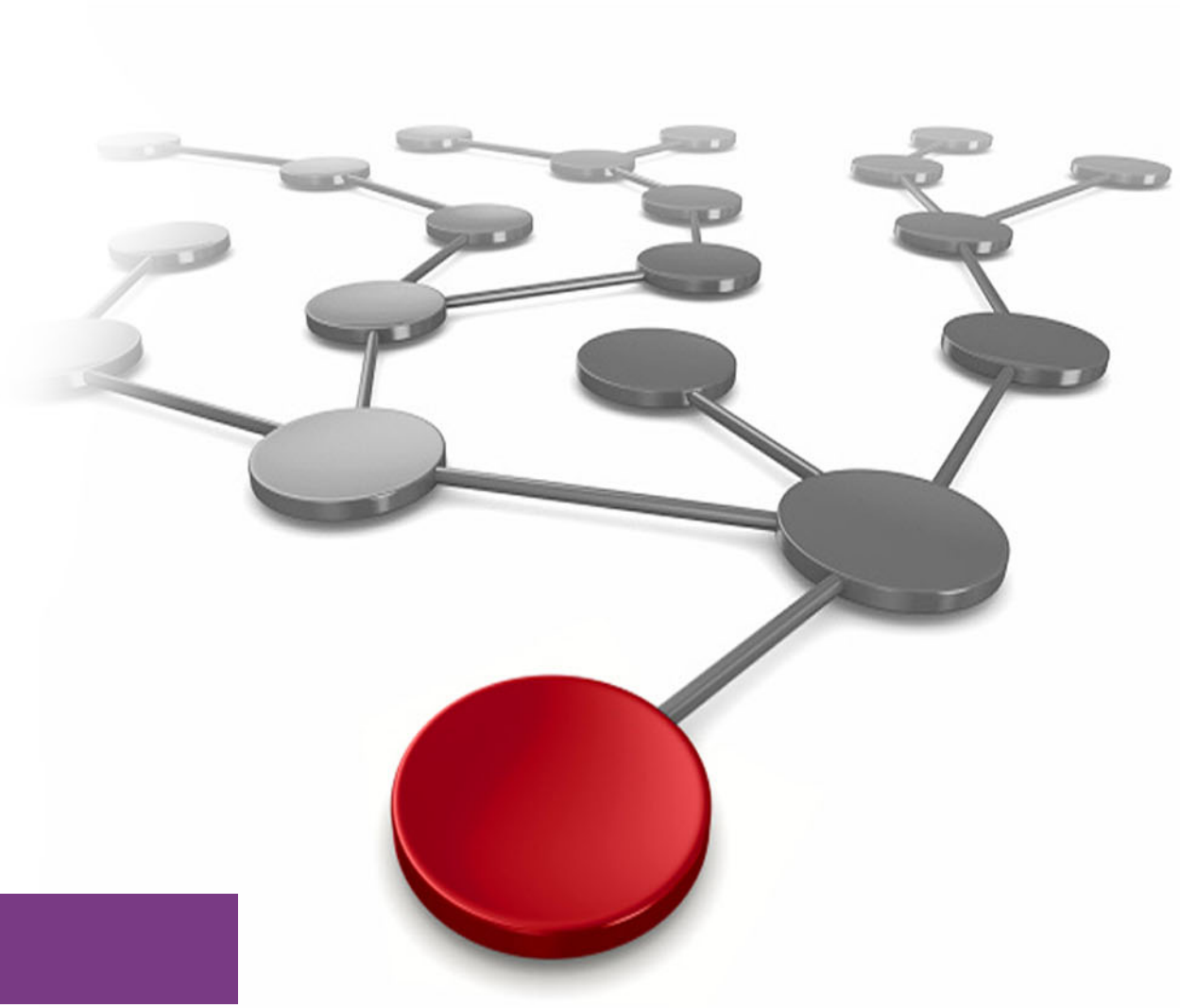
Redbooks
ibm.com/redbooks

SAN and Fabric Resiliency Best Practices for IBM c-type Products

David Green
David Lutz
Lyle Ramsey
Bhavin Yadav



Storage

IBM

Redpaper

Fabric best practices for using IBM SAN-c switches

This IBM Redpaper publication describes best practices for deploying and using advanced Cisco NX-OS features to identify, monitor, and protect Fibre Channel (FC) Storage Area Networks (SANs) from problematic devices and media behavior.

NX-OS: This paper focuses on the IBM c-type SAN switches with firmware Cisco MDS NX-OS Release 8.4(2a).

Introduction

Faulty or improperly configured devices, misbehaving hosts, and faulty or substandard FC media can significantly impact the performance of FC fabrics and the applications they support. In most real-world scenarios, these issues cannot be corrected or completely mitigated within the fabric itself. Instead, the behavior must be addressed directly. However, with the proper knowledge and capabilities, the fabric can often identify and in some cases, mitigate or protect against the effects of these misbehaving components to provide better fabric resiliency.

This document concentrates specifically on *Port-Monitor* function (and related capabilities) that help provide optimum fabric resiliency using Cisco Data Center Network Manager (DCNM) for IBM c-type and Cisco MDS 9000 series switches.

For more information about the features that are described in this publication, see the product documents that are appropriate for your NX-OS release. Cisco documentation can also be found by searching the Cisco website.

- *NX-OS Administrator's Guide*

  https://www.cisco.com/c/en/us/support/storage-networking/mds-9000-nx-os-san-os-software/products-installation-and-configuration-guides-list.html

- *NX-OS Command-Line Reference*

  https://www.cisco.com/c/en/us/td/docs/switches/datacenter/mds9000/sw/8_x/command/cisco_mds9000_command_ref_8x.html

- *Cisco DCNM Administrator's Guide*

  https://www.cisco.com/c/en/us/support/cloud-systems-management/prime-data-center-network-manager/products-installation-and-configuration-guides-list.html

Factors that affect fabric resiliency

Several common types of abnormal behavior originate from fabric components or attached devices:

- Faulty media (fiber-optic cables and Small Form-Factor Pluggables [SFPs]/optics): Faulty media can cause frame loss due to excessive Cyclic Redundancy Check (CRC) errors, Forward Error Correction (FEC) errors, invalid transmission words, and other conditions, which can result in I/O failure and application performance degradation.
- Misbehaving devices, links, or switches: Occasionally, a condition arises where a device (server or storage array) or link (inter-switch link (ISL)) behaves erratically and causes disruptions in the fabric. If not immediately addressed, this situation might result in severe stress on the fabric.
- Congestion: Congestion is caused by latencies or insufficient link bandwidth. End devices that do not respond as quickly as expected can cause the fabric to hold frames for excessive periods, which can result in application performance degradation or, in extreme cases, I/O failure.
- Credit loss: Credit loss occurs when either of the following conditions occur:
  - The receiving port does not recognize a frame (usually due to bit errors), so it does not return an R_RDY.
  - The sending port does not recognize the R_RDY (usually due to link synchronization issues).

Faulty media

In addition to high-latency devices that cause disruptions to data centers, fabric problems are often the result of faulty media. Faulty media can include bad cables, SFPs, extension equipment, receptacles, patch panels, improper connections, and so on. Media can fault on any SAN port type and fail, often unpredictably and intermittently, making it even harder to diagnose. Faulty media that involves server/host and storage device ports (F_Ports) results in an impact to the end device attached to the F_Port and to devices that communicate with this device.

Failures on ISLs or E_Ports can have an even greater impact. Several flows (host and target pairs) can simultaneously traverse a single E_Port. In large fabrics, this can be hundreds or thousands of flows. If a media failure occurs that involves one of these links, it is possible to disrupt some or all of the flows that use the affected path. Severe cases of faulty media, such as a disconnected cable, can result in a complete failure of the media, which effectively brings a port offline. This situation is typically easy to detect and identify. When it occurs on an F_Port, the impact is specific to flows that involve the F_Port. E_Ports are typically redundant, so severe failures on E_Ports typically only result in a minor drop in bandwidth because the fabric automatically uses redundant paths. Also, error reporting that is built into the operating system readily identifies the failed link and port, which allows for simple corrective action and repair. With moderate cases of faulty media, failures occur, but the port can remain online or transition between online and offline. This situation can cause repeated errors, which can occur indefinitely or until the media fails. When these types of failures occur on E_Ports, the result can be devastating because there can be repeated errors that impact many flows. This can result in significant, long-lasting impacts to applications.

Signatures of these types of failures include:

- CRC errors on frames
- Invalid Transmission Words

- State Changes (ports that go offline or online repeatedly)
- Credit loss on an E_Port prevents traffic from flowing on that E_Port, resulting in frame loss and I/O failures for devices that are crossing that link

Misbehaving devices

Another common class of abnormal behavior originates from high-latency end devices (host or storage). A high-latency end device is one that does not respond as quickly as expected and thus causes the fabric to hold frames for excessive periods. This situation can result in application performance degradation or, in extreme cases, I/O failure. Common examples of moderate device latency include disk arrays that are overloaded and hosts that cannot process data as fast as requested. Misbehaving hosts, for example, become more common as hardware ages. Bad host behavior is usually caused by defective host bus adapter (HBA) hardware, bugs in the HBA firmware, and problems with HBA drivers. Storage ports can produce the same symptoms due to defective interface hardware or firmware issues. Some arrays deliberately reset their fabric ports if they are not receiving host responses within their specified timeout periods. Severe latencies are caused by badly misbehaving devices that stop accepting, or acknowledging frames for excessive periods. However, with the proper knowledge and capabilities, the fabric can often identify and, in some cases, mitigate or protect against the effects of these misbehaving components to provide better fabric resiliency.

Congestion

Congestion occurs when frames that are carried on a link cannot be immediately transmitted.

The following situations can be a source of congestion:

1. Links, hosts, or storage respond more slowly than expected. Therefore, they do not return buffer credits to the sender quickly enough. This is commonly called a *slow-drain condition*.
2. More data arrives for a port than the port can transmit at the current link speed. This is called *overutilization*.

Congestion results in fabric latencies. As FC-link bandwidth increases from 1 - 64 Gbps, instances of insufficient link bandwidth capacities radically decreases.

Latency occurs when devices respond more slowly than they should. These devices are the major source of congestion in today's fabrics due to their inability to promptly return buffer credits to the switch.

Slow-drain devices

A device that is *slow drain* responds more slowly than expected. The device does not return buffer credits (through R_RDY primitives) to the transmitting switch fast enough to support the *offered load*, even though the offered load is less than the maximum physical capacity of the link that is connected to the device.

Figure 1 illustrates the condition where a buffer backup on ingress port 6 on B1 causes congestion upstream on S1, port 4. When all available credits are exhausted, the switch port that is connected to the device must hold additional outbound frames until a buffer credit is returned by the device.

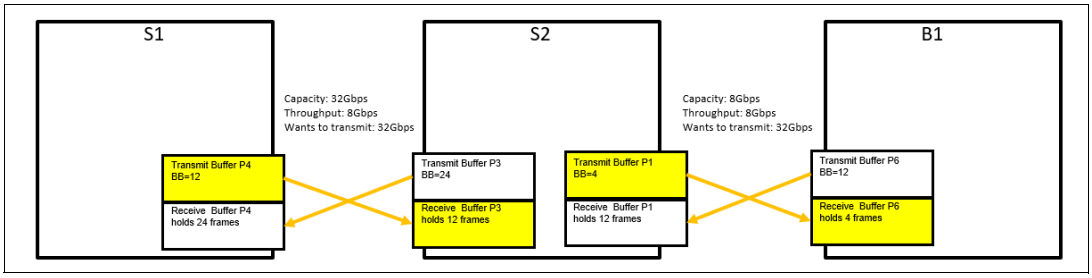


Figure 1 Device slow drain example

When a device does not respond in a timely fashion, the transmitting switch is forced to hold frames for longer periods, which result in high buffer occupancy. In turn, this results in the switch lowering the rate at which it returns buffer credits to other transmitting switches. This effect propagates through switches (and potentially multiple switches, when devices attempt to send frames to devices that are attached to the switch with the high-latency device) and ultimately affects the fabric.

Figure 2 shows how latency on a switch can propagate through the fabric.

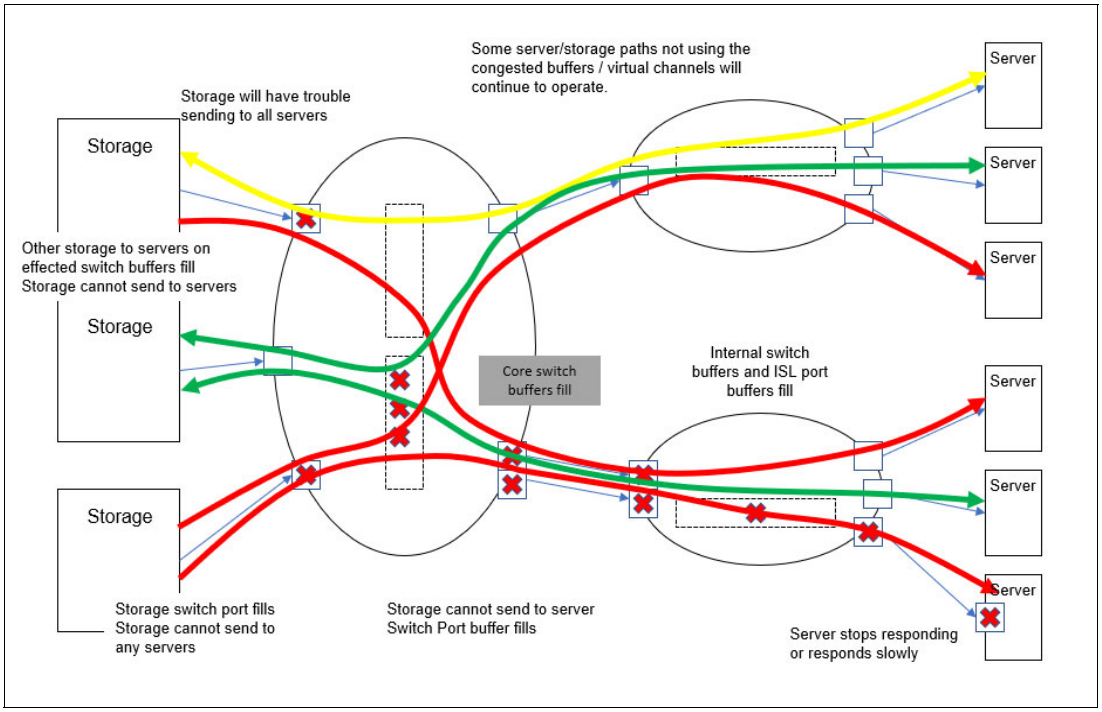


Figure 2 Latency on a switch can propagate through the fabric

Note: The impact to the fabric (and other traffic flows) varies based on the severity of the latency that is exhibited by the device. The longer the delay that is caused by the device in returning credits to the switch, the more severe the problem.

Moderate device latencies

Moderate device latencies from the fabric perspective are defined as those not severe enough to cause frame loss. If the time between successive credit-returns by the device is between a few-hundred-microseconds to tens-of-milliseconds, the device exhibits mild-to-moderate latencies because this delay is typically not enough to cause frame loss.

This situation causes a drop in application performance, but typically does not cause frame drops or I/O failures.

The effect of moderate device latencies on host applications might still be profound, based on the average disk service times that are expected by the application. Mission-critical applications that expect average disk service times of, for example, 3 ms, are severely affected by storage latencies in excess of the expected service times. Moderate device latencies have traditionally been difficult to detect in the fabric. Advanced monitoring capabilities, such as Cisco SAN Analytics, that are implemented in the 32 Gbps platforms make it much easier to detect and send alerts for the latency.

Severe device latencies

Severe device latencies result in frame loss, which triggers the host Small Computer System Interface (SCSI) stack to detect failures and to retry I/Os. This process can take tens of seconds (possibly as long as 30 to 60 seconds), which can cause a noticeable application delay and potentially results in application errors. If the time between successive credit returns by the device is in excess of 500 ms, the device is exhibiting severe latency. When a device exhibits severe latency, the switch is forced to hold frames for excessively long periods (on the order of hundreds of milliseconds). When this time becomes greater than the established timeout threshold, the switch drops the frame (per FC standards).

Because the effect of device latencies often spreads through the fabric, frames can be dropped due to timeouts, not just on the F_Port to which the misbehaving device is connected, but also on E_Ports carrying traffic to the F_Port. Dropped frames typically cause I/O errors that result in a host retry, which can result in significant decreases in application performance. The implications of this behavior are compounded and exacerbated by the fact that frame-drops on the affected F_Port (device) result not only in I/O failures to the misbehaving device (which are expected), but also on E_Ports. This might cause I/O failures for unrelated traffic flows that involve other hosts (and typically are not expected).

Latencies on ISLs

Latencies on ISLs are usually the result of back pressure from latencies elsewhere in the fabric. The cumulative effect of many individual device latencies can result in slowing the link. The link itself might be producing latencies if it is a long-distance link with distance delays or there are too many flows that use the same ISL. Although each device might not appear to be a problem, the presence of too many flows with some level of latency across a single ISL or trunked ISL might become a problem. Latency on an ISL can ripple through other switches in the fabric and affect unrelated flows.

NX-OS can provide alerts and information that indicate possible ISL latencies in the fabric, through one or more of the following items:

- Time-out discards on the device E_Port or TE_Port carrying the flows to and from the affected F_Port or device
- Port Monitor alerts, if they are configured for timeouts
- Elevated time Tx buffers are 0 on the affected E_Port, which also might indicate congestion

Credit loss

Buffer credits are a part of FC flow control and the mechanism that FC connections use to track the number of frames that are sent to the receiving port. Each time a frame is sent, the credit count is reduced by one. When the sending port runs out of credits, it is not allowed to send more frames to the receiving port. When the receiving port successfully processes a frame and frees up the buffer where the frame was stored, it tells the sending port that it has

the frame by returning an *R_RDY* primitive. When the sending port receives an R_RDY, it increments the credit count.

Credit loss occurs when either of these conditions occur:

- The receiving port does not recognize a frame (usually due to bit errors), so it does not return an R_RDY
- The sending port does not recognize the R_RDY (usually due to link synchronization issues).

FC links are never perfect, so the occasional credit loss can occur, but it becomes an issue only when all available credits are lost and congestion becomes severe. When credit loss occurs on links, it is usually caused by faulty media. The switch automatically tries to recover from a complete loss of credits on links by resetting the link. It resets the link after a credit loss for 1 second on a device port and 1.5 seconds on an ISL.

High-performance networks

With the use of low latency Solid State Drives (SSD) and Flash controllers, the performance of the SAN becomes critical to achieving the full performance potential from those technologies. Eliminating latency from the SAN requires a level of planning and consideration that is often above what is necessary for traditional enterprise class storage, given the nominal operating ranges of those devices.

Poorly constructed and maintained SANs can add latency to the SCSI exchange completion times to varying degrees. This additional latency can often go undetected or be considered insignificant for “spinning disk” subsystems as it is often a small percentage of the response time (that can be in the 10’s to 100’s of milliseconds) that those devices are capable of achieving. This is not true of SSD and Flash storage where the latency contribution from sub-optimal SAN conditions can equal or exceed the capable response time for those technologies.

The Fabric Resiliency best practices that are discussed in this paper are especially critical as they pertain to maintaining a high-performance SAN. However, in addition to those practices, SAN-design considerations exist, regarding the use of mixed speed-devices and ISLs.

Mixed speed SANs

To enable the technology to be refreshed from one generation to the next, it is generally required for multiple device speeds to exist in the SAN. The existence of mixed speed devices cannot be avoided; however, mixed speed devices that span more than one generation of technology should be avoided. For example, mixing 4 Gbps and 8 Gbps devices is generally acceptable; however, mixing 4Gbps and 16 Gbps is not. The speed-matching that is required to accommodate these large speed differentials introduces latency and potential congestion points that can significantly degrade the performance and stability of SAN.

ISLs and multi-hop ISLs

Many flows between servers and storage or storage-to-storage devices need to flow across the ISLs. As a result, ISLs are notorious for introducing latency into the transmission flows. The size and, more importantly, the number of ISLs that is required between switches must consider both response time requirements and bandwidth requirements. With storage devices getting into sub-millisecond response times, ISLs need to make sure that credits are always available, so frames will not be delayed. This might require multiple ISL trunks instead of fewer larger-bandwidth trunks.

If frames need to traverse multiple switches to reach their destination, delays can be introduced with each hop that is required between the source and destination switches. IBM recommends that the number of hops not exceed one for devices crossing ISLs, especially where strict performance requirements exist. The use of ISLs for access to SSD or Flash should be avoided altogether.

Best Practice:

- Avoid introducing devices to the SAN that span more than one generation of technology.
- Avoid traversing multiple hops between switches when you access SSD or Flash for high-performance use cases.

Designing resiliency into the fabric

This document is not intended to cover the general set of design considerations that are required for designing a SAN, but design elements must be considered to ensure that the fabric is resilient by design.

This section includes preferred practices for each of the following design elements:

- Fabric topology, including Core versus Edge switches and fabrics that span multiple sites
- Zoning recommendations
- Port-Channels
- VSAN trunking
- Meaningful naming convention

Fabric topology

The topology of your fabric and where devices are connected to it can affect the overall resiliency of your fabric.

The following basic designs apply to fabrics:

- Core/Edge
- Edge/Core/Edge
- Dual-Core/Edge
- Mesh

Some variances to each of these designs exist, but all fabrics fall into some combination of these designs. Occasionally fabrics are implemented as a hybrid, such as a core/edge design, but the edge switches are interconnected to each other. This is generally not recommended because it complicates the fabric and creates opportunities for mistakes to be made. A hybrid fabric also makes it more difficult to diagnose problems.

