IBM zEnterprise BladeCenter Extension

Network Connectivity Options

- **Learn about zBX connectivity options**
- **Use technologies from multiple vendors for zBX connections**
- **See how zBX achieves high availability**

Patty Driver
Gwen Dente
Friedrich Welter

Redpaper

ibm.com/redbooks

International Technical Support Organization

IBM zEnterprise BladeCenter Extension: Network Connectivity Options

May 2014

REDP-5036-00

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

First Edition (May 2014)

This edition applies to zEnterprise BladeCenter Extension Model 003.

This document was created or updated on May 6, 2014.

© Copyright International Business Machines Corporation 2014. All rights reserved.
Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

© Copyright IBM Corp. 2014. All rights reserved.

Notices

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

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

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

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

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

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

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

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

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

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

COPYRIGHT LICENSE:

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

© Copyright IBM Corp. 2014. All rights reserved.

Trademarks

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

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

BladeCenter®
BNT®
eServer™
IBM®
PR/SM™
Redbooks®
Redpaper™
Redbooks (logo) ®
Resource Link®
System z®
VTAM®
z/OS®
zEnterprise®

The following terms are trademarks of other companies:

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

Preface

This IBM® Redpaper™ publication describes the configuration of the networking equipment that attaches to the IBM zEnterprise® BladeCenter® Extension (zBX), which allows communication with the applications that reside on the intraensemble data network (IEDN). In most cases, the IEDN remains a closed Layer-2 network to maintain a highly available and secure environment that IBM can support. Therefore, when connecting to the IEDN, Layer-3 *routed* connectivity is still the preferred method. However, now the zBX top-of-rack (TOR) switches support Layer-2 *switched* connections that can provide an easier migration path when moving data center workloads to the zBX environment.

This paper includes a brief introduction to the IEDN architecture and configuration and how these types of connections work. It also introduces the zBX architecture and explains the implications that network connections can have on the redundancy and high availability setup for this system. Finally, this paper provides concrete examples for connecting the IEDN and external data network through zBX for both Layer-3 routed and Layer-2 switched connection configuration options.

This paper is intended for network architects and network administrators who are responsible for designing and implementing zBX network configurations. It is assumed that you have a basic background in IBM zEnterprise and network concepts.

Authors

This paper was produced by a team of specialists from around the world working at the IBM International Technical Support Organization, Poughkeepsie Center.

Patty Driver is an I/O Technologist and Firmware Engineer for zFirmware Core Technologies, IBM Systems and Technology Group, Power and z Systems, IBM System z® Hardware Brand Technology Development in Poughkeepsie, New York.

Gwen Dente is a Certified IT Specialist IBM eServer™, zSeries, Communications (IBM VTAM®, IP, EE), Security - ATS, Americas, Client Technical Specialist: SYS.System z, IBM Sales and Distribution, Technical Sales Support in Gaithersburg, MD.

Friedrich Welter is with the IBM Systems and Technology Group, Enterprise Systems and Technology Development in Boeblingen, Germany.

Thanks to the following people for their contributions to this project: Ella Buslovich, Gary Fisher, Gus Kassimis, Angelo Macchiano, Irena Slywkanycz, and Jerry Stevens.

© Copyright IBM Corp. 2014. All rights reserved.

Now you can become a published author, too!

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

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

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

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

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

  ibm.com/redbooks

- Send your comments in an email to:

  redbooks@us.ibm.com

- Mail your comments to:

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

Stay connected to IBM Redbooks publications

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks

- Follow us on Twitter:

  http://twitter.com/ibmredbooks

- Look for us on LinkedIn:

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

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

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

- Stay current on recent Redbooks publications with RSS Feeds:

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

1

Introduction

IBM is expanding the requirements for Layer-3 connectivity between the external data network and the intraensemble data network (IEDN) top-of-rack (TOR) of zBX. Test cases show that IBM can support a wider set of connectivity alternatives than those that were originally recommended.

IBM continues to recommend that the IEDN should remain a closed Layer-2 network, but now you can attach external zBX ports with either Layer-2 or Layer-3 definitions. The eased Layer-3 routed connectivity requirements allow a Layer-2 switched connection to the zBX. If you choose this option, you cannot use the centralized control of Unified Resource Manager (zManager). If you choose to use the Layer-2 connectivity, make sure that you follow these rules:

1. Loop detection and elimination protocols that require using IEEE standard Bridge Protocol Data Unit (BPDU) packets to detect and eliminate switching loops, such as Spanning Tree Protocol (STP), Rapid Spanning Tree Protocol (RSTP), Multiple Spanning Tree Protocol (MSTP), and so on, are not permitted on the links between customer networking equipment and the zBX TOR switches.
2. Loop detection and elimination protocols that do not rely on IEEE standard BPDU packets, such as Per VLAN Spanning Tree Plus (PVST+), Simple Loop Prevention Protocol (SLPP), are permitted.
3. The allowable IEDN VLAN range is limited to VLANs 10 - 1030.

1.1 Overview of IEDN architecture and design

The zEnterprise zManager is the single management entity for all the heterogeneous resources that comprise a zEnterprise ensemble. zManager orchestrates platform management and virtualization, providing monitoring, provisioning, and control for the ensemble resources.

Network Virtualization Management (NVM), a zManager function, provides a single point of control for provisioning a secure private data network called the *intraensemble data network* (IEDN) which is used to communicate between applications running on hosts in an ensemble. The security and isolation of network traffic between hosts on the IEDN is provided through

© Copyright IBM Corp. 2014. All rights reserved.

NVM in collaboration with other firmware and hardware elements of the zEnterprise system. The IEDN spans the entire ensemble, which can be formed by a collection of up to eight zEnterprise systems, also called *nodes*, managed by a single instance of the zManager that is on the Hardware Management Console (HMC), as shown in Figure 1-1.

