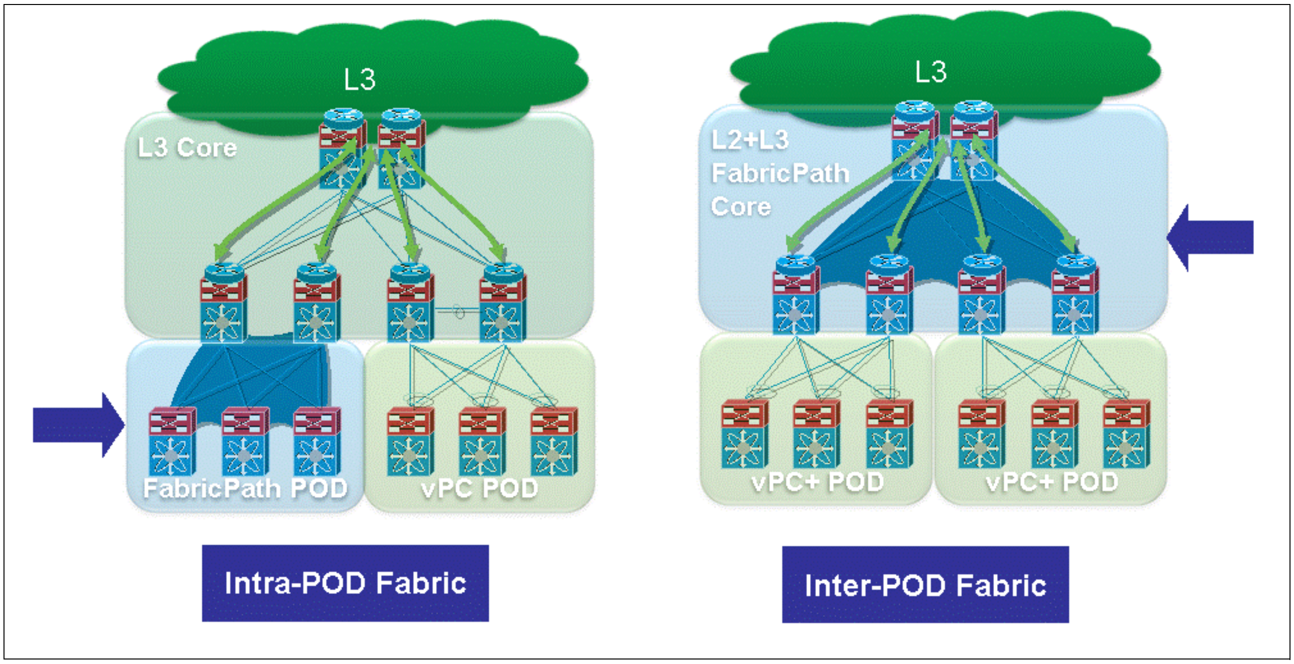


Figure 8-16 Intra-PoD Fabric versus Inter-PoD Fabric

Regarding the Inter-PoD scenario, this could also allow to reduce the number of layers in the data center network hierarchy, for example, from three to two layers. In fact, there is enough bandwidth and port density on the core Nexus 7000s for aggregating the whole network so there is no need to deploy a dedicated layer for aggregating and isolating PoDs. In the next section, we focus on two tiered FabricPath designs, as these will be the most common deployment models, especially for greenfield installations.

8.3.3 FabricPath design considerations

In this section, we capture some key considerations when designing data center networks with Cisco FabricPath. As discussed during this chapter there are a lot of different options when planning for a FabricPath solution because of the different hardware platforms that support the technology and the different designs that can be used to satisfy a given set of requirements.

In this section, we cover the following topics:

- Main factors guiding the FabricPath design
- Layer 3 integration considerations
- Server attachment considerations
- Legacy access layer integration considerations
- Comprehensive high level design sample so that the reader can find a starting point that can be used in different situations

Main factors guiding the FabricPath design

We now start by listing and discussing the main factors guiding the FabricPath solution design process:

- Port density and oversubscription:
  - What is the current requirement in terms of edge ports?
    - Nexus 5500, Nexus 6000 and Nexus 7000 platforms can accommodate different port densities at the leaf layer, and can be combined with the Nexus 2000 Fabric Extenders to provide a high number of edge ports.
  - What is the expected growth in terms of edge ports?
    - Nexus 5500,Nexus 6000 and Nexus 7000 platforms can accommodate different port densities at the spine layer, and can be combined with the Nexus 2000 Fabric Extenders to provide a high number of edge ports. Nexus 5000 and 6000 can support up to 24 FEXs and the Nexus 7000 up to 48 FEXs.
  - What level of oversubscription is acceptable?
    - More switches at the spine layer and more links belonging to the EtherChannel uplinks between spine and leaf layers can result in different oversubscription ratios.
- Routing integration and L2/L3 boundary options:
  - Where will the L2/L3 boundary be placed?
    - It can be placed both at the spine layer or at the leaf layer.
  - How much capacity for inter-VLAN and North-South traffic is needed?
- Flexibility and scope:
  - How flexible is the design with respect to VLAN extension and VM mobility?
- Host scale per FabricPath domain:
  - What is the total number of unique MAC entries in the CE domain and/or ARP entries the design can support?
    - M1/M2 cards can handle up to 128K MAC addresses while F1 and F2 cards can learn up to 16K MAC addresses per System on a Chip SOC (two ports or four ports respectively), as described in Chapter 2. The Nexus 5500 can learn up to 32K MAC addresses and the Nexus 6000 support for up to 256,000 MAC addresses.
  - What is the number of VLANs and how many hosts per VLAN?

Layer 3 integration considerations

We now discuss in more detail the possibilities in terms of routing integration for FabricPath based data centers. We focus on two key scenarios:

- Routing integration with two spine switches
- Routing integration with four spine switches

Note: It is possible to include up to 16 spine switches within a FabricPath domain.

Figure 8-17 shows the available options in terms of Layer 3 integration for a two spine switch design.

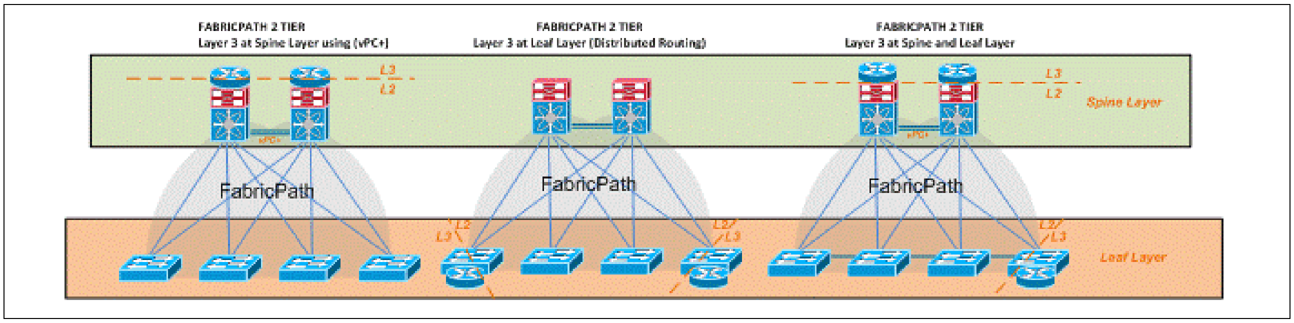


Figure 8-17 Layer 3 integration for a two spine FabricPath design

Going from left to right in the diagram, we have the following models:

- L2/L3 boundary at the spine layer:

  This model is the most consistent with current data center network deployments in terms of having the L2/L3 boundary at the aggregation layer, or at the core layer in case of a collapsed two tiered design, as shown in the diagram. This design allows to have active active default gateways using vPC+ and provides any VLAN anywhere at the access layer. The Layer 3 function is located on the Nexus 7000 chassis. Network services appliances such as firewalls and application delivery controllers would be connected at the spine layer using this model, and would connect to CE ports at the FP spine layer.

- Distributed routing (L2/L3 boundary at the leaf layer):

  This model is unique to FabricPath because the spine layer will become a pure spine (just FP core ports), forwarding only FabricPath traffic. If there is only one leaf switch (usually two to provide redundancy) in charge of performing routing function this becomes a “centralized routing” design because all North-South traffic flows will pass through the L3 gateways. Using this model network services appliances such as firewalls and application delivery controllers would be connected at the leaf layer. This model simplifies the spine layer design because the L2/L3 boundary and the L4-7 network services are decoupled from the spine. The drawback of having a single pair of L3 gateways at the leaf layer is that all traffic that needs Layer 3 will pass through those and this may not be scalable from a bandwidth perspective. Also latency may be higher because there are additional hops compared to the previous model. An option to address those concerns is to have multiple SVIs, spreading the load across different pairs of leaf switches as shown in Figure 8-17.

- Hybrid model (L2/L3 boundary at both layers):

  This model mixes the two models that were presented earlier so it is possible to have L2/L3 boundaries at both layers in the FabricPath hierarchy.

Figure 8-18 shows the available options in terms of Layer 3 First Hop Routing Protocol integration for a four spine switch design. All these options assume that Layer 3 will be integrated at the spine layer.

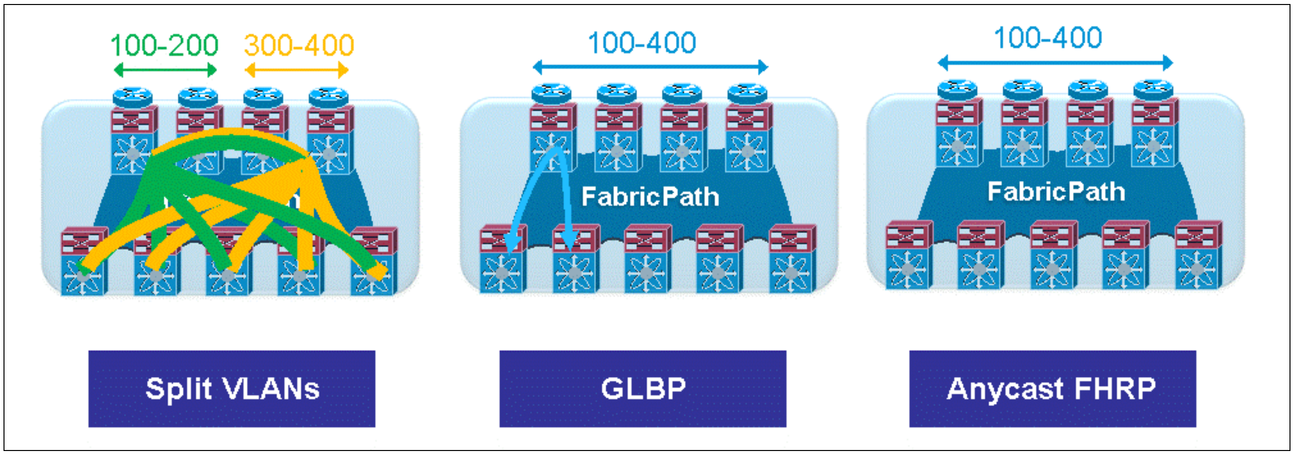


Figure 8-18 First hop routing protocol integration for a four spine FabricPath design

Going from left to right, the diagram shows the following cases:

- Active/Active HSRP with split VLANs: traffic is polarized to a pair of spine switches on a VLAN basis. This is similar to creating two Spanning Trees for two different sets of VLANs: traffic will be load balanced in a static fashion. Using this approach inter-VLAN traffic can be suboptimal
- Gateway Load Balancing Protocol (GLBP): each host is pinned to a single gateway, which is determined via round robin. In fact all the gateways share a common virtual IP address that is then mapped to the multiple virtual MAC addresses (up to 4). Dynamic load sharing can be achieved toward the L3 exit points based on which MAC each server learns through ARP but only one gateway will be active from the host point of view
- Anycast FHRP: All gateways will be active for a given host, allowing multipathing. This possibility will be available in the future for routing integration with FabricPath, with NX-OS 6.2.

8.3.4 vPC+

The vPC+ enhancement was built to improve multipathing when vPC is combined with FabricPath. In essence it allows the creation of a virtual switch ID for the Edge switches running a vPC down to a Classical Ethernet (CE) domain as shown in Figure 8-19.

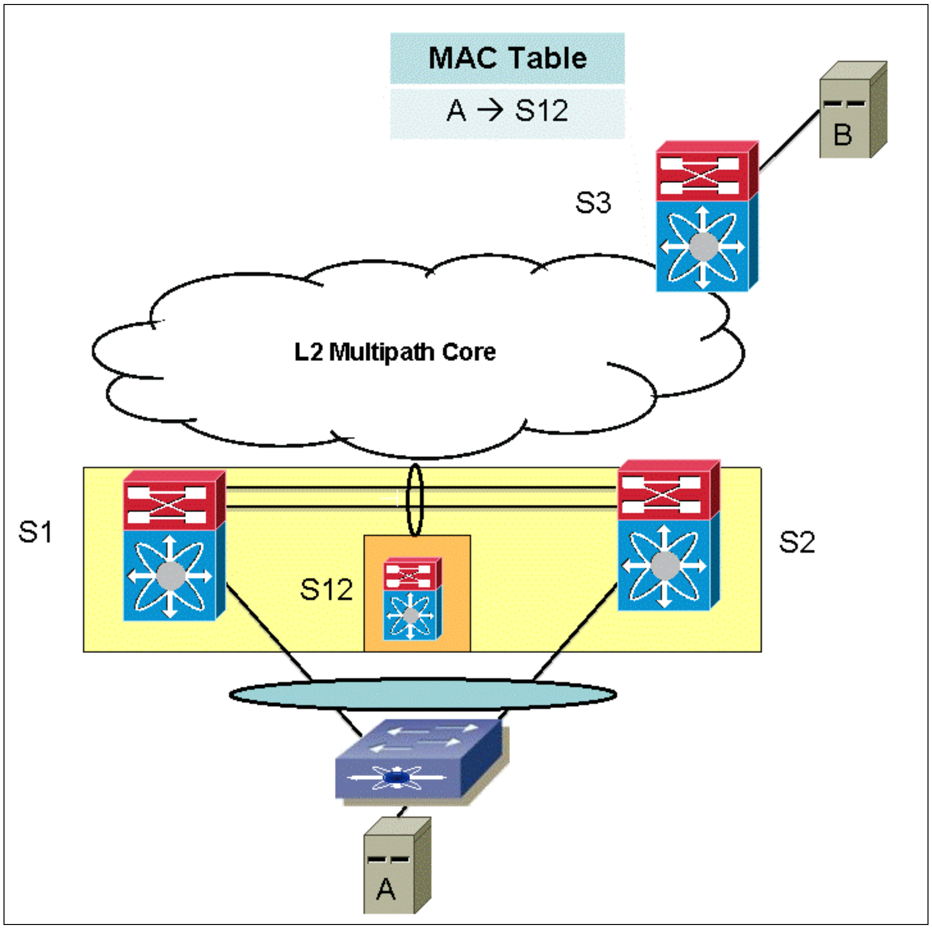


Figure 8-19 The vPC+ enhancement

The example in Figure 8-19 works well to illustrate how vPC+ works. Without vPC+ when S3 receives the traffic from Host A it will add an entry in its MAC table pointing at either S1 or S2, whichever received and forwarded the packet from Host A. That is not ideal because it minimizes the multipath capability of FabricPath. On the other hand if S1 and S2 could also advertise a path to virtual switch, in this example S12, and they tagged the packets from Host A with that switch ID, then S3 would add MAC A as pointing at S12. Since S3 has 2 different paths to S12, via S1 and S2, it can now use both paths to Host A (via S1 and S2).

The vPC+ feature provides these capabilities:

- Allows CE switches to interoperate with FP switches
- Provides active/active L2 paths for dual-homed CE devices
- Offers resilient architecture for CE/FP gateways
- Provides Active/Active HSRP for routed traffic
- Brings enhancements to multicast traffic support compared to vPC
- Simplifies connectivity to non FabricPath capable devices or your legacy Classical Ethernet (CE) environment

Server attachment considerations

Another critical aspect to consider when designing data center networks with FabricPath will be the available options in terms of host connectivity, as depicted in Figure 8-20.

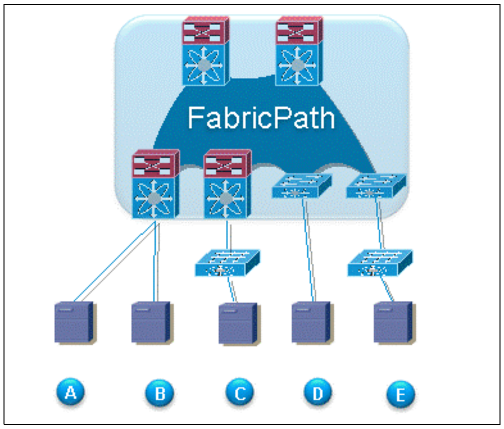


Figure 8-20 Host connectivity options with FabricPath

Going from left to right in the diagram:

- A) 1/10 GE connectivity to a Nexus 7000 F1 module. CX-1 cabling provides a cost-effective solution.
- B) 1/10 GE connectivity to a Nexus 7000 F2 module. CX-1 cabling provides a cost-effective solution.
- C) 1/10 GE connectivity to a FEX module attached to a Nexus 7000 F2 I/O module
- D) 1/10 GE connectivity to a Nexus 5500
- E) 1/10 GE connectivity to a FEX module attached to a Nexus 5500

Figure 8-21 provides additional details on the possibilities for option C (at the bottom of the diagram) and for option E (at the top of the diagram) with a generic IBM System x Series or POWER server.

The approaches shown at the far right hand side of Figure 8-21 are usually the preferred ones because all links will be active using these topologies.

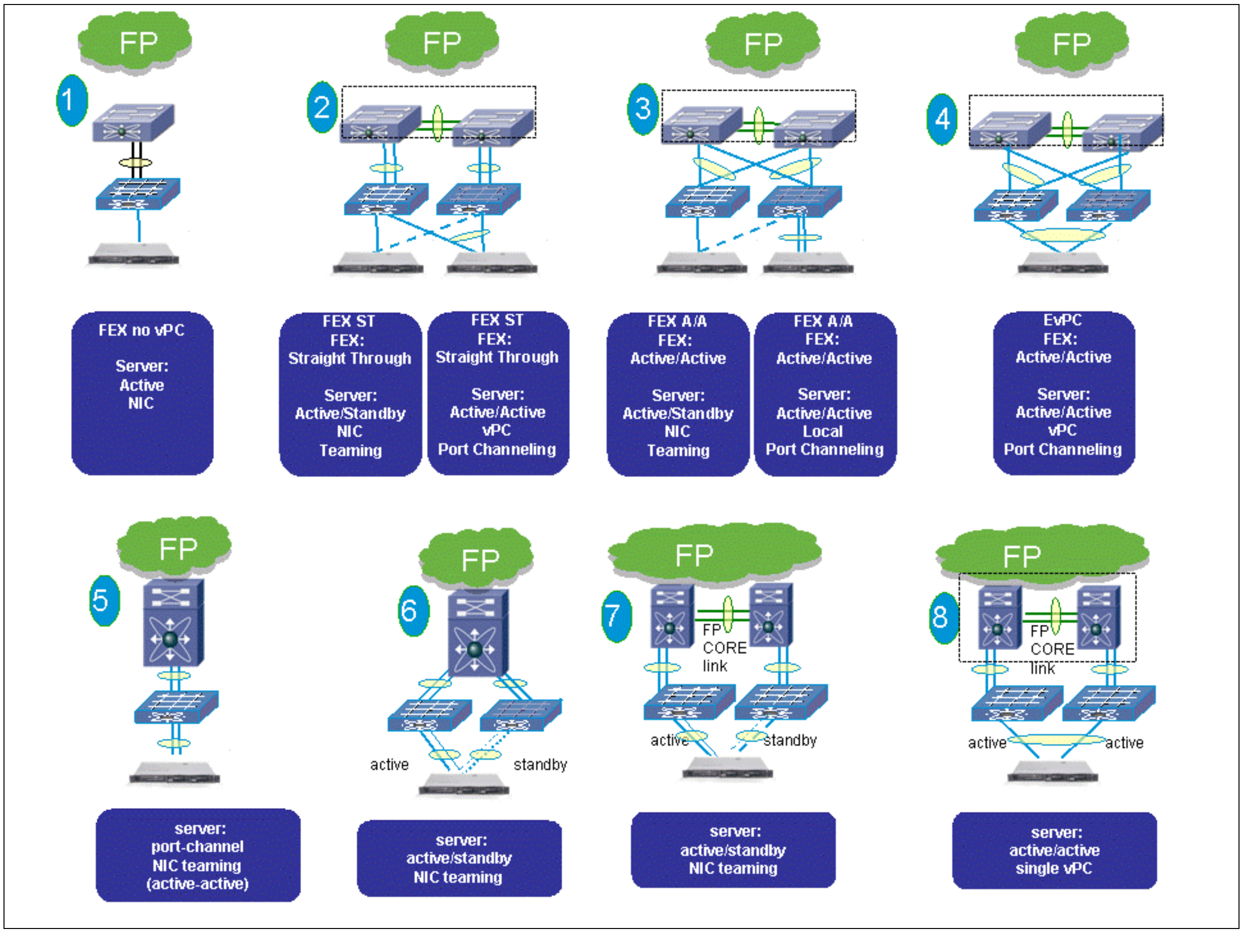


Figure 8-21 FEX connectivity options for IBM System x or POWER

Legacy access block integration

When connecting legacy access blocks (made of devices that do not support FabricPath) to FabricPath domains it is important to highlight that the FabricPath leaf layer has to be the root for spanning tree protocol because superior level BPDUs will be discarded. In fact no BPDUs are forwarded across the FabricPath domain. Nevertheless an optional mechanism allows to propagate Topology Change Notifications (TCNs) if needed. When connecting a Catalyst based legacy access layer, for example, this has to be connected through CE ports and a port channel can be used from the downstream access block to obtain active/active dual homing.

Another common use case may be integrating a Catalyst 6500 VSS with a Nexus 7000 vPC. Additional information on this scenario can be found here:

http://www.cisco.com/en/US/prod/collateral/switches/ps5718/ps708/white_paper_c11_589890.html

Figure 8-22 provides an overview diagram that captures some of the connectivity options discussed in this section. The most important design decision when designing a FabricPath data center network is the spine switch platform selection. Even if theoretically possible it is highly recommended to deploy homogeneous platforms for the spine layer. Features and port density guide the decision in this case.

On the other hand, it may be rather common to mix and match different devices at the leaf layer to accommodate different server platforms requirements, for example, as shown in Figure 8-22.

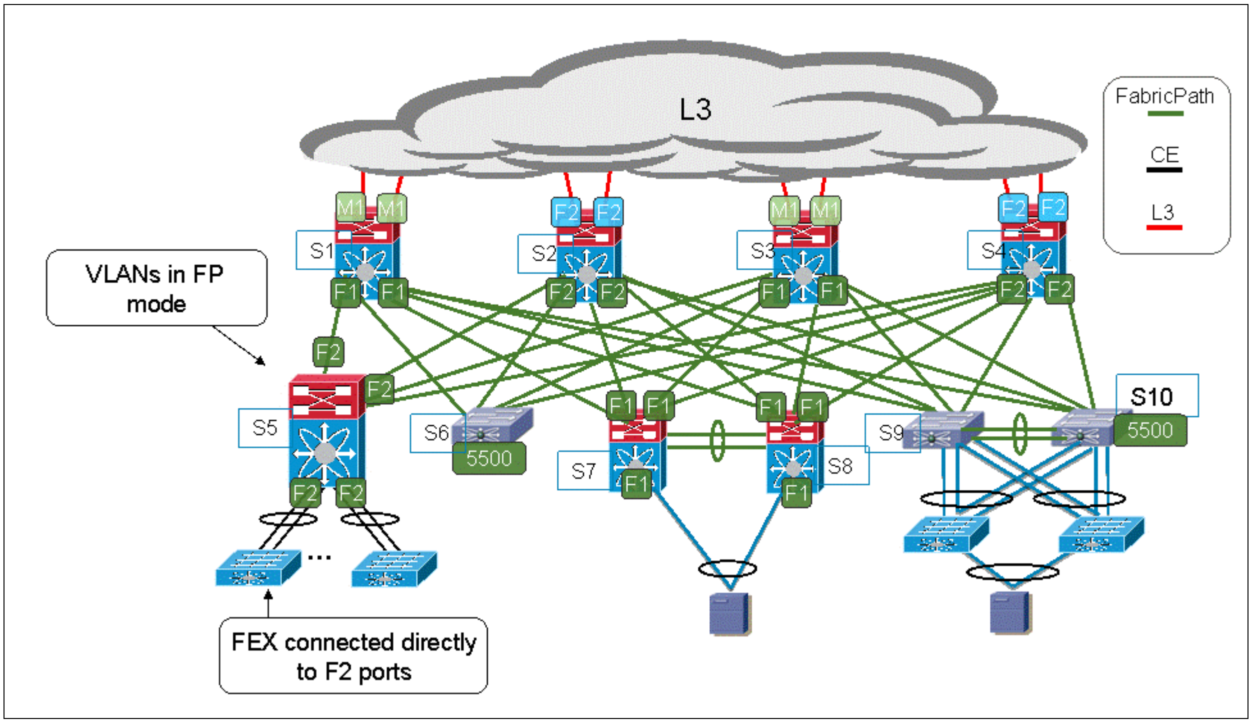


Figure 8-22 High level FabricPath architecture

It is always important to keep the FabricPath scalability limits in mind in order to future proof the network for future requirements and expected growth.

You can find the verified FabricPath scalability limits on the Nexus 7000 platform at this link:

`http://www.cisco.com/en/US/docs/switches/datacenter/sw/verified_scalability/b_Cisco_Nexus_7000_Series_NX-OS_Verified_Scalability_Guide.html#reference_3AD0536C32FF4B499A0936409729951D`

You can find the FabricPath scalability limits on the Nexus 5000 platform at this link (specifying the NX-OS release level):

`http://www.cisco.com/en/US/products/ps9670/products_installation_and_configuration_guides_list.html`

Note: The Nexus 6000 scalability limits are initially the same as the Nexus 5500, however, for the latest information, check the Cisco website:

`http://www.cisco.com`

8.3.5 IETF TRILL and FabricPath

Transparent Interconnection of Lots of Links (TRILL) is an IETF standard[2] for Layer 2 multipathing. The standard is driven by multiple vendors, including Cisco, and has now moved from the draft state to the proposed standard state.

[2] For more information on TRILL visit `http://datatracker.ietf.org/doc/rfc5556/`

TRILL is also known as RBridges, invented and proposed by Radia Perlmann, who invented the spanning tree protocol in the 1980s. In 2004 she proposed a new solution that could replace the spanning tree protocol. A short introduction of RBridges is available from her paper at this website:

http://www.ieee-infocom.org/2004/Papers/26_1.PDF

FabricPath capable hardware is also TRILL capable - Cisco is planning to support TRILL in the future with a software upgrade.

Nevertheless, there are some key differences between FabricPath and TRILL:

- vPC+ allows the host to use active/active paths to the leaf switches, while TRILL only specifies active/standby connectivity.
- FabricPath can leverage active/active FHRP while TRILL leverages active/standby HSRP.
- With FabricPath, the fabric is the root bridge for the CE region, while for TRILL the CE region is part of the TRILL fabric.
- FabricPath will allow multiple topologies within the fabric, while TRILL does not support that function.

Table 8-1 summarizes the main differences between TRILL and FabricPath in terms of supported features.

Table 8-1 FabricPath and TRILL feature comparison

Feature	Cisco FabricPath	IETF TRILL
Ethernet Frame Routing	Yes	Yes
vPC+	Yes	No
FHRP Active-Active	Yes	No
Multiple topologies	Yes	No
Conversational MAC learning	Yes	No
Interswitch links	Point-to-point only	Point-to-point or shared

Cisco will continue to propose features that are unique to FabricPath and therefore outside of the TRILL standard in the coming years.

However, Cisco will continue to support TRILL at the edge of the fabric for interoperability with IBM and other vendor networks. FabricPath, being a proprietary protocol, means that it will not function outside a Cisco-only network.

8.4 Data Center Network overlays

This section focuses on Data Center Network overlays as a way of connecting Data Center Networks via Layer 2 and making sure the client to server traffic patterns are always optimized regardless of the physical location of the VM itself, which may vary (IP mobility or path optimization).

In general, an overlay network is created between two devices (overlay termination endpoints) that encapsulate the overlay protocol over a network that does not need to understand the overlay protocol. In this sense the overlay is transparent to the transport network that is not able to de-capsulate the overlay protocol.

The following technologies can be classified as overlays in the context of data center networking:

- Overlay Transport virtualization (OTV) is covered in 8.4.1, “Overlay Transport Virtualization”.
- Virtual Private LAN Services (VPLS) is covered in 8.4.4, “VPLS”.
- Location and Identity Separation Protocol (LISP) is covered in 8.4.5, “LISP”.
- Virtual eXtensible LAN (VXLAN) is described in 5.3.10, “VXLAN: Virtual eXtensible Local Area Network” on page 277.

8.4.1 Overlay Transport Virtualization

Overlay Transport Virtualization (OTV) is an IP-based overlay that has been designed from the ground up to provide Layer 2 extension capabilities over any transport infrastructure: Layer 2 (L2) based, Layer 3 (L3) based, IP switched, MPLS, and so on. The only requirement from the transport infrastructure is providing IP connectivity between the remote data center sites. OTV provides an overlay that enables L2 connectivity between separate L2 domains while keeping these domains independent and preserving the fault-isolation, resiliency, and load-balancing benefits of an IP-based interconnection.

OTV introduces the concept of “MAC routing,” which means that a control plane protocol is used to exchange MAC reachability information between network devices providing LAN extension functionality. OTV also introduces the concept of dynamic encapsulation for L2 traffic flows that need to be sent to remote locations. Each Ethernet frame is individually encapsulated into an IP packet and delivered across the transport network. This eliminates the need to establish virtual circuits, called pseudowires, between the data center locations. Immediate advantages include improved flexibility when adding or removing sites to the overlay, more optimal bandwidth utilization across the WAN (specifically when the transport infrastructure is multicast enabled), and independence from the transport characteristics (Layer 1, Layer 2 or Layer 3).

OTV provides a native built-in multi-homing capability with automatic failure detection, critical to increasing high availability of the overall solution. Two or more devices can be leveraged in each data center to provide LAN extension functionality without running the risk of creating an end-to-end loop that would jeopardize the overall stability of the design. This is achieved by leveraging the same control plane protocol used for the exchange of MAC address information, without the need of extending the Spanning Tree Protocol (STP) across the overlay.

The following sections detail the OTV technology and introduce alternative design options for deploying OTV within, and between, data centers.

Components and terminology

A high level view of OTV is shown in Figure 8-23.

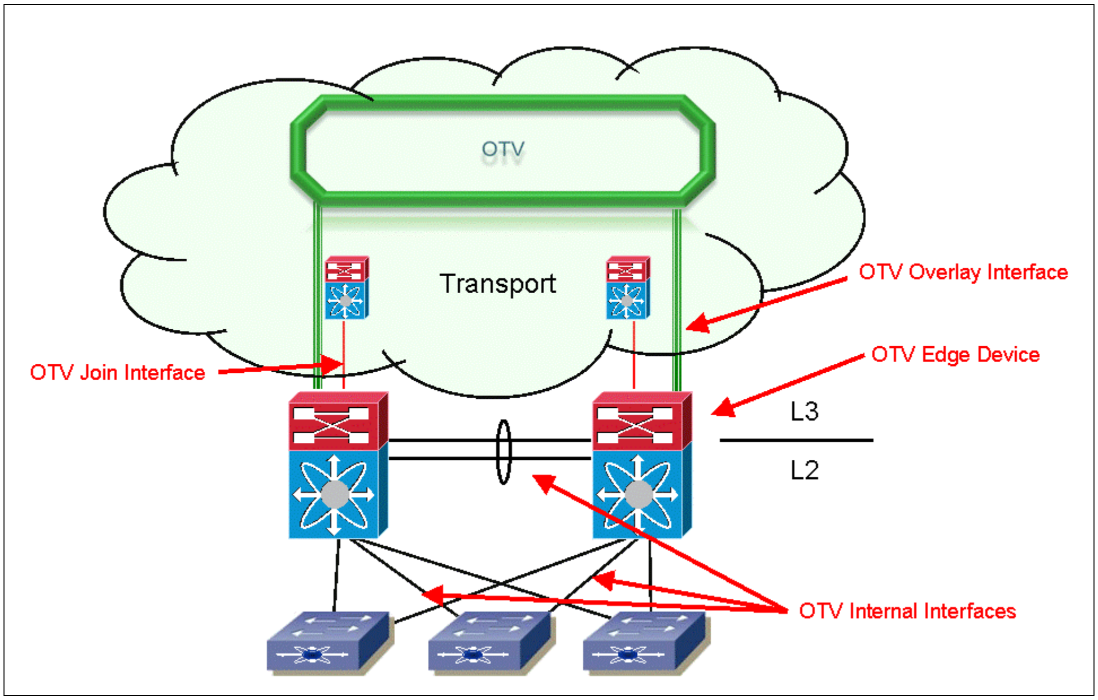


Figure 8-23 OTV overview

OTV edge device

The edge device is responsible for performing all of the OTV functions. It receives the Layer 2 traffic for all VLANs that need to be extended to remote locations and dynamically encapsulates the Ethernet frames into IP packets that are then sent across the transport infrastructure. The edge device can be located at the aggregation layer as well as at the core layer depending on the network topology of the site. The OTV edge devices are the only ones that are aware of the existence of the overlay network. All other devices can continue to function and be configured as they always were. It is recommended that at least two OTV edge devices are deployed at each data center site to improve resiliency. The OTV edge devices can be positioned in different parts of the data center, the choice depends on network topology of the site.

OTV internal interfaces

To perform OTV functionality, the edge device must receive the Layer 2 traffic for all VLANs that need to be extended to remote locations. The Layer 2 interfaces, where the Layer 2 traffic is usually received, are named internal interfaces. Internal interfaces are regular Layer 2 interfaces configured as access or trunk ports. Trunk configuration is typical given the need to concurrently extend more than one VLAN across the overlay. There is no need to apply OTV-specific configuration to these interfaces. Also, typical Layer 2 functions (like local switching, spanning-tree operation, data plane learning, and flooding) are performed on the internal interfaces.

OTV Join interface

The Join interface is used to source the OTV encapsulated traffic and send it to the Layer 3 domain of the data center network. A single Join interface can be defined and associated with a given OTV overlay. Multiple overlays can also share the same Join interface.

The Join interface is used by the edge device for different purposes:

- “Join” the Overlay network and discover the other remote OTV edge devices.
- Form OTV adjacencies with the other OTV edge devices belonging to the same VPN.
- Send/receive MAC reachability information.
- Send/receive unicast and multicast traffic.

OTV Overlay interface

The Overlay interface is a logical multi-access and multicast-capable interface that must be explicitly defined by the user and where the entire OTV configuration is applied. Every time the OTV edge device receives a Layer 2 frame destined for a remote data center site, the frame is logically forwarded to the Overlay interface. This instructs the edge device to perform the dynamic OTV encapsulation on the Layer 2 packet and send it to the Join interface toward the routed domain.

OTV control plane considerations

One fundamental principle on which OTV operates is the use of a control protocol running between the OTV edge devices to advertise MAC address reachability information instead of using data plane learning. However, before MAC reachability information can be exchanged, all OTV edge devices must become “adjacent” to each other from an OTV perspective. This can be achieved in two ways, depending on the nature of the transport network interconnecting the various sites. If the transport network connecting the OTV sites together is multicast enabled, a specific multicast group can be used to exchange the control protocol messages between the OTV edge devices. If the transport is not multicast enabled, an alternative deployment model is available, where one (or more) OTV edge device can be configured as an “Adjacency Server” to which all other edge devices register and communicates to them the list of devices belonging to a given overlay.

From a MAC learning perspective:

- OTV proactively advertises MAC reachability using the control plane.
- MAC addresses are advertised in the background once OTV has been configured.
- No specific configuration is required.
- IS-IS is the OTV Control Protocol running between the edge devices. There is no need to learn how IS-IS works.

Neighbor discovery

Before any MAC address reachability can be advertised, the OTV edge devices must discover each other and build a neighbor relationship.

Neighbor discovery using a multicast enabled transport infrastructure

Assuming the transport is multicast enabled, all OTV edge devices can be configured to join a specific Any Source Multicast (ASM) group where they simultaneously play the role of receiver and source. If the transport is owned by a Service Provider, for example, the enterprise will have to negotiate the use of this ASM group with the SP.

Each OTV edge device sends an IGMP report to join the specific ASM group used to carry control protocol exchanges. All of the edge devices join the group as hosts, leveraging the Join interface. This happens without enabling PIM on this interface. The only requirement is to specify the ASM group to be used and associate it with a given Overlay interface. The OTV control protocol running on an OTV edge device sends Hellos to all other edge devices using this multicast group.

For this to happen, the original frames must be OTV-encapsulated, adding an external IP header. The source IP address in the external header is set to the IP address of the Join interface of the edge device, whereas the destination is the multicast address of the ASM group dedicated to carry the control protocol. The resulting multicast frame is then sent to the Join interface toward the Layer 3 network domain.

The multicast frames are carried across the transport network and optimally replicated to reach all the OTV edge devices that joined that ASM multicast group. Since all of the OTV edge devices join the ASM group all of them receive the OTV Hello messages. The receiving OTV edge devices de-capsulate the packets and the Hellos are passed to the control protocol process.

The same process is repeated by the other devices and the end result is the creation of OTV control protocol adjacencies between all the edge devices. The use of the ASM group as a vehicle to transport the Hello messages allows the edge devices to discover each other as if they were deployed on a shared LAN segment. The LAN segment is basically implemented via the OTV overlay.

Two important considerations for OTV control protocol are as follows:

- This protocol runs as an “overlay” control plane between OTV edge devices, which means there is no dependency with the routing protocol (IGP or BGP) used in the Layer 3 domain of the data center, or in the transport infrastructure.
- The OTV control plane is transparently enabled in the background after creating the OTV Overlay interface and does not require explicit configuration. Tuning parameters, like timers, for the OTV protocol is allowed, but this is expected to be more of a corner case than a common requirement.

Neighbor discovery using a unicast-only transport infrastructure

OTV can be deployed with unicast-only transport. As previously described, a multicast enabled transport infrastructure lets a single OTV update or Hello packet reach all other OTV devices by virtue of leveraging a specific multicast control group address.

The OTV control plane over a unicast-only transport works exactly the same way as OTV with multicast mode. The only difference is that each OTV devices would need to create multiple copies of each control plane packet and unicast them to each remote OTV device part of the same logical overlay. Because of this head-end replication behavior, leveraging a multicast enabled transport remains the recommended way of deploying OTV in cases where several DC sites are involved. At the same time, the operational simplification brought by the unicast-only model (removing the need for multicast deployment) can make this deployment option very appealing in scenarios where LAN extension connectivity is required only between a few DC sites.

To be able to communicate with all the remote OTV devices, each OTV node needs to know a list of neighbors to replicate the control packets to. Rather than statically configuring in each OTV node the list of all neighbors, a simple dynamic means is used to provide this information. This is achieved by designating one (or more) OTV edge device to perform a specific role, named Adjacency Server. Every OTV device wishing to join a specific OTV logical overlay, needs to first “register” with the Adjacency Server (by start sending OTV Hello messages to it). All other OTV neighbor addresses are discovered dynamically through the Adjacency Server. Thereby, when the OTV service needs to be extended to a new DC site, only the OTV edge devices for the new site need to be configured with the Adjacency Server addresses. No other sites need additional configuration.

The reception of the Hello messages from all the OTV edge devices helps the Adjacency Server to build up the list of all the OTV devices that should be part of the same overlay (named unicast-replication-list). This list is periodically sent in unicast fashion to all the listed OTV devices, so that they can dynamically be aware about all the OTV neighbors in the network.

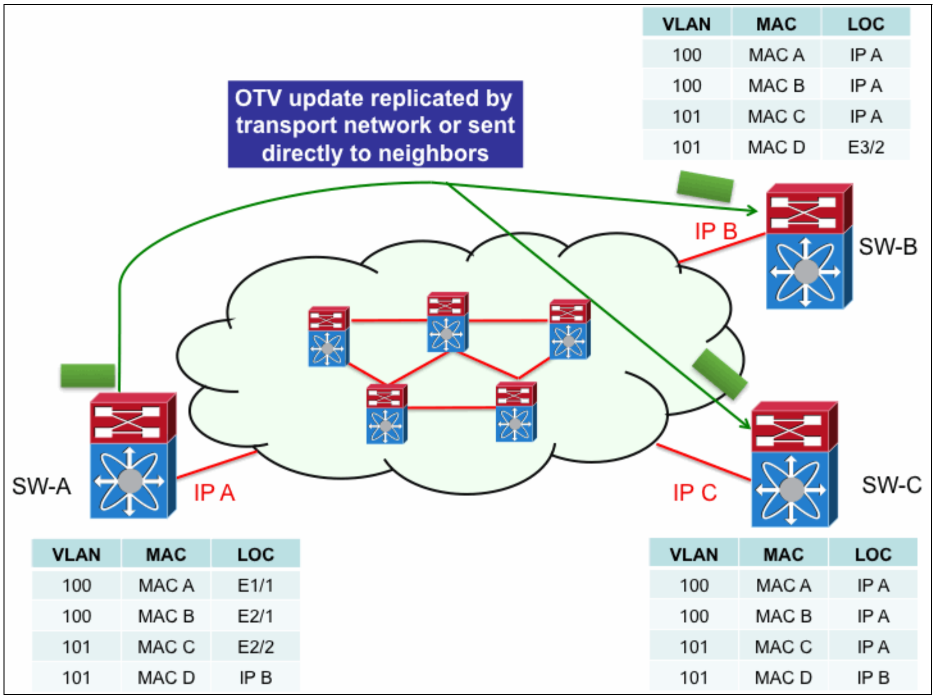
A pair of Adjacency Servers can be deployed for redundancy purposes. These Adjacency Server devices are completely stateless between them, which implies that every OTV edge device (OTV clients) should register its existence with both of them. For this purpose, the primary and secondary Adjacency Servers are configured in each OTV edge device. However, an OTV client will not process an alternate server's replication list until it detects that the primary Adjacency Server has timed out. Once that happens, each OTV edge device will start using the replication list from the secondary Adjacency Server and push the difference to OTV.

Note: The routing protocol used to implement the OTV control plane is IS-IS. There is no need to learn how IS-IS works. It is configured automatically without requirements for any user intervention.

From a security perspective, it is possible to leverage the IS-IS HMAC-MD5 authentication feature to add an HMAC-MD5 digest to each OTV control protocol message. The digest allows authentication at the IS-IS protocol level, which prevents unauthorized routing message from being injected into the network routing domain. At the same time, only authenticated devices will be allowed to successfully exchange OTV control protocol messages between them and hence to become part of the same Overlay network.

8.4.2 OTV MAC address advertisement

Figure 8-24 depicts how the OTV new MAC advertisements work.



OTV update replicated by transport network or sent directly to neighbors
IP B
SW-B
SW-A
IP A
IP C
SW-C

VLAN	MAC	LOC
100	MAC A	IP A
100	MAC B	IP A
101	MAC C	IP A
101	MAC D	E3/2

VLAN	MAC	LOC
100	MAC A	E1/1
100	MAC B	E2/1
101	MAC C	E2/2
101	MAC D	IP B

VLAN	MAC	LOC
100	MAC A	IP A
100	MAC B	IP A
101	MAC C	IP A
101	MAC D	IP B

Figure 8-24 OTV MAC advertisement

When the OTV edge device named SW-A in Figure 8-24 on page 462 learns three new MAC addresses (MAC A, MAC B, and MAC C) on its internal interface (this is done via traditional MAC address learning) an OTV Update message is created containing information for MAC A, MAC B, and MAC C. The message is OTV encapsulated and sent into the Layer 3 transport. If the transport is using multicast the IP destination address of the packet in the outer header is the multicast group used for control protocol exchanges. On the other hand if unicast is being used then an update is sent to each individual neighbor using its unicast address as the destination. A single OTV update can contain multiple MAC addresses for different VLANs.

The OTV Update is delivered to all remote edge devices which de-capsulate it and hand it to the OTV control process.

The MAC reachability information is imported in the MAC Address Tables (CAMs) of the edge devices. As noted in Figure 8-24 on page 462, the only difference with a traditional CAM entry is that instead of having associated a physical interface, these entries refer the IP address of the Join interface of the originating edge device.

The same control plane communication is also used to withdraw MAC reachability information. For example, if a specific network entity is disconnected from the network, or stops communicating, the corresponding MAC entry would eventually be removed from the CAM table of the OTV edge device. This occurs by default after 30 minutes on the OTV edge device. The removal of the MAC entry triggers an OTV protocol update so that all remote edge devices delete the same MAC entry from their respective tables.

OTV forwarding

Once the control plane adjacencies between the OTV edge devices are established and MAC address reachability information is exchanged, traffic can start flowing across the overlay.

Local traffic will be switched normally. For example, when the device with MAC address A in Figure 8-24, attached to SW-A E1/1, needs to communicate with the device with MAC address B, attached to SW-A E2/1, a typical Layer 2 lookup happens. SW-A sees that MAC B is local and forwards the packet to interface E2/1.

On the other hand, if SW-A in Figure 8-24 on page 462 receives a packet destined to MAC D a traditional Layer 2 lookup is performed, but this time the MAC D information in the MAC table does not point to a local Ethernet interface but to the IP address of the remote OTV edge device that advertised the MAC reachability information, in this case SW-B (with its Join interface IP B).

The SW-A OTV edge device encapsulates the original Layer 2 frame: the source IP of the outer header is the IP address of its Join interface (IP A), whereas the destination IP is the IP address of the Join interface of the remote edge device (IP B).

The OTV encapsulated frame (a regular unicast IP packet) is carried across the transport infrastructure and delivered to the SW-B remote OTV edge device.

SW-B de-capsulates the frame exposing the original Layer 2 packet. It then performs another Layer 2 lookup on the original Ethernet frame and discovers that MAC D is reachable through a physical interface, which means it is a MAC address local to the site.

The frame is sent to the interface where MAC D is seen as connected, in this case is interface E3/2.

It is important to note that the host with MAC D does not have to be physically connected to E3/2 on SW-B, it only has to be reachable over that interface. In other words there could be another switch connected to SW-B E3/2 to which the host with MAC D is ultimately connected.

Note: Given that Ethernet frames are carried across the transport infrastructure after being OTV encapsulated, some considerations around MTU are necessary. In the first implementation, the OTV encapsulation increases the overall MTU size by 42 bytes.

OTV forwarding of broadcast traffic

It is important to highlight that a mechanism is required so that Layer 2 broadcast traffic can be delivered between sites across the OTV overlay. While OTV failure isolation features limit the amount of broadcast traffic across the transport infrastructure, some protocols like Address Resolution Protocol (ARP), would always mandate the delivery of broadcast packets.

When a multicast enabled transport infrastructure is available, the broadcast frames are sent to all remote OTV edge devices by leveraging the same ASM multicast group in the transport already used for the OTV control protocol. Layer 2 broadcast traffic will then be handled exactly the same way as the OTV Hello messages.

For unicast-only transport infrastructure deployments, head-end replication performed on the OTV device in the site originating the broadcast would ensure traffic delivery to all the remote OTV edge devices part of the unicast-only list.

8.4.3 OTV failure isolation

One of the main requirements of every LAN extension solution is to provide Layer 2 connectivity between remote sites without giving up the advantages of resiliency, stability, scalability, and so on, obtained by interconnecting sites through a routed transport infrastructure.

OTV achieves this goal by providing four main functions: Spanning Tree (STP) isolation, Unknown unicast traffic suppression, ARP optimization, and broadcast policy control.

OTV STP isolation

OTV, by default, does not transmit STP Bridge Protocol Data Units (BPDUs) across the overlay. This is a native function that does not require the use of an explicit configuration, such as BPDU filtering, and so on. This allows every site to become an independent STP domain: STP root configuration, parameters, and the STP protocol type can be decided on a per-site basis. This is illustrated in Figure 8-25. This feature fundamentally limits the fate sharing between data center sites: a STP problem in the control plane of a given site would not produce any effect on the remote data centers.

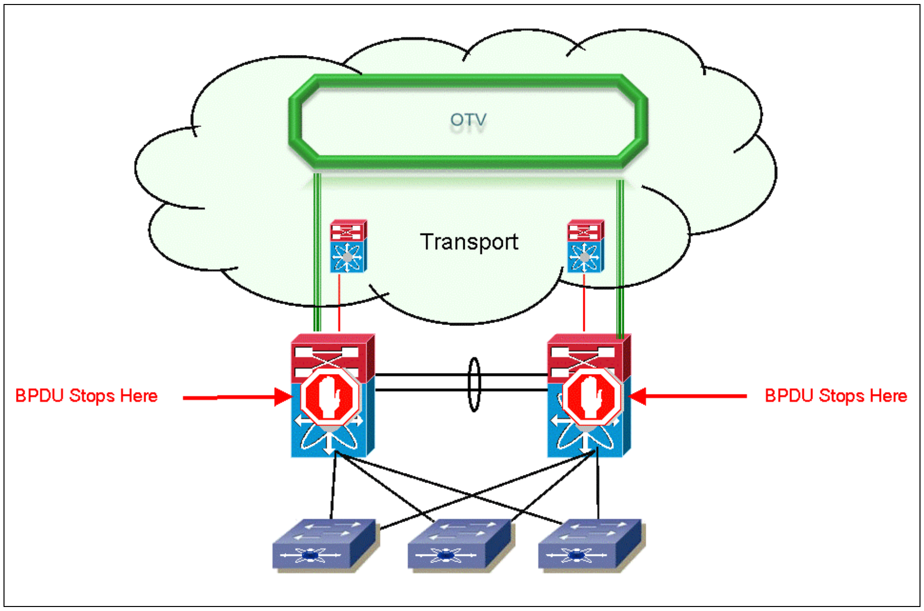


Figure 8-25 OTV STP isolation example

OTV unknown unicast handling

The introduction of an OTV control protocol allows advertising MAC address reachability information between the OTV edge devices and mapping MAC address destinations to IP next hops that are reachable through the network transport. The consequence is that the OTV edge device starts behaving like a router instead of a Layer 2 bridge, since it forwards Layer 2 traffic across the overlay if it has previously received information on how to reach that remote MAC destination. When the OTV edge device receives a frame destined to a MAC address it does not have in its forwarding table, the traffic is flooded out the internal interfaces, since they behave as regular Ethernet interfaces, but not via the overlay.

To support specific applications which require the flooding of Layer 2 traffic to function, a configuration knob is provided to enable selective flooding. Individual MAC addresses can be statically defined so that Layer 2 traffic destined to them can be flooded across the overlay, or broadcast to all remote OTV edge devices, instead of being dropped. The expectation is that this configuration would be required in very specific corner cases, so that the default behavior of dropping unknown unicast would be the usual operation model.

ARP optimization

Another function that reduces the amount of traffic sent across the transport infrastructure is ARP optimization.

This optimization happens when a device in an OTV site sources an ARP request to determine the MAC of the host with a particular IP address, lets use IP A as an example. The ARP request is a Layer 2 broadcast frame and it is sent across the OTV overlay to all remote sites eventually reaching the machine with address IP A, which creates an ARP reply message. The ARP reply is sent back to the originating host. The OTV edge device in the original site is capable of snooping the ARP reply and it caches the contained mapping information (MAC A --> IP A for example) in a local data structure named ARP Neighbor-Discovery (ND) Cache. Any subsequent ARP request from this site for the same IP A address is no longer forwarded to the remote sites but is locally answered by the local OTV edge device on behalf of the remote device IP A.

Because of this ARP caching behavior, you should consider the interactions between ARP and CAM table aging timers, since incorrect settings may lead to black-holing traffic. This is also a consequence of the OTV characteristic of dropping unknown unicast frames as discussed in an earlier section. The ARP aging timer on the OTV edge devices should always be set lower than the CAM table aging timer, and it is by default.

OTV multi-homing

One key function built in the OTV protocol is multi-homing where two (or more) OTV edge devices provide LAN extension services to a given site. The detection of the multi-homing is fully automated and it does not require additional protocols and configuration.

A redundant node deployment, combined with the fact that STP BPDUs are not sent across the OTV overlay, may lead to the creation of an end-to-end loop. The concept of Authoritative edge device (AED) is introduced to avoid this situation. The AED has two main tasks:

1. Forwarding Layer 2 traffic (unicast, multicast and broadcast) between the site and the overlay (and vice versa).
2. Advertising MAC reachability information to the remote edge devices.

To do this, each OTV device maintains dual adjacencies with other OTV edge devices belonging to the same DC site. OTV edge devices use a configurable “otv site-VLAN” for discovering and establishing adjacency with other OTV edge device in the same site. This adjacency is called a Site Adjacency. In addition to the Site Adjacency, OTV devices also maintain a second adjacency, named “Overlay Adjacency,” established via the Join interfaces across the Layer 3 network domain.

In order to enable this functionality, it is mandatory to configure each OTV device with a site-identifier value. All edge devices that are in the same site must be configured with the same site-identifier. This site-identifier is advertised in IS-IS hello packets sent over both the overlay as well as on the site VLAN. The combination of the site-identifier and the IS-IS system-id is used to identify a neighbor edge device in the same site. The dual site adjacency state then used to determine the Authoritative Edge Device (AED) role for each extended data VLAN. VLANs are split between the OTV edge devices belonging to the same site. This is achieved via a very deterministic algorithm (which is not configurable today).

Each OTV edge device can proactively inform its neighbors in a local site about their capability to become an Authoritative Edge Device (AED) and its forwarding readiness. In other words, if something happens on an OTV device that prevents it from performing its LAN extension functionalities, it can now inform its neighbor about this and let itself excluded from the AED election process.

It is important to note that the first immediate consequence of the AED election is that all Layer 2 multicast and broadcast streams need to be handled by the AED device for the specific VLAN, leading to a per-VLAN load-balancing scheme for these traffic flows. The exact same happens to unicast traffic. Only the AED for the VLAN is allowed to forward unicast traffic in and out a given site, for that particular VLAN.

FHRP isolation

The last capability introduced by OTV is the ability to filter First Hop Redundancy Protocol (FHRP), such as HSRP or VRRP, messages across the logical overlay. This is required to allow for the existence of the same default gateway in different locations and optimize the outbound traffic flows (server to client direction).

Given that the same VLAN/IP subnet is available in different sites, the free exchange of FHRP messages across the OTV connection would lead to the election of a single default gateway for all of the combined sites. This would force traffic to follow a suboptimal path to reach the default gateway each time it is required to be routed outside the subnet and the server is located in a different site.

It is important that you enable the filtering of FHRP messages across the overlay because it allows the use of the same FHRP configuration in different sites. The end result is that the same default gateway is available, and characterized by the same virtual IP and virtual MAC addresses, in each data center. This means that the outbound traffic will be able to follow the optimal and shortest path, always leveraging the local default gateway.

This is illustrated in Figure 8-26, which shows how two sites can have an active HSRP router. In this example Host A would use SW-A as its default gateway and Host B would use SW-C. If we were not filtering HSRP between the two sites and had a single (not localized) HSRP active, for example at Site West, then the traffic from Host B would have to travel across the OTV overlay to Site West before it can exit, as shown by the blue dashed line. In the worse case scenario traffic between two devices in Site East (Host B and Host C), may have to travel to Site West for them to communicate with each other.

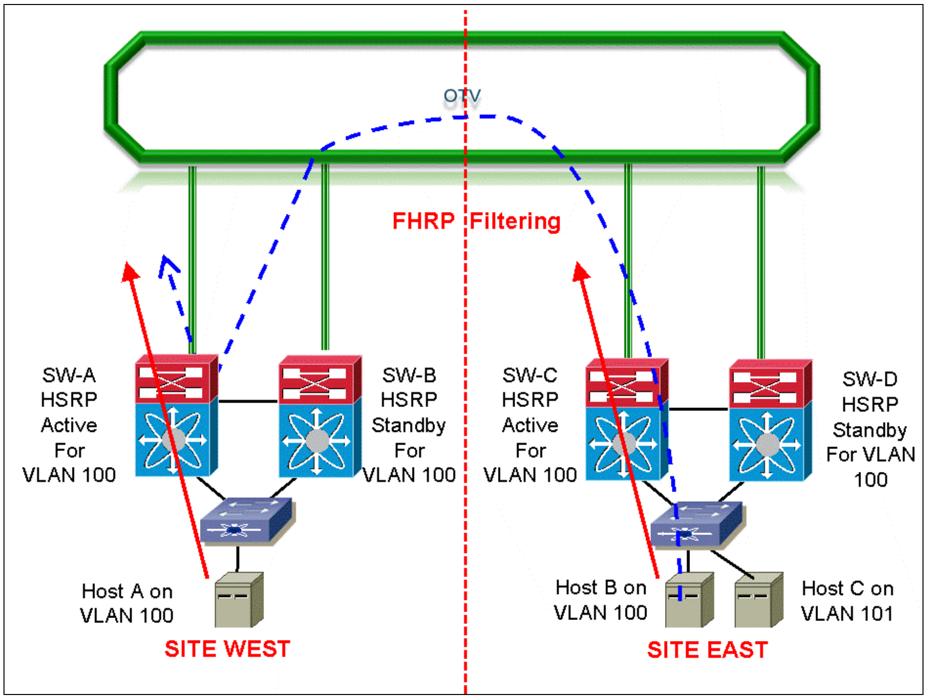

OTV
FHRP Filtering
SW-A HSRP Active For VLAN 100
SW-B HSRP Standby For VLAN 100
SW-C HSRP Active For VLAN 100
SW-D HSRP Standby For VLAN 100
Host A on VLAN 100
Host B on VLAN 100
Host C on VLAN 101
SITE WEST
SITE EAST

Figure 8-26 OTV FHRP Localization

OTV and SVI coexistence

A VLAN can either have Layer 3 connectivity through a VLAN interface (SVI) or the VLAN can be extended over OTV. If you have a design scenario that requires the VLAN to be both extended over OTV to a remote site and have Layer 3 connectivity through a VLAN interface, you must configure OTV in a separate device, or VDC, from the VDC that contains the VLAN interfaces. Figure 8-27 depicts using a separate VDC for OTV. In this design the OTV VDCs (VDC-OTV-1 and 2) have their own separate OTV internal (red color) and OTV join (green color) interfaces. The “regular” VDCs (VDC-SW-1 and 2) run the SVIs and HSRP for the site and have a regular L3 (blue color) interface for forwarding non OTV traffic.

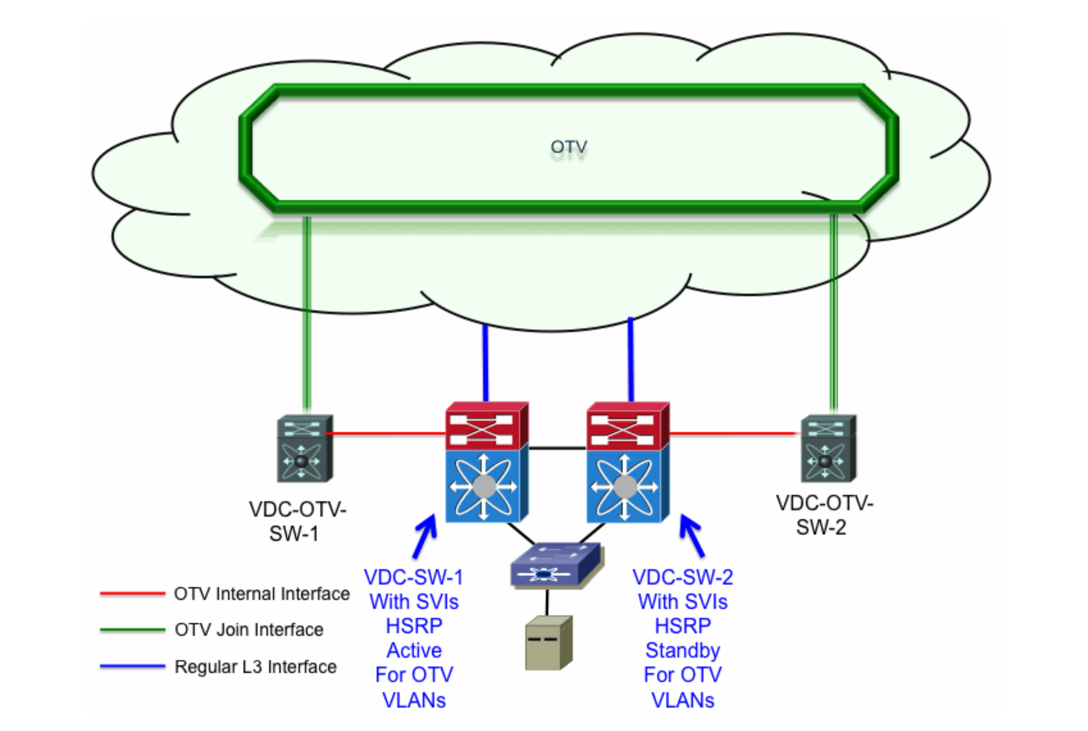

OTV
VDC-OTV-SW-1
VDC-OTV-SW-2
VDC-SW-1 With SVIs HSRP Active For OTV VLANs
VDC-SW-2 With SVIs HSRP Standby For OTV VLANs
OTV Internal Interface
OTV Join Interface
Regular L3 Interface

Figure 8-27 Using separate VDCs for SVIs and OTV

OTV hardware and license requirements

At the time of writing this book, OTV is supported on the Nexus 7000 and the ASR 1000 product line, with more products coming in the future.