Core/Edge

In a core/edge fabric design, there is a clear hierarchy of switches. The core switches sit at the center of a fabric. Critical hosts and the storage systems are connected to the core switch. Edge switches with less critical hosts are then connected to the core switch. It is recommended that your core switches be director-class switches with redundant, hot-swappable components. Smaller switches that do not have redundant components should be relegated to be edge switches. Figure 3 shows a typical core/edge fabric design.

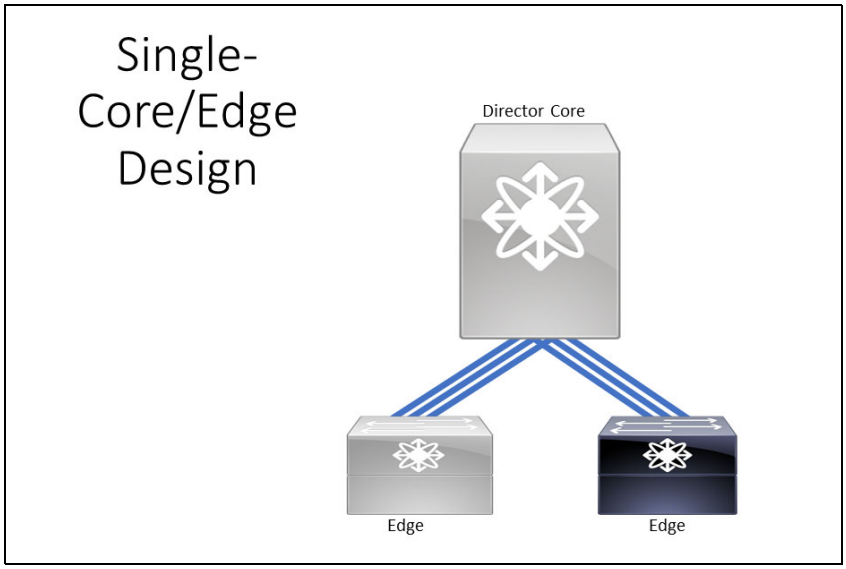


Figure 3 Core/Edge Design

There are two variants on the Core/Edge design: Edge/Core/Edge and Dual-Core Edge.

Edge/Core/Edge

In this design, the core switch is a Director-class switch. The edge switches can be smaller edge switches or director-class switches. The topology diagram for an Edge/Core/Edge fabric looks similar to Figure 3. However, in this design the core switch does not have any end devices connected to it. It is used as backbone switch. End devices are connected to the edge switches. Figure 3 depicts only two edge switches connected to the core. However in this topology, there are usually several edge switches connected to the core switch.

This design has the advantage of having multiple edge-switches that are installed closer to the attached devices, which reduces cabling. This design has the disadvantage of more traffic crossing the ISLs and more points of congestion.

Dual-Core/Edge

In this design, a fabric has two core switches. The core switches are connected by using high-bandwidth ISLs. Edge switches are connected to one or, more commonly, both of the core switches. Critical hosts and storage ports are connected to both core switches. This is not a common design, but some fabrics are implemented in this manner.

When you use this design, take care with zoning to not allow unnecessary traffic across the ISLs between the core switches. Examples of this include hosts that take extra hops to get to storage ports on the other core switch, or storage virtualization products that access storage ports across the ISLs between the core switches.

Because of the risk of unnecessarily zoning devices across the ISLs between the cores, it is preferred that dual-core fabrics be used only during technology refreshes to migrate to newer switches. After the migration is complete, split the dual-core fabrics into separate fabrics.

An advantage to a dual-core edge design is that the impact of core-switch failure is reduced because the second core switch continues to allow flows to occur on the fabric. However, for similar availability, IBM prefers a quad-fabric design to a dual-core design.

Mesh design

In a mesh fabric, the switches are all interconnected. Some issues with a mesh design are that it can be harder to troubleshoot because there is not a clear hierarchy of switches as there is in a core/edge design, and often devices on different switches must traverse multiple hops.

In a core/edge design that is implemented properly, as in Figure 3, a host attached to the edge switch is only one hop away from its storage.

Figure 4 shows a typical mesh fabric. Some of the switches are two hops away from each other. Therefore, the preferred topology is a core/edge fabric design.

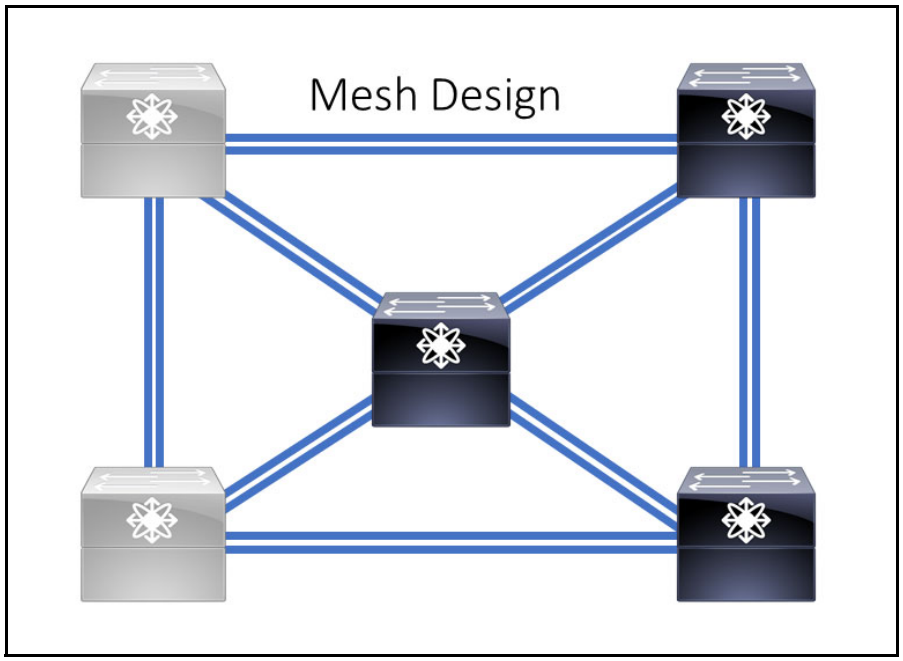


Figure 4 Mesh Design

Best Practice: A core/edge design is preferred for most fabrics.

Multi-site fabrics

Fabrics frequently span multiple sites. This is done for redundancy or, more commonly, to provide a means for storage replication between sites. The links between these sites are typically lower bandwidth and use technologies such as Fibre-Channel over IP (FCIP).

The underlying infrastructure for these links is more vulnerable to issues and congestion than the local links in each of the local data centers. As a result, several items must be considered in the design to ensure maximum resiliency.

Traffic separation

Isolating different types of traffic to separate virtual SANs (VSANs) should be done on all fabrics, but it is especially important on multi-site fabrics. The most common use for inter-site links is storage replication. This traffic should be in its own VSAN and should ideally traverse separate links than production traffic, which traverses the links between sites. Keeping replication traffic in its own VSAN also ensures that storage-based replication is using

dedicated ports for replication, which are in the dedicated VSAN and cannot be zoned to hosts.

For some Storage System implementations (such as Spectrum Virtualize clusters in a Hyperswap Topology), separate ISLs is a requirement for the internode traffic between the sites.

For details on fabric implementation for Hyperswap clusters, see:

http://www.redbooks.ibm.com/abstracts/redp5597.html?Open

Distance link design

Links between sites typically traverse long distances and are often implemented by using Fibre-Channel over IP (FCIP). This protocol allows FC traffic to traverse existing IP networks between sites and does not require installing costly dedicated links for the FC traffic. However, FC is a lossless protocol that demands a reliable and speedy network. IP networks typically have much higher latencies than FC networks, and wide area network (WAN) providers can give two different routes between your sites, where one has significantly higher latency than the other. This is much less common when you use long-distance FC links because those tend to be more direct routing.

A resilient multi-site fabric design should have inter-site links where the latency is nearly the same across all the links. If there are latency differences, they should be within at least several milliseconds of each other. Use multiple vendors for your inter-site links. IBM favors implementing the fabrics so that one redundant fabric uses the links from one vendor, and the other fabric uses the links from the other vendor. This approach ensures that if one vendor has a complete outage, you will still have one fabric that can transmit data between your sites.

Cisco MDS and IBM c-Type switches support either having both vendors carry both fabrics, or implementing the inter-site links so that one fabric is on one vendor and the other fabric is on the second vendor. This implementation is preferred over mixing the vendors on each fabric because some storage systems and hosts have better tolerance for a complete failure on one fabric instead of partial failure on both. Replication continues more reliably on the surviving fabric rather than having both fabrics impacted by a failure.

If the fabrics are implemented so that the links from both vendors are carrying both fabrics, the devices will have connections to a given port on a remote device that uses both vendors. If there is a loss of site connectivity because of one of the vendors, the device connectivity to that port will be partial. Only some connections to that port will be lost.

Using IVR for multi-site fabrics

By default, devices in different Virtual SANs (VSAN) on an IBM c-Type or Cisco fabric cannot communicate with each other. Each VSAN is its own logical fabric with its own zoning, name server and other fabric services. Cisco *Inter-VSAN Routing* (IVR) is a Cisco feature that allows you to specify specific devices to export into another VSAN. With IVR, the devices can communicate without a full fabric merge between the two VSANs, which keeps the VSANs isolated and prevents problems in one from propagating to the other. The most common use case for this feature is multi-site fabrics.

The links between the sites are usually lower bandwidth links. IVR enables devices at each site to communicate while preventing a full fabric merge. This keeps the fabric-management related data to a minimum, which then prevents changes at one fabric from propagating to the other site and isolates the sites from disruptions at another site.

It is recommended that you implement IVR for devices in your production VSANs that need to communicate with devices at the remote site instead of allowing the fabrics to merge.

VSANs that are dedicated to only replication will typically have only a few storage ports with no other traffic using them. These VSANs are also typically set up and not changed frequently after initial setup. Devices that should not communicate to remote sites will not be logged into replication VSANs. Since all devices should communicate with remote devices, and because there are few devices that are logged into replication VSANs and because they do not change frequently, fabric management propagation is also infrequent. Implementing IVR in these cases is not recommended. IVR adds complexity, which usually results in a decrease in resiliency.

For details on IVR, see the following configuration guide:

https://www.cisco.com/c/en/us/td/docs/switches/datacenter/mds9000/sw/8_x/config/ivr/cisco_mds9000_ivr_config_guide_8x.html

Best Practice: Separate replication traffic from production traffic on mult-site links. For production VSANs, use the IVR feature to prevent fabrics from merging across sites.

Port-channels

Port-channels are used on IBM c-Type and Cisco switches to aggregate separate ISLs together into a single logical link. All members of the port-channel must be of the same technology (FC or FCIP) and must have the same connected speed.

A port-channel optimizes the use of bandwidth by allowing a group of ISLs to merge into a single logical link. Traffic is distributed evenly and in order over a port-channel, achieving greater performance with fewer links. Within the port-channel, multiple physical links appear as a single link, which simplifies management. A port-channel can be assigned to a VSAN, or it can be shut down without having to configure each individual link that is a member of the port-channel. N-Ports that are connected to N Port Virtualization (NPV) devices can also be added to port-channels.

Port-channels provide excellent protection from credit that is lost on ISLs as each link in the port channel maintains and manages its own pool of buffer credits. If credit loss occurs on an ISL, frames continue to flow by using other links in the port-channel until the switch can detect the credit loss (typically 2 seconds) and perform a link reset to recover the credits.

More IT environments are relying on server virtualization technologies that can share host adapter connections. Specifically, N_Port ID Virtualization (NPIV) allows many clients (servers, guest, or hosts) to use a single physical port on the SAN.

Each of these communications paths from server, virtual, or otherwise, is a data flow that must be considered when planning for how many interswitch links are needed. These virtualized environments often lead to a situation where there are many data flows from the edge switches, potentially leading to frame-based congestion if there are not enough ISL or trunk resources.

Best Practice: Group the ISLs into a port-channel between switches instead of using individual ISLs.

VSAN trunking

On IBM c-Type and Cisco fabrics, an ISL, port-channel, or FCIP tunnel is trunking when it is allowing traffic from multiple VSANs to traverse the link. Example 1 shows an FCIP tunnel that is configured for trunking and allowing multiple VSANs across the tunnel.

Example 1 Trunking Interface

```
switch1# show int fcip4
fcip4 is trunking
    Hardware is IPStorage
    Port WWN is 20:5a:00:2a:6a:a4:c2:00
    Peer port WWN is 20:2b:00:2a:6a:1b:50:80
    Admin port mode is auto, trunk mode is on
    snmp link state traps are enabled
    Port mode is TE
    Port vsan is 1
    Operating Speed is 10000 Mbps
    Belongs to port-channel2
    Trunk vsans (admin allowed and active) (1,3-4,10)
    Trunk vsans (up)                       (1,3-4,10)
    Trunk vsans (isolated)                 ()
    Trunk vsans (initializing)             ()
    Using Profile id 12  (interface IPStorage2/1)
```

Trunking allows multiple VSANs to traverse the same set of port-channels or ISLs but still maintain separation of the VSANs. The default setting is for any interface, port-channel, or FCIP tunnel to have a trunking mode of auto. This means that if there are multiple VSANs configured on a switch, when that switch is connected to another switch, it attempts to allow the VSANs to merge across the link. If VSANs are not defined on both switches, they become isolated on that trunking ISL or port-channel.

Care must be taken when you enable trunking mode and configure which VSANs are allowed across a port-channel or ISL, especially for lower-bandwidth long-distance links. Even if devices are not zoned to each other across the links, VSANs attempt to merge if the same VSAN is defined on each switch. Fabric changes, zoning changes, and so on, will propagate to the other site. Additionally, if separate VSANs have been implemented to carry replication and production traffic between sites, they should be routed across different port-channels and trunking should either be turned off or carefully configured to not allow the two VSANs to traverse the same links.

Best Practices: Only configure trunking on an ISL or Port-Channel if the ISL or Port-Channel needs to carry traffic for multiple VSANs. For ISLs and Port-Channels that are trunking, configure trunking to allow the VSANs that need to access the ISL or Port-Channel.

For example, replication traffic should not be trunked with production traffic and FICON® traffic should not be mixed with open system traffic on the same ISL or port-channel.

Routing policies for open-systems fabrics

The routing policy determines the route or path that frames take when they traverse the fabric.Two routing policies are available for IBM c-Type and Cisco fabrics.

- Exchange-Based Routing (EBR)
- Source-Destination Routing (SDR)

EBR is the default routing policy and is always the preferred routing policy for Fibre-channel protocol (FCP) fabrics. The basic unit of transmission for FCP is frames. A group of frames is called a *sequence*. One or more sequences is called an *exchange*. When a sender starts a new exchange, the fabric decides how to route that exchange through the fabric and which ISLs to use. Different exchanges between the same two endpoints can traverse different physical ISLs in a port-channel.

SDR uses the Source ID and Destination ID in the frames to make the routing decision. The fabric uses the same route through the fabric for all traffic between a Source and Destination. This routing policy can cause traffic to stack up on the same few ISLs in a port-channel if traffic from multiple Source/Destination pairs is routed across the same ISLs. EBR is the preferred routing policy because it provides for better load-balancing across the links in a port-channel and more consistent performance.

One of the few use cases for SDR might be distance links, where links between switches on a fabric have latencies that are significantly different. As an example, a fabric with multiple FCIP tunnels might have one tunnel with 25 ms latency while another tunnel has 45 ms latency. In these cases with EBR, commands between two end-devices (most commonly replication between storage systems) can arrive out of order. If this happens, the devices wait for the preceding commands before processing the first command. This means that the fabric is performing at the speed of the link with the highest latency.

The use of SDR ensures in-order delivery of commands because all traffic between two end-device ports would follow the same route through the fabric. However, there is a risk of increased latency on the links. Also, there is a risk of some links being under-utilized if the switches decide to route most or all of the traffic between multiple port pairs down the same link or links. The recommended fix is to implement the fabric so that differing latency does not exist on the distance links.

Best Practice: Keep the default exchanged-based routing policy in place unless your implementation has a specific requirement for source-destination routing. For the case involving different latencies on distance links, correct the issues with the distance links to minimize the difference in latency between the links.

Meaningful naming convention

Cisco fabrics have features that allow SAN administrators to assign meaningful names or descriptions to switches, zones and zonesets, switch ports (interfaces), and the devices that are attached to the switches. It is critical to use a consistent and meaningful naming convention to maintain a reliable storage network. A naming convention enhances the ease of administration of the fabric and makes troubleshooting easier. Increasing the ease of administration and increasing the serviceability of the fabric makes the fabric more reliable. Meaningful names that are assigned to devices should be considered and implemented in a well thought-out manner. A naming scheme that is both documented and consistent should be developed.

The naming convention will likely vary for each fabric depending on the needs of the administrators. However, a good naming convention includes consistent naming for both the switches and the devices on the fabric. It is also recommended that VSANs be named. Assign meaningful names to both VSANs and the switches to greatly reduce confusion when connecting devices to the fabric.

The naming convention will likely vary for each fabric depending on the needs of the administrators. However, a good naming convention includes consistent naming for both the switches and the devices on the fabric. It is also recommended that VSANs be named. Assign meaningful names to both VSANs and the switches to greatly reduce confusion when connecting devices to the fabric.

For devices, Cisco fabrics have three options for creating a naming convention.

- fcalias: this is applied to devices World-Wide Port Name (WWPN)
- device-alias: this is applied to devices (WWPN)
- port description: this is applied to switch ports

The port description should be used with either the fcalias or the device alias. It is recommended that device-aliases be used.

fcalias

The *fcailias* can be used only for zoning devices together. It can contain multiple WWPNs under the same alias, and can span multiple switches but it is VSAN-specific. This means that it is configured per VSAN. it is possible that the same device WWPN can have different fcaliases in different VSANs. The fcalias has limited use since it can only be used for zoning. It is only distributed when full-zoneset distribution is enabled. If you are only distributing the active zoneset, fcaliases will not be distributed to the VSAN. This means that if you are zoning using fcalias, you must enable full zoneset distribution for that VSAN. Depending on how many zonesets are in the full zoneset, it can be large. This can be a concern if you have smaller switches in your fabric that have limited memory to store the full zone database for all VSANs. Example 2 shows a listing of fcaliases defined on a fabric.

Example 2 Listing of fcaliases defined on a fabric

```
fcalias name SVC_Node1_Port3 vsan 1
  pwwn 50:05:07:68:01:10:40:2d

fcalias name SVC_Node1_Port4 vsan 1
  pwwn 50:05:07:68:01:20:40:2d

fcalias name SVC_Node2_Port3 vsan 1
  pwwn 50:05:07:68:01:10:40:24

fcalias name SVC_Node2_Port4 vsan 1
  pwwn 50:05:07:68:01:20:40:24

fcalias name SVC_Node3_Port3 vsan 1
  pwwn 50:05:07:68:01:10:40:03

fcalias name SVC_Node3_Port4 vsan 1
  pwwn 50:05:07:68:01:20:40:03

fcalias name SVC_Node4_Port3 vsan 1
  pwwn 50:05:07:68:01:10:40:df
```

device-alias

Unlike the fcalias, the device-alias is not VSAN-specific. It is distributed via Cisco Fabric Services (CFS). It can be used for multiple Cisco fabric functions, such as port security and Cisco IVR, and can also be used for zoning. When device aliases are used, the full zoneset does not have to be distributed.

Example 3 shows a device-alias database listing.

Example 3 device-alias database listing

```
device-alias name tim pwwn 22:22:22:22:44:44:44:44
device-alias name tom pwwn 11:22:33:44:11:22:33:44
device-alias name test1 pwwn 11:22:33:44:44:33:22:11
device-alias name test2 pwwn 11:22:33:44:44:33:44:44
device-alias name dev_fa pwwn 21:00:00:e0:8b:05:8b:7c
device-alias name dev_fb pwwn 21:00:00:e0:8b:03:b5:29
device-alias name fast900a1 pwwn 20:02:00:a0:b8:12:0f:13
device-alias name testalias pwwn 21:00:00:24:ff:22:ae:f6
device-alias name fastt900b1 pwwn 20:03:00:a0:b8:12:0f:13
device-alias name change_name pwwn 21:00:00:e0:8b:04:d2:51
device-alias name furthermore pwwn 55:55:55:55:55:55:55:55
```

Note: The device alias can only be used to identify a single WWPN, unless the enhanced device alias feature is enabled.

Because of the limitations when using fcaliases to name devices, it is better to implement enhanced device aliases.

One example of a use-case for device aliases is the NPIV feature on IBM SVC, Spectrum Virtualize, and FlashSystem clustered storage systems. When this feature is enabled, the storage system is allowed to log into the fabric using two WWPNs on each port.

- The first WWPN is physical and tied to the adapter port.
- The second WWPN is virtual and can float.

Hosts should be zoned to the virtual WWPN. If an SVC node port goes offline because of an issue, or a cluster node is taken down for maintenance, the virtual WWPNs will float to another node in the cluster. If you are using WWPN-based zoning without device aliases, you must update all of your zoning to rezone the hosts to the new virtual WWPNs when this feature is implemented.

Port description

When enhanced device aliases are used, you can simply add the new virtual WWPN to the existing alias. Then after verifying host connectivity, remove the physical WWPN from the alias and leave the zoning unchanged for the hosts. The port description is a plain-text description of the interface that a device is plugged into. This naming can be used to name a port for a specific device, or name a port that is part of a Port-channel. At a minimum, you should add port descriptions for ports that are used in Port-channels. This makes troubleshooting and fabric maintenance easier and reduces the chances of cabling or other errors. While devices might get moved to different ports or different switches, ISLs between switches rarely change after they are connected.