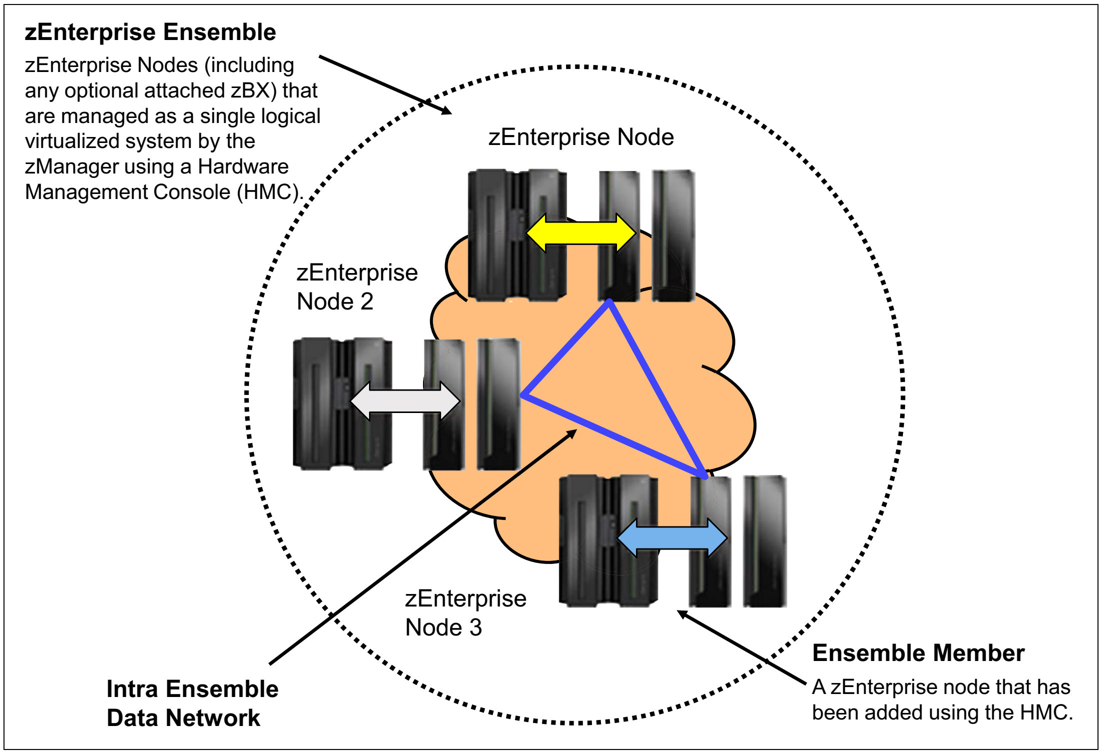
zEnterprise Ensemble
zEnterprise Nodes (including any optional attached zBX) that are managed as a single logical virtualized system by the zManager using a Hardware Management Console (HMC).
zEnterprise Node
zEnterprise Node 2
zEnterprise Node 3
Ensemble Member
A zEnterprise node that has been added using the HMC.
Intra Ensemble Data Network

Figure 1-1 View of zEnterprise ensemble and IEDN

zManager communicates with the Support Element (SE) of each zEnterprise node to manage various functions, including network virtualization management. The SE of each node provides the necessary management controls for System z and any zBX attached to it. An independent 1 GbE management network, called the *Intra Node Management Network* (INMN), is established to communicate with the various virtual servers and hypervisors in the node, as shown in Figure 1-2 on page 3. Communication on this network is restricted to zManager functions and cannot be directly used or accessed by customer applications, including management applications. You can use a Network Management Application that uses the zManager application programming interfaces (APIs) under secured conditions to access the INMN indirectly through the Hardware Management Console (HMC).

The INMN is not affected by connectivity options that are described in this paper for platforms outside the ensemble that access the virtual servers on the IEDN inside the zBX, and the rest of the ensemble on the CPC.

Figure 1-2 on page 3 depicts a single zEnterprise node composed of a z196 CPC and a zBX. zManager performs its management functions over the INMN, and virtual servers in the ensemble communicate with each other over the IEDN.

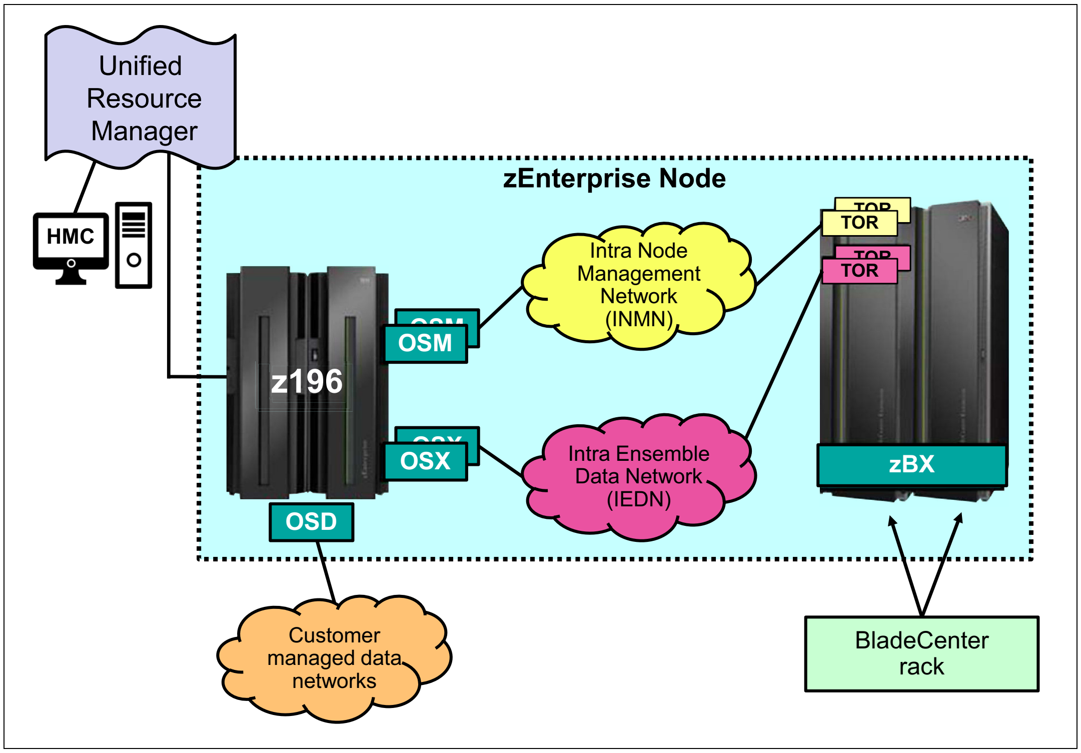
Unified Resource Manager
HMC
zEnterprise Node
z196
OSM
OSM
OSX
OSX
OSD
Intra Node Management Network (INMN)
Intra Ensemble Data Network (IEDN)
TOR
TOR
TOR
TOR
zBX
Customer managed data networks
BladeCenter rack

Figure 1-2 zEnterprise networks

Hosts residing on the System z CPC access the IEDN through the configuration of a 10 GbE OSA card as channel type OSX. To physically form the IEDN, Ethernet cables connect the OSX cards to a set of pre-specified ports in the 10 GbE TOR switch inside the zBX. Up to eight such connections from OSX cards in System z CPCs are supported.

Hosts residing in the zBX access to IEDN through a 10 GbE network interface card (NIC) on each blade, a pair of redundant 10 GbE High Speed Switches (HSSs) in each chassis, and a pair of redundant 10 GbE TOR switches in the first zBX rack, as shown in Figure 1-3 on page 4. You can read a more detailed description of how these redundant network components are interconnected in 2.1.2, “High availability design at the node level” on page 15.