On the Nexus 7000 line, OTV requires the Transport Services license. OTV is supported on all existing M1 and M2 line cards. The F-series line cards can be used as internal interfaces.

For the ASR 1000, at least the Advanced IP Services feature set is required to run OTV.

OTV advanced features and multicast support

OTV supports multicast and has many advanced features that were not discussed in this chapter. You can find out more about them searching on Cisco's CCO website.

For more information about OTV on the ASR 1000 platform, visit this website:

http://www.cisco.com/en/US/docs/ios-xml/ios/wan_otv/configuration/xe-3s/wan-otv-confg.html

For more information about OTV on the Nexus 7000 platform, visit this website:

http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DCI/whitepaper/DCI3_OTV_Intro_WP.pdf

OTV standardization process

OTV has been submitted to the IETF for standardization. See the following link for more information:

http://tools.ietf.org/html/draft-hasmit-otv-00

8.4.4 VPLS

Virtual Private LAN Services (VPLS) creates a multipoint-to-multipoint topology over an MPLS network, built upon a set of full mesh pseudowires between all sites to where the VPLS instance needs to be deployed. Usually it is deployed if the data center sites requiring Layer 2 connectivity are more than two. In fact if the sites are just two a simple Ethernet over MPLS tunnel can be created to encapsulate Layer 2 traffic over a MPLS network.

Alternatives for physical transport between data center in metro distance (less than 100 km) are dark fiber, CWDM or DWDM. VPLS over an MPLS overcomes the distance limitations of these techniques.

VPLS is an IETF standard included in the Layer 2 VPN working group. More information on the VPLS standard can be found here:

http://datatracker.ietf.org/wg/l2vpn/charter/

At a basic level, VPLS can be defined as a group of Virtual Switch Instances (VSIs) that are interconnected using EoMPLS circuits in a full mesh topology to form a single, logical bridge, as represented in Figure 8-28. In concept, a VSI is similar to the bridging function found in IEEE 802.1q bridges in that a frame is switched based upon the destination MAC and membership in a Layer 2 VPN (a virtual LAN or VLAN). VPLS forwards Ethernet frames at Layer 2, dynamically learns source MAC address to port associations, and frames are forwarded based upon the destination MAC address. If the destination address is unknown, or is a broadcast or multicast address, the frame is flooded to all ports associated with the virtual bridge. Therefore in operation, VPLS offers the same connectivity experienced if a device were attached to an Ethernet switch by linking virtual switch instances (VSIs) using MPLS pseudowires to form an “emulated” Ethernet switch.

Because VPLS emulates a virtual switch over the MPLS network, loops can be created, so Spanning Tree isolation and end-to-end loop prevention need to be considered.

There are several techniques to achieve loop prevention in VPLS networking, ranging from EEM scripting to MC-LAG. Covering those options is outside of the scope of this book, see the Cisco Validated Design Zone for MPLS/VPLS Data Center Interconnect, available here:

http://www.cisco.com/en/US/netsol/ns749/networking_solutions_sub_program_home.html

The most important use case for the purpose of this publication is enabling Layer 2 connectivity across data centers to enable VM mobility and geo-clustering, for example for Disaster Recovery purposes. Another common use case for VPLS is data center migration or relocation. The benefits of using a standard mechanism such as VPLS is that the PE routers involved in the VPLS instance can be heterogeneous and this may be a requirement if different sites use different data center edge routing platforms.

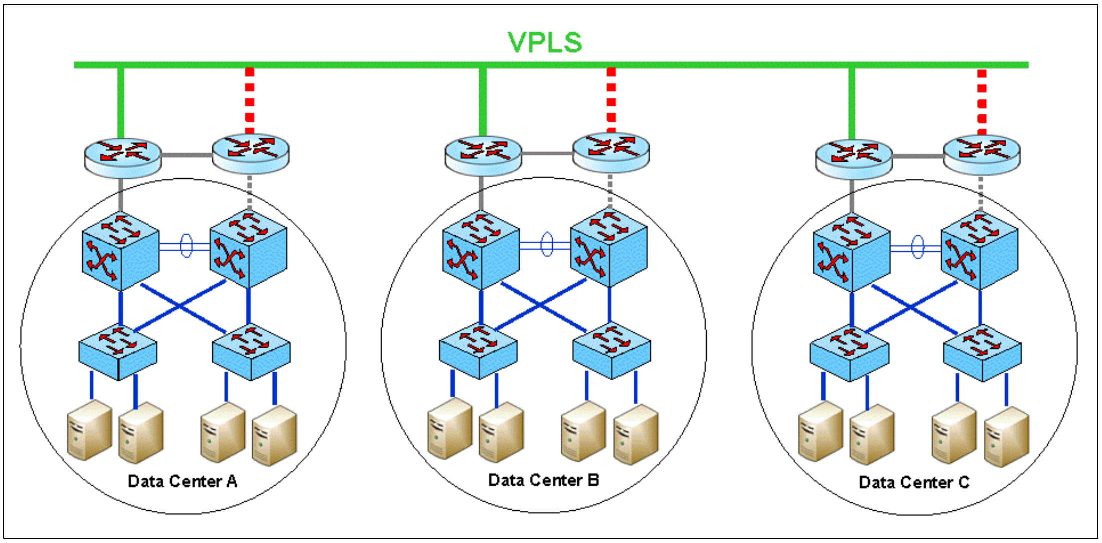


Figure 8-28 Virtual Private LAN Services architecture overview

Virtual Private LAN Services technology is supported on the following Cisco platforms:

- Cisco 7600 Series
- Cisco ASR 9000
- Cisco ASR 1000
- Cisco Catalyst 6500 (with Sup720 and a WAN card or with Sup2T natively)
- Nexus 7000
- Service Provider platforms such as the Cisco GSR and CRS Routers

Several examples, design and implementation guides are available on the Cisco Validated Design Zone for MPLS/VPLS Data Center Interconnect, available here:

`http://www.cisco.com/en/US/netsol/ns749/networking_solutions_sub_program_home.html`

8.4.5 LISP

The current Internet routing and addressing architecture uses a single namespace - the IP address, to simultaneously express two things about a device: its identity and its physical location. This is created by using a subnet mask which allows us to divide the IP address (IPv4 or IPv6) into the network part, which is used to describe the physical location, and the host part, which is used to describe the devices identity.

Locator/ID Separation Protocol (LISP) creates a new paradigm by splitting the device identity, known as an Endpoint Identifier (EID), and its location, known as its Routing Locator (RLOC), into two different numbering spaces. Splitting EID and RLOC functions yields many benefits including:

- Improved routing system scalability by using topologically-aggregated RLOCs
- Provider-independence for devices numbered out of the EID space (IP portability)
- Low-OPEX multi-homing of end-sites with improved traffic engineering
- IPv6 transition functionality
- Path optimization allows traffic to be sent directly to the location (RLOC) where the end device (EID) is located
- IP mobility (EIDs can move without changing; only the RLOC changes)

This processing is shown in Figure 8-29.

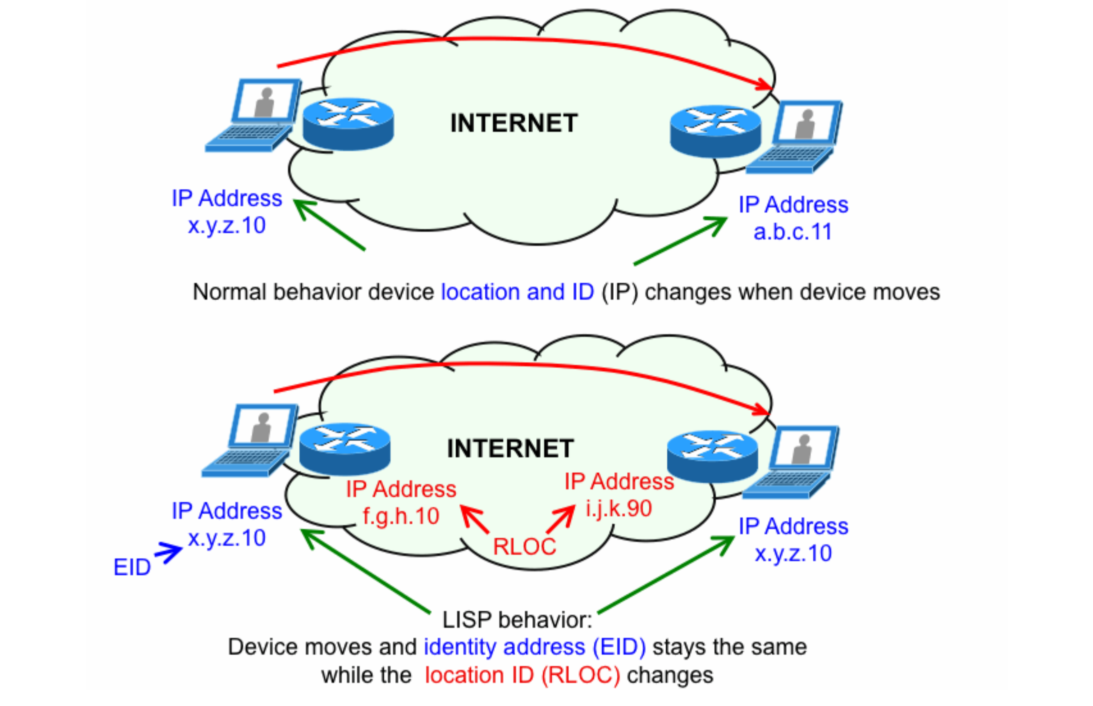

INTERNET
IP Address
x.y.z.10
IP Address
a.b.c.11
Normal behavior device location and ID (IP) changes when device moves
INTERNET
IP Address
f.g.h.10
IP Address
i.j.k.90
RLOC
IP Address
x.y.z.10
EID
IP Address
x.y.z.10
LISP behavior:
Device moves and identity address (EID) stays the same
while the location ID (RLOC) changes

Figure 8-29 The difference between normal and LISP behavior

LISP is a simple, incremental, network-based implementation that is deployed primarily in network edge devices.

It requires no changes to host stacks, DNS, and little to no major changes to existing network infrastructures.

At the time of writing this book, LISP standards are being developed within the IETF LISP Working Group, and are in various states of approval and review.

8.4.6 LISP components and terminology

As illustrated in Figure 8-30, the EID namespace represents customer end points (PC or laptop for example) in exactly the same way that end points are defined today. The only difference is that the IP addresses used within these LISP sites are not advertised within the RLOC namespace. LISP functionality is deployed exclusively on Customer Edge (CE) routers, which function in the LISP roles of Ingress Tunnel Router (ITR) and Egress Tunnel Router (ETR) devices.

To fully implement LISP on an Internet-scale and in order to provide interoperability between LISP and non-LISP sites, additional LISP infrastructure components are required to be deployed as well. These additional LISP infrastructure components, as illustrated in Figure 8-30, support the LISP roles of Map-Server (MS), Map-Resolver (MR), Proxy Ingress Tunnel Router (PITR), and Proxy Egress Tunnel Router (PETR) devices. The specific functions implemented by each of the LISP devices illustrated in Figure 8-30 are discussed in the following sections.

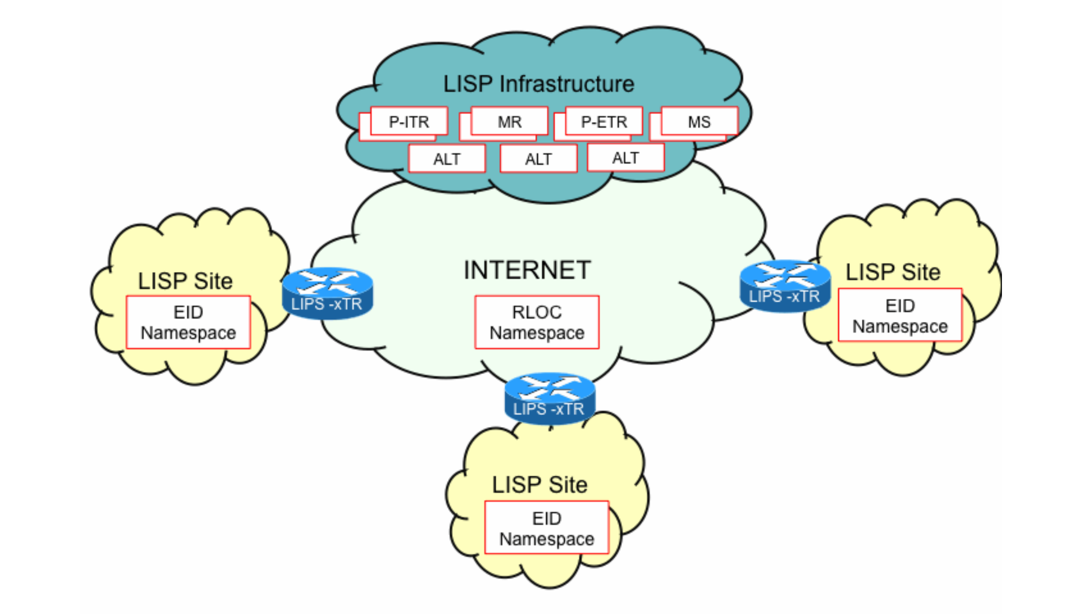


Figure 8-30 LISP components

LISP site edge devices

- **ITR:** Ingress Tunnel Router is deployed as a CE device. It receives packets from site-facing interfaces, and either encapsulates packets to remote LISP sites or natively forwards packets to non-LISP sites.
- **ETR:** Egress Tunnel Router is deployed as a CE device. It receives packets from core-facing interfaces and either de-capsulates LISP packets or natively delivers non-LISP packets to local EIDs at the site.

It is common for LISP site edge devices to implement both ITR and ETR functions, because traffic is usually flowing in and out of the site over the same device. When this is the case, the device is referred to as an xTR. However, the LISP specification does not require that a device perform both ITR and ETR functions.

LISP infrastructure devices

- **MS:** The Map-Server is deployed as a LISP Infrastructure component. It configures the LISP site policy for LISP ETRs that register to it. This includes the EID prefixes for which registering ETRs are authoritative, and an authentication key, which must match the one also configured on the ETR. Map-Servers receive Map-Register control packets from ETRs. When the MS is configured with a service interface to the LISP ALT, it injects aggregates for the EID prefixes for registered ETRs into the ALT. The MS also receives Map-Request control packets from the ALT, which it then encapsulates to the registered ETR that is authoritative for the EID prefix being queried.
- **MR:** The Map-Resolver is deployed as a LISP Infrastructure device. It receives Map-Requests encapsulated to it from ITRs, and when configured with a service interface to the LISP ALT, forwards Map-Requests to the ALT. Map-Resolvers also send Negative Map-Replies to ITRs in response to queries for non-LISP addresses.
- **ALT:** An Alternative Topology device is deployed as part of the LISP Infrastructure to provide scalable EID prefix aggregation. Because the ALT is deployed as dual-stack (IPv4 and IPv6) BGP over generic routing encapsulation (GRE) tunnels, ALT-only devices can be implemented using basic router hardware, or other off-the-shelf devices capable of supporting BGP and GRE.

As it will be demonstrated in the server mobility example, described later in this section, an MS and a MR are functions that can be enabled on edge devices, and do not require dedicated hardware.

LISP interworking devices

- **PITR:** A Proxy ITR is a LISP Infrastructure device that provides connectivity between non-LISP sites and LISP sites. A PITR does this by advertising coarse-aggregate prefixes for LISP EID namespace into the Internet, thus attracting non-LISP traffic destined to LISP sites. The PITR then encapsulates and forwards this traffic to LISP sites. This not only facilitates LISP/non-LISP interworking, but also allows LISP sites to see LISP ingress traffic engineering benefits from non-LISP traffic.
- **PETR:** A Proxy ETR is a LISP Infrastructure device that acts as an intermediate Egress Tunnel Router (ETR) for LISP sites which need to encapsulate LISP packets destined to non-LISP sites.

It is likely that a LISP interworking device will implement both PITR and PETR functions. When this is the case, the device is referred to as a PxTR. This happens because usually there is bi-directional traffic between LISP and non-LISP sites. However, the LISP specification does not require that a device perform both PITR and PETR functions.

8.4.7 LISP functionality

LISP functionality consists of LISP data plane and control plane functions.

LISP data plane

Figure 8-31 highlights the steps required to establish communication between devices deployed in LISP enabled sites.

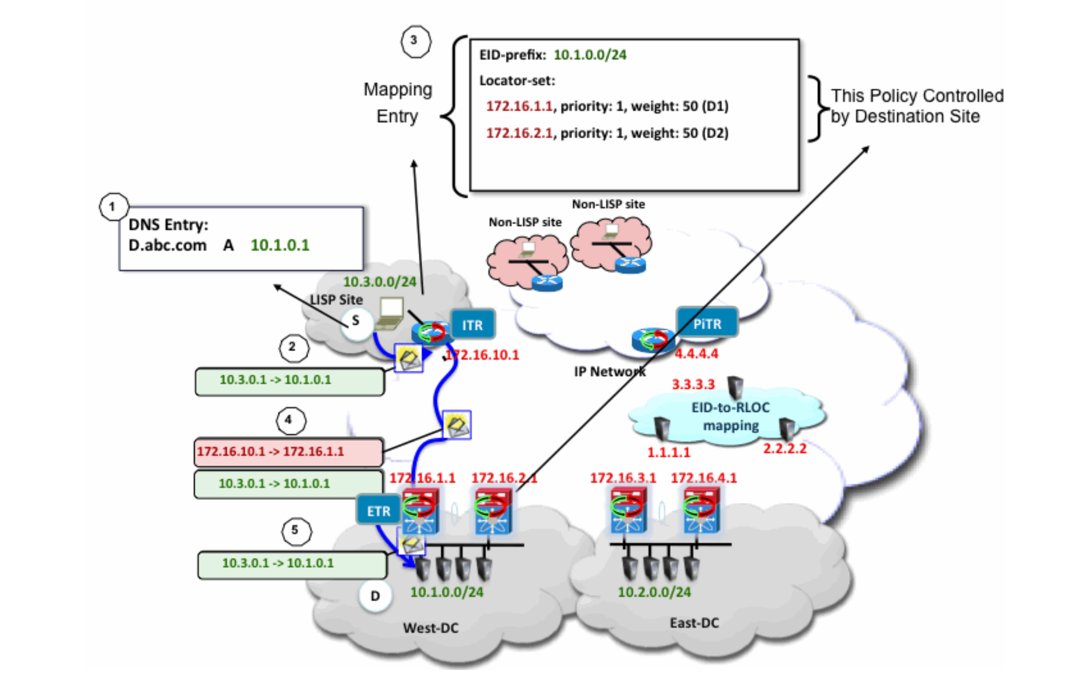


Figure 8-31 Communications between LISP enabled sites

The following operations take place as numbered in Figure 8-31 on page 473:

1. The client sitting in the remote LISP enabled site queries DNS for the IP address of the destination server deployed at the LISP enabled data center site. In this example DNS returns IP address 10.1.0.1 for the server.
2. Traffic originated from the client is steered toward the local LISP enabled device (usually the client's default gateway). The LISP device performs first a lookup for the destination (10.1.0.1) in its routing table. Since the destination is an EID subnet, it is not present in the RLOC space so the lookup fails, triggering the LISP control plane. In the current IOS and NX-OS LISP implementation, the LISP control plane is triggered if the lookup for the destination address produces no results (no match) or if the only available match is a default route.
3. The ITR receives valid mapping information from the mapping database and populates its local map-cache. Notice how the destination EID subnet (10.1.0.0/24) is associated to the RLOCs identifying both ETR devices at the data center LISP enabled site (both routers in the West DC). This happens because the two routers in the West DC have the subnet (10.1.0.0/24) configured as an EID, and both their interfaces (172.16.1.1 and 172.16.2.1) as the RLOC addresses to be registered with the mapping database. Also, each entry has an associated priority and weights values that are controlled by the destination site (configured in the West DC routers) to influence the way inbound traffic is received from the transport infrastructure. The priority is used to determine if both ETR devices can be used to receive LISP encapsulated traffic destined to a local EID subnet (load-balancing scenario). The weight allows tuning the amount of traffic received by each ETR in a load-balancing scenario (hence the weight configuration makes sense only when specifying equal priorities for the local ETRs).
4. On the data-plane, the ITR performs LISP encapsulation of the original IP traffic and sends it into the transport infrastructure, destined to one of the RLOCs of the data center ETRs, in this case the source IP would be 172.16.10.1 and the destination would be 172.16.1.1. Assuming the priority and weight values are configured the same on the ETR devices, the selection of the specific ETR RLOC is done on a per flow basis based on hashing performed on the 5-tuple L3 and L4 information of the original client's IP packet. The entire packet from 10.3.0.1 (source) to 10.1.0.1 (destination) remains “unchanged” inside the encapsulation. The transport network forwards the packet based on the LISP encapsulation information and is unaware of the inner packet’s IP information.
5. The ETR receives the packet, de-capsulates it and sends the original packet into the site toward the destination EID. In this case the server’s IP Address/EID of 10.1.0.1.

While Figure 8-31 on page 473 shows only the North-to-South flow, a similar mechanism would be used for the return traffic originated by the West DC EID and destined to the remote client, where the LISP devices would exchange their roles of ITRs and ETRs.

On the other hand, Figure 8-32 shows the use of PxTR devices to establish communication between devices deployed in non-LISP sites and EIDs devices in LISP enabled sites.

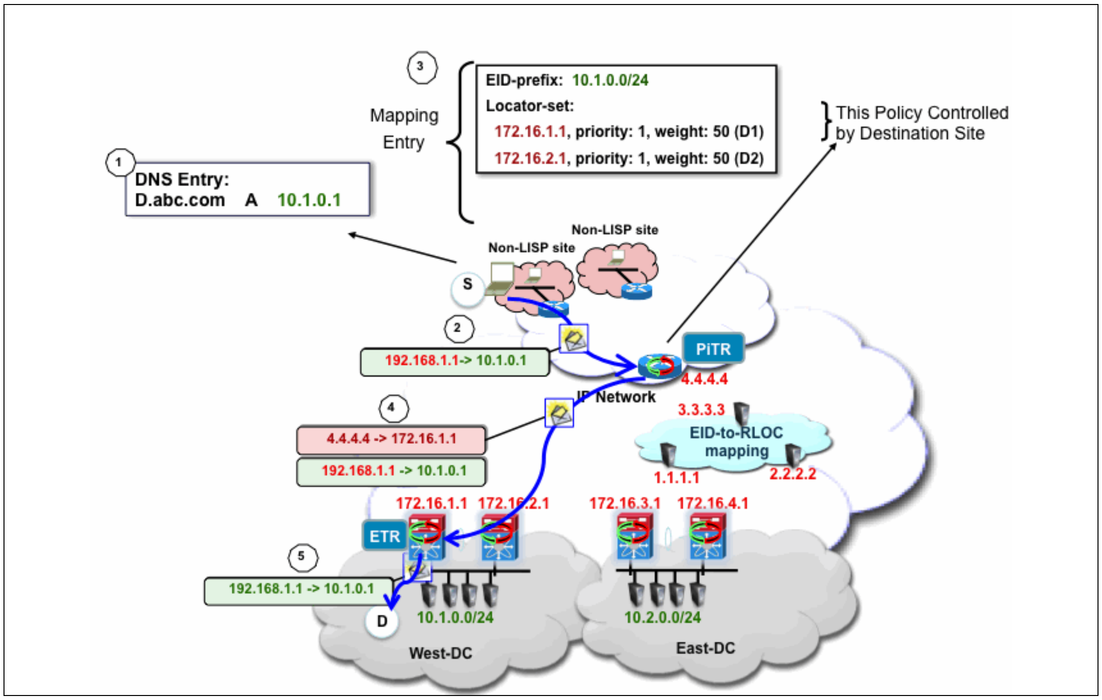


Figure 8-32 Communication between non LISP and LISP enabled sites

Once the traffic reaches the PITR device, the mechanism used to send traffic to the EID (devices) in the data center is identical to what previously discussed. For this to work, it is mandatory that all the traffic originated from the non-LISP enabled sites be attracted to the PITR device. This is ensured by having the PITR injecting aggregated routing information for the data center EIDs into the network connecting to the non-LISP sites.

The LISP control plane

Figure 8-33 depicts the steps required for an ITR to retrieve valid mapping information from the mapping database.

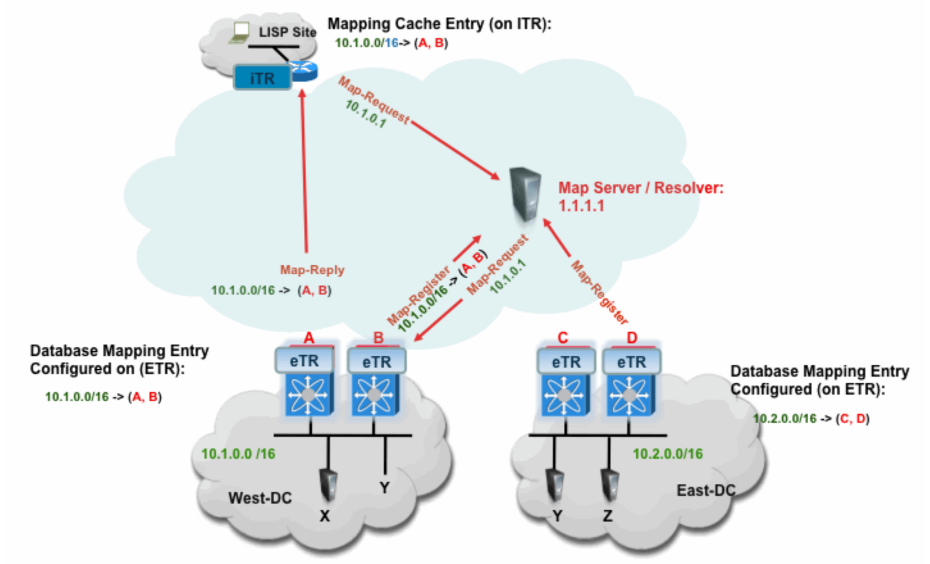


Figure 8-33 Diagram of the LISP Control Plane

1. The ETRs register with the MS the EID subnet(s) that are locally defined and which they are authoritative. In this example the EID subnet is 10.1.0.0/16. Map-registration messages are sent periodically every 60 seconds by each ETR.
2. Assuming that a local map-cache entry is not available, when a client wants to establish communication to a data center EID, a Map-Request is sent by the remote ITR to the Map-Resolver, which then forwards the message to the Map Server.
3. The Map Server forwards the original Map-Request to the ETR that last registered the EID subnet. In this example it is ETR with locator B.
4. The ETR sends to the ITR a Map-Reply containing the requested mapping information.
5. The ITR installs the mapping information in its local map-cache, and starts encapsulating traffic toward the data center EID destination.

8.4.8 LISP use case: Server mobility

While LISP offers several great use cases, one of the most relevant for data centers is to enable server and Virtual Machine (VM) mobility. As discussed earlier the traditional IP addressing model associates both Location and Identity to a single IP address space, making mobility a very cumbersome process since identity and location are tightly bundled together.

Because LISP creates two separate name spaces, that is, separates IP addresses into Route Locators (RLOC) and End-point Identifiers (EID), and provides a dynamic mapping mechanism between these two address families, EIDs can be found at different RLOCs based on the EID-RLOC mappings. RLOCs remain associated with the topology and are reachable by traditional routing. However, EIDs (servers or VMs) can change locations dynamically and are reachable via different RLOCs, depending on where an EID attaches to the network. In a virtualized data center deployment, the VM IP addresses are the EIDs and hence are free to migrate between data center sites preserving their IP addressing information.

The decoupling of Identity from the topology is the core principle on which the LISP VM-Mobility solution is based. It allows the End-point Identifier space to be mobile without impacting the routing that interconnects the Locator IP space.

In the LISP VM-Mobility solution, VM migration events are dynamically detected by the LISP Tunnel Router (xTR) based on data plane events. When a move is detected, as illustrated in Figure 8-34, the mappings between EIDs and RLOCs are updated by the new xTR. By updating the RLOC-to-EID mappings, traffic is redirected to the new locations without causing any churn in the underlying routing.

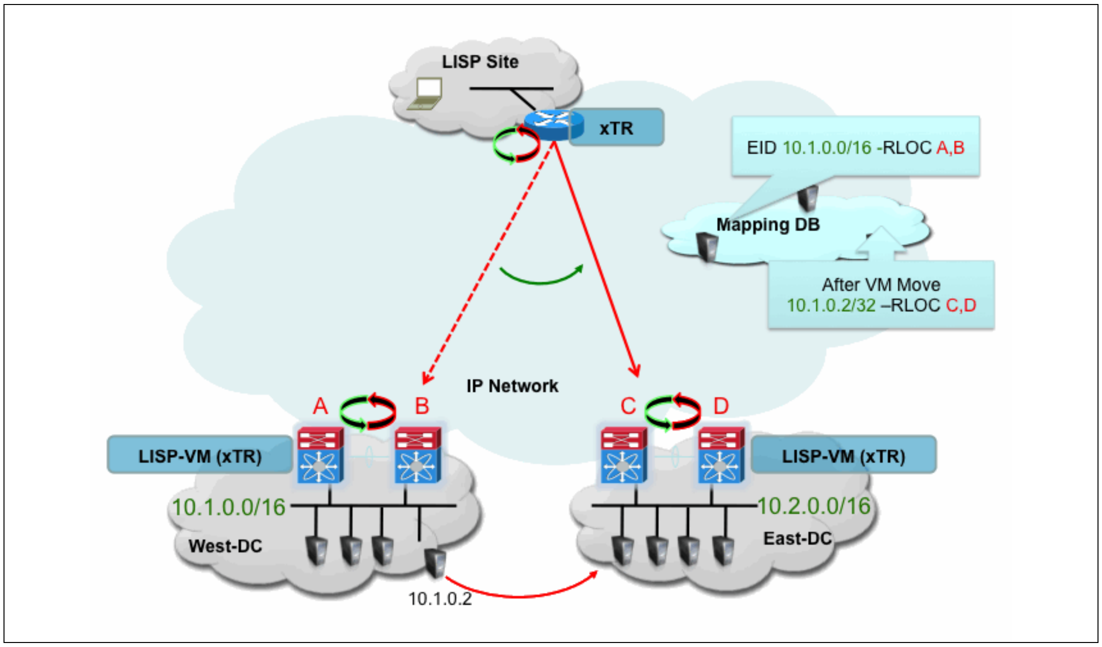


Figure 8-34 An example of LISP VM Mobility

The xTR of the data center where the VM was migrated detects the move by comparing the source IP address of the traffic it receives from the VM against a range of prefixes that are allowed to roam. These prefixes are defined as Dynamic-EIDs in the LISP VM Mobility solution. When deployed at the first hop router, LISP VM-Mobility provides adaptable and comprehensive first hop router functionality to service the IP gateway needs of the roaming devices that relocate.

The LISP xTR configured for LISP Host Mobility detects a new dynamic-EID move event if:

1. It receives an IP data packet from a source (the newly arrived workload) that is not reachable from a routing perspective via the interface on which the packet was received.
2. The source matches the dynamic-EID configuration applied to the interface, two kinds of packets can trigger a dynamic EID discovery event on the xTR:
   - IP packets sourced from the EID IP address. A discovery triggered by this kind of traffic is referred to as “data plane discovery.”
   - ARP packets (which are L2 packets) containing the EID IP address in the payload. A discovery triggered by this kind of traffic is referred to as “control plane discovery”. A typical scenario where this can be leveraged is during the boot-up procedure of a server (physical or virtual), since it would usually send out a Gratuitous ARP (GARP) frame to allow duplicate IP address detection, including in the payload its own IP address.

Once the LISP xTR has detected an EID, it is then responsible to register that information to the mapping database. This registration happens immediately upon detection, and then periodically every 60 seconds for each detected EID.

Figure 8-35 provides a high level example of LISP mobility. The first hop routers for LISP mobility are A and B in the West-DC and C and D in the East-DC.

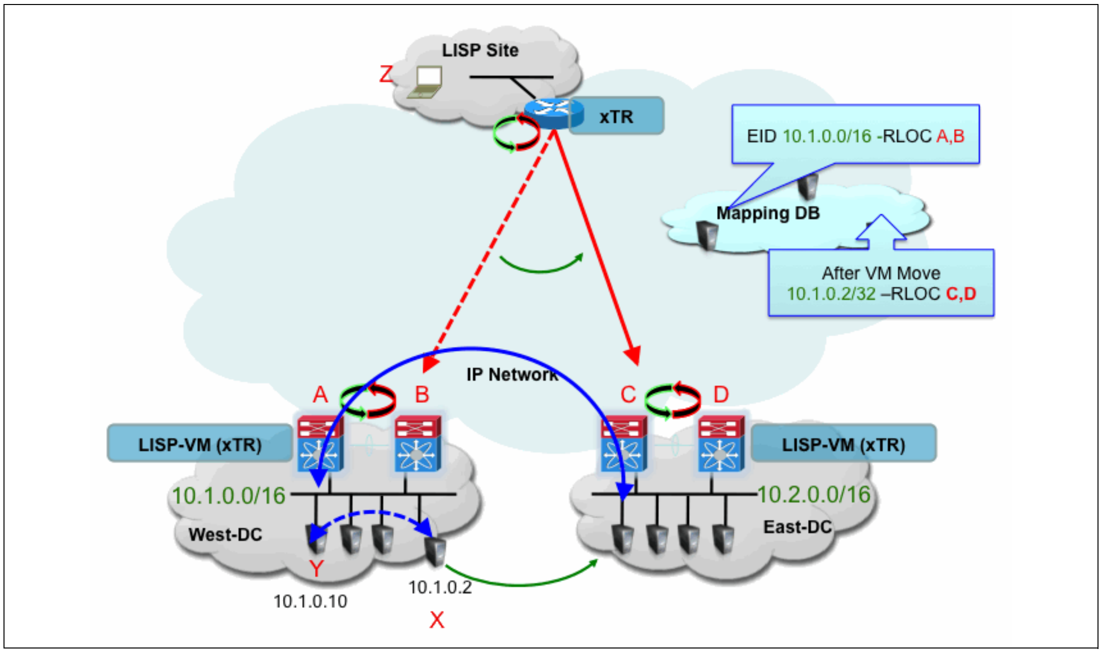


Figure 8-35 An example of LISP Mobility in Across Subnet Mode

Once Host X in Figure 8-35 moves from the West-DC to the East-DC, several things happen:

1. When the VM that is migrated from the West-DC (with subnet 10.1.0.0/24) to the East-DC (with subnet 10.2.0.0/24). The VM retains its IP address and MAC address, and sources an IP packet that reaches one of the two DC xTR devices in the East data center. This triggers an “URPF-like” failure event, since the packet is received on a local SVI belonging to the 10.2.0.0/24 subnet (as an example) and the source of the packet is from the 10.1.0.0/24 subnet. The result of the check failure is that the packet is punted to the CPU, causing the dynamic discovery of the EID.
2. The xTR that discovered the EID installs a local /32 route associated to the EID in its routing table. This is important to allow exchange of traffic (like ARP) to happen with the VM.
3. The discovering xTR sends out a Map-Notify-Group message on the local interface where the discovery of the EID happened.
4. The multicast message reaches the peer xTR in the East-DC side, which also install a valid /32 route for the EID.
5. The discovering xTR sends a Map-Register messages for the /32 EID address to the Map-Server.
6. The Map-Server adds to the database the entry for the specific EID, associated to the RLOCs (C and D) assigned to the xTRs in the East-DC.
7. The Map-Server sends a Map-Notify message to the last xTR in the West DC that registered the 10.1.0.0/16 prefix. This message is to notify the xTR that a specific VM belonging to that subnet has just been discovered in a remote DC site.
8. The xTR in the West-DC receives the Map-Notify message from the Map-Server and adds a /32 Null0 route associated to the 10.1.0.10 EID to its routing table.

9. The West-DC xTR also notifies its peer by leveraging a site local Map-notify message.
10. The West-DC peer xTR also installs in the routing table the /32 Null0 route associated to the 10.1.0.10 EID.

LISP hardware and license requirements

Next we discuss some applicable requirements for LISP.

EID: Once an EID is discovered in the remote data center site, a periodic liveliness check is performed by the xTR that discovered it. The check is performed by pinging the EID every 60 seconds. If three consecutive PINGs are not responded to, the LISP xTR removes the EID from its dynamic EID table and stops registering it with the Map-Server. This is done to handle scenarios where EIDs are silently disconnected from the network.

LISP: We have presented a use case of how LISP enables server migrations across subnets, meaning that there are no Layer 2 extensions between the sites. LISP can also be used for server mobility cases where there is a Layer 2 extension between sites. This use case is called Extended Subnet Mode, and allows ingress traffic optimization to occur.

More information on LISP use cases can be found at this website:

http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DCI/5.0/LISPmobility/DCI_LISP_Host_Mobility.html

At the time of writing this book, LISP is supported in the following Cisco hardware and license combinations:

- ISR 800 Series Advanced IP Services or Advanced Enterprise Services
- ISR 1800 Series Advanced IP Services or Advanced Enterprise Services
- ISR 1900 Series Data license and above
- ISR 2800 Series Advanced IP Services or Advanced Enterprise Services
- ISR 2900 Series Data license and above
- ISR 3800 Series Advanced IP Services or Advanced Enterprise Services
- ISR 3900 Series Data license and above
- 7200 Series Advanced IP Services or Advanced Enterprise Services
- ASR 1000 Series Data license and above
- N7000 Series Requires the M1 32-port 10-GE SFP+ I/O module and the Transport Services License (N7K-TRS1K9)

LISP advanced features and additional information

Many of the advanced features that exist in LISP were not covered in this chapter. You can find additional information on LISP at the following websites:

http://lisp.cisco.com

http://www.cisco.com/go/lisp

http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DCI/5.0/LISPmobility/DCI_LISP_Host_Mobility.html

8.5 Summary

In this chapter, we discussed various technologies for modern data center networks, such as multi-chassis EtherChannels, data center fabrics, and data center overlays. These technologies do not need to exist independently. They can be used together in a complementary way to address challenging requirements driven by IT trends such as virtualization, cloud computing, and server mobility.

Table 8-2 shows how these technologies can be used together to meet certain requirements.

Table 8-2 Network overlays and network fabrics relationship

Requirement	Intra-DC	Inter-DC
Layer 2 connectivity	vPC/ FabricPath/TRILL/VXLAN	OTV/VPLS
IP mobility	LISP	LISP

The types of features and technologies from Table 8-2 that can be leveraged to create a solution will naturally have software, license, and hardware dependencies. Always check the hardware and software requirements while considering the best solution to meet your requirements.

9

NX-OS network operating system

In this chapter, we introduce the NX-OS network operating system and the ways it enables us to build world class data centers. NX-OS was built with self-healing, modularity, and scalability in mind. NX-OS is the operating system for Cisco's Nexus line of products, as well as Cisco's MDS product line. Since NX-OS provides administrators with a consistent operating system for both network and storage devices, it allows them to simplify the design and operation of the entire data center fabric.

9.1 NX-OS building blocks

To meet the needs of the modern data center, a network device, or more particularly, the operating system that powers that device must have the following features:

- Resilient: To provide critical business-class availability
- Modular: To be capable of extension to evolve with business needs and provide extended life-cycles
- Portable: For consistency across platforms
- Secure: To protect and preserve data and operations
- Flexible: To integrate and enable new technologies
- Scalable: To accommodate and grow with the business and its requirements
- Easy to use: To reduce the amount of learning required, simplify deployment, and provide ease manageability

Cisco NX-OS Software is designed to meet all these criteria across all Cisco platforms that run it.

Cisco NX-OS is a highly-evolved modular operating system that builds on more than 15 years of innovation and experience in high-performance switching and routing. It finds its roots in the Cisco SAN-OS operating system used worldwide in business-critical loss-intolerant SAN networks. As a direct result of having been deployed and evolving from nearly a decade in the extremely critical storage area networking space, NX-OS can deliver the performance, reliability, and lifecycle expected in the data center.

Cisco NX-OS is built on a Linux kernel. By using a version 2.6 based Linux kernel as its foundation, NX-OS gains the following benefits:

- Established and widely field-proven core kernel code
- Efficient multi-threaded preemptive multitasking capabilities
- Native multiprocessor and multicore support

The choice of Linux over other operating systems and of the Linux 2.6 kernel over other versions of Linux was strategic. Here are some of the reasons for this choice:

- By inheritance, this implies that NX-OS shares the largest installed base of any UNIX-like operating system in existence. The community development model and widespread use of Linux provide the benefits of rigorous code review, rapid community-based defect resolution, and exceptional real-world field testing.
- The kernel is a near-real-time OS kernel, which is actually preferred over a true-real-time OS for applications such as networking, which may have many parallel critical activities running on a platform.
- This particular kernel version currently provides the best balance of advanced features and maturity and stability. It is currently the most widely deployed version of the Linux kernel.
- As of version 2.6 the kernel introduced an advanced and scalable kernel architecture leveraging multiple run-queues for handling multi-core and multiple-CPU system configurations.

These characteristics provide the solid foundation of resilience and robustness necessary for any network device OS powering the mission-critical environment of today's enterprise-class data centers.

The multi-threaded preemptive multitasking capability provides protected fair access to kernel and CPU resources. This approach helps ensure that critical system processes and services are never starved for processor time. This feature, in turn, helps preserve system and network stability by helping ensure that routing protocols, spanning tree, and internal service processes get access to the CPU cores as needed.

To improve scalability NX-OS incorporates a highly scalable CPU queue and process management architecture which employs multiple processor thread run-queues. This in turn enables more efficient use of modern multi-core CPUs. Combined with its memory mapping techniques and path to 64-bit, NX-OS provides simplified scalability both upwards and downwards to accommodate both control plane growth and multiple platforms.

9.1.1 Modular code base

Several categories of modular system code are built on top of the Linux kernel. These can be generally described as follows:

- Platform-dependent hardware-related modules
- System-infrastructure modules
- Feature modules

The platform-dependent hardware-related modules consist of subsystems such as hardware and chipset drivers specific to a particular hardware platform on which Cisco NX-OS runs. This portion of the OS is the part that must be ported across hardware platforms and allow the other subsystems within the OS to communicate with and tie into the specific hardware features of a platform. The platform-dependent modules typically provide standardized APIs and messaging capabilities to upper-layer subsystems. The modules essentially constitute a hardware abstraction layer to enable consistent development at higher layers in the OS, improving overall OS portability. The defined nature of the platform-dependent modules enables the overall reduction of the code base that specifically requires porting to deliver Cisco NX-OS on other hardware platforms. The result is greater consistency in implementation, reduced complexity in defect resolution, and faster implementation of cross-platform features across the various Cisco NX-OS platforms.

The system infrastructure modules provide essential base system services that enable system process management, fault detection, fault recovery, and inter-service communication. The system management component of the system infrastructure provides service management for other features of the OS. It is also the component responsible for fault detection for the feature services, and it is fully capable of performing fault recovery of a feature service as needed. Working together with other infrastructure modules services, it can provide stateful fault recovery of a feature, enabling recovery of a fault within a specific feature in less than a second, while preserving the runtime state of that feature. This capability enables transparent and nondisruptive fault recovery within the system, increasing overall network stability and service uptime.

The individual feature modules consist of the actual underlying services responsible for delivering a particular feature or protocol capability. Open Shortest Path First (OSPF), Enhanced Interior Gateway Routing Protocol (EIGRP), Intermediate System-to-Intermediate System (IS-IS) Protocol, Border Gateway Protocol (BGP), Spanning Tree Protocol, Fibre Channel over Ethernet (FCoE), the routing information base (RIB), Overlay Transport Virtualization (OTV), and NetFlow export are all examples of system-level features embodied in modular components.

Each feature is implemented as an independent, memory-protected process spawned as needed based on the overall system configuration. This approach differs from that of traditional network operating systems in that only the specific features that are configured are automatically loaded and started.

This highly granular approach to modularity enables benefits such as these:

- Compartmentalization of fault domains within the OS and its services, resulting in significantly improved overall system resiliency and stability
- Simplified portability for cross-platform consistency through reusable components, or building blocks. and little use of platform-specific code
- More efficient defect prioritization and repair through the isolation of specific functions to particular modules
- Improved long-term platform extensibility through the capability to easily integrate new feature modules into the OS infrastructure through established and consistent OS interfaces
- More efficient resource utilization because only features specifically enabled through configuration are loaded into memory, present command-line interface (CLI) elements, and consume CPU cycles
- Improved security because features that are not configured or enabled do not run, thus reducing the exposure of the OS to attacks

9.1.2 Intelligent fault detection and recovery

In addition to the resiliency gained from architectural improvements, Cisco NX-OS provides internal hierarchical and multi-layered system fault detection and recovery mechanisms. No software system is completely immune to problems, so an effective strategy for detecting and recovering from faults quickly and with as little effect as possible is essential. Cisco NX-OS is designed from the start to provide this capability.

Individual service and feature processes are monitored and managed by the Cisco NX-OS System Manager, an intelligent monitoring service with integrated high-availability logic. The system manager can detect and correct a failure or lockup of any feature service within the system. The system manager is in turn monitored and managed for health by the Cisco NX-OS kernel. A specialized portion of the kernel is designed to detect failures and lockups of the Cisco NX-OS System Manager. The kernel itself is monitored through hardware. A hardware process constantly monitors the kernel health and activity. Any fault, failure, or lockup at the kernel level is detected by hardware and will trigger a supervisor switchover.

The System Manager is responsible for these tasks:

- Maintaining the overall high availability (HA) states for services
- Enforcing HA policies
- Starting, monitoring, and restarting process, services, and protocols

Along with the System Manager NX-OS services also rely on the Persistent Storage Service (PSS). The persistent storage service (PSS) is used by services to store and manage their operational run-time information. The PSS component works with system services to recover states in the event of a service restart. PSS functions as a database of state and run-time information, which allows services to make a checkpoint of their state information whenever needed. A restarting service can recover the last known operating state that preceded a failure, which allows for a stateful restart.

Each service that uses PSS can define its stored information as private (it can be read only by that service) or shared (the information can be read by other services). If the information is shared, the service can specify that it is local (the information can be read only by services on the same supervisor) or global (it can be read by services on either supervisor or on modules).

For example, if the PSS information of a service is defined as shared and global, services on other modules can synchronize with the PSS information of the service that runs on the active supervisor.

In short, the HA process can recover services in one of three ways, all of which are attempted in the following order:

1. Stateful restart: The service can retrieve its old state from the internal database (PSS). This method does not require the restarting service to exchange messages with any other service to recover its state. This is non traffic impacting way of recovering the service. Some processes that support stateful restarts (as of the writing of this book) are:
   - Layer 3: NETSTACK, OSPFv2, GLBP, VRRP, HSRP, ARP, DHCP_SNOOP, NTP, ICMPv6 and CTS
   - Layer 2: vPC, STP, LACP, CDP, IGMP and Dot1.X
2. Stateless restart: The service loses its state upon being restarted. It needs to exchange messages with other services to recover. This can be traffic impacting
3. Reset: This can happen as a reload for single supervisor (or non modular) system such as a Nexus 5500. A reload can cause an impact on traffic forwarding. For a dual supervisor system such as a Nexus 7000, an MDS 9500 or an MDS 9700 a reset can take place as a supervisor switchover which most commonly does not affect traffic.

Dual supervisor systems

The Nexus 7000, the MDS 9500, and the MDS 9700 systems with dual supervisors provide hitless redundancy for the supervisor module. The active supervisor constantly synchronizes its state to the standby supervisor so that the standby is ready to take over if required. The CPU watchdog process monitors supervisor responsiveness. A switchover occurs after the other restart mechanisms have failed, when the kernel fails or panics, or when the supervisor experiences a hardware failure. A supervisor switchover is hitless and there is no traffic forwarding impact.

9.1.3 Continuous operation and high availability

Cisco NX-OS is designed from the start to provide consistent, predictable, and reliable high availability. The design goal for the data center is continuous operation: no service disruption. Cisco NX-OS provides a high-availability architecture that moves toward this goal with fully nondisruptive stateful supervisor switchover (SSO) for control-plane redundancy in modular platforms, and nondisruptive In-Service Software Upgrade (ISSU) for all Cisco Nexus platforms.

When running on platforms that offer redundant control-plane hardware, Cisco NX-OS is designed to provide efficient event-based state synchronization between active and standby control-plane entities. This approach allows the system to rapidly perform a fully stateful control-plane switchover with little system disruption and no service disruption.

For platforms without redundant control planes, Cisco NX-OS can implement ISSU by retaining the software state throughout the upgrade process and retaining packet-forwarding intelligence through its hardware subsystems, preventing service disruption.

Cisco NX-OS is also designed to take full advantage of the distributed environment on platforms with distributed hardware forwarding, so that data-plane forwarding is not affected during redundant control-plane switchover. This architecture effectively delivers true nondisruptive control-plane failover.

We cover some of these features in more depth later in this chapter.

9.1.4 Enhanced usability and familiar operation

The Cisco IOS Software is already the recognized leader in internetworking device operating systems.

For decades, Cisco IOS Software has been the foundation for routing and switching configuration in many environments. The Cisco IOS CLI has essentially become the standard for configuration in the networking industry.

To reduce the amount of time needed to learn Cisco NX-OS and to accelerate adoption, Cisco NX-OS maintains the familiarity of the Cisco IOS CLI. Users comfortable with the Cisco IOS CLI will find themselves equally comfortable with Cisco NX-OS. In addition, Cisco NX-OS has integrated numerous user interface enhancements on top of the familiar Cisco IOS CLI to make configuration and maintenance more efficient. These are just some of the simple but effective UI enhancements found in Cisco NX-OS:

- Non-hierarchical CLI: Almost any command can be run from any mode. You can run show commands from the interface and global configuration modes. Global commands can be run from the interface configuration mode. The system is intelligent enough to determine the nature of a command and process it regardless of whether the current mode the configuration or execution mode.
- Configuration mode contextual CLI history: Separate CLI command histories are maintained for each configuration mode. Reentry of commands in a given configuration mode is simplified; you can use the up and down arrow to cycle through the command history stored for that particular configuration mode.
- Advanced multilevel and multi-command output piping: The capability to stream CLI output through advanced filters and parsing commands enables complex formatting and manipulation of information for easier parsing and processing.
- More verbose and descriptive status output: The show command output tends to be more informative and less obscure or opaque in Cisco NX-OS, allowing more effective troubleshooting and status monitoring.

Cisco IOS Software users will quickly find themselves familiar with the Cisco NX-OS CLI and its enhancements. Typically, most networking professionals will also quickly find themselves seeking those functional enhancements in other operating systems.

9.2 Virtual device contexts overview

Cisco NX-OS also provides the capability to virtualize the platform on which it is running. Using Cisco NX-OS virtual device contexts (VDCs), a single physical device can be virtualized into many logical devices, each operating independently, effectively approximating separate physical devices as shown Figure 9-1.

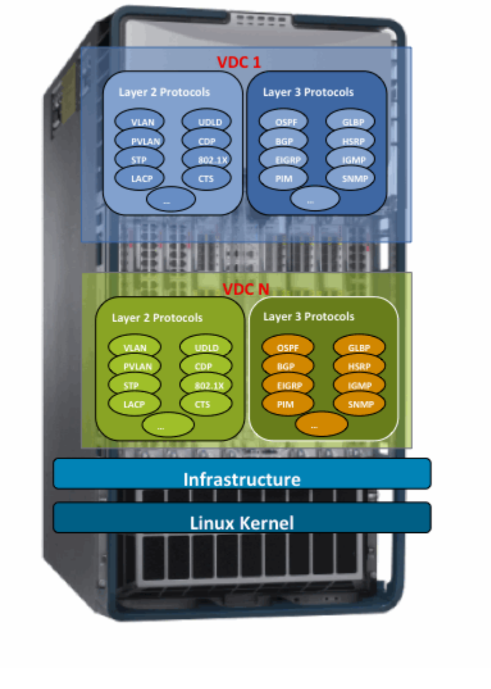


Figure 9-1 VDC diagram

Cisco NX-OS VDCs differ from other implementations of virtualization found in most networking devices by applying in-depth virtualization across multiple planes of operation:

- Virtualization at the data plane: Physical interfaces are associated with a specific VDC instance. Data-plane traffic transiting the physical device can be switched only from one interface in a VDC to another within the same VDC. This virtualization is internalized in the switching system and is not subject to external influence, providing very strong data-plane separation between VDCs. The only means of integrating traffic between VDCs is through a physical cross-connection of ports between two or more VDCs.
- Virtualization at the control plane: All control-plane functions are virtualized within the operating system at the process level. This approach effectively creates separate failure domains between VDCs, reducing the fate-sharing between them. Network or system instability within the domain of a single VDC does not affect other VDCs or the network domain in which they are operating.
- Virtualization at the management plane: Virtualization of the management plane is where VDCs truly stand out compared to other network device virtualization solutions. VDCs virtualize the configuration of each logical device, and they also virtualize all supporting management environment services and capabilities. Each VDC maintains separate configurations and operational relationships with typical common support and security services such as:
  - Separate independent syslog servers configurable per VDC
  - Separate independent authorization, authentication, and accounting (AAA) servers configurable per VDC
  - Independent addressable management IP addresses per VDC
  - Separate independent NetFlow export targets per VDC
  - Independent per-VDC local authentication user lists with per-VDC role-based access control (RBAC)

The end result is extremely effective separation of data traffic and operational management domains suitable for cost-effective infrastructure consolidation in security-sensitive environments.

9.3 Security

Cisco NX-OS provides the tools required to enable the advanced security features needed to protect the network infrastructure as well as the actual platform on which it is running.

Security is designed using a two-pronged approach: security at the data plane and security at the control plane. This approach effectively secures transient traffic passing through a Cisco NX-OS device as well as the traffic destined for the device itself.

Cisco NX-OS currently enables the deployment of both common and more advanced data-plane and infrastructure-level security features, including:

- IEEE 802.1ae and Cisco TrustSec platform
- IP source guard (IPSG)
- Dynamic Host Control Protocol (DHCP) snooping
- Unicast reverse-path forwarding (uRPF)
- Hardware-based IP packet validity checking
- Port, router, and VLAN access control lists (ACLs)
- IEEE 802.1x
- Bridge Protocol Data Unit (BPDU) guard

These controls enable effective protection against most man-in-the-middle, common resource, and spoofing attacks.

At the control plane, Cisco NX-OS provides a robust security toolset to prevent attacks directed at the Cisco NX-OS device itself or active sessions on the device. Tools include capabilities such as:

- Control-plane policing (CoPP)
- RBAC with RADIUS and TACACS+ integration
- Strong password checking
- Secure Shell (SSH) Protocol and Secure FTP (SFTP)

More importantly, the overall Cisco NX-OS system architecture provides the interfaces and structures necessary to easily implement future security features consistently and cleanly.

> **Note:** NX-OS is Common Criteria Certified for Cisco Nexus 5000 Series Switch (5010, 5020, 5548P, 5596UP) with Nexus 2000 Series Fabric Extenders (2148T, 2224TP, 2248TP, 2232PP), Cisco Nexus 7000 Series Switch (7010, 7018), Cisco MDS 9000 and Secure Access Control Server (ACS).
>
> For more information on Cisco Common Criteria certifications, visit this website:
>
> `http://www.cisco.com/web/strategy/government/security_certification/net_business_benefit_seccert_common_criteria.html`

9.4 IPv6 support

From the start, Cisco NX-OS was designed to comprehensively support IPv6. Its broad portfolio of IPv6 features and extensible architecture allows it to integrate flexibly into world-class enterprise and service provider IPv6 environments.

Support for common IPv6 routing protocols, features, and functions and the capability to easily add more make Cisco NX-OS an excellent OS for IPv6 critical deployments.

9.5 Advanced system management

The operation, administration, and monitoring (OAM) of network elements are critical to long-term data center infrastructure sustainability. Cisco NX-OS includes a number of traditional and advanced features to ease OAM in the data center.

- Simple Network Management Protocol (SNMP): Traditional SNMP Versions 1, 2c, and 3 are supported for read operations, allowing integration into existing systems for effective monitoring. Cisco NX-OS on the Cisco Nexus family is also certified in the latest versions of the EMC Ionix management platform.
- NETCONF and XML: Cisco NX-OS provides integration with IETF NETCONF-compliant systems through XML transactions over a secure SSH interface.
- Cisco Generic Online Diagnostics (GOLD): Cisco NX-OS supports Cisco GOLD online diagnostics for active component and subsystem testing.
- Cisco Embedded Event Manager (EEM): Cisco NX-OS provides a scripted interface that enables the configuration of automated event-triggered actions to be run by the system autonomously. Embedded Event Manager supports more than 20 event detectors that are highly integrated with different software components to trigger actions in response to network events. These policies are programmed using either simple command-line interface (CLi) or using a scripting language called Tool Command Language (Tcl).
- Single-image download: Because of its simplified licensing and image structure, every image of Cisco NX-OS contains all the features of Cisco NX-OS available at that release. Individual features are loaded, enabled, and made available to the system platform based on Cisco NX-OS electronic licensing. Therefore, only a single image is available for download for a given version on a given platform. No decoder or version and feature chart is required to determine which image is appropriate for a given environment.
- Scheduler: Cisco NX-OS includes a generic system scheduler that can be configured to run CLI-based system commands at a given time, on a one-time or recurring basis.
- CallHome: The CallHome feature in Cisco NX-OS allows an administrator to configure one or more contact email addresses that are notified when the CallHome function is triggered. The notification process is triggered during certain events that are configurable, such as Cisco GOLD test results and scripts that run based on Cisco EEM events. CallHome enables rapid escalation of events and proactive prevention of a pending failure.
- Configuration checkpoint and rollback: Cisco NX-OS incorporates an advanced configuration and rollback facility to preserve and protect the configuration state. Configuration snapshots, or checkpoints, can be created manually at the CLI or initiated automatically by the system at major configuration events (such as the disabling of a feature). Checkpoints can be stored locally on the device in the local checkpoint database or in a file in integrated or removable storage. Using the rollback capability, the current running configuration can be restored to a particular state stored in a checkpoint.

9.6 In-service software upgrade

Non-disruptive In-Service Software Upgrade (ISSU) is the ability to upgrade (or downgrade) the OS without the need of taking the system out from the network's production environment.

NX-OS provides the ability to perform in-service software updates (ISSUs). The overall modular software architecture of NX-OS supports plug-in-based services and features. This framework makes it possible to perform complete image upgrades, non-disruptively without impacting the data forwarding plane. This capability allows for nonstop forwarding during a software upgrade, including upgrades between full image versions (for example, from 5.0 to 5.1).

An in-service software upgrade is initiated manually either through the command-line interface (CLI) by an administrator, or via the management interface of the Datacenter Network Manager (DCNM) software platform. When initiated, an in-service software upgrade updates (as needed) the following components on the system:

- Supervisor BIOS, Kickstart Image, System Image
- I/O module BIOS and Image
- Connectivity Management Processor (CMP) BIOS and Image

Once initiated, the ISSU Installer service begins the ISSU cycle. The upgrade process is composed of several phased stages designed to minimize overall system impact with no impact to data traffic forwarding.

It is important to note that an ISSU-based upgrade is a system-wide upgrade and not a virtual device context (VDC)-based upgrade. Therefore, the initiation of an ISSU-based upgrade applies the same image and versions across the entire system, to all configured VDCs. Virtual device contexts are primarily a control-plane and user-interface virtualization and cannot currently run independent image versions per virtualized resource.