The naming convention should consist of a well-defined schema for each type of object (switch, VSAN, device, port) that is being named. It must be consistent, documented, and user-friendly.

While the exact schema that you create will best fit your needs, a suggested schema is:

- Switch Name: <Fabric><Location>SwitchName>
- VSAN: <SwitchName><VSANName>
- Port (ISL):<SwitchName><PortChannelNumber><RemoteSwitchName+Interface>
- Port (Device) - <SwitchName><DeviceGroup><Device><DevicePort>
- Device Alias - <DeviceGroup><Device><DevicePort>

Using the above schema, the Device Alias for Port 1 on HBA 1 of Node 3 of a 4-Node VMware Cluster that is connected to a switch named FabricA_ProdDC_Switch_1 would be:

```
FabricA_ProdDC_Switch1_ESXCluster4_Node3_P1_1
```

The above device alias contains much information. A robust naming convention makes troubleshooting easier. For instance, when you are trying to resolve an issue such as a connectivity problem, the switch output (such as the device login data or zoning) can be filtered by the meaningful names to quickly determine whether zoning is missing or a port is offline.

Best Practice: Define and implement a meaningful naming convention. This is a disaster prevention option available that does not require software. It provides storage and SAN administrators with the ability to visually determine the site, detail, and attached device. This helps to quickly identify devices when troubleshooting.

N Port Virtualization

Most fibre-channel networks are deployed as a core-edge design with smaller edge switches connected to one or a few core switches. Larger directors have a much higher cost-per-port than smaller edge switches. However, such deployment leads to more complexity in the fabric because:

- As the number of switches increases the number of switch domain IDs also increases. This is further exacerbated by having embedded switches in server chassis
- If these embedded switches are not Cisco or IBM C-Type switches, there are interoperability issues

N Port Virtualization (NPV) reduces the number of fibre-channel domain IDs that are required on a fabric. A switch that is operating in NPV mode does not join a fabric or participate in any fabric services, such as zoning. Instead, it appears as a host to the switch it is connected to.

Figure 5 shows a typical NPV configuration. The three servers are connected to an NPV device, which is then connected to the switch. The servers log into the fabric at the switch, not the NPV device.

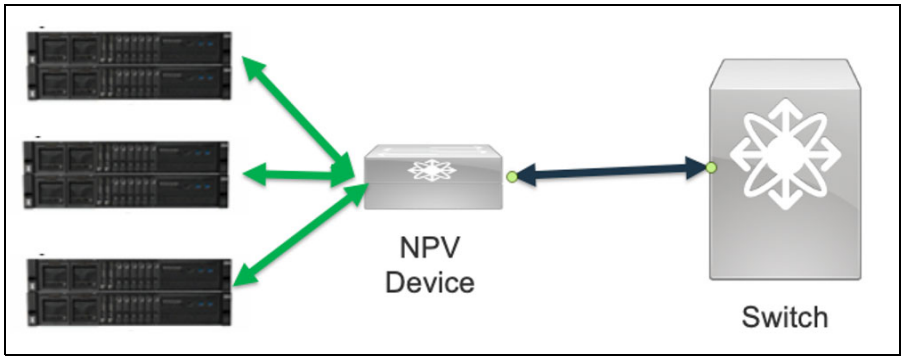


Figure 5 NPV configuration

An NPV device looks much like a hypervisor running virtual machines. This can reduce the complexity of the fabric. For smaller top of rack or switches embedded in server chassis with limited resources, this means that they are not storing the zoning database, which can be large. Smaller switches also do not need to maintain other fabric services, such as a name server or domain controller. NPV mode also removes the interoperability issues since fabric services are not running on the NPV devices.

For a full list of the features that are unavailable in NPV mode, see:

https://www.cisco.com/c/en/us/td/docs/switches/datacenter/mds9000/sw/8_x/config/interfaces/cisco_mds9000_interfaces_config_guide_8x/configuring_npv.html

NPV mode is currently supported on the following Cisco switches:

- Cisco MDS 9132T 32-Gbps 32-Port Fibre Channel Switch
- Cisco MDS 9148T 32-Gbps 48-Port Fibre Channel Switch
- Cisco MDS 9396T 32-Gbps 96-Port Fibre Channel Switch
- Cisco MDS 9148S 16-Gbps Multilayer Fabric Switch
- Cisco MDS 9396S 16-Gbps Multilayer Fabric Switch

NPV is not supported on the director-class switches, such as the MDS 9700 family switches or the MDS 9250i.

N Port Identifier Virtualization (NPIV) is sometimes confused with NPV. NPIV enables a switch to assign multiple FCIDs to a single N-Port. This capability allows multiple applications or hosts to connect to a fabric on the same N-Port. NPIV also allows for access control and port security for individual devices attached to a single N-Port.

Note: Connecting a Cisco or third-party switch in NPV mode to a Cisco director or another switch requires the NPIV feature be enabled on the director or other switch. This feature is enabled by default on NX-OS versions 8.4(2) and later.

F Port Channels

F Port Channels between a switch and NPV device provide the same benefits of performance and fault tolerance that E Port Channels provide between switches. F Port channels can be used between Cisco NPV devices or Cisco UCS Interconnects and fabric switches.

Be aware of the access control list ternary content-addressable memory (ACLTCAM) programming with NPV. Ternary content-addressable memory (TCAM) is a special region of memory on each FC line card in the Cisco MDS family. The TCAM memory on each linecard is divided into sections and used for various fabric services and NX-OS features.

The ACL region programs which devices can talk to which other devices, based on the zoning configuration in the active zoneset. When N ports are configured in a port channel, ACL programming in TCAM is repeated and can become exhaustive. For example, If there are 30 servers and they are zoned with eight targets each, then there will be 30 x 8 = 240 ACL TCAM entries that are programmed on each member of the F Port-Channel. If there are 8 members in the
F Port-Channel, then there will be a total of 1920 ACL TCAM entries programmed for the F Port-Channel. On a large fabric, this situation can result in the ACL TCAM maximum size being exceeded. The following options reduce TCAM usage:

- Distribute the F port channel links across different line cards on the director
- Connect the F port channel links to switches with lower TCAM usage
- Split the F port channel into multiple port channels with fewer links in each port channel. Distribute the devices that are attached to the NPV device across the port channels

- Use single initiator/single target zones or use Smart Zoning.

NPV traffic management

NPV traffic management is responsible for balancing servers across the uplinks between the NPV device and the switch. Traffic management has three modes, which are described below.

Auto

By default, traffic management automatically assigns traffic for each server to one of the links between the NPV device and the switch when the server logs into the fabric. NPV devices frequently include multiple links to a fabric, such as if a fabric has dual-core switches. In these cases, the NPV device assigns the server to the external link that has the fewest number of servers that are assigned to it.

However, this can lead to unequal utilization on links because the individual servers have different link-utilization rates, so one external link might have much higher utilization than another. Also, a server cannot be non-disruptively moved to another link. The server must log out, then log back in on the new link. Because of the risk of unequal utilization in a dual-core fabric design, the NPV device should have sufficient bandwidth to each core switch to carry all the load for all of the NPV devices. This makes the fabric as resilient as possible.

Traffic Map

Traffic Map is the preferred option to load-balance your servers across uplinks. This is a work-around for the problem of unequal link-utilization. This option allows you to manually specify which servers use which uplinks to the switches. If it is enabled and configured, the server only uses the uplinks that are configured for it. Other uplinks are not used, even if the uplink fails. When you use Traffic Map, you should configure redundant uplinks for a resilient fabric.

Use the following command to list the suggested mapping based on measured loads:

```
show npv traffic-map proposed
```

Use the following command to manually map devices to specific external links to better load-balance your servers:

```
npv traffic-map server-interface
```

Disruptive

Disruptive traffic management is the third option. This mode works independently of the auto and traffic-map options. If this feature is enabled, the NPV device forces the server ports to reinitialize, which moves them to new uplinks. The NPV device continues to do this until load balancing is achieved. This feature forces the ports to reinitialize each time a new uplink comes up. This feature should only be used during initial setup to achieve an initial load balancing.

Best Practice: Enable NPV mode on smaller embedded switches and non-IBM c-Type or non-Cisco switches. On large fabrics, consider enabling NPV mode on smaller 1U switches. Configure traffic management to use Traffic Map and ensure that each device can connect to each fabric on at least two links.

Zoning

Zoning plays an important role in restricting device-to-device and server-to-device communications. A smaller number of paths for servers or devices results in improved performance and reliability. This is because the servers and devices have fewer paths to discover and perform recovery or retry on.

Switch port or WWPN zoning

Zones contain a list of switch ports or WWPNs that are allowed to communicate with each other. When switch-port zoning is used, switch ports are zoned together, which is acceptable if there is only a single device that is attached to each of the ports in a zone. However, if a switch port contains multiple devices that are logged in using NPIV, then all of the devices that are attached to that switch port are able to communicate with all of the devices attached to the other switch ports in the zone.

WWPN zoning allows you to specify the WWPNs that can communicate, so it provides more control over which devices can communicate with which other devices, regardless of how many devices are logged into the fabric on a given switch port. Switch-port zoning is not supported in some configurations, such as when you use the IBM Spectrum® Virtualize NPIV feature.

For this reason, we recommend WWPN zoning, which is required when virtualized NPIV connections are used, and allows devices to be plugged into different ports without requiring zoning changes.

Best Practice: Use WWPN zoning rather than switch-port based zoning. WWPN zoning provides more control over which devices can communicate with other devices.

Zoning types

Several zoning types are available, as follows:

- **Group-of-hosts to Device**

  A group of hosts, based on the operating system type (such as Microsoft Windows / VMware, IBM AIX®), is zoned to the devices they need to connect to. This style of zoning provides the advantage of requiring few zones to provide connectivity. This style **is not recommended** because it enables too many connection paths for every server and device and allows servers to have paths to other servers and devices to have paths to other devices.

- **Single-Initiator to Target**

  This is a common zone type, which has all the ports or WWPNs for a server in the zone with the storage device it needs to communicate with. **This is the recommended zoning for small fabrics.**

- **Single-Host port to Device-port**

  This is the preferred zoning type because it provides the greatest isolation and control of server-to-storage paths. The main issue with this zoning style is that it requires several zones, which can be unmanageable with any medium-to-large fabric.

- **Smart zoning**

  Smart zoning is a newer zoning type that combines host type-to-device and host port-to device zoning types (see , “Smart zoning”). **This is the recommended zoning types for medium and large fabrics.**

Best Practice: Use Single Host port to Device zoning for small to medium fabrics. Use Smart Zoning for larger fabrics. The Autozone feature can also be used for single switch fabrics.

Smart zoning

Zoning is the method that is used to control which devices in a fabric are allowed to communicate (pass frames) with each other. It is important to have a zoning methodology that allows only the intended devices to communicate with each other. Some zoning methods allow servers to talk to other servers, or storage devices to talk to other storage devices. Although these devices should ignore this type of traffic, resources are used during probing. In some cases, two devices that are not intended to communicate will try to establish a communication path, which can cause a problem.

Currently, the most common zoning method is the *initiator target zoning method* where a zone contains one initiator (host HBA) and one target (storage port). In today's devices, servers can have multiple HBAs, and storage devices can have multiple ports that are connected to provide additional data paths for performance and redundancy. In this type of configuration, the number of zones that are required to establish this level of connectivity can be massive, which introduces greater potential for errors in host-to-storage connections.

To reduce the number of zones, it is common to see a server HBA and multiple storage ports in a single zone, which allows for storage-to-storage communication. However, most storage devices recognize this and ignore the connection.

A newer zoning method is called *peer zoning*, where a zone can contain one or more HBA ports that are tagged as initiators with several storage ports that are tagged as target ports. When this type of zone is activated, the switch does not allow initiator-to-initiator communication or storage-to-storage communication. This allows a single zone to contain a storage device and the host HBAs that need to communicate with that storage device.

CISCO peer zoning is Smart Zoning, which implements a flexible peer zoning method. When you create a zone and add HBA or storage ports to the zone, tag them if they are an initiator port, target port, or a port that is both initiator and target.

References

References for additional details on smart zoning can be found in the following technical documents:

- Smart Zoning - Cisco:

  https://www.cisco.com/c/en/us/support/docs/storage-networking/zoning/116390-technote-smartzoning-00.html

- Implementing Smart Zoning on IBM c-Type and Cisco Switches

  https://www.ibm.com/support/pages/node/689261

The above documents can also be found with an internet search argument of *Implementing Smart Zoning.*

Maintaining an optimal FC SAN environment

This section discusses how to maintain an optimal FC SAN environment.

Switch firmware levels

Maintaining a resilient FC SAN environment includes the scheduling of regular updates to the switch firmware on the switches in the SAN. Updates should be done at least every 18 months, but not more frequently than every six months. Significant planning goes into each upgrade, so six months is the recommended minimum time between upgrades. The maximum time between upgrades is 18 months. Additionally, you should never be more than one major revision behind the current recommended version of firmware. If you get too far behind, the upgrade often requires interim upgrades to reach the target version, thereby increasing the risks during the upgrade. It is also possible to fall so far behind that you are running an unsupported version of firmware. The exception to the minimum six-month interval is if you encounter a bug that impacts your operations, and you need to upgrade to fix the bug. Before any upgrades, you should back up the startup and running configurations on your switch, both locally on the switch and to an off-switch location.

When you are planning an upgrade, see the Cisco Recommended Releases information:

https://www.cisco.com/c/en/us/td/docs/switches/datacenter/mds9000/sw/b_MDS_NX-OS_Recommended_Releases.html

The planning for your upgrade must include a review of the release notes for the NX-OS version that you plan to upgrade to:

http://www.cisco.com/en/US/products/ps5989/prod_release_notes_list.html

The release notes list the supported non-disruptive upgrade paths and bugs that are fixed (and known bugs that are not fixed) in the new release. The release notes also list potential disruptions that can occur during the upgrade. Upgrades carry a risk of disruption. However, some upgrades might cause certain types of links, such as FCIP tunnels, to drop during an upgrade. If this is expected during an upgrade, the release notes include this information.

Best Practice: Consult the Cisco Recommended Releases guide to determine which release you should upgrade to. Review the release notes for the target release to determine the supported upgrade path and any required interim NX-OS versions.

Port-Monitor

Maintaining a resilient fabric includes implementing an effective port monitoring configuration. Port-Monitor enables each of the switches in the fabric to continuously monitor the switch ports for congestion or link errors. If a switch detects a problem, it adds a timestamped entry to the logging log. Optionally, it can use *Remote Monitoring* to alert administrators of the problem. The timestamped entries are critical to pinpoint the source of a problem and correlate any issues with other devices if they are impacted.

For example, if monitoring is configured to detect CRC errors on links, then a quick scan of the logging log regularly would show errors. The entries are timestamped, so you can see how frequently they occur and the effectiveness of remediation steps. On a Cisco MDS series switch, the monitoring is accomplished through the Port-Monitor function.

Port-Monitor is configured by using policies. Policies must be defined per switch, so different switches can have different policies that are defined. However, it is strongly recommended that all switches have the same policies. The exception is fabrics that are designed with core switches that only have ISLs connected to them. These switches do not need a port-monitor policy that uses logical-type Edge ports because they do not include edge devices. However, all the core switches in this type of fabric should have the same policies that are applied to them.

Policies can be defined for three logical types of ports:

- **All**: This includes all ports on the switch.
- **Edge**: This includes only ports where devices are attached.
- **Core**: This includes only ISLs to other switches.

Only one Port-Monitor policy that includes a given port type can be active on a switch. If a port-monitor policy is active for port-type All, other policies cannot be active.

If a policy of port-type Edge is active, an administrator can also activate a policy for port type Core, because Core ports are not included in the Edge policy. However, a policy for port type All cannot be activated. Because of this restriction, it is recommended that you configure and activate policies for port-types Edge and Core. Edge ports and Core ports have different thresholds for detection, and include some features (such as Congestion Isolation) that work only on Core ports. Some actions, such as shutting down a port, should not be enabled on Core ports because this would affect all the devices that are traversing that ISL.

Figure 6 lists the default active Port-Monitor policy for Cisco MDS series switches. This policy should not be used because it includes only Edge ports and does not monitor Core ports. It also does not monitor enough of the counters that should be monitored.

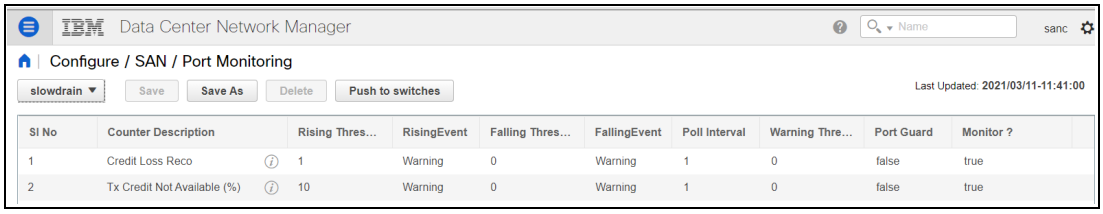


Figure 6 Default slow-drain policy

Figure 7 lists the default available policies in IBM Data Center Network Manager (DCNM). It includes policies for Access (Edge) ports, Trunk (Core) ports, and All ports.

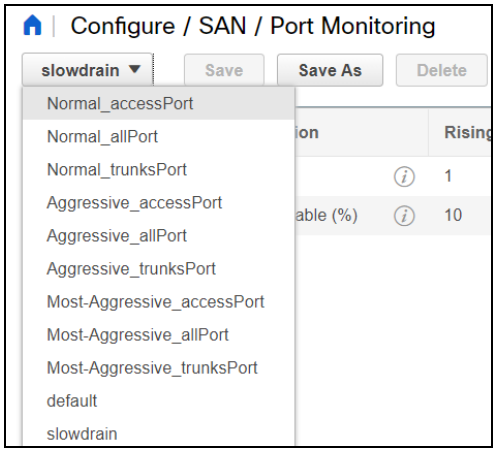


Figure 7 Default policies

There are three policies (Normal, Aggressive, and Most-Aggressive) for each port type. The differences in the policies are the counter-trigger levels.

Recommended port monitoring policies

You should deactivate the default slow-drain policy that is active on the switches and activate the Normal_accessPort and Normal_trunksPort policies that are listed in Figure 7. You can also copy these policies to custom policies and implement those custom policies. It is strongly recommended that you copy the existing policies to custom policies and implement separate policies for Access (Edge) ports and Trunk (Core) ports. You should not implement a single policy for all ports. When you implement custom policies, you can include the counters from the default Slow-Drain policy in the policies that you implement.

These policies will likely need to be tuned over time.

- If the thresholds are set too aggressively, you might initially get too many alerts.
- If thresholds are not aggressive enough, you might miss an event when it happens.

For details on implementing port-monitor policies, see the following whitepaper:

https://www.cisco.com/c/dam/en/us/products/collateral/storage-networking/mds-9700-series-multilayer-directors/white-paper-c11-736963.pdf

Best Practice: Deactivate the default slow-drain policy and copy the existing Core and Edge policies to custom policies. Add the counters in the slow-drain policy to those new policies and deploy those policies.

Slow-drain device detection

IBM c-Type switches include several features that are used to detect slow-drain devices, such as:

- TxWait: Realtime counter for 0 Tx buffer credits. It indicates the number of times credits have been at 0 for at least 2.5 microseconds when there is a frame waiting to be sent. This counter increments in 2.5 microsecond intervals.
- Slowport-Monitor: Allows for monitoring of ports that are in a continuous 0 Tx buffer credit state for a user-specified number of milliseconds.
- LR Rcvd B2B: Indicates that the adjacent side of the link sent a Link Reset (LR) to the switch, but the switch received frames that it has not been able to clear out. 100 ms after receiving the LR, the port is failed with 'Link Reset failed due to nonempty receive queue' (or LR Rcvd B2B). This normally indicates that another port, which is meant for receiving frames, is congested.

  LR Rcvd B2B is an important statistic. If a device attached to the switch is out of transmit (Tx buffers) for too long, it resets the link to recover credits. This indicates that the switch is unable to forward frames through the fabric. The device resetting the link is not the source of the slow drain. There is a device elsewhere in the fabric that is the source of the slow drain.

Congestion isolation

Congestion isolation is a feature on IBM c-Type switches that can be used to isolate slow devices to low-priority virtual links to prevent them from affecting traffic to other devices. This feature is available only on ISLs. This means that traffic moving between two ports on the same switch cannot be isolated because of congestion on one of the ports.

For details on how this process works, see:

https://www.cisco.com/c/en/us/td/docs/switches/datacenter/mds9000/sw/8_x/config/interfaces/cisco_mds9000_interfaces_config_guide_8x/congestion_avoidance_isolation.html

When congestion is detected, the slow device can be isolated manually or using Port-Monitor and a PortGuard action. To maintain a resilient fabric, you should configure your port-monitoring policy to include port-guard actions to isolate slow devices. Congestion Isolation can be triggered by port-monitoring on these counters:

- credit-loss-reco
- tx-credit-not-available
- tx-slowport-oper-delay
- txwait

Remote Monitoring

IBM c-Type Remote Monitoring (RMON) is an SNMP-based specification that allows various network agents and console systems to exchange network monitoring data. NX-OS supports RMON alarms, events, and logs to monitor NX-OS devices. You can use the RMON alarms and events to monitor IBM c-Type and Cisco MDS 9000 Family switches that run NX-OS.

When you configure an RMON alarm, it monitors a specific SNMP MIB object for a specified interval. When the MIB object value exceeds the value that is specified in the alarm (the rising threshold), the alarm condition is set and a single event is triggered. It does not matter for how long the condition exists. When the MIB object value falls below the alarm value (falling threshold), the condition is cleared.