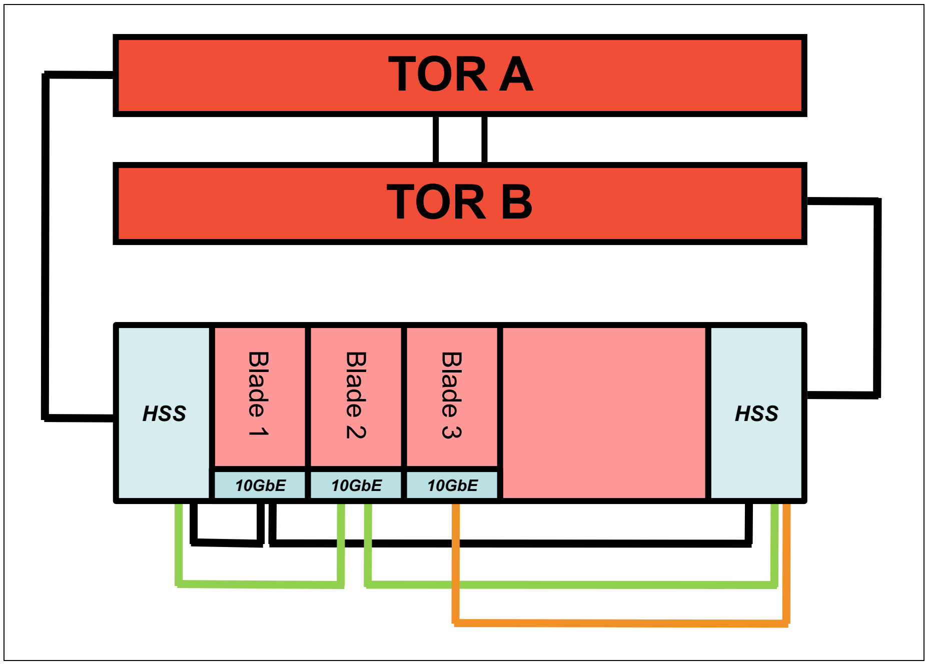

Figure 1-3 zBX blade networking

In an ensemble with multiple CPCs with zBXs, the IEDN across these nodes is formed by connecting the TOR switches of the zBXs in a tree structure, as shown in Figure 1-4 on page 5. You must connect these cables point-to-point, but you can insert a System z qualified dense wavelength division multiplexing (DWDM) device in the path for distance extension.[1]

[1] Registered users can visit the IBM Resource Link® library for current information about qualified WDM vendor products:
https://www.ibm.com/servers/resourcelink/lib03020.nsf/pages/systemzQualifiedWdmProductsForGdpsSolutionDocument&pathID=

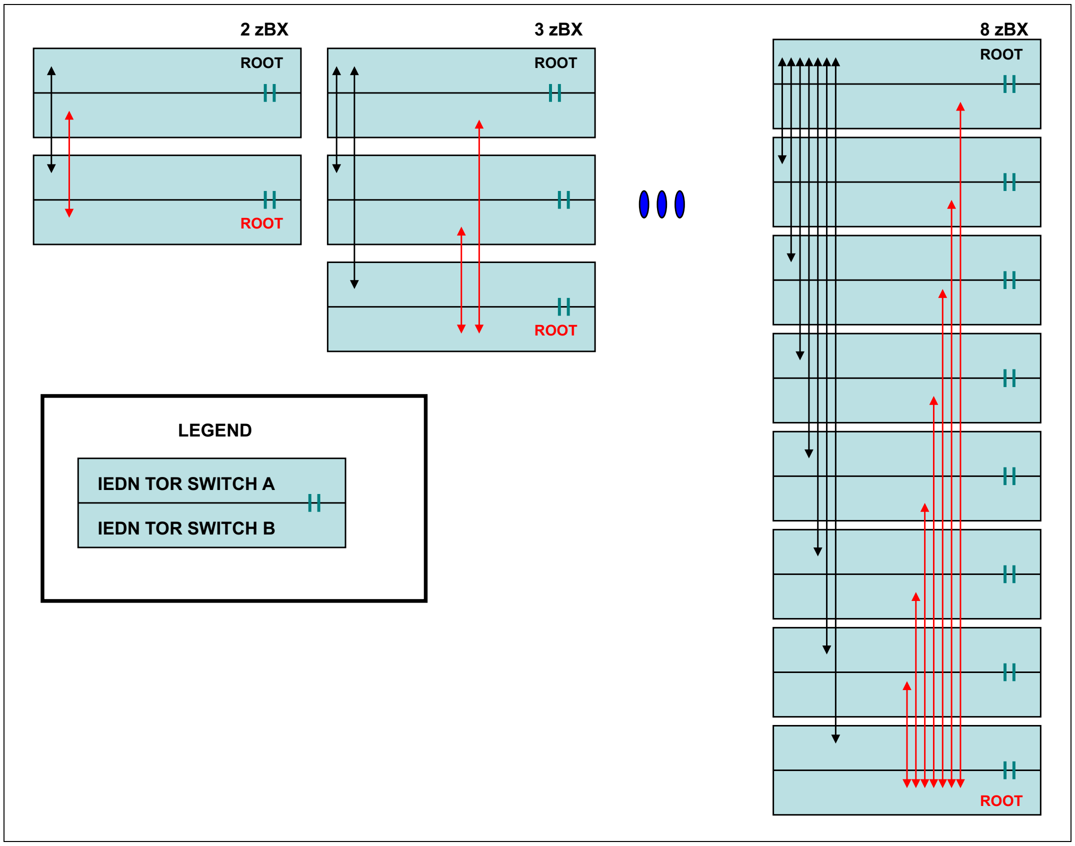

Figure 1-4 Tree structure for connecting multiple CPC/zBX nodes

1.2 Communicating with IEDN members

Ensemble configuration provides a topology that you can use to maintain an entirely isolated IEDN, or to interconnect the IEDN to the external network.

The IEDN is an entirely isolated Layer-2 network when you limit communication of hosts that reside in the zBX only to other hosts directly reachable over the IEDN. That is, the virtual servers in the zBX are designed to communicate only with other virtual servers in the zBX and with other virtual servers in CPC logical partition (LPAR) members of the ensemble. Thus, the isolated Layer-2 network of the IEDN has no need to share networking hardware, such as NICs, switches, cables, with other data center Layer-2 networks.

To communicate between the IEDN and external networks, for example, for virtual servers and their host operating systems, in the zBX systems that need to communicate with other virtual servers that are not hosted in the ensemble, you can use the following options:

- Connect to the IEDN through a zEnterprise member of the ensemble that resides both on the customer network and on the IEDN, such as LPAR.
- Reach the IEDN by connecting directly to virtual servers behind the TORs in the zBX.

1.2.1 Connecting to an IEDN over an LPAR

Figure 1-5 shows the first option in a scenario where a IBM z/OS® image connects through an OSA (OSD) port to the external network and through OSX ports to the ensemble's IEDN. The z/OS image resides in both networks. Figure 1-5 illustrates Client A, which resides on the customer data center network, communicating with VS1 that resides on another LPAR in the ensemble. The figure also shows how Client A communicates with a server on a blade in Blade Chassis 1 (BC1) in the zBX.