Additionally, in order to minimize overall network impact, an in-service software upgrade should only be initiated when the network topology is stable, as control-plane components necessary to participate in topology state changes may be in the process of upgrading during a transition.

An ISSU of ISSD (In-Service Software Downgrade) should be performed between supported images. The supported image information can be found in the NX-OS Software release notes. ISSU is supported on both the Nexus 7000, Nexus 5500, MDS 9500, and MDS 9700 platforms.

9.7 Control plane and data plane separation

A completely separate data plane helps ensure that traffic forwarding continues even if there is a disruption in the control plane, as long as the software is intelligent enough to program the hardware for nonstop operation. To understand how this works, we must first understand that NX-OS divides the traffic that it manages into three functional components or planes:

- Data plane: Handles all the data traffic. The basic functionality of a NX-OS device is to forward packets from one interface to another. The packets that are not meant for the switch itself are called the transit packets. These packets are handled by the data plane.

- Control plane: Handles all routing protocol control traffic. These protocols, such as the Border Gateway Protocol (BGP), Open Shortest Path First (OSPF) Protocol, and Protocol Independent Multicast (PIM) Protocol, send control packets between devices. These packets are destined to router addresses and are called control plane packets.
- Management plane: Runs the components meant for NX-OS device management purposes such as the command-line interface (CLI) and Simple Network Management Protocol (SNMP).

The relationship between the control and data planes is that the control plane programs the data plane's forwarding logic. It tells the data plane how to forward the packet. As an example let us consider how a Nexus 7000 does its forwarding. The supervisor, or control plane, copies the entire forwarding table it has calculated into the forwarding engines on each of its line cards. This includes security information and several other attributes. When a packet enters a Nexus 7000 line card interface, a lookup happens on the local card's forwarding engine and the packet is forwarded based upon the information contained in it. If the packet fails a security check (an ACL drop rule, for example) it will be dropped. If the packet has to be rewritten, it is done locally.

Since the data plane does not have to consult with the control plane to forward packets, the control plane is free to handle everything else. Also, if the control plane is very busy (running a high CPU) for any reason, data forwarding is not affected.

9.7.1 Control Plane Policing (CoPP)

Control Plane Policing (CoPP) protects the control plane and management planes and separates them from the data plane, thereby ensuring network stability, reachability, and packet delivery.

This feature allows a policy-map to be applied to the control plane. This policy-map looks like a normal QoS policy and is applied to all traffic entering the switch from a non-management port. A common attack vector for network devices is the denial-of-service (DoS) attack, where excessive traffic is directed at the device interfaces.

The Cisco NX-OS device provides CoPP to prevent DoS attacks from impacting performance. Such attacks, which can be perpetrated either inadvertently or maliciously, typically involve high rates of traffic destined to the supervisor module or CPU itself.

The supervisor module has both the management plane and control plane and is critical to the operation of the network. Any disruption or attacks to the supervisor module will result in serious network outages. For example, excessive traffic to the supervisor module could overload and slow down the performance of the entire Cisco NX-OS device. Attacks on the supervisor module can be of various types such as DoS that generates IP traffic streams to the control plane at a very high rate. These attacks force the control plane to spend a large amount of time in handling these packets and prevents the control plane from processing genuine traffic.

Examples of DoS attacks are as follows:

- Internet Control Message Protocol (ICMP) echo requests
- IP fragments
- TCP SYN flooding

These attacks can impact the device performance and have the following negative effects:

- Reduced service quality (such as poor voice, video, or critical applications traffic)
- High route processor or switch processor CPU utilization

- Route flaps due to loss of routing protocol updates or keep-alives
- Unstable Layer 2 topology
- Slow or unresponsive interactive sessions with the CLI
- Processor resource exhaustion, such as the memory and buffers
- Indiscriminate drops of incoming packets

To protect the control plane, the Cisco NX-OS device segregates different packets destined to the control plane into different classes. Once these classes are identified, the Cisco NX-OS device polices the packets, which ensures that the supervisor module is not overwhelmed.

Here are some control plane packet types:

- Receive packets: Packets that have the destination address of a router. The destination address can be a Layer 2 address (such as a router MAC address) or a Layer 3 address (such as the IP address of a router interface). These packets include router updates and keep-alive messages. Multicast packets can also be in this category where packets are sent to multicast addresses that are used by a router.
- Exception packets: Packets that need special handling by the supervisor module. For example, if a destination address is not present in the Forwarding Information Base (FIB) and results in a miss, then the supervisor module sends an ICMP unreachable packet back to the sender. Another example is a packet with IP options set.
- Redirected packets: Packets that are redirected to the supervisor module. Features like Dynamic Host Configuration Protocol (DHCP) snooping or dynamic Address Resolution Protocol (ARP) inspection redirect some packets to the supervisor module.
- Glean packets: If a Layer 2 MAC address for a destination IP address is not present in the FIB, the supervisor module receives the packet and sends an ARP request to the host.

All of these different packets could be maliciously used to attack the control plane and overwhelm the NX-OS device. CoPP classifies these packets to different classes and provides a mechanism to individually control the rate at which the supervisor module receives these packets. For effective protection, the NX-OS device classifies the packets that reach the supervisor modules to allow you to apply different rate controlling policies based on the type of the packet. For example, you might want to be less strict with a protocol packet such as Hello messages but more strict with a packet that is sent to the supervisor module because the IP option is set. The following parameters that can be used for classifying a packet:

- Source IP address
- Destination IP address
- Source MAC address
- Destination MAC address
- VLAN
- Source port
- Destination port
- Exception cause

Once the packets are classified, the NX-OS device has different mechanisms to control the rate at which packets arrive at the supervisor module. Two mechanisms control the rate of traffic to the supervisor module. One is called policing and the other is called rate limiting.

Using hardware policies, you can define separate actions for traffics that conforms to, exceeds, or violates certain conditions. The actions can transmit the packet, mark down the packet, or drop the packet.

When you bring up your NX-OS device for the first time, the NX-OS Software installs a default CoPP policy to protect the supervisor module from DoS attacks. You can change the CoPP policies as needed from the CLI. You can also remove the default CoPP policy using the CLI.

Notes:

- The Nexus 7000 line has four default CoPP policies to choose from: Strict, Moderate, Lenient or None.
- NX-OS physical devices support only hardware-based CoPP which does not support the management interface (mgmt0). The out-of-band mgmt0 interface connects directly to the CPU and not pass through the in-band traffic hardware where CoPP is implemented. To limit traffic on the mgmt0 interface you can ACLs.

Since CoPP runs in special hardware for the Nexus 7000, Nexus 5500, and the MDS 9500. This keeps the CoPP process from negatively affecting the CPU of the supervisor module.

9.8 Graceful Restart and Non Stop Routing

The terms Graceful Restart (GR) and Non Stop Routing (NSR) are used to represent the controlled restart operations and signaling used between protocol neighbors for the recovery of a control plane process. Graceful Restart is described in several IETF RFCs such as RFC 3623 titled “Graceful OSPF Restart”. Graceful Restart is sometimes also knows as Non-Stop Forwarding (NSF). In NX-OS Non Stop Routing is also referred to as Stateful Restart.

Graceful Restart includes the ability for a network device to notify its neighbors of its “Non-Stop Forwarding” capability so that data is continually forwarded during a restart operation. The GR extensions used in NXOS are based on the IETF RFCs except for EIGRP, which is Cisco proprietary and it can interoperate with Cisco NSF. This implies that for OSPFv2, OSPFv3, and BGP the GR extension are compatible with RFC based extensions and would work with other compliant devices.

Traditionally, when a networking device restarts, all routing peers associated with that device detect the device has gone down and routes from that peer are removed. The session is re-established when the device completes the restart. This transition results in removal and re-insertion of routes. This “churn” of routes can affect multiple routing domains. This was required because of the inability of the restarting device to forward traffic during the reload period. With the capabilities of separating the control and forwarding planes of devices as described in the previous section a device with a restarting supervisor can still forward traffic.

In that case, route removal and insertion caused by routing protocol restarts is no longer necessary, creating unnecessary routing instabilities. Graceful Restart and Non Stop Routing suppress routing changes on peers to SSO-enabled devices during processor switchover events, reducing network instability and downtime. SSO devices are described in the section titled “Continuous operation and high availability” on page 485. GR and NSR, when coupled with SSO provide the foundation for fast recovery from a processor failure and allow the use of ISSU to perform software upgrades with little downtime. SSO is necessary to handle other non routing protocol related items needed for the router to operate following a switchover. These include syncing the complete router configuration, Cisco Express Forwarding (CEF) forwarding entries and other needed information to the standby processor.

Graceful Restart and Non Stop Routing both allow for the forwarding of data packets to continue along known routes while the routing protocol information is being restored (in the case of Graceful Restart) or refreshed (in the case of Non Stop Routing) following a processor switchover.

When Graceful Restart is used, peer networking devices are informed, via protocol extensions prior to the event, of the SSO capable routers ability to perform Graceful Restart. The peer device must have the ability to understand this messaging. When a switchover occurs, the peer will continue to forward to the switching over router as instructed by the GR process for each particular protocol, even though in most cases the peering relationship needs to be rebuilt. Essentially, the peer router will give the switching over router a grace period to re-establish the neighbor relationship, while continuing to forward to the routes from that peer. Graceful Restart is available today for OSPF, ISIS, EIGRP, LDP and BGP. Standards are defined for OSPF, ISIS, BGP and LDP to ensure vendor interoperability.

When Non Stop Routing is used, peer networking devices have no knowledge of any event on the switching over router. All information needed to continue the routing protocol peering state is transferred to the standby processor so it can pick up immediately upon a switchover. The device does not inform its peers it is going through an SSO event.

Unlike Graceful Restart, Non Stop Routing uses more system resources due to the information transfer to the standby processor. Standards are not necessary in the case of Non Stop Routing, as the solution does not require any additional communication with protocol peers. NSR is used in cases where the routing protocol peer does not support the RFCs necessary to support Graceful Restart.

NX-OS offers a multilayer approach to routing protocol high availability by using both Non Stop Routing as well as Graceful Restart. It first attempts to use NSR. As an example if OSPFv2 experiences problems, NX-OS attempts to restart the process from its previous run time state. The neighbors would not register any neighbor event in this case. If the first restart is not successful and another problem occurs, NX-OS attempts a Graceful Restart.

Non Stop Forwarding is used in the following scenarios:

- First recovery attempt after process experiences problems
- ISSU
- User-initiated switchover using the system switchover command

A Graceful Restart is used in the following scenarios:

- Second recovery attempt after the process experiences problems within a 4-minute interval
- Manual restart of the process using the restart isis command
- Active supervisor removal
- Active supervisor reload using the reload module active-sup command

In NX-OS Non Stop Routing and Graceful Restart are available for BGP, IS-IS, OSPFv2, OSPFv3, and EIGRP.

9.9 NX-OS fast failover features

Fast convergence is a large and important topic in any modern data center network design. Typically, fast convergence refers to a collection of features as well as sub-features of particular protocols used for faster-than-default network convergence in response to a failure. Fast convergence features can be categorized into two general types:

- Fast Failure Detection (FFD): Features that affect the speed with which a neighbor failure event can be detected
- Fast Failure Reaction (FFR): Features that affect the speed with which protocols notify the rest of the network that a change has occurred and the speed with which protocols running on network devices recalculate the new state based on an event that was received

9.9.1 Fast Failure Detection

Most modern Layer 3 protocols have a concept of neighbor relationships and exchange keep-alives, heartbeats, or hellos, to maintain state. All Layer 3 protocols supported in Cisco NX-OS with exception of Routing Information Protocol Version 2 (RIPv2) use this mechanism. Hellos are used to establish initial communication with a neighbor and maintain a relationship.

Typically, each protocol has a certain interval that neighbors use to exchange hellos. If no hellos are received within a period specified by a particular time-out value, referred to as dead time or hold time by some protocols, the neighbor is considered to be dead, and any routes whose next hop was pointing to the dead neighbor have to be recalculated. It is important that the neighbor down detection occur as quickly as possible and modifying hello and dead intervals may be a way to achieve that. For example, default hello and dead timers for OSPF are 10 seconds and 40 seconds, respectively. Waiting for 40 seconds to realize that a neighbor is down would not be an acceptable behavior for many customers. Hence, modifying the default hello and dead time-out values may be the way to achieve Fast Failure Detection (FFD), and so these modified timers can be referred to as FFD timers.

As with all features and settings, you should apply them only if they are needed. In deciding whether applying FFD timers will make a difference in neighbor-down detection time, consider a few general scenarios:

- Data-path failures: If Layer 3 connectivity between two peers is point-to-point and a physical link fails completely, applying FFD timers typically will not improve the convergence time. Convergence time will not be improved because on the Cisco Nexus 7000 Series and many other products, link-down notification to protocols will always be faster than default dead-timer expiration. However, if Layer 3 connectivity between two peers is not point-to-point, applying FFD timers will make a difference. Typically, this scenario occurs when a Layer 2 switch or coarse or dense wavelength-division multiplexing (CWDM or DWDM) multiplexer is in the data path of hellos. In a failure in your DWDM infrastructure the two end devices could still have interface links up while traffic is not passing between them.
- Control-path failures: As mentioned in prior sections, Cisco NX-OS provides a comprehensive high-availability infrastructure in which the system manager monitors processes and executes a process restart if it detects a failure. In an extremely rare and unlikely case, the system manager may not detect any problems when a protocol running on the supervisor fails to send or receive hellos provided that the data path is operational. In such a case, reducing the hello interval and dead time-out value can reduce the time the system takes to detect a peer-down event.

This discussion applies to hello and dead-timer modifications for each protocol that needs to achieve FFD. When the number of protocols and interfaces grows large, you should consider other solutions to achieve FFD such as Bidirectional Forwarding Detection (BFD), discussed later in this chapter.

9.9.2 Fast Failure Reaction

The capability to detect a peer failure quickly is only a part of the solution to the problem of reducing network convergence time. The other important part of solving this problem is helping ensure that after a failure is detected, the network can recover quickly. Fast Failure Reaction (FFR) can be addressed by answering two questions. First, how quickly can the protocols running in network devices affected by a failure inform the rest of the network about what happened?

For example, in the case of OSPF, you can modify the speed with which LSAs are generated. Second, how quickly can protocols running in all network devices recalculate their states based on a received notification? OSPF allows you to modify the speed with which the shortest-path first (SPF) tree is recalculated based on received LSAs. Depending on the particular Layer 3 protocol supported by Cisco NX-OS, features exist to address one or both of these questions. These features typically consist of timers, which can be referred to as FFR timers, and other settings.

9.9.3 Bidirectional Forwarding Detection

As described earlier, FFD is a very important component of Layer 3 network convergence. However, modifying FFD timers for each protocol and for each interface on which that protocol runs can be problematic. Here are some important drawbacks of this approach:

- Increased CPU load because the supervisor CPU must generate hellos more often for each protocol and for each interface on which that protocol runs. For example, if the CPU use goes up to 100 percent, a switch may miss generation or processing of hellos. Hence, more false positives will be seen in the network due to shorter hello and dead intervals.
- Operational challenges because the administrator must configure modified timers for many protocols on multiple devices and keep track of FFD timer configuration best practices for each protocol.
- Wasted link bandwidth because multiple sets of hellos are sent on the wire at a short interval by different protocols.

Bidirectional Forwarding Detection (BFD) is a protocol that tries to solve these problems. It implements a lightweight hello communication mechanism and builds a neighbor relationship state.

BFD is described in RFC 5880. BFD is a detection protocol designed to provide fast forwarding-path failure detection times. BFD provides sub-second failure detection between two adjacent devices and can be less CPU-intensive than protocol hello messages because some of the BFD load can be distributed onto the data plane on supported modules.

Cisco NX-OS can distribute the BFD operation to compatible modules that support BFD. This process offloads the CPU load for BFD packet processing to the individual modules that connect to the BFD neighbors. All BFD session traffic occurs on the module CPU. The module informs the supervisor when a BFD failure is detected. In order to provide security, BFD does TTL checking on received packet, and you can also configure SHA-1 authentication of BFD packets.

BFD supports stateless restarts and in-service software upgrades (ISSUs). ISSU allows you to upgrade software without impacting forwarding. After a reboot or supervisor switchover, Cisco NX-OS applies the running configuration and BFD immediately sends control packets to the BFD peers. Also, BFD supports virtual routing and forwarding instances (VRFs).

You can configure the BFD session parameters for all BFD sessions on the entire device, or per interface. The per interface configuration overrides the entire device parameters. The BFD session parameters are negotiated between the BFD peers in a three-way handshake. The following BFD session parameters are included:

- Desired minimum transmit interval: The interval at which this device wants to send BFD hello messages.
- Required minimum receive interval: The minimum interval at which this device can accept BFD hello messages from another BFD device.
- Detect multiplier: The number of missing BFD hello messages from another BFD device before this local device detects a fault in the forwarding path.

In NX-OS, the minimum transit interval and the required minimum receive interval range is from 50 to 999 milliseconds and the default is 50. The multiplier range is from 1 to 50. The multiplier default is 3.

Once BFD is configured you can point several L3 protocols to take advantage of it. In other words if you configure BGP, OSPFv2 and HSRP to use BFD over an interface running it, if BFD declares its neighbor on that interface down all three of those protocols will declare their neighbor over that interface down and re-converge. Since BFD uses very fast timers failure discovery can happen very quickly.

General advantages of BFD include:

- Reduced CPU load because fewer per-protocol hellos need to be generated; each protocol sends initial hellos for neighbor discovery and synchronization, and after neighbors have been discovered, BFD hellos are used for keep-alive and failure detection purposes.
- Increased link bandwidth availability because only BFD sends hellos at a shorter interval while other protocols continue sending hellos at the default interval.
- Failure detection in less than a second can be provided; this BFD capability is especially important for certain protocol implementations that do not support sending hellos at intervals of less than a second.
- Reduced operational complexity because hello and dead timers do not have to be configured individually for each L3 protocol.

Several protocols can take advantage of BFD for fast failover in Cisco NX-OS, including OSPFv2, EIGRP, HSRP, Protocol-Independent Multicast (PIM) and BGP. BFD can also run on many types of interfaces, including physical interfaces, Layer 2 and Layer 3 PortChannels, and sub-interfaces.

You can get more details about configuring BFD here:

http://www.cisco.com/en/US/docs/switches/datacenter/sw/5_x/nx-os/interfaces/configuration/guide/if_bfd.html

Conclusion

The NX-OS Software is purpose built with data center operations in mind. It helps ensure continuous availability and provides design flexibility for a mission critical networks. It is one of the foundational tools that can be used to build a modern data center.

10

Data center network management and automation

Today’s data center network infrastructure is evolving from static to dynamic, and technologies such as cloud combined with highly virtualization and distributed workload applications are dominating the network backbone infrastructure. Aside, threats due to malicious attachments are increasing, and both businesses and consumers are looking at service providers to incorporate more defenses into the network. All these factors are pushing organizations towards a complex network infrastructure and as the data center environment becomes more distributed, complexity increases exponentially, and manageability becomes an even greater challenge.

A converged, dynamic infrastructure can help you address today's challenges and tomorrow's opportunities. Smarter data center solutions from IBM and Cisco can increase resource utilization, support business continuity, provide higher availability and faster provisioning, and help you achieve productivity gains through virtualization, optimization, energy stewardship, and flexible sourcing. IBM and Cisco together can help you to plan and deploy a responsive networking infrastructure that adapts quickly, cost effectively, and securely to rapidly changing demands.

In this chapter, we discuss how recent efficient, standards based innovations in data center networking can meet the agile requirements, such as security and performance, of the next generation data center. We describe how improved economics of IT can be achieved by connecting servers and storage with a high-speed and intelligent network fabric that is smarter, faster, greener with more simplified management. We also describe some of the products available from both IBM and Cisco that will lead to better management and automation, and how they work together to provide an effective, all-encompassing solution that harnesses the creative and intellectual abilities of both companies. In addition, we briefly introduce some of the important considerations for comprehensive data center management.

10.1 Data center network management requirements

IT departments today are challenged to look beyond traditional silos of networking and storage to manage this converged, virtualized data center, and modern data centers are becoming increasingly complex and massive. Proliferation of new technologies such as virtualization is adding yet another level of complexity while enabling higher workloads to be placed on the network. Thus, it becomes more imperative to continuously monitor and manage the enterprise network in a comprehensive manner. In the following topics we introduce some of the important characteristics and the associated solutions for effective and efficient data center network management.

10.1.1 Topology discovery

Given the complexity of the traditional data center network, being able to access a logical representation of the topology that is discovered by management tools is invaluable. This discovery can be done in a variety of ways to gather the total topology (SNMP Sweep, Seed Switch, IP address range). There also needs to be a facility which allows the administrator(s) to get an overview of the environment based on multi-tenancy, co-location facilities, or other compliance or regulatory needs. We discuss some of the network management tools able to provide this detail.

10.1.2 Syslog

As a monitoring or troubleshooting facility there are few more important components than the syslog. The syslog records messages to the switch console (and, optionally, to a logging server on another system). Not all messages indicate a problem with the system. Some messages are purely informational, while others might help diagnose problems with links, internal hardware, or system software. Any management facilities which provide a single "pane of glass" management for the data center network, or portions of it, should gather and interpret these syslog messages and provide notification capabilities based on the severity of the event.

10.1.3 Event management

While we already mentioned the Syslog capabilities as being important, which to a certain degree includes event management, the extension of that is to collect this data from multiple switches and provide interpretation of those events. Event messages (syslog, SNMP, or other types) are often cryptic, so having intelligent interpretation and providing resolution guidance if applicable can help to speed along resolution.

10.1.4 Role Based Access Control (RBAC)

Managing multiple administrators is key for any network infrastructure. Just having multiple users is really not enough; being able to delineate the allowed functionality per login is just as important. With the trend towards I/O convergence where multiple groups share configuration, troubleshooting, and monitoring capability on a single physical switch it is important that roles and responsibilities are maintained. RBAC with permissions fulfills this requirement.

10.1.5 Network configuration management

Configuration management is the most fundamental feature provided by any network management tool, whether this is on the LAN or the SAN side. Many network teams choose to use the command line interface (CLI) usually driven by the ability to script a large number of changes. On the other hand, GUI configuration tools provide one incredibly valuable configuration feature, feedback. GUI configuration tools can provide consistency checks and potentially judge the impact of the change. The best example of this is when implementing a zoning change in a storage area network. Before committing the zoning change most GUI management tools will provide a delta of what is being committed. Another advantage of the GUI is to simplify the configuration of advanced or complex features.

10.1.6 Network change management

Network change management and network configuration management are two sides of the same coin. After a feature is implemented, mechanisms must be in place to validate the configuration or to back out a configuration back to the last known good configuration. The other feature that needs to be available when managing change windows in the data center is the ability to figure out “what changed.” That ability to compare the current configuration to recently backup, or last known good, running configuration can help to eliminate parsing through seemingly countless lines of configuration.

10.1.7 Network performance management

Performance monitoring and management within the data center infrastructure sounds fairly straightforward and yet the complexity of monitoring the hundreds or thousands of ports makes it a very complicated task. Add to that complexity the realization that performance monitoring is not about an individual port but is more about that traffic flow through a network end-to-end and how that impacts the applications performance and responsiveness. Given those facts network performance management requires multiple pieces. Alerting based on the appropriate thresholds for the environment, real-time performance statistics plus historical data to provide context to the data, and predictive capabilities to forecast hotspots in the environment.

10.1.8 Provisioning

Provisioning is not the same as configuration. The aim of provisioning is to be able to provide dynamic template based configuration which can be used as a stand-alone feature or leveraged for more complex automation tasks. The move towards cloud architectures leverage dynamic provisioning of resources within the data center makes ability to deliver this increasingly important.

10.2 Cisco Prime Data Center Network Manager LAN and SAN

With the increase in virtualization, data centers are being transformed and IT departments must be empowered to effectively manage this transformation. The new release of Data Center Network Manager joins Cisco Prime Infrastructure family of management products on the journey to unify management of Cisco Data Center. Cisco Prime Data Center Network Manager (DCNM) supports Unified Fabric by combining management of both Ethernet and storage networks in a single dashboard. The dashboard enables network and storage administrators to troubleshoot health and performance across the whole range of Cisco NX-OS capable platforms, including the Cisco Nexus and Cisco MDS 9000 Families. DCNM also provides separate GUI tools for the LAN and SAN administrative teams so that traditional disciplines of managing the unique requirements of the LAN and SAN are maintained. The separate tools can be further leveraged with roles based access control to add an additional level of security and change management control.

10.2.1 Features and benefits of Cisco Prime DCNM

Cisco DCNM provides a robust framework and comprehensive feature set that meets the routing, switching, and storage administration needs of present and future virtualized data centers. Cisco Prime DCNM offers visibility into the virtualized hosts by integrating with industry standard hypervisors and provides host-tracking capability to easily manage and diagnose virtual and physical servers. Cisco DCNM streamlines the provisioning of the fabric and proactively monitors the LAN and SAN components. Offering an exceptional level of visibility and control through a single management pane for the Cisco Nexus and Cisco MDS 9000 family products, Cisco DCNM is the Cisco recommended solution for managing mission-critical data centers.

10.2.2 DCNM automated discovery

Cisco DCNM discovers the topology of the environment based off of a seed switch IP address. The administrator can control how many switch hops we wants to discover relative to the LAN topology or scope which VSANs will be manageable on the SAN side. After initial discovery, DCNM continuously captures any changes in the network after discovery to allow for a real-time accurate representation of the network.

- Using automated network discovery, provides up-to-date physical and logical inventory information
- Tracks inventory and performance information in real time; information can be used as a source of truth for asset tracking or as a data source for a configuration management database (CMDB)
- Discovery has the ability to scope what is discovered based on VSAN and VLAN

10.2.3 DCNM Web dashboards

DCNM web dashboards offer a real-time fabric and network health summary with detailed views of individual network components, enabling IT operations staff to respond quickly to events based on their severity. A single dashboard combines the management of both Ethernet and storage networks. The dashboard enables network and storage administrators to troubleshoot health and performance across the whole range of Cisco NX-OS Software platforms, including the Cisco Nexus and Cisco MDS 9000 Families, regardless of protocol type.

Web based DCNM dashboards provides the following capabilities:

- Operational monitoring views of SAN, LAN, and server environments
- Domain driven dashboards for host, storage and switch
- Context driven searches launch within domain dashboards
- Top Talkers: Shows heaviest host and storage ports to give quick data points for potential hot spots
- Capable of copying dashboard to a relevant view based on LAN, SAN, or custom defined environment

10.2.4 DCNM Web Topology View

Detailed topology views provide end-to-end path information across multiple fabric clouds, including the shortest path. DCNM Web Topology Views provide the following information:

- Displays real-time operationally focused topology of the data center infrastructure
- Offers Layer 2 overlay topology maps to streamline the troubleshooting process and reduce the mean time to repair; roll the cursor over the topology to view detailed information about paths and switch attributes
- Topology view showing ISL, switch CPU, and switch memory utilization

10.2.5 DCNM event management

The Event Browser and feature-specific Events tabs in Cisco DCNM enable IT operations staff to view and manage recent status events. Events include status-related system messages that Cisco DCNM retrieves from managed devices and messages generated by the Cisco DCNM server.

DCNM event management provides the following capabilities:

- A real-time network health summary with detailed view of individual network components, enabling operations staff to respond quickly to events based on their severity
- Ability to acknowledge working on the alert and when resolved, delete it
- Forward syslog alerts based on monitored facility
- Combines Event and Syslog for all monitoring switches for one “event” log to aid in determining the root of a chain of events

Event forwarding

Event forwarding provides the following capabilities:

- Enables integration with enterprise operations console (NOC) for alerts and events
- Uses email and traps to notify operations staff of service disruptions
- Add context to path alert by identifying name of host, ISL and storage entity

10.2.6 DCNM web templates

A template based provisioning and configuration feature of Cisco DCNM simplifies the deployment of SAN and LAN components. In addition, DCNM provides the predefined template scripts for common tasks, which IT administrators also can customize, also can import their own scripts into template formats and store them in the template repository.

These are some of the benefits of using DCNM Web templates:

- Pre-built templates for provisioning LAN and SAN components
- Pre-built template deployment scheduler and rollback mechanism
- Customizable templates with conditional statements
- Creation of new templates using a template editor
- Ability to import a configuration script and turn it into a template

10.2.7 DCNM performance and capacity management

The DCNM performance and capacity management feature is used for performance management enhancements, including the capability to trend and correlate across multiple performance indicators on a single chart, interactive zooming options, and predictive analysis. It provides the following capabilities:

- Detailed visibility into real-time and historical performance statistics in the data center
- Provides insight into port and bandwidth utilization, error count, traffic statistics, and so on
- Includes scheduled custom reports that can be off-loaded for post processing
- Provides capability to overlay several ports for performance comparison. Add in validation deployment scheme, multipathing, or load balancing as examples
- Provides predictive analysis based on historical data gathered to identify potential congestion points and to aid in capacity planning

10.2.8 DCNM capacity manager

Understanding resource consumption is essential to increase throughput of applications and projects into the data center. The DCNM capacity manager provides the following capabilities:

- Track port utilization by port tier and predict when an individual tier pool will be consumed
- Chart view of port consumption based on custom groupings

10.2.9 DCNM configuration and change management

DCNM configuration and change management enables pre-deployment validation of configuration changes, reducing the opportunities for misconfiguration; also provides the capability to compare current configuration files and copy back previously backed up files to the selected switches. It provides the following capabilities:

- Pre-deployment validation of configuration changes, reducing opportunities for human error
- Using historical configuration archive coupled with configuration comparison, enables a network administrator to identify the last-known good state if configuration problems occur
- Provides capability to back up configuration files from all switches

10.2.10 DCNM: VMpath analysis for LAN and SAN

DCNM extends network visibility to the virtual infrastructure by mapping the entire path from the virtual machine and switch to the physical storage and Ethernet networks. The Cisco VMpath feature that is part of the host dashboard, provides views to help troubleshoot the performance of a virtual machine and virtual host while also reporting on the health and performance of the network and storage paths associated to that virtual machine. The virtual machine-aware dashboard displays performance charts, path attributes, topology, path alerts, and information about the utilization of virtual machines and virtual hosts. The increased visibility into virtualized infrastructure helps IT administrators locate performance anomalies behind or in front of the hypervisor that may be causing service degradation as well as eliminate virtual computing and networking as a root cause of the problem.

Following are some of the key highlights:

- Provides a view of the virtual machine path through the physical network to the storage array and to the data store
- Provides the capability to view performance for every switch hop, all the way to the individual VMware vSphere server and virtual machine
- VM Performance stats on CPU, memory, latency, traffic, and errors

10.2.11 DCNM reports

The DCNM Web Based Reports menu allows the administrator to create customized reports based on historical performance, events, and inventory information gathered by the Cisco DCNM. You can create aggregate reports with summary and detailed views. Also, you can view previously saved reports in the included repository or you can choose to have DCNM email the reports in web or CSV formats.

These are a few key highlights:

- Allows network administrators to build custom reports from predefined templates
- Provides easy-to-schedule reports that can be exported for postprocessing or sent by email

10.2.12 DCNM image management

This feature enables easy-to-perform, nondisruptive (In-Service Software Upgrade (ISSU) mass deployment of Cisco NX-OS Software images, which can be scheduled or run on demand.

10.3 Cisco Virtual Network Management Center

Cisco Virtual Network Management Center (VNMC) provides centralized multi-device and policy management for Cisco network virtual services. The product addresses those issues by automating processes, freeing staff to focus on optimizing the network environment. Cisco VNMC supports greater scalability along with standardization and consistent execution of policies.

When combined with the Cisco Nexus 1000V Switch, ASA 1000V Cloud Firewall, or the Cisco Virtual Security Gateway (VSG), the IT team network team can implement the solution to provide:

- Rapid and scalable deployment through dynamic, template-driven policy management based on security profiles
- Easy operational management through XML APIs to help enable integration with third-party management and orchestration tools
- A nondisruptive administration model that enhances collaboration across security and server teams while maintaining administrative separation and reducing administrative errors
- Support for edge firewalls to enable:
  - Adding and configuring edge firewalls
  - Creating and applying edge security profiles that contain access control list (ACL) policy sets (ingress and egress), connection time-out, Network Address Translation (NAT) policy sets, TCP intercept, VPN interface policy sets, and more
  - Site-to-site IPsec VPNs

10.3.1 Improving security and enhance collaboration

Cisco VNMC operates in conjunction with the Cisco Nexus 1000V Virtual Supervisor Module (VSM) to improve operations and collaboration across IT. It streamlines the services performed by security, network, and server administrators.

10.3.2 Security administrator

This solution allows the security administrator to author and manage security profiles and Cisco Virtual Security Gateway (VSG) instances through the VNMC programmatic interface with Cisco Nexus 1000V. Cisco VSG provides trusted multi-tenant access with granular, zone-based, and context-aware security policies.

Cisco VNMC also manages the Cisco ASA 1000V Cloud Firewall to enable rapid and scalable security at the edge through dynamic, template-driven policy management.

10.4 Cisco Access Control System (ACS)

To meet new demands for access control management and compliance, and to support the increasingly complex policies that this requires, the Cisco Secure Access Control System can fulfill these needs.

The enterprise can gain the following benefits from the Cisco Secure Access Control System:

- Receive support for two distinct protocols: RADIUS for network access control and TACACS+ for network device access control
- Use multiple databases concurrently for maximum flexibility in enforcing access policy. Cisco ACS can also interact with LDAP services providing a comprehensive authentication framework.
- Enjoy increased power and flexibility for access control policies that may include authentication for various access requirements including wireless and VPN use cases.
- Get integrated monitoring, reporting and troubleshooting components, accessible through an easy-to-use, web-based GUI

Use Cisco Secure Access Control System to complement your existing infrastructure and enhance visibility and control across the domain. It offers central management of access policies for device administration and for wireless and wired 802.1X network access scenarios.

10.5 Cisco Prime Infrastructure

Cisco Prime Infrastructure combines the wireless functionality of Cisco Prime Network Control System (NCS) and the wired functionality of Cisco Prime LAN Management Solution (LMS), with the rich application performance monitoring and troubleshooting capabilities of Cisco Prime Assurance Manager.

This single solution enables IT organizations to consolidate tools and converge workflow, reducing operational overhead and increasing productivity. It provides an operational model based on lifecycle processes aligned with the way network operators do their jobs.

Cisco Prime Infrastructure provides these capabilities:

- Wired lifecycle functionality such as inventory, configuration and image management, automated deployment, compliance reporting, integrated best practices, and reporting
- Integration with Cisco knowledge base helps to ensure optimal service and support, product updates, best practices, and reports to improve network availability
- End-to-end network visibility of applications and services to quickly isolate and troubleshoot performance issues
- Configuration compliance auditing against best practices baselines or regulatory standards such as Payment Card Industry Data Security Standard
- All of the existing Cisco Prime NCS capabilities for converged user access visibility, wireless lifecycle management, troubleshooting, radio frequency planning, and optimization
- A wide variety of predefined reports for up-to-date information on the network including detailed inventory, configuration, compliance, audit, capacity, end-of-sale, security vulnerabilities, and many more
- Powerful REST-based APIs to gather and distribute network information for operations, capacity planning, and business intelligence

You can learn more about Cisco Prime Infrastructure from the datasheet located here:

http://www.cisco.com/en/US/prod/collateral/netmgtsw/ps6504/ps6528/ps12239/data_sheet_c78-715144.html

10.6 Cisco Network Services Manager

The Network Services Manager network abstraction layer helps in provisioning and deploying the numerous individual network components as sophisticated network “containers.”

IT administrators can create these containers:

- Across single and multi-PoD cloud computing deployments
- Much more easily and quickly than with template- and script-based systems

In addition, Cisco Network Services Manager:

- Dramatically reduces network operational costs and potential misconfiguration
- Optimizes capacity use and accelerates service delivery

10.6.1 Deploying flexible, policy-directed management

Cisco Network Services Manager offers a flexible, policy-directed approach to managing and controlling cloud computing network services. Through a configuration user interface, Network Services Manager helps administrators dynamically define and control an array of behaviors in their cloud computing environment, including these:

- Creating different levels of service capability or “service tiers” for tenant use
- Defining the capabilities and resources available in each tier
- Structuring a system of “containment” tailored to tenant application and deployment model needs.

Table 10-1 shows the Cisco network services manager’s features and their description.

Table 10-1 Cisco Network Manager Services features and description

Feature	Description
Abstraction and virtualization	Network Services Manager virtualizes network services by abstracting a logical representation of the physical network that it manages. The abstraction and virtualization allow you to provision and deploy numerous individual network components quickly, thereby reducing network operations costs and accelerating service delivery.
Policy implementation and adherence	Policies enforce topology models, network features, and network behaviors, thereby ensuring adherence to organizational or architectural requirements, such as security or access to resources.
Dynamic topology support	Network Services Manager retrieves abstract information from a business model, converts that information into device-specific configurations, and implements the resulting configurations on the appropriate devices. If the physical topology changes and the topology model is updated, Network Services Manager evaluates the defined policies against the new topology model to determine the best option for providing the required services while adhering to the rules and policies that you defined. Network Services Manager automatically disseminates the updated device-specific configurations to the affected devices to ensure ongoing adherence to policy.
Use of network containers	A network container is a logical group of virtual network resources that are created and managed as a unit. With network containers, Network Services Manager enables you to quickly and easily configure physical and virtual network infrastructure and network services to interoperate with computing and storage resources.
Administration user interface (UI)	Network Services Manager provides a browser-based Administration UI that enables you to view multiple virtual platforms (domains) and domain-specific policies.
Northbound (NB) Application Programming Interface (API)	The Network Services Manager NB API provides an integration interface for automated network provisioning as part of a larger cloud environment. The NB API enables you to instantiate cloud service models and topologies, create new cloud service offerings, modify existing offerings, and support containers and workloads. Additions, changes, or deletions made using the NB API are reflected in the Administration UI.

10.7 Cisco Security Manager (CSM)

Cisco Security Manager (CSM) helps to enable consistent policy enforcement and rapid troubleshooting of security events, offering summarized reports across the security deployment. Using its centralized interface, organizations can scale efficiently and manage a wide range of Cisco security devices with improved visibility.

Cisco Security Manager provides a comprehensive management solution for:

- Cisco ASA 5500 Series Adaptive Security Appliances
- Cisco intrusion prevention systems 4200 and 4500 Series Sensors
- Cisco AnyConnect Secure Mobility Client

The key benefits of CSM are described in the following sections.

10.7.1 Policy and object management

Policy and object management provides these benefits:

- Helps enable reuse of security rules and objects
- Facilitates process compliance and error-free deployments
- Improves ability to monitor security threats

10.7.2 Event management

Event management provides these benefits:

- Supports syslog messages created by Cisco security devices
- Facilitates viewing of real-time and historical events
- Provides rapid navigation from events to source policies
- Includes pre-bundled and customizable views for firewall, intelligent protection switching (intrusion prevention systems), and VPN

10.7.3 Reporting and troubleshooting

Reporting and troubleshooting provides these benefits:

- Provides system and custom reports
- Offers export and scheduled email delivery of reports in CSV or PDF format
- Provides advanced troubleshooting with tools such as ping, traceroute, and packet tracer

10.8 Cisco Prime Network Registrar (CNR)

Cisco Prime Network Registrar provides integrated, scalable, reliable Domain Name System (DNS), Dynamic Host Configuration Protocol (DHCP), and IP Address Management (IPAM) services for both IPv4 and IPv6.

Figure 10-1 shows the Cisco Prime Network Registrar functions.

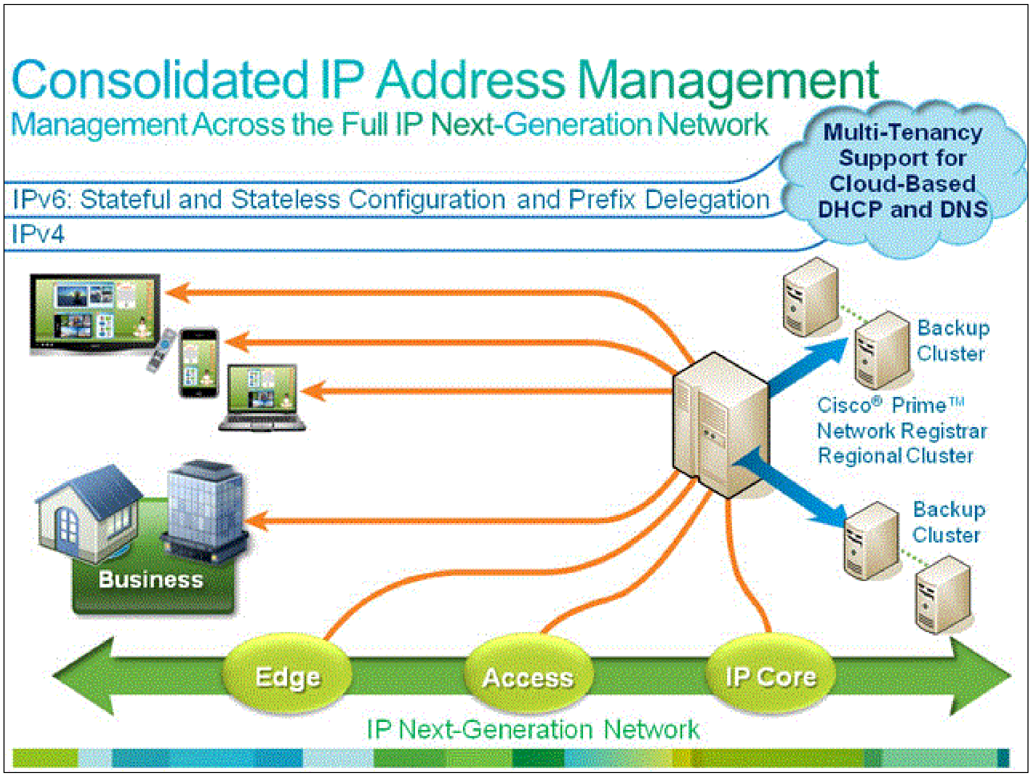

Consolidated IP Address Management
Management Across the Full IP Next-Generation Network
IPv6: Stateful and Stateless Configuration and Prefix Delegation
IPv4
Multi-Tenancy Support for Cloud-Based DHCP and DNS
Backup Cluster
Cisco® Prime™ Network Registrar Regional Cluster
Backup Cluster
Business
Edge
Access
IP Core
IP Next-Generation Network

Figure 10-1 Cisco Prime Network Registrar

Organizations can only be as agile as their IP infrastructure can support. For example, managing thousands of IP addresses manually creates bottlenecks when provisioning, as well as increasing the possibility of service outages caused by human error.

The growing complexity of networks further increases the difficulty of managing today's networks. Operators must accommodate new types of servers and clients, potentially from multiple vendors. TCP/IP continues to connect more devices, resulting in a higher cost to manage each new device as the number of devices added to the network increases. Furthermore, new technologies like IPv6, virtualization, cloud services, and mobile connectivity increase management complexity and drive the need for a comprehensive, integrated, and feature-rich IP address management (IPAM) solution.

Agility helps enable organizations to cut operating expenses by facilitating new ways to identify waste, expense, and risk in everyday network management processes. Networks that are smarter through automation and simplification of the allocation, tracking, and reclaiming of IP addresses can break the vicious cycle of rising costs, increased downtime, and endless stress to network staff. To increase operational efficiency and build a network that responds quickly to changing business needs, operators need a solution with these capabilities:

- Centralizes IP address management for the entire network
- Maintains a complete and up-to-date IP address inventory
- Provides comprehensive visibility into IP usage with inventory assurance
- Eliminates manual processes prone to human error
- Simplifies IP address management and accelerates troubleshooting to help enable operators to spend less time managing the network and more time focused on strategic initiatives

- Dynamically scales to match changing IP needs
- Allows multiple administrators with individual and group access controls
- Helps enable a seamless migration from IPv4 to IPv6 without affecting service availability
- Supports user-defined IP blocks and address types with custom fields
- Reduces security and compliance risks associated with unauthorized use of addresses
- Provides extensive reporting capabilities
- Reduces overall operating expenditures

Cisco Prime Network Registrar IPAM is designed to optimize operator workflow efficiency by reducing task steps and eliminating manual input where possible. This is achieved by centralizing IP address space management while automating many common management tasks, including discovery of devices, IP address allocation, reconciliation of IP addresses, and support for both IPv4 and IPv6 within the same network. Operators are also able to manage all IP allocation-related tasks from a single point of access.

To learn more about Cisco Prime Network Registrar, see this website:

http://www.cisco.com/go/cnr

10.9 Network capabilities enabling application performance management

This section discusses various network capabilities that affect application performance.

10.9.1 Netflow

NetFlow is embedded within Cisco IOS and NX-OS Software to characterize network operation. Visibility into the network is an indispensable tool for IT professionals. In response to new requirements and pressures, network operators are finding it critical to understand how the network is behaving, including these:

- Application and network usage
- Network productivity and utilization of network resources
- The impact of changes to the network
- Network anomaly and security vulnerabilities
- Long term compliance issues

Cisco IOS and NX-OS NetFlow fulfills those needs, creating an environment where administrators have the tools to understand who, what, when, where, and how network traffic is flowing. When the network behavior is understood, business process will improve and an audit trail of how the network is utilized is available. This increased awareness reduces vulnerability of the network as related to outage and allows efficient operation of the network. Improvements in network operation lower costs and drives higher business revenues by better utilization of the network infrastructure.

NetFlow facilitates solutions to many common problems encountered by IT professionals.

Analysis of new applications and their network impact

Identify new application network loads such as VoIP or remote site additions.

Reduction in peak WAN traffic

Use NetFlow statistics to measure WAN traffic improvement from application-policy changes; understand who is utilizing the network and the network top talkers.

Troubleshooting and understanding network pain points

Diagnose slow network performance, bandwidth hogs and bandwidth utilization quickly with command line interface or reporting tools.

Detection of unauthorized WAN traffic

Avoid costly upgrades by identifying the applications causing congestion.

Security and anomaly detection

NetFlow can be used for anomaly detection and worm diagnosis along with applications such as Cisco CS-mars.

Validation of QoS parameters

Confirm that appropriate bandwidth has been allocated to each Class of Service (CoS) and that no CoS is over- or under-subscribed.

10.9.2 Switch Port Analyzer (SPAN)

Network administrators can analyze network traffic passing through ports by using SPAN to send a copy of the traffic to another port on the switch that has been connected to a SwitchProbe device or other Remote Monitoring (RMON) probe or security device. SPAN mirrors received or sent (or both) traffic on one or more source ports to a destination port for analysis. The network analyzer can be a Cisco SwitchProbe, a Fibre Channel Analyzer, or other Remote Monitoring (RMON) probes. This ability to mirror traffic non-disruptively also trace analysis to be done without contaminating the “crime” scene. When troubleshooting or root cause analysis gets to the point of needing packet captures the last thing the administrative team wants to do is reset the links by placing a packet capture tool in-path which may inadvertently make it difficult to find the true problem by resetting the link.

10.9.3 Algo Boost

Cisco Algorithm Boost or Algo Boost technology is a networking innovation with patents pending, that offers the highest speed, visibility and monitoring capabilities in the networking industry.

Ideal for high performance trading, Big Data and high performance computing environments, this new technology offers network access performance as low as 190 nanoseconds.

The first switch to integrate the Cisco Algo Boost technology is the Cisco Nexus 3548 full-featured switch which pairs performance and low latency with innovations in visibility, automation, and time synchronization. It is tightly integrated with the feature set of the Cisco Nexus Operating System, a proven operating system used in many of the world’s leading data centers, creating a truly differentiated offering.

To give customers a view into how much latency is being introduced by microbursts, Cisco Algo Boost analytical framework includes active buffer monitoring that provides highly granular data and pro-active notifications on switch congestion. Compared to software-driven techniques that take periodic snapshots of the packet buffer, Algo Boost utilizes hardware to provide millisecond-level polling to form rich histograms, charting buffer usage over time.

10.9.4 PoAP

PowerOn Auto Provisioning (PoAP) automates the process of upgrading software images and installing configuration files on Cisco Nexus switches that are being deployed in the network for the first time.

When a Cisco Nexus switch with the PoAP feature boots and does not find the startup configuration, the switch enters PoAP mode, locates a DHCP server and bootstraps itself with its interface IP address, gateway, and DNS server IP addresses. It also obtains the IP address of a TFTP server or the URL of an HTTP server and downloads a configuration script to run on the switch to download and install the appropriate software image and configuration file.

NOTE: The DHCP information is used only during the PoAP.

PoAP requires the following network infrastructure:

- A DHCP server to bootstrap the interface IP address, gateway address, and DNS server
- A TFTP or HTTP server containing the configuration script used to automate the software image installation and configuration process
- One or more servers containing the desired software images and configuration files

10.9.5 Network Timing Protocol (NTP)

Network time synchronization, to the degree required for modern performance analysis, is an essential exercise. Depending on the business models, and the services being provided, the characterization of network performance can be considered an important competitive service differentiator. In these cases, great expense may be incurred deploying network management systems and directing engineering resources towards analyzing the collected performance data. However, if proper attention is not given to the often-overlooked principle of time synchronization, those efforts may be rendered useless.

Time is inherently important to the function of routers and networks. It provides the only frame of reference between all devices on the network. This makes synchronized time extremely important. Without synchronized time, accurately correlating information between devices becomes difficult, if not impossible. When it comes to security, if you cannot successfully compare logs between each of your routers and all your network servers, you will find it very hard to develop a reliable picture of an incident.

NTP is the standard, ubiquitous mechanism used to ensure a common clock and timestamp across the data center. It should be configured on all available devices (applications, servers, storage, network), ideally with a common, reliable clock service (NTP Server)

To learn more about NTP, view the best practices paper at this website:

http://www.cisco.com/en/US/tech/tk869/tk769/technologies_white_paper09186a0080117070.shtml

10.9.6 Cisco Embedded Event Manager

Cisco IOS and NX-OS Embedded Event Manager (EEM) is a powerful and flexible subsystem that provides real-time network event detection and onboard automation. It gives you the ability to adapt the behavior of your network devices to align with your business needs.

Your business can benefit from the capabilities of Embedded Event Manager without upgrading to a new version of Cisco Software. It is available on a wide range of Cisco platforms.

Embedded Event Manager supports more than 20 event detectors that are highly integrated with different Cisco IOS and NX-OS Software components to trigger actions in response to network events. Your business logic can be injected into network operations using IOS Embedded Event Manager policies. These policies are programmed using either simple command-line interface (CLI) or using a scripting language called Tool Command Language (TCL).

Harnessing the significant intelligence within Cisco devices, IOS and NX-OS Embedded Event Manager helps enable creative solutions, including automated troubleshooting, fault detection, and device configuration.

10.9.7 Device performance and capacity

Capacity planning is the process of determining the network resources required to prevent a performance or availability impact on business-critical applications.

Performance management is the practice of managing network service response time, consistency, and quality for individual and overall services.

Performance problems are often related to capacity. Applications are sometimes slower because bandwidth and data must wait in queues before being transmitted through the network. Capacity and performance management also has its limitations, typically related to CPU and memory. The following are potential areas for concern:

- CPU
- Backplane or I/O
- Memory and buffers
- Interface and pipe sizes
- Queuing, latency, and jitter
- Speed and distance
- Application characteristics

To learn more about capacity and performance management, read the best practices white paper, available at this website:

http://www.cisco.com/en/US/tech/tk869/tk769/technologies_white_paper09186a008011fde2.shtml

> Cisco and IBM have a long history of collaborating to create enhanced, valuable solutions for our customers. The two companies intend to continue to work together to build upon this initial “energy management” integration and to deliver advanced energy management solutions and capabilities. In addition, Cisco is playing a vital role in the IBM Green Sigma™ Coalition. For more information about the Green Sigma Coalition, go to this website:
>
> http://www.ibm.ibm.biz/Bdx2C7

10.10 IBM Tivoli solutions

As discussed previously, worldwide data centers are evolving significantly and moving towards more centralized dynamic model. Operating costs, service level agreements along with regulatory policies are becoming the critical points for any organization in manageability of these data centers. In the following sections, we discuss IBM solutions for efficient enterprise network management.

10.10.1 IBM Tivoli Netcool Network Management

Whether it is a small to large enterprise business, small to global communication service provider, government body or utility business, the uninterrupted availability and performance of mission-critical services is intrinsically tied to its organization's success. Gaining real-time visibility into infrastructure configuration and health is a key factor in optimizing operational efficiency and maximizing asset utilization. IBM Netcool® Network Management is a service assurance solution for real-time network discovery, network monitoring, and event management of IT domains and next-generation network environments. It delivers event, network, and configuration management enabling network device security and control in a single solution. The customizable web based user interface enabled through the Tivoli Integrated Portal infrastructure allows users to achieve end-to-end visualization, navigation, security and reporting (real-time and historical) across Tivoli and Cisco management tools.

Following are few key highlights about IBM Tivoli Netcool Network Management:

- Automate the configuration and management of network devices
- Improve the performance of cloud network
- Visualize services across a heterogeneous infrastructure
- Reduce operational costs
- Improve visibility of network bottlenecks

With event management, network management, access management, configuration change management and compliance management in a single solution with one intuitive interface, IT organizations can improve time to value for management of new network devices and improve the robustness of the network. Designed for use in both the smallest and largest deployments, this solution offers around the clock visibility, control and automation to help IT staff to deliver high-network availability and performance, driving uptime across the infrastructure and across the business. IBM Tivoli Netcool Network Management is one of the critical component in effective delivery of IT services. In the following sections, we discuss more about IBM Tivoli Netcool Configuration Manager and IBM Tivoli Network Performance Management.

Figure 10-2 shows a conceptual overview of the Network Manager functional layers and depicts a standard Network Manager installation.

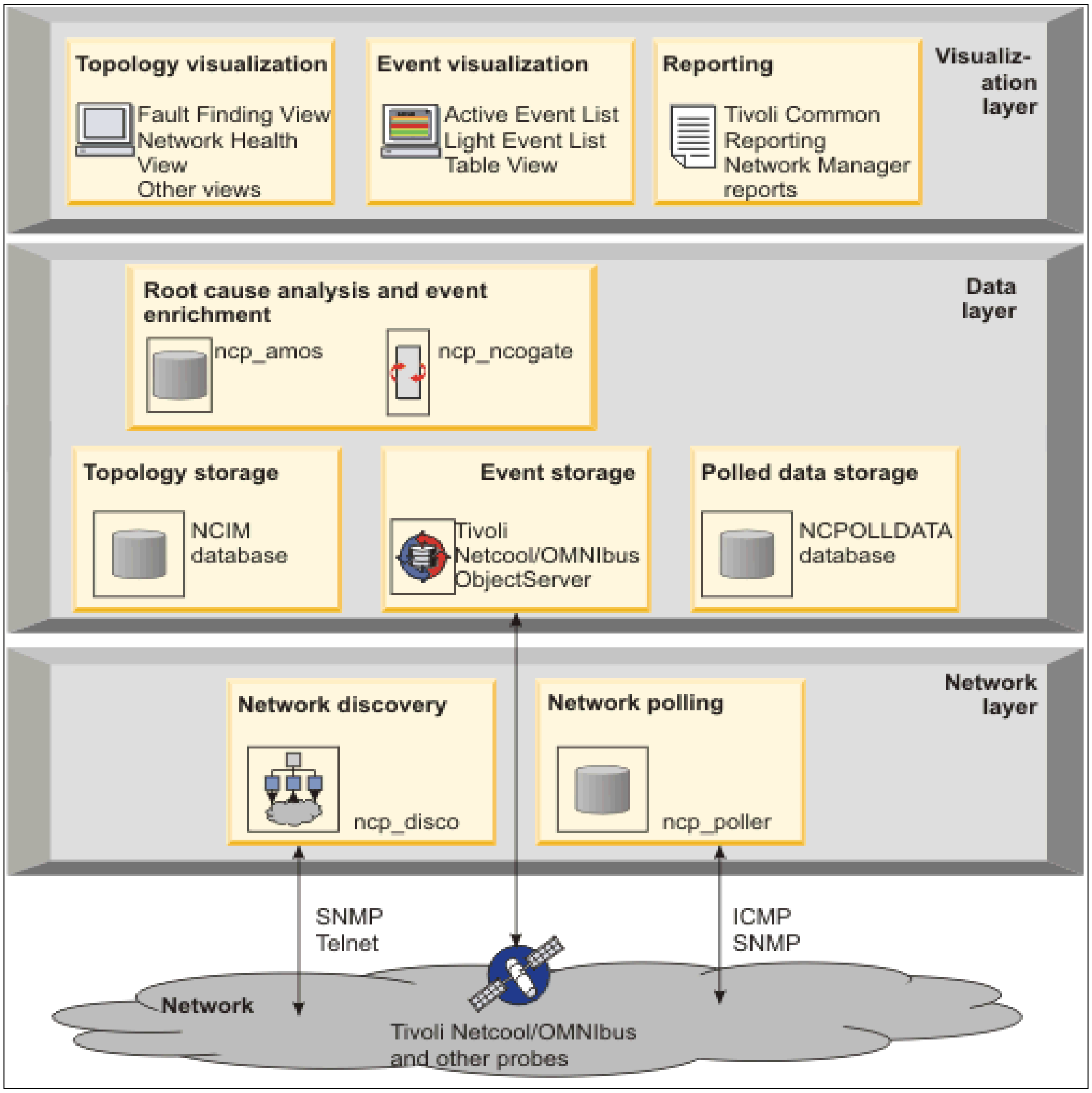


Figure 10-2 IBM Tivoli Network Manager functional layers

The Network Manager architecture can be divided into three layers: network layer, data layer, and visualization layer.

Network

The network layer interacts directly with the network. This layer contains network discovery and polling functionality. Network discovery retrieves topology data and network polling retrieves event data.

Data

The data layer stores the topology data retrieved by network discovery and the event data retrieved by network polling. Network polling also includes storage of polled SNMP data for reporting and analysis. This layer also provides root cause analysis functionality that correlates topology and events to determine the source of network faults, and event enrichment functionality that adds topology data to events.

Visualization

The visualization layer provides the tools operators and administrators need to view topology, view events, and run network troubleshooting tools.

> **Note:** Tivoli Netcool/OMNIbus is a separate product and It is possible to configure the Network Manager to include failover. Network Manager is designed to be installed with Tivoli Netcool/OMNIbus to enhance fault management, including root cause analysis, and correlation of alerts with the network topology.

10.10.2 IBM Tivoli Netcool Configuration Manager

Change is constant, whether within an individual network or reaching across the entire instrumented, interconnected, and intelligent world and managing change, especially when it comes to device configuration, is essential. Without configuration management, errors introduced during change can damage a host of critical operations from provisioning to performance, and from availability to security. So how do you manage this change? How do you ensure that changes are made efficiently and correctly? How do you track improper changes back to their source so you can correct them now and prevent more errors in the future?

The Tivoli Netcool configuration management solutions enable organizations to integrate data and capabilities from multiple tools to get the most from the Tivoli management implementation.

Following are the unique advantages of Tivoli Netcool configuration management solutions:

- Automation of network configuration tasks, from small networks to large complex networks, driving operational cost savings and improving network availability with the reduction of manual configuration errors
- SmartModel automation, which enables standardized representation of native device configuration
- A unified platform that provides a single point of access for all changes and policies
- Accurate configuration capabilities with nondisruptive rollback and reusable templates
- Extensive visibility into security, with access control based on roles, devices, and commands
- Always on compliance to help efforts to enforce regulatory, security and operational policies
- Support for approval and scheduling of unit of workflow
- Upgrade operating systems for the devices being managed

Configuration management solutions can extend management reach into devices and network areas that need greater attention, to help provide for increased network reliability and reduced chance of performance degradation due to error. Whether the need is to repair a problem that was introduced during an official or unofficial change, to thoroughly investigate the impact of a new configuration before making a change, to confirm the configurations already in place in a network, to automate changes to ensure they are made in a standardized and timely manner, or to roll back a change that has resulted in a problem, configuration management is an important component of the overall data center management solution stack. Network and configuration management solutions bring together data on topology, connectivity, performance, and other aspects of the network, adding value across the organization's entire management portfolio by increasing insight and enhancing management capabilities.

A configuration management solution is valuable in itself, especially best of breed solutions such as those provided by Tivoli Netcool software, but organizations can extend the value of Tivoli Netcool configuration management solutions even further by deploying them as part of the larger Tivoli network management portfolio (such as Tivoli Provisioning Manager (TPM), Tivoli Application Dependency Discovery Manager (TADDM), and Tivoli Change and Configuration Management Database (CCMDB)), integrating with other key functional areas such as fault management, network management, and performance management.