RMON works with Port-Monitoring (PMON) to detect problems and forward alerts to the SAN administrators. If RMON is enabled and configured, PMON alerts are automatically forwarded by RMON to administrators.

RMON notification options

RMON supports the following options for notification when alarms are generated for RisingThreshold or FallingThreshold events:

1. SNMP notification: This sends an SNMP notification.
2. Log an entry to the local RMON log on the switch.
3. Both 1 and 2. In this configuration, an SNMP notification is sent and the event is logged to the local switch.

It is recommended that you enable RMON in conjunction with PMON. This enables the switches in your fabrics to notify you if PMON detects an error condition. For more details on implementation, see:

https://www.cisco.com/c/dam/en/us/products/collateral/storage-networking/mds-9700-series-multilayer-directors/white-paper-c11-736963.pdf

Best Practice: Enable remote and port monitoring for improved link monitoring and notifications.

Credit recovery on IBM c-Type and CISCO fabrics

IBM c-Type and Cisco fabrics include several mechanisms for recovering from congestion due to credit loss.

Virtual Output Queues

Virtual Output Queues (VOQs) are an inherent part of MDS 9000 switches. These queues can be used to set priority on fibre-channel traffic. They are also used to prevent head-of-line blocking on a receiving port that can occur when a destination port is congested.

Figure 8 illustrates how VOQs can prevent blocking. Port 1 is the ingress port. It has frames that need to go to egress ports 4,5 and 6 but port 4 is congested. With multiple queues, the frames from ports 5 and 6 can move through the switch and the only frames that are held up on port 1 are the frames to port 4. It is important to note that VOQs are intrinsic to MDS 9000 family switches and this is not a feature that can be enabled or disabled. Port 1 is the ingress port.

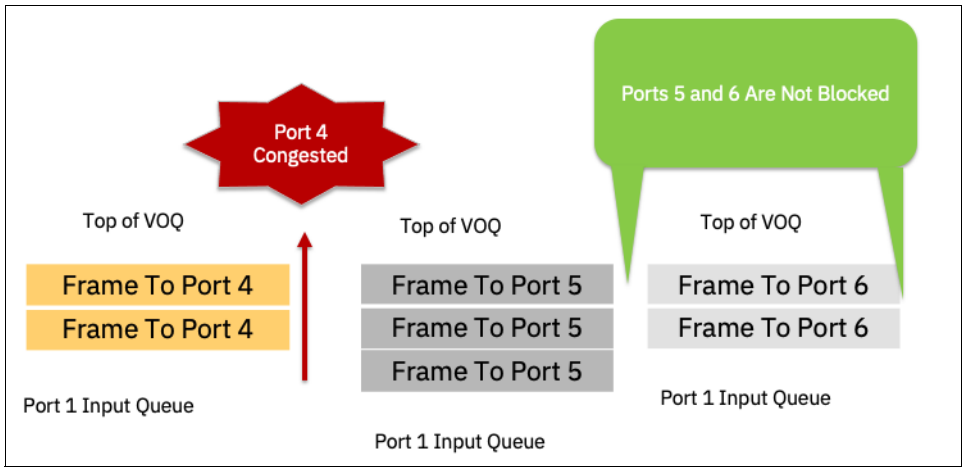


Figure 8 Using VOQ to prevent blocking

SNMP traps

We have already looked at “Port-Monitor” on page 21, which monitors multiple counters at a very low granularity. An SNMP trap is generated if any of the counters in the active port-monitor policy exceeds the configured thresholds over the configured duration. An SNMP trap can be forwarded to a remote system if Remote Monitoring is enabled.

Congestion-drop timeout

A fabric with congestion cannot deliver frames to destination ports in a timely fashion. In a congestion situation, a frame spends much more time than usual switching latency traversing the fabric. However, frames do not remain in a switch forever. IBM c-Type and Cisco switches drop frames that are not delivered to their egress ports within a certain time. This value is called *congestion-drop timeout*. The congestion-drop timeout is enabled and the value is set to 500 ms by default. Changing the congestion-drop timeout to a lower value causes frames that are stuck in a switch to be dropped more quickly. This action frees up the buffers faster in the presence of a slow-drain device. It is recommended that you retain the default value for core ports (ISLs) and set the edge port value to a value not less than 200 ms.

No-credit-drop timeout

No-credit-drop timeout is a reactive mechanism that is available on Cisco MDS 9000 Family switches to automatically recover from slow drain. If Tx B2B credits are continuously unavailable on a port for a duration longer than the configured no-credit-drop timeout value, all frames that consume the egress buffers of the port are dropped immediately. Also, all frames that are queued at ingress ports that are destined for the port are dropped immediately. While the port remains at zero Tx B2B credits, new frames received by other

ports on the switch to be transmitted out of this port are dropped. The default frame-timeout value is 500 ms. No-credit-drop timeout is disabled by default. If you choose to enable it, it is recommended that you leave the core ports at the default value and set the edge ports to a lower value. The recommended timeout value for edge ports is 300 ms.

Credit loss recovery

There are two mechanisms that are defined by the FC specifications to deal with credit loss.

- The first mechanism is Buffer to *Buffer Credit Recovery (BB Credit Recovery)*. This mechanism functions by negotiating a value in the FLOGI/ACC(FLOGI) or ELP/ACC(ELP). This value equates to a count of frames and credits that are sent by each side. Each side keeps track of both the frames and credits it transmitted and received.

  Each time the count of frames hits the agreed-to value, a special primitive (called a BB_SC_S) is sent. When the other side of the link receives the BB_SC_S, it looks at how many frames it received. If it received less than the agreed-to value, *extra* R_RDYs are transmitted back to the sender of the BB_SC_S to make up for the frames that were presumably lost.

  Also, each time the count of R_RDYs-sent reaches the agreed-to value, a special primitive (called a BB_SC_R) is sent. When the other side of the link receives the BB_SC_R, it looks at how many R_RDYs it has received. If it received less than the agreed-to value, it increments the received R_RDY-count to make up for the R_RDYs that were lost.

  Consequently, BB Credit Recovery can recover lost BB credits without much impact on performance if all the BB credits are not lost before the agreed-to count of frames and R_RDYs are sent.

- The second mechanism is *Credit Loss Recovery*. This mechanism detects when there are 0 Tx BB credits for a continuous period of 1 second (F ports) or 1.5 seconds (E ports). When this occurs, the detecting side sends a Link Reset (LR) FC Primitive to the adjacent side of the link. The adjacent side should send back a Link Reset Response (LRR) back. If this completes successfully, then both sides are at their full complement of BB credits and normal traffic returns.

  It is important to note that the LR/LRR, despite its name, does not really reset the link itself. It resets the BB credits on the link. The link stays up and does not bounce or flap.

Port Flap or Error-Disable

One option in the Port-Monitor function is to flap ports if any of the monitored counters exceed the configured thresholds over a specified duration. The port-flapping should restore the link to normal operating condition. However, if a device or an HBA malfunctioned permanently or the port is frequently flapping, it is better to shut the port and leave it down. This can be done by error-disabling the switch port. Error-disabled ports can then be brought back online after the error conditions are resolved. It is not recommended that you error-disable ISLs for slow-drain or congestion indications since they are almost always caused by end devices utilizing the ISL. However, it is acceptable to error-disable ISLs when there are physical link issues, link input errors, invalid transmission words, invalid CRCs, sync loss, signal loss, link loss, and similar errors.

Performance monitoring

Monitoring the performance of your fabrics is a critical part of maintaining a resilient SAN. Performance monitoring can alert you to problems occurring on your fabric and provide critical data that is needed for resolving problems. It can also identify potential sources of congestion, such as overworked storage ports. You can address these issues before they impact your production.

Datacenter Network Manager

Datacenter Network Manager (DCNM) is a software suite that is used to manage your IBM c-Type or Cisco switches and fabrics. While it is used to configure fabrics and implement features such as zoning, it also contains a performance-monitoring component.

SAN Analytics and Telemetry Streaming

SAN Analytics is a licensed feature that is offered on the 32 Gbps IBM c-Type and Cisco products. It is included in the architecture of the products and therefore does not require additional components in the data center. It is possible to create flows to monitor specific initiator or target pairs. This can be especially useful when troubleshooting congestion, where the congested port is an NPV device or host with multiple virtual hosts that are logged into the same port.

Telemetry Streaming is a component that is related to SAN Analytics that can stream the metrics that are generated by SAN Analytics to an external receiver. The receiver then provides long-term history of the metrics for performance and trend analysis. A receiver can collect metrics from several switches at the same time, and can correlate the information that is collected from initiators and targets.

For more information on both features, see:

https://www.cisco.com/c/en/us/products/collateral/storage-networking/mds-9700-series-multilayer-directors/solution-overview-c22-740197.html

IBM Storage Insights

IBM Storage Insights is a cloud-based software tool that can collect performance data and metadata from IBM storage systems. The licensed version of the tool can also collect data from non-IBM storage. Storage Insights has extensive alerting and reporting capabilities. Storage Insights also allows you to open tickets against monitored IBM storage directly from Storage Insights if you need technical support, and it enables IBM Storage Support to view performance data and collect support log data. It is recommended that you register for Storage Insights and use it as part of your monitoring solution.

For more information about Storage Insights, see:

https://www.ibm.com/support/knowledgecenter/SSQRB8/com.ibm.spectrum.si.doc/f_saas_prd_overview.html

Storage Insights can monitor your storage and send you alerts for performance-related problems and issues with specific devices. Some common alerts are response times for volumes or the status of a storage system.

Figure 9 shows the Storage Insights Dashboard. In this dashboard, the monitored storage systems with problems are moved to the top of the dashboard.

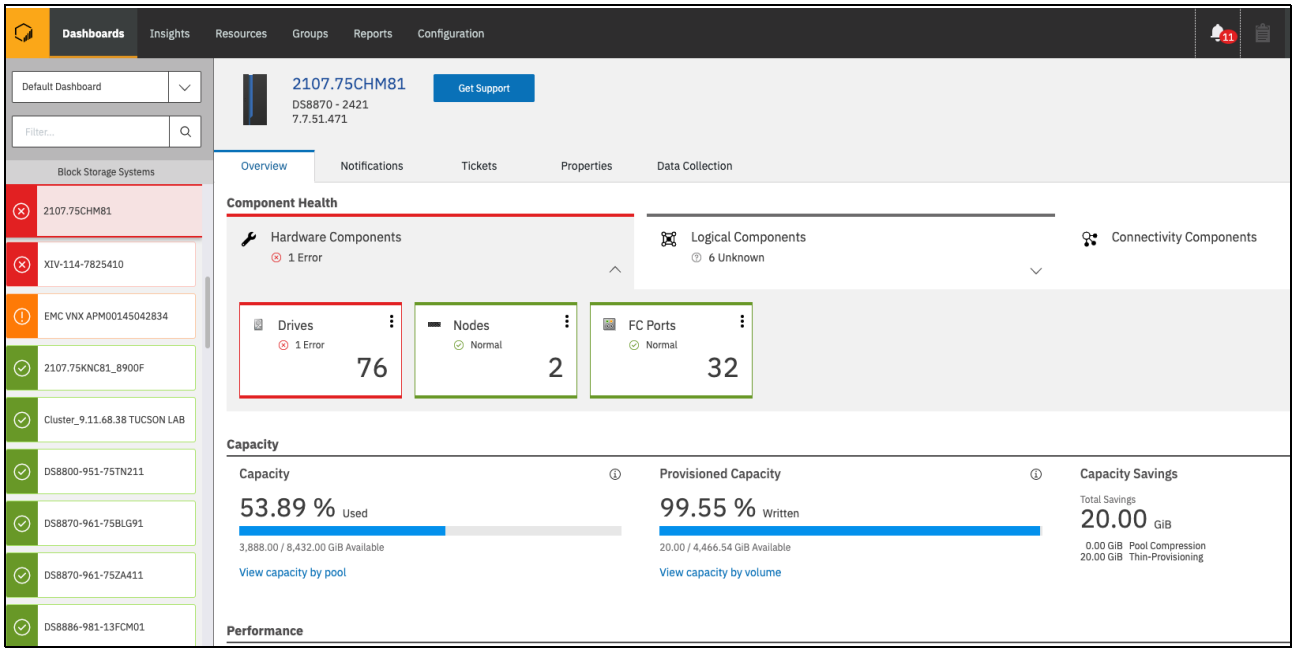


Figure 9 Storage Insights Dashboard

The dashboard provides information on which component of the storage system is a problem. You can use Storage Insights to start troubleshooting the problem and, if necessary, open a ticket against the storage system. You can also configure alerts for that storage system so that you are notified of future problems.

Storage Insights also supports the monitoring of Cisco and Brocade fabrics. Storage Insights and Storage Insights Pro can collect metadata and performance data from your switches similar to storage systems. You can configure alerts for your switches, fabrics, and switch ports to be alerted on error counters.

Storage Insights is not intended to replace DCNM as a management tool and does not have the capabilities that are required to configure a fabric, such as switches and zoning. However, it should be a part of your monitoring strategy. Figure 10 shows a preview example of the list of physical switches in Storage Insights.

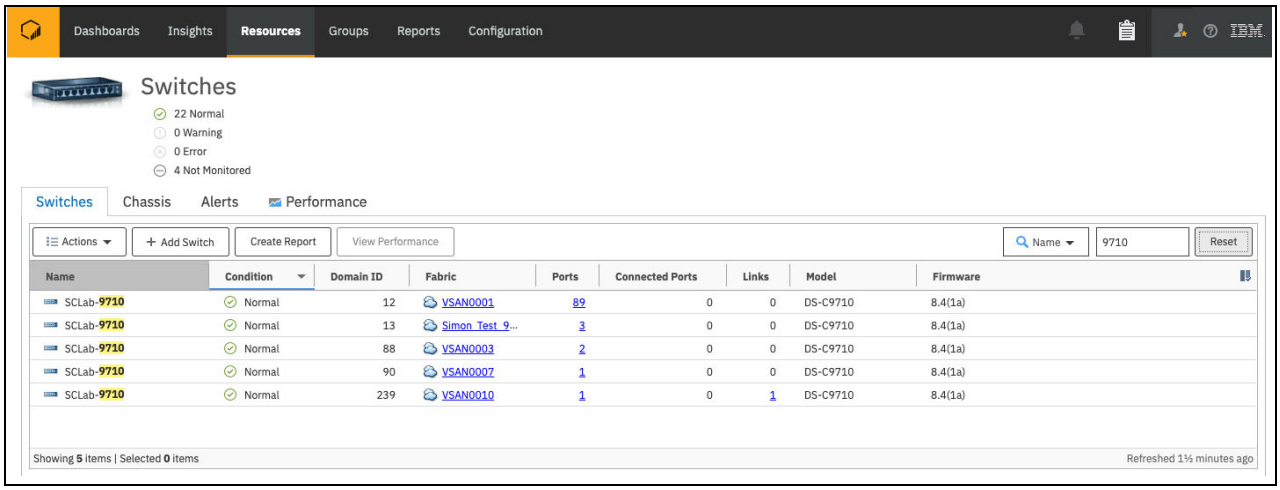


Figure 10 Storage Insights physical switches view

Figure 11 shows an example of drilling down into a switch in Storage Insights.

Figure 11 Storage Insights - detailed view

Note: Figure 10 and Figure 11 show a preview of an expected feature. The screen captures that show fabric support might not be the same as in the final product.

Best Practice: For switches maintained by IBM, implement Storage Insights.

Implementing the free version of Storage Insights enhances the IBM support teams capabilities to provide quick problem resolutions by having quick access to information on switches.

The pro version of Storage Insights provides additional access to performance history and the ability to set up real time alerting.

Performance monitoring tools

This section provides insight into some standard and advanced performance monitoring features and tools, which are used with c-type SAN switches and directors that are offered by IBM and can be leveraged to aid in SAN Fabric resiliency. The performance features and tools discussed in this section are a subset of the available tools.

To view performance information in your SAN environment, IBM c-type switches leverage Data Center Fabric Manager (DCNM) as a standard-base tool to achieve performance monitoring. In tandem, Device Manager (DM) can provide several mechanisms that allow you to monitor and view real-time and lightweight, high-level historical data for IBM c-type Family performance and troubleshooting. Data can be graphed over time to give a real-time insight into the performance of the port. Data includes, but is not limited to:

- Real-time SAN Inter-Switch Links (ISLs) statistics
- SAN modules, ports, and a host of additional SAN elements
- Entire SAN Fabric Health

- Ingress and egress fibre channel traffic errors
- Class-2 traffic showing buffer-to-buffer and end-to-end credit flow control statistics
- Checking for Overutilization
- Threshold monitoring
- Rx and Tx Utilization percentages
- Link Failures, InvalidCrcs, InvalidTxWaitCounts, Sync Losses
- FICON data fabrics

Real-time performance statistics allow administrators to configure custom polling-interval settings for statistical data collection that can help you to troubleshoot IBM® c-type SAN fabric issues. The results can be displayed in the DM Java UI.

DM is used for monitoring and configuring ports on the IBM c-type Family switches. As with real-time performance, when you gather DM statistics you can configure selective polling intervals. This allows you to monitor performance of your SAN environment and troubleshoot any potential problems that exceed specified thresholds.

the polling interval can be set at one hour, thirty minutes, or as low as ten seconds. The results that are displayed are based on several menu drop-down options as follows:

- Absolute value or Value per second
- Minimum or maximum value per second

DM provides the following two types of performance views:

- The Device view tab, which can be used to configure the monitor option per port
- The Summary tab

To configure the monitor option per port, login into DM as shown in Figure 12.

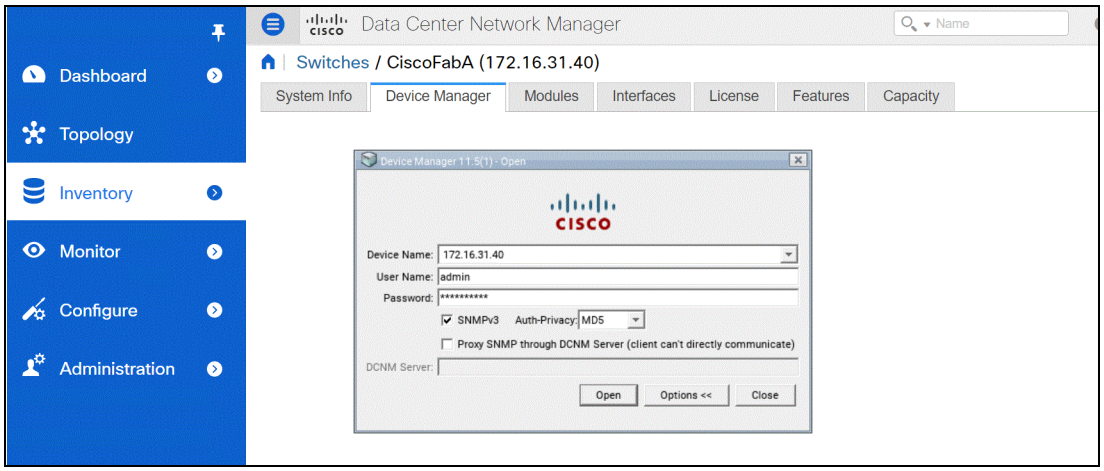


Figure 12 DCNM Device Manager Tab

The per-port monitoring option provides a large amount of statistics. In this step, we select the Device tab view, and right-click on fc1/1, and select **MONITOR** to view the real-time monitor dialog box as shown in Figure 13.

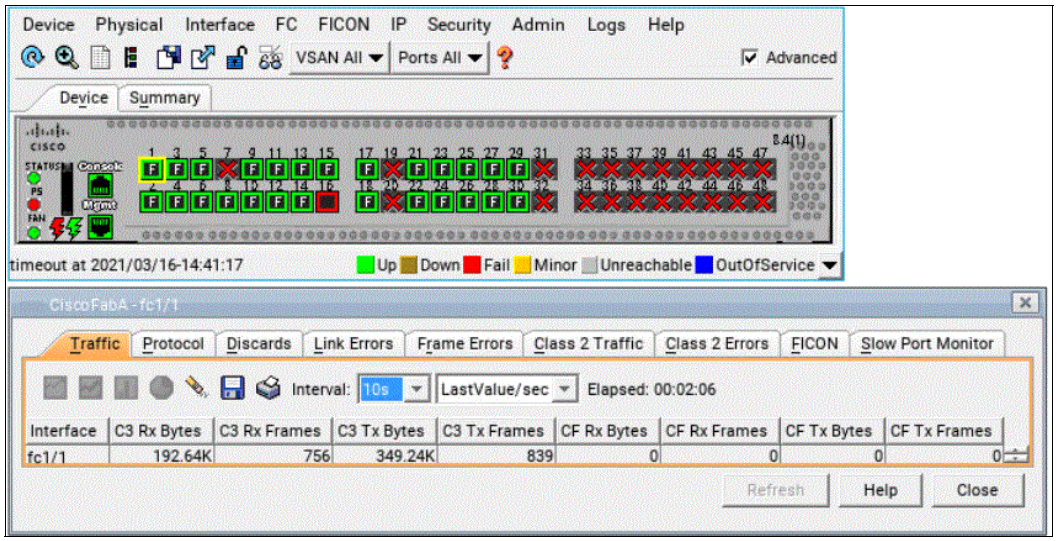


Figure 13 Traffic Monitor View

The Summary view displays active connected devices, port speed and the option to configure parameters, as shown in Figure 14.