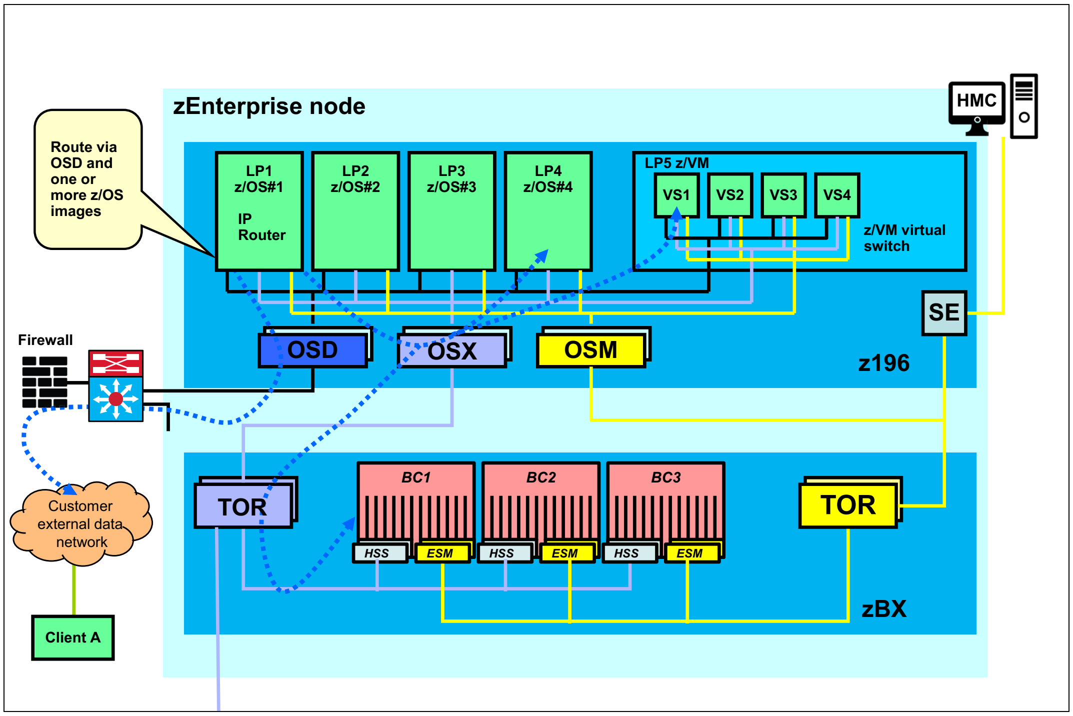

Figure 1-5 Communicating through an LPAR

Client A's route to the ensemble leads through a firewall/router next to an LPAR (zOS#1) that has interfaces in both the external network and the IEDN. Client A's communication path enters zOS#1 over a LAN interface on the OSD channel and is then routed by the TCP/IP stack in zOS#1 over the second LAN interface, on the OSX channel, into the IEDN. There, the associated VLAN ID connects the path to VS1 on an ensemble LPAR and to a server on a blade in the zBX. It is a requirement to use VLANs to communicate between virtual servers across the IEDN.

Figure 1-6 on page 7 shows in more detail how VLAN isolation is achieved across these LAN interfaces by using a distinct VLAN ID (#10) in the customer's external data network, which is separated from the IEDN VLAN ID (#22).

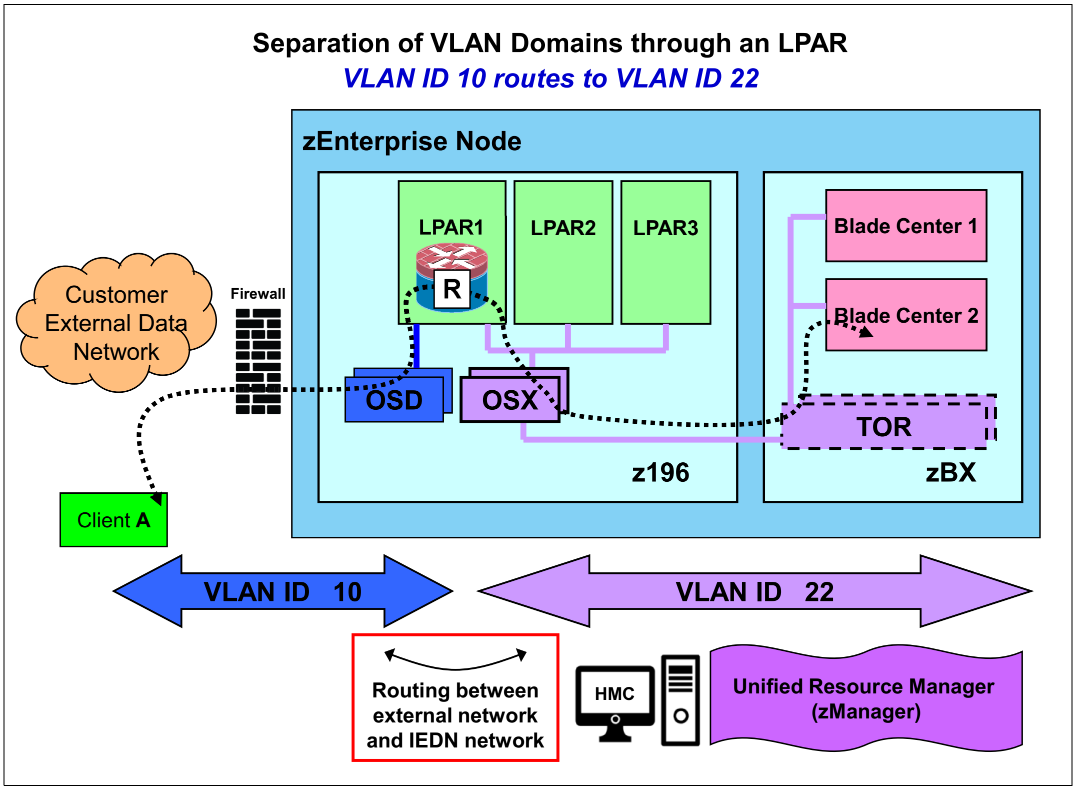
Separation of VLAN Domains through an LPAR
VLAN ID 10 routes to VLAN ID 22
zEnterprise Node
LPAR1
LPAR2
LPAR3
R
Blade Center 1
Blade Center 2
Customer External Data Network
Firewall
OSD
OSX
TOR
z196
zBX
Client A
VLAN ID 10
VLAN ID 22
Routing between external network and IEDN network
HMC
Unified Resource Manager (zManager)

Figure 1-6 Routing between external VLAN ID and IEDN VLAN ID

LPAR1's routing table routes Client A's traffic in to and out of the IEDN. The traffic from BladeCenter 2 is routed through the IEDN to LPAR1, which owns the gateway address for the subnet on VLAN22, where it is processed for further routing. In this case, there is still physical isolation between data center Layer-2 networks and the IEDN, as unique networking hardware (NICs, cables, and switches) is traversed over each network. Thus, the VLAN Domains are segregated from each other. This type of connectivity with an LPAR on zEnterprise is well-known to zEnterprise users and requires no further explanation. However, the second type of connectivity, which is shown in Figure 1-7 on page 8, represents a new type of interconnection over a zBX and thus merits a detailed examination.