Note: Go to this website for the complete packaged solution:
`http://www.ibm.biz/Bdx2Bg`

10.10.3 IBM Tivoli Network Performance Flow Analyzer

Analysis and visualization of network traffic is important for optimizing and protecting the operation of networked IT infrastructures. The IBM Tivoli Netcool Performance Flow Analyzer is a flow based network traffic performance analysis system that provides detailed visibility on the composition of network traffic. It helps to plan for capacity and troubleshoot problems and provides rapid time to value with small footprint and easy installation. It gives network operators detailed visibility of network traffic behavior to quickly identify and resolve root causes of performance problems. It also helps proactively manage with detailed visibility of network traffic composition.

Tip: Tivoli Network Performance Flow Analyzer is a one of its kind solution that provides graphical display of both inflows and outflows on the same report. This helps users to quickly identify the root cause of the performance issue on a network device.

Highlights of IBM Tivoli Network Performance Flow Analyzer include these:

- Collects Cisco NetFlow versions v5, v7, v9 and the Internet Engineering Task Force (IETF) defined Internet Protocol Flow Information eXchange protocol (IPFIX) flow records
- Is designed to be highly scalable and cost effective   appropriate for basic enterprise environments through Tier 1 Communications Service Providers (CSP)
- Can improve capacity planning and network redesigns   helps optimize use of network resources so network designers are able to be more judicious in making plans knowing the composition and sources of network traffic
- Assists efforts to ensure mission critical applications are provided the necessary resources
- Identifies unanticipated traffic and isolates to source
- Identifies most of the applications that are consuming most network bandwidth and by what hosts within the network
- Has early detection of traffic pattern changes, changes that could impact service performance
- Locates business impacting network problems that are caused by incorrect configuration, unexpected customer traffic
- Validates network architecture updates and new network and traffic management configurations as well as confirms applications' traffic is being forwarded correctly
- Supports user defined aggregates that enable network operators to define the sets of information needed
- Delivers extremely fast report generation and response times due to in memory database

Figure 10-3 illustrates the IBM Tivoli Network Performance Flow Analyzer architecture and subsystems.

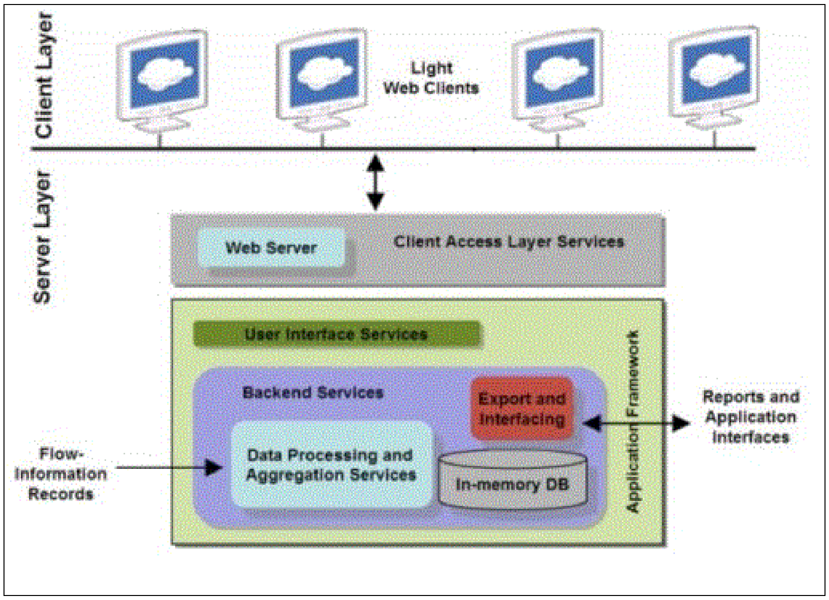


Figure 10-3 IBM Tivoli Network Performance Flow Analyzer architecture and subsystems

IBM Tivoli Network Performance Flow Analyzer is designed for high performance. The system uses parallelism on multicore architectures and uses a fast in memory aggregation database (ADB). A single installation can accommodate flow information records that are exported from many routers, switches, and interfaces. On a typical dual core server with default configuration, a processing speed of 50,000 flows per second can be achieved.

Note: Higher flow rates require more resources, a distributed setup, or a less complex configuration.

It also offers an API to the flow analyzer daemon. All configuration, control, and database access functions are supported by the API. The API can also be used for scripted output (for example, in CSV and PDF format). Periodically, advanced users can write custom scripts to export into a desired output format. The scripting language can also be used for event notification. Events can be signalled as syslog messages.

Note: For complete IBM Tivoli Network Management solutions and deployment steps, see this website:

`https://www.ibm.biz/Bdx2HW`

IBM Tivoli Netcool GUI Foundation Server

The IBM Tivoli Netcool GUI Foundation is a server application that runs GUIs from different IBM Tivoli products. The IBM Tivoli Netcool GUI Foundation provides single sign-on, consolidated user management, and a single point of access for different Netcool applications. The IBM Tivoli Netcool GUI Foundation also provides the ability to create customized pages and administer access to content by user, role, or group. The IBM Tivoli Netcool GUI Foundation is installed automatically with the first IBM Tivoli Netcool GUI Foundation-enabled product. Subsequent products may install updated versions of the IBM Tivoli Netcool GUI Foundation.

10.10.4 IBM Tivoli OMNIbus

IBM Tivoli OMNIbus enables real-time consolidation of enterprise-wide events and or management consoles, including events from a very-broad spectrum of IBM and CISCO software and management systems. OMNIbus delivers real-time, centralized monitoring of complex networks and IT domains (server, network, security, storage, and other IT management).

As mentioned OMNIbus bridges the gap across multiple networks and IT silos to help organizations improve the end-to-end availability of their applications and services. When OMNIbus software detects faults, they are processed in the ObjectServer, a high-speed, in-memory database that collects events from across the infrastructure in real time. OMNIbus then eliminates duplicate events and filters events through an advanced problem escalation engine. Thus, OMNibus enables IT staff to hone in on the most critical problems and even automate the isolation and resolution of those problems (Also, enables the sophisticated cross-domain correlation).

Enterprises can rely on OMNIbus to consolidate the management of networks and multiple management systems and tools under a “single pane of glass” view. OMNIbus make it easier for enterprises to manage problems across large heterogeneous networks and IT, and thereby reduce costs and improve overall productivity.

In addition, enterprises can deploy the IBM Tivoli Monitoring family that integrates with Netcool OMNIbus to proactively measure user experiences and performance across devices and services and generates alarms based on established thresholds. Example of such devices and service can be Layer 1, 2 and 3 network routers and switches, multiprotocol label switching (MPLS), virtual private network (VPN), asynchronous transfer mode (ATM), frame relay, synchronous optical network (SONET), Voice over IP (VoIP) and legacy Private Branch Exchange (PBX) based services

10.10.5 IBM Tivoli NetView for z/OS

The IBM Tivoli IBM NetView® for z/OS program provides functions to help maintain the highest degree of availability for IBM System z networks. The NetView program offers an extensive set of tools for managing and maintaining complex, multivendor, multi-platform networks and systems from a single point of control. It provides advanced correlation facilities to automate any network or system event, support for both TCP/IP and SNA networks, a set of user interfaces to meet the needs of any user, and management functions that work with other products to provide a complete picture of the networks and systems.

With open application programming interfaces, the NetView program can be an integration point for both z/OS and distributed vendors. The NetView program enables the management of networks and systems through graphical display and automation and can quickly guide support personnel to an appropriate response or even respond automatically. It reduces the need for manual resource definition and complex automation setup through production-ready automation and extends centralized management into multiple network environments. The NetView program can be used in an enterprise as a centralized manager, a mid-level manager, or a z/OS management endpoint.

These NetView capabilities results in the following benefits:

- Increased network and system efficiency and availability
- Centralized management for TCP/IP and SNA network environments, which reduces the need for duplicate network management systems

- Enhanced operations and message management support to improve and simplify operator interactions and to provide more control in automating and managing day-to-day operations
- Management of larger networks, more resources, and more systems with fewer resources and personnel

Below we discuss more in particular about added capabilities with release of IBM Tivoli NetView for z/OS V6.1 and below are some key highlights

- Analysis and automation of z/OS system, NetView, and job log messages via a consolidated log (Canzlog) for both real time and archived messages.
- Built-in analysis of packet trace data that reduces diagnostics time.
- Support for Dynamic Virtual IP Address Hot Standby distribution.
- Tighter integration with IBM Tivoli Network Manager.
- Enhanced NetView Discovery Library Adapter (DLA) for sysplex, z/OS system, and IP data.
- Support for IBM GDPS® Active/Active continuous availability solution.
- Usability and consistency updates for the NetView Management Console.
- Updated configuration style sheet report generator to aid in migration.

System automation

Tivoli NetView allows IBM MVS™ (Multiple Virtual Storage) messages to come to NetView for automation and browsing without having to make difficult configuration choices. A shared data space is established to capture all z/OS, NetView, and job messages into a consolidated log for automation. In addition, all message attributes are exposed and available for automation. Users can create their own attributes that can be passed on for further automation or for viewing by operators.

Enhanced IP management

Packet trace: The NetView IP packet trace function has been enhanced to analyze the data in a packet trace to identify potential network and system problems. This enhancement significantly reduces the time needed for problem identification and diagnosis. From within the packet trace session analysis, individual connections can be dropped and the packet trace data can be saved for later retrieval or further analysis.

Usability updates

The NetView management console user interface is updated to improve usability and provide consistency with other products across the Tivoli portfolio. The NetView Tivoli Enterprise Management Agent and associated workspaces have added support for new data types and support of 64-bit data.

IBM Tivoli NetView for z/OS V6.1 provides enhanced management for any business-critical IT network infrastructure and NetView's support of and participation in Integrated Service Management demonstrates the importance of network and systems management in providing visibility, control, and automation across both business and IT. By fully participating and working with both the CCMDB (Change Configuration Management Database) and TADDM (Tivoli Application Dependency Discovery and Manager) product sets, NetView for z/OS can preserve a company's investments in NetView and exploit its new functionality to better manage not just organizations IT resources but also the technical and business processes that govern how an enterprise functions.

In addition, the tighter integration between IBM Tivoli OMNIbus and Tivoli Network Manager allows the better view, creation, and display of network relationships and improved visibility to organizations distributed environment from z/OS and the MultiSystem Manager agent for IBM Tivoli OMNIbus and IBM Tivoli Network Manager in NetView gathers views and relationships and loads them into the NetView for z/OS Resource Object Data Manager (RODM) data cache. These resources can then be viewed and managed from the NetView Management Console (NMC) to provide improved distributed management.

10.10.6 IBM Security Network Intrusion Prevention System

In this new era of a digital world where IT infrastructure and networks are reaching new levels of complexity, it is essential to deal with growing Internet security threats proactively. Internet security breaches can result in financial losses, unauthorized access to confidential information, theft of intellectual property and service disruption. Around the world, organizations are looking for solutions that can offer innovative features to preemptively protect them from Internet threats. With clear understanding of the critical balance between performance and protection and, with a comprehensive line of high-performance models, the IBM Security Network Intrusion Prevention System is designed to deliver uncompromising protection for every network layer, protecting organization from both internal and external threats. Listed here are a few key highlights:

- Unmatched levels of performance without compromising breadth and depth of security
- Protection for business-critical assets such as networks, servers, endpoints and applications from malicious threats
- Reduced cost and complexity by consolidating point solutions and integrating with other security tools
- Evolving protection powered by IBM X-FORCE[1] research to stay ahead of the threat
- Secure web applications from threats like SQL injection and cross-site scripting attacks
- Monitor data security risks throughout network with integrated data-loss prevention
- Prevent network abuse and data loss from instant messaging and peer-to-peer file sharing

On other side, organizations often struggle to balance and optimize the following six key areas: performance, security, reliability, deployment, management and confidence and IBM Security Network IPS solutions deliver in all these areas.

Delivering superior performance without compromise

Security should enhance network performance, not detract from it. Purpose-built IBM Security Network Intrusion Prevention System solutions offer high throughput, low latency and maximum availability to maintain efficient network operations. This includes the ability to utilize the full spectrum of security protection and eliminates the need to choose between the highest levels of security and the performance required to maintain service levels for business-critical applications. By offering industry-leading performance beyond 20 Gbps of inspected throughput, IBM Security Network Intrusion Prevention System solutions provide the required performance, while delivering high levels of security.

Delivering high levels of availability

Devices placed in the flow of network traffic must be extremely reliable. IBM Security Network Intrusion Prevention System solutions offer the highest levels of reliability and availability through high-availability configurations (active/active or active/passive), hot-swappable redundant power supplies and hot-swappable redundant hard drives. In addition, the geographic high-availability option can use the management port to share quarantine-blocking decisions to ensure secure failover to a geographically remote standby device, if needed.

[1] http://www.ibm.com/security/xforce/

Provides ease of deployment

Each IBM Security Network Intrusion Prevention System comes preconfigured with the proven X-FORCE default security policy, which provides immediate security protection out of the box.It also features a Layer 2 architecture that does not require network reconfiguration. Network and security administrators can easily choose one of three operating modes: active protection (intrusion prevention mode), passive detection (intrusion detection mode) and inline simulation (simulates inline prevention).

Centralizing security management

IBM Security Network Intrusion Prevention System appliances are centrally managed by the IBM Security IBM SiteProtector™ system. The IBM Security SiteProtector System provides simple, powerful configuration and control of IBM agents, along with robust reporting, event correlation, and comprehensive alerting. Also included is IPv6 management support of IBM Security Network Intrusion Prevention System appliances, including the ability to display IPv6 events and IPv6 source/destination IP addresses.

10.10.7 IBM QRadar Network Anomaly Detection

In this section we brief you about the capabilities of a latest member of IBM Network Security systems portfolio, the IBM QRadar Network Anomaly Detection appliance. It analyzes complex network activity in real-time, detecting and reporting activity that falls outside normal baseline behavior. The analytics can look not only at inbound attacks, but also can detect outbound network abnormalities and using advanced behavioral algorithms, the QRadar Network Anomaly Detection appliance analyzes disparate data that can collectively indicate an attack network and traffic flows, intrusion prevention system (IPS) alerts, system and application vulnerabilities, and user activity. It quantifies several risk factors to help evaluate the significance and credibility of a reported threat, such as the business value and vulnerabilities of targeted resources.

By applying behavioral analytics and anomaly detection, the application can flag abnormal events such as these:

- Outbound network traffic detected to countries where the company does not have business affairs
- FTP traffic observed in a department that does not regularly use FTP services
- A known application running on a non-standard port or in areas where it is not allowed (such as unencrypted traffic running in secure areas of the network)

The QRadar Network Anomaly Detection appliance leverages the QRadar Security Intelligence Platform and is designed to complement IBM SiteProtector[2] and IBM Security Network IPS[3] deployments. This new appliance also receives a threat intelligence feed from IBM X-Force research, providing insight into suspect entities on the Internet based upon knowledge of more than 15 billion Web pages and images. The X-Force IP Reputation Feed provides QRadar Network Anomaly Detection with a real-time list of potentially malicious IP addresses – including malware hosts, spam sources and other threats. If the product sees any traffic to or from these sites, it can immediately alert the organization and provide rich contextual information about the activity.

[2] https://www.ibm.biz/Bdx2X9

[3] https://www.ibm.biz/Bdx2XL

10.11 Integration between Cisco and IBM solutions

Nowadays, every organization must leverage its IT infrastructure to remain competitive and meet the challenges of an on demand world. This requires extending the use of IT investments in networks, systems and applications to connect with customers, suppliers and business partners. While increased connectivity drives immense benefits, it can yield corresponding risks. Viruses, worms and Internet attacks can cause significant IT infrastructure damage and loss of productivity. Combined, IBM and Cisco solutions provide a holistic approach that allows customers to integrate and instrument their IT infrastructure.

The data center infrastructure management tools from Cisco and IBM complement each other. As the industry moves from device management to system management, the interaction between data center tools is critical to maintain real time visibility and health of the entire data center.

Typically, the Cisco tools (DCNM, VNMC, CSM, and so on) will provide information that the IBM Tivoli tools can leverage. The Cisco tools will provide the best and most granular interface to the network infrastructure, whereas the Tivoli tools offer a consolidated view of the infrastructure, and in many cases a “manager of manager” view.

The IBM Tivoli tools cover a broad technology set, including facilities, servers, applications, storage and mainframe. Combined with the deep visibility that the Cisco tools offer into the network and security, the IBM breadth ensures that a holistic end-to-end integrated management framework can be provided.

A consolidated management view provides accurate visibility of interdependencies, capacity and status. Cisco and IBM constantly strive to improve management and automation solutions for real time monitoring/reporting, asset/lifecycle management and maximum flexibility, uptime and availability.

Figure 10-4 shows the IBM and Cisco Data Center tooling architecture framework.

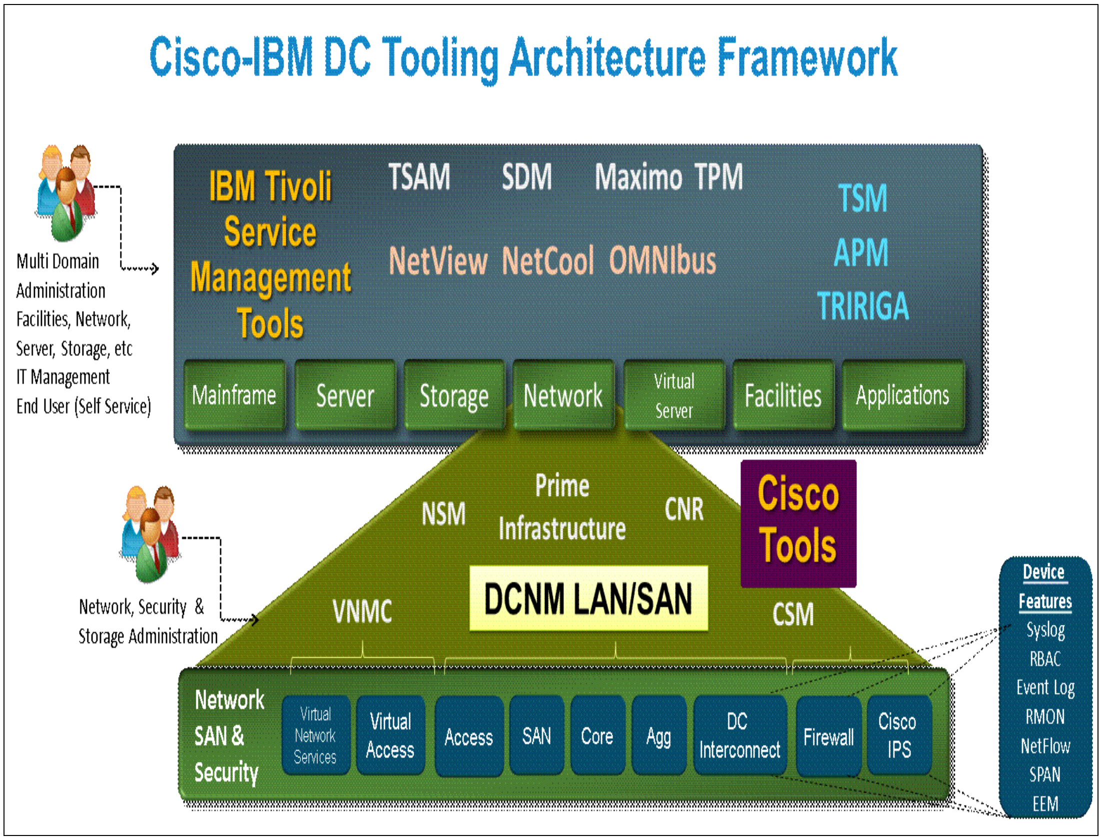


Figure 10-4 Cisco and IBM Data Center tooling architecture framework

10.11.1 Security

The IBM integrated security solution for Cisco networks is designed to offer lower risk and leverages our combined experience in building data networks and successfully managing large, complex systems integration programs. If a business wants to improve identity management and address security compliance on a Cisco network infrastructure, our solution establishes security access policies for devices connecting to the corporate network. It can also provide or deny admission based on operating system, patch or antivirus software levels or configuration settings.

The following solutions can be made available to meet respective requirements.

- IBM Tivoli Identity Manager and IBM Tivoli Access Manager
  - Cisco Secure Access Control Server to provide network admission control for authorized users
- IBM systems with pre-loaded Cisco Trust Agent to scan devices and communicate with Tivoli Security Compliance Manager
  - Cisco Secure Access Control Server to provide network admission control for authorized users
- IBM Tivoli Security Compliance Manager for policy creation, implementation, remediation and reporting
- IBM Global Services systems management consulting and design services with systems management services

Advantage: Address your security challenges

The IBM integrated security solution for Cisco networks helps enterprises address security breaches in two key areas.

- Managing the identities of users who connect to the enterprise
- Dealing with the “wellness” of a device connecting to the network
- Helping ensure the connecting devices are in compliance with defined security policies

Enterprises can have better control over user access by knowing the identity of the individual connecting to the corporate network. Protect the IT infrastructure and assets from unauthorized access and deny access to devices which could infect other parts of the business.

The IBM solution enables businesses to select the appropriate service offering based on the respective specific security objectives. This is accomplished in three steps.

- Verification of identities for network access
- Device security compliance checking for network access
- Automated remediation of non-compliant devices

Enterprises can choose to implement one, two or all of these functions.

The benefits: Keep critical information safe

The IBM integrated security solution for Cisco networks simplifies IT security management, helps in responding more quickly and accurately to reduce risk. Unlike other basic “scan and block” products, the solution autonomically quarantines and remedies noncompliant devices to help protect the business from security challenges. With IBM and Cisco Systems tools and expertise, the organization can potentially:

- Ensure security policies across the IT infrastructure by centralizing control and protecting local autonomy.
- Prevent the loss of valuable information and intellectual capital.
- Boost productivity by reducing errors and automating routine administration tasks.
- Grant access to cleared users, systems or devices that want to engage the business, while protecting organization from unauthorized access to the IT infrastructure.

10.11.2 Storage

Cisco offers a comprehensive portfolio of storage networking products that includes directors, fabric switches, IP storage routers and metro-optical networking solutions. When this is combined with IBM System Storage disk and tape storage, IBM System p, System x, and System z servers as well as planning, design and implementation services from IBM Global Services, it provides organizations with end-to-end storage networking solutions.

As an example, the Cisco MDS 9000 for IBM System Storage switches combined with IBM SAN Volume Controller (SVC) provides a solution that offers customers the ability to securely “virtualize” their storage, anywhere in the storage network. Whether organizations require cost-effective storage networking solutions for the enterprise data center, department/workgroup and small/medium-size business or high-performance business continuance implementations in the metropolitan-area network (MAN), IBM and Cisco meet their requirements with the following examples of intelligent, highly available storage networking solutions.

- Integrated storage networking intelligence delivers the scalability, performance, security and manageability required to progress from discrete SAN silos to interconnected SANs
- Cisco MDS 9000 for IBM System Storage switches qualified with the IBM System Storage disk and tape and IBM System p, System x, and System z servers
- IBM SAN Volume Controller (SVC) and the Cisco MDS 9000 providing network-based virtualization, point-in-time copy, and replication
- Planning, design, and implementation services from IBM Global Services and IBM Business Partners

10.11.3 Networking

The long-standing strategic partnership between IBM and Cisco provides the basis for a leadership role in this changing environment. This relationship enables both companies to have a real and lasting impact by co-creating the next generation of solutions for data center, communication, and collaboration. IBM and Cisco are jointly providing an end-to-end data center solution by offering specialized computing, storage, and networking technology portfolios for the data center. Joint solutions feature Cisco products integrated into, or working with, IBM products.

Additionally, with IBM's integration of networking back into its portfolio, many solutions are also possible using IBM System Networking switches connected into Cisco enterprise networks.

Cisco’s Data Center Networking products are tightly integrated into the IBM data center designs at all levels. Cisco submits itself to strenuous Open and Mainframe systems testing, provided by IBM, to meet the rigorous demands of our clients needs taking into account their current requirements and infrastructure. Cisco and IBM are jointly leading the industry with the emergence of new technologies and solutions in the data center. IBM and Cisco offer a complete and integrated solution for a virtualized data center and delivers the following unparalleled value to organizations.

- IBM has vast industry-specific experience in business process and transformation, technology leadership, and world class design and implementation services
- Cisco’s highly acclaimed hardware and software has established the company as the world leader in networking for the Internet
- Cisco MDS 9000 for IBM System Storage switches qualified with the IBM System Storage disk and tape and IBM System p, System x, and System z servers
- IBM and Cisco data center solutions enable communication across systems for faster deployment of applications
- IBM's networking hardware is built on open standards with proven interoperability and joint validation testing into Cisco networks
- IBM and Cisco jointly develop, integrate, and deliver productive and cost-effective data center networking solutions to our clients

10.12 Comprehensive data center management

The next generation data centers have to be immensely more efficient to meet new requirements while keeping costs in check as the demand for and price of resources continue to rise. However, the realization of this efficiency requires a deep and pervasive transformation in how data centers are designed, managed, operated, populated, and billed, mandating a unified and coordinated effort across organizational and functional boundaries toward a common set of goals. In this section, we briefly introduce you to some of the important considerations when you plan your next generation data center management strategy.

10.12.1 Fragmented data availability

Knowledge of a data center has to start with its contents. However, this information is often spread across organizations and lacks a detailed level of granularity that would include asset ownership, operational purpose, physical asset location down to rack slot placement, and server inventories to a level of granularity that includes virtualized servers. Even defined, this information is difficult to maintain with an acceptable degree of accuracy.

Availability of monitoring data can also vary considerably regarding what operational characteristics of assets are being monitored and trended over time using a myriad of varying tools. Data on the environmental conditions of the facility can be fragmented as well. This fragmentation of both asset inventory and monitoring data makes it impossible to obtain a unified picture of a data center's current contents, asset placement, ownership, and operational conditions that integrates both IT and facilities. Without this picture, effective decisions and efficiency analytics based on actual data become impossible as well. As is often said, you cannot manage what you do not measure.

Tip: You can take advantage of IBM tools such as IBM Maximo® Asset Management (MAM) to track asset characteristics, process associated work orders, and place assets on the data center floor. And, Tivoli Application Dependency Discovery Manager delivers automated discovery and configuration tracking capabilities to build application maps that provide real-time visibility into application complexity.

10.12.2 Defining your efficiency metrics

A standardized data architecture includes managed capacity providers and consumers. Such an architecture facilitates the creation of related metrics, reports, and role-based dashboards that can be used to assess efficiency of use, determine overall costs, and perform consumption-based billing. A standardized data architecture also enables comparisons across capacity types, different physical sites, and so on. Differentiating between facility and IT capacity consumers enables the determination and tracking of industry-standard efficiency metrics such as Power Usage Effectiveness (PUE) or its inverse metric, Data Center Infrastructure Efficiency (DCiE), both of which consider the ratio of facility versus IT power consumption. You might also consider additional efficiency and sustainability metrics introduced by industry consortium The Green Grid[4], such as Carbon Usage Effectiveness (CUE), Water Usage Effectiveness (WUE), Energy Reuse Effectiveness (ERE), and Data Center Compute Efficiency (DCcE).

[4] http://www.thegreengrid.org/

Associating consumption measurements with logical asset characteristics enables reporting and analysis beyond just efficiency, including cost recovery based on associated services provided by the consuming assets, viewing consumption based on asset ownership, and so on. Taking advantage of the standard data architecture, facility level measurements can be rolled up to gain regional or business-wide perspectives on capacity provision, consumption, and efficiency.

10.12.3 Assigning efficiency architects

IT solution architects often have little insight into or, for that matter, concern about the operations of the facility providing the hosting for the servers, storage, and networking deployed as part of their deliverables. As long as there is adequate space, power, cooling, and network bandwidth to operate the solution effectively, the intricate details of the provisioning of that capacity is outside of their scope. Conversely, facility architects might have limited knowledge of the purpose of the servers being deployed within their hosting environments.

As long as their facility can accommodate the spatial, power, cooling, and networking requirements of the equipment to be deployed, what logical solution is operating on those servers is beyond their purview. However, these singular deployment scopes can miss considerable efficiency gains. Utilization knowledge across capacity types at the facility level, coupled with awareness of server operational purpose at the IT level, enables placement of hardware within the data center at the most efficient possible location, considering both facility capacity and the grouping of like-purposed equipment.

Tip: IBM Tivoli Business Service Manager allows business and operations teams to manage the complex relationships between services and the supporting technology infrastructure

10.12.4 Expanding the scope of capacity management

In order to gain large efficiency gains, we must expand our treatment of capacity, both in breadth and in depth. At the macro level, as increasingly powerful servers occupy increasingly smaller physical footprints and generate larger heat signatures, data centers can easily run out of available cooling or power long before physical space is exhausted. Consequently, it becomes imperative to track and manage multiple capacity types concurrently, constantly striving for a balance in consumption across them. Exhausting any one type of capacity leaves other capacities stranded, limiting the growth, increasing the operational costs for the work performed within the environment, and resulting in lost opportunities for additional business without incurring additional construction costs.

Environmental concerns and governmental regulations often necessitate the inclusion of additional capacities, such as water usage and carbon emissions, and larger data transmission demands can require the inclusion of network bandwidth capacity as well. While the breadth of capacities to manage increases, it deepens as well. Individual servers can be as little as 20% utilized, while data storage demands continue to escalate. Effective utilization of consolidation, virtualization, and cloud computing demands the awareness and management of asset level capacities like CPU and storage. As a result, a comprehensive view of data center efficiency necessarily spans facilities and IT.

Tip: IBM Tivoli Monitoring is the tool to monitor the status of the IT environment. It allows you to monitor the operating systems, databases, and servers throughout distributed environments with a flexible, customizable portal.

10.12.5 Integrating across IT and facilities for the bigger picture

To achieve the greatest degree of efficiency from the data center environment, IT staff must have a comprehensive perspective that spans both the facilities and IT equipment that are present on the data center floor. Such a perspective starts with the data center inventory and the operational lifecycle that keeps it up-to-date. That inventory is supplemented with instrumented metering that gathers data about the ongoing operational conditions of the data center. That data can then be consolidated, enabling monitoring across all aspects of data center operations for potential problem conditions. That consolidated data set can also provide the foundation for detailed analytics, role-based dashboards, historical trend reporting, and future planning.

For example, IBM Tivoli Maximo for asset management is used for asset life cycle management and IBM Tivoli for monitoring (ITM) to get to know the asset utilization and alerts for both IT and Facilities. Further integration with IBM products like, NetCool OMNIbus and Tivoli Business Service Manager organizations can not only manage the IT Services at common platform such as ITIL, they will also be leveraging IBM analytics for a healthy strategy formulation for business success.

10.12.6 Optimizing based on analytics

As IT personnel gains a good handle on the data center inventory and the ability to gather data about the facility’s operational conditions, then IT staff can take the advantage of detailed analytics to determine the best options for improving the overall efficiency.

Tip: IBM and Cisco offer a broad, integrated portfolio of information and analytics capabilities, spanning software, hardware and services. These capabilities can be deployed on a solid foundation with an enterprise-class Big Data platform to take advantage of all of the data sources.

Figure 10-5 depicts a reference architecture for comprehensive data center management.

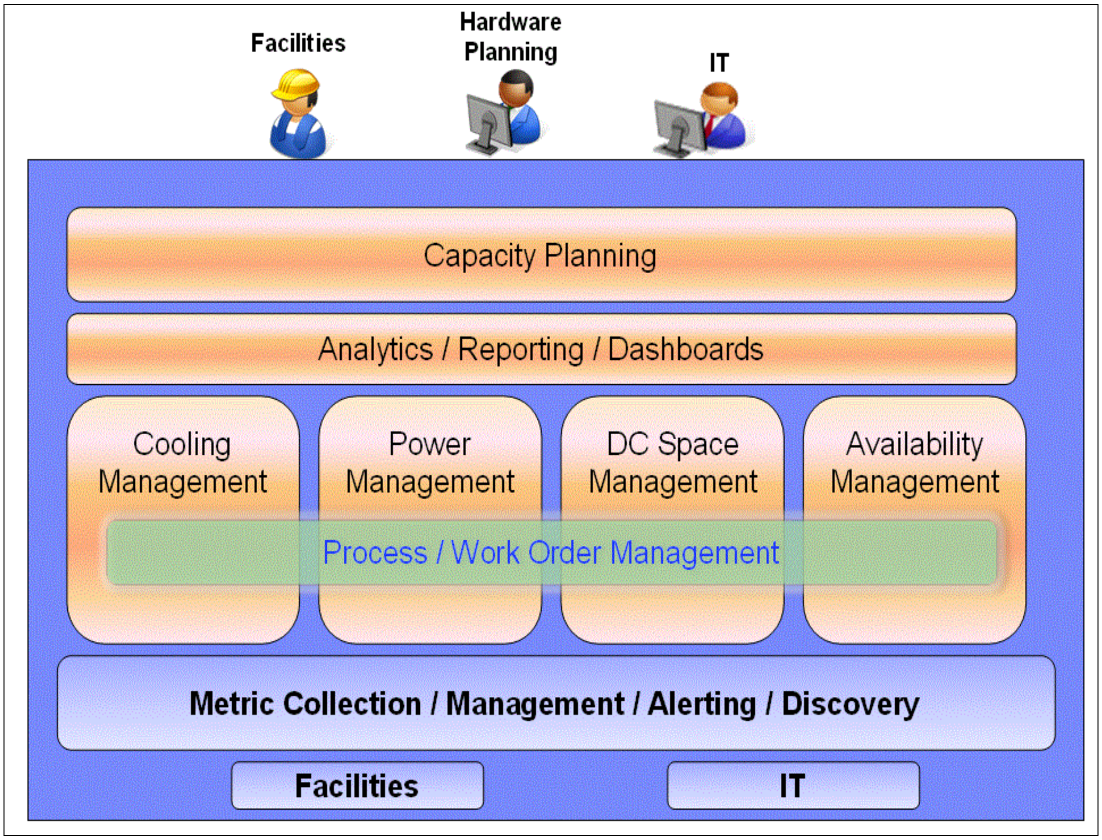


Figure 10-5 A reference architecture for comprehensive data center management

10.12.7 Using knowledge for future planning

As IT organizations assess how available capacities are consumed and determine operational and environmental pain points, they can plan better for future capacity needs because they will know which types of capacity will be needed where and when.

10.12.8 Effective integration into daily business

One-time assessments and the performance of associated one-time actions might yield several measurable benefits, but they will likely be short-term because the same business processes that resulted in the inefficiency continue to be used. The necessary lifestyle changes come when we integrate a detailed efficiency focus with the processes that govern the daily operations.

10.13 Data center automation

In order to improve service agility, maintain data center stability and drive standards, there has been a massive trend towards automation in the data center.

Cloud computing has been one of the greatest catalysts for this trend. According to the National Institute of Standards and Technology (NIST) definition of cloud computing, it introduces the following characteristics; On-demand self-service, Broad network access, Resource pooling, Rapid elasticity and Measured service.

Note: The NIST definition of cloud computing is available here:

`http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf`

Here are some other challenges that are addressed by the cloud model:

- Lack of agility
- High cost of IT staff
- Business-IT dissonance
- Low capacity utilization
- High operating costs
- Overcrowding of data center
- Labor-intensive, manual processes for service management
- High error rates due to disconnected processes
- Infrastructure sprawl

As a consequence, system and management vendors have been working diligently to deliver the infrastructure and tools to bring cloud computing to the mainstream data center.

Additionally, the mass adoption of data center virtualization has opened up a new set of capabilities for service automation and “lights out management”. Another recent technology enabler is the rapid adoption and maturity of XML APIs and REST based interfaces. Vendors such as Cisco and IBM have introduced new capabilities to optimize and leverage the “wire once” infrastructure.

As we move forward, new tools will be introduced to the data center that treat the network as a “Big Data” source, and can collect massive amounts of information from the infrastructure, parse through analytics engines and make intelligent decisions in real time to reconfigure and optimize the data center infrastructure. This will only be possible if the network is ready to be automated.

Both IBM and Cisco has embraced to notion of network automation with a combination of APIs to drive the infrastructure and software capability and features. Third party upstream management tools will be able to take advantage of these published APIs and tools by “driving” the network infrastructure through their automation and orchestration capability.

Network Administrators and data center managers will benefit from this new environment by reducing the time for service enablement by orchestrating complex tasks across multiple platforms concurrently. Network service provisioning can be intricate as it requires coordination of multiple devices and significant configuration to ensure a secure and stable application environment.

10.13.1 Cisco Network Services Manager

Cisco Network Services Manager (NSM) is an example of this new capability to accelerate service agility for network provisioning. Cisco Network Services Manager is designed to help enable customers to organize their network resources into a flexible multi-tenant infrastructure that integrates the network with their existing IT operational tools and processes. Network Services Manager's network abstraction layer allows customers to automatically provision a set of network features into an end-to-end topology, or “network container,” much more easily and quickly than previously possible with template - and script-based systems, dramatically reducing network operations costs and the potential for misconfiguration while optimizing capacity utilization and accelerating service delivery.

Figure 10-6 shows the Cisco NSM architecture.

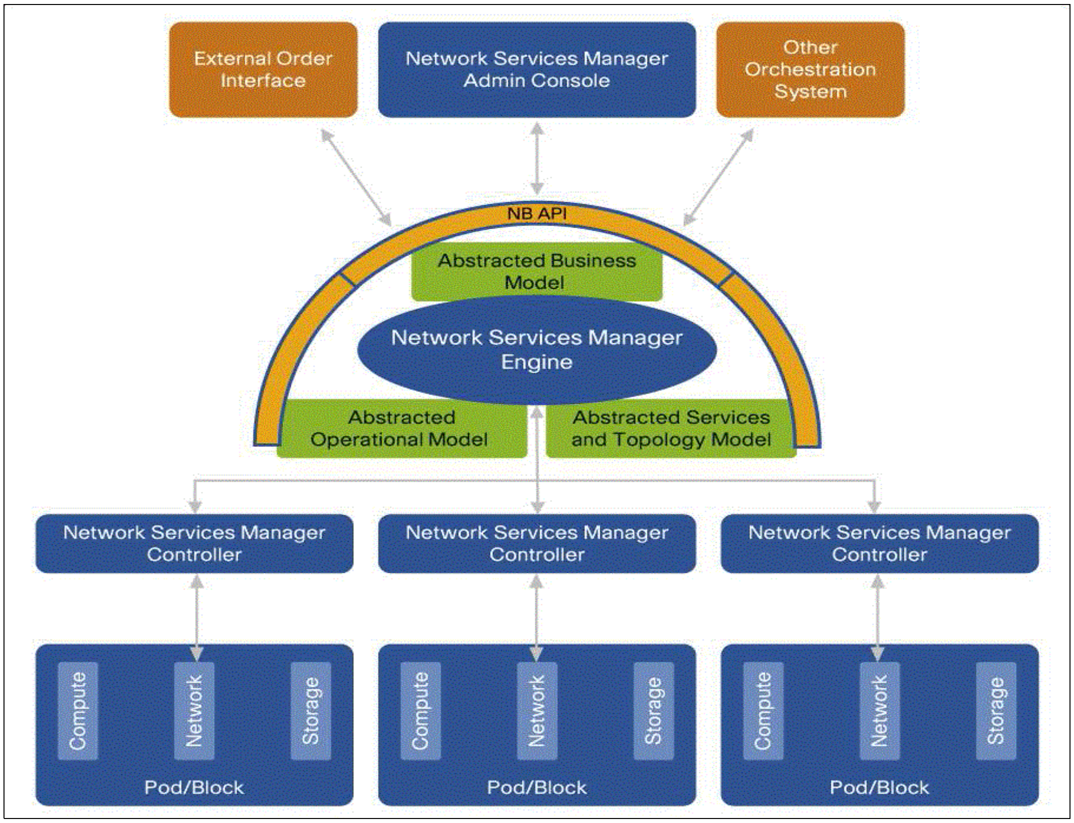

External Order Interface
Network Services Manager Admin Console
Other Orchestration System
NB API
Abstracted Business Model
Network Services Manager Engine
Abstracted Operational Model
Abstracted Services and Topology Model
Network Services Manager Controller
Network Services Manager Controller
Network Services Manager Controller
Compute
Network
Storage
Pod/Block
Compute
Network
Storage
Pod/Block
Compute
Network
Storage
Pod/Block

Figure 10-6 CISCO Network Service Manager

NSM can be used to reduce provisioning time and human error by orchestrating the synchronisation of routers, switches, firewalls, load balancers, and WAN optimization platforms.

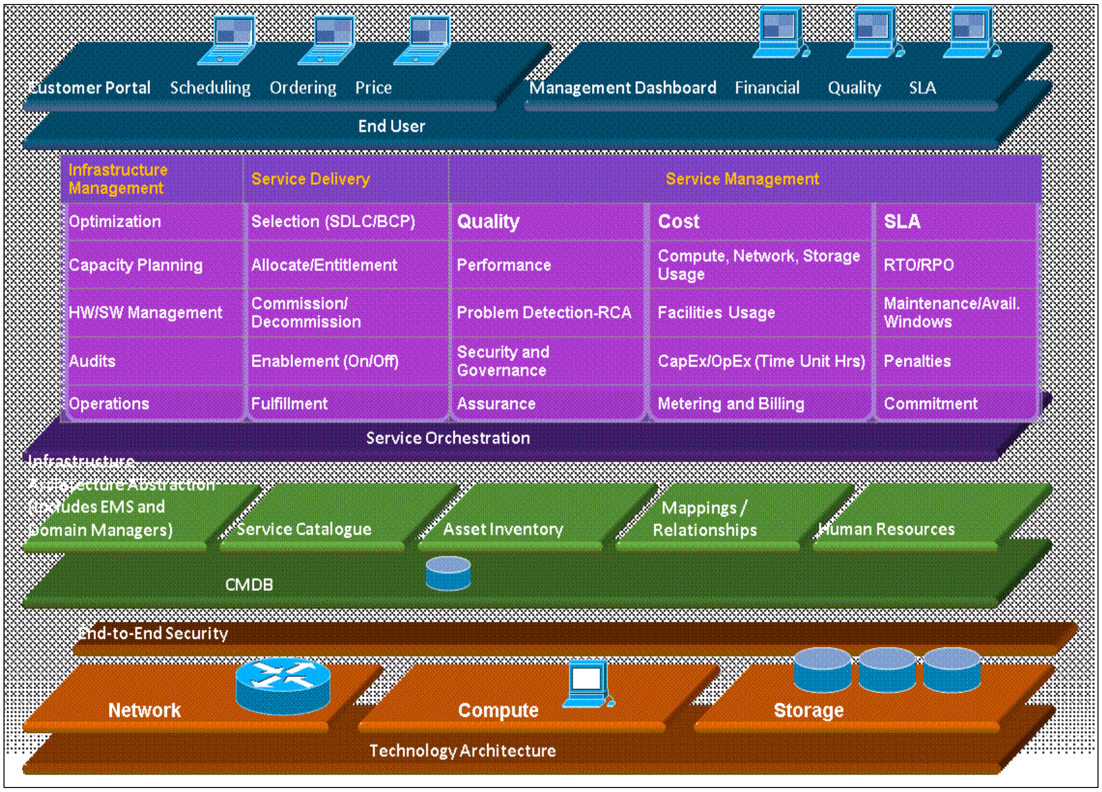
10.13.2 Automated reporting

Data center automation is not just about provisioning. One of the easiest and most effective ways to introduce automation is to start with automated reporting. Reports can be made into a template and scheduled so that data center SLAs can be monitored and scheduled with self service portals so that end users or individual tenants can access their own SLA reports.

Cloud management framework

A complete cloud management framework includes much more than automation. Depending upon the cloud use case and customer maturity, it should also include capabilities such as billing/penalties, governance, capacity planning, self service portal, service catalog, and so on.

Figure 10-7 shows an example of a cloud reference architecture framework.



Infrastructure Management	Service Delivery	Service Management		
Optimization	Selection (SDLC/BCP)	Quality	Cost	SLA
Capacity Planning	Allocate/Entitlement	Performance	Compute, Network, Storage Usage	RTO/RPO
HW/SW Management	Commission/ Decommission	Problem Detection-RCA	Facilities Usage	Maintenance/Avail. Windows
Audits	Enablement (On/Off)	Security and Governance	CapEx/OpEx (Time Unit Hrs)	Penalties
Operations	Fulfillment	Assurance	Metering and Billing	Commitment

Figure 10-7 Reference example for cloud architecture framework

In a fully automated environment as seen at a mature cloud service provider, there will be a software stack that performs the following functions:

- Self service portal for service requests: Can be by IT administrators or end consumers
- Service catalog: Contains the standard build environments, typically including server, storage, OS, network, and security attributes.
- Configuration Management Database (CMDB): To track and manage resources and service lifecycle.
- Approvals system: To control policy and governance
- Orchestration engine: To coordinate the configuration and deployment across multiple devices
- Reporting tool: For capacity management, consumption modeling, and perhaps chargeback/billing capability

As can be seen from the cloud framework diagram, there are likely to be many other tools in the complete stack.

However, it should be noted that data center automation does not need to implement a complete stack. There are many ways to introduce automation to the data center in a very graceful manner, allowing gradual change to strategy and operations teams. Cisco and IBM have professional services that can assist customers with IT Consultancy Services to help our customers on the automation journey.

Another nascent approach to cloud automation is the growing use of open source tools, software stacks such as those proposed by the OpenStack community.

The Openstack mission is “to produce the ubiquitous open source cloud computing platform that will meet the needs of public and private cloud providers regardless of size, by being simple to implement and massively scalable.” However, in its current state of development (November 2012) it is limited in functionality and is unsuitable for diverse, complex enterprise environments without significant commitment of resources for development and integration. It is used today in a few specific cloud deployments. The promise of OpenStack is exciting and both Cisco and IBM are active contributors to this valuable initiative.

11

Software Defined Networking and Cisco Open Network Environment

In this chapter, we introduce the Software Defined Networking paradigm.

Traditionally, the network architects and engineers have struggled to offer a consistent, easy to understand and leverage interface to the network that could be used by application developers and IT service owners, who often feel that the network is way too complex and rigid for rapid infrastructure deployment, quick and reliable IT infrastructure changes. The Software Defined Networking (SDN) approach is aimed at overcoming those limitations so that the network can become flexible, responsive and consistent through easy to use, reliable and standard programmable interfaces.

We explore the following topics:

- We first introduce the main SDN concepts, its definition and its objectives. Also, we introduce a framework showing the areas of collaboration between IBM and Cisco.
- At the end of the chapter, we introduce the Cisco Open Network Environment and provide additional details on its objectives and evolution, discussing the Cisco One Platform Kit, Controller and Agent model and Virtual Network overlays.

11.1 Software Defined Networking overview

In this section, we start by introducing the Software Defined Networking (SDN) paradigm, describing its objectives and providing a high level definition.

OpenDaylight Project: Since the writing of this book, the OpenDaylight project has been announced.

SAN FRANCISCO, April 8, 2013: The Linux Foundation, the nonprofit organization dedicated to open source development and technologies, today announced the founding of the OpenDaylight Project, a community-led and industry-supported open source framework that will accelerate adoption, foster new innovation and create a more open and transparent approach to Software-Defined Networking (SDN).

Why OpenDaylight? The adoption of new technologies and pursuit of programmable networks has the potential to significantly improve levels of functionality, flexibility and adaptability of mainstream data center architectures. To leverage this abstraction to its fullest requires the network to adapt and evolve to a Software-Defined architecture. One of the architectural elements required to achieve this goal is a Software-Defined-Networking (SDN) platform that enables network control and programmability.

IBM and Cisco are Platinum members of this project. For more information about the mission, see this website:

`http://www.opendaylight.org/`

11.1.1 SDN objectives and definition

Software Defined Networking aims at enhancing the programmability of the networking infrastructure so it can be easily consumed by application and IT services developers across different networking areas, vendors and implementations without requiring custom software development and usage of scripting languages.

While networks have supported significant innovations in compute and storage, networking technologies have not tracked the expanding levels of virtualization and programmability of these technologies. Networks thus are increasing in complexity due to the increased demand for multi-tenancy, higher bandwidths, and the emergence of on-demand resource and application models in the Cloud. Software-Defined Networking (SDN) tries to address these challenges by altering the traditional paradigm for network control.

By decoupling the control and data planes, introducing programmability, centralizing intelligence, and abstracting applications from the underlying network infrastructure, highly scalable and flexible networks can be designed that readily adapt to changing business needs.In order to achieve this level of virtualization and programmability, though, the network has to overcome traditional barriers and evolve so that the whole network can be exposed as an easy to use, programmable interface.

With increased network programmability more granular network control (at the session, user, or device level), and improved automation and management can be achieved.

Integral to the SDN value proposition is not just transforming the way networks are designed and operated but also ensuring a tight coupling between the network and the rest of the IT infrastructure (Servers and Storage), IT Service Management and Applications. This way there can be a closer synchronization between the network, rest of IT and the business, turning the network into an enabler instead of being perceived as a show-stopper: this is the high level goal of the Software Defined Networking paradigm.

It is difficult to define Software Defined Networking (SDN) in an agreed upon fashion among all the involved actors as different industry players will have different viewpoints and approaches to the subject, but from a high level point of view, SDN can be defined as a set of tools, protocols, and technologies.

These SDN tools enable the following concepts:

- Separation of the network data plane (packet forwarding) and the control plane (logic, network state and policies) so that the latter can be centralized, usually as software running on a server.
- Abstraction of network topology and functions so these can be consumed by and orchestrated together with servers, storage and applications.
- The network to become more programmable, via open APIs - Application Programming Interface.

Another important benefit of the SDN architecture is enhanced automation, allowing networks to accommodate highly elastic and rapidly changing demands of users or cloud-based applications. Cloud-based applications can now be managed through intelligent orchestration and provisioning systems, beyond the compute and storage space and including the network.

SDN can be used for many purposes, including simplifying network control and management, automating network virtualization services, and providing a platform from which to build agile network services. To accomplish these goals, SDN leverages IETF network virtualization overlays and the ONF OpenFlow standards, but it is not limited to them. SDN is an evolving concept and the technologies and standards enabling SND are evolving rapidly.

11.1.2 Areas of collaboration between IBM and Cisco

As introduced in the previous section, SDN is not just a new networking technology, but it is promising a radical shift in terms of the degree of integration between networks and the rest of IT infrastructure and applications.

Following this definition, we can try to expand it further by defining the common IBM and Cisco building blocks to enable this transformation: a possible framework to simplify the explanation of these building blocks is provided in Figure 11-1.

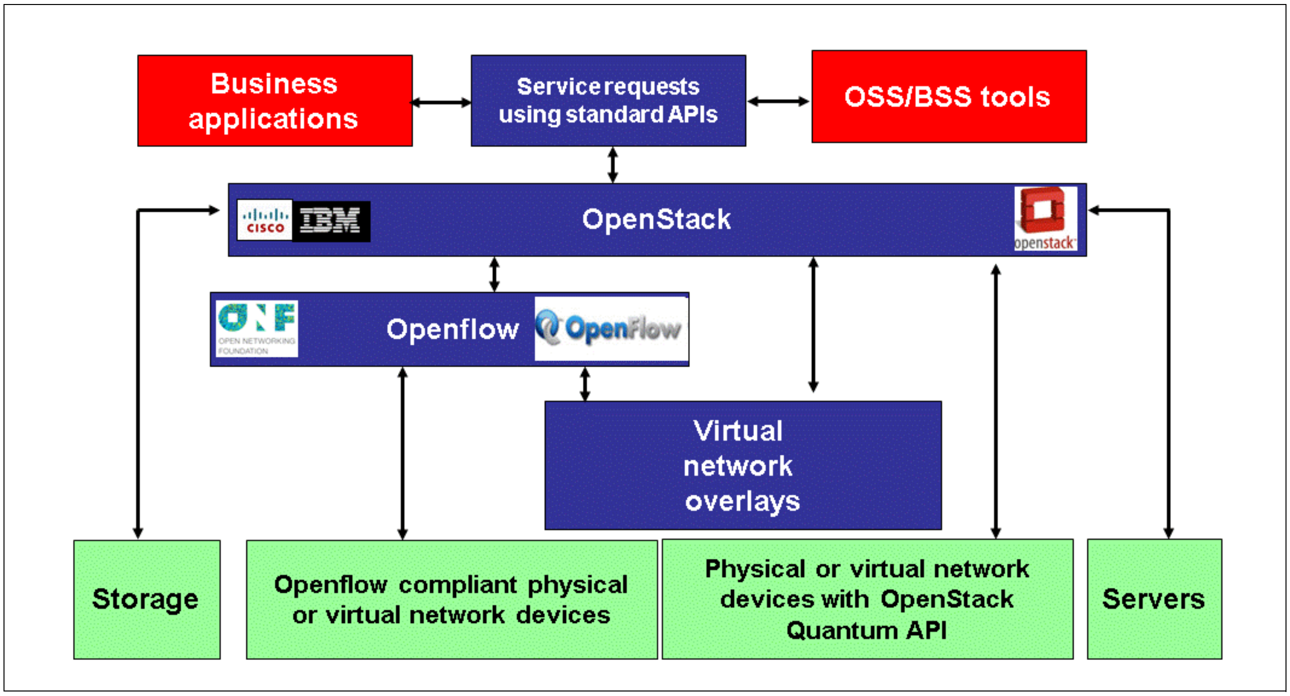

Business applications
Service requests using standard APIs
OSS/BSS tools
OpenStack
cisco
IBM
openstack
Openflow
ONF
OPEN NETWORKING FOUNDATION
OpenFlow
Virtual network overlays
Storage
Openflow compliant physical or virtual network devices
Physical or virtual network devices with OpenStack Quantum API
Servers

Figure 11-1 Software defined networking framework

- The red blocks at the top of Figure 11-1 represent the systems that consume the SDN standard network and data center interface: mainly business applications (Front-end) and Operations Support Systems (also called operational support systems or OSS) and Business support systems (BSS) Tools (Back-end) as both could greatly benefit from having an easy to program network. An example of an OSS application is a Trouble Ticketing System while an example of a BSS application is a Billing System. The green blocks at the bottom of Figure 1-1 represent the infrastructure elements that the SDN stack executes upon, both in a direct way (Switches, Routers, Firewalls, Application Delivery Controllers for example) and in an indirect way (Servers and Storage end points).
- The blue building blocks are layers of software that are in scope for Software Defined Networking. From bottom to top:
  - Virtual network overlays implement some networking functions via software so they make SDN approaches easier than in case of traditional tightly coupled software-hardware implementations of networking functions.
  - OpenFlow is a protocol standardized by the Open Network Foundation (ONF[1]) that allows to decouple forwarding and control planes. It includes two main components: OpenFlow compliant network devices (virtual or physical) and a controller which centralizes the control plane functions and can program the forwarding plane of the OpenFlow compliant network devices via a secure channel. Both Cisco and IBM have announced OpenFlow support and product development initiatives and are committed to the evolution of the standard as members of the ONF.
  - OpenStack[2] is a platform for developing and deploying cloud infrastructures encompassing servers, storage, and networks. Both IBM and Cisco are part of the OpenStack ecosystem. Moreover, IBM and Cisco both have representatives in the Board of Directors of the OpenStack Foundation.

[1] For more information on the ONF and OpenFlow visit https://www.opennetworking.org/

[2] For more information on OpenStack visit http://www.openstack.org/

- As also indicated on Figure 11-1, cloud infrastructures that utilize OpenStack do not require the use of OpenFlow to program the network devices. The OpenStack automation software can use the OpenStack Quantum (Network) API to directly program the network devices.
- Standard APIs are necessary to enable the consumption of the IT infrastructure resources (including the network) by both front-end and back-end applications. Standard APIs can be consumed both at the OpenStack layer and at the OpenFlow layer. An application programming interface (API) is a specification intended to be used as an interface by software components to communicate with each other and makes abstractions available to programmers.

11.2 Cisco Open Network Environment

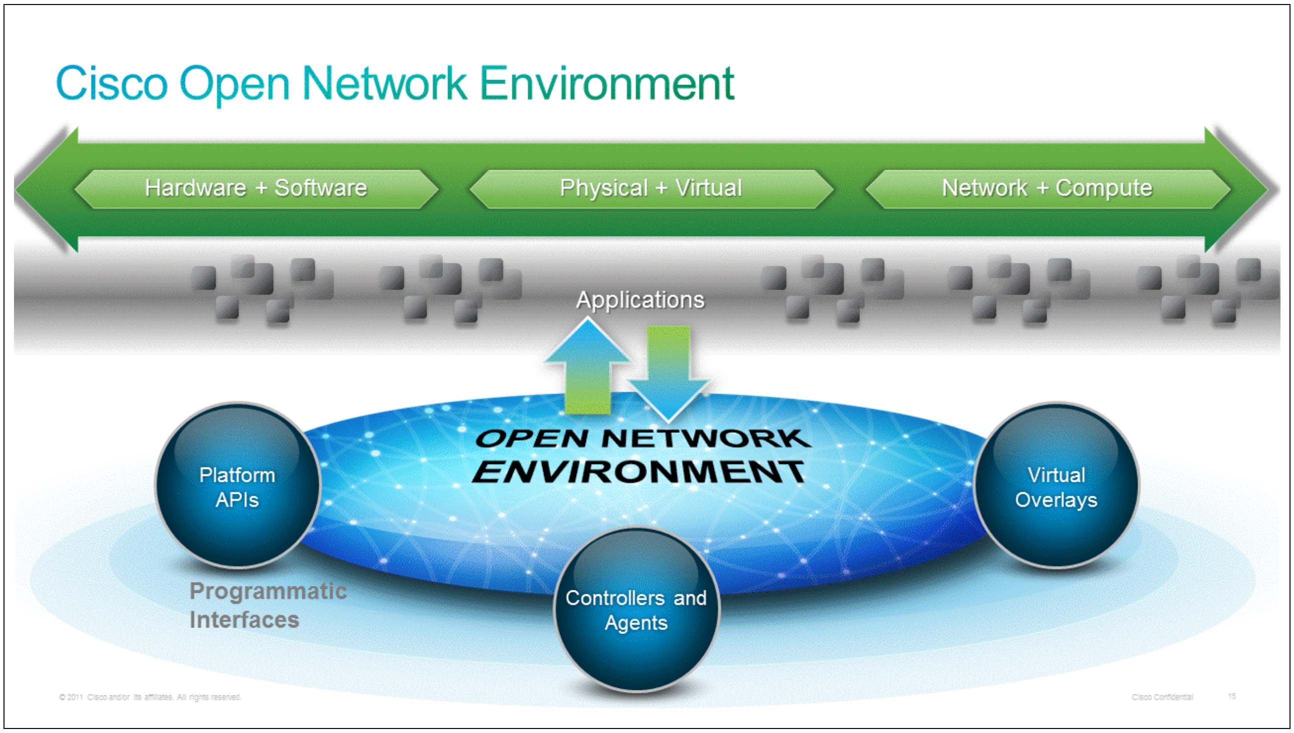
This section focuses on describing Cisco Open Network Environment or Cisco ONE. As mentioned previously, there are several definitions in the industry today for Software-Defined Networking, one common mistake of some of those definition is to use the OpenFlow protocol as a synonym for Software-Defined Networking.

Software-Defined Networking is more than the OpenFlow protocol. The OpenFlow Protocol is an element of an SDN architecture, it is the protocol used for communication between the control (controller) and data plane of network devices that support OpenFlow but Software-Defined Networking is broader than that.

Cisco ONE complements current approaches to software-defined networking while expanding to the entire solution stack, from transport to management and orchestration.

Cisco ONE, via its three pillars, has the objective of creating a multifunction, continuous feedback loops that deliver network intelligence and optimization across infrastructure and applications; for example, between analytics engines, policy engines, and network infrastructure. By replacing a monolithic view of the network and opaque application requirements with a multidimensional model that spans from the application to the ASIC, it provides a deeper and richer interaction between applications and the network.

The next sections will focus on defining the three pillars of Cisco ONE: Programmatic Interfaces, Controllers and Agents, and Virtual Network Overlays.



Cisco Open Network Environment

Hardware + Software

Physical + Virtual

Network + Compute

Applications

OPEN NETWORK ENVIRONMENT

Platform APIs

Virtual Overlays

Controllers and Agents

Programmatic Interfaces

© 2011 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential 15

Figure 11-2 Cisco Open Network Environment

As illustrated in Figure 11-2, Software-Defined Networking concepts are a component of the Open Network Environment and the OpenFlow protocol is also part of the ONE but it is not the only option to link agents and controllers.