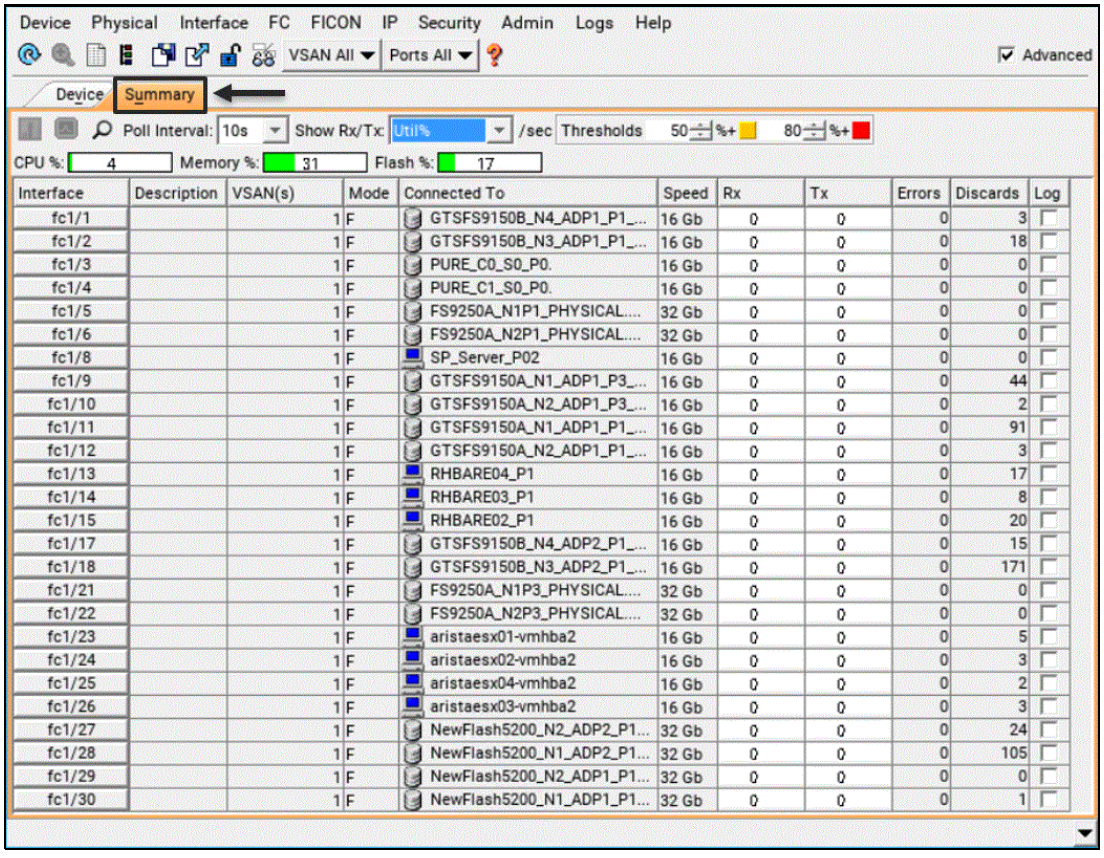


Figure 14 Summary View Tab

When deciding how you want data to be interpreted, make sure that you set the required polling intervals, Rx/Tx and Thresholds settings, as shown in Figure 15.

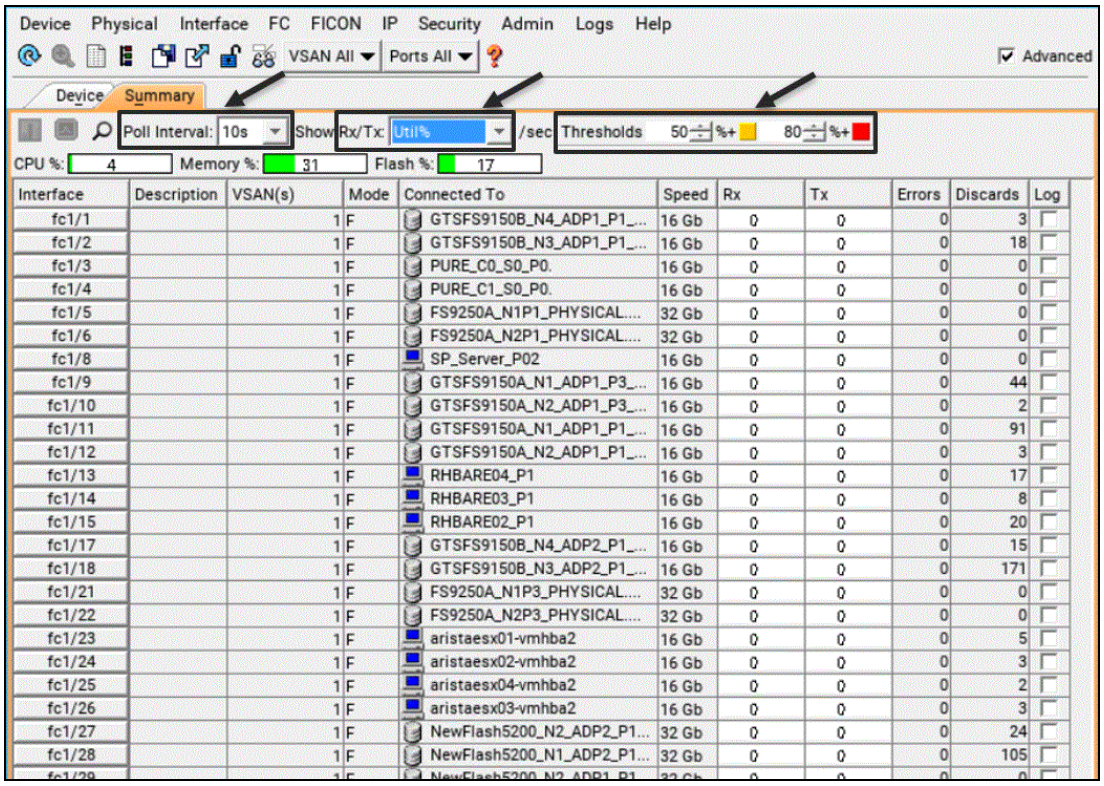


Figure 15 Summary View Configured Settings

The **Poll Interval** options can be selected from the drop-down list, as shown Figure 16.

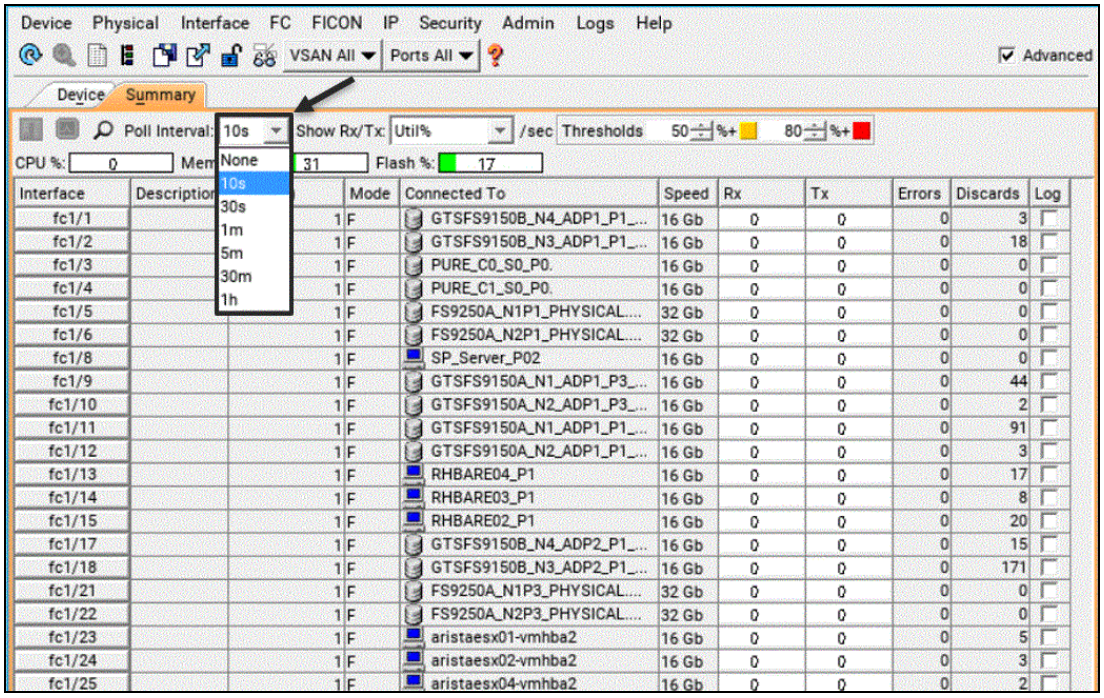


Figure 16 Poll Interval

When using Device Manager to set recommended error thresholds, select **Threshold Manager,** as shown in Figure 17.

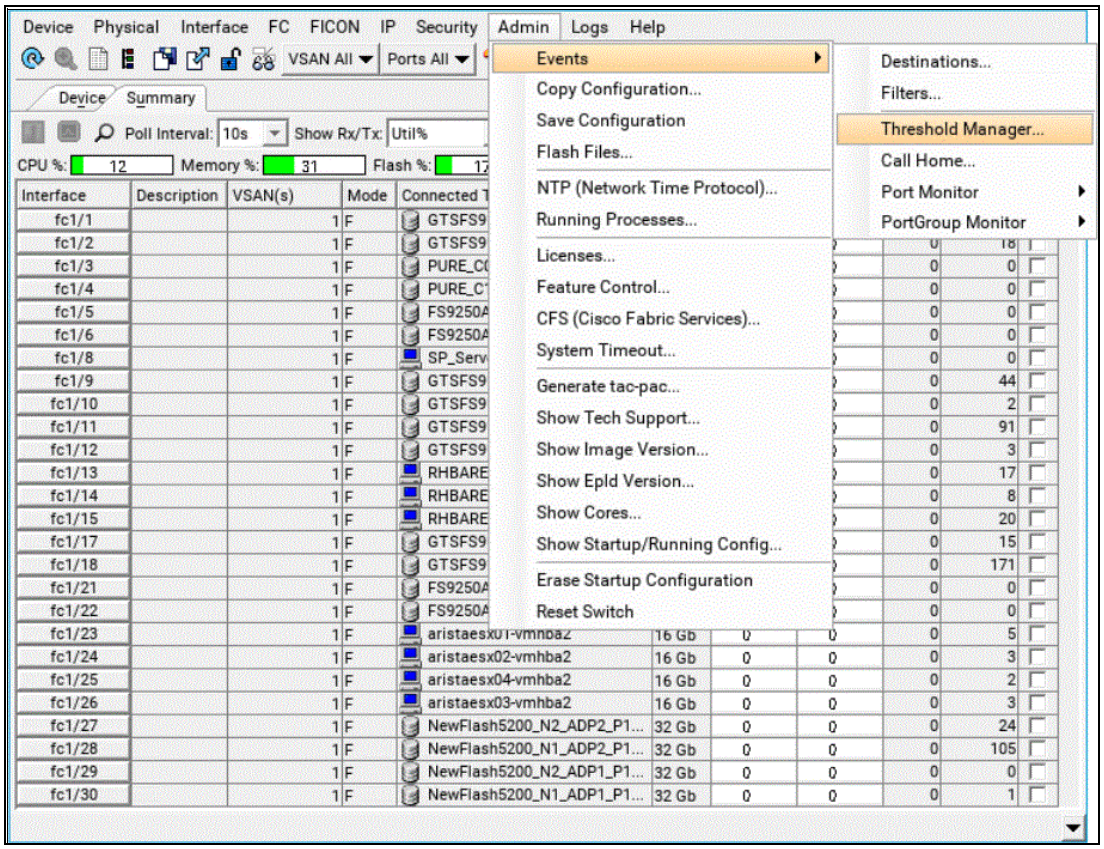


Figure 17 Threshold Manager selection

Threshold Monitor can trigger an SNMP alert and log messages when a selected statistic reaches its configured threshold value.

Best Practice: Configure Device Manager thresholds on your IBM c-type Family switches. Doing so allows you to monitor performance of your SAN environment and troubleshoot potential problems that exceed specified thresholds.

The following values are considered industry best practices:

- Link Failures: Value =1 Sample = 60
- Sync Losses: Value =1 Sample = 60
- InvalidTxWords: Value =1 Sample = 60
- InvalidCrcs: Value =1 Sample = 60

Additional variables and thresholds can be selected and applied to a single port, multiple ports, or all ports, as shown in Figure 18.

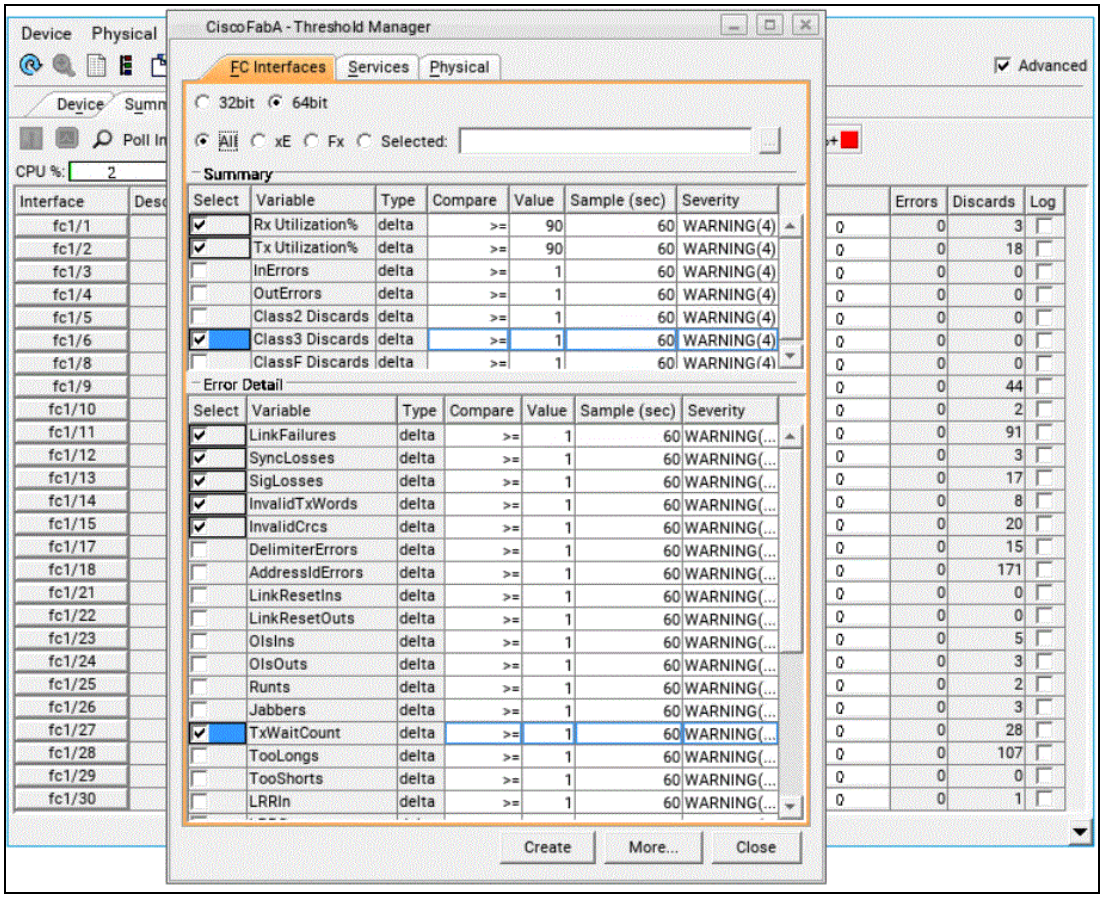


Figure 18 Threshold Manager

Cisco DCNM or other third-party switch-centric tools will only collect data from the switches. Storage Insights and Storage Insights Pro currently can monitor storage and will soon be able to monitor both switches and storage. You can also open tickets against Storage Insights and have the option to allow IBM Storage Support to remotely collect support data.

If legal or other corporate or regulatory requirements prevent you from using a cloud-based service such as Storage Insights, consider using IBM Spectrum Control, an on-premises offering from IBM.

Best Practice: Implement Storage Insights or Storage Insights Pro as a performance-monitoring tool to capture performance data from your storage and switches.

Data Center Network Manager

IBM c-type family provides the following advanced licensed features that can be used for analytics and telemetry streaming to help you sustain resiliency in your environment:

- Data Center Network Manager (DCNM) Advanced
- SAN Insights analytics and telemetry data streaming

DCNM Advanced

DCNM is a management tool used for provisioning, monitoring, and troubleshooting IBM c-type Family SAN environments. It provides a command- and control-style structured regime that provides you with complete visibility into your entire c-type Family Fabric infrastructures.

This is a centralized high-level web-based view, which includes a complete feature set that meets administrative requirements in data centers by streamlining c-type management, provisioning, monitoring, and troubleshooting SAN devices.

Important: IBM c-type DCNM Advanced is the recommended WebUI for the SAN Insights feature.

Figure 19 shows the DCNM Advanced login screen.

Figure 19 DCNM Login screen

After you log into DCNM, the dashboard summary is displayed. The summary provides storage administrators with a 24-hour snapshot of their SAN fabric and the ability to focus on key health and performance metrics on your IBM c-type SAN fabric.

Various default dashlets can be customized to provide a display of your SAN environment. These dashlets include Inventory - Switches, Inventory - Modules, Top CPU, Top ISLs, Link Traffic, and Alerts, as shown in Figure 20.

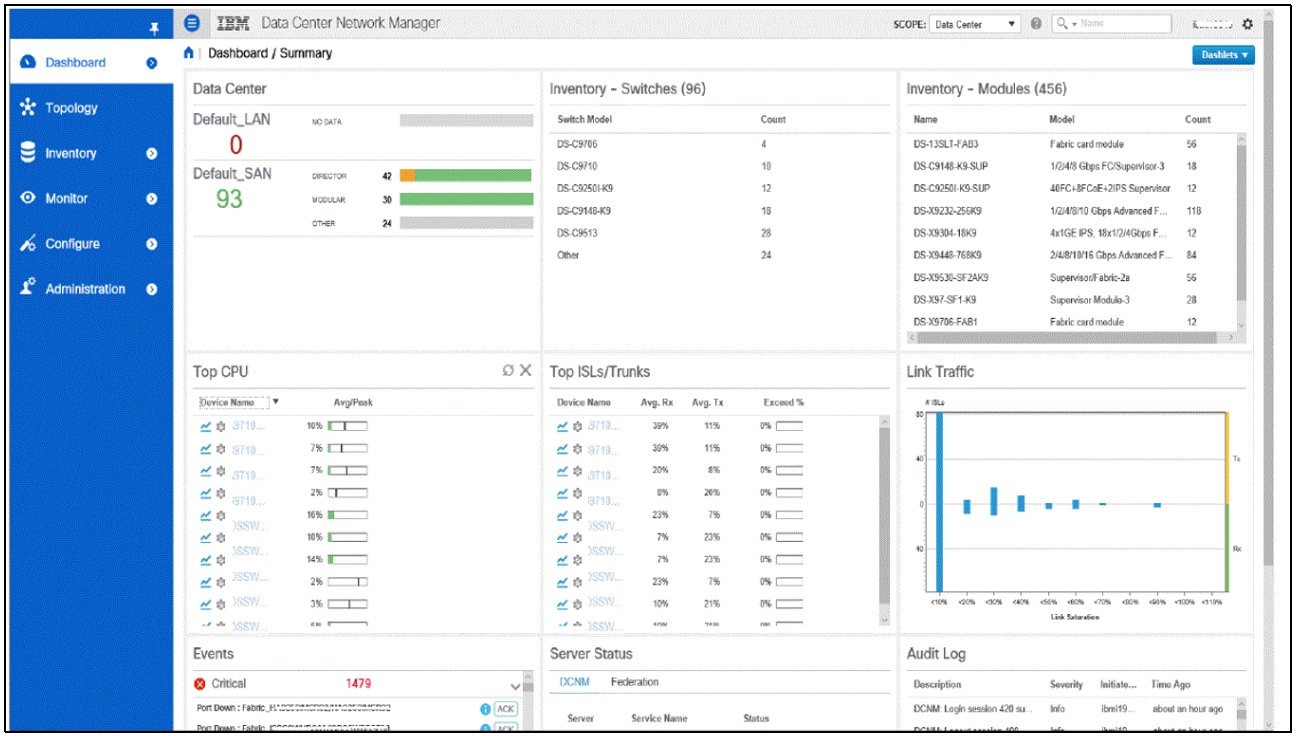


Figure 20 DCNM Summary Dashboard

Various scopes are available in the DCNM web interface. In this example, the Default_SAN scope is shown. The Topology item is selected on the left-hand side, which displays the fabric topology view, as shown in Figure 21.

Figure 21 Topology View

DCNM SAN Management Configuration Guide11.5(1) is available at:

https://www.cisco.com/c/en/us/td/docs/dcn/dcnm/1151/configuration/san/cisco-dcnm-san-configuration-guide-1151/overview.html

SAN Insights analytics and telemetry data streaming

During the last decade, the Storage industry continued to experience monumental changes by adopting all-flash arrays, non-volatile memory express (NVMe), and NVMe over fabrics (NVMeOF) as emerging technologies. These technologies provide unprecedented access to NVMe flash storage, servers, and the applications that run on them.

These high-performance technologies are key motivators that drive storage trends. Emerging solutions are handling millions of Input/Output Operations Per Second (IOPS) and providing lower microsecond response times.

SAN Insights provides deep visibility and understanding of how the components interact within your enterprise storage infrastructures.

The SAN Insights dashboard provides a high-level overview with scope metrics, fabric, and switch views, as shown in Figure 22.