1.2.2 Connecting to the IEDN over zBX TORs

Figure 1-7 on page 8 illustrates external network connections that reach the IEDN by connecting through the external ports of the TOR switches. The figure shows Client B communicating with z/OS#2 in an ensemble LPAR and a server that resides on a blade in the zBX. Client B's route leads over a node (router) next to the zBX in the IEDN. Client B’s communication path leads over the external ports of the IEDN TOR switches to reach z/OS#2 in the System z CPC or to reach a server in the zBX.

The zBX connection that is shown in Figure 1-7 on page 8 merits more clarification than the LPAR connection shown in Figure 1-5 on page 6 and Figure 1-6. Although the zBX connection looks like other connections to switches in a BladeCenter, it is subject to a set of rules that differentiate zBX connections from connections to standard network equipment.

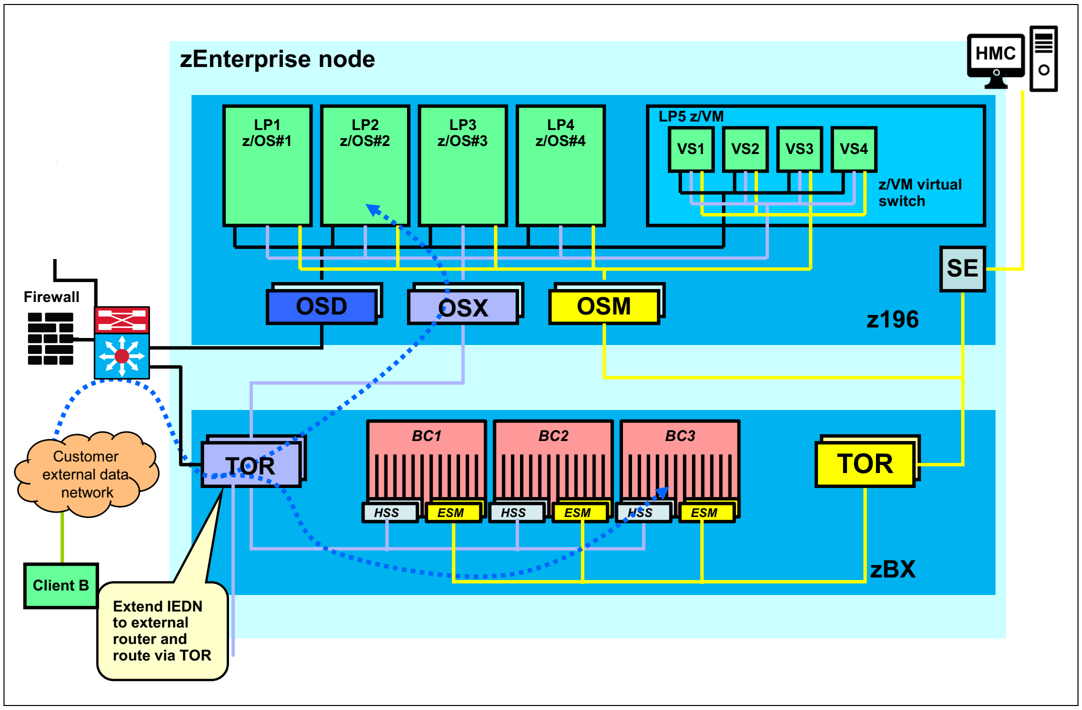

Figure 1-7 Routed (Layer-3) communication path through external TOR ports of the zBX

When a virtual server in a zBX that resides on BladeCenter #3 (BC3) in Figure 1-7 attempts to communicate with a virtual server that resides on a host that is outside the ensemble, such as Client B, the traffic traverses the 10 GbE NIC that is attached to the BladeCenter. The same single 10 GbE NIC attaches to the IEDN over the redundant High Speed Switch Modules or Switches (HSM or HSS), and then connects to the redundant IEDN TORs. The traffic on the IEDN TORs is sent either over the internal IEDN ports to reach an IEDN target, or sent over the external, non-IEDN ports that face the external customer network because these routers contain the gateway for the subnet on the virtual servers. So other than the specific set of TOR ports that are reserved for external facing communications, the same networking hardware is traversed in the zBX for both IEDN and non-IEDN communications.

This lack of ability to physically isolate the IEDN from other data center Layer-2 networks in the zBX is one reason why Layer-3 routing to enter the TOR switch of the zBX has been required in the past. The IEDN Layer-2 was terminated at the attached Layer-3 router, and joined with data center Layer-2 networks in that Layer-3 equipment. This maintained the physical distinction and isolation between the IEDN and data center Layer-2 networks. In addition, the complex interaction and exchange of intra-switch messages and protocols makes maintaining a highly available IEDN difficult. The data center network configuration influences the IEDN's high availability settings and characteristics.

In Figure 1-7, the VLANs and Layer 2 domains used in the IEDN remain segregated from the external customer network and any of its VLANs, that is, Layer-2, domains. This means that the IEDN VLAN domains are separated from the external network VLAN domains: Communication across the two types of domains requires Layer-3 routing. A firewall is inserted between the external router and the TORs to ensure isolation of the networks through permit and deny operations.

Figure 1-8 shows in greater detail the separation of the external and IEDN VLAN domains when Client B connects to IEDN servers over the TORs of the zBX.