11.2.1 Programmatic Interfaces: Cisco One Platform Kit (onePK)

The Cisco One Platform Kit (onePK) provides a set of APIs for programmatic access to Cisco operating systems and hardware platforms, including IOS, IOS-XR, and NX-OS operating systems. It will be supported on ASR, ISR G2, CRS, Catalyst and Nexus switches for example.

The onePK can be used to more easily reach into the network and extract the information needed and also can be used to directly manipulate flows and modify packets. This allows the creation of customized services and optimizes routine operational tasks with improved automation.

The challenge that is being addressed with the onePK is that today's networks are provisioned, monitored, and maintained by a large toolkit of protocols and techniques, from command-line interface (CLI) to Extensible Markup Language (XML), Simple Network Management Protocol (SNMP) to syslog (see Figure 11-3). While powerful and effective for their prescribed tasks, these tools are not programmatic, thereby preventing most application programmers from more effectively interacting with the network.

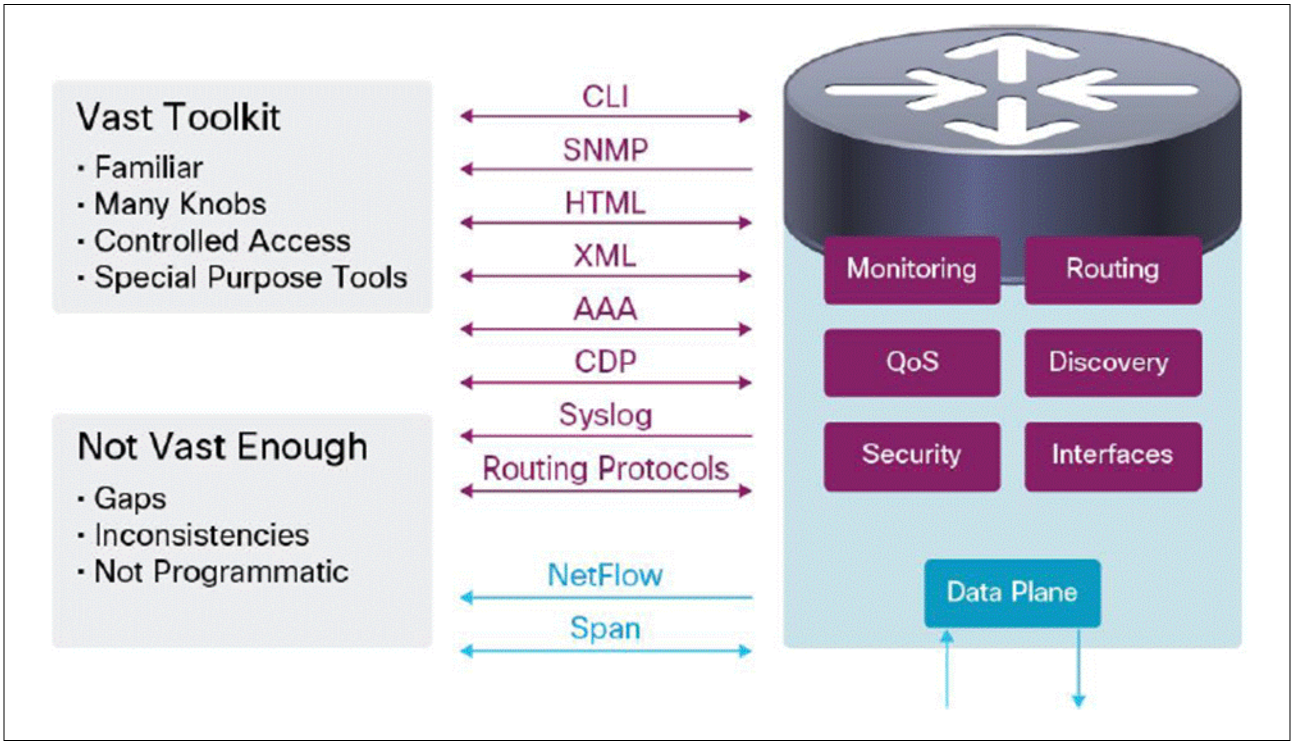


Figure 11-3 How we interact with network software today

An application programming interface (API), which is a specification intended to be used as an interface by software components to communicate with each other, makes abstractions available to programmers. In the case of the Cisco onePK is will enable direct, programmatic access to the network and will provide a consistent interface to all Cisco routers and switches.

The One Platform Kit will enable programmers to write applications that directly access the routers and switches through the onePK API, as shown in Figure 11-4.

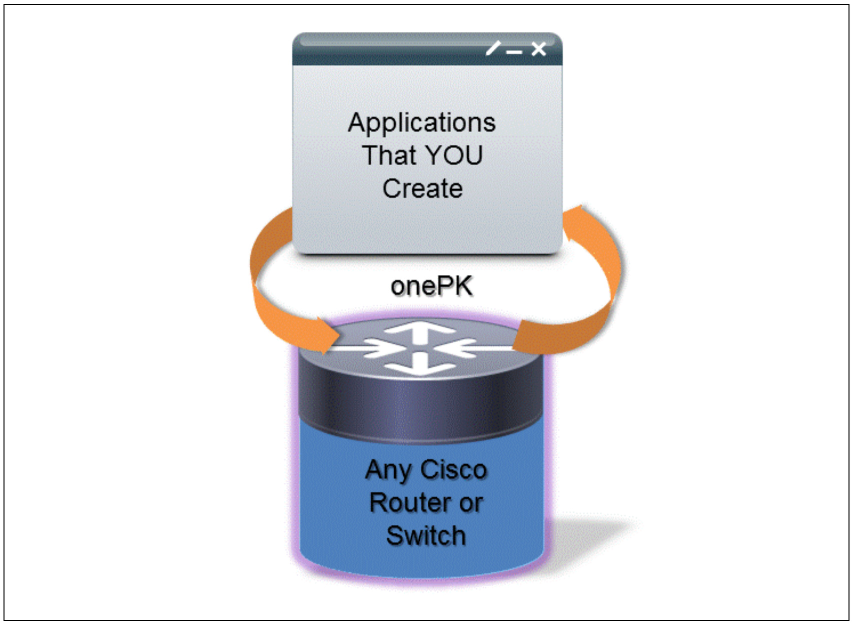
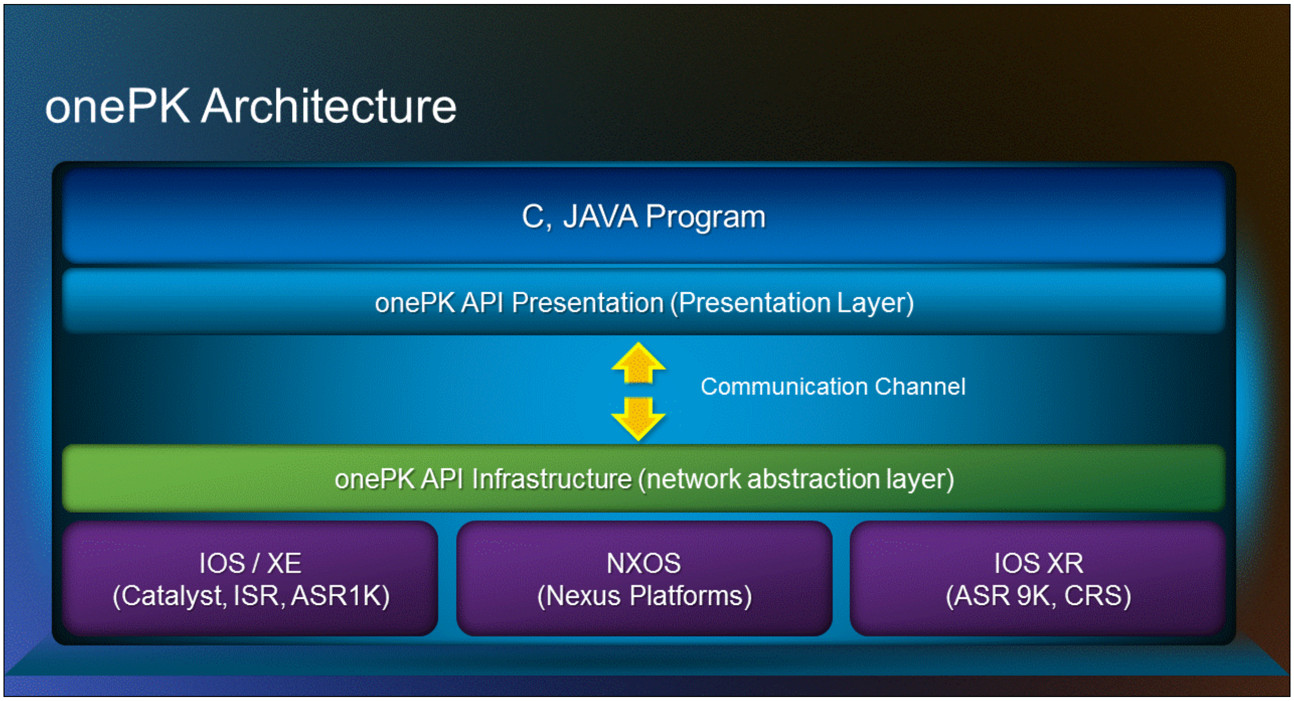


Figure 11-4 Applications access the Routers and Switches via the onePK API

The onePK architecture is composed of three major elements: the network abstraction layer, the communication channel, and the presentation layer.

- The network abstraction layer provides access to functions that are internal to a router or switch. One of the primary values of this layer is that it abstracts underlying differences between operating systems and platforms. For example, if your application uses a onePK function call to read interface statistics, that same function call will work across all Cisco networking software platforms (IOS, IOS XR, IOS XE, and Cisco NX-OS Software).
- The presentation layer consists of the API libraries that programmers can use within their applications. There are currently many separate APIs available for separate network tasks. With onePK, application programmers get a universal network programming toolkit. The onePK libraries are initially available in C and Java. The architecture is extensible, so additional language bindings (such as Python) can be added. The presentation library was designed with very few dependencies, so it can be easily integrated with existing tooling and development requirements.
- The communication channel provides a fast, safe, extensible channel between the application and the network element. Applications are required to authenticate before being allowed to access the functions of the network abstraction layer.

Figure 11-5 shows the onePK architecture, its three major elements, the operating systems where it will be supported as well as the programmable languages initially supported by onePK.


onePK Architecture

C, JAVA Program

onePK API Presentation (Presentation Layer)

Communication Channel

onePK API Infrastructure (network abstraction layer)

IOS / XE
(Catalyst, ISR, ASR1K)

NXOS
(Nexus Platforms)

IOS XR
(ASR 9K, CRS)

Figure 11-5 onePK Architecture

Cisco One Platform Kit (onePK) use case

The Cisco onePK will enable a wide range of use cases. Following is a simple example; several other uses cases exist.

On this example the network operator needs to direct (route) traffic using unique or external decision criteria; for example, route long lived elephant flows, backup traffic using the lowest $ (dollar) cost path, or have trading information follow the path with the lowest delay.

The traditional routing protocols do not consider those unique and variable business logics when calculating the route (path) that packets are directed so they are not a viable option, and to program routes manually is not scalable and dynamic enough.

In this case a custom routing application augments the network with a custom routing policy. This policy redirects traffic using the customer's own business logic. Those business logics will vary between customers. The custom routing application is built and deployed using onePK, communicating directly with the forwarding plane. The unique data forwarding algorithm is highly optimized for the network operator's application.

The application can reside on a central server and communicate with multiple devices using the onePK API.

Figure 11-6 shows how an external application could program the network using the onePK API to customize the traffic flow based on business logic.

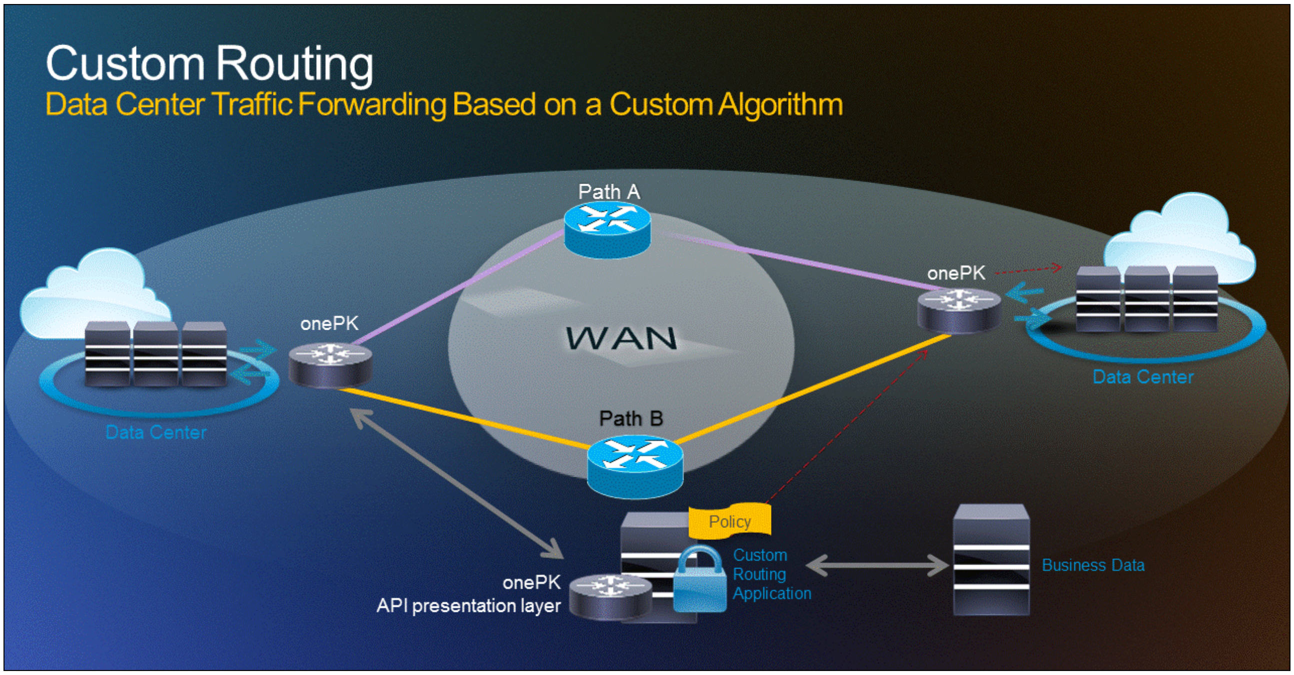


Figure 11-6 onePK used to achieve a custom routing policy

Note: For more details on the Cisco onePK, go to http://www.cisco.com/go/onepk

11.2.2 Controllers and Agents

Some functionalities delivered by the network benefit from a logically centralized coordination across multiple network devices. The agent and controller model consolidate state from multiple network elements into a centralized controller. The centralized controller function referenced here can be a single instance or can be implemented in a hierarchical model.

A Controller is a process or application on a device, a server for example, interacting with a set of other devices using APIs or protocols.

An Agent is a process or library on a device that understands requests from the controller and then leverages device APIs to deliver a task or domain specific function.

The Controller and Agent model has been used in networking for quite some time however those Agent-Controller pairs have always served a specific task (or set of tasks) in a specific domain. One of many examples of a Controller and Agent that serve a specific task is the Cisco Wireless LAN Controller (WLC), the WLC centrally manage, secure, and configure access points so serves a set of tasks in the wireless domain.

In the Cisco ONE model the Controller and Agent pillar is focused on providing a more generic controller-agent infrastructure that can be leveraged across multiple domains.

The Cisco “ONE” Controller is a Platform for generic control functions and state consolidation across multiple entities.

The ONE controller can communicate with Cisco network devices using the Cisco onePK API described on previous section and when using the onePK API it can access the service sets supported by the onePK. The ONE controller can use the OpenFlow Protocol to communicate with Cisco and non-Cisco devices that support an OpenFlow Agent.

The ONE controller core infrastructure is responsible for several functions, including processing a Dijkstra SPF (shortest path first) algorithm. It also provides a Topology Manager that can present a centralized topology view of the physical and virtual networks, a Forwarding Rules Manager for forwarding entries management among other core functions.

The controller also has some built-in Applications, for example the Network Slicing Manager, discussed in more detail on the next section, the Network Troubleshooting Manager and a Custom Routing application. The framework for those “built-in applications” is extensible and those are just examples of applications planned for the first release.

The controller than presents a set of northbound APIs (REST, WebSockets, OSGi) that can be used by applications (Cisco, customer-developed or 3rd party) to interact with the controller and ultimately with the infrastructure that the controller controls.

This controller and agent model, based on the Cisco ONE controller, presents a centralized view of the infrastructure to the applications so that the applications can directly perform changes to network infrastructure to meet the application needs. By providing those standard based northbound API the ONE Controller exposes in a programmatic way all the capabilities of the infrastructure that it controls but it also adds value by providing some built-applications discussed above. The Cisco ONE controller is a JAVA based application.

Figure 11-7 shows the ONE Controller Architecture and its components.

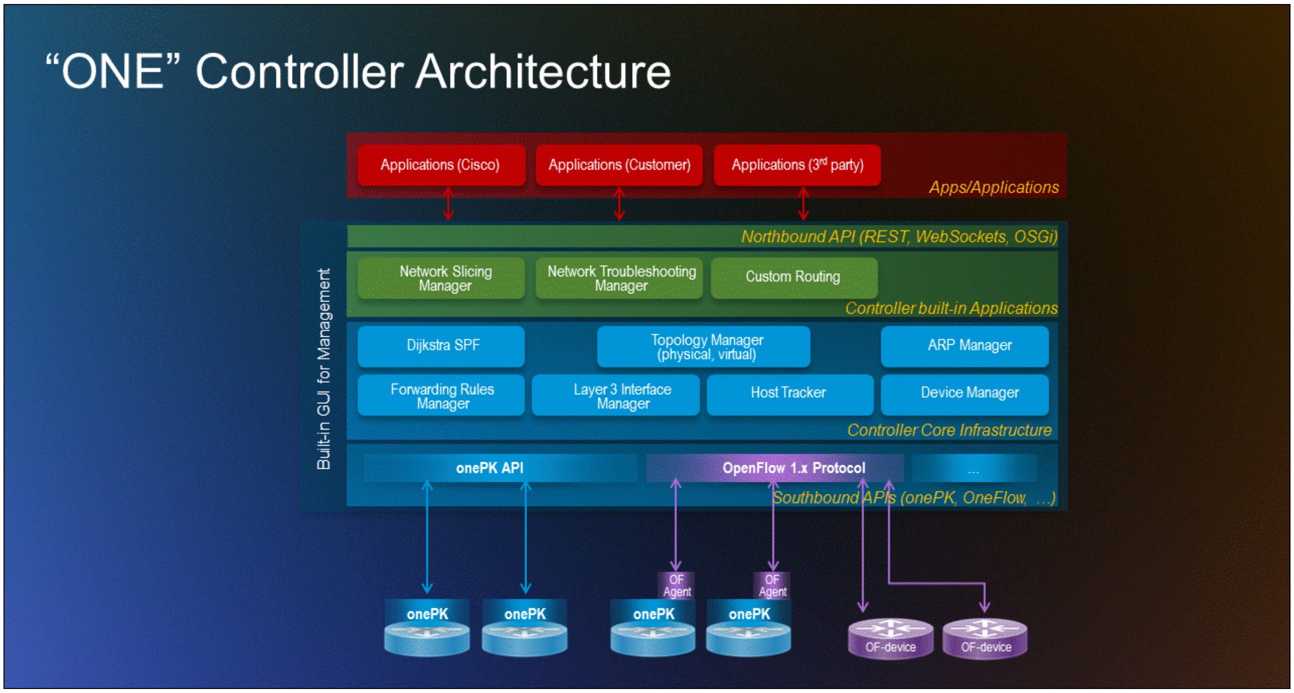

"ONE" Controller Architecture
Applications (Cisco)
Applications (Customer)
Applications (3rd party)
Apps/Applications
Built-in GUI for Management
Northbound API (REST, WebSockets, OSGi)
Network Slicing Manager
Network Troubleshooting Manager
Custom Routing
Controller built-in Applications
Dijkstra SPF
Topology Manager (physical, virtual)
ARP Manager
Forwarding Rules Manager
Layer 3 Interface Manager
Host Tracker
Device Manager
Controller Core Infrastructure
onePK API
OpenFlow 1.x Protocol
...
Southbound APIs (onePK, OneFlow, ...)
OF Agent
OF Agent
onePK
onePK
onePK
onePK
OF-device
OF-device

Figure 11-7 ONE controller Architecture

OpenFlow

OpenFlow is a protocol used for the communication between a controller and a network device that understand the OpenFlow protocol, usually a switch. It gives the remote controller the ability to modify the behavior of network devices through a well-defined "forwarding instruction set." The network devices that can support OpenFlow include routers, switches, virtual switches, and access points.

According to the definition from the ONF (Open Network Foundation), OpenFlow is an open standard that enables researchers to run experimental protocols in the campus networks and it provides a standardized hook to allow researchers to run experiments without requiring vendors to expose the internal workings of the network devices. This definition and use case (researchers to run experimental protocols) has some historical reasons given that OpenFlow was initially developed by universities and to be used within the university campus. In the context of this book, the applicability of OpenFlow needs to be considered within the Data Center.

In a classical router or switch, the fast packet forwarding (data path) and the high level routing decisions (control path) occur on the same device. The way OpenFlow works is that it separates these two functions. The data path portion still resides on the switch, while high-level routing decisions are moved to a separate controller, typically a standard server. The OpenFlow Switch and Controller communicate via the OpenFlow protocol, which defines messages, such as packet-received, send-packet-out, modify-forwarding-table, and get-stats. Figure 11-8 illustrates the OpenFlow model.

The data path of an OpenFlow Switch presents a clean flow table abstraction; each flow table entry contains a set of packet fields to match, and an action (such as send-out-port, modify-field, or drop). When an OpenFlow Switch receives a packet it has never seen before, for which it has no matching flow entries, it sends this packet to the controller. The controller then makes a decision on how to handle this packet. It can drop the packet, or it can add a flow entry directing the switch on how to forward similar packets in the future.

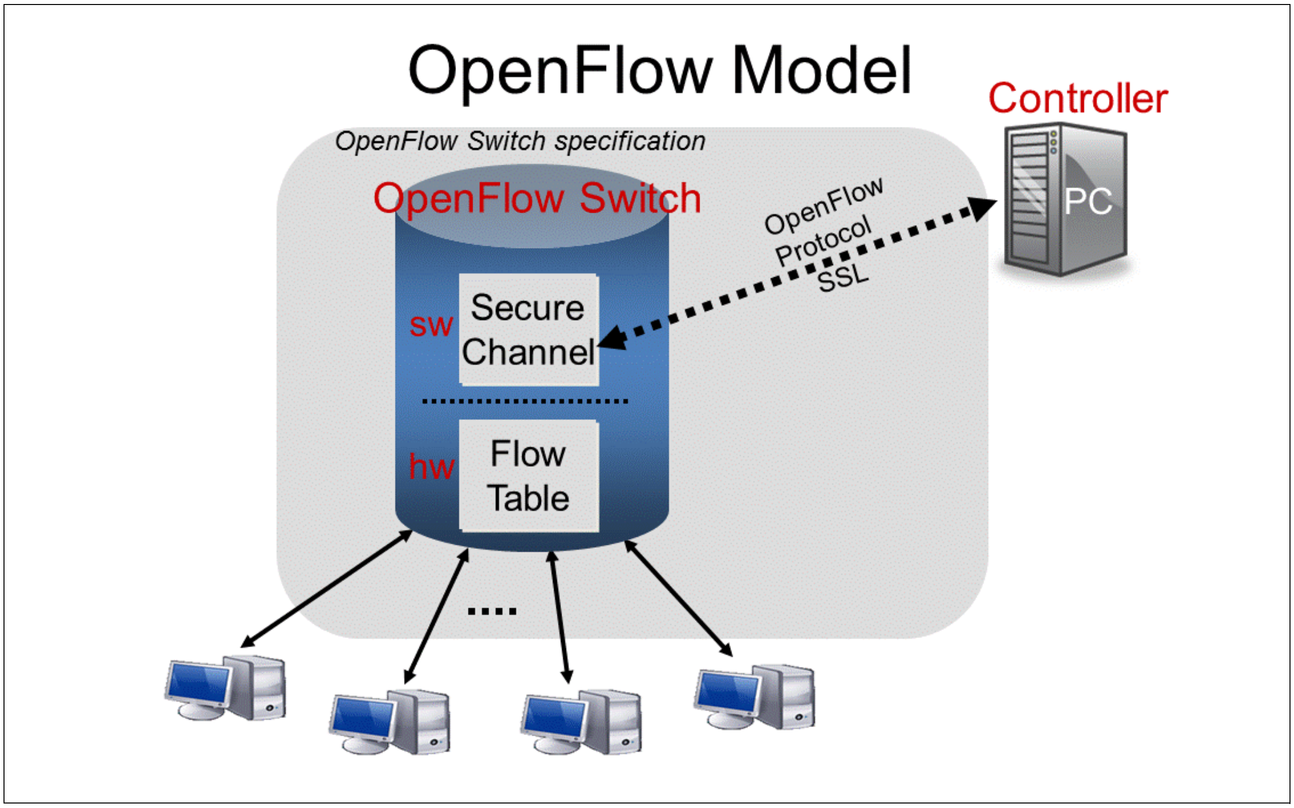


Figure 11-8 The OpenFlow mode. The Flow Table is controlled by a remote controller via the Secure Channel

The OpenFlow protocol is being standardized at the Open Networking Foundation (ONF). It has been going through a rapid evolution since version 1.0 released in December 2009.

As described previously, the OpenFlow protocol is part of Cisco Open Network Environment, and specifically will be supported on the Cisco ONE controller and OpenFlow agents are available on some selected Cisco switches.

Note: For more information about OpenFlow, visit: `https://www.opennetworking.org`

11.2.3 Virtual network overlays

Server virtualization brings with it new data center networking requirements. In addition to the regular requirements of interconnecting physical servers, network designs for virtualized data centers have to support the following needs:

- Huge number of endpoints: Today physical hosts can effectively run tens of virtual machines, each with its own networking requirements. In a few years, a single physical machine will be able to host 100 or more virtual machines.
- Large number of tenants fully isolated from each other: Scalable multi-tenancy support requires a large number of networks that have address space isolation, management isolation, and configuration independence. Combined with a large number of endpoints, these factors will make multi-tenancy at the physical server level an important requirement in the future.
- Dynamic network and network endpoints: Server virtualization technology allows for dynamic and automatic creation, deletion and migration of virtual machines. Networks must support this function in a transparent fashion, without imposing restrictions due to IP subnet requirements, for example.

Virtual overlay networks can be used to meet some of those requirements in a way similar to creating virtual servers over a physical server. They independent of physical infrastructure characteristics, ideally isolated from each other, dynamic, configurable and manageable.

Hypervisor based overlay networks (which take advantage of virtual Ethernet switches) can provide networking services to virtual servers in a data center. Virtual Ethernet switches form part of the platform for creating agile network services; they can also aid in simplifying network control and management and automating network virtualization services. Overlay networks are a method for building one network on top of another.

The major advantage of overlay networks is their separation from the underlying infrastructure in terms of address spaces, protocols and management. Overlay networks allow a tenant to create networks designed to support specific distributed workloads, without regards to how that network will be instantiated on the data center's physical network. In standard, TCP/IP networks, overlays are usually implemented by tunneling. The overlay network payload is encapsulated within an overlay header and delivered to the destination by tunnelling over the underlying infrastructure.

Within Cisco ONE, VXLAN, which has been submitted to IETF for standardization, is the technology adopted to enable Virtual Network Overlays. VXLAN is described in more details in Chapter 5, “Cisco virtual networking solutions” on page 253.

IBM has proposed Distributed Overlay Virtual Ethernet (DOVE) as the technology to enable Virtual Network Overlays.

At the moment, multiple solutions for Virtual Network Overlays have recently been proposed by networking product providers. The industry has begun work to find common areas of standardization. The first step towards this goal has been to publish a common problem statement through the IETF and forming a working group to standardize on solutions.

12

IBM and Cisco next generation architectures for Small, Medium, and Large Data Centers

In this chapter, we present reference architectures for Small, Medium, and Large Data Centers. The architectures described can be deployed today and are based on the Cisco and IBM technologies that were introduced in the previous chapters.

The chapter explores the following topics:

- In section 12.1, “IBM and Cisco Data Center objectives” on page 552, we describe the objectives of the modern data center.
- In section 12.2, “IBM and Cisco Data Center component model” on page 552, we introduce the data center component model, pinpointing which of the previously introduced hardware and software components can be used for which functional area.
- In section 12.3, “Data center requirements” on page 560 modern data center, we explore the technical, non-functional, and operational requirements.
- In section 12.4, “Data center use cases” on page 562, we present three use cases (Small, Medium, and Large Data Center) that differ in terms of requirements to give the reader several reference designs that can be used as starting points for specific data center designs.
- At the end of the chapter, we discuss the IBM and Cisco Services that can be leveraged to plan, design, and implement the modern data center building blocks and architecture, which have been described throughout this book.

12.1 IBM and Cisco Data Center objectives

As introduced in Chapter 1, “The modern data center” on page 1, many different IT and business requirements are shaping modern data center architectures. From a business perspective, the modern data center has to deliver IT services efficiently, optimizing the usage of capital and assets. From a high level point of view, we can identify four main operational objectives for modern data centers:

- Control: The ability to monitor and govern the IT and Network infrastructure is a fundamental business requirement. Obtaining IT infrastructure visibility is the first stepping stone to achieve the required level of control and the correct tools need to be deployed to achieve this.
- Scalability: The ability of expanding the IT infrastructure with minimal risk and effort is a very important point to make sure the data center design is future-proof and able to meet the objectives in terms of data and computing capacity growth we are currently witnessing
- Stability: The ability to minimize risk of faults and the ability to quickly and seamlessly recover from faults is also one of the key objectives of modern data centers which cannot tolerate any kind of downtime
- Flexibility: The ability to meet present and future requirements and quickly leverage new possibility is becoming more and more important as budget constraints and fast paced technology evolutions are common characteristics of the IT industry.

12.2 IBM and Cisco Data Center component model

In this section, we introduce the IBM and Cisco Data Center component model, which highlights the functional areas a modern data center needs to address. This model is used as a basis to develop the use cases (in 12.4, “Data center use cases” on page 562). The infrastructure building blocks presented in the previous chapters of this book are mapped to the components so that the reader can understand which products and technologies can be leveraged to deliver a certain function.

Figure 12-1 shows the IBM and Cisco Data Center component model.

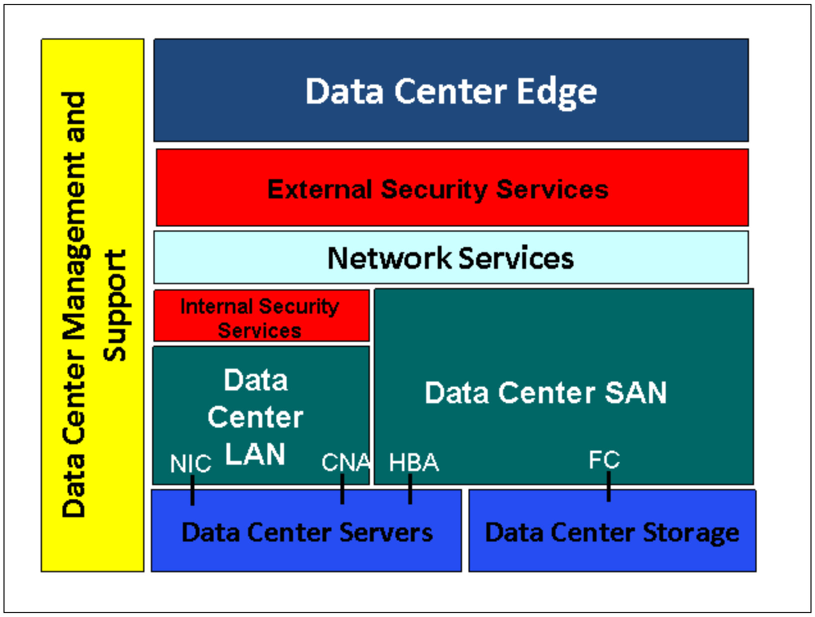


Figure 12-1 Data Center component model

From top to bottom the functional areas are:

- Data Center Edge: This block allows external access to the data center resources.
- External Security Services: This block enforces security policies on traffic to and from the data center.
- Network Services: This block provides all the networking services required for the data center except packet forwarding and security services.
- Internal Security Services: This block enforces security policies on traffic within the data center.
- Data Center LAN: This block provides packet forwarding services between clients and servers and between servers in the data center.
- Data Center SAN: This block provides connectivity for block level I/O (as opposed to file level I/O) between server and storage end points within the data center.
- Data Center Servers: This block provides data center computing power resources. Server attach to the LAN and SAN via NICs, CNAs or HBAs.
- Data Center Storage: This block provides data center data storage resources. Storage is connected to the SAN using the FC protocol.
- Data Center Management and Support: This block enables to control, monitor and remotely support data center resources.

In the following sections, we provide additional information on all the building blocks depicted in Figure 12-1. The subcomponents are also covered in greater detail in Chapter 6, “Application networking solutions and data center security, physical and virtual” on page 287.

12.2.1 Data Center Edge

The Data Center Edge block functional components are shown in Figure 12-2.

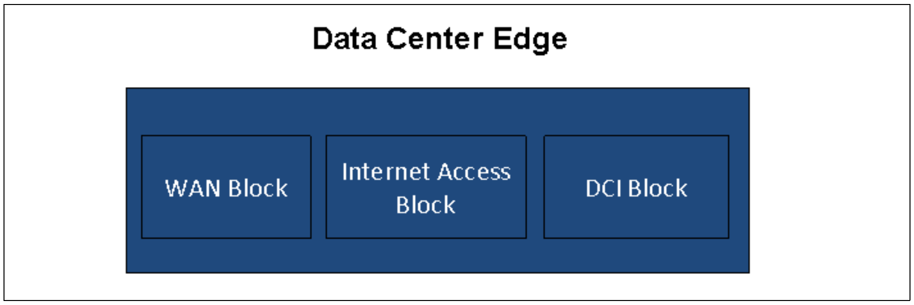


Figure 12-2 Data Center Edge component

The Data Center Edge block allows external access to the data center resources. The data center edge provides the remote connection point from branch offices, internet users, other data Centers and partners. The Data Center Edge connects directly to the core switches with 1 Gbps or 10 Gbps Ethernet. Depending on the customer and deployment requirements the data center edge may be broken into multiple physical blocks. The Data Center Edge function can be broken down in three main functional sub-components:

- WAN Block: Provides access via private networks to branch offices, campuses and extranets usually procuring circuits from Service Providers. The WAN Block is made up by Layer 2/ Layer 3 switches or routers directly connected to the core. The customer's WAN networks are generally “trusted” and typically consist of Virtual Private Networks such as MPLS networks, Point-2-Point private lines, and Metropolitan Area Networks (MAN). The WAN block may contain a Layer 3 routing instance and sometimes make use of Virtual Routing and Forwarding (VRFs) instances to provide Layer 3 segmentation. WAN connections are typically circuits such as MPLS, Metro Ethernet, Frame Relay, SDH/SONET and ATM.
- Internet Access Block: Provides Internet access for clients accessing the data center from the Internet. Usually this block has dedicated hardware so that the connections coming from the Internet (not trusted) can have a dedicated LAN segment. This block also usually hosts the internet facing (external-facing) servers and they are usually placed in a security zone called DMZ (“demilitarized zone”).
- Data Center Interconnect (DCI) Block: Provides access to other data centers within a Metropolitan Area or Geographically further away. This provides high availability by enabling Active/Active data centers and also Disaster Recovery by enabling Active/Standby data centers. Usually metro connectivity is provided by deploying dark fibers or DWDM optical connectivity using a point-to-point or ring topology.

12.2.2 External Security Services

The External Security Services block enforces security policies on traffic to and from the data center. Some of these services may be deployed on dedicated hardware while others may be collapsed on the same devices. An example of a device capable of providing all these functions is the Cisco ASA, which has been introduced in Chapter 6, “Application networking solutions and data center security, physical and virtual” on page 287.

Figure 12-3 shows the functional sub-components in the External Security Services block.

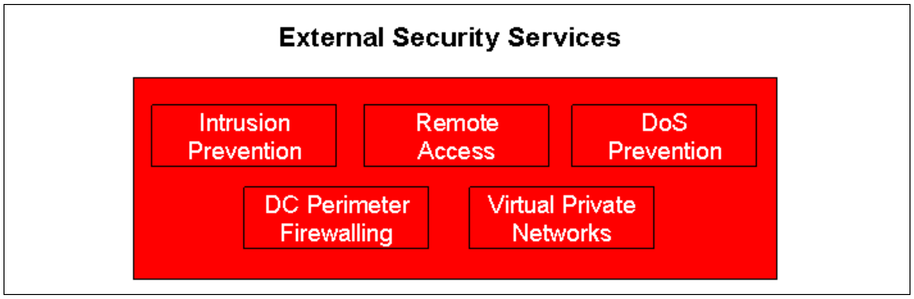


Figure 12-3 External Security Services component

- Intrusion prevention: Provides traffic inspection capabilities. This can be used to inspect all the traffic coming to and from the data center or can be used to selectively inspect the traffic coming from or going to the Internet.
- Remote access: Provides secure access (VPN) for clients accessing the data center from the Internet: Usually this is placed in the DMZ and may be paired with the Internet Access Block.
- DoS prevention: Provides Denial of Service mitigation capabilities. Usually this is placed in the DMZ.
- DC Perimeter firewalling: Provides security policies enforcement for traffic coming in and going out of the data center.
- Virtual Private Networks: Provides virtual private network capabilities for clients on other sites to access the data center via site to site VPNs, those site to site VPNs are usually established over the Internet.

12.2.3 Network Services

The Network Services block provides for the deployment of network aware functions and applications that require leachability to all layers of the data center LAN environment. The services block enables a backbone “point of service” used for DC services aggregation. Often these functions are deployed through dedicated, purpose-built appliances, services modules that are inserted into Cisco switches, and in some cases multiple services can be collapsed on the same hardware. External compliance regulations may require the need to deploy security services, for example, through physical separation, while DNS and DHCP services can be consolidated on the same physical appliance.

Figure 12-4 shows the sub-components within the Network Services block.

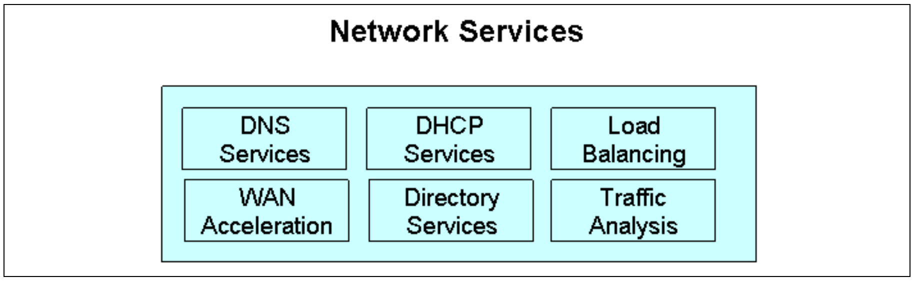


Figure 12-4 Network Services component

- DNS Services: An Internet service that translates domain names into IP addresses. Because domain names are alphabetic, they're easier to remember. The Internet however, is really based on IP addresses. Every time you use a domain name, therefore, a DNS service must translate the name into the corresponding IP address. The DNS system is, in fact, its own network. If one DNS server does not know how to translate a particular domain name, it asks another one, and so on, until the correct IP address is returned.
- DHCP Services: Dynamic Host Configuration Protocol, a protocol for assigning dynamic IP addresses to devices on a network. With dynamic addressing, a device can have a different IP address every time it connects to the network. In some systems, the device's IP address can even change while it is still connected. DHCP also supports a mix of static and dynamic IP addresses. Dynamic addressing simplifies network administration because the software keeps track of IP addresses rather than requiring an administrator to manage the task. Usually data center servers use statically assigned IP addresses or rely on the Load Balancer IP address.
- Load Balancing: Load balancers distribute traffic over multiple potential paths and servers so that data path overloads are minimized. Cisco ACE may be used to provide this functionality.
- WAN Acceleration: WAN acceleration devices provide data transfer efficiencies improving user experience and reducing bandwidth requirements. Some techniques to accomplish this are: Transport flow optimization, data redundancy elimination, compression, protocol acceleration. Cisco WAAS can be used to provide this functionality.
- Directory Services: Provides secure access to data center resources only to authorized users. Active Directory, RADIUS and TACACS are commonly used examples. Cisco ACS can be used to provide this functionality.
- Traffic Analysis: Allows administrators to mirror and inspect live traffic flows for monitoring or troubleshooting purposes. Cisco NAM can be used to provide inspection functionality, while the SPAN feature in NXOS can be used for mirroring of traffic.

12.2.4 Internal Security Services

The Internal Security Services block enforces security policies on traffic within the data center. Some of these services may be deployed on dedicated hardware while others may be collapsed on the same devices. Virtual security solutions may be leveraged at this layer. Examples of devices capable of providing all these functions is the Cisco ASA, and the Virtual Security Gateway or the Cisco ASA 1000v virtual firewall. Figure 12-5 shows the functional sub-components within the Internal Security Services block.

Internal Security Services
Intrusion Prevention
Virtual Security Services
DC Internal Firewalling

Figure 12-5 Internal Security Services component

- Intrusion prevention: Provides deep traffic inspection capabilities. Usually this is placed in the DMZ.
- Virtual Security Services: Provide security policies enforcement for traffic from Virtual-Machines and usually deployed within a hypervisor.
- DC Internal firewalling: Provides security policies enforcement for traffic within the data center.

12.2.5 Data Center LAN

The Data Center Local Area Network (LAN) block is a multi-layer hierarchical design that moves data traffic back and forth from the Data Center Edge to the DC Server block that handles client to server traffic and within the DC Server block that handles server to server traffic. The IBM and Cisco products introduced in Chapter 2, "IBM and Cisco building blocks for the data center" on page 29 can be used to provide these functions.

Figure 12-6 shows the functional sub-components within the DC LAN block.

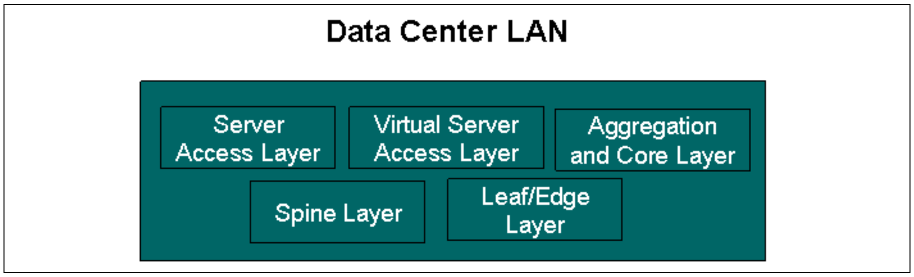


Figure 12-6 Data Center LAN component

- Server access layer: This block is where the servers physically attach to the network. Server components consist of standalone or virtualized servers, blade servers with integral switches or pass-through cabling, clustered servers, and mainframes with OSA adapters. The access layer network infrastructure consists of modular or fixed form factor switches and integral blade server switches. In case of a Fabric data center network this layer is called Leaf/Edge Layer, as introduced in 8.3.1, "FabricPath" on page 442.
- Aggregation and core layer: The core layer provides the high-speed packet switching backplane for all flows going in and out of the data center. The core layer provides connectivity to multiple distribution modules and provides a resilient Layer 3 routed fabric with no single point of failure. The aggregation layer serves as access port aggregation point and provides functions such as network services integration (FW, LB), Layer 2 domain definitions, spanning tree processing, and default gateway redundancy. For smaller or collapsed designs the Core/Distribution can be combined. In case of a Fabric data center network this layer is called Spine Layer, as introduced in the 8.3.1, "FabricPath" on page 442 and in the case of using FabricPath the Spine layer may or may not have all the functions of traditional aggregation/core layer but instead it may be much simpler.
- Virtual server access layer: When Server virtualization is used virtual switching functions may also be provided at the hypervisor level. In case of x86 virtualization, the Cisco Nexus 1000v 5.3, "Nexus 1000v" on page 258 can be leveraged to provide this function, instead of VMware virtual switches. In case of POWER platform the VIOS server will be used to virtualize the Server I/O.

12.2.6 Data Center SAN

The Data Center Storage Area Network (SAN) component provides connectivity for block level I/O between server and storage end points within the data center

Figure 12-7 shows the functional subcomponents for the Data Center SAN block:

- FC/FCoE Server Edge Layer: Provides SAN access to the Data Center Server block, allowing connectivity to the Storage block through the FC/FCoE Core Layer. At this layer both FC and FCoE are supported, through the use of both MDS fixed factor switches and Nexus fixed form factor switches.

- FC/FCoE Core Layer: Provides connectivity for the FC/FCoE Server Edge Layer, aggregating traffic through components such as the MDS and Nexus modular platforms. If FCoE is supported also in the core of the network then sometimes this is referred to as "multi-hop" FCoE. At this layer FC traffic replication is deployed through the Data Center Edge block using technologies such as FCIP or DWDM.

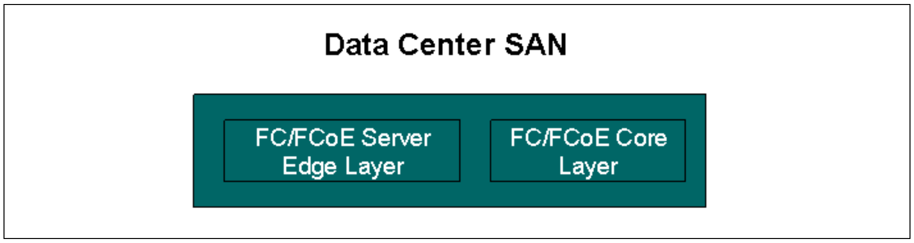


Figure 12-7 Data Center SAN component

Leveraging FCoE the DC LAN component and the DC SAN component can be merged, as described in Chapter 7, "Convergence of LAN and SAN: Fibre Channel over Ethernet" on page 369 and as depicted in Figure 12-8. In this case the data center will have a Converged Server access layer, common to both the SAN and the LAN, while Storage end points will support a mix of FC and FCoE protocols for network connectivity.

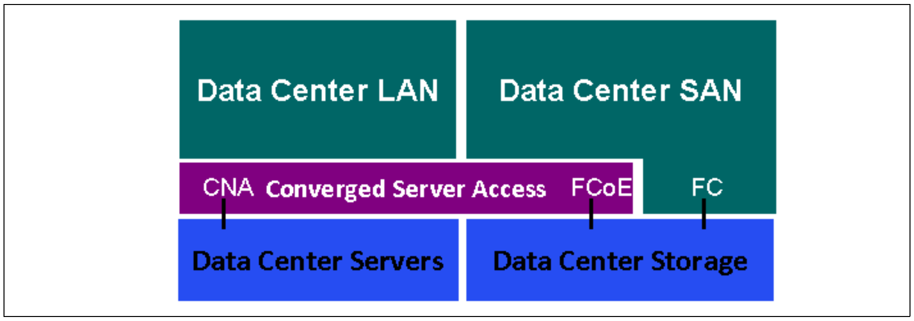


Figure 12-8 Converged Server Access

The choice of FCoE, NAS, or traditional Fibre Channel infrastructure is influenced by several factors. Chapter 7, "Convergence of LAN and SAN: Fibre Channel over Ethernet" on page 369 of this book provides a deeper insight into the SAN design choices. The adoption of FCoE brings several benefits to data center managers, including these:

- Reduce costs (many IBM servers have FCoE capable 10G NICs (zero cost) and the Cisco Nexus LAN infrastructure can support FCoE with just a software license.
- Reduced capital, energy and cooling costs with fewer adapters, cables and switches.
- Less hardware means reduced space and fewer rack units.
- Transitioning towards a converged infrastructure can reduce cost in environmental spend.
- FCoE virtualizes the cabling infrastructure and enables a more agile infrastructure.
- Optimized asset utilization with a unified, efficient DC fabric that is more flexible to the needs of the business.
- Lower management overhead by maintaining a single fabric.
- Increased application availability by simplifying the network and server configuration.
- Wire-once infrastructure that accommodates SAN, or iSCSI and can be dynamically reconfigured.

The most common inhibitor to implementing FCoE is not technical but operational. Many customers are not ready to take advantage of a converged infrastructure due to their siloed operational environment and separation of duties between SAN and LAN teams. However, most DC Managers and CIOs are working to converge IT operations whereby tools such as event management and asset registers are common across the server, storage, network and facilities siloes.

One key benefit of the proposed Cisco-IBM DC architecture is that customers who do not have a converged operational model today can gracefully migrate to a hybrid fabric with both FC and FCoE. The Nexus architecture allows customers to transition to a converged infrastructure at their own pace, one server at a time.

12.2.7 Data Center Servers

The Data Center Servers block provides data center computing power resources.

Figure 12-9 shows the functional subcomponents.

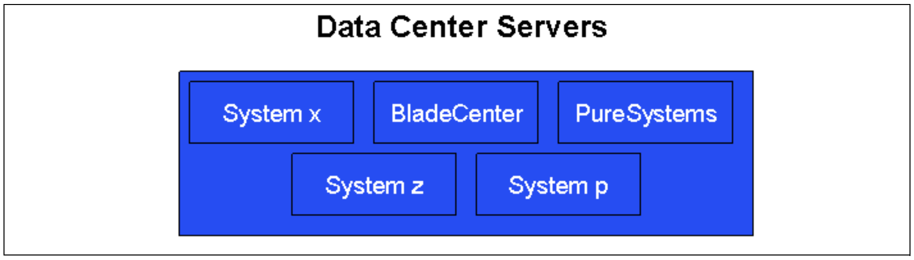


Figure 12-9 Data Center Servers component

The IBM building blocks for this functional block have been described in Chapter 2, “IBM and Cisco building blocks for the data center” on page 29.

The recommended server platform connectivity options have been described in Chapter 4, “Data center physical access layer evolution” on page 191.

12.2.8 Data Center Storage

The Data Center Storage block provides data center data storage resources.

Figure 12-10 shows the subcomponents for this functional component.

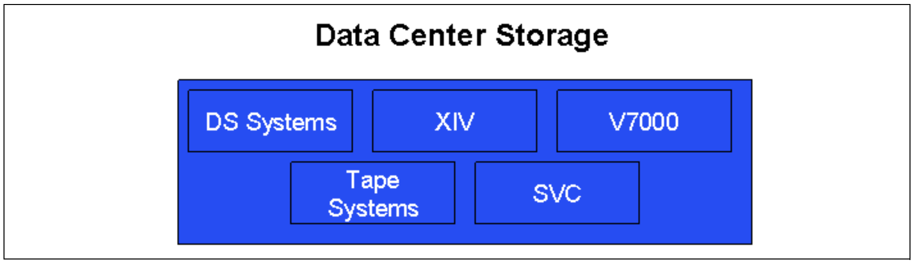


Figure 12-10 Data Center Storage component

The IBM building blocks for this functional block have been described in Chapter 2, “IBM and Cisco building blocks for the data center” on page 29.

12.2.9 Data Center Management and Support

The Data Center Management and Support block enables administrators to control, monitor and remotely support data center resources.

The Cisco and IBM solutions described in Chapter 10, “Data center network management and automation” on page 499 can be leveraged to provide these functions.

Figure 12-11 shows the subcomponent of this functional block.

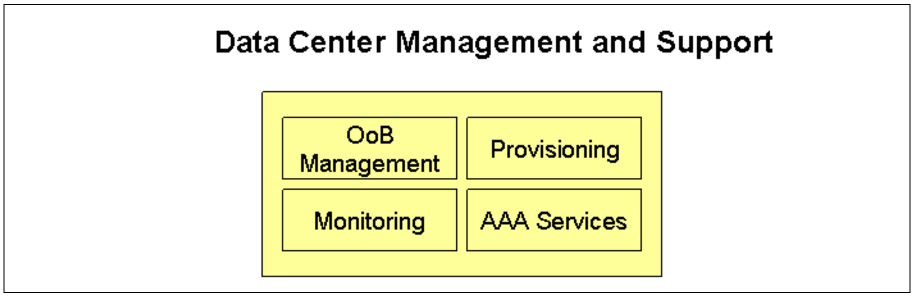


Figure 12-11 Data Center Management and Support component

- Out of Band (OoB) Management for Data center devices provides remote console access to the network devices via a Terminal Server and IP out of band management for network, SAN, and other devices.
- Provisioning refers to configuration and change management on the data center infrastructure.
- Monitoring refers to basic management capabilities on the data center infrastructure, usually using the SNMP protocol, syslog and ICMP.
- AAA Services refers to Authorization, Authentication, and Accounting Services to ensure the right users are authorized to perform management tasks and those operations are accounted for.

12.3 Data center requirements

This section discusses the various types of requirements applicable to a data center.

12.3.1 Technical requirements

These are the technical requirements:

- Network:
  - Logical design will enable 802.1Q trunking to facilitate zone and security model separation requirements
  - Multiple link redundancy has the functional capability of load-sharing traffic across links while not employing Spanning Tree Protocol (STP) interface blocking
  - Support inter-switch connectivity at 1Gb, 10Gb, 40Gb or 100GB with link aggregation for higher throughput
  - Must be able to provide converged switching fabric for both data and storage protocols

- Server requirements:
  - Design will enable both NIC teaming and LAG server and appliance connectivity
  - Provide the ability in hardware for each individual host to connect to two different Top-of-Rack switches
  - Server and appliance attachment will use Top-of-rack (ToR), End-of-Row (EoR), or Center of Row (CoR) designs in order to leverage low cost NIC connections using standard copper jumpers and minimize custom cable installations
  - Server PoDs will allow for a variety of server attachment patterns (1 GE and 10 GE) and a combination of physical and virtualized hosts
  - Support 100 Mb (for legacy), 1Gb, and 10 Gb host connectivity
- WAN requirements:
  - For Medium to Large Data Center environments, the WAN block should also contain a Layer 3 routing instance to provide both static and dynamic routing functions facing the WAN providers,
  - Provides Access to customer private MPLS network,
  - Provides Internet VPN customer access / business partner/ B2B,
  - The WAN block should include a Layer 2 switching function that acts as a landing point for trusted WAN termination. This Layer 2 domain allows for deployment of WAAS technologies,
  - WAN block must be able to re-distribute BGP into OSPF and OSPF or static routes into the BGP routing protocol,
- MAN
  - Network Cores and last mile connectivity can be split between 2 sites to provide true and real disaster recovery capabilities.
  - Provides standard for MPLS MAN connectivity to separate security zones over a single MAN circuit between data centers.
- Internet
  - Capable of enforcing security function zones and perimeters, including to virtualized switching layer
  - Logical separation using VLANs
  - Must have the capability of creating multiple security functional zones and perimeters
  - Must be able to provide security services (IDS/IPS)
  - Provides external Public resource availability (that is, Hosting)
- Network general:
  - Must enable secure remote management, monitoring and control capabilities
  - Must support VLAN spanning between data centers
  - Quality of Service (QOS), Voice, Real-time services, video
  - Shall have the capability to expand and support (IPv6) as required
  - Shall provide and support Dynamic Routing Protocols (BGP, OSPF, EIGRP) for IPv4 and IPv6
  - Shall support IP Overlay Methodologies
  - Solution should include redundant power supplies

12.3.2 Non-functional requirements

These are the non-functional requirements:

- Must not have any single points of failure in hardware component. Hardware redundancy within and across DC switches.
- Shall be designed as Modular Building Blocks
- Shall be Highly Available
- Shall provide rapid deployable connectivity
- Solution platform should have hot-swappable hardware replacement capability
- Solution must provide all L3 (IP) routing functions and services in a highly available manner
- Solution should provide upgrades to higher speed interfaces as available (40Gb, 100Gb) without require chassis upgrade

12.3.3 Operational requirements

These are the operational requirements:

- Must enable ability to perform system upgrade/maintenance without down time
- Provide for a centralized management solution that controls all data center switching, be it Top-of-Rack, End-of-Row, or virtualized into server chassis
- Solution should allow for as little cabling between racks as possible. Cost-effectiveness for cabling should allow for copper-based options as well as fiber-optics where needed
- Solution should include redundant power supplies
- Enable switch logical consolidation (“stacking”) to reduce number of management entities
- Must support out-of-band management
- Management function that is capable of meeting security best practices such as Authentication, Authorization, and Accounting (AAA)

12.4 Data center use cases

This section describes three sample use cases that will include the building blocks introduced earlier in this book at that will adapt to the component model introduced in 12.2, “IBM and Cisco Data Center component model” on page 552.

This chapter describes the Small, Medium, and Large Data Center use cases that address the requirements described in Table 12-1.

Table 12-1 Use case requirements

REQUIREMENT	SMALL	MEDIUM	LARGE
Max # 1 G ports	768	3072	6144
Max # 10 G ports	512	2048	4096
# FEXs	16	64	128
# Aggregation blocks	collapsed	4	8
# Core devices	2 Nexus 5K or 6K*	2 Nexus 7K	4 Nexus 7K or 6K*
Fabric support	NO	YES	YES
vPC support	YES	YES	YES
NAS	YES	NO	NO
FCoE	optional on host and N Series	YES	YES
FC	optional	YES	YES
Legacy integration	NO	YES	YES
OTV	NO	YES	NO
1000v support	optional	optional	optional
Network Services	YES	YES	YES

Note: Cisco has recently (January 2013) announced the Nexus 6000 product family as well as the Nexus 2248-10G. Those are new products that were announced after the writing of this chapter. All the calculations and design decisions were made based on the capabilities and port density of the Nexus 5500 product family as well the Nexus 2232PP FEX.

Given that the Nexus 6000 and Nexus 2248-10G are in most cases a superset of the Nexus 5500 and Nexus 2232PP respectively, the reader is encouraged to consider the capabilities of those new products and use them if they further improves the designs and solutions proposed on this chapter.

The Nexus 6000 and Nexus 2248-10G capabilities are covered in Chapter 2, “IBM and Cisco building blocks for the data center” on page 29.

12.5 Small Data Center use case

In this section, we describe the Small Data Center use case.

- We first define the use case in terms of the characteristics that the use case must support, in 12.5.1, “Definition/introduction”
- We then introduce important assumptions behind this use case in terms of server, network, storage, security and legacy integration. Assumptions are defined with a broader scope than specific requirements so that the reader can understand which varying requirements can be addresses by the Small Data Center use case, in 12.5.2, “Assumptions”.

- In 12.5.3, “Design concept” on page 566, we define the design concept, which is then detailed in 12.5.4, “PoD design” on page 568 and in 12.5.5, “Service Zone design” on page 569.

12.5.1 Definition/introduction

The Small Data Center use case will have the following characteristics (see Table 12-1 on page 563 for more details):

- Maximum number of 1 G ports: 768 using only Nexus 2248TP-E.
- Maximum number of 10 G ports: 512 using only Nexus 2232PP. Usage of the Nexus 2248-10G is optional to improve 10 GE port density at the access layer and oversubscription.
- Number of FEXs: 16: 8 Nexus 2248s and 8 Nexus 2232s to obtain a mixed 1 GE / 10 GE access. Also the 2248-10G can be used to provide higher 10 GE port density at the access layer
- Number of aggregation blocks: 1 - collapsed core design
- Number of PoDs: 4, with 3 racks in each PoD for a total of 12 server racks.
- Number of Service Zones: 1 with a single aggregation block
- Number of Core devices: 2 Nexus 5596s to obtain a high port count at the collapsed core/aggregation layer. The Nexus 6000 can also be used.
- No FabricPath support
- vPC support
- NAS based storage with an optional FC attachment
- No Legacy integration (Greenfield Data Center)
- No Layer 2 extension (Single Site)
- Optional Nexus 1000v support
- Network Services included: WAAS, ACE, ASA
- WAN connection

Note: In order to leverage 40 GB connectivity between access and core layers, the Nexus 6000 platform can be leveraged.

Note: In order to improve 10 GE port density and the oversubscription ratio, the Nexus 2248-10G with 48 10 GE ports and 40 GE uplinks can be used in conjunction with the Nexus 6004/6001 platform.

12.5.2 Assumptions

This section introduces the most important assumptions for this use case. These will be broken down into Servers, Network, Storage, Security, and Legacy integration assumptions.