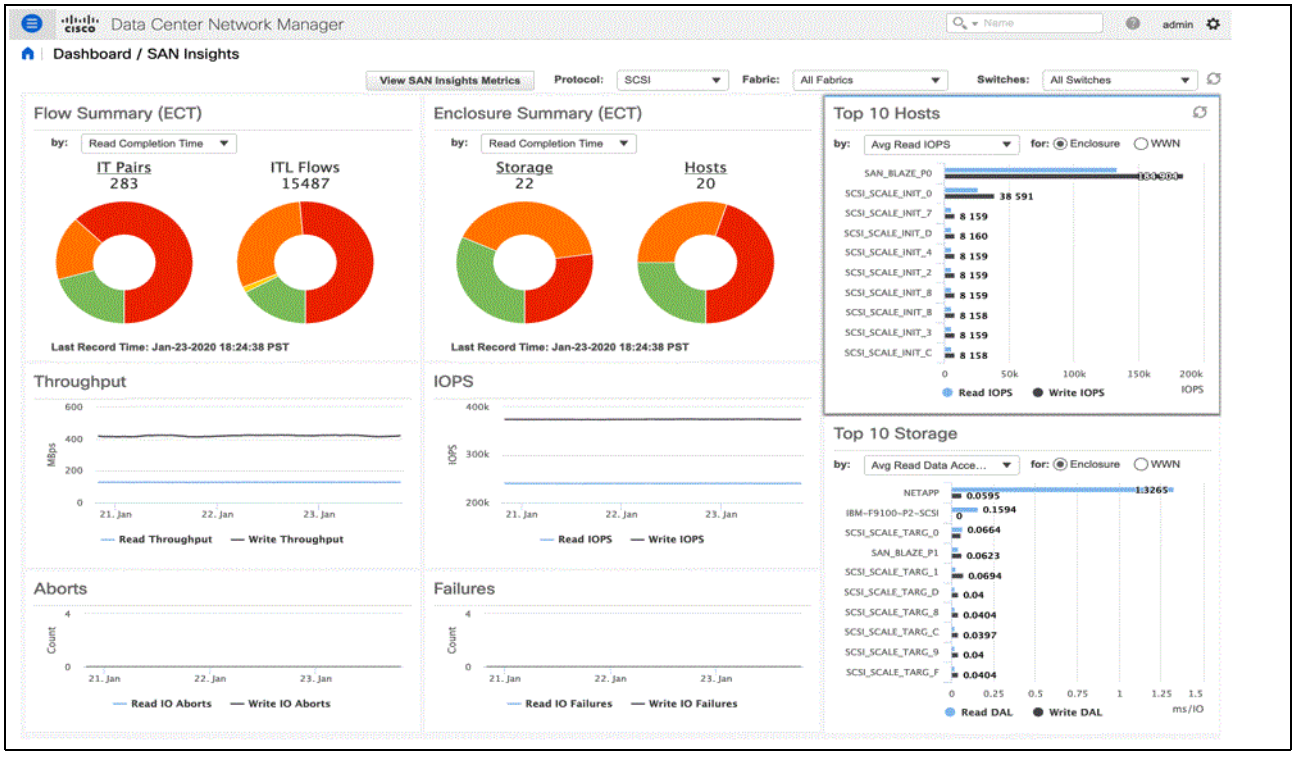


Figure 22 SAN Insights Dashboard

In general, SAN Insights is easy to deploy because the analytics are integrated into IBM c-Type switches.

This integration makes deployment into your new or existing SAN simple and it can scale natively, based on the size of your SAN. SAN Analytics grow with your fabric. Therefore, limitations to geographical scaling do not exist. The end-to-end visibility and capabilities can adhere to hybrid environments, because SAN Insights is agnostic to both compute- and storage-architectural design. SAN Insights inspects I/O flows to provide a unified view of an environment regardless of the architecture or manufacture of disk or flash arrays, servers, and operating systems.

You can monitor host, storage, and IT pairs and drill-down on flow and interface metrics, as shown in Figure 23. You can view health scores for all connected devices and enclosures.

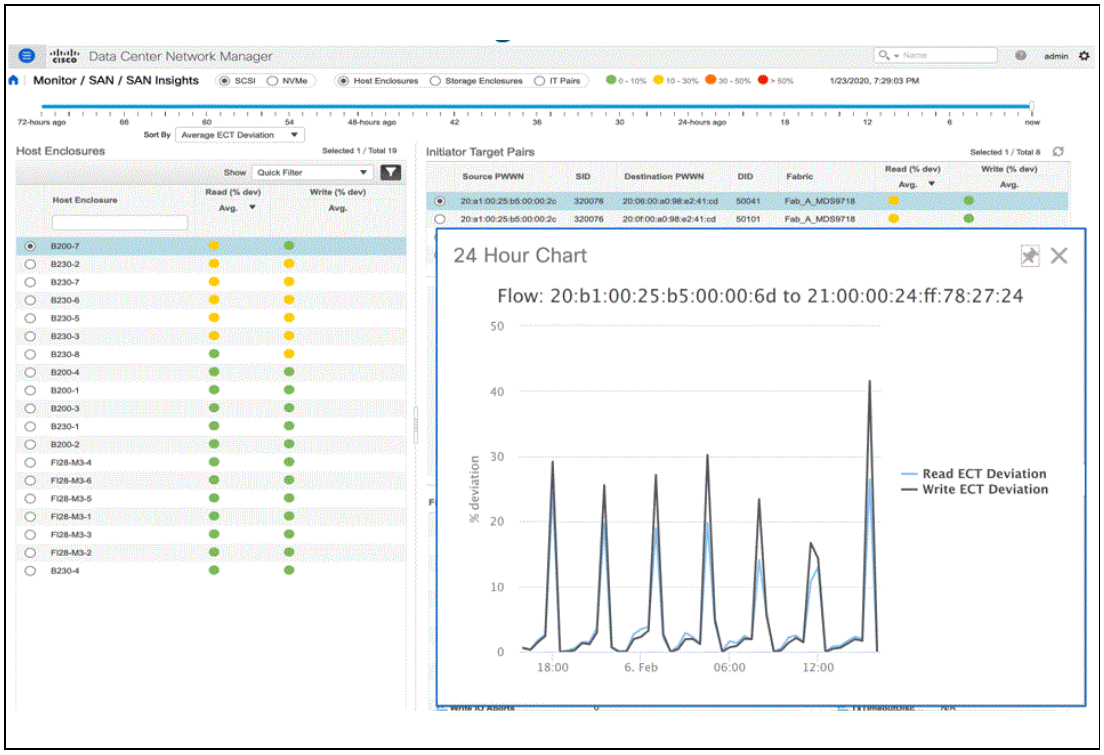


Figure 23 Average ECT Devices

SAN Insights is a licensed feature and analytics engine in DCNM Advanced that provides insights into your SAN fabric by allowing you to monitor, analyze, identify, and troubleshoot performance issues. SAN Telemetry Streaming is used to stream the data of interest to DCNM SAN Insights for analysis and is displayed in the DCNM WebUI.

To obtain the maximum benefit of the c-type SAN Insights features, it is recommended that you consider the following requirements:

- Deploy SAN Insights with DCNM SAN Advanced Edition as they both need to reside on the same server.
- SAN Insights Analytics feature is supported on 32 Gbps and faster line cards with a minimum level of Cisco MDS NX-OS Release 8.3(1) and later.
- SAN Insights is a term-based license that is valid for a minimum of three years or 5-year maximum subscription if you prefer a longer term.
- Cisco SAN Insights Hardware can be deployed on a Physical Clustered server, dedicated VMware management chassis or a Cisco DCNM Hardware Appliance to ensure resiliency.

Figure 24 shows the IBM c-type 32G portfolio with hardware-integrated telemetry capabilities that provide an end-to-end metrics view.

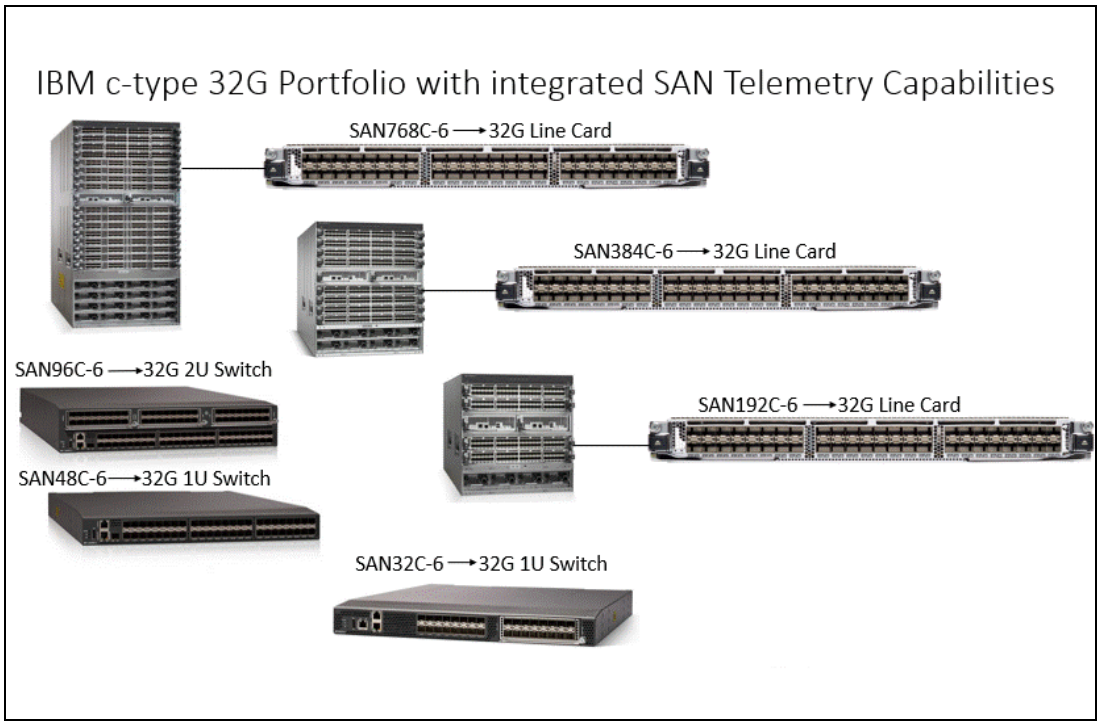


Figure 24 c-type 32G portfolio

SAN Insights can provide real-time visibility into SCSI and NVMe fabrics, as shown in Figure 25.

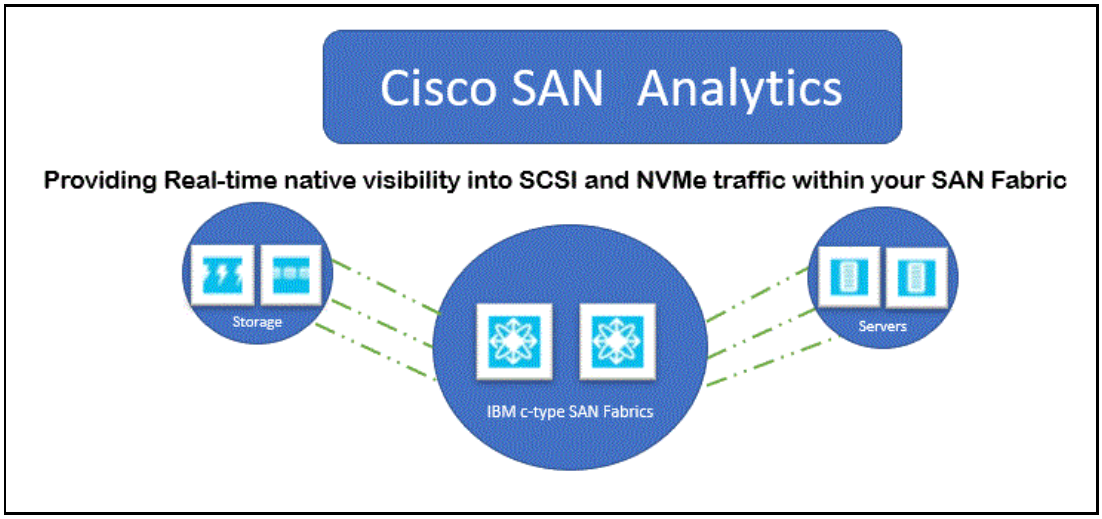


Figure 25 SAN Analytics

SAN Analytics is one of the industry's first solutions that provides insights into FC SAN traffic by inspecting FC frames natively on c-type switches without the use of external devices like appliances, taps, or probes. SAN Analytics help you to maintain performance, evaluate, and proactively troubleshoot issues across your organizations.

Figure 26 shows a snapshot of the slider that can be used to go back in time to identify patterns from 72 hours ago. Each dash is interpreted in one-hour increments. Custom graphing can be implemented to take a deeper dive into trending over longer periods of time.

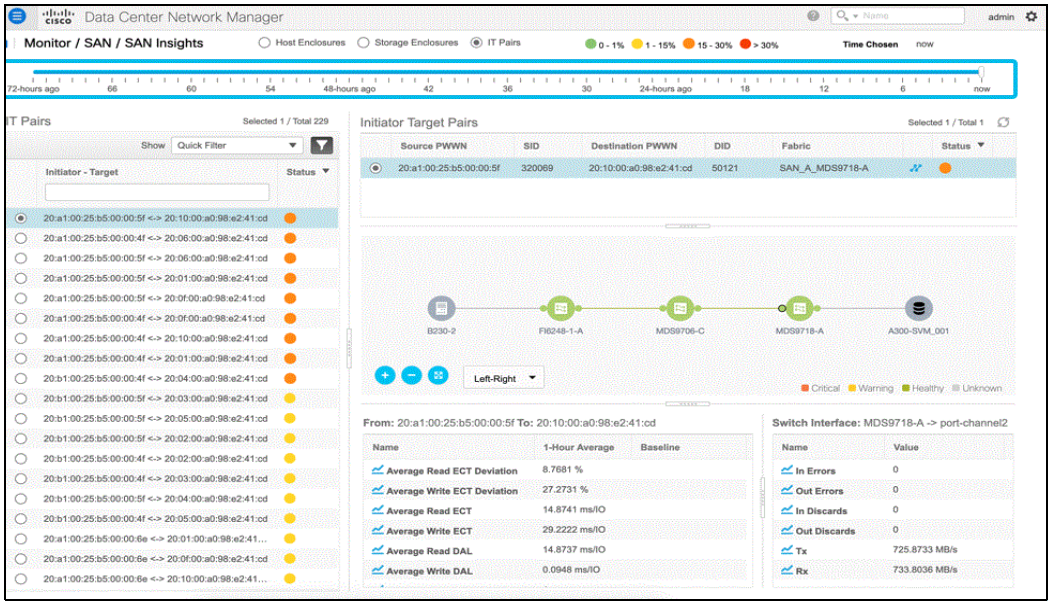


Figure 26 Slider snapshot

Important: For details, refer to:

SAN Analytics and SAN Telemetry Streaming Configuration Guide

SAN Insights Discovery

SAN Insights Discovery (SID) 2.0(1) is a free next-level storage area network health-check tool that is architecturally designed to provide detailed data for new and existing SAN fabrics. SID can support c-Type directors and switches and provide insights into SAN health, port usage, power consumption, licenses, zoning, and migrations. The information provided by SID reports allows you to review your environment for potential problems and proactively prepare for End Of Life (EOL) and End Of Support (EOS) conditions that occur throughout the life of your fabric. Performance of your switches, topology deployment, and inventory helps you make the best decision when managing c-Type fabrics.

The discovery mechanism includes two main components:

- SID collector, which is responsible for the collection of switch information
- Analysis Center, which is the cloud-based portal hosted on https://csid.cisco.com

Note: SID collector can be downloaded from https://www.cisco.com/ and requires a business single sign on (SSO) or a Cisco.com account to access and download the software.

SID collector runs as a standalone binary script locally on your Windows or Linux system and does not require software installation. SID leverages Secure Shell (SSH) sessions to connect to the seed switch in your fabric then runs **show** commands to congregate information about hardware inventory, ports, and performance data from the SAN switches it is discovering.

Then, it generates a data collection zip file that can be uploaded to the SID cloud portal for analysis, which will convert switch data that was collected into a visual display of your SAN fabric. You can run an SID collector on fabric A and fabric B in tandem to capture your entire environment over a period of time and upload multiple collections simultaneously to the cloud portal.

Note: To upload your SAN Insights data collection for analysis, see `https://csid.cisco.com` cloud portal.

Important: SID accounts are single tenant, dedicated to serving one individual customer so data is not shared unless the account owner provides users permissions to access the SID reports. SID security is defined in Cisco's `security compliance documentation`.

Figure 27 shows the secure cloud access portal used to access the SID analysis application.

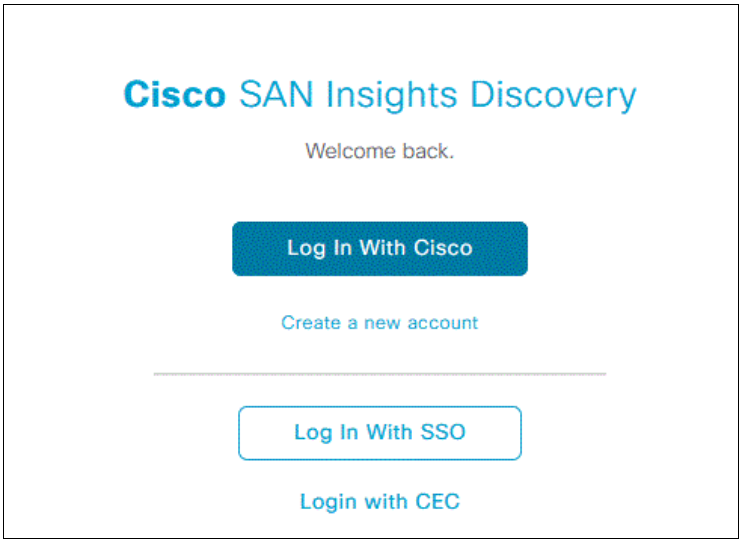


Figure 27 SID secure access portal

Figure 28 shows SID Users and Accounts menu option.

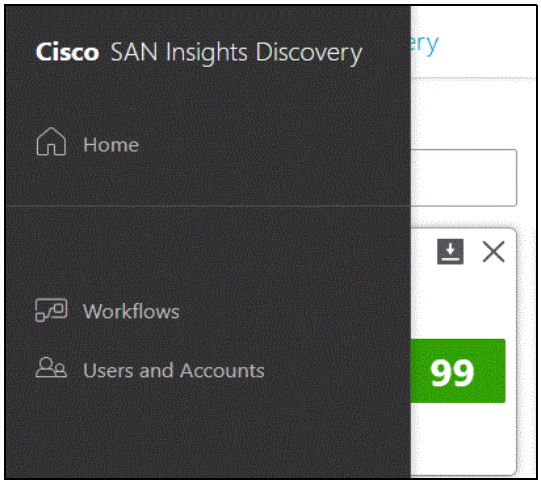


Figure 28 Users and Accounts

Figure 29 on page 42 shows general account details.

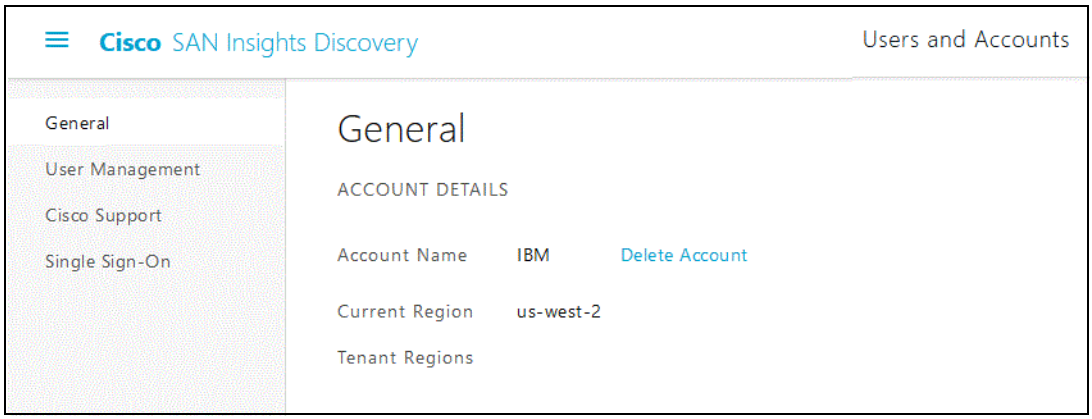


Figure 29 Account Details

Figure 30 shows the User Management panel that can be used to add, remove, or change member roles from account admin, network admin or simply to an observer. When providing access to your accounts reports, you can add users' email addresses to the SID cloud portal User Management section. This option provides users with the ability to download reports in CSV. format.

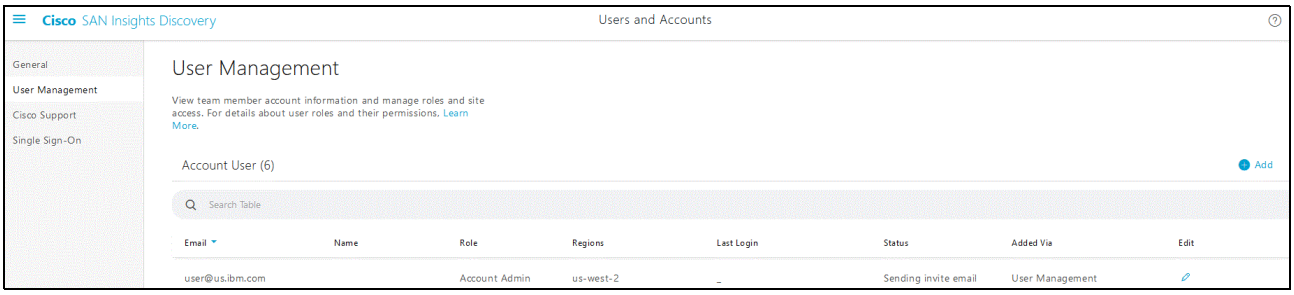


Figure 30 User Management

Figure 31 shows the My Reports dashboard, which allows you to search for previously-uploaded SAN fabric data collections. You can also upload, view, or download reports and delete reports that are analyzed by SAN Insights Discovery.