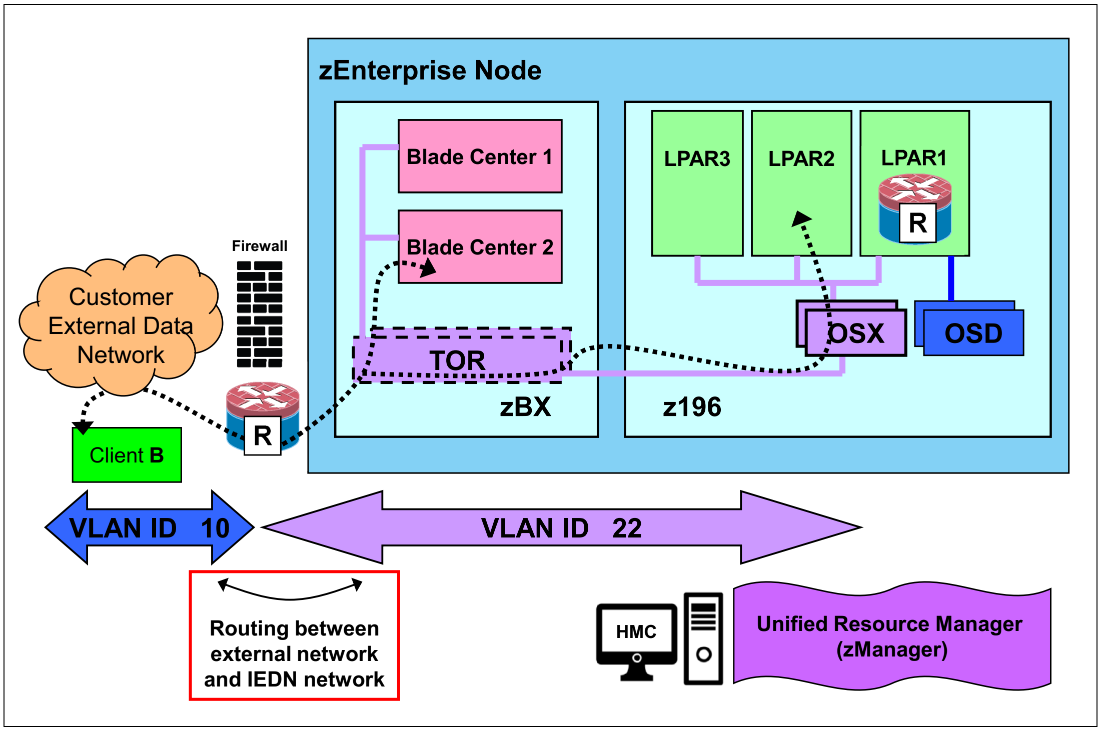

Figure 1-8 Separating VLAN domains through external TOR ports of the zBX

When an ensemble LPAR is used to interconnect the external network to the IEDN network, as Figure 1-5 on page 6 and Figure 1-6 on page 7, the connection is ROUTED through the LPAR. When, as in Figure 1-7 on page 8 and Figure 1-8, the adjacent node connected to the zBX external ports provides a communication path to virtual servers in the ensemble, this connection is also ROUTED. That is, the adjacent node provides a ROUTED or Layer-3 connection point. Communication with partners who are not next to the zBX must be routed by the adjacent node into the customer's external network. Routing between VLAN segments maintains separation of the networks.

The ensemble was originally designed to maintain this separation of the IEDN from external customer networks. Thus, a node next to the IEDN was required to present a ROUTED endpoint, with the IEDN VLAN terminating in the node next to the IEDN. As you saw in Figure 1-5 on page 6 and Figure 1-6 on page 7, the adjacent node might have been an LPAR with connections to both the IEDN and to the outside world. Alternatively, as you saw in Figure 1-7 on page 8 and Figure 1-8, the adjacent node might have been any node (router) next to the zBX that was capable of terminating the IEDN VLAN ID, and then routing out into the external network. The design where the IEDN VLANs remain in a private network is enforced through centralized zManager and TOR controls.

1.2.3 Separate network management boundaries

Using routing, as opposed to switching, also establishes a clear network management and administration boundary, in which existing data center network management tooling covers the data center Layer-2 networks, and zManager fully assumes those responsibilities for the IEDN. In the closed IEDN network, zManager takes responsibility for configuring, monitoring, and managing the networks, and also for:

- Ensuring that no vMAC or VLAN collisions can occur.

  zManager controls all dynamic MAC address generation by assigning a MAC address prefix to all hypervisors and virtual switches. OSX is also considered a virtual switch of the IBM PR/SM™ hypervisor. This central configuration approach eliminates MAC address conflicts and unauthorized virtual MAC generation.
- Isolating the Layer-2 by preventing TOR switches from processing switch protocol messages, for example, STP sent by externally connected nodes.
- Ensuring that virtual servers can communicate with each other from both physical and logical configuration points of view, but only when authorized to do so.
- Providing for high availability in the Layer-2 network.
- Assuming network diagnostic responsibilities across the IEDN. Call IBM support if network communication paths are not working inside the IEDN.

zManager is no longer the sole authority for the entire Layer-2 domain if the IEDN is no longer an isolated Layer-2 network. Two distinct management entities must be coordinated to ensure that consistent configuration, security, and reliability, availability, and serviceability (RAS) are provided. The advantages of the functions performed by zManager alone, as described above, are compromised.

Note: When you enable Layer-2 connectivity to zBX, you may enable many combinations of Layer-2 functions across multiple vendors on the ports, some of which may be incompatible with the settings used inside zBX switches. While it is impossible to test all possible combinations, we feel that the guidelines in this paper will help you develop a configuration that will be compatible with zBX design and that will work for your data center.

1.2.4 Maintaining security over the path to IEDN

If you permit communication between the external customer data network and the ensemble, you will probably want to keep this path secure. Before you decide how to implement this communication path, you should first understand the security provisions built into the IEDN design.

The ensemble contains its own enforcement points in the TOR, for external connections on the ingress/egress ports, and also in the hypervisors. All virtual servers and VLANs must pass through the enforcement points, the access points of hypervisors and TORs, where their authorization is confirmed. The access points contain security enforcement that has been defined through zManager. However, you can also continue to implement other security services in the ensemble by using traditional security mechanisms that are outside the control of zManager, such as IP Filtering (a firewall function), encryption, access control lists, user ID, and password authentication.

Partly for security purposes, a Layer-2 connection from the external network to the TOR has not been supported in the past. It was formerly an IBM support requirement to enforce a Layer-3 termination point on any node next to the external ports of the IEDN TORs, as shown in Figure 1-7 on page 8 and Figure 1-8 on page 9. This Layer-3 Routing function would most typically be installed in a dedicated routing platform (router) but could be in any node capable of terminating a connection in Layer-3 mode. The external Layer-3, or Layer-2/Layer-3, platform could not bridge, that is switch, into the external network but was required to ROUTE to it. This requirement for a Layer-3 termination point (a routing termination point) avoids VLAN ID collisions by ensuring that the IEDN VLAN IDs are not merged with the external customer network VLAN IDs, and allows the IEDN security provisions described above to provide strict access controls for the entire IEDN.

Note: When connectivity requirements are eased to allow Layer-2 switched identity connections, you still cannot exchange Bridge Protocol Data Units (BPDUs) at Layer-2 between an external Layer-2 switch and the IEDN TOR.

STP or RSTP messages (BPDUs) received from external switches are filtered out at the TOR to avoid network topology or STP topology changes. Other BPDUs, like those for VLAN registration protocols, are also filtered out so that only zManager can affect VLAN IDs permitted in the IEDN. This TOR filtering protects the security of the IEDN network and avoids network topology changes that could affect the high availability configuration of switches inside the zBX, which we describe in the next chapter.

2

High availability architecture of zBX switches and blades

This chapter introduces the IBM zEnterprise BladeCenter Extension (zBX) high availability configuration, including physical cabling and internal deterministic use of Rapid Spanning Tree Protocol (RSTP). We describe the implications this has for connecting to the zBX at Layer-2, including why connectivity to the zBX is not compatible with using Spanning Tree Protocol (STP) in data center networks.

The intraensemble data network (IEDN) was designed to allow out-of-the-box, highly available communication between virtual servers in an ensemble. All hardware components in the IEDN are duplicated to provide for full redundancy. The redundancy and high availability setup are described in this chapter, beginning at the blade and chassis level, moving to the ensemble node level, and then moving to the ensemble level composed of a maximum of eight nodes.

© Copyright IBM Corp. 2014. All rights reserved.

2.1 High availability design

High availability design occurs at the chassis, node, and ensemble levels and also in connecting to a data center network.

2.1.1 High availability design at the chassis level

Each zBX frame consists of two chassis, each with two 10 GbE switch modules plugged into it. They offer connectivity to two 10 Gbit Ethernet NICs on each blade. That is, each blade has two 10 Gb Ethernet interfaces, with each of them connected to one of the 10 GbE chassis internal switches.