Number of servers

We use the following assumptions for the Data Center Server component:

- Number of servers: 76 total physical servers, of which roughly 75% are virtualized, with 50 VMs per server as a baseline. The design will have to scale up to 92 physical servers and can also include servers with a higher VM density. This means that the access switches need to account for this growth in terms of server access ports. We use rack mountable server instead of Blades because usually Small Data Centers do not require very high density computing and this allows to use different platforms with more flexibility and the environment is easier to maintain from an operational perspective even if the space footprint will be increased.
- Number of NICs and speed: Each server will typically require 3x1 Gigabit Ethernet ports and a minimum of 2x10 Gigabit Ethernet ports. Server with higher I/O requirements can also include 3x or 4x10 Gigabit. Also a single 10 GE port can be used to consolidate the 3 1 GE ports, if needed, as this is in line with server virtualization best practices. Having redundant connectivity for VMware VCenter gives greater reliability at the expense of higher 1 GE port count at the access layer.
- Number of images: The design will have to accommodate 2,850 VMs (50 VMs x 76 physical servers x 75% virtualization ratio) and 11 physical servers. With additional servers and more aggressive VM densities the design can scale up to higher number of VMs depending on many different factors such as workload characteristics, security and availability requirements, compliance requirements, operational requirements. In real life scenarios VM densities around 30 VM per physical server are more typical.
- Server platform mix: The server infrastructure will be fully based on x86 based IBM x Series. In order to accommodate higher VM densities the x3850 X5 have been chosen. This is a reasonable assumption for small and virtualized data center environments.

Network

We use the following assumptions for the Data Center LAN component; some are derived from the server infrastructure characteristics:

- Number of ports and spare ports: The Small Data Center use case will scale up to 512 10 GE ports or 768 1 GE ports. In this use case, we include 4 PoDs that will consume 228 1 GE ports (60% port utilization) and 152 10 GE ports (60% port utilization). Adding additional servers this can scale up to 384 1 GE ports and 256 10 GE ports to adapt to virtualized x Series best practices (2x10 GE and 3x10 GE from each physical host).
- Growth: We include 4 PoDs that will consume 228 1 GE ports (60% port utilization) and 152 10 GE ports (60% port utilization), leaving space for growth in terms of additional servers or additional ports to accommodate increase I/O requirements and / or and increase VM density per physical server. Also additional ports can be added on the Nexus 5596s or Nexus 6000.
- Power/Rack: The maximum power allowed per rack will be 12kW. The power consumption of the active equipment installed in the racks has to be kept below this limit.

Storage

We use Network Attached Storage for the Small Data Center use case. This is a common design option for Small Data Centers as it converges the servers and storage end points network. We assume each VM will have on average 45 GB of disk space. Assuming that non-virtualized server will have directly attached storage (DAS) the total storage requirement for the Small Data Center use case will be roughly 135 TB.

If also FC support is needed from the host side, FC native ports or converged FCoE ports (connected via CNAs on the server side) are available on the Nexus 5596 platform. This allows you to use HBAs and CNAs in the server block and FC/FCoE storage arrays in the storage block.

Security

We use a Cisco ASA as the consolidated Security Enforcement point for internal security services. Also, we assume that the data center will be accessible only from a private WAN so there is no need to deploy dedicated External Security Services appliances. Using this model it is possible to mix and match VM placement in different PoDs regardless of their security requirements (Application, Web, Database, Infrastructure Services) because using the FEX model all the ports will appear in the same configuration file (as if the data center network is made up by one single switch) allowing easy VLAN provisioning regardless of the physical server location and easy security policy configuration and enforcement on the Cisco ASA platform.

Legacy integration

We assume that the Small Data Center use case will be a greenfield environment, so no legacy integration is required.

12.5.3 Design concept

This section provides an overview of the Small Data Center use case design, elaborating on the design concept and the design decisions.

Overview

Figure 12-12 shows the Small Data Center use case design concept, based on the aforementioned assumptions. The design will include 4 PoDs (3 racks each) for servers (19 servers in each PoD) and access switches plus another 2 racks for the network collapsed core/aggregation layer, the NAS arrays and the network services appliances. The Routers in the WAN edge block are assumed to be in a separate Telecom room and will be outside the scope of this use case. For additional details on the WAN edge block refer to 12.8, “WAN Edge Block Design for Small, Medium, and Large Data Centers” on page 615.

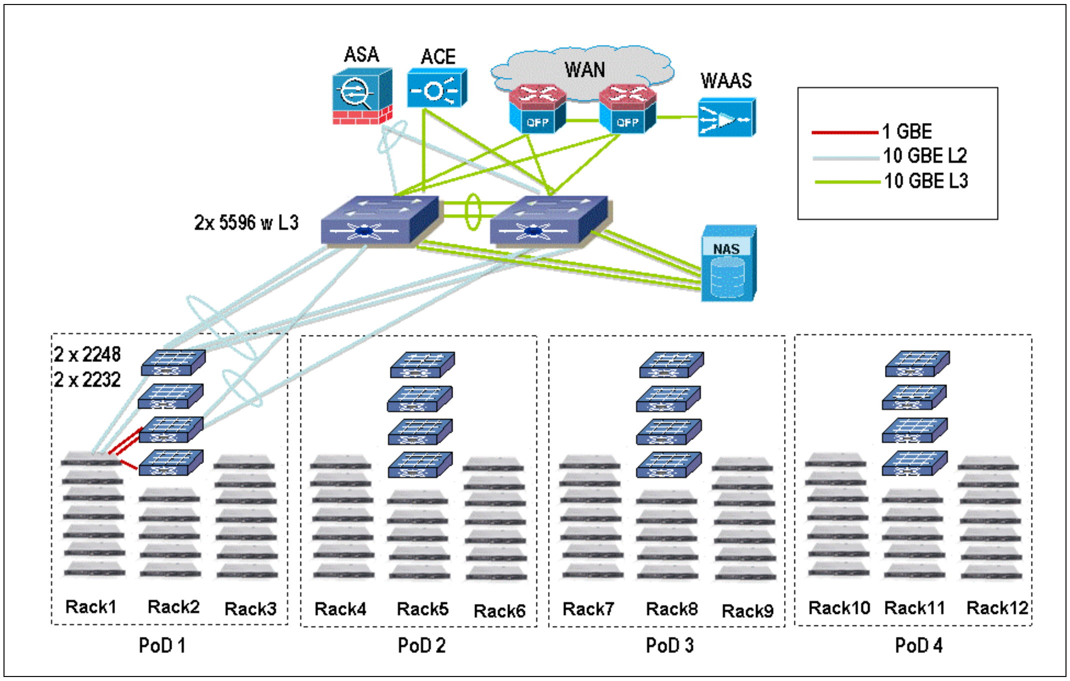


Figure 12-12 Small Data Center use case design concept

- The collapsed core/aggregation model will be used, with a mixed access layer made up by Nexus 2248s for 1 GE connections and Nexus 2232s for 10 GE connections. These FEXs will have uplinks to the collapsed core/aggregation block which will be deployed with 2 Nexus 5596s with Layer 3 expansion modules to allow having SVIs. In fact this will also become the Layer 2 / Layer 3 boundary for the Small Data Center use case.
- In order to accommodate virtualized x Series platforms, the servers will have connections to both access switches, as shown in Chapter 4, "Data center physical access layer evolution" on page 191.
- The Network Attached Storage arrays will be attached to the collapsed core/aggregation layer via Layer 3.
- The Network Services appliances will be also attached to the collapsed core/aggregation layer.
- The Edge block will also be connected to the collapsed core/aggregation layer via redundant Layer 3 connections, as shown in Figure 12-12. For more information on the WAN Edge design refer to 12.8, "WAN Edge Block Design for Small, Medium, and Large Data Centers" on page 615.

Design decisions

The FEXs have been chosen to scale the number of ports without having additional active switches in the racks to manage. This simplifies network operations. The Nexus 2200 are dual attached to provide additional resiliency in case of node failures. Another option would have single homing between the access and the collapsed core/aggregation layer - this would allow to include 32 FEXs instead of 16, but with lower resiliency. All the network appliances are deployed in an active/standby fashion even if Figure 12-12 shows just one for simplicity purposes.

The Nexus 6000 platform can be chosen in conjunction with the 2248-10G to obtain higher 10 GE density at the access layer and improve oversubscription through the use of 40 GE uplinks.

12.5.4 PoD design

This section further describes the PoD design chosen for the Small Data Center use case scenario. We provide additional details on the number of racks, the physical layout, oversubscription and design decisions.

Number of racks

As described in the previous section the overall solution will be made up of 14 racks - 12 for server and access switches and 2 for the collapsed network core/aggregation layer, storage and network services devices.

Each PoD will be deployed using 3 racks with access switches in the middle rack. This allows to reduce the number of orphan ports that would be associated with a 2 rack PoD design. Using higher computing densities (for example deploying BladeCenters) a 2 rack configuration may be worth exploring as an additional option.

Physical layout

Figure 12-13 shows the physical layout of the equipment in the 3 rack PoD. The shown links are for illustration only purposes and just refer to the top server in the middle rack. The 19 physical server are visible across the 3 racks. These consume 4 RUs each. Also the 2200 pairs (1 RUs each) are shown from the rear to simplify the drawing.

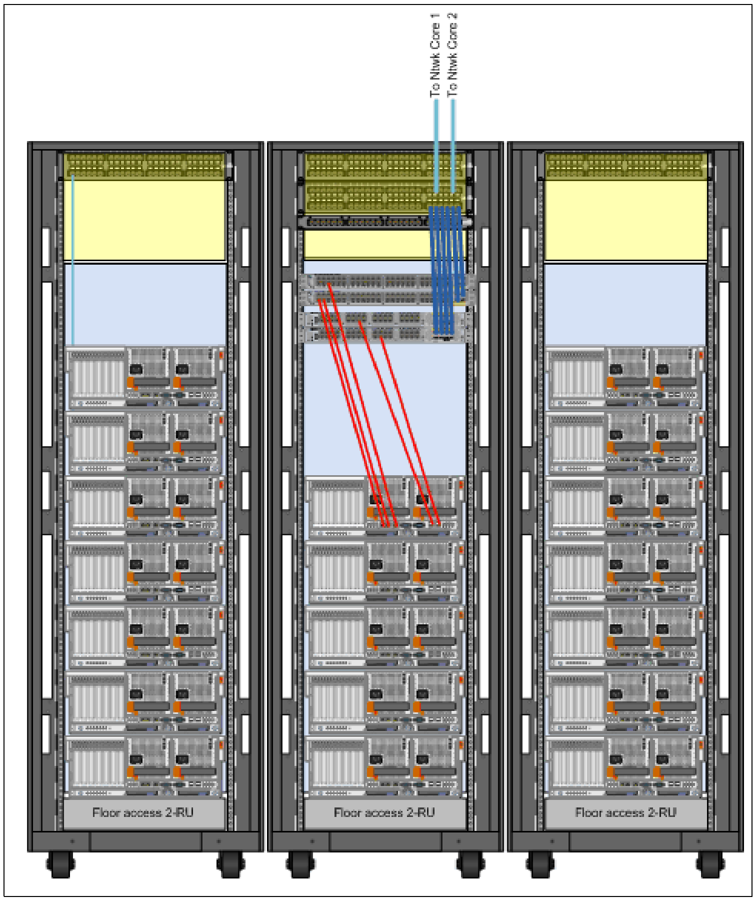


Figure 12-13 Small Data Center use case PoD rack layout

The red links are server to access switch links: 2x10G (production traffic), one to each of to the 2232 and 3x1G (2 for VMware VCenter and 1 for the RSA Adapter for Network Management), 2 to one 2248 and 1 to the other. In case the Nexus 1000v is used the 10 GE ports can be bundled in a LAG to provide additional resiliency.

The blue links represent uplinks to the core switches: the 2248s will typically have 2x10 GE while the 2232s will have 4x10G. This will allow us to calculate the oversubscription ratios in the next section.

As mentioned on the introduction section there is additional room for growth in terms of server density within the deployed racks. In fact 4 physical servers can be added in each PoD, resulting in 4x4=16 additional physical servers. This way the Small Data Center use case design can scale from 76 up to 92 physical servers.

Oversubscription

At the access layer the oversubscription ratio will be different for 10 GE ports and 1 GE ports. Also if servers have higher I/O requirements the design allows to add additional 10 GE ports, so we also measure the worst case oversubscription scenario.

- 10 GE ports at the access layer will have a 4.75: 1 oversubscription ratio (19 used ports vs 4 available uplinks, all 10 GE). This is acceptable as not all VMs will be using the links at all times. Adding additional servers and/or adding 10 GE interfaces to existing servers can get to worst case oversubscription ratio of 8:1 (using all 32 server access ports). If lower oversubscription is needed servers could also be attached directly to the Nexus 5500 at the collapsed core/aggregation layer, which has additional room for growth in terms of available 10 GE ports. Also in order to improve oversubscription all the 8x10 GE uplinks can be used from the 2232s - with 19 used ports this in fact generates a 2.375: 1 oversubscription ratio.
- 1 GE ports at the access layer will have a 1.9: 1 oversubscription ratio (38 used ports vs 2 10 GE uplinks). This is acceptable as 1 GE connections will be used only for systems and network management purposes and will not generate a lot of traffic. Also the FEXs aggregating the 2x1 GE ports from each physical server (instead of 1 GE port) can be mixed to reduce the oversubscription ratio and simplify cabling, if needed. Also the 3x1 GE ports can be consolidated on a single 10 GE port, if needed.

12.5.5 Service Zone design

The Small Data Center use case design will just include one Service Zone, as it will only include one distribution block and one network services block. This means that it will not be possible to address differently non functional requirements (for example availability) on different parts of the data center infrastructure.

Aggregation and core

The collapsed aggregation/core layer will have 80 x 10 GE ports for each Nexus 5596, with Layer 3 capabilities. Based on the design each PoD will aggregate 12x10 GE access ports (4x10 GE from each Nexus 2232 and 2x10 GE from each Nexus 2248) so the overall 10 GE ports consumed by PoD connectivity will be 48 on two Nexus 5596s (24 on each switch).

Additional ports will be consumed to attach the ACE devices (2x10 GE on each Nexus 5000), ASA devices (2x10 GE on each Nexus 5000) and NAS arrays (2x10 GE on each Nexus 5000) and the Edge Routers (1x10 GE on each Nexus 5000). The numbers include the active and the standby appliances where this is applicable. At least 2 x 10 GE links (vPC peer link) need to be used on both Nexus 5596s for the Inter Switch Link at the aggregation/core layer.

So only 66x10 GE ports (33 on each Nexus 5596) will be consumed at the collapsed core/aggregation layer out of the available 160. This gives more flexibility to accommodate more FEXs or additional 10 GE access ports.

It is always important to check the scalability limits of the network infrastructure, and in this use case, the following values apply:

- ARP Table size on the Nexus 5596 is limited to 16k entries: The Small Data Center use case will host between 3k and 7k VMs so the ARP limit is not an issue with this design
- MAC Address Table size on the Nexus 5596 is limited to 30k entries: The Small Data Center use case will host between 3k and 7k VMs so the MAC table size limit is not an issue with this design, even if each VM needs 3 MAC Addresses, which can be a requirement
- Number of FEXs: The design can scale up to 16 FEXs (4 PoDs) if the FEXs are dual homed to the aggregation layer. If the FEXs are single attached the FEXs can scale up to 32. In this case 48 additional ports will be consumed at the collapsed core/distribution layer, requiring a total of 112 10 GE ports, out of the available 160.
- Number of 10 GE access ports: The design with 16 FEXs can scale up to 32x8 = 256 10 GE ports while the design with 32 FEX can scale up to 512 10 GE ports. Another option that can be used to scale up to 512 10 GE ports but keeping the benefits of dual homing the FEXs is to use only Nexus 2232s as the access layer, instead of mixing 1 GE and 10 GE ports in each PoD. Additional 10 GE access ports are also available on the Nexus 5k.

Network services

The Small Data Center use case includes various network services that will be also deployed in a different fashion:

- Load balancing and application delivery: ACE will be deployed in Active/Standby mode via Layer 3 using Source NAT. Active/Active deployment is also an option and it is the preferred option in case of multi-tenancy requirements using ACE virtual contexts. For more information see 6.2, “Cisco Application Control Engine (ACE) and SLB: Considerations and solutions” on page 288.
- WAN acceleration: WAAS will be deployed in Active/Active mode via Layer 3, as an appliance on a stick (or one armed configuration). WAAS deployments require an additional appliance to function as the WAAS manager - this requires additional 10 GE ports if it is located in the Small Data Center use case. For more information, see 6.4, “Wide Area Application Services (WAAS)” on page 300.
- Firewalling and Intrusion prevention (internal security services): ASA will be deployed in Active/Standby mode in Bridge Group mode via Layer 2. Active/Active deployment is also an option and it is the preferred option in case of multi-tenancy requirements using ASA virtual contexts. For more information see 6.7, “ASA series security in the data center” on page 325.
- Cisco ISE/ACS are optional components for AAA
- Cisco NAM is an optional component for traffic analysis
- Cisco DCNM, Tivoli NM, NCM and OMNIbus are optional components for Data Center Network Management.

12.5.6 Storage and storage network design

For the storage subsystem we are assuming each VM will need 45 GB of disk space, the total storage requirement is 135 TB with 3k VMs. In order to deploy this 2 NAS units can suffice plus an expansion unit - allowing to scale up to 192 TB.

The NAS arrays will be connected to the collapsed core/aggregation layer via Layer 3 and each controller will be attached using 2x10 GE ports (one to each Nexus 5596). The units will consume a total of 12 RU so they can be hosted in the same 2 racks where the Nexus 5596 are placed to minimize cabling distance. Also the network services appliances (ACE, ASA and WAAS) can be hosted on this rack.

The possibility of adding FC and FCoE ports on the Nexus 5596 gives additional flexibility in terms of supported storage protocols in the data center. In fact using CNAs it is possible to support NFS, iSCSI, FC and FCoE or a mix of these, as required, as there is room for additional Ethernet, FCoE and FC ports (Unified ports) at the collapsed, converged aggregation/core layer.

12.6 Medium data center use case

In this section, the Medium Data Center use case is discussed.

- The use case is first defined in terms of the characteristics that it must support, in 12.5.1, “Definition/introduction” on page 564.
- Then important assumptions behind this use case are introduced in terms of server, network, storage, security and legacy integration. Assumptions are defined with a broader scope than specific requirements so that the reader can understand which varying requirements can be addressed by the Medium Data Center use case, in 12.6.2, “Assumptions”.

12.6.1 Definition/introduction

The Medium Data Center use case is defined in general to be up to four times as large as the Small Data Center use case. It is in also approximately half the size of the Large Data Center use case. The Medium Data Center use case will have the following characteristics (see Table 12-1 on page 563 for more details on the use cases):

- Maximum number of 1 Gbps ports: 3072 using only Nexus 2248s
- Maximum number of 10 Gbps ports: 2048 using only Nexus 2232s.Number of FEXs: 64: 32 Nexus 2248s and 32 Nexus 2232s to obtain a mixed 1 GE / 10 GE access layer.
- Number of aggregation blocks: 4 - (aggregation blocks are Layer 2 only)
- Number of Core devices: 2 Nexus 7000s to provide Layer 3 routing and VDC capability.
- Number of PoDs: 16

> **Tip:** 10G direct attach copper (DAC) cables (also known as Twinax Cables) have a maximum length of 10M. Three racks in a PoD enable the use of the 10M or less DAC cables. Pod sizes can be adjusted to account for server sizes and data center power restrictions.

- Number of Service Zones: 5 total (comprised of 1 per aggregation block and 1 that spans the DC.) This number maybe increased based on requirements.

> **Tip:** Multiple VLANs can be configured to span the data center if required. However, it is strongly recommended to limit the number of VLANs that span everywhere.

- Optional but recommended use of FabricPath to replace classical spanning tree or vPC.
- Optional NAS or FCOE based storage
- Legacy integration using a separate VDC at the core
- Optional Layer 2 extension between sites using OTV support (dual site)
- Optional Nexus 1000v support
- Network Services included: WAAS, Load Balancers, ASA

> **Note:** In order to improve 10 GE port density and the oversubscription ratio the Nexus 2248-10G with 48 10 GE ports and 40 GE uplinks can be used in conjunction with the Nexus 6000 platform.

12.6.2 Assumptions

This section introduces the most important assumptions for this use case. These will be broken down into Servers, Network, Storage, Security and Legacy integration topics.

The assumption section represents the requirements that were used when designing the data center solutions for this use case; the assumptions do not represent the maximum capacity or scale of any of the areas within the data center, including servers, network, and storage.

Number of servers

A medium enterprise data center typically has a mix of x86 Servers and UNIX Servers.

xSeries

In this use case, the x86 will be an IBM x3850 Series Server (4RUs) for both virtualized and non-virtualized workloads (smaller servers such as X3690 could also be used as non-virtualized servers).

An xSeries server can contain 50 virtual machines or server images. For this use case, a 50%/50% mix of xSeries virtual and non-virtual servers will be used. If fifty servers are virtualized with 50 virtual machines per server then the total amount of virtual images is 2500. Fifty non-virtualized servers represents 50 operating system instances running on a server. Therefore, in an one hundred server example with a 50%/50% mix of virtual and non-virtual servers 98% (2500/2550) of the server images would be virtualized.

pSeries

The UNIX Servers will be IBM System p770 Series Server with 4 CECs (16RUs) for virtualized workloads. A pSeries virtualized server can contain approximately 13 images or LPARs. (This is a conservative number of LPARs.)

For this use case, all pSeries servers will be virtualized.

Number of servers per PoD

Each physical PoD has three consecutive racks and is either fully equipped with System x Series servers or with System p Series servers.

System x Series server PoDs are each equipped with 19 Servers regardless if the servers are virtualized.

System p Series server PoDs are each equipped with 5 Servers with 4 CECs for virtualized pSeries. All pSeries are virtualized.

Number of servers in the data center

The Medium Data Center network design contains the following number of servers:

- xSeries Servers: 171 physical servers (86 non-virtualized and 85 virtualized servers). Using the calculation of 171 servers / 19 servers per PoD results in a total of 9 xSeries server PoDs.
- pSeries Servers: 35 physical (all virtualized servers). Using the calculation of 35 servers / 5 servers per PoD results in a total of 7 pSeries server PoDs.

Each of the 4 aggregation blocks contains 4 PoDs which results in a data center total of 16 PoDs. The Medium Data Center design requires 9 xSeries PoDs and 7 pSeries PoDs.

In an ideal world all servers in a PoD would be of the same type and function. In reality environments tend to be mixed. For conceptual purposes, the PoDs are assumed to be all of the same type of servers. As long as the port counts are not exceeded in a PoD, server types could be mixed.

- Number of NICs (LAN interfaces) and speed:

  We assume a 50/50 mix of virtualized and non-virtualized System x Series Servers. The System p Series Servers are all virtualized in this use case. The following number of NICs are required for LAN connectivity only:

  Virtualized System x:

  - 2x 10GE as Production/Admin/Backup & Recovery
  - 2x 1GE for VCC (Hypervisor Management)
  - 1x 1GE for RSA/IMM

  Non-virtualized System x:

  - 2x 10GE as Production/Admin/Backup & Recovery
  - 1x 1GE for RSA/IMM

    or

  - 2 x 1 Gbps for prod/Admin
  - 2 x 1 Gbps for Backup
  - 1 x 1 Gbps for RSA/IMM.

  Virtualized System p with 2 LAN VIOs:

  - 4x 10GE for Production/Admin
  - 4x 10GE for Installation, Backup and Recovery
  - 2x 1GE for VIO Management
  - 2x 1GE for VIO Installation, Backup and Recovery
  - 4x 1GE Service Processor interfaces for redundant HMC connectivity

With the NIC numbers above, their quantities are calculated for the data center and included in Table 12-2.

Table 12-2 Medium Data Center port counts

Server	Virtualized	# of Servers	1 Gbps ports	10 Gbps ports
IBM X-Apps	No	43	215	0
IBM X -DB	No	43	43	86
IBM X -VM	Yes	85	255	170
IBM X - Total		171	513	256
IBM P-Apps	No	0	0	0
IBM P -DB	No	0	0	0
IBM P-VM	Yes	35	280	280
IBM P - Total		75	280	280
Total Ports			793	536

- Number of images

  The maximum number of virtualized images per physical server is the following:

  – Virtualized System x Series Server: maximum of 50 VMs

  – Virtualized System p Series Server: maximum of 13 LPARs

Network

The following assumptions were used for the data center LAN component:

- Oversubscription:

  For the initial setup, a reasonable oversubscription is acceptable (this design will target a max 5:1 per network device and 30:1 for worst case scenario for end-to-end oversubscription) as it is very unlikely that the servers will use all the available link capacity. Potential future bottlenecks identified via a proper capacity management can be mitigated by adding links, which reduces the over subscription.

- Traffic flows:

  Approximately 70% of the traffic is expected to be East<->West so it remains local within the data center, and 30% of the traffic is North <-> South and leaves the data center via the WAN (Wide Area Network)

- Number of ports and spare ports:

  With 64 FEXs the Medium Data Center use case can scale up to a maximum of 2048 10 GE ports or 3072 1 GE ports. In this specific use case, we include 4 aggregation blocks, each one with 4 PoDs with a mix of 10GE and 1GE network ports. Overall this will provide 1024 10GE ports (52% port utilization) and 1536 1 GE ports (52% port utilization).

  The power density and cooling tend to be the limiting factor in the number of servers and utilization of network port capacity.

We include 4 PoDs per Access Block and 4 aggregation blocks in the data center. The Medium Data Center use case can grow into the Large Data Center use case by adding aggregation blocks to the core, and by adding core devices. If there is the desire to grow the solution proposed for Medium Data Center into the Large Data Center solution, then FabricPath has to be adopted. Adopting FabricPath when first deploying the Medium Data Center solution will allow an easier transition to the Large Data Center solution.

- Power/Rack:

  The maximum power allowed per rack will be 12kW. The power consumption of the active equipment installed in the racks has to be kept below this limit.

- Virtual machines:

  The number of virtual machines for the Medium Data Center design is 4 x the number of virtual machines for the small design. In theory this number is 28000. In reality, other limits based on the number of ports will be reached first.

Storage

Fibre Channel over Ethernet (FCoE) and native Fibre Channel (FC) are used for the Medium Data Center use case. The FC attached storage arrays will be connected using multiple 8G connections to a dual fabric setup to ensure high availability and load balancing between the storage controllers. From a storage array perspective the connections are FC but from host side they are seen as FCoE as seen in Figure 12-18 on page 581.

Security

Cisco ASAs are used as the consolidated Security Enforcement point for internal and external security services.

- The data center will be remotely accessible from a private WAN and from the internet.
- Internal security services are attached at the core layer.
- External security services are included in the internet access component of the data center edge and discussed in section 12.9, “DMZ design” on page 617.

Service Zones

This Medium Data Center design creates 5 Service Zones. Four zones that are limited to just a single Access Block and a fifth Service Zone that spans the entire data center. This will be discussed in greater detail in the Design Concept Section.

Using this model it is possible to mix and match VM and LPAR placement in different PoDs within an Access Block regardless of their security requirements (Application, Web, Database, Infrastructure Services).

Legacy integration

Most of the time when a new data center architecture is implemented it must be connected to an already existing data center environment. In the context of the Medium Data Center described in this chapter the engineering team can take advantage of the Virtual Device Context (VDC) feature of the Nexus 7000 switch to interconnect the new and legacy data centers. VDCs were described in Chapter 9 (NX-OS Network Operation System).

See section “Interconnecting the new and legacy data centers” on page 588 for a description of the legacy integration.

12.6.3 Design concept

This section provides an overview of the Medium Data Center use case design, elaborating on the design concept and the design decisions. This is only one example of how to construct a Medium Data Center design.

> **Tip:** In order to understand and make design decisions about routing (L3) and VLANs (L2) it is important to understand Virtual Port Channels and FabricPath. VPC and FabricPath are discussed in Chapter 8, “Overcoming the limitations of traditional data center networks” on page 421.
>
> An excellent source is also located at these websites:
>
> For vPC:
>
> `http://www.cisco.com/en/US/docs/switches/datacenter/sw/design/vpc_design/vpc_best_practices_design_guide.pdf`
>
> For FabricPath:
>
> `http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9402/white_paper_c11-687554.html`

Overview

Figure 12-14 shows the high level concept of the Medium Data Center design. It depicts a single core (pair) of Nexus 7000s with up to four Access Blocks. Each Access Block is made up of a single pair of Nexus 5596s with up to 4 PoDs. Each PoD consists of 4 FEXs.

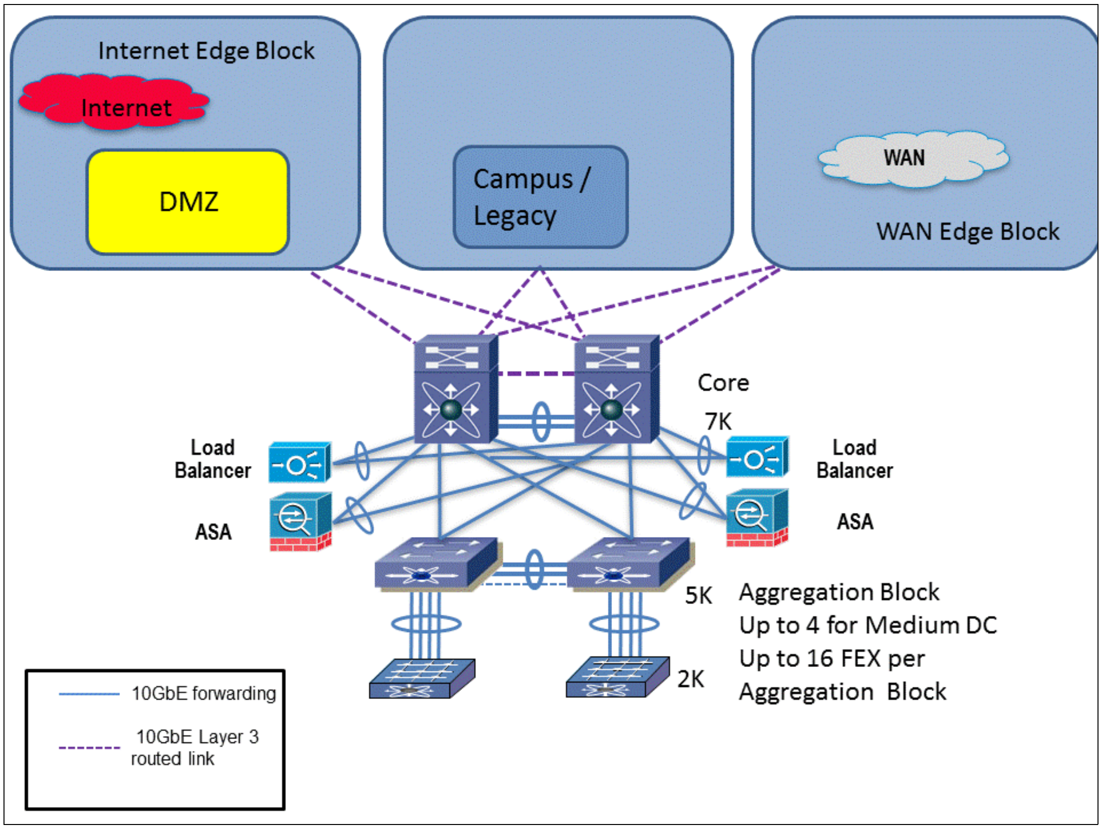


Figure 12-14 High level concept of Medium Data Center design

The Access Block design is discussed in Chapter 4, “Data center physical access layer evolution” on page 191.

The Medium Data Center design uses a routed core (Layer 3) model with Layer 2 switching only at the Access Blocks. A pair of Nexus 7000s configured as a vPC domain is used as the core. For simplicity, the Nexus 7000 core provides routing capability for the entire data center design. Only Layer 2 switching is configured on the Access Blocks. The aggregation blocks are pairs of Nexus 5500s as vPC Domains with attached FEXs. (The Nexus 5500s access switches could support a Layer 3 routing instance if required.)

The Nexus 7000 core runs a dynamic protocol with the WAN edge block. It also functions as the HSRP or VRRP default gateway for the servers.

The medium Data Center design uses a Nexus 7000 vPC domain connected to a Nexus 5K vPC domain. This design is depicted in Figure 12-15.

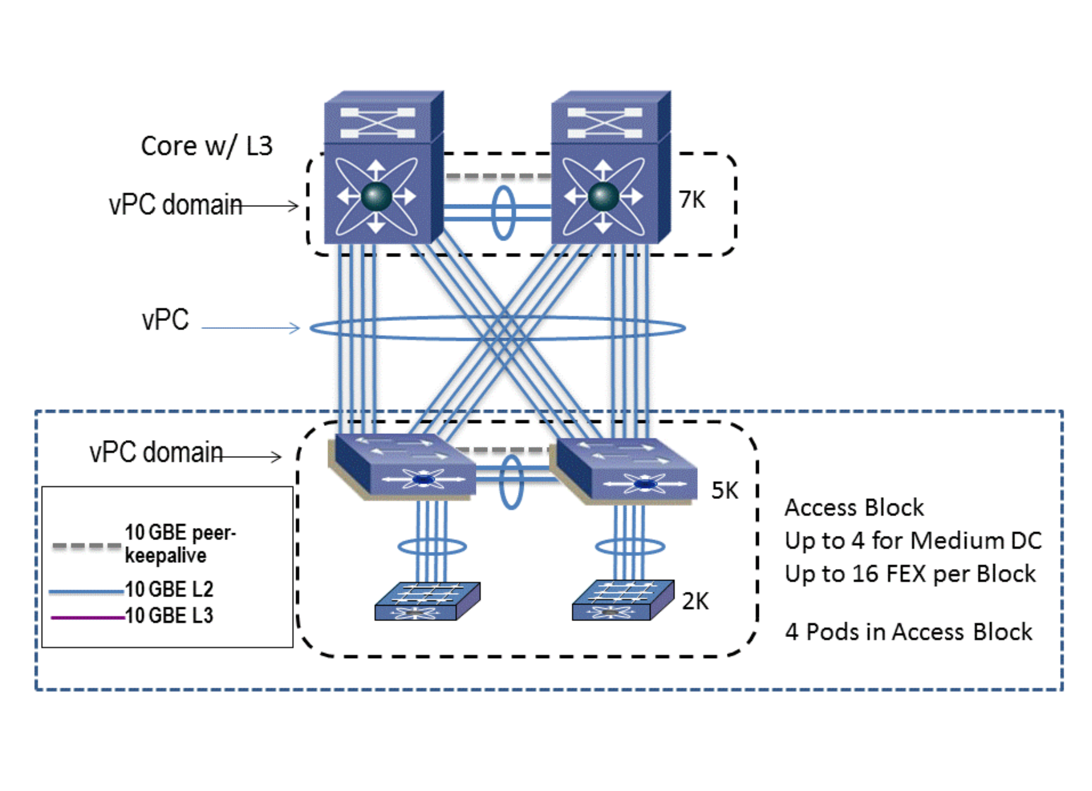


Figure 12-15 vPC Design for Medium Data Center

This topology superposes two layers of vPC domain and the bundle between vPC domain 1 and vPC domain 2 is by itself a vPC. This allows all of the links between the Nexus 5500s and Nexus 7000s to be active concurrently.

In the Medium Data Center design the following links are needed between the Nexus 7000s. These are depicted in Figure 12-16 on page 579.

- A vPC peer-link used to synchronize the state between vPC peer devices.
- A vPC peer-keep alive link is also needed between vPC peer devices.
- A dedicated L3 point-to-point link between the two Nexus 7000s is used to form L3 adjacency for the routing protocol and to pass traffic between them in the case of an L3 uplink/device down. The L3 link provides a backup routed path for the vPC domain in order to increase network resilience and availability.

L2/L3 routing boundary

The Nexus 7000 core layer is used for both active/active connectivity to the Nexus 5500 Access Block (s) and provides an active/active default gateway for L2/L3 boundary (i.e hosting interface VLAN and HSRP or VRRP feature).

M2 Series line cards provide scalable Layer 2, Layer 3, OTV capabilities. F2e Series line cards provide high-density cost-effective Layer 2/3 10-Gigabit Ethernet connectivity.

The vPC peer link can use F or M card type ports as long as both ends are identical, but it is better to use M ports, if available, because they support more MAC addresses.

HSRP/VRRP active/active with vPC

HSRP (Hot Standby Router Protocol) or VRRP (Virtual Router Redundancy Protocol) should be used to provide high availability for the servers IP default gateway.

The vPC domain at the core layer in the medium Data center design performs the L2/L3 boundary. Each vPC VLAN should be configured with an interface VLAN (or Switched Virtual Interface-SVI). HSRP or VRRP runs on top of this interface.

The Edge block will also be connected to the core layer via redundant Layer 3 connections The links to the Internet Edge, Campus/Legacy, and the WAN Edge block will be described in different sections. For additional details on the WAN edge block refer to 12.8, “WAN Edge Block Design for Small, Medium, and Large Data Centers” on page 615.

Network Service appliances

The Network Services appliances will also be attached to the core layer. Some customers may elect to have multiple pairs of Network service appliances and attach them at the aggregation layer or create a separate network services distribution block.

Server connectivity

Figure 12-16 depicts an xSeries Virtual server attached to the Medium Data Center design. The server has 2 x 10 Gbps ports for production and backups VLANs. It also has 2 x 1 Gbps ports for vCenter Consoles and an 1 Gbps port for an IMM port. Due to the implementation of vPC, the Nexus 7000 pair looks like a single switch and the Nexus 5500 pair look like a single switch from a STP perspective. All eight uplinks from each 5500 to the Nexus 7000 cores are active. Four cross links are used between the 5500s.

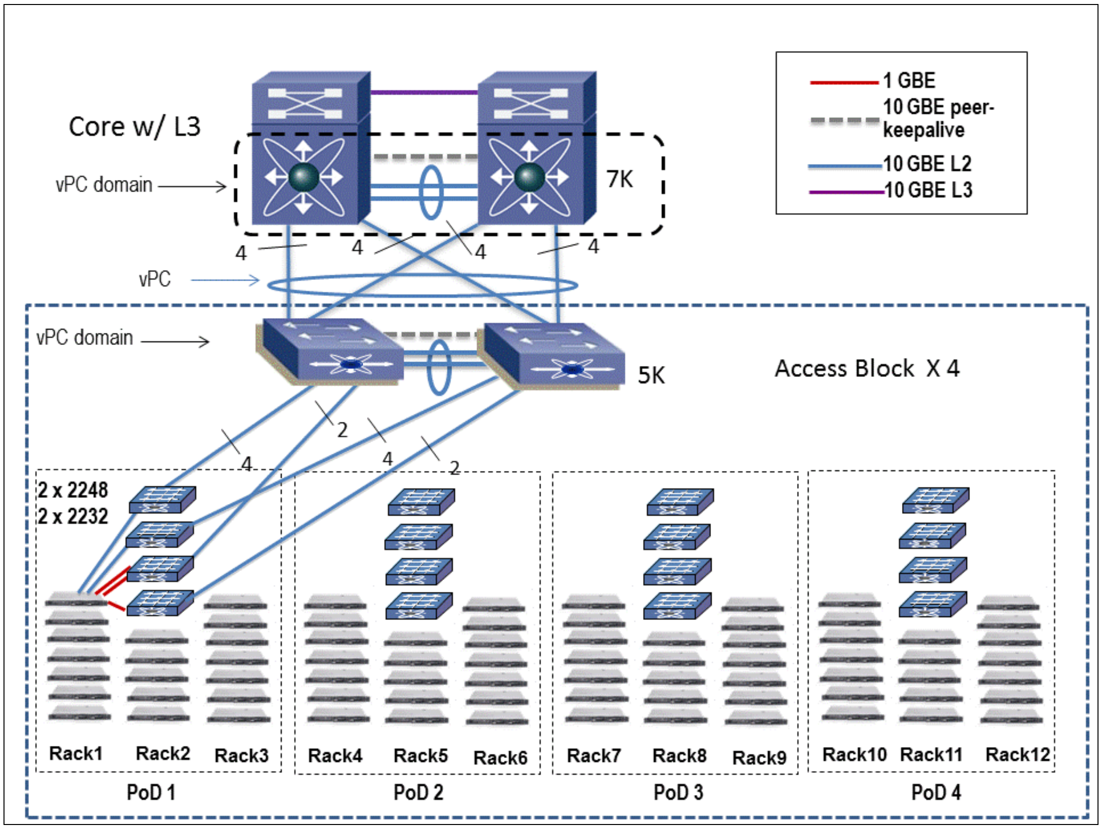


Figure 12-16 Medium data Center Design with Server PoDs

12.6.4 Service Zone design

The Medium Data Center use case design will include five Service Zones. This means that it will be possible to address different non functional requirements (for example availability) on different parts of the data center infrastructure. Figure 12-17 depicts the VLAN numbering scheme for the Medium Data Center design. Service Zones 1 through 4 are limited to a single Access Block. Service Zone 5 spans the entire data center. This hybrid approach to Service Zones allows creating VLANs to accommodate VM mobility anywhere in the data center while also allowing classical VLAN isolation per physical block as necessary.

The VLAN scheme allows for 250 VLANs per Service Zone. This number can be increased if needed.

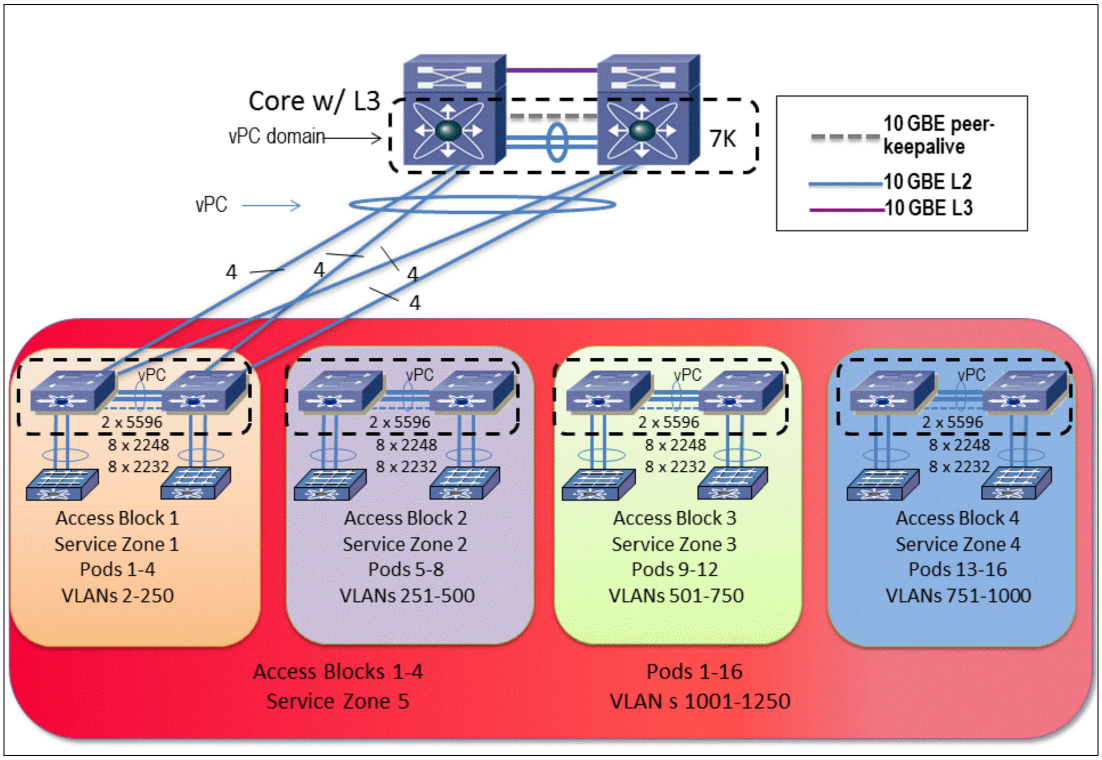


Figure 12-17 Medium Data Center design VLAN scheme

12.6.5 FCOE design of the Medium Data Center

Fibre Channel Over Ethernet (FCoE) is an option for this environment. Figure 12-18 depicts Servers attached via FCoE. The server will attach to the 2232 FEXs using CNA cards. The SAN switches will attach via Fibre Channel to the Nexus 5596s.

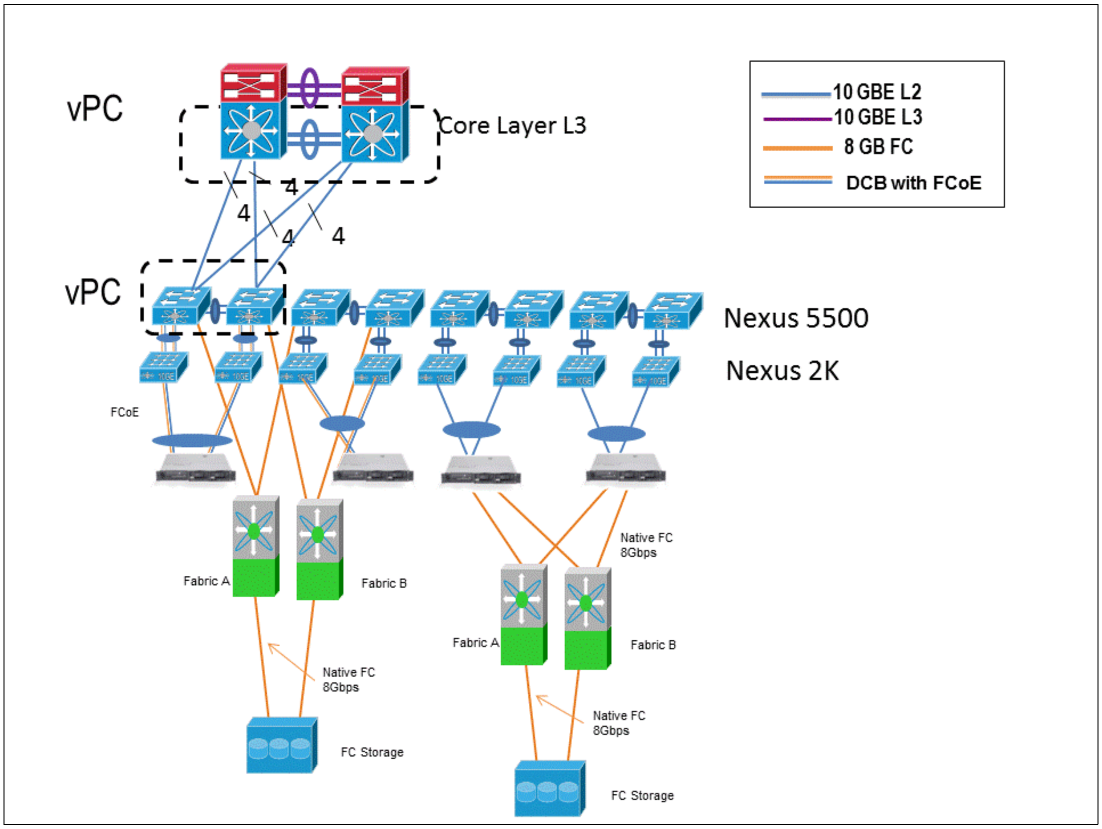


Figure 12-18 Medium Data Center design with FCOE

12.6.6 FabricPath design for the Medium Data Center

Figure 12-19 depicts the same Medium Data Center with FabricPath enabled instead of a classical Spanning Tree environment using vPC. Greater detail on FabricPath can be found in the next use case Large Data Center design.

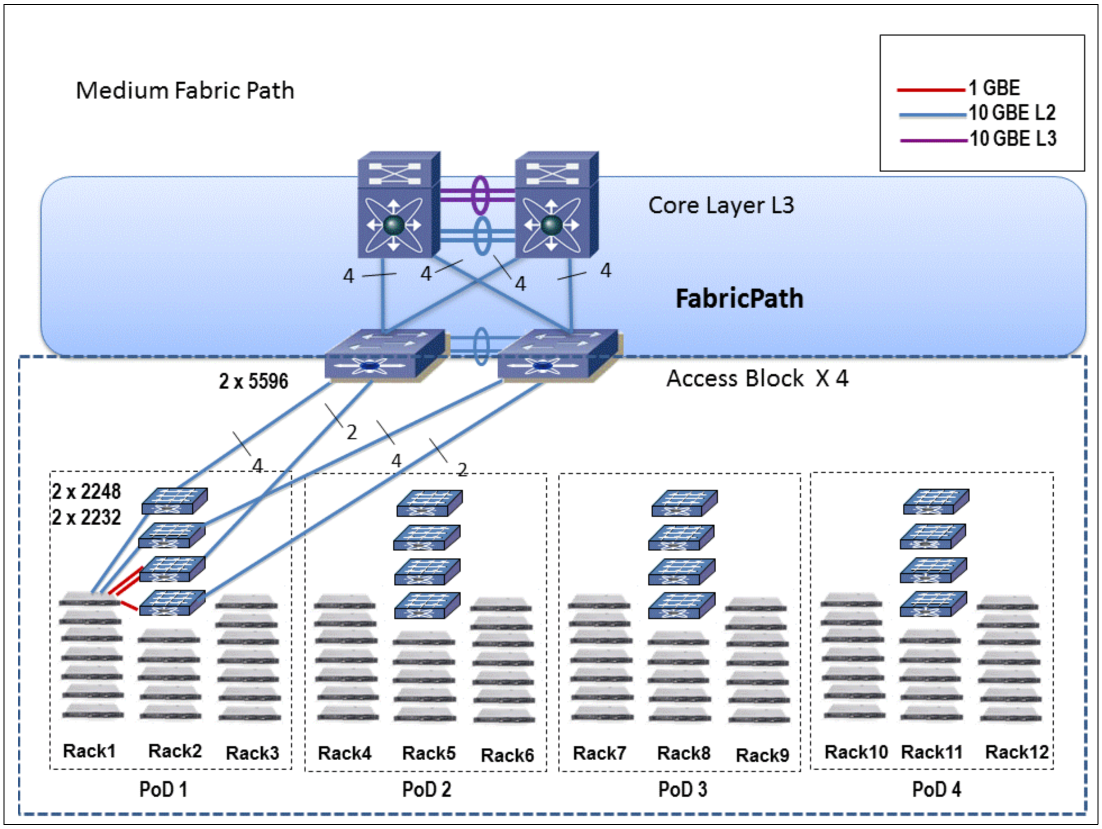


Figure 12-19 Medium Data Center with Fabric Path instead of spanning tree

12.6.7 FabricPath and FCoE for the Medium Data Center

Figure 12-20 depicts the Medium Data Center design with FabricPath and FCOE. Greater detail on FabricPath and FCOE can be found in the next use case Large Data Center design.

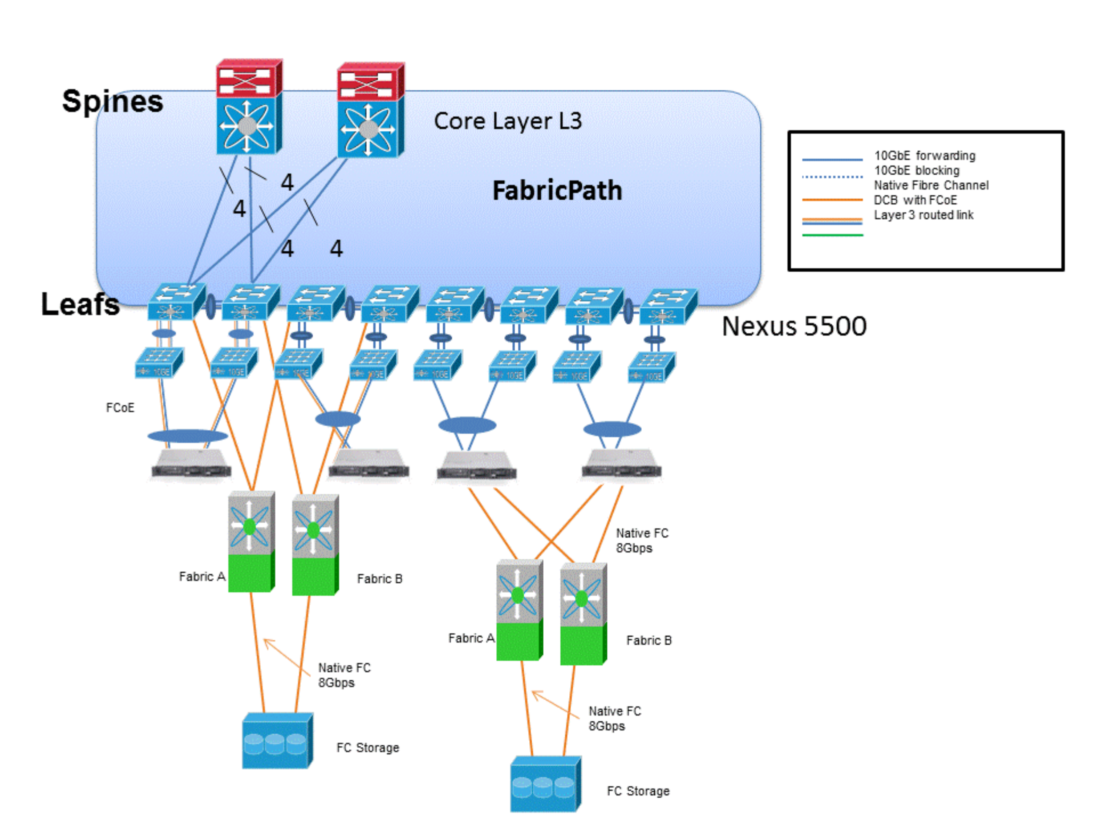


Figure 12-20 Medium Data Center with FCOE and Fabric Path

12.6.8 Design decisions for the Medium Data Center

The following considerations apply:

- The exact models and configurations of Nexus 7K, 5K, and 2K will need to be decided at the time of the design. Cisco regularly updates the product families with faster speeds and better density.
- The security policy for the network will dictate the need for internal security services. This use case depicts a a pair of firewalls in transparent mode to create a more secure internal zone. The firewalls could have also been configured in routed mode.
- A decision needs to be made on Spanning Tree, vPC, and FabricPath Design. Both FabricPath and vPC enable a loop free design. vPC is more familiar to those with traditional data center design and does not require additional software licenses. FabricPath will scale better and enable the data center to grow with ease.
- This use case depicts the network services attached directly to the core. This design optimizes cost. Many customer may elect to push network services out to the aggregation blocks.
- The number of server racks in the PoD design will determine the ability to drive the port utilization up on the PoDs. Many factors will determine the size of PoDs. The three critical factors are cabling standards for the DC, Power Density for the Racks, and NIC requirements for the server. This design uses a three rack PoD design to optimize the use of copper 10 Gbps cabling.
- Consistent FEX selection. Some server configurations require multiple10 Gbps Ethernet ports while some Pods only require a a single 1 Gbps Ethernet port for the servers. PoDs may be optimized for the exact server configurations or standardized with the two 2248 and two 2232 design shown in this use case.

12.6.9 PoD design

This section further describes the PoD design chosen for the Medium Data Center use case. We provide additional details on the number of racks, the physical layout, oversubscription and design decisions.

Number of racks

As described in the previous section the overall solution will be made up of 54 racks - 48 for server and access switches and 6 for the network core/aggregation layer, storage and network services devices.

Each PoD will be deployed using 3 racks with FEXs in the middle rack. This allows the reduction of the number of unused ports that would be associated with a 2 rack PoD design since more compute resources will be able to reach the FEXs. Using higher computing densities (for example, deploying BladeCenters), a 2-rack configuration may be worth exploring as an additional option.

Physical layout

Figure 12-21 shows the physical layout of the System x PoD, which is equipped with 19 System x Series Servers. Each x3850 Server is 4 RUs. The Nexus 2248 and 2232 switches are placed in the middle rack to minimize the distance from server port to network port.

- Each virtualized System x Series Server has 2x 10Gb and 3x 1Gb interfaces
- Each non-virtualized System x Series Server has
  - 2x 10Gb and 1x 1Gb interface, or
  - 5x 1Gb interfaces

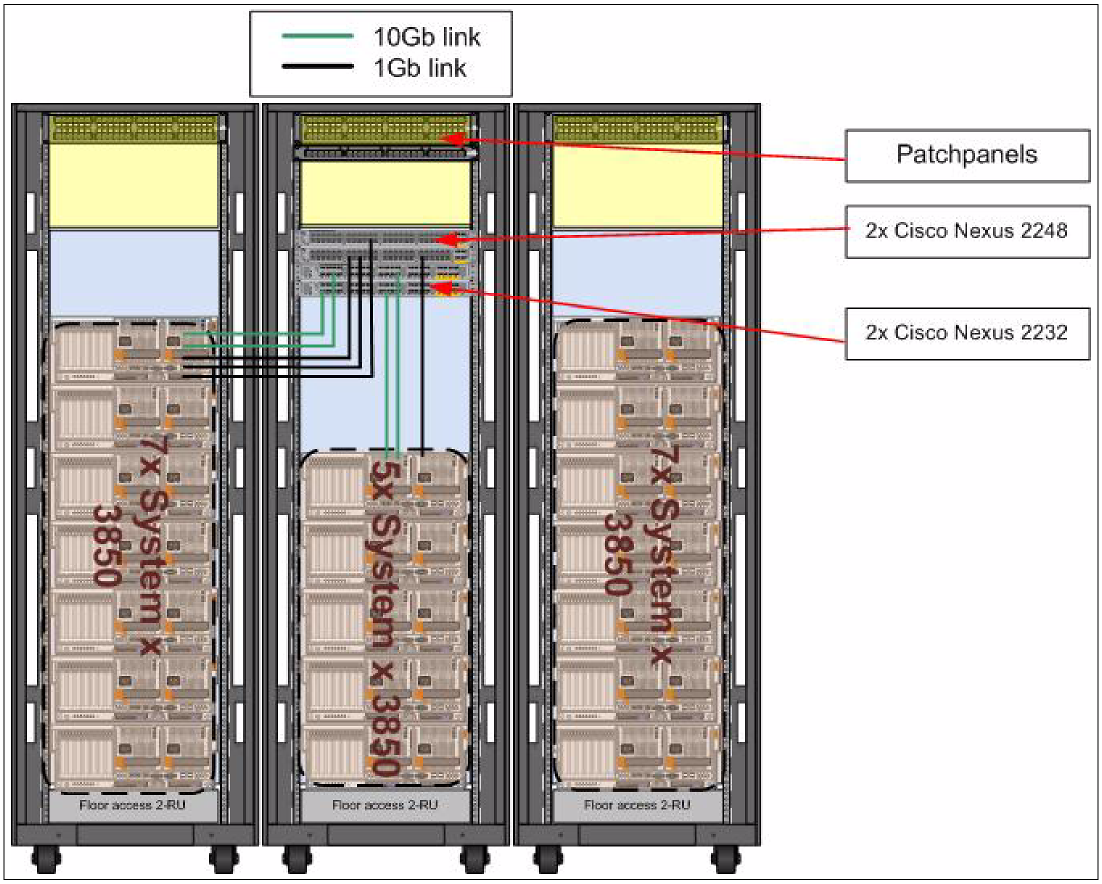


Figure 12-21 System x PoD

Figure 12-22 shows the physical layout of the System p PoD, which is equipped with 5 virtualized System p Series Servers. Each p770 Server has 4 CECs and is 16 RUs. The Nexus 2248 and 2232 switches are placed in the middle rack to minimize the distance from server port to network port.

Each virtualized System p Series Server has 8x 10Gb and 8x 1Gb interfaces.