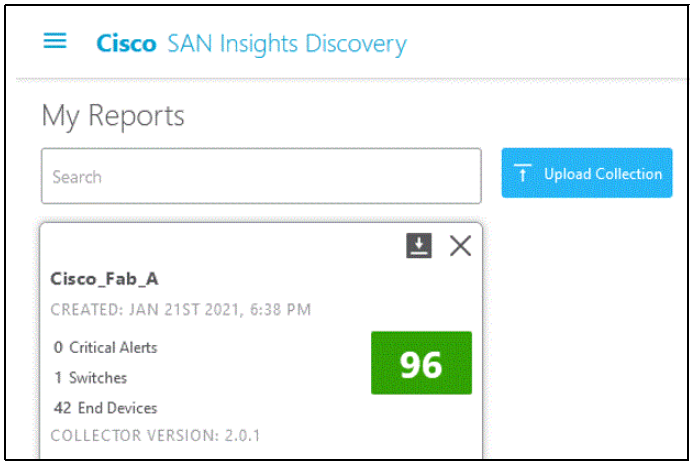


Figure 31 My Reports dashboard

Figure 32 shows the My Reports dashboard search for previously-uploaded SID reports.

Figure 32 Search Reports

Figure 33 illustrates how to delete SID reports.

Figure 33 Delete Reports

Figure 34 shows how to upload the SID data collection zip file from your local workstation.

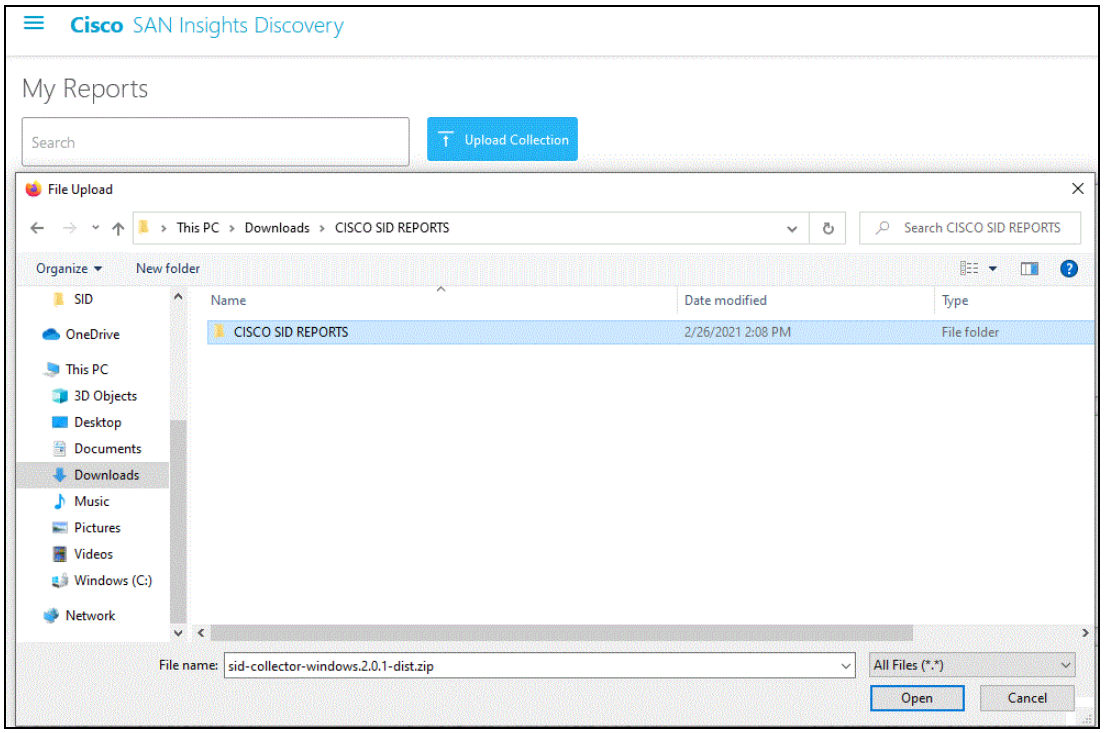


Figure 34 Upload Collection

Figure 35 shows information such as the analyzed report overview of fabric score, alerts, and EOS/EOL alerts.

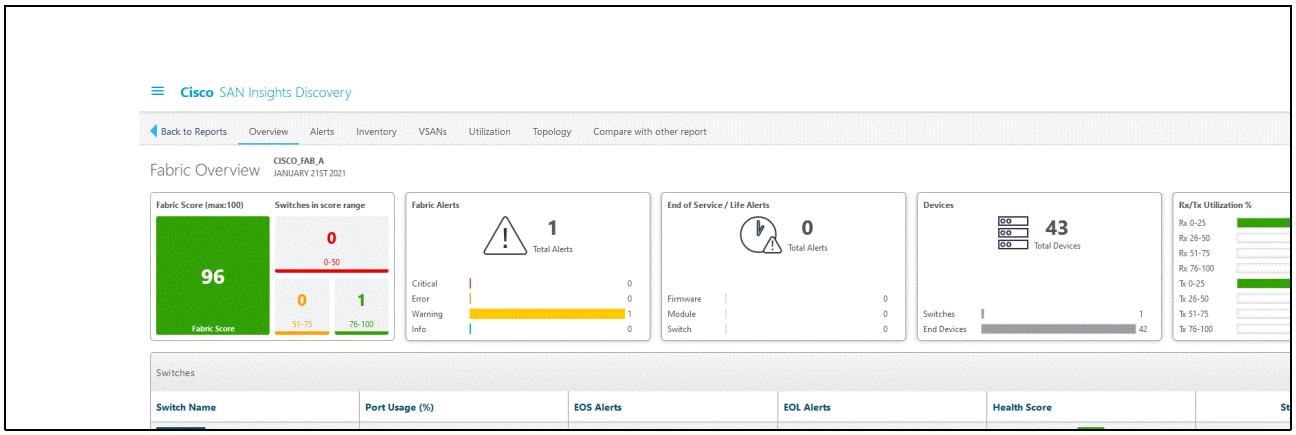


Figure 35 SAN Insights Discovery overview

Figure 36 shows the inventory view.

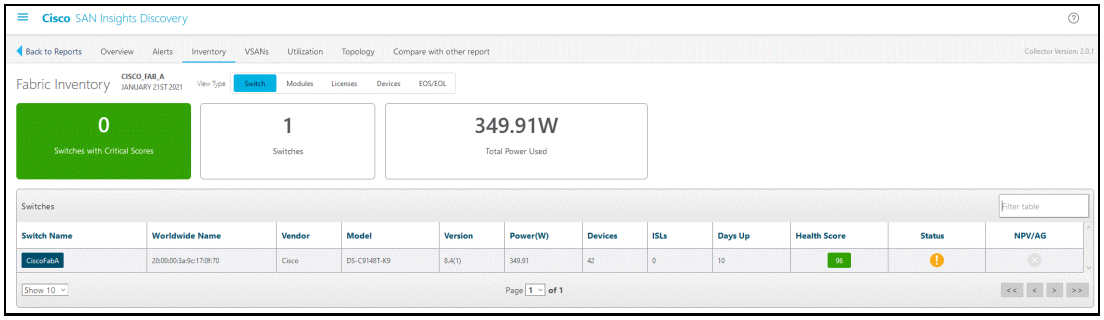


Figure 36 SAN Insights Discovery - Inventory view

Figure 37 shows the SAN fabric topology view and selected end devices.

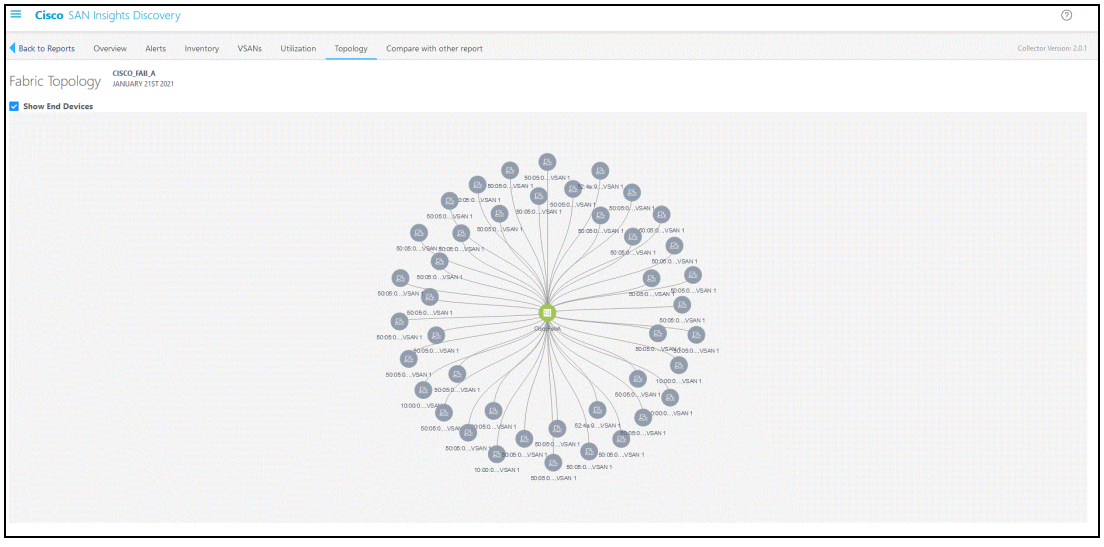


Figure 37 Topology view

SAN Insights Discovery provides the ability to share reports with your administrators. This capability provides a huge advantage when you proactively plan to prepare for firmware

upgrades, quality assurance measures, and IBM c-type hardware and fabric health checks prior to change activities.

The SID collector captures raw output from the switches and requires that you upload the collection to csid.cisco.com, where the data is analyzed, converted, and formatted into a report that provides a detailed evaluation of your SAN fabrics health.

Note: For more information, see:

https://www.cisco.com/c/en/us/products/storage-networking/index.html#~assess-fabric-health

Backup IBM c-type configuration

When you perform configuration changes, upgrades, or hardware replacements on IBM c-type switches, the best practice is to first perform a backup of the switch configuration. IBM c-type family can leverage Cisco NX-OS software that resides on the switches to create backups of the switch configuration. The backup copy of a configuration file stored in the internal memory can be sent to a remote server as a backup or to be used to configure other IBM c-type devices in your fabric. The commands are executed by the software when the device is started by the startup-config file or when commands are entered at the command prompt in configuration mode.

Cisco NX-OS software has two types of configuration files:

- Startup configuration: Used during device boot to configure the software features
- Running configuration: Contains the current configuration changes that must be made to the startup-configuration

These two configuration files can be different in instances where you want to change the device configuration temporarily without saving the running configuration changes to the startup-configuration.

Before you change the startup configuration file, save the running-configuration file to the startup configuration using the **`copy running-config`** and **`copy startup-config`** commands. You can also copy a configuration file from a backup copy located on a file server to the startup configuration.

To change the current running configuration, use the **`configure terminal`** command to enter configuration mode. When you enter global configuration mode, commands generally execute immediately and are then saved to the running configuration file immediately when the command is issued or when you exit configuration mode.

Best Practice: Backup your switch configuration using Secure File Transfer Protocol (SFTP) or SCP and save a copy to an external location before making changes.

Example 4 shows how to save the running -config by issuing the **`copy running-config`** and **`copy startup-config`** commands to store your current running configuration. Then, the saved configuration is copied to an SFTP server.

Example 4 Backing up the configuration from the command line

```
switch1# copy running-config startup-config
[########################################] 100%
Copy complete.
switch1# copy startup-config sftp://10.201.215.28/tmp/switch1_startup-config-date.txt
Enter username: user1

user1@10.201.215.28's password:
Connected to 10.201.215.28.
sftp> put  /var/tmp/vsh/switch1-startup-config  /tmp/switch1_startup-config-date.txt
Uploading /var/tmp/vsh/switch1-startup-config to /tmp/switch1_startup-config-date.txt
/var/tmp/vsh/switch1-startup-config                     100%  196KB 194.3KB/s   00:00
sftp> exit
Copy complete.
switch1#
```

The actions shown in Example 4 accomplish three things:

- Verify that you have an operational SFTP server in your environment.
- Verify that you can communicate to the server over the IP network.
- Allow you to store a copy of the configuration in a location that is external to the switch so that you have a backup in the event of a switch failure.

Alternatively you can perform a configuration backup of your switch using DCNM. This feature allows you to backup device configurations from the running configuration, using the command line interface. The backup files can be stored on the DCNM server or as recommended, to an external location.

Figure 38 shows how to access switch configuration by selecting **Configure** → **Backup** → **Switch Configuration**.

Figure 38 Switch Configuration

Figure 39 shows a list containing the Running and Startup configurations.

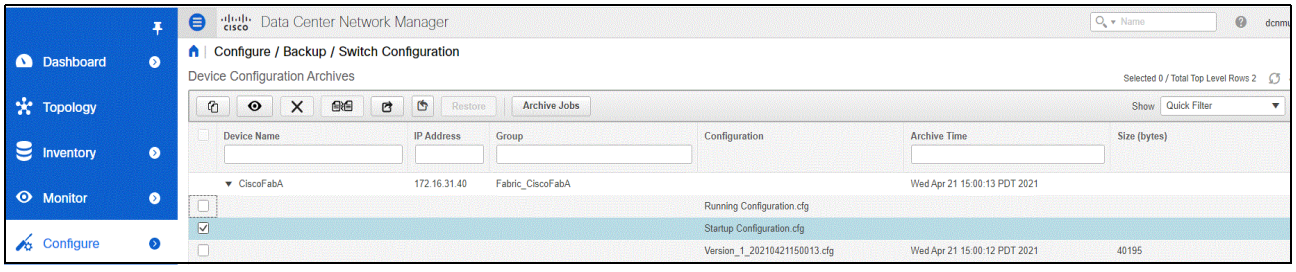


Figure 39 Running and Startup configurations

Figure 40 on page 47 shows the display of the Startup configuration.

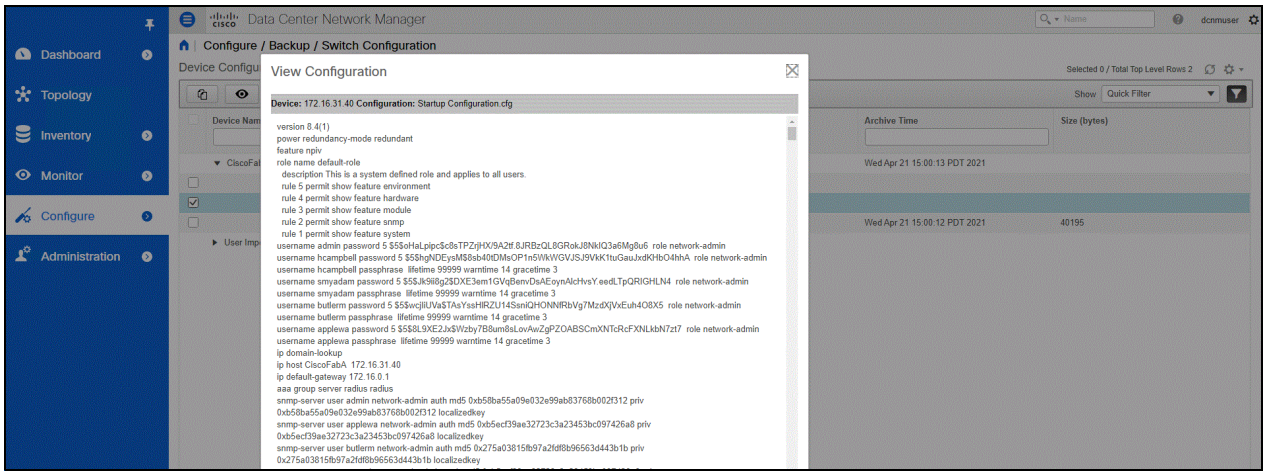


Figure 40 View configuration

Important: For more information on how to perform configuration backups, see: https://www.cisco.com/c/en/us/td/docs/dcn/dcnm/1151/configuration/san/cisco-dcnm-san-configuration-guide-1151/configure.html#topic_eks_lxx_pcb

Maintaining an optimal FCIP SAN environment

This section describes how to maintain an optimal FCIP SAN environment.

Ensure proper FCIP bandwidth

The task of determining the necessary FCIP bandwidth consists of two general objectives.

- Define the bandwidth that is required to meet the needs of the business.
- Validate that the actual bandwidth that is delivered matches the projected and expected bandwidth of the solution as implemented.

Latency or round-trip time (RTT) plays an important role in high-performance environments and when long-distance links are involved. Although not much can be done to reduce RTT, it is important to understand the role it can play in long-distance solutions.

This section is not intended to be a broad architectural review, but rather more specific to the type of questions that should be considered, and the tests and methodologies that can be performed to achieve the objectives. Several implementation and design considerations exist for an FCIP SAN environment.

Amount of bandwidth

There are a number of basic questions to be answered when determining the amount of bandwidth that is necessary for the needs of a business.

- The amount of data that will be transferred or replicated needs to be calculated.
- The amount of time for how quickly this data must be transferred from the local to remote site needs to be determined

When these values are known, you can determine a starting point for the minimum amount of bandwidth necessary by dividing the total data amount by the time.

To determine the current amount of data to be transferred or replicated, the following questions need to be considered:

- How much data needs to be initially transferred for synchronized copies at both the local and remote data sites?
- How frequently is the initial data being changed?
- What is the profile of the change-rate over a period of time?
- What is the maximum amount of the data change-rate?

The first question, listed above, is the starting point for data-sizing. Using a data replication solution as an example, the initial amount of data to be synchronized to a remote data site must be determined. To do this, you must determine which local data sets, volumes, and consistency groups need to be replicated to the remote data site. Then, you can calculate the total size of all the data to be transferred.

If multiple storage systems replicate data between the local and remote data sites, then all the individual replication streams must have a reliable answer for the amount of data to be initially transferred. Therefore, the characteristics of data transfer and replication applications that share a given FCIP tunnel must be understood.

Data sets, volumes, and consistency groups are rarely consistent in size. Therefore, most storage systems apply a fairness algorithm so that each item to be transferred is given equal portions of the bandwidth. With a mixture of large and small volumes and consistency groups, the balanced transfer rates result in the smaller volumes or consistency groups being synchronized before the larger ones.

After a smaller volume or consistency group is synchronized, most systems begin to transfer data that was changed in the source volume or consistency group, while the larger volumes and consistency groups are still being synchronized. This sizing scope is determined with questions about change frequency, workload profile, maximum change, and change rate over a period of time.

When the amount of initial data to be transferred or replicated is understood, the next step involves determining the rate of change of the data to be mirrored. This value is not a percentage of how much of each volume or consistency group is changing, but the amount of data that changes in terms of size, such as bytes. Depending on the type of data and applications that are using the data, the change rate might be consistent over time, or it might vary greatly over a given time period.

The best answer for the change rate of each data set, volume, or consistency group is the maximum change rate for a set time period, such as 24 hours. The change rates for the various data units to be transferred or replicated should be determined based on a common time period.

The combination of the initial amount of data to be synchronized with the amount of change data is the total scope of the data to be transferred or replicated. The next step is to determine what the business needs, or the requirements in terms of how quickly the data can be initially transferred. Then, you must determine the recovery point objective (RPO).

When the time factor is known, the bandwidth that is needed for the FCIP tunnel can be calculated to determine the bandwidth value in bits per second. This bandwidth setting is needed for the current amount of data to be replicated.

At this point, growth-over-time is the one additional factor that must be considered. Over time, most businesses experience growth of their replication needs. The exercises and calculations are for current needs; data growth has not yet been considered.

As a business grows and expands, the amount of data to be synchronized and the change rate will likely increase over time. Therefore, meeting the current bandwidth needs is likely to be insufficient for the future operations of the replication solution.

The current bandwidth requirement must be adjusted accordingly for future bandwidth-needs based on trend metrics. If a business has been experiencing data growth of approximately 25% per year, then the replication needs in a year will likely see similar growth. There are no rigid rules for “future proofing” bandwidth needs. However, future bandwidth-needs must be considered and evaluated and will result in an adjustment to the current bandwidth-needs to account for future data growth.

For more information, see the Cisco MDS 9000 Series IP Services Configuration Guide. Throughput for any compression mode is dependent on the compressibility of the data to be replicated. Compression will not provide a reduction of bandwidth needs.

Actual versus allocated bandwidth

Continuing with the replication solution scenario, the following simple example for bandwidth-needs is provided. A company has 17 TB of data that needs to be replicated, and the maximum change rate is 3 TB per day for a total of 20 TB. The business needs of the company state that initial synchronization with changed data during the synchronization period is to be completed within 24 hours.

Trending data shows that the growth rate of the data is just under 10% per year. Therefore, the calculations are as follows:

```
((17 TB + 3 TB) * 8 bits/ B) / (24 hours * 3600 second/hour) = 1.852 Gbps
```

To account for the growth over one year, the bandwidth needed should be adjusted:

```
1.852 Gbps * 1.10 = 2.037 Gbps
```

By rounding down, the company should plan on approximately 2 Gbps total bandwidth between the local and remote data sites to meet the current replication requirements and remain viable for almost a year into the future. With redundant fabrics, one implementation design for this replication solution example could be comprised of a single 1 Gbps link per fabric across two redundant fabrics.