The chassis internal switch modules provide uplink connectivity to the TOR switches. Chassis-to-TOR connectivity is described in the following section. The internal switches of the chassis are themselves interconnected by a redundant pair of Ethernet links. The links are aggregated together by using IEEE 802.3ad link aggregation; however, this connection between chassis switches is deactivated by default by RSTP based on intelligent selection of STP weights and costs. More details about the use of RSTP in the zBX are described in a later section.

The blade has a connection to each of the chassis switches, as shown in Figure 2-1. However, because the two blade network interface cards (NICs) are not connected to the same switch, no active/active link aggregation can be used. Therefore, the NICs are set up in a primary/backup configuration by the hypervisor. If the active NIC goes down, the backup NIC is activated automatically in a few milliseconds. The active/backup NIC group serves as an uplink to a virtual switch inside the hypervisor on each blade.

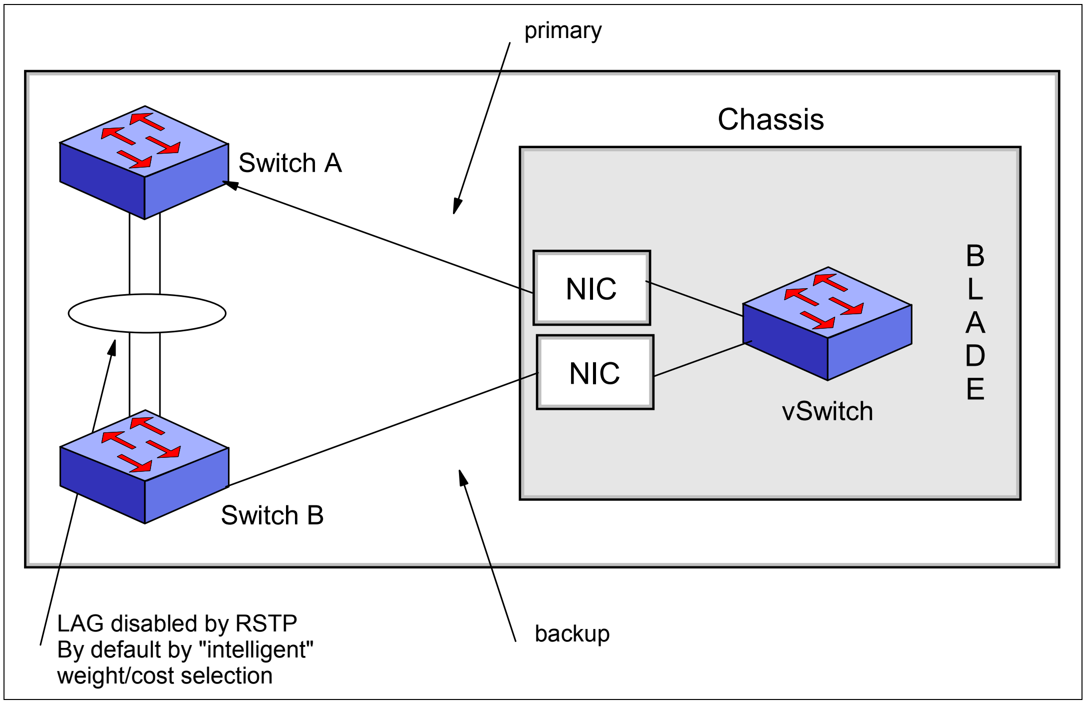

Figure 2-1 Chassis in a zBX frame

2.1.2 High availability design at the node level

Up to eight chassis can be grouped together to build a zBX in an ensemble node. A zBX has a pair of redundant TOR switches that offer the needed connectivity for all chassis. The chassis internal High Speed Switches (HSSs) are connected to the TOR switches, which themselves are interconnected by a pair of links. These links are set up using IEEE link aggregation and interconnect the two root switches of the zBX RSTP tree.

As shown in Figure 2-2, switch A of the chassis is connected by a pair of links to TOR A of the node while switch B of the chassis is connected to TOR B. The chassis-to-TOR switch links are aggregated with IEEE link aggregation, allowing for a 20 Gb uplink per chassis switch. Because the physical interconnections between chassis and TOR switches do not provide for a loop-free topology when all connections are enabled (for example, there would be a loop consisting of TOR A → chassis switch A → chassis Switch B → TOR B and back to TOR A), RSTP is used to resolve and deactivate the loop. By selecting fixed cost settings for each link, the RSTP algorithm always results in the link between chassis switch A and chassis switch B being chosen to be deactivated. This link is only activated if a TOR switch fails or if both links to that TOR switch from the chassis switch fail.