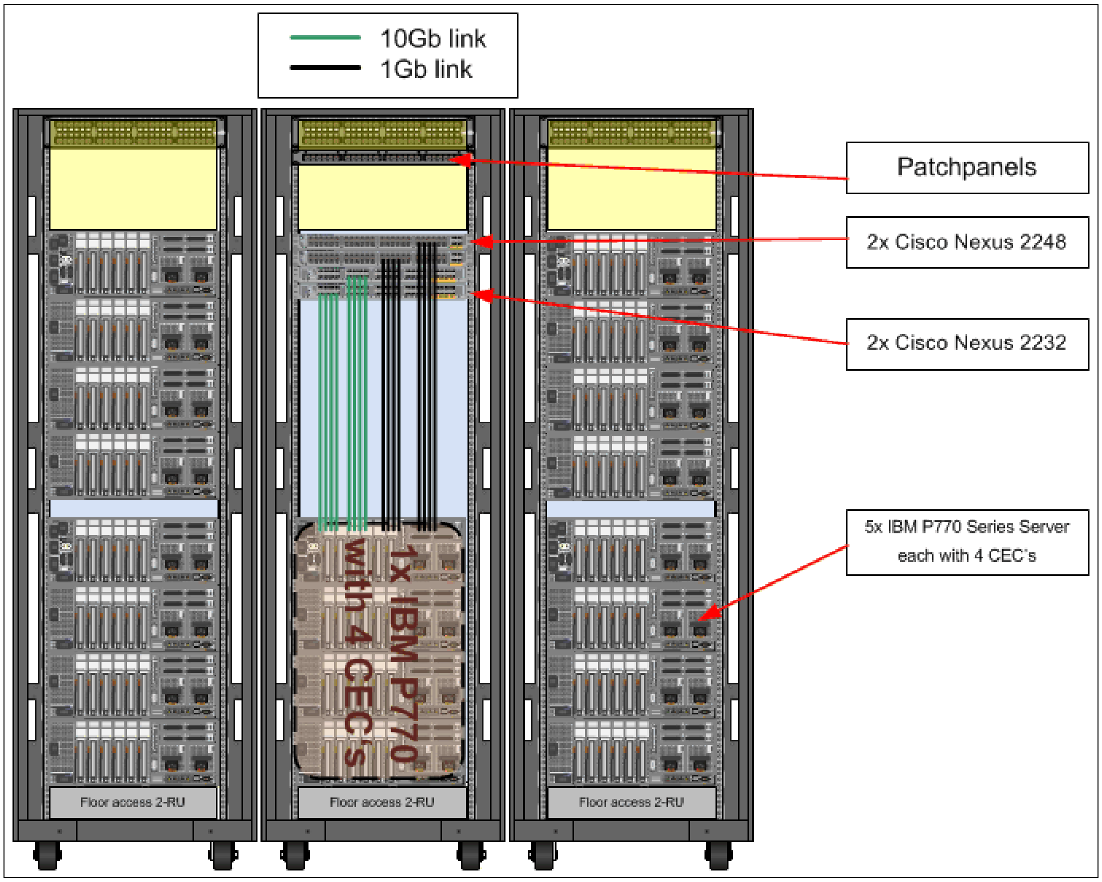


Figure 12-22 System p PoD

For illustration, the links are only depicted conceptually for one server. Independent of the type of PoD, the server links must be equally distributed over all Nexus 2k in order to achieve maximum redundancy.

The patch panels are used for the uplinks from the Nexus 2232 and Nexus 2248 switches to the Nexus 5596 switches. The 2248s will have 2x10 GE uplinks, while the 2232s will have 4x10GE uplinks. This will allow us to calculate the oversubscription ratios in the next section.

Oversubscription

The oversubscription ratio within the aggregation block is different for 10 GE ports and 1 GE ports and a System x PoD has a different oversubscription than a System p PoD based on the number of required interfaces.

Access Level Oversubscription

- Nexus 2248 in System x PoD - Apps:
  - Each Nexus 2248 will have 2x 10GE uplink ports to the Nexus 5596
  - Each Nexus 2248 will have 48 x 1GE used for System x servers
  - Oversubscription = 48/20 = 2.4:1
- Nexus 2232 in System x PoD - DB:
  - Each Nexus 2232 will have 4x 10GE uplink ports to the Nexus 5596
  - Each Nexus 2232 will have 19x 10GE used for System x servers
  - Oversubscription = 190/40 = 4.75:1
- Nexus 2248 in System x PoD - DB:
  - Each Nexus 2248 will have 2x 10GE uplink ports to the Nexus 5596
  - Each Nexus 2248 will have 10 x 1GE used for System x servers
- Oversubscription = 10/20 = 0.5:1
- Nexus 2232 in System x PoD - Virtualized:
  - Each Nexus 2232 will have 4x 10GE uplink ports to the Nexus 5596
  - Each Nexus 2232 will have 19x 10GE used for System x servers
  - Oversubscription = 190/40 = 4.75:1
- Nexus 2248 in System x PoD - Virtualized:
  - Each Nexus 2248 will have 2x 10GE uplink ports to the Nexus 5596
  - Each Nexus 2248 will have 24x 1GE used for System x servers
  - Oversubscription = 24/20 = 1.2:1
- Nexus 2232 in System p PoD - Virtual:
  - Each Nexus 2232 will have 4x 10GE uplink ports to the Nexus 5596
  - Each Nexus 2232 will have 20x 10GE used for System p servers
  - Oversubscription = 200/40 = 5:1
- Nexus 2248 in System p PoD - Virtual:
  - Each Nexus 2248 will have 2x 10GE uplink ports to the Nexus 5596
  - Each Nexus 2248 will have 20x 1GE used for System x servers
  - Oversubscription = 20/20 = 1:1

Aggregation level oversubscription

Each Nexus 5596 uses 8x 10 GE up links to the core and has 24 x 10 Gbps downlinks to the Nexus 2ks so that the oversubscription ration is 3:1 for each Nexus 5596.

The calculated oversubscription ratios above are considered acceptable as not all virtual or physical servers will be using the links at all times and the 1 GE connections which are used mainly for systems and network management purposes are generating only limited traffic volumes.

Adding additional servers and/or adding 10 GE interfaces to existing servers will increase the oversubscription ratio, but in order to improve oversubscription all 8x10 GE uplinks can be used from the Nexus 2232 and all 4x10GE uplink can be used from the Nexus 2248.

Also as required to control and reduce the oversubscription, more links between the Nexus 5596 and the core switches (Nexus 7K) can be added.

Network Services

The Medium Data Center use case will include different network services. Figure 12-23 on page 589 depicts the load balancing and firewall solutions.

The Network Services appliances will be attached to the core layer. Some customers may elect to have multiple pairs of Network services and attach them at the aggregation layer or create a separate network services distribution block.

The Network service device can be of any type. It can be a switch, a server, a firewall, a load-balancer or a NAS for example.

Network service devices may be attached in a Layer 3 routed mode or may be attached in transparent mode. Cisco ASA firewalls can be configured in transparent or routed mode. Both modes are supported for integrating ASA firewalls when attached to a vPC domain.

The following are some examples of network services:

- Firewalling and Intrusion prevention (internal security services): ASA will be deployed in Active/Standby mode.
- Load balancing and application delivery: A load balancer will be deployed in Active/Standby mode via Layer 3 using Source NAT. Active/Active deployment is also an option and it is the preferred option in case of multi-tenancy requirements using load balancing virtual contexts.
- WAN acceleration: WAAS will be deployed in Active/Active mode via Layer 3.
- DNS and DHCP IP services
- Cisco ISE/ACS are optional components for AAA
- Cisco NAM is an optional component for traffic analysis
- Cisco DCNM, Tivoli NM, NCM, and OMNIbus are optional components for data center network management.

12.6.10 Interconnecting the new and legacy data centers

Most of the time when a new data center architecture is implemented it must be connected to an already existing data center environment. In the context of the Medium Data Center described in this chapter the engineering team can take advantage of the Virtual Device Context (VDC) feature of the Nexus 7000 switch to interconnect the new and legacy data centers. VDCs were described in Chapter 9, “NX-OS network operating system” on page 481.

A new VDC is created in each Nexus 7000 core switch to act as a Legacy Connection Router. Each VDC is created with 12 x 10 Gbps M2 linecard ports for full Layer-3 and security capabilities. F2 linecard ports could be used as well, but since we plan to implement OTV (to be discussed later in this chapter) M2 linecards are required and can be split between the VDCs without adding any extra costs. This will allow for redundant Layer 3 connectivity between the environments as shown in Figure 12-23.

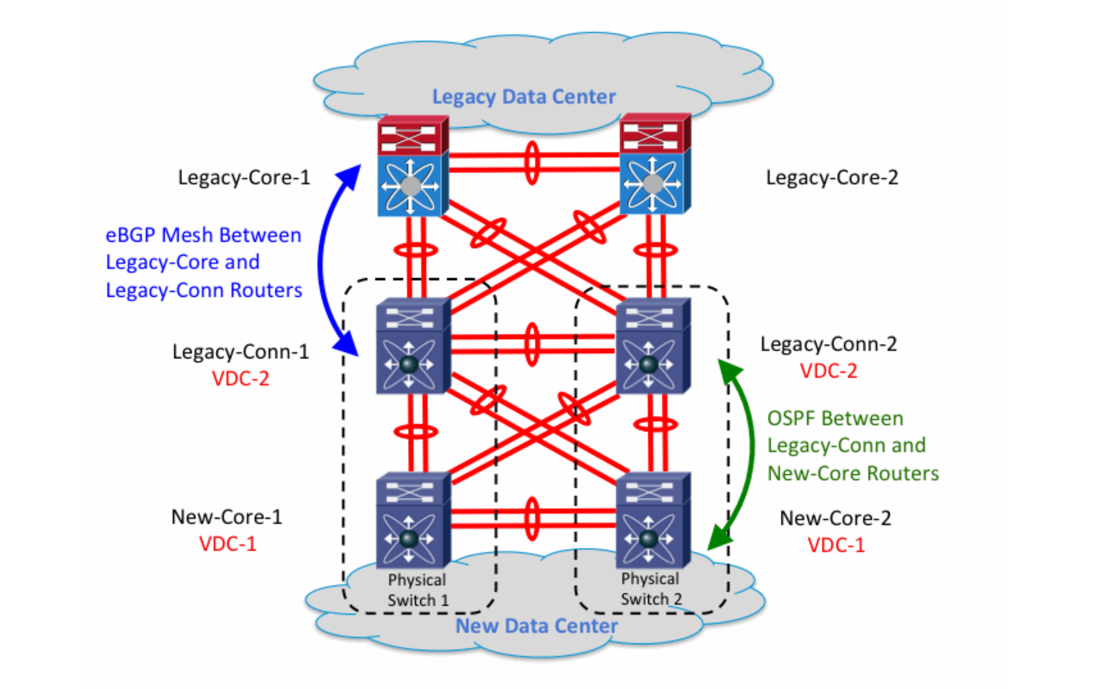


Figure 12-23 Connectivity between the new and legacy data center environments

Physical connectivity

As shown in Figure 12-23, each Nexus 7000 core is split into VDCs. The core VDC is connected to the Nexus 5500 as described earlier in this chapter and provides the layer-3 edge for all of the new data center VLANs using SVIs.

The Legacy Connectivity VDCs are created with 12 10 Gbps interfaces and are connected as Layer-3 EtherChannels to both New Core routers as well as to both Legacy Core routers and to each other. This simplifies routing configurations and provides additional bandwidth as required.

> **Note:** While using some routing protocols, such as BGP, interface bandwidth is not taken into account therefore other methods of failover or traffic redirection may be required when using layer-3 EtherChannels. For example you can setup eBGP neighborships using loopback interfaces, and use static routes to the neighbors loopback address across several single layer-3 interfaces to load balance traffic between the eBGP neighbors. Also, if multiple paths exist, you can use the minlink feature for the EtherChannel and have it go down if a physical link goes down. In more complex cases, you can use Embedded Event Manager (EEM) scripts to affect routing in case of a link failure.

The Legacy Core routers do not have to be the Layer-3 Edge for the existing data center, they must however have layer-3 reachability to all of the subnets of the legacy environment.

The New Core routers are connected to each other to provide a secondary path in case of a Legacy Connectivity router failure, as are the Legacy Core routers.

Routing between the legacy and new data centers

The routing protocols running to connect the legacy and new data center environments are depicted in Figure 12-23 on page 589. eBGP is running between the legacy core routers and the legacy connection routers in a full mesh. This means that Legacy-Core-1 has an eBGP neighborship to both Legacy-Conn-1 and Legacy-Conn-2 and so on. iBGP is running between Legacy-Core-1 and Legacy-Core-2 as one AS, and Legacy-Conn-1 and Legacy-Conn-2 as another AS.

eBGP was chosen since it allows for very granular and simple filtering of routes between the two environments. It also allows us to use network statements to easily redistribute OSPF routes to our eBGP neighbors. This is true for other routing protocols that could be running in the legacy environment.

OSPF is running between the new core and legacy connection routers using point-to-point OSPF interfaces. Since the Medium Data Center design discussed in this chapter is relatively small from a routing protocol perspective, we can use a single OSPF area. We redistribute eBGP routes from the legacy data center to the new data center core routers on the legacy connection routers, using appropriate metrics.

Separation between the new and legacy environments

With this inteconnections design, we can easily install firewalls between the two environments, or even use simple ACLs on the legacy connection routers. Since all of the links are layer-3 there are no spanning-tree issues to be concerned about.

If multiple VRFs are required in each requirement for separation of traffic this design can still be utilized. The Layer-3 point-to-point interfaces would need to be changed to Layer-3 trunks (using sub-interfaces), each sub-interface would have to be added to a separate VRF and the routing protocols would need to be configure accordingly.

12.6.11 Extending data center VLANs using OTV

Overlay Transport Virtualization (OTV) is a technology that allows extending Layer 2 VLANs across distributed data centers. OTV is covered in depth in Chapter 8, “Overcoming the limitations of traditional data center networks” on page 421. In this section, we describe how we would extend some of the VLANs of our new Medium Data Center design to a second data center.

Just as with the legacy connectivity described earlier in this chapter, we use the Virtual Device Context (VDC) feature to create a new device to be used for OTV functionality. In each New Core Nexus 7000, we create a VDC with 4 M2 10 Gbps ports (we only really need two ports but have allocated extras for future growth). Once the new OTV VDCs are created the architecture will look like Figure 12-24.

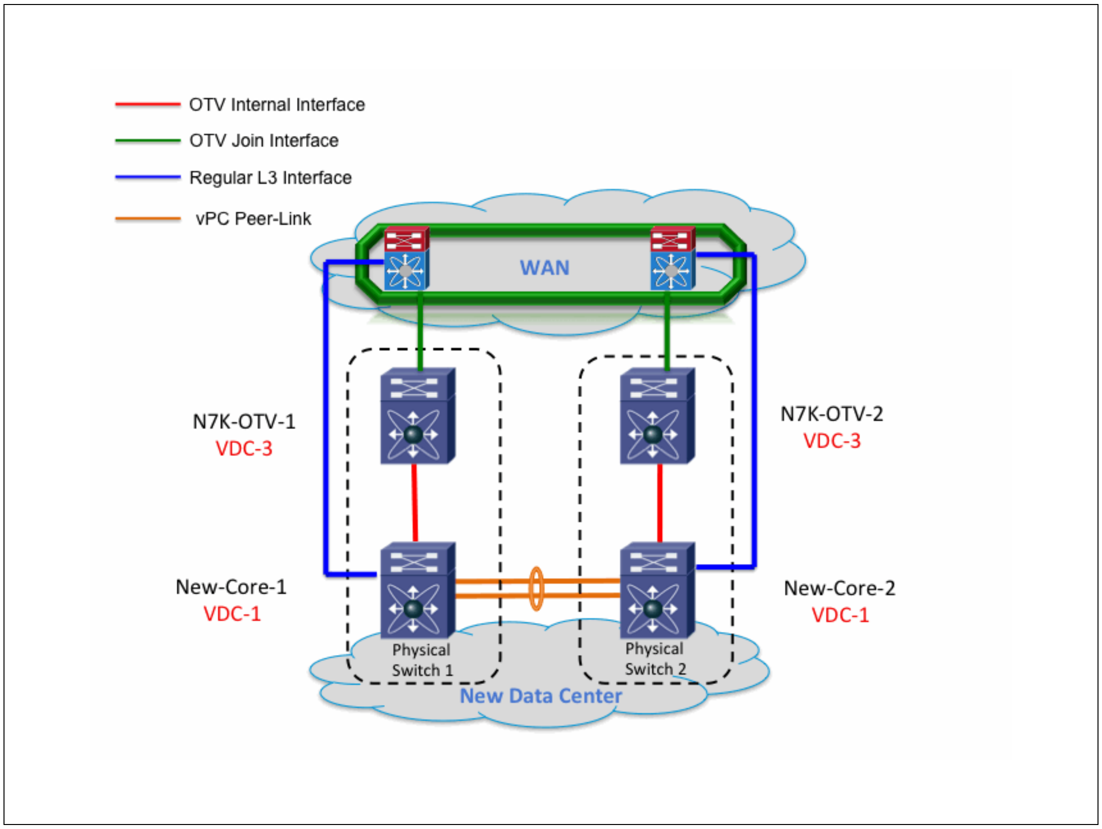


Figure 12-24 Using OTV to extend VLANs across the WAN

Physical connectivity for OTV

Since an M1 or M2 linecard is required for the OTV join interface functionality, as described in chapter 8, in the Medium Data Center design, we decided to use a N7K-M224XP-23L for each Nexus 7000 chassis. That means there are 24 total 10 Gbps M2 ports per chassis. Some of which we can use for our Legacy Interconnect VDC as described earlier. We can leave the entire Core VDC with only F2 ports, and distribute the M2 ports among the other VDCs. If all you have are regular F2 linecards (the non E models), then you cannot mix them with M series linecards, and assigning the M2 linecard to the other VDCs is the best option. When we create the OTV VDC, we assign 4 10 Gbps M2 ports to it. As shown in Figure 12-24 the OTV VDCs will be connected to the Core VDCs in the same physical switch.

The connections between the OTV and Core VDCs are configured as regular Layer 2 trunk ports. An OTV site VLAN should be configured and allowed on the trunks along with all the VLANs that will be extended between the OTV sites. All of these VLANs should be allowed to cross Layer 2 port-channel (this can be the vPC peer-link) between the two New Core VDCs. We are using a single 10 Gbps internal interface per OTV VDC because the WAN between the two data centers we are extending the VLANs between will not be able to carry more than 10 Gbps of traffic.

The OTV Join Interfaces for both OTV VDCs are configured as Layer-3 and connect to the separate WAN routers, if possible, to provide higher availability. They can be configured for dynamic or static routing depending on how the WAN router is setup. See Chapter 8 to learn more about using multicast or Unicast Adjacency Servers to create OTV adjacencies and extend the Layer 2 domain.

12.6.12 Logical view of Nexus 7000 VDCs

As described in chapter 8, VDCs partition a single physical device into multiple logical devices that provide fault isolation, management isolation, address allocation isolation, service differentiation domains, and adaptive resource management. You can manage a VDC instance within a physical device independently. Each VDC appears as a unique device to the connected users. A VDC runs as a separate logical entity within the physical device, maintains its own unique set of running software processes, has its own configuration, and can be managed by a separate administrator. Since VDCs are used as part of the Medium Data Center design, each physical Nexus 7000 chassis is made up of three logical devices as depicted in Figure 12-25.

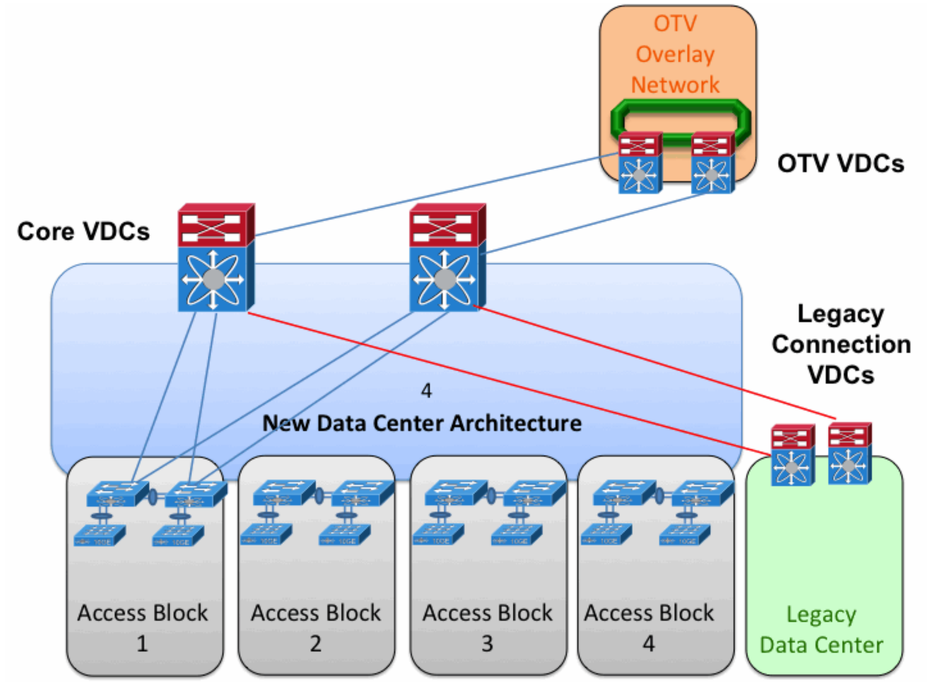


Figure 12-25 VDC Logical Devices

The three VDCs used in this design are Core, Legacy Connection and OTV. Each one plays a different and critical role and as such is a separate logical entity. One thing that needs to be taken into consideration is that all 3 VDCs in a single physical chassis run the same version of NX-OS code, and as such all three are affected in the case of a software upgrade or downgrade. While Nexus 7000s with VDCs can still do In Service Software Upgrades (ISSU) it is important to understand the operational and change management requirements when upgrading the physical device, since it affects 3 logical devices at once.

12.6.13 Storage and storage network design for Medium Data Center use case

In the Medium Data Center use case, Fibre Channel over Ethernet (FCoE) as well as traditional Fibre Channel is used to provide connectivity from the Servers to the Storage Arrays. The storage arrays used on the Medium Data Center only have Fibre Channel interfaces.

The Storage requirements and solution to be used on the Medium Data Center use case is the same as described for the Large Data Center use case. For the details on the solution, see 12.5.6, “Storage and storage network design” on page 571.

12.7 Large data center use case

In this section, we describe the Large Data Center use case.

12.7.1 Definition/introduction

The Large Data Center use case will have the following characteristics (see Table 12-1 on page 563 for more details):

- Max number of 1 G access ports: 6144 if only Nexus 2248 is used on the ToR
- Max number of 10 G access ports: 4096 if only Nexus 2232 is used on the ToR
- Number of FEXs (Nexus 2232 or 2248): 128
- Number of aggregation blocks: 8
- Number of Core devices: 4 Nexus 7010
- Number of PoDs: 32
- Number of Service Zones: 3
- FabricPath support
- vPC support
- FCoE and native FC support
- Legacy integration
- No Layer 2 extension (Single Site)
- Optional Nexus 1000v support
- Network Services Integration, including Security and Load Balancing.
- WAN Connection.

12.7.2 Assumptions

This section introduces the most important assumptions for this use case. Those will be broken down into Servers, Network, Storage, Security and Legacy integration assumptions.

The assumption section represents the requirements we used when designing the Data Center solutions for this use case; the assumptions do not represent the maximum capacity or scale of any of the areas within the data center, including servers, network, storage.servers

The following assumptions are used for server components.

Number of servers

A large enterprise data center typically has a mix of x86 servers and AIX servers. In this use case, the x86 will be an IBM x3850 Series Server (4RUs) and the AIX server will be an IBM System p770 Series Server with 4 CECs (16RUs). For this use case, the following numbers of servers are used:

Number of servers per PoD:

Each physical PoD has three consecutive racks and is either fully equipped with System x Series servers or with System p Series servers.

System x Series server PoD, each equipped with:

- Number of virtualized System x Series servers: 14
- Number of non-virtualized System xSeries servers: 5

System p Series servers PoD each equipped with:

- Number of virtualized System p Series servers: 5

Number of servers per Access Block:

Each Access Block combines 3 System x PoDs and 1 System p PoD with the following number of servers:

- System x Series servers: 57 (42 virtualized and 15 non-virtualized)
- System p Series servers: 5

Number of servers in the data center:

The complete data center network design combines 8 Access Blocks with the following number of servers:

- System x Series servers: 456 (336 virtualized and 120 non-virtualized)
- System p Series servers: 40

Number of NICs (LAN interfaces) and speed

We assume a mix of virtualized and non-virtualized System x Series Servers. The System p Series servers are all virtualized in this use case. The following number of NICs are required for LAN connectivity only:

Virtualized System x:

- 2x 10 GE as Production/Admin/Backup and Recovery
- 2x 1 GE for VCC (Hypervisor Management)1x 1 GE for RSA/IMM

Non-virtualized System x:

- 2x 10 GE as Production/Admin/Backup and Recovery
- 1x 1 GE for RSA/IMM

Virtualized System p with 2 LAN VIOs:

- 4x 10 GE for Production/Admin
- 4x 10 GE for Installation, Backup and Recovery
- 2x 1 GE for VIO Management

- 2x 1 GE for VIO Installation, Backup, and Recovery
- 4x 1 GE Service Processor interfaces for redundant HMC connectivity

With the NIC numbers above, the following NIC numbers can be calculated:

- Virtualized System x NIC numbers:
  - 10 GE ports per PoD: 28, per Access Block: 84, per data center: 672
  - 1 GE ports per PoD: 42, per Access Block: 126, per data center: 1008
- Non-Virtualized System x NIC numbers:
  - 10 GE ports per PoD: 10, per Access Block: 30, per data center: 240
  - 1 GE ports per PoD: 5, per Access Block: 15, per data center: 120
- System p NIC numbers:
  - 10 GE ports per PoD: 40, per Access Block: 40, per data center: 320
  - 1 GE ports per PoD: 40, per Access Block: 40, per data center: 320
- Total number of NICs used in the data center: 2680
- Number of images

  The maximum number of virtualized images per physical server is the following:

  - Virtualized System x Series Server: maximum of 80 VMs
  - Virtualized System p Series Server: maximum of 50 LPARs

  The maximum number of virtualized images per Access Block is the following:

  - Virtualized System x Series Server: maximum of 3360 VMs per Access Block
  - Virtualized System p Series Server: maximum or 250 LPARs per Access Block

  Total number of virtual images of the whole data center fabric: 28880

  Total number of physical images of the whole data center fabric: 120

  Total number of virtual and physical images on the data center fabric: 29000

- Server platform mix

  A PoD is always either installed with System x or System p Series Servers, but never a mix of both. An Access Block is a combination of one System p PoD and three System x PoDs.

Network

We use the following assumption for the data center LAN component:

- Oversubscription:

  For the initial setup, a reasonable oversubscription is acceptable (this design will target a max 5:1 per network device and 30:1 for worst case scenario for end-to-end oversubscription) as it is very unlikely that the servers will use all the available link capacity. Potential future bottlenecks identified via a proper capacity management can be mitigated by adding links, which reduces the oversubscription.

- Traffic flows:

  Approximately 70% of the traffic is expected to be East<->West so remains local within the data center and 30% of the traffic is North <-> South and leaves the data center to the WAN (Wide Area Network)

- Number of ports and spare ports:

  With 128 FEXs the large use can scale up to 4096 10 GE ports or 6144 1 GE ports. In this specific use case, we include 8 Access Blocks, each one with 4 PoDs with a mix of 10 GE and 1 GE network ports. Overall this will provide 2048 10 GE ports (60% port utilization) and 3072 1 GE ports (47% port utilization).

  Those limits are not dictated by the hardware limit of the platforms selected (Nexus 5596, Nexus 2232 or 2248) but instead those limits have been selected to provide an example of a network that takes into consideration more than the maximum hardware limit of the switches (that is, power consumption and rack space)

- Growth:

  We include 4 PoDs per Access Block and 8 Access Blocks in the data center. The number of uplink interfaces can be increased and further Access Blocks can be added. That will leave space for growth in terms of additional servers or additional ports to accommodate increase I/O requirements and / or an increase on VM density per physical server. The leaf switches Nexus 5596s and the spine switches Nexus 7010 have spare ports for future growth available.

- Power/Rack:

  The maximum power allowed per rack will be 12kW. The power consumption of the active equipment installed in the racks has to be kept below this limit.

Storage

We assume that there is a mix of FCoE and classical FC required. The network and storage design will provide three different storage capabilities. The first network part will use FCoE and DCB up to the Nexus 5596 to utilize the cost reduction benefits. A second part of the network and storage design will be based on classical FC and a third part of the network and storage design will enable a mix of FCoE and FC.

Security

Both, external and internal security services are required. All physical security devices and appliances will be connected to a dedicated Service Block, which is also used as connectivity point for other network services and WAN routers.

Legacy integration

We assume that the Large Data Center use case will require a connectivity point for the integration of legacy network devices into the new infrastructure. Layer 2 and Layer 3 capabilities between the new and legacy infrastructure are required for server migration scenarios.

12.7.3 Design concept

This section provides an overview of the Large Data Center use case design, elaborating on the design concept and the design decisions.

Overview

Figure 12-26 shows the Large Data Center use case design concept, based on the aforementioned assumptions. The design will include 8 Access Blocks and each Access Block will consist of 4 PoDs. Each PoD will have 3 racks for servers. An additional Service Block will be introduced as connectivity point for legacy network devices, security, Layer 2 and 3 Boundary, network services and WAN routers.

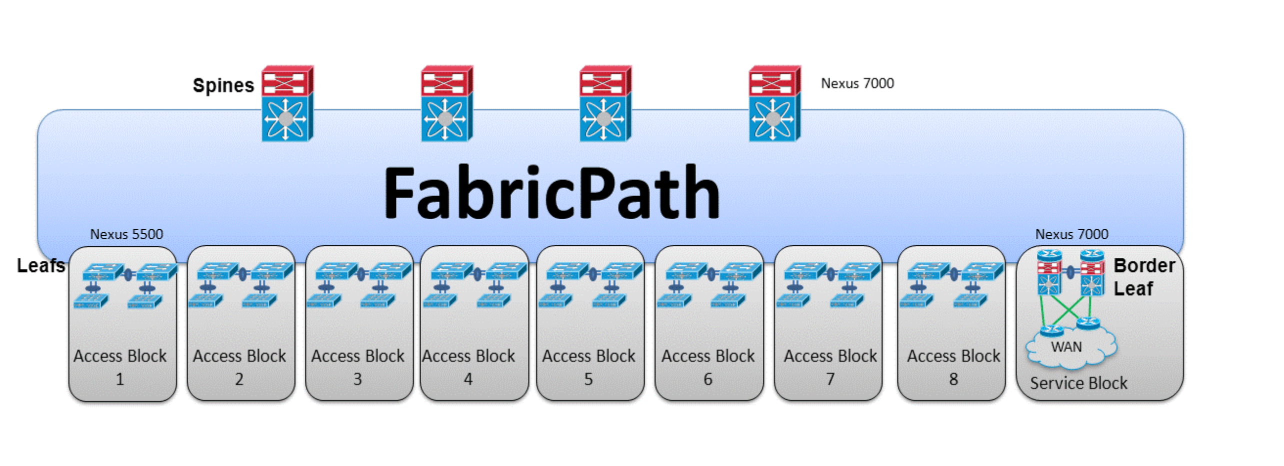


Figure 12-26 Large Data Center architecture overview diagram

Fabric Path technology will be used to design a data center wide network fabric. The spine switches are four Cisco Nexus 7010 switches, whereas the leaf switches are Nexus 5596. The border leaf will be Cisco Nexus 7010 switches. The spine and leaf switches have Layer 2 functionality only, because the border leaf switches will provide the Layer 3 functionality.

As shown in Figure 12-27, each Nexus 5K leaf switch has 8 uplinks to the Nexus 7K spine switches with two channelized links to each Nexus 7k. The Nexus 7k switch used as border leaf will have a 8-port channelized links to each Nexus 7k spine switch, means overall 32 connections per Nexus 7k border leaf to the spine switches.

The Nexus 7k switches used as Border Leaf will provide the connectivity point for legacy network devices depicted as Catalyst 6500 switches with VSS functionality and for Service Switches/Appliances. All the connected switches can utilize the vPC+ feature. The WAN routers for Internet connectivity and/or dedicated leased lines will be connected via redundant Layer 3 connections.

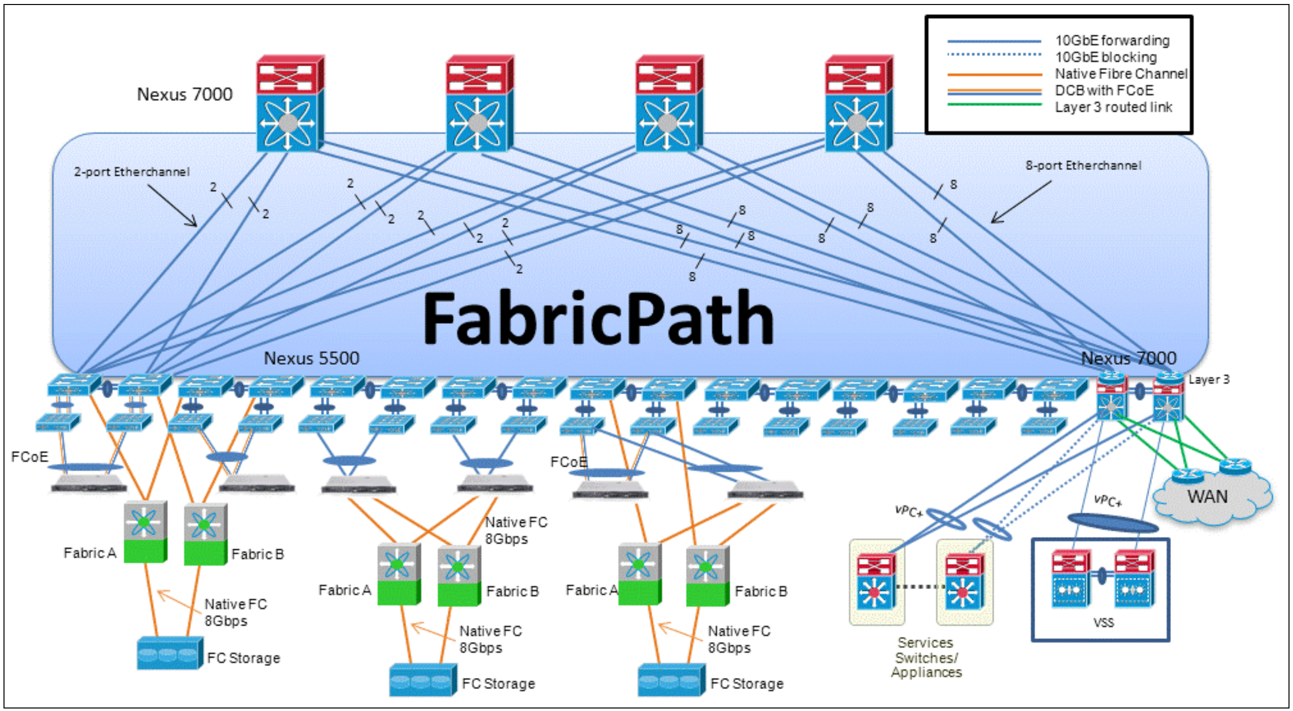


Figure 12-27 Large Data Center network design

Cisco Nexus 2000 Fabric Extenders are used as ToR in the PoD design to provide 1 GE and 10 GE network ports. Each Nexus 2k is only connected to one Nexus 5k parent switch. The enhanced vPC design can also be used instead especially in cases where single attached servers are required.

As depicted in Figure 12-27 the servers labeled with FCOE uses FCoE from the Server up to the Nexus 5k as the SAN technology. The SAN switches are connected to the Nexus 5k, which has native FC ports available. Servers connected to Access Block 3 or Access Block 4 will require additional HBAs for FC connectivity, because it is not intended to use FCoE. Any other server connected to Access Block 5-8 can either use FCoE or native FC, because both technologies are supported.

Design decisions

Here we discuss considerations in making design decisions:

- Cisco FabricPath with its leaf and spine architecture creates a scalable data center fabric ideally for this Large Data Center use case. It has multiple advantages which includes:
  - Spreading risk from two core devices to multiple devices. The Leaf/Spine architecture is constructed so that failure of a spine switch will cause minimum traffic disruption due to the use of Equal-cost multi-path routing (ECMP) and the very fast (sub-second) convergence capabilities of FabricPath. Also on this architecture with four spines, if one of the spine devices fail, this would result in only 25% reduction on the network capacity, compared with 50% reduction on the traditional model with just two aggregation/core switches.
  - Increased capacity in the core via Layer 2 Multipathing which is essential for increased East-West traffic in the data center
  - STP free design with increased availability
  - Increased VM Mobility radius based on large Layer 2 domains

  - Easier services provisioning with any VLAN anywhere
  - Conversational Learning – A feature in FabricPath that enables the MAC address tables of the switches to be used more effectively by not learning unnecessary mac-addresses. This effectively means that an unlimited number of mac-addresses can be connected to the fabric.
- Cisco FabricPath plus Layer 3 integration at the border leaf
  - Provides a “cleaner” spine design.
  - Traffic distributed equally across spines (no hot spot)
- Cisco Nexus 7010 as border leaf to provide the links to the spine switches plus all the additional links to the legacy network devices, service switches/appliances and WAN devices.
- Spare ports on all devices are available in order to increase the number of links based on capacity demands and scalability requirements to improve oversubscription.

12.7.4 PoD design

This section further describes the PoD design chosen for the large Data Center use case scenario. We provide additional details on the number of racks, the physical layout, oversubscription and design decisions.

Number of racks

The Large Data Center use case has 8 Access Blocks for server connectivity. As shown in Figure 12-28, each Access Block is a combination of 4 PoDs.

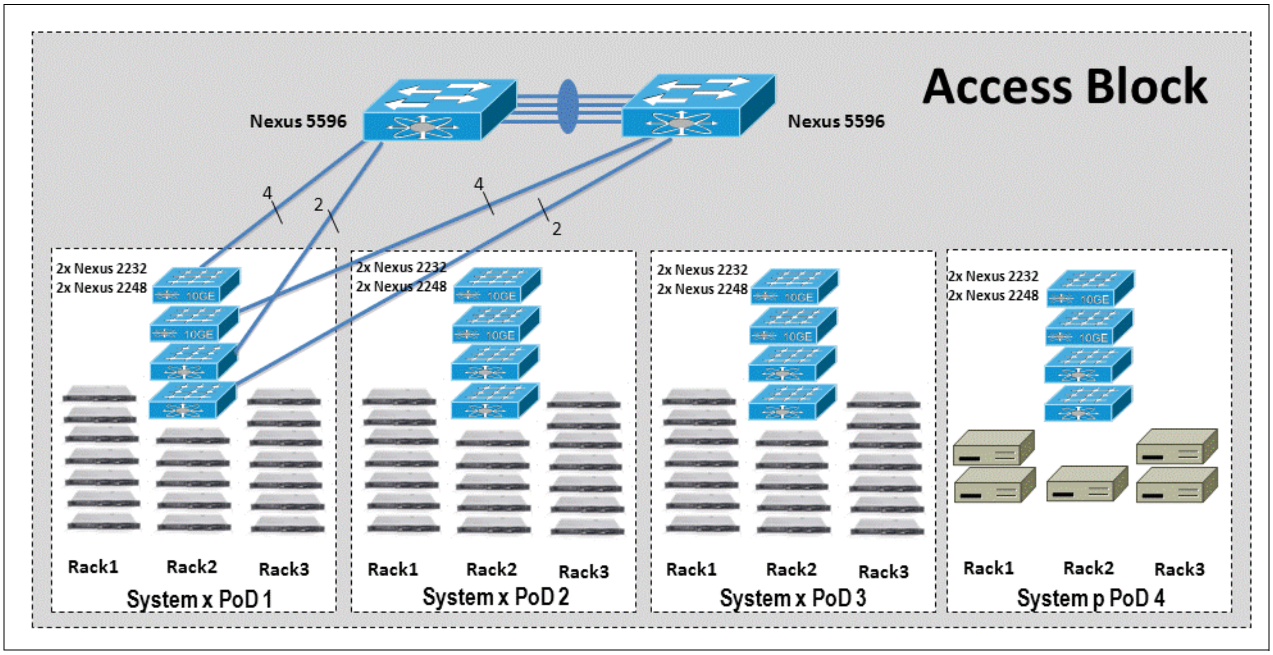


Figure 12-28 Access Block design

Each PoD will be deployed using 3 racks for servers with Nexus 2k access switches in the middle rack. Using higher computing densities per rack (for example 1RU servers) a 2 rack configuration may be worth exploring as an additional option. Overall there are 12 server racks required for the server placement in an Access Block. All other devices like the spine/leaf switches, storage and network services devices are in separate network racks which are required additionally.

Physical layout

Figure 12-29 shows the physical layout of the System x PoD, which is equipped with 14 virtualized System x Series Servers and 5 non-virtualized System x Series Servers. Each x3850 Server is 4 RUs. The Nexus 2248 and 2232 switches are placed in the middle rack to minimize the distance from server port to network port.

- Each virtualized System x Series Server has 2x 10Gb and 3x 1Gb interfaces
- Each non-virtualized System x Series Server has 2x 10Gb and 1x 1Gb interface

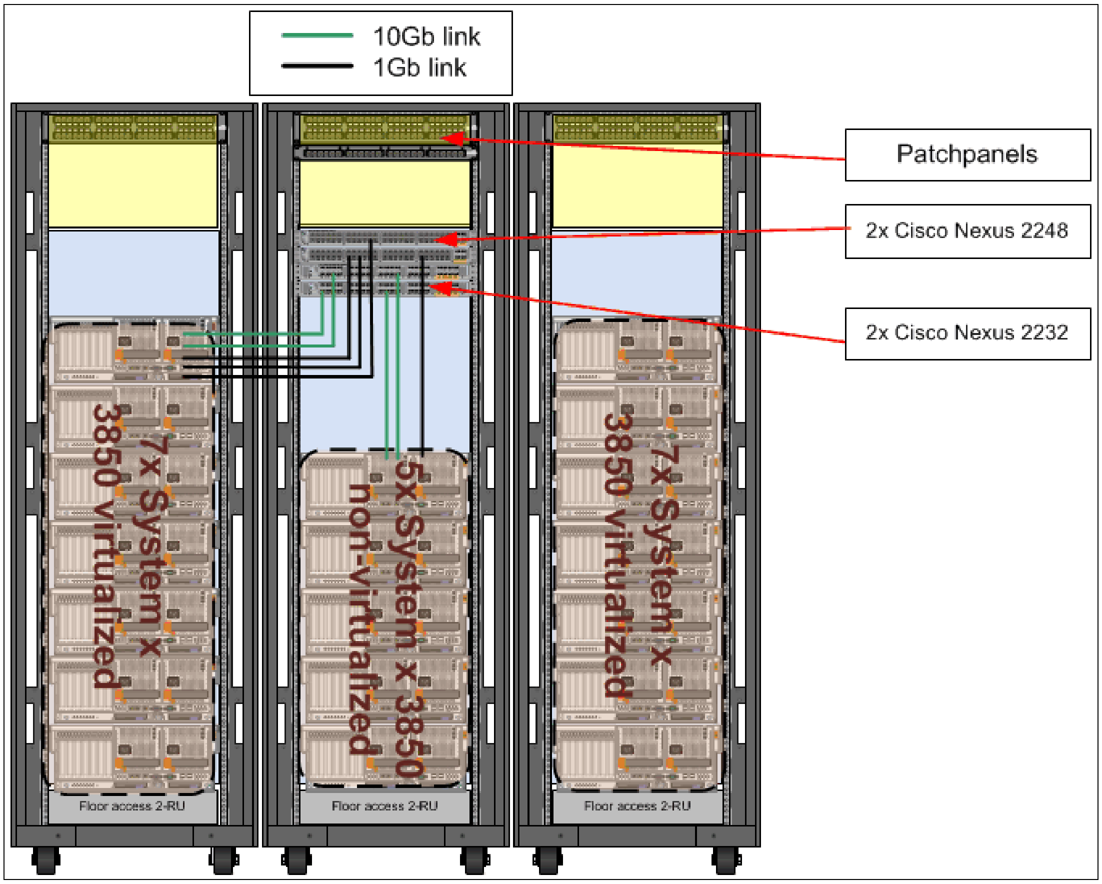


Figure 12-29 System x PoD

Figure 12-30 shows the physical layout of the System p PoD, which is equipped with 5 virtualized System p Series Servers. Each p770 Server has 4 CECs and 16 RUs. The Nexus 2248 and 2232 switches are placed in the middle rack to minimize the distance from server port to network port. Each virtualized System p Series Server has 8x 10Gb and 8x 1Gb interfaces.

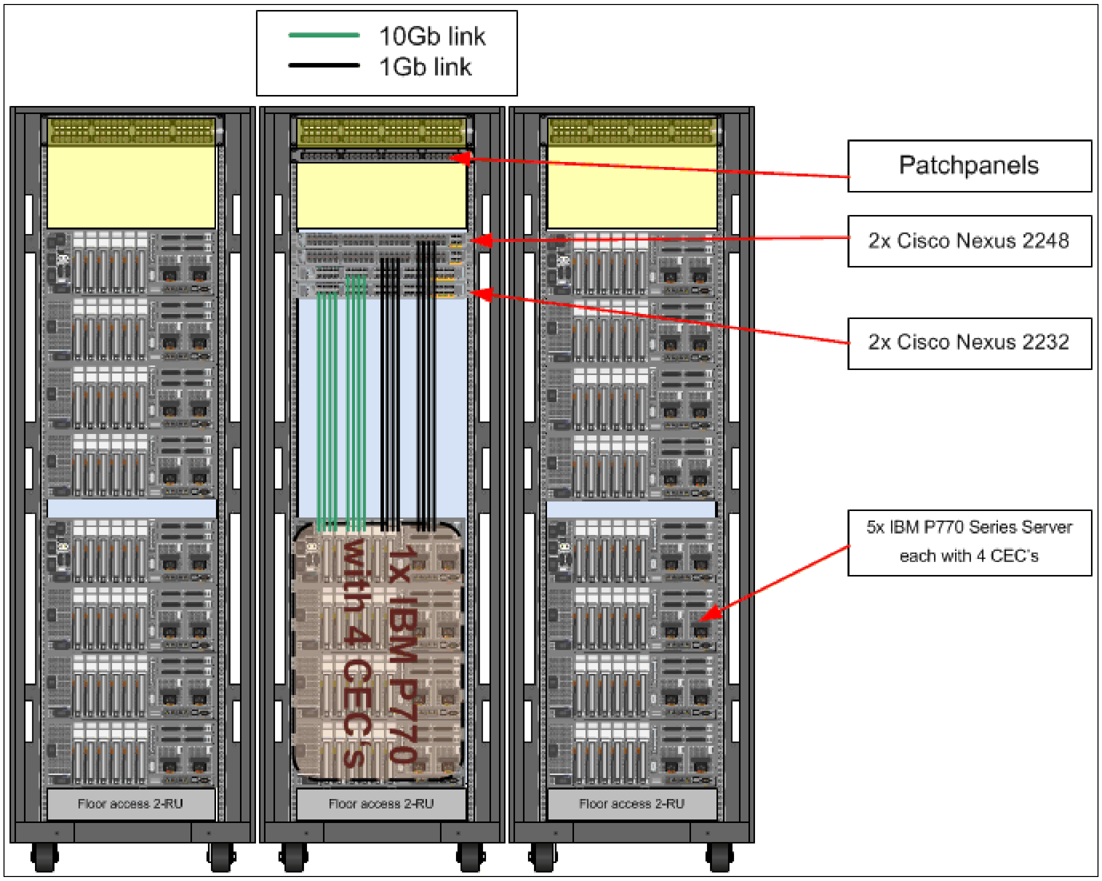


Figure 12-30 System p PoD

For illustration, the links are only depicted conceptual for one server. Independent of the type of PoD, the server links must be equally distributed over all Nexus 2k in order to achieve maximum redundancy.

The patch panels are used amongst others for the uplinks from the Nexus 2232 and Nexus 2248 switches to the Nexus 5596 switches. The 2248s will have 2x10 GE links, while the 2232s will have 4x10 GE links. This will allow us to calculate the oversubscription ratios in the next section.

Oversubscription

The oversubscription ratio within the Access Block is different for 10 GE ports and 1 GE ports and a System x PoD has a different oversubscription than a System p PoD based on the number of required interfaces.

- Nexus 2232 in System x PoD:
  - Each Nexus 2232 will have 4x 10 GE uplink ports to the Nexus 5596
  - Each Nexus 2232 will have 19x 10 GE used for System x servers
  - Oversubscription = 190/40 = 4.75

- Nexus 2248 in System x PoD:
  - Each Nexus 2248 will have 2x 10 GE uplink ports to the Nexus 5596
  - Each Nexus 2248 will have 24x 1 GE used for System x servers
  - Oversubscription = 24/20 = 1.2
- Nexus 2232 in System p PoD:
  - Each Nexus 2232 will have 4x 10 GE uplink ports to the Nexus 5596
  - Each Nexus 2232 will have 20x 10 GE used for System p servers
  - Oversubscription = 200/40 = 5
- Nexus 2248 in System p PoD:
  - Each Nexus 2248 will have 2x 10 GE uplink ports to the Nexus 5596
  - Each Nexus 2248 will have 20x 1 GE used for System x servers
  - Oversubscription = 20/20 = 1

Each Access Block has in total 16 Nexus 2k (8x Nexus 2232 and 8x Nexus 2248) connected to one pair of Nexus 5596 parent switches. Each Nexus 5596 uses 8x 10 GE links for spine connectivity and 240 GE are coming from the Nexus 2k so that the oversubscription ration is 3:1 for each Nexus 5596.

The calculated oversubscription ratios above are considered acceptable as not all virtual or physical servers will be using the links at all times and the 1 GE connections used mainly for systems and network management purposes are generating only limited traffic volumes.

Adding additional servers and/or adding 10 GE interfaces to existing servers will increase the oversubscription ratio, but in order to improve oversubscription all the 8x10 GE uplinks can be used from the Nexus 2232 and all the 4x10 GE uplink can be used from the Nexus 2248.

Also as required to control and reduce the oversubscription, more links between the Nexus 5596 and the spine switches (Nexus 7K) can be added.

12.7.5 Service Zone design

As shown in Figure 12-31 on page 603 the Large Data Center use case design will include three Service Zones. Each Service Zone contains network, storage and compute resources. The introduction of the three Service Zones as a “logical” entity will address the following topics:

- Manageability:

  Each Service Zone has a pre-defined range of VLANs and storage capabilities, which simplifies the daily operations and capacity management/planning. A proper capacity management per Service Zone helps to avoid under or over provisioned resources in the data center; that is, compute resources can only be added if there are enough storage resources available.

- Maintainability:

  Limited maintenance and lifecycle activities can be performed in a Service Zone without impacting other Service Zones or customers in a multi-tenant environment. A large percentage of possible common failures or maintenance tasks are within one Service Zone only and maintenance tasks outside of the Service Zone are performed less often. The compute resources must only be connected to network and storage resources in the same Service Zone to create consistent maintenance domains.

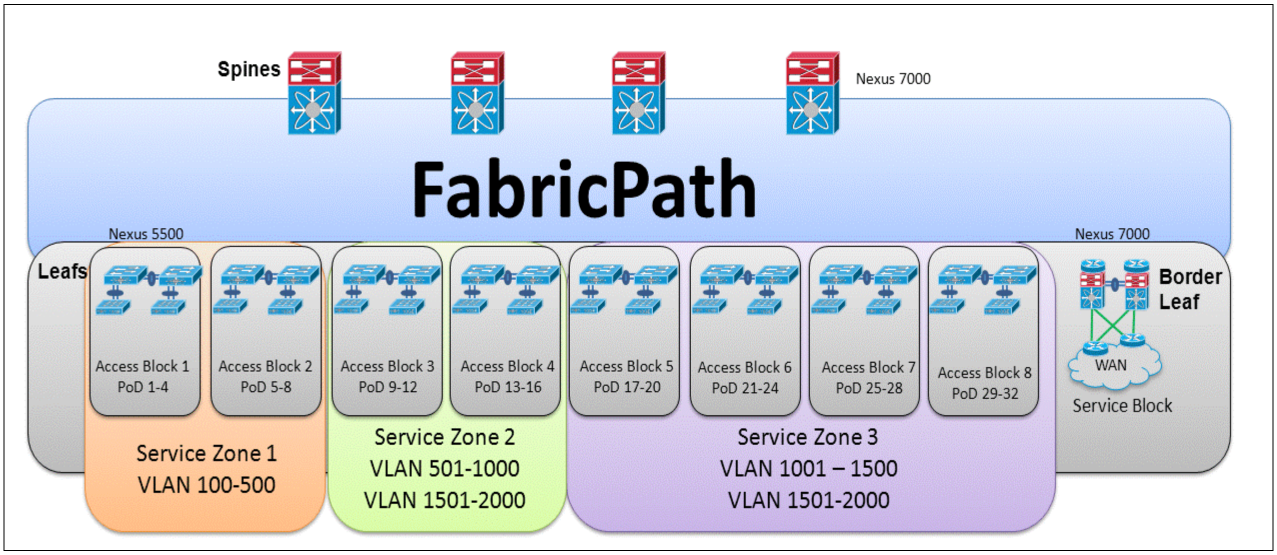


Figure 12-31 Large Data Center use case - service zone design

Overall the introduction of Service Zones ensure the stability of the overall data center infrastructure with the combined view of network, server and storage. This avoids the possibility that the added-value of flexibility and scalability based on new technologies and features can be un-intentionally exploited or jeopardized via ignorance:

- Service Zone 1:
  - VLAN 100-500
  - FCoE enabled
- Service Zone 2:
  - VLAN 501-1000 and VLAN 1501-2000
  - Native Fibre Channel
- Service Zone 3:
  - VLAN 1001-1500 and VLAN 1501-2000
  - Native Fibre Channel and FCoE enabled

As illustrated on Figure 12-31, there will be VLANs that are shared / extended between the security zones, for examples VLANs 1501 to 2000 are present on security zone 2 and 3. This allows Layer 2 connectivity between servers or virtual-machines present on both zones. Virtual-Machine mobility may not be possible because there will not be access to a shared storage between both zones; however new technologies are available that allows VM and Storage Mobility to occur simultaneously so if those new server virtualization technologies are adopted, then Virtual-Machine mobility between the zones will be possible.

The Border Leaf and Spine devices do not belong to a specific Service zone, because the functions performed from these devices are relevant for all Service Zones and hence are all in the same fault domain.

12.7.6 Network scale and design considerations

The Network Fabric using Cisco Fabric Path replaces a classical collapsed aggregation/core design. The Spine switches have only Layer 2 functionalities, whereas all the Layer 3 routing and forwarding will be performed on the Border Leaf.

Each spine switch will use 48 10 GE ports. (2x 10 GE ports to each leaf and 8x 10 GE ports to each border leaf switch. Based on this port numbers, a Cisco Nexus 7004 would be sufficient, but the used Nexus 7010 ensures better scalability.

Each border leaf switch has 8x 10 GE links to the spine switches, because all inter-VLAN traffic will be routed on the border leaf. An oversubscription rate of 2:1 in the Network Fabric for all traffic going to the border leaf is considered acceptable. If an oversubscription of 1:1 is needed, each border leaf switch will require 16x 10 GE links to each spine switch.

The Layer 3 function in the border leaf switch is required for inter-VLAN routing, but also to forward traffic to and from the WAN. The use of dynamic routing protocols or static routing is dependent of the WAN service provider.

It is always important to check the scalability limits of the network infrastructure, and in this use case, the following values apply:

- ARP Cache Size: The ARP cache size is a critical value in the border leaf devices. The ARP cache is a table which stores mappings between the Data Link Layer (MAC) addresses and Network Layer (IP) addresses. The border leaf devices have the Layer 3 interface of all connected subnets and hence need to resolve and store the MAC addresses of the locally attached subnets in the ARP cache. The Nexus 7010 supports a maximum of 400k ARP entries. If the maximum is reached, each additional ARP packet receiving the device will be dropped until there is new space available due to an ARP time-out.

  The maximum number of virtual and physical images on the Network Fabric in the Large Data Center use case is 29000, so the ARP cache size is not a problem at all even if each virtual image will have 3 virtual NICs with 3 MAC addresses.

- MAC Address Table size: The MAC address table size on the Nexus 5596 is limited to 32000 entries. Each Access Block can host a maximum of round about 3600 images. With Cisco Fabric Path, each Nexus 5596 sees only the MAC addresses of the servers that are actively communicating on the network and have a bidirectional flow with servers connected to the local Access Block, this is the feature called conversational learning on FabricPath. Each Nexus 5596 does not need to learn the MAC addresses of all other Access Blocks. Hence the MAC address table size limit is not an issue with this design, even if each VM needs 3 MAC Addresses, which can be a requirement.
- Number of FEXs: The design start with 16 FEXs, 8 FEXs connected to each Nexus 5596. With 8 FEXs per Nexus 5596 and the ports required for uplinks to the spines and for connection between the two Nexus 5596, there are still up to 60 ports available on each Nexus 5596. Those available ports can be used for increasing the number of uplinks between the Nexus 5596 and the Nexus 7010s on the spine or those ports can used to add additional FEXs / PoDs as long as the number of FEXs per Nexus 5596 does not exceed 24 FEXs which is the maximum supported by Cisco.
- Number of 10 GE access ports: The design with 4 PoDs and 16 FEXs can scale up to 32x8 = 256 10 GE ports while the design with 8 PoDs and 32 FEXs can scale up to 512 10 GE ports. Another option that can be used to scale up to 512 10 GE ports is to use only Nexus 2232s as the access layer, instead of mixing 1 GE and 10 GE ports in each PoD.

Network services

The Large Data Center use case will include different network services. All network services, either as integrated modules in services switches or as a specific appliance will be connected to the border leaf. Dependent of the type of network service, they will be deployed in different fashions:

- Load balancing and application delivery will be deployed in Active/Standby mode via Layer 3 using Source NAT. Active/Active deployment is also an option and it is the preferred option in case of multi-tenancy requirements using virtual contexts.
- WAN acceleration: WAAS will be deployed in Active/Active mode via Layer 3, as an appliance on a stick (or one armed configuration). For more information see 6.4, “Wide Area Application Services (WAAS)” on page 300.
- Firewalling and Intrusion prevention (internal and external security services) will require a Cisco ASA which will be deployed in Active/Standby mode in Layer 3. Active/Active deployment is also an option and it is the preferred option in case of multi-tenancy requirements using ASA virtual contexts. For more information see 6.7, “ASA series security in the data center” on page 325.
- Cisco ISE/ACS are optional components for AAA
- Cisco NAM is an optional component for traffic analysis

Cisco DCNM, Tivoli NM, NCM and OMNIbus are optional components for data center network management.

Legacy integration

Existing legacy network devices can be connected to the Nexus 7010 border leaf via either 10 GE or 1 GE interfaces. The vPC+ functionality enables the connectivity via channelized links and can extend VLANs from the legacy network into the Network Fabric. Hence, VLAN translations or other technologies to extend VLANs for hot server migrations are not required which simplifies the migration scenario.

12.7.7 Storage and Storage Area Network design

In the Large Data Center use case, Fibre Channel over Ethernet (FCoE) as well as traditional Fibre Channel is used to provide connectivity from the Servers to the Storage Arrays. The storage arrays used on this solution only have Fibre Channel interfaces. FCOE is described in detail in Chapter 7, “Convergence of LAN and SAN: Fibre Channel over Ethernet” on page 369.

Fibre Channel over Ethernet (FCoE) design

The approach to adopt FCoE in the Large Data Center use case is based on the topologies shown in Figure 12-32 and Figure 12-33.