Therefore, the bandwidth-needs for replication have been determined and implemented as 1 Gbps FCIP tunnel-per-fabric between the local and remote data sites. The next suggested step is to verify that the actual bandwidth meets the design target before the FCIP tunnels are put into production.

Best Practice: Ensure that the FCIP links are sized in both bandwidth and latency to meet the expected workloads.

Link quality

The characteristics of the FC and IP protocols are different regarding packet-loss:

- FC is based on lossless connections between device ports with a high emphasis on in-order delivery of frames.
- IP is based on the assumption that some degree of packet-loss will occur.

The FC-SCSI protocol is sensitive to response times. The higher the response times (round-trip times), the lower the throughput will be. FC-SCSI is sensitive to fluctuating latencies, and tends to experience issues when latency is not consistent. This means that the FCIP links have a much lower tolerance to out-of-order, slow start, and retransmits than typical IP links.

Jitter

Jitter is the measurement of how much the round-trip time (RTT) changes from its nominal value. If the RTT value is 50 ms, then a jitter of 5% indicates that the RTT values were ranging from 47.5 ms to 52.5 ms. There are always conditions that cause the RTT to fluctuate. This is expected but we recommend that the fluctuations (jitter) be under 20% and, if possible, closer to 10% for sustained periods. This means that having a high jitter for 15 mins is likely ok, but having high jitter all the time or for periods of hours is likely going to impact through put and cause timeouts.

Best Practice: Jitter should not vary by more than 15 to 20%.

Retransmits

Packet-loss is more common in IP networks than frame-drop is in FC networks. The protocol accommodates this situation by having the receiving-end request a retransmit when packet-loss is detected. The need to retransmit packets increases the latency to the end-devices. In SCSI, this can lead to timeouts or replication suspensions due to increased latency.

The acceptable retransmit-levels in an FCIP network are typically much lower than in an FC networks. In an FCIP network, retransmit levels should be under 0.05% and typically closer to 0.01%.

Onboard logging can be configured to record when retransmits exceed a given level, with the *tcp logging onboard tcp-retransmission-threshold* parameter and has a default value of 0.05%. The retransmission is set in each FCIP profile and can be displayed by using the **`show FCIP profile`** command.

Example 5 show fcip profile command

```
FCIP Profile 12
    Internet Address is 10.1.1.100  (interface IPStorage2/1)
    Listen Port is 3225
    TCP parameters
        SACK is enabled
        PMTU discovery is enabled, reset timeout is 3600 sec
        Keep alive is 60 sec
        Minimum retransmission timeout is 200 ms
        Maximum number of re-transmissions is 4
        Retransmission rate of OBFL logging threshold is 0.05%
        Maximum allowed bandwidth is 10000000 kbps
        Minimum available bandwidth is 8000000 kbps
        Configured round trip time is 1000 usec
        Congestion window monitoring is enabled, burst size is 50 KB
        Auto jitter detection is enabled
```

Best Practice: Retransmits should be under 0.05%

Out of Order

Like jitter and retransmits, Out-of-Order packets are not handled well by FCIP implementation. Out-of-Order packets tend to create issues at the FC and SCSI layers, which can create timeouts or equipment checks.

IBM recommends limiting Out-of-Order packets to under 0.05% and ideally to under 0.01% similar to the retransmit rates

Best Practice: Out of Order packets should be under 0.05%

Use multiple network links

For high availability, it is important to use multiple network connections between the local and remote switches. You can accomplish this in several ways, such as having multiple links in a port channel, FSPF-Based Load Balancing, or Virtual Router Redundancy Protocol (VRRP). The most common and simple solution is to have multiple FCIP links in a port channel.

The physical connections to the network are done via the IPStorage ports, which are the network interface connections to the network and have the properties of link speed, media type, full or half duplex.

An FCIP profile is created and provides the IPStorage ports with the TCP layer information, such as IP address, retransmit thresholds, and some FC information. The FC link is the bridge between the FC layer and the TCP/IP layers. This link is known as the *FCIP link*.

The two FCIP links are then associated with the port channel, which is then defined to the different VSANs that will be using it.

The example shown uses two Ethernet links connected to one port channel. In this configuration, FC traffic is sent to the port channel and then spread across the two FCIP links, encapsulated into packets, and sent out over the two IPStorage links. Figure 41 shows two IPStorage ports associated with two FCIP links and one port channel.

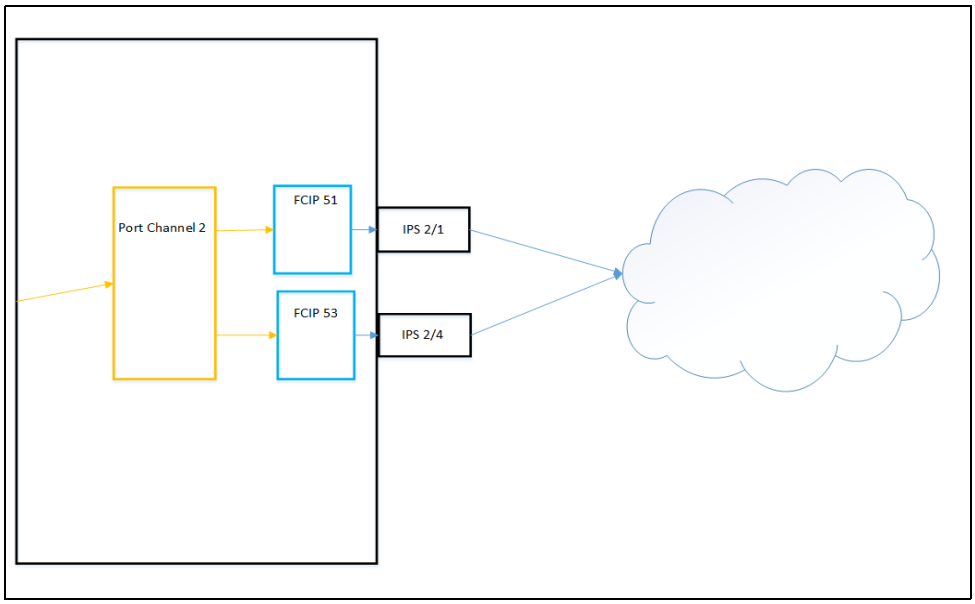


Figure 41 FCIP setup with a port channel and two FCIP network links

To configure and set up FCIP on the CISCO switches, use the Cisco MDS 9000 Series IP Services Configuration Guide, which has step-by-step procedures and other details.

Best Practice: Use multiple network links to connect local and remote switches that use FCIP. In most setups, it is appropriate to associate multiple FCIP links to a port channel.

Accelerators

Response times play a major role in the amount of data that can be sent over an extended link. To help in this area, accelerators are available where the local FCIP switch sends responses immediately before the remote switch / devices have sent them to allow the local device/host to continue to send more data immediately, which increases throughput.

Sending these early responses can create issue when the actual response is not the same as the early response and not all devices can deal with this condition. You need to check which accelerators are recommended for use with the host software, and devices you are using.

Best Practice: Do not use FCIP Write Acceleration with IBM disk storage devices. Use FCIP Tape Acceleration for tape devices if they do not share the same port channel as the disk storage.

Traffic optimizers

Several network optimizers that can be used to modify the network traffic flows to maximize the network's bandwidth. An optimizer like Silver Peak is not recommended and, for most storage devices, is not supported because Small Computer System Interface (SCSI) is a latency-sensitive protocol and the use of these types of programs can introduce variable latencies.

Monitoring

Monitoring the IP links in an FCIP network is an excellent way to identify issues before they have an impact on the network. The ability to monitor the IP network from the FCIP switches is limited and can usually be best achieved by the network-monitoring tools. However, you can also use commands to display the key values from the switch.

Round-Trip Time

The **`show interface fcip counters`** command contains the round-trip time and the jitter values.

Example 6 show interface fcip counters command

```
sc9706Cisco3# show int fcip 51 counters
fcip51
    TCP Connection Information
      2 Active TCP connections
        2 Attempts for active connections, 0 close of connections
        Path MTU 1500 bytes
        Current retransmission timeout is 200 ms
        Current Send Buffer Size: 758 KB, Requested Send Buffer Size: 0 KB
        CWM Burst Size: 50 KB
 CONN<0>
```

```
    Data connection: Local 10.1.1.100:65534, Remote 10.1.1.101:65534
    TCP Parameters
      Round trip time: Smoothed 16 ms, Variance: 8 Jitter: 150 us
      Advertized window: Current: 758 KB, Maximum: 24580 KB, Scale: 5
      Peer receive window: Current: 2047 KB, Maximum: 2047 KB, Scale: 5
      Congestion window: Current: 593 KB, Slow start threshold: 1950 KB
      Measured RTT : 500000 us Min RTT: 6440 us Max RTT: 0 us
    TCP Connection Rate
      Input Bytes: 0.00 MB/sec, Output Bytes: 0.00 MB/sec
      Input Frames: 0/sec, Output Frames: 0/sec
 CONN<1>
    Control connection: Local 10.1.1.100:65533, Remote 10.1.1.101:65533
    TCP Parameters
      Round trip time: Smoothed 1 ms, Variance: 1 Jitter: 151 us
      Advertized window: Current: 749 KB, Maximum: 24580 KB, Scale: 5
      Peer receive window: Current: 2045 KB, Maximum: 2045 KB, Scale: 5
      Congestion window: Current: 50 KB, Slow start threshold: 1947 KB
      Measured RTT : 500000 us Min RTT: 17 us Max RTT: 0 us
```

Interface errors

To display interface errors (drops and collisions), use the **show interface ipStorage counters** command.

Example 7 show interface ipStorage counters command

```
sc9706Cisco3# show int ipStorage 2/1 counters
IPStorage2/1
    5 minutes input rate 8 bits/sec, 1 bytes/sec, 0 frames/sec
    5 minutes output rate 8 bits/sec, 1 bytes/sec, 0 frames/sec
    10716 packets input, 675728 bytes
      10248 multicast frames, 512 broadcast frames
      0 errors, 0 queue drops, 1 if-down drops, 0 RED drops
      0 bad ether type drop, 0 bad protocol drops
    10717 packets output, 639034 bytes, 0  underruns
      0 multicast, 2 broadcast
      0 errors, 0 collisions, 0 arp drops, 0 if-down drops
```

Retransmits

Retransmits can be see in the Onboard Failure Log (OBFF). To display retransmits, use the **show logging onboard** command.

Example 8 Logging logfile for TCP MAX transmit messages

```
sc9706Cisco3# show logging logfile | grep TCP
%PORT-5-IF_DOWN_TCP_MAX_RETRANSMIT: %$VSAN 1%$ Interface fcip51 is down(TCP conn. closed - retransmit
failure) port-channel2
%PORT-5-IF_DOWN_TCP_MAX_RETRANSMIT: %$VSAN 1%$ Interface fcip53 is down(TCP conn. closed - retransmit
failure) port-channel2
%PORT-5-IF_DOWN_TCP_MAX_RETRANSMIT: %$VSAN 1%$ Interface fcip51 is down(TCP conn. closed - retransmit
failure) port-channel2
%PORT-5-IF_DOWN_TCP_KEEP_ALIVE_EXPIRED: %$VSAN 1%$ Interface fcip53 is down(TCP conn. closed - Keep alive
expired) port-channel2
%PORT-5-IF_DOWN_PEER_CLOSE: %$VSAN 1%$ Interface fcip51 is down(TCP conn. closed by peer) port-channel2
%PORT-5-IF_DOWN_PEER_CLOSE: %$VSAN 1%$ Interface fcip53 is down(TCP conn. closed by peer) port-channel2
```

FICON

FICON attachment is a licensed feature and requires the MAINFRAME_PKG license. To extend the FICON attachment feature using FCIP, the SAN_EXTN_OVER_IP license is also required. By default, FICON is disabled. Therefore, you must enable the FICON feature with the **feature ficon** command.

To use the FICON setup wizard, enter the **setup ficon** command.

For information on configuring switches for FICON, see:

- Cisco MDS 9000 Series Fabric Configuration Guide
- Cisco FICON Basic Implementation, REDP-4392

FICON VSANs

Although open systems (sometimes called distributed systems) can be mixed in the same VSAN as FICON, it is not a recommended practice. FICON and open-systems traffic have different characteristics and often do not work well together. When FICON and open-systems traffic are in different VSANs, there is excellent traffic isolation between the two workloads.

FICON typically uses some different settings from open systems, such as in-order delivery enabled and default zoning enabled. When FICON is in its own VSAN, the settings for FICON and open systems can be different.

Best Practice: Use separate VSANs for FICON and Open Systems.

In-Order delivery

In most cases, fabrics deliver data in the same order that it is sent. However, in some situations data can be delivered out of order, especially when multiple interswitch links or marginal links are present. Unlike open systems, FICON is sensitive that the order received is the same as the order sent. To ensure that every possible in-order delivery occurs, the in-order delivery setting must be enabled in all FICON VSANs. To enable in-order delivery, use the **in-order-guarantee vsan #** command. In Example 9, in-order delivery is enabled for VSAN5.

Example 9 in-order-guarantee vsan # command

```
sc9706Cisco3# show in-order-guarantee
global inorder delivery configuration:not guaranteed

VSAN specific settings
vsan 1 inorder delivery:not guaranteed
vsan 3 inorder delivery:not guaranteed
vsan 4 inorder delivery:not guaranteed
vsan 5 inorder delivery:guaranteed
vsan 10 inorder delivery:not guaranteed
vsan 12 inorder delivery:not guaranteed
```

Best Practice: Enable in order delivery for all FICON VSANs

FICON zoning

FICON uses a configuration file to define how the FICON channels are connected to the storage device. Unlike open systems, FICON does not rely on the switch to provide connection information or require the switch to enforce the allowed connections.

Typically in a FICON environment, we enable the VSAN to allow all ports to communicate with each other by permitting the default-zone. This is done with the **`zone default-zone permit vsan #`** command. In Example 10, the default-zone is enabled for VSAN5.

Example 10 zone show policy command

```
sc9706Cisco3# show zone policy
Vsan: 5
   Default-zone: permit
   Distribute: full
   Broadcast: unsupported
   Merge control: allow
   Generic Service: read-write
   Smart-zone: disabled
```

Best Practice: Enable default-zone on all FICON VSANs

Fabric binding

When the FICON channel is attached to one switch and the end device (usually storage) is attached to a different switch, this is referred to as *cascading switches*. For cascading switches, fabric binding must be enabled. FICON channels query the attached switch port to verify that fabric binding is enabled and will not allow the channel to be activated if fabric binding not enabled. To enable fabric binding, use the **`fabric-binding activate vsan # force`** command.

The switches that have the FICON channels and FICON devices attached must be added to the binding table. To display the list of switches in the binding table, use the **`show fabric-binding database`** command. In Example 11 fabric-binding is activated on VSAN5.

Example 11 show fabric-binding status command

```
VSAN 1 :No Active database
VSAN 3 :No Active database
VSAN 4 :No Active database
VSAN 5 :Activated database
VSAN 10 :No Active database
VSAN 12 :No Active database
```

Best Practice: Configurations that contain multiple switches must use the fabric-binding feature.

FICON tape accelerator

Tape workloads are sequential in nature. When they flow over long distances, high latencies exist due to the high round-trip time in obtaining acknowledgments from the remote device. The FICON Tape Accelerator can be enabled so that the local switch will send the acknowledgments to allow the FICON host to continue sending data to the remote device.

The FICON tape accelerator can be used in long-distance extensions, such as dark fiber or FCIP. The FICON tape accelerator should not be confused with the FCIP tape accelerator, which is for open-systems tape that uses an FCIP extension link. For more information, see “Accelerators” on page 52.

Best Practice: Use the FICON Tape accelerator for tape flows over an extended distance.

FICON XRC accelerator

Future IBM storage devices will not support XRC. All new installations should consider using device-based Metro or Global Mirror, instead of XRC, to replicate data.

Summary of best practices

Table 1 Summary of best practices

Section	Practice	Reference
General	Avoid introducing devices to the SAN that span more than one generation of technology.	“High-performance networks” on page 6
General	Avoid traversing ISLs when accessing SSD/Flash for high-performance use cases.	“High-performance networks” on page 6
Design	A core-edge design is preferred for most fabrics. If you are using a dual-core design, ensure that devices are not zoned across the ISLs between the core switches unless it is necessary to do so.	“Fabric topology” on page 7
Design	Separate replication traffic from production traffic on mult-site links and use the IVR feature to prevent fabrics from merging across sites.	“Multi-site fabrics” on page 9
Design	Use a port-channel between switches with multiple ISL for redundancy and enough ISLs to meet at least 80% of the bandwidth requirements.	“Port-channels” on page 11
Design	Only configure multiple VSANs to use the same ISLs and Port-Channels where necessary.	“VSAN trunking” on page 12
Design	Leave the default exchanged-based routing policy in place	“Routing policies for open-systems fabrics” on page 13
Design	Have a meaningful naming convention.	“Meaningful naming convention” on page 13
Design	Use NPort Virtualization (NPV) on embedded switches and smaller switches.	“N Port Virtualization” on page 16
Zoning	Use WWPN zoning.	“Zoning types” on page 19
Zoning	Use target initiator zoning for small to medium fabrics. Use smart zoning for large fabrics.	“Smart zoning” on page 19
Maintain	Update firmware every 6-18 months. Check the Cisco recommended releases for target firmware levels.	“Switch firmware levels” on page 21
Monitoring	Deactivate the default slow-drain policy and copy the existing Core and Edge policies to custom policies. Add the counters in the slow-drain policy to those new policies and deploy those policies.	“Port-Monitor” on page 21

Section	Practice	Reference
Monitoring	Enable remote and port monitoring for improved link monitoring and notifications.	“Remote Monitoring” on page 24
Monitoring	For switches maintained by IBM implement Storage Insights.	“IBM Storage Insights” on page 27
Backup	Back up the configuration on all of your switches on a regular basis.	“Backup IBM c-type configuration” on page 45
FCIP	Ensure the FCIP links are sized in both bandwidth and latency to meet the expected workloads.	“Ensure proper FCIP bandwidth” on page 47
FCIP	Jitter should not vary by more than 15 to 20%.	“Jitter” on page 50
FCIP	Retransmits should be under 0.05%.	“Retransmits” on page 50
FCIP	Out of Order packets should be under 0.05%.	“Out of Order” on page 51
FCIP	Use multiple network links to connect local and remote switches using FCIP. Associating multiple FCIP Links to a port channel is the most appropriate for most setups.	“Use multiple network links” on page 51
FCIP	Do not use FCIP write acceleration with IBM disk storage devices. Use FCIP Tape Acceleration for tape devices if they do not share the same port channel as the disk storage.	“Accelerators” on page 52
FICON	Use separate VSANs for FICON attach.	“FICON VSANs” on page 54
FICON	Requirement: Enable in order delivery for all FICON VSANs.	“In-Order delivery” on page 54
FICON	Requirement: Enable default-zone on all FICON VSANs.	“FICON zoning” on page 55
FICON	Requirement: Configurations with multiple switches must use the fabric binding feature.	“Fabric binding” on page 55
FICON	Use the FICON Tape accelerator for any tape flows over an extended distance.	“FICON tape accelerator” on page 56

Authors

This paper was produced by a team of specialists from around the world.

David Green works with IBM SAN Central troubleshooting fibre-channel networks and storage performance problems. He is the author of multiple IBM Redbooks® and a frequent speaker at IBM Technical University.

David Lutz is a Consulting Remote Technical Support Specialist in Canada. He has 41 years of experience in service and support of mainframe and storage products. David is a member of the Field Assist Support Team, which provides storage resilience services.

Lyle Ramsey is a veteran IT management executive and consultant and currently collaboratively oversees the IBM GTS CISCO SAN Architecture Global Strategy. He has more than two decades of experience leading large-scale information technology programs within private and government sectors. Lyle is a Cisco MDS and DCNM SME in addition over the years holding qualification certifications from SNIA, Brocade, NetApp, EMC, Microsoft, and Academy of Business. Lyle is based in Phoenix, Arizona.

Bhavin Yadav is a Technical Marketing lead within Cisco's SAN DC Switching unit., based out of Silicon Valley. He has been with Cisco SAN switching unit for a decade. He is an avid blog writer, has written a number of white papers and technical articles, and is a Cisco Live recognized speaker on SAN migration technology. Currently, he focuses mainly on technical enablement activities, along with designing various Cisco SAN training courses delivered through Cisco Learning Academy.

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

Mike Blair
Technical Leader, Cisco Systems

Edward Mazurek
Americas Data Center Technical Leader, Cisco Systems

Bert Dufrasne
Project Leader, IBM Redbooks, San Jose Center

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 IBM Redbooks 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

Stay connected to IBM Redbooks

- Find 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

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

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, MD-NC119, Armonk, NY 10504-1785, US

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 jurisdictions 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 provide in any way it believes appropriate without incurring any obligation to you.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

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.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

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 actual people or business enterprises 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. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at http://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks or registered trademarks of International Business Machines Corporation, and might also be trademarks or registered trademarks in other countries.

Redbooks (logo)  ®	FICON®	IBM Spectrum®
AIX®	IBM®	Redbooks®

The following terms are trademarks of other companies:

The registered trademark Linux® is used pursuant to a sublicense from the Linux Foundation, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis.

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.

VMware, and the VMware logo are registered trademarks or trademarks of VMware, Inc. or its subsidiaries in the United States and/or other jurisdictions.

Other company, product, or service names may be trademarks or service marks of others.

REDP-5632-00

ISBN 0738460079

Printed in U.S.A.