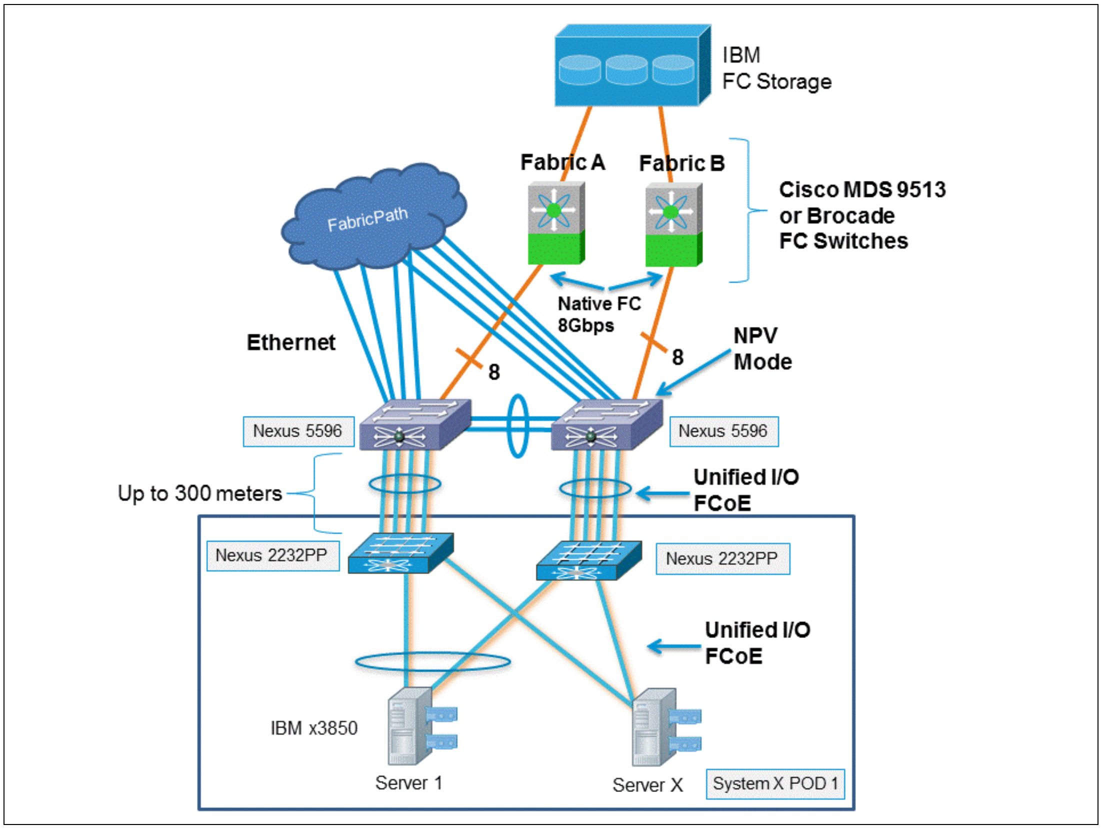


Figure 12-32 FCoE topology for the Large Data Center use case: System x PoD

The topology shown in Figure 12-32 shows that each System x connects with a 10 GE interface to the two Nexus 2232PP installed on the ToR within the PoD. The servers can be connected using Virtual Port-Channel (server 1) or with NIC teaming (server X). It is important to remember that vPC or NIC teaming is only applicable for the Ethernet traffic. As far as Fibre Channel over Ethernet traffic is concerned, one link is connected to Fabric A and the other link is connected to Fabric B. The servers use their multi-pathing software for load balancing traffic between the two paths to the storage array.

Each Nexus 2232 is single homed to one Nexus 5596, this keep very clearly the Fabric A and Fabric B separation for the SAN traffic. There are four 10 GE links between each Nexus 2232PP and its parent Nexus 5596.

When the Fibre Channel over Ethernet (FCoE) traffic arrives on the Nexus 5596, the FCoE encapsulation is removed and the frames are sent in native Fibre Channel connections to the SAN core. The Nexus 5596 on the left is connected to Fabric A and the Nexus 5596 on the right is connected to Fabric B. There are eight 8 Gbps Fibre Channel links between the Nexus 5596 and the SAN Core. Those FC links can be channelled (grouped together) or not, depending if the SAN core is made of Cisco MDS 9513s or 9710s (recommended because the links can then be channelled) and it is possible to use Virtual SANs (VSANs), or can be made of Brocade FC switches.

As described in 2.15.2, “Nexus 5500” on page 102 to make the Fibre Channel connections from the Nexus 5596 to the SAN, we can utilize the Nexus 5596 Unified Ports that can run as 8 Gbps native FC ports. This increases the flexibility of the design because more FC ports can be easily added to reduce the oversubscription or to address any application specific requirement.

In the topology shown in Figure 12-33, we illustrate how an IBM System p770 would connect to the network. The server has eight (8) 10 GE interfaces, four connecting to each Nexus 2232. Two of those connections are used for Production/Admin traffic and the other two are used for Installation, Backup, and Recovery.

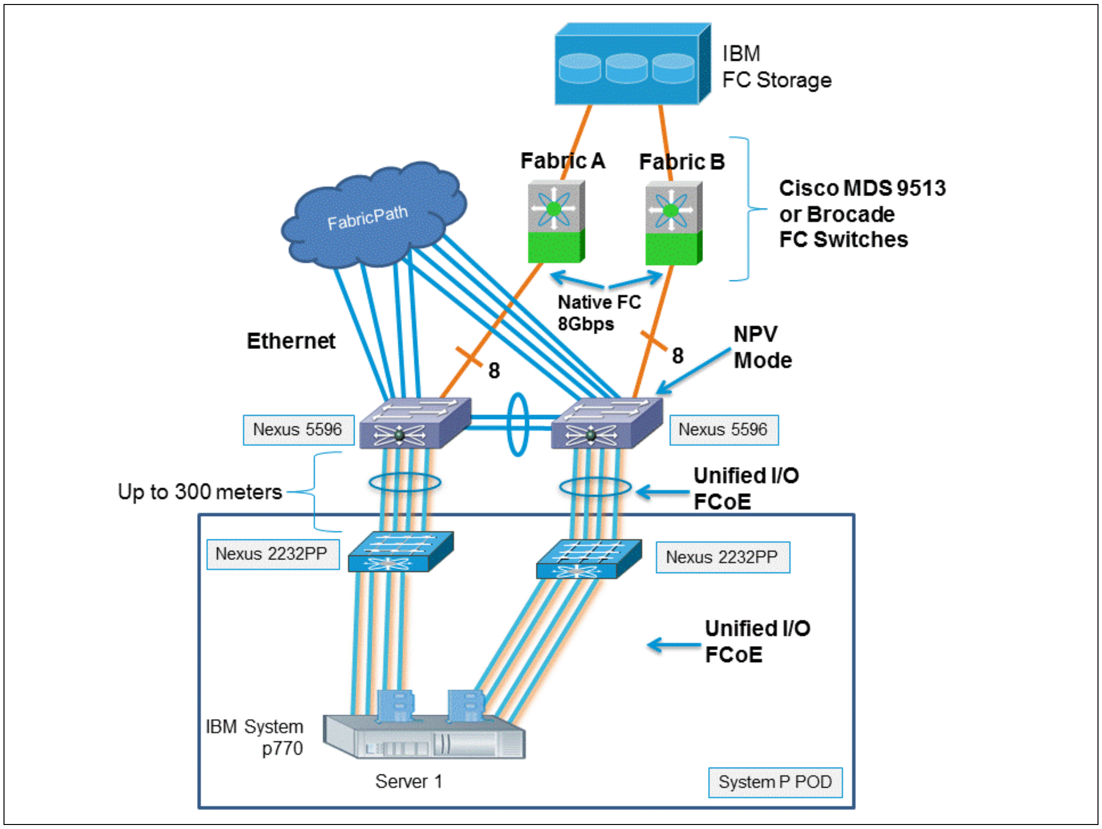


Figure 12-33 FCoE topology for the Large Data Center use case: System p PoD

Oversubscription on the Fibre Channel over Ethernet (FCoE) design

Each Nexus 5596 has four Nexus 2232PP connected to it, as per “Oversubscription” on page 601. With each Nexus 2232PP connected with four 10 GE links, then each Nexus 5596 can receive a maximum of 160 Gbps of Fibre Channel over Ethernet traffic. This assumes that there are no regular Ethernet traffic on the uplinks from the Nexus 2232PP to the Nexus 5596, which will not be the case in a real implementation but this scenario needs to considered when considering oversubscription between the Servers and the disk Arrays.

With a maximum of 160 Gbps of Fibre Channel over Ethernet arriving on the Nexus 5596, then with 64 Gbps of native Fibre Channel uplinks going towards the SAN - eight 8 Gbps Fibre Channel links between Nexus 5596 and the SAN Core - there is 2.5:1 (160/64) oversubscription between the Nexus 5596 and the SAN core.

This delivers an end-to-end oversubscription ratio of 11.8:1 for System x and 12.5:1 for System p to access the Fibre Channel SAN. Those oversubscription ratios are within the acceptable range for most customers however specific application requirements must be considered when designing your environment.

Fibre Channel design

In the Large Data Center use case there will be servers that have native Fibre Channel connectivity to the Storage Area Network (SAN) to access the disk arrays. Those servers are equipped with 8 Gbps Fibre Channel Host Bus Adapters (HBAs).

The Fibre Channel design selected for the Large Data Center use case is based on a collapsed Core design, where Cisco MDS 9500s or MDS 9710s are used as the Directors. Servers as well as Disk Arrays connect directly to the Cisco MDS Directors. Each MDS 9513 can provide up to 528 FC ports and the MDS 9710 can provide up to 384 16G ports. Therefore, each platform has enough port density to enable this collapsed design, as shown in Figure 12-34.

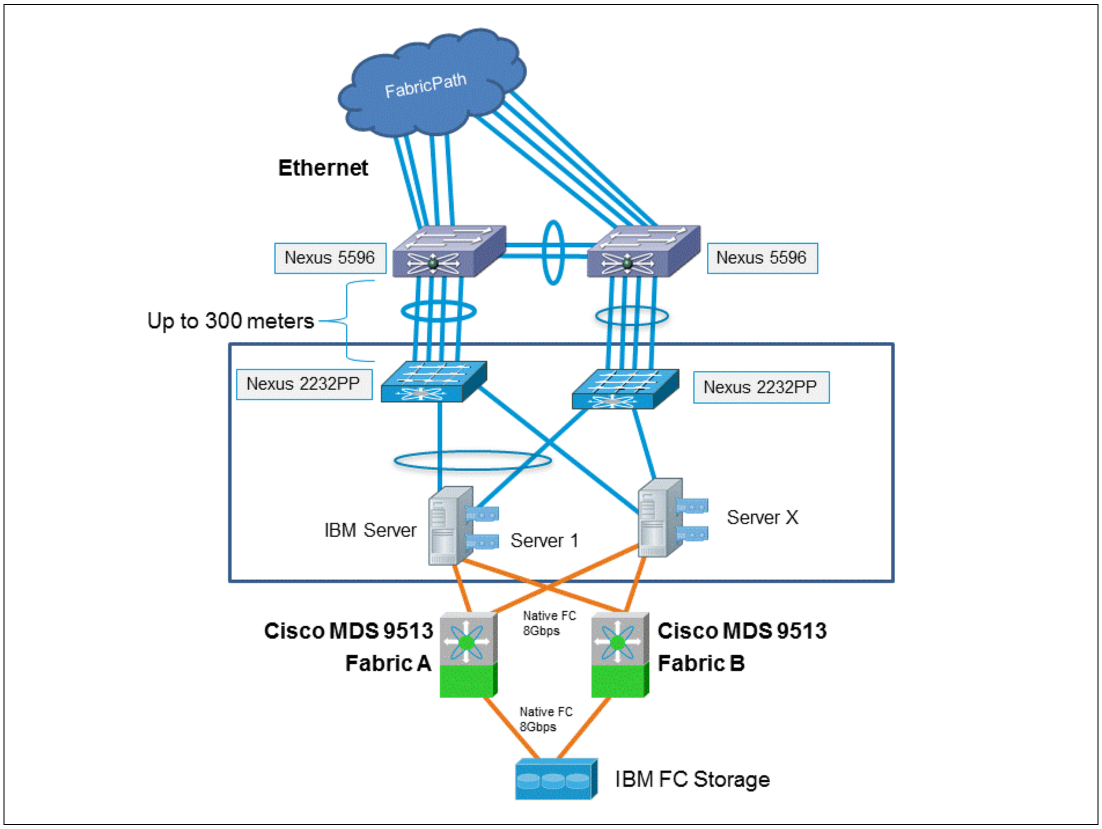


Figure 12-34 Fibre Channel Design based on Cisco MDS 9513: Collapsed Core design

There will be one Cisco MDS 9513 or MDS 9710 for Fabric A and another for Fabric B. The servers and the disk arrays are connected to both MDS 9513 or 9710s. However the MDS 9513 or MDS 9710 themselves are not directly connected, in order to keep Fabric A and Fabric B completely isolated. A second Cisco MDS 9513 or MDS 9710 can be added to each Fabric if more than 528 or 384 ports, respectively, are required per Fabric.

Note: For more details on the Cisco MDS 9500 Series Multilayer Directors, see this website:

http://www.cisco.com/en/US/products/ps5990/index.html

The use of Virtual SANs (VSANs) is recommended on large environments, particularly if multi-tenancy is required. Given that the requirements specified for this use case are not specific enough that allow a full recommendation on how the VSANs should be used, this level of detail will not be covered, however it is highly recommended that you consider using this feature when planning your SAN.

There are multiple use cases for VSANs, such as creating a VSAN for each type of operating system, or utilizing them on the basis of business functions (such as a VSAN for development, for production, or for lab). VSAN 1 is created on the Cisco MDS switch by default and cannot be deleted. As a best practice, VSAN 1 should be used as a staging area and another VSAN should be created for the production environment.

Each Cisco MDS 9513 or 9710 supports up to 256 VSANs. Of these, one is a default VSAN (VSAN 1), and another is an isolated VSAN (VSAN 4094). User-specified VSAN IDs range from 2 to 4093.

Note: For more details on a Large SAN Design Best Practices, see this website:

http://www.cisco.com/en/US/prod/collateral/ps4159/ps6409/ps5990/white_paper_C11-515630.html

Service Zone 1 SAN design with FCoE

As defined on the requirements, within Service Zone 1 (refer to 12.6.4, “Service Zone design” on page 580), FCoE will be used to provide connectivity from the servers to the storage arrays. This is implemented following the design approach introduced on “Fibre Channel over Ethernet (FCoE) design” on page 606.

As illustrated in Figure 12-35, all the servers within Service Zone 1 access their storage arrays by using FCoE from the server to the Nexus 2232PP, which then connects to a single Nexus 5596. Each Nexus 5596 then have eight 8 Gbps native Fibre Channel links (channelled together) to one MDS 9513 or MDS 9710- Fabric A or Fabric B. The IBM Storage array is then directly connected to the MDS 9513 or MDS 9710 via multiple 8 Gbps FC links.

Figure 12-35 shows Access Blocks 1 and 2 that make Service Zone 1, however within each Access Block only one of the three System x PoDs is represented and the System p PoD. Also only a single server is represented, the exact number of the servers per PoD is covered on “Number of servers” on page 565.

Each Server within Service Zone 1 has access to any of the storage arrays connected to the SAN, this provides flexibility to the server administrators because the servers can exist anywhere within Service Zone 1, and also Virtual Machine mobility is possible given that the servers have access to a shared storage.

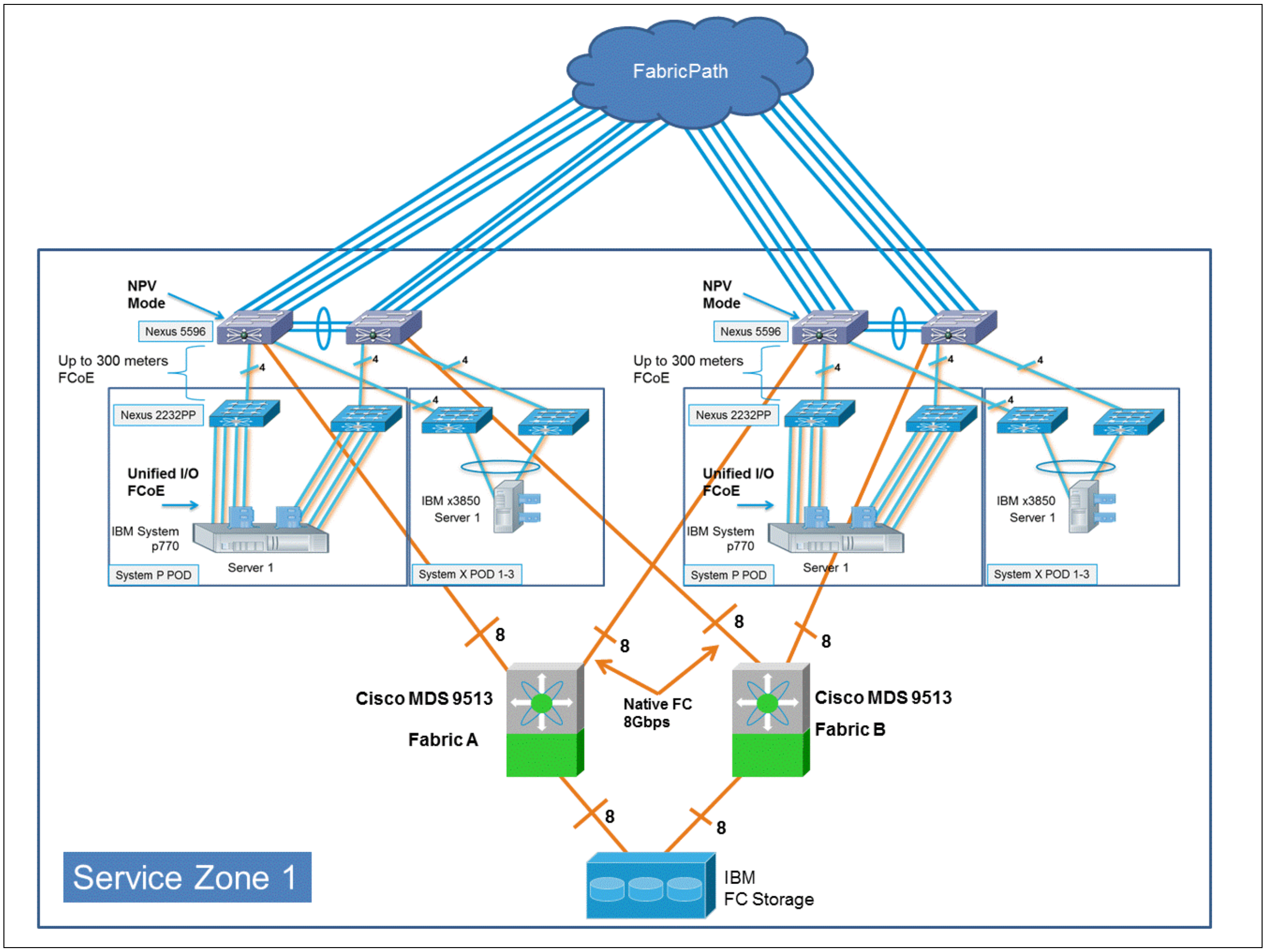


Figure 12-35 Service Zone 1 SAN Design with FCoE

Service Zone 2 SAN design with native Fibre Channel

As defined on the requirements, within Service Zone 2 (refer to 12.6.4, “Service Zone design” on page 580), native Fibre Channel will be used to provide connectivity from the servers to the storage arrays. This is implemented following the design approach represented on “Fibre Channel design” on page 608.

As illustrated in Figure 12-36, all the servers within Service Zone 2 access the storage arrays by using native Fibre Channel.

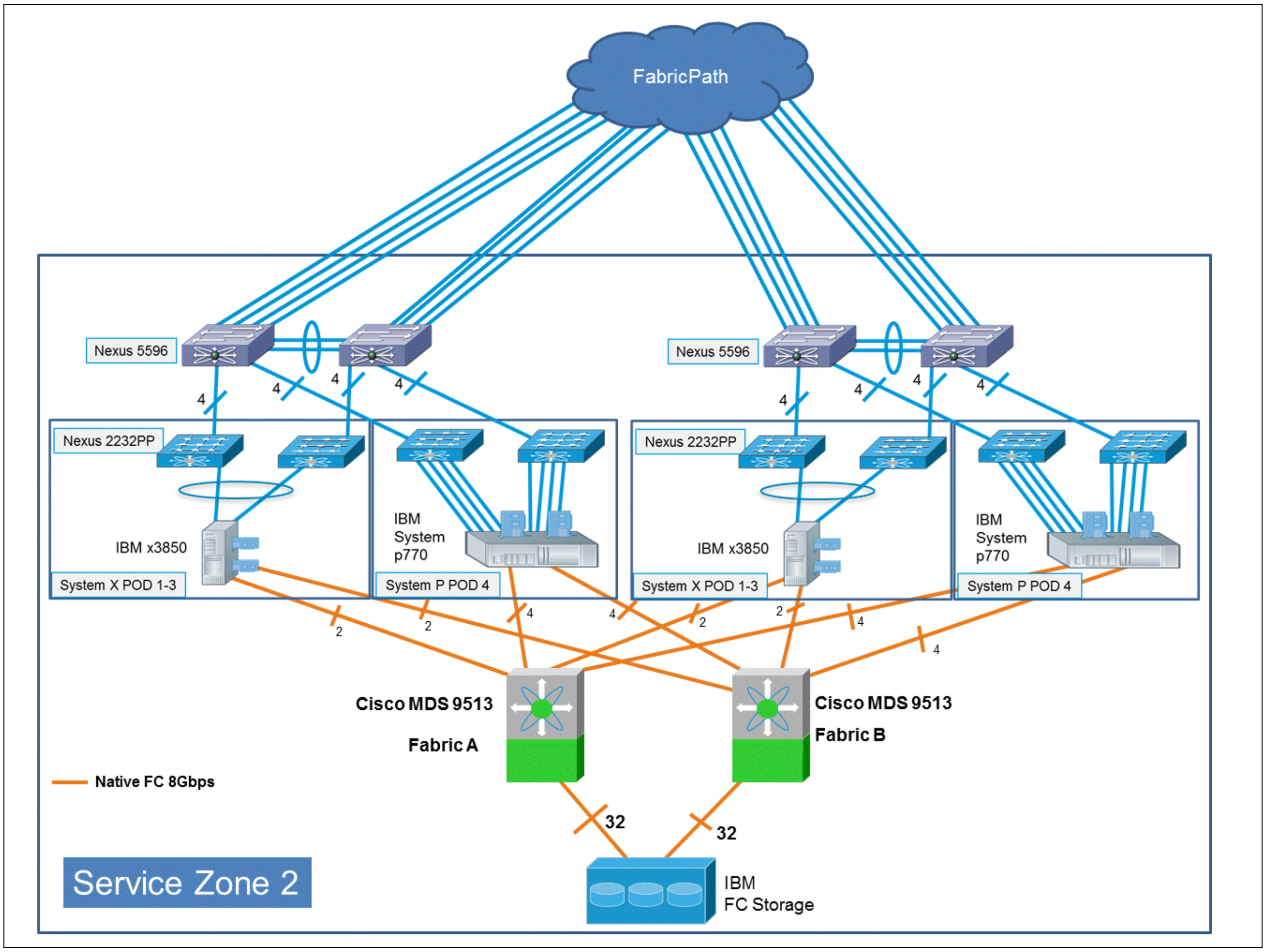


Figure 12-36 Service Zone 2 SAN Design with native Fibre Channel

The total number of Fibre Channel connections within each SAN Fabric on Service Zone 2 can be calculated based on the following:

- Each System x has two 8 Gbps connections to each SAN Fabric;
- Each System p has four 8 Gbps connections to each SAN Fabric;
- There are 57 System x servers and 5 System p servers per Access Block (each pair of Nexus 5596 defines a Access Block);

Based on the above there are:

- 57 System x * 2 connection per server = 114 FC connections for System x per Fabric;
- 5 System p * 4 connection per server = 20 FC connections for System p per Fabric;
- Total of 134 FC connections per Fabric for one Access Block;

Service Zone 2 is made of two Access Blocks, so the total number of FC connections for the servers within Service Zone 2 is 134 * 2 = 268 Fibre Channel connections for servers per SAN Fabric.

As shown in Figure 12-36 each server connects directly to both Cisco MDS 9513s or MDS 9710s, one on each Fabric. The IBM Storage Array also connects to the MDS 9513s or MDS 9710s.

As per above, there will be up to 268 host (server) connections on each Fabric. To keep the fan-out ratio within acceptable levels the IBM storage array will connect to each Fabric with thirty-two 8 Gbps FC links. This will provide an 8.3:1 fan-out ratio – 268 host connections / 32 target connections = 8.3:1 host HBAs to one storage port. More links can be added between the IBM storage array and the MDS 9513s to reduce oversubscription if needed, for example an IBM DS8870 supports up to 128 host ports.

In this design each MDS 9513 (IBM part number 2054-E11) will be equipped with 6 * 48-port 8 Gbps Advanced Fibre Channel Switching Module (IBM FC#: 2820) and 2 * 32-port 8 Gbps Advanced Fibre Channel Switching Module (IBM FC#: 2818). This will provide each MDS 9513 with 352 Fibre Channel ports. The servers should be connected to the 48-port modules, while the Storage Array should have its connections on the 32-port modules. There are still three slots left available on the MDS 9513 for future growth. In addition, the design with MDS 9710s will be equipped with 8 * 48-port 16 Gbps Advanced Fibre Channel Switching Module. This will provide each MDS 9710 with 384 Fibre Channel ports. Servers and storage arrays can connect to any port on the MDS 9710s.

Figure 12-36 on page 611 shows Access Blocks 3 and 4 that make Service Zone 2, however within each Access Block only one of the three System x PoD is represented and the System p PoD. Also only a single server is represented, the exact number of the servers per PoD is covered on “Number of servers” on page 565.

Each Server within Service Zone 2 has access to any of the storage arrays connected to the SAN, this provides flexibility to the server administrators because the servers can exist anywhere within Service Zone 2, and also Virtual Machine mobility is possible given that the servers have access to a shared storage.

Service Zone 3 SAN design with native Fibre Channel and FCoE

As defined on the requirements, within Service Zone 3 (refer to 12.7.5, “Service Zone design” on page 602), native Fibre Channel and FCoE will be used to provide connectivity from the servers to the storage arrays. This is implemented combining the design approaches presented on “Fibre Channel over Ethernet (FCoE) design” on page 606 and “Fibre Channel design” on page 608.

As illustrated in Figure 12-37, the servers within Service Zone 3 can access the storage arrays by using native Fibre Channel or FCoE. Figure 12-37 shows Access Blocks 5, 6, 7 and 8 that make Service Zone 3, however within each Access Block only one of the three System x PoDs is represented and the System p PoD. Also only two servers are represented within each PoD, one is accessing the SAN via FCoE and the other is accessing the SAN via native Fibre Channel - to keep the diagram simple, the only FC connections shown in the diagram are for the System p within the first Access Block and the System x on the right on the last Access Block. The exact number of the servers per PoD is covered on “The assumption section represents the requirements we used when designing the Data Center solutions for this use case; the assumptions do not represent the maximum capacity or scale of any of the areas within the data center, including servers, network, storage.servers” on page 593.

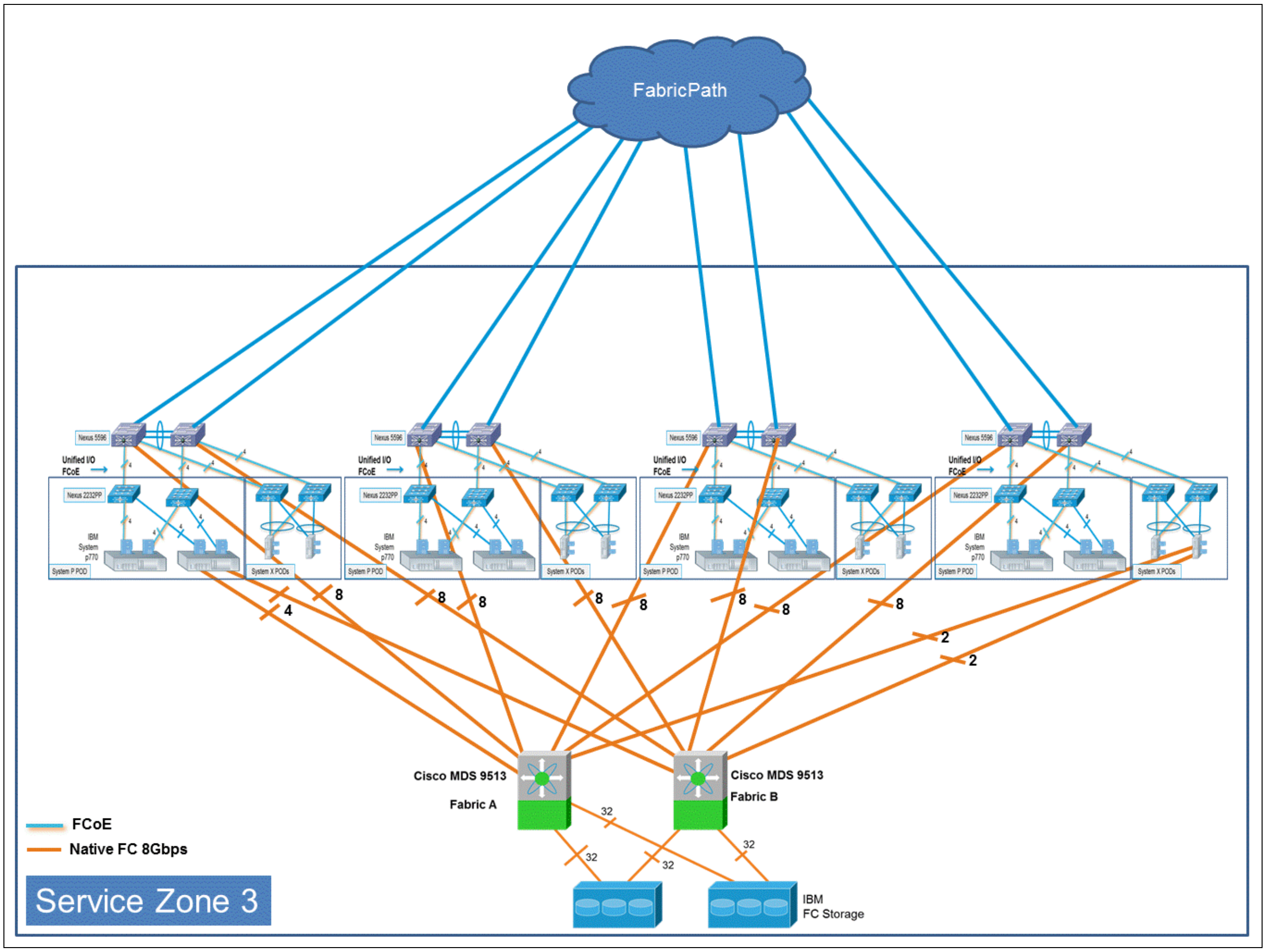


Figure 12-37   Service Zone 3 SAN Design with native Fibre Channel and FCoE

For our use case, we assume that half of the servers within Service Zone 3 are accessing the storage array using FCoE and the other half will be using native Fibre Channel. In a real world implementation those numbers will vary however the design principle used within Service Zone 3 is applicable for any environment where you will find some servers equipped with Converged Network Adapters (CNAs) and others equipped with Host Bus Adapters (HBAs). This may happen because of integrating existing servers that have Host Bus Adapters (HBAs) in a new environment where the new servers have Converged Network Adapters (CNAs). The fact is that Fibre Channel over Ethernet (FCoE) and native Fibre Channel can co-exist in the same environment side by side and access the same storage array that is connected to the SAN via native Fibre Channel or via Fibre Channel over Ethernet (FCoE), in our use case all the storage arrays use native Fibre Channel.

To provide the number of FC ports required, the native FC SAN will have a pair of Cisco MDS 9513s, one per Fabric. Each MDS 9513 will be equipped with 6 * 48-port 8 Gbps Advanced Fibre Channel Switching Module (IBM FC#: 2820) and 3 * 32-port 8 Gbps Advanced Fibre Channel Switching Module (IBM FC#: 2818). This will provide a total of 384 8 Gbps Fibre Channel ports per fabric. The servers should be connected to the 48-port modules, while the Storage Arrays and uplinks between the MDS 9513 and the Nexus 5596 should be connected to the 32-port modules. There are still two slots available on each of the MDS 9513 for future growth.If the design with MDS 9710s is used, they will be equipped with 8 * 48-port 16 Gbps Advanced Fibre Channel Switching Module. This will provide each MDS 9710 with 384 Fibre Channel ports. Servers, Storage arrays and uplinks can connect to any port on the MDS 9710.

Connecting the FC SANs from the multiple Service Zones

The sections above describe the design for the SAN for each of the Service Zones that form the Large Data Center use case. The SANs are aligned with the Service Zone that they serve. As previously described, the Service Zone concept has the objective of providing this alignment between LAN, SAN and Compute resources to facilitate operational processes.

There are cases where it may be desirable to interconnect the SANs of the different Service Zones; this would be needed for example if Virtual Machine mobility is required, given that access to a shared storage is generally needed although some recent developments by hypervisor vendors allow Virtual Machine and Storage Mobility to happen at the same time.

Figure 12-38 shows how the SAN from Service Zone 2 and Service Zone 3 should be interconnected. With this interconnection as well with the fact that VLANs 1501 to 2000 are present on Service Zones 2 and 3 (as in Figure 12-31 on page 603), then it is possible to have Virtual Machine mobility between those Service Zones.

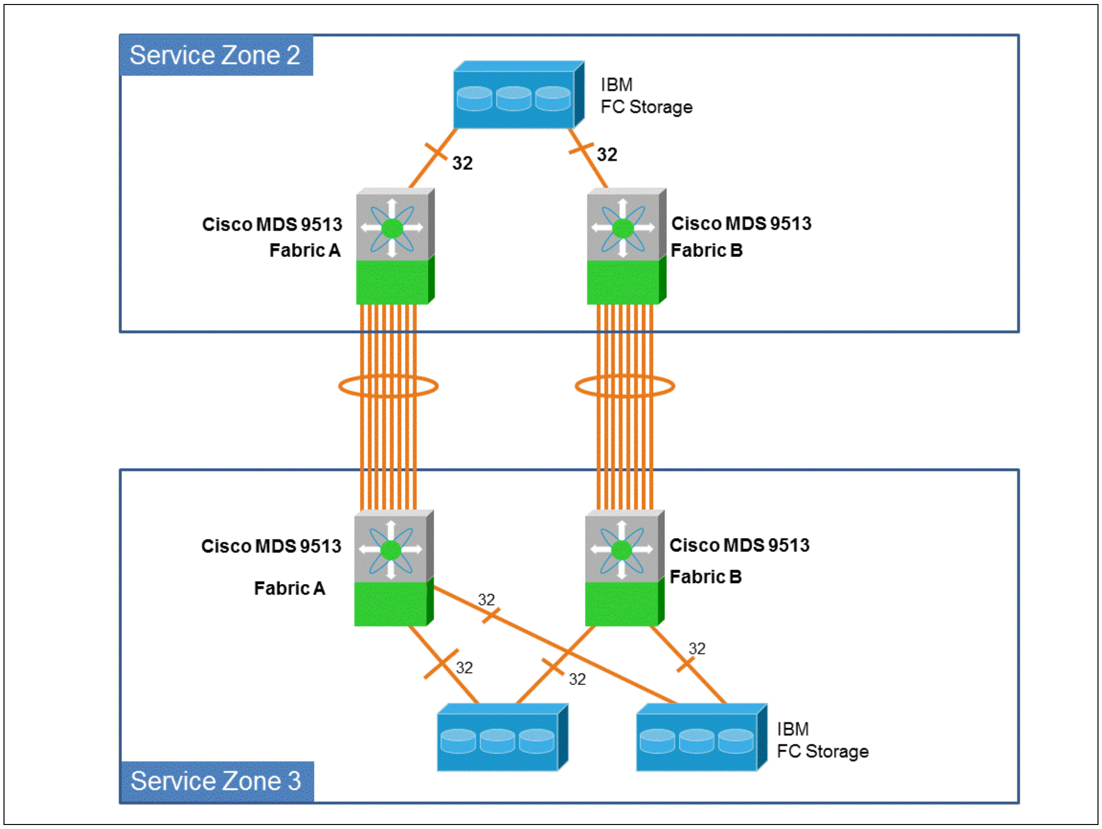


Figure 12-38 Interconnection between the SANs on Service Zone 2 and Service Zone 3 using a port-channel

The interconnection of the SANs is simple; each Fabric on each of the Services Zones is made of a single MDS 9513 or MDS 9710 so to interconnect them a single port-channel made of up to 16 8 Gbps FC links is recommended. Figure 12-38 shows an 8-port, port-channel, however more links (up to 16 in total) can be added for additional capacity as needed. To increase availability, it is recommended to distribute the physical links across switching modules, so a failure of a switching module cannot bring down the PortChannel link.

12.8 WAN Edge Block Design for Small, Medium, and Large Data Centers

This section presents a design for the WAN Edge Block that could be used on the Small, Medium and Large Data Center use cases with small variations, primarily on the areas of performance and throughput.

The WAN Edge forms the data center perimeter to the enterprise wide area or provider IP/NGN backbone and to the public Internet. It provides customers access to their Wide Area Network. Most enterprise customers will connect to their own MPLS Wide Area network.

These perimeter nodes may be dedicated to Layer 3 routing functions or may be multi-service in nature, providing Layer 2 interconnects between data centers as well as Layer 3 services. The Cisco platforms recommended as WAN Edge routers include: the Cisco CRS, Cisco ASR 9000 and ASR 1000. The WAN edge will attach to the core using a dynamic routing protocol. See Figure 12-39 for a diagram of the WAN Edge.

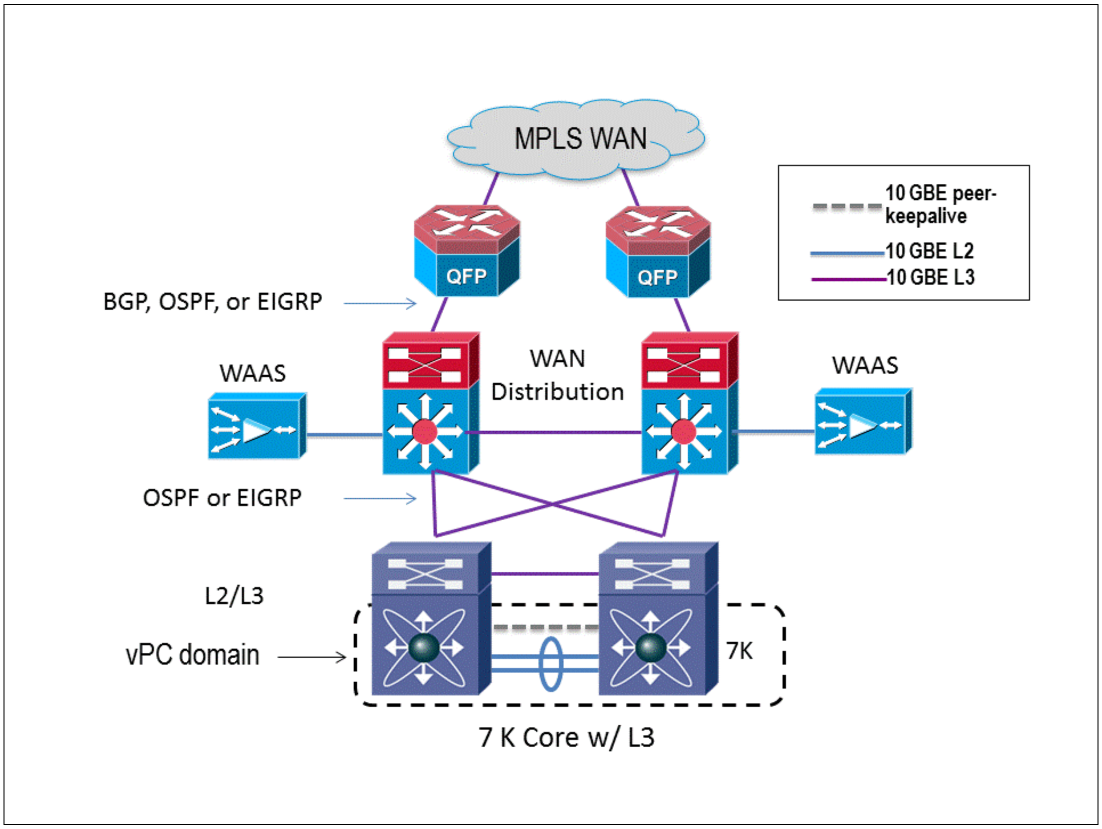


Figure 12-39 WAN Edge Design

On the design above a pair of routing capable switches is used as a WAN Distribution block to connect the WAN routers to the core. This has several advantages. If a different dynamic routing protocol is used in the data center than the WAN, the WAN distribution block can export/import the routing protocols into the data center. However the real advantage is the WAN distribution provides a place to connect WAN accelerator (like the Cisco WAAS) or IPS devices.

The WAN distribution block, although it provides those advantages, is optional. It is a valid design to connect the WAN routers directly to the data center core in the Small and Medium Data Center use cases, and directly to the Border Leaf Switches in the Large Data Center use case.

A functionality that is important for the WAN Edge Block is Quality of service (QoS), this is because the WAN circuits are of much lower speed than the links inside the data center and also the WAN routers are the point of entrance for traffic into the data center. If we consider that QoS is used within the data center the WAN routers are at the edge of the QoS domain and they will mark traffic when the traffic enters the data center, and these markings will be trusted at the nodes within the data center infrastructure

The WAN routers need to be able to:

- Classify and Mark the traffic coming into and leaving the data center;
- Manage congestion and try to avoid it by using Queuing, Scheduling, and Dropping features;

Provide Shaping and Policing capabilities to condition the speed the traffic enters or leaves the data center.

12.9 DMZ design

Internet connectivity is a requirement for all commercial enterprise data centers. Connectivity to the internet is through the Internet Access Block of the data center edge. The four most common uses of internet connectivity for an enterprise customer are:

- Hosting services: Servers with interfaces sitting on the internet for public access. A secure segment is created that contains web facing servers. DNS, Servers, appliances, load balancers, SMTP mail servers, and public FTP servers may also reside at this level.
- Remote Access: Businesses have a remote VPN client service for their employees or sub-contractors to be able to access the company's intranet remotely. To create this function a secure segment is also set up to support IPSec and/or SSL tunnel terminating devices.
- Internet VPN: These are typically site to site connections with partners or remote offices. This also uses IPSec and/or SSL tunnel terminating devices.
- Outbound internet access: The Internet Access Block provides outbound access to the internet. The services located on this segment can vary widely from IDS/IPs to forward proxy services, SMTP gateway and News feeds.

In order to accomplish internet connectivity a semi-trusted LAN segment is created called a demilitarized zone (DMZ). A DMZ is constructed with multiple firewall that add layers of security between the Internet and a company's critical data and business logic. Figure 12-40 illustrates a typical DMZ configuration.

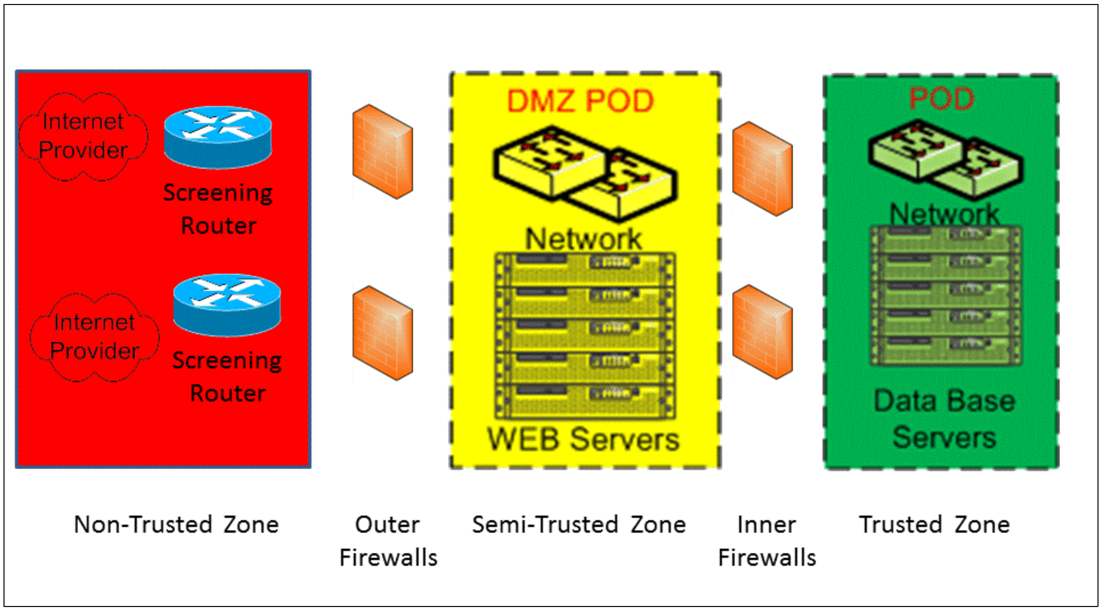


Figure 12-40 DMZ Design

The main purpose of a DMZ configuration is to protect the critical data and business logic in the data center environment from unauthorized access.

An outer firewall pair is used between the public Internet and the Web facing server or servers processing internet originated requests to the company Web site.

An inner firewall pair is also used between the Web facing server and the application servers to which it is forwarding requests. Company data resides behind the inner firewall.

The area between the two firewalls gives the DMZ configuration its name. Additional firewalls can further safeguard access to databases holding administrative and application data. Load Balancers and other appliances may also be placed at this level.

A company's security policy will dictate what is allowed from a network architecture standpoint. This security policy will dictate the physical security requirements for Internet Access and any external security services.

A more detailed example of the internet Access Block is shown in Figure 12-41.

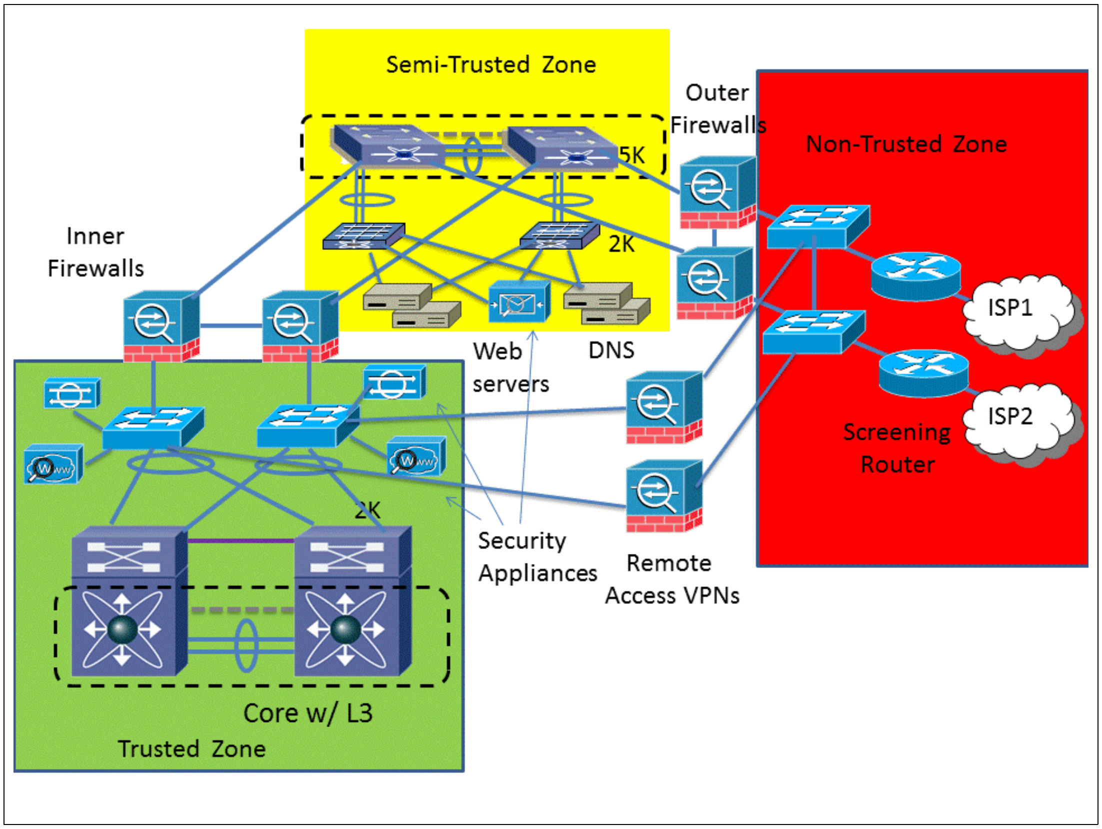


Figure 12-41 Detailed DMZ design

The Internet edge has to be designed to provide service availability and resiliency, to be compliant with regulations and standards, to provide flexibility in accommodating new services and adapt with the time, to be secure, and to facilitate administration.

12.9.1 Public-facing services

Public-facing services are placed in the DMZ zone for security and control purposes. The DMZ acts as a middle stage between the Internet and organization's private resources, preventing external users from direct access to internal servers and data.

E-mail and web server farms may be located in the trusted or semi trusted zones. Other services implemented in the DMZ often include the e-mail and web security appliances.

Tip: It is recommended that a separate internal or private DMZ be implemented to support internal clients and services. The separation of public-facing and internal services is recommended in order to ensure that appliances that provide services for internal use are not vulnerable to outside attacks.

12.9.2 Corporate Internet access

Users at the campuses access e-mail, instant messaging, web browsing, and other common services via the existing Internet edge infrastructure at the headquarters or regional offices. Depending on the organization's policies, users at the branches may also be forced to access the Internet over a centralized Internet connection, typically at the headquarters.

12.9.3 Remote access

The Internet edge infrastructure may also provide mobile users and teleworkers with access to private applications and data residing in the organization's network. This remote access is authenticated and secured with SSL or IPSec VPNs. Access control policies may also be enforced to limit access to only the necessary resources according to the user's role. Typical services provided to mobile users and teleworkers include e-mail, access to intranet websites, business applications, video on demand, IP telephony, instant messaging, and many others.

12.9.4 Internet VPN

Site to site VPN connections with partners or remote offices can also be set up through the internet Access Block. These connections also use IPSec and/or SSL tunnel terminating devices.

More information on DMZ design can be found at this website:

http://www.cisco.com/en/US/docs/solutions/Enterprise/Security/IE_DG.html

12.10 Services supporting next generation data center architectures

This section provides additional information on IBM and Cisco Professional and Managed Services that support the data center architectures that have been described earlier in this chapter.

12.10.1 IBM Global Technology Services for modern data centers

As a total solutions provider, IBM can offer a single source for the components that are necessary for a turnkey data center solution, including architecture, design or design validation, integration, logistics, ordering support and procurement, site preparation, cabling configuration, installation, system testing, and project management.

IBM Global Technology Services® (GTS) has a complete suite of services to help you assess, design, implement, and manage your data center network, server and storage.

IBM Data Center strategy, assessment, optimization, and integration services combine the IBM IT and business solutions expertise, proven methodologies, highly skilled global resources, industry-leading management platforms and processes, and strategic partnerships with other industry leaders to help you create an integrated communications environment that drives business flexibility and growth. IBM Data Center strategy, assessment and optimization services help identify where you can make improvements, recommend actions for improvements and implement those recommendations. In addition, you can:

- Resolve existing data center availability, performance, or management issues.
- Establish a more cost-effective networking and communications environment.
- Enhance employee and organizational productivity.
- Enable and support new and innovative business models.

IBM Services for data center help you position you to better meet the high-availability, high-performance and security requirements you need to stay competitive. It helps you understand, plan for and satisfy dynamic data center demands with a flexible, robust and resilient data center design and implementation. Whether you are upgrading, moving, building, or consolidating your data centers, the goal of IBM services is to help improve the success of your projects.

To help you deliver a reliable infrastructure, we offer services that include:

- architecture and design based on your data center requirements.
- A comprehensive design that integrates with your data center servers, storage, existing networking infrastructure and systems management processes.
- Project management, configuration, implementation, network cabling and system testing of server-to-network connectivity, routers, switches, acceleration devices and high-availability Internet Protocol (IP) server connections.

Options for proactive monitoring and management of your enterprise data center infrastructure to help achieve optimal levels of performance and availability at a predictable monthly cost.

For more information visit:

`http://www.ibm.com/services/us/en/it-services`

12.10.2 Cisco Data Center Networking Services

Today, the data center is a strategic asset in a world that demands better integration among people, information, and ideas. Your business and your data center work better when technology products and services are aligned with your business needs and opportunities.

Cisco and our industry-leading partners deliver intelligent, personalized services that accelerate the transformation of your data center. Using a unique combination of network and application based perspective and a unified view of data center assets, Cisco takes an architectural approach to help you efficiently consolidate, virtualize, and manage data center resources. Cisco Data Center Services help transform, optimize, and protect your data center to reduce costs, deliver high availability, and improve application performance. Cisco Data Center Networking and Virtualization teams work closely together to provide overall data center services across all business segments.

Cisco Data Center Networking Services are divided into the following categories:

- Data Center Strategic IT and Architecture Services
- IT Infrastructure Planning and Deployment Services
- Efficiency and Facilities Services
- Technical Support and Operations Management
- Data Center Optimization Services

Cisco and our industry-leading partners use best practices and proven methodologies to help you quickly and efficiently plan and deploy a high-performance, resilient, and scalable data center network for your business.

The Data Center Services are delivered by Cisco experts who hold a wide array of industry certifications and are subject matter experts in business and technology architecture and data center technologies. They have direct experience in planning, designing, and supporting data centers and virtualization solutions. Cisco product and technology expertise is continually enhanced by hands-on experience with real-life networks and broad exposure to the latest technology and implementations.

For more information, visit this website:

`http://www.cisco.com/go/services`

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this book.

IBM Redbooks publications

The following IBM Redbooks publications provide additional information about the topic in this document. Note that some publications referenced in this list might be available in softcopy only.

- *IBM Flex System and PureFlex System Network Implementation*, SG24-8089
- *Implementing an IBM/Cisco SAN*, SG24-7545
- *Implementing an IBM b-type SAN with 8 Gbps Directors and Switches*, SG24-6116
- *IBM System Storage SAN Volume Controller Best Practices and Performance Guidelines*, SG24-7521
- *IBM System Storage SAN Volume Controller and Storwize V7000 Replication Family Services*, SG24-7574
- *Implementing the IBM System Storage SAN Volume Controller V6.3*, SG24-7933
- *Implementing the Storwize V7000 and the IBM System Storage SAN32B-E4 Encryption Switch*, SG24-7977
- *IBM SAN and SVC Split Cluster and VMware Solution Implementation*, SG24-8072
- *IBM Flex System FC3171 8Gb SAN Switch and Pass-thru*, TIPS0866
- *IBM Flex System FC5022 16Gb SAN Scalable Switches*, TIPS0870
- *IBM Flex System Networking in an Enterprise Data Center*, REDP-4834
- *Moving to IBM PureFlex System x86-to-x86 Migration*, REDP-4887
- *Choosing IBM Flex System for Your Private Cloud Infrastructure*, REDP-4920
- *Positioning IBM Flex System 16 Gb Fibre Channel Fabric for Storage-Intensive Enterprise Workloads*, REDP-4921
- *IBM Flex System Manager Demonstration Guide*, REDP-4989
- *IBM Flex System Interoperability Guide*, REDP-FSIG
- *IBM PureFlex System and IBM Flex System Products and Technology*, SG24-7984
- *IBM Flex System p260 and p460 Planning and Implementation Guide*, SG24-7989
- *IBM Flex System V7000 Storage Node Introduction and Implementation Guide*, SG24-8068
- *Implementing SmartCloud Entry on an IBM PureFlex System*, SG24-8102
- *IBM Flex System EN2024 4-port 1Gb Ethernet Adapter*, TIPS0845
- *IBM Flex System x240 Compute Node*, TIPS0860
- *IBM Flex System EN2092 1Gb Ethernet Scalable Switch*, TIPS0861
- *IBM Flex System Manager*, TIPS0862
- *IBM Flex System Enterprise Chassis*, TIPS0863
- *IBM Flex System Fabric EN4093 and EN4093R 10Gb Scalable Switches*, TIPS0864
- *IBM Flex System EN4091 10Gb Ethernet Pass-thru Module*, TIPS0865

- *IBM Flex System FC3171 8Gb SAN Switch and Pass-thru*, TIPS0866
- *IBM Flex System FC3172 2-port 8Gb FC Adapter*, TIPS0867
- *IBM Flex System CN4054 10Gb Virtual Fabric Adapter and EN4054 4-port 10Gb Ethernet Adapter*, TIPS0868
- *IBM Flex System FC3052 2-port 8Gb FC Adapter*, TIPS0869
- *IBM Flex System FC5022 16Gb SAN Scalable Switches*, TIPS0870
- *IBM Flex System IB6131 InfiniBand Switch*, TIPS0871
- *IBM Flex System IB6132 2-port FDR InfiniBand Adapter*, TIPS0872
- *IBM Flex System EN4132 2-port 10Gb Ethernet Adapter*, TIPS0873
- *IBM Flex System p24L, p260 and p460 Compute Nodes*, TIPS0880
- *ServeRAID M5115 SAS/SATA Controller for IBM Flex System*, TIPS0884
- *IBM Flex System x220 Compute Node*, TIPS0885
- *IBM Flex System x440 Compute Node*, TIPS0886
- *IBM Flex System IB6132 2-port QDR InfiniBand Adapter*, TIPS0890
- *IBM Flex System FC5022 2-port 16Gb FC Adapter*, TIPS0891
- *IBM Flex System PCIe Expansion Node*, TIPS0906
- *IBM Flex System CN4058 8-port 10Gb Converged Adapter*, TIPS0909
- *IBM Flex System Fabric CN4093 10Gb Converged Scalable Switch*, TIPS0910
- *IBM Flex System EN4132 2-port 10Gb RoCE Adapter*, TIPS0913
- *IBM Flex System Storage Expansion Node*, TIPS0914
- *IBM SmartCloud Desktop Infrastructure: VMware View on IBM Flex System*, TIPS0920
- *IBM SmartCloud Desktop Infrastructure: Citrix XenDesktop on IBM Flex System*, TIPS0928
- *IBM SmartCloud Desktop Infrastructure: IBM Virtual Desktop on IBM Flex System*, TIPS0929
- *Connecting an IBM PureFlex System to the Network*, TIPS0941
- *Implementing the IBM Storwize V7000 Unified*, SG24-8010

You can search for, view, download or order these documents and other Redbooks, Redpapers, Web Docs, draft and additional materials, at the following website:

ibm.com/redbooks

Other publications

These publications are also relevant as further information sources:

Towards an Open Data Center with an Interoperable Network (ODIN), Volume 3: Software Defined Networking and OpenFlow:

http://www.ibm.com/common/ssi/cgi-bin/ssialias?subtype=WH&infotype=SA&appname=STGE_QC_QC_USEN&htmlfid=QCW03021USEN&attachment=QCW03021USEN.PDF

Cisco Announces Open Network Environment to Unleash Application-Driven Network Programmability - See more at this website:

http://newsroom.cisco.com/uk/press-release-content?articleId=892363&type=webcontent#sthash.LKyRV8An.dpuf

Cisco Open Network Environment:

http://www.cisco.com/en/US/prod/collateral/iosswrel/content/at_a_glance_c45-707979.pdf

OpenFlow: Enabling Innovation in Campus Networks:

http://www.openflow.org/documents/openflow-wp-latest.pdf

Introduction to OpenFlow:

https://www.opennetworking.org/standards/intro-to-openflow

The Rise in Campus Networking Slicing Among Universities:

http://www.cisco.com/en/US/prod/collateral/iosswrel/content/campus_networking.pdf

Nexus 1000V and Virtual Network Overlays Play Pivotal Role in Software Defined Networks:

http://blogs.cisco.com/datacenter/nexus-1000v-and-virtual-network-overlays-play-pivotal-role-in-software-defined-networks/

Cisco Nexus 3548 Switch Data Sheet:

http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps11541/ps12581/data_sheet_c78-707001.html

Cisco Nexus 2000 Series Fabric Extenders Data Sheet:

http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps10110/data_sheet_c78-507093.html

CISCO MDS 9000 family SAN OS 2.0(X):

http://www.cisco.com/en/US/prod/collateral/ps4159/ps6409/ps5989/ps6217/product_data_sheet09186a00801bcfd8.pdf

Cisco Nexus 7000 Series switches:

http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9402/ps9512/Data_Sheet_C78-437762.pdf

Cisco Nexus 5548P, 5548UP, 5596UP, and 5596T switches:

http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9670/data_sheet_c78-618603.pdf

Cisco Nexus 3000 Series switches overview:

http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps11541/ps12581/data_sheet_c78-707001.html

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

IBM

Redbooks

IBM and Cisco: Together for a World Class Data Center

(1.0” spine)
0.875”<->1.498”
460 <-> 788 pages

IBM and Cisco

Together for a World Class Data Center

Read about how IBM and Cisco are collaborating in the data center

Learn about the technologies and products that are leveraged

See practical use case examples

This IBM Redbooks publication is an IBM and Cisco collaboration that articulates how IBM and Cisco can bring the benefits of their respective companies to the modern data center.

It documents the architectures, solutions, and benefits that can be achieved by implementing a data center based on IBM server, storage, and integrated systems, with the broader Cisco network.

We describe how to design a state-of-the art data center and networking infrastructure combining Cisco and IBM solutions. The objective is to provide a reference guide for customers looking to build an infrastructure that is optimized for virtualization, is highly available, is interoperable, and is efficient in terms of power and space consumption. It will explain the technologies used to build the infrastructure, provide use cases, and give guidance on deployments.

SG24-8105-00 ISBN 0738438421