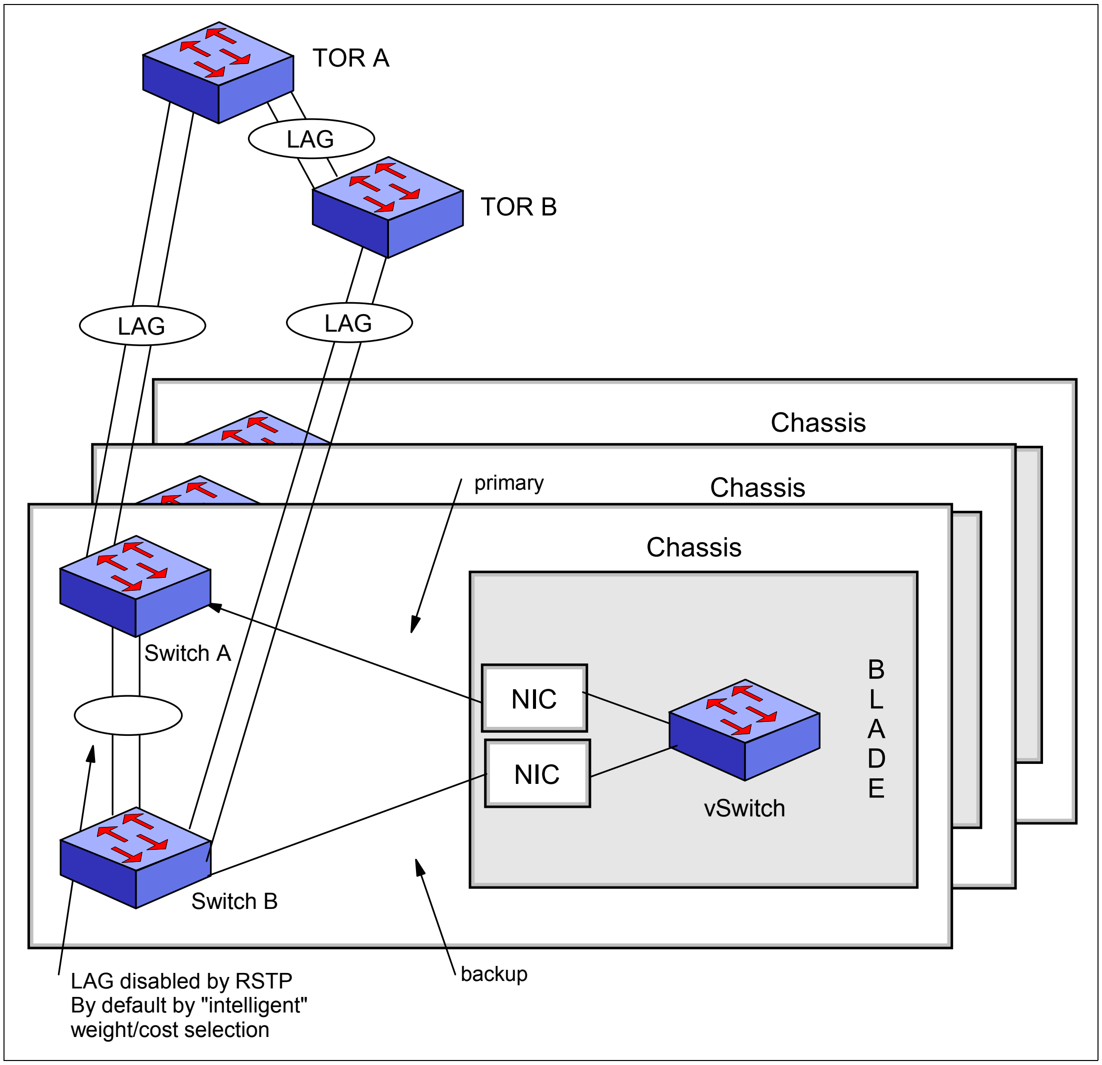

Figure 2-2 Link Aggregation Group operation in a zBX chassis

2.1.3 High availability design at the ensemble level

An ensemble consists of up to eight ensemble nodes. Each ensemble node consists of an IBM System z CPC and its optionally connected zBX. The zBX contains a pair of TOR switches and up to eight chassis.

The TOR switches of the zBX in each node are interconnected in order to group nodes containing zBXs to build an ensemble, as shown Figure 1-4 on page 5, which illustrates a tree structure. The first two installed zBXs are anchored as the first and last zBXs, and as additional zBXs are added to the ensemble they are inserted in between these two. Two spanning trees of interconnected Layer-2 switches are built to interconnect the zBXs. In the first zBX, TOR A is chosen as the switch serving as the connection point. In the last zBX, TOR B is chosen as the connection point. These chosen switches serve as root switches for two spanning trees. The two root switches are purposefully chosen for high availability purposes to be in independent zBXs because different zBXs are most likely connected to different power domains and may even be in different data centers.

A connection is made from TOR A in each of the zBXs to TOR A in the first zBX, while a similar connection is made from TOR B in each zBXs to TOR B in the last zBX, as shown in Figure 2-3. This connectivity establishes two spanning trees, depicted as independent red and green trees in the figure. Only the red tree is active while the green links are deactivated by RSTP, again achieved by establishing intelligent cost settings for the network segments. For any zBX in the ensemble, the path from TOR B to TOR A and then to the first node is always cheaper than the direct path from TOR B to the first node. The green tree only gets activated if node 1 fails completely.

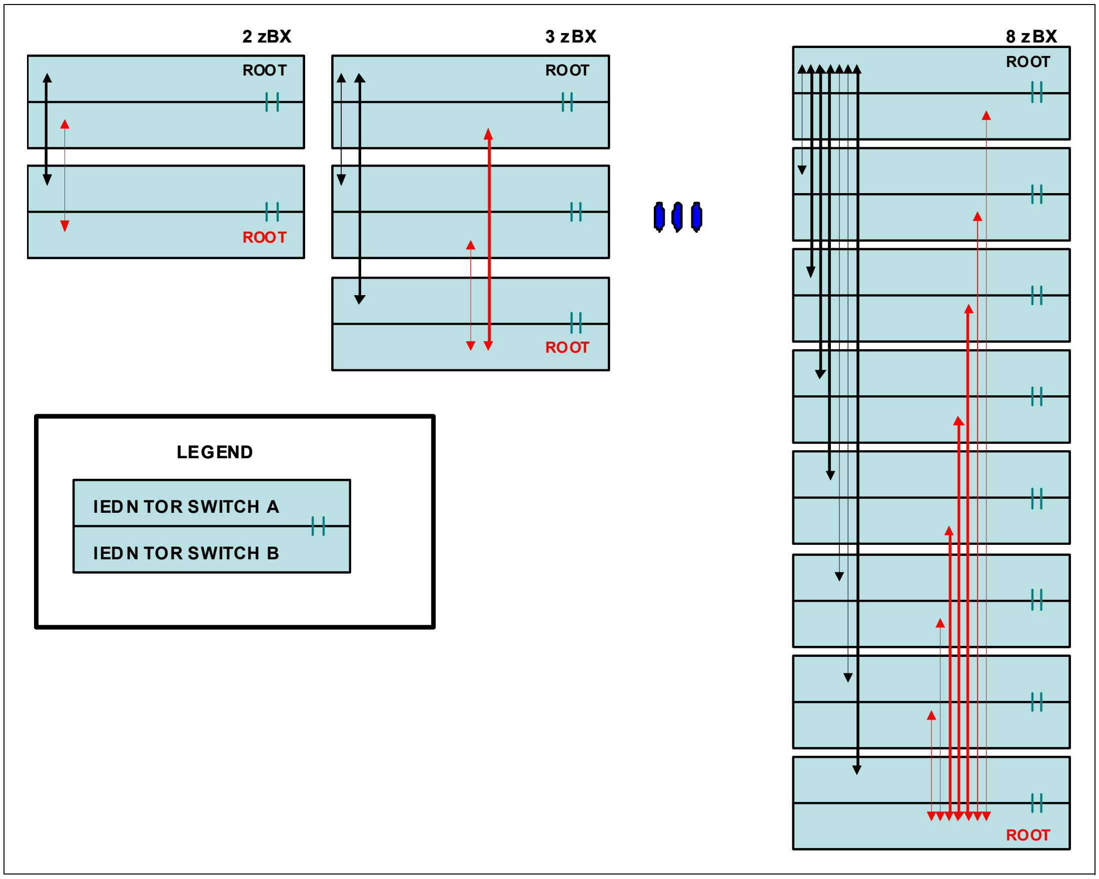

Figure 2-3 IEDN interconnection over eight zBXs

2.1.4 High availability design for connecting to a data center network

We noted that the IEDN is using RSTP to achieve and maintain high availability. A direct implication of this is that in order to maintain the HA characteristics of the IEDN, you need to prevent switch-interaction with the customer's network. Failure to do so could cause STP settings to interfere with root bridge and active/blocked link selection, and so on.

The following sections present several ways to connect the IEDN to the data center network without causing interference on switch management and at the intra-switch protocol level. The examples maintain the isolation of the two networks from a high availability perspective as well as from an error propagation perspective. Errors in the data center network do not affect the IEDN and vice versa. The highest level of isolation is provided if the two networks are separated by a router. However, switched connection solutions that do not rely on the use of the STP/RSTP protocol can also offer good or acceptable isolation.

The IEDN was configured to appear like a single switch to the user. It does not matter which node in an ensemble and which TOR switch in an ensemble is chosen to connect to the customer's networks. In fact, all are possible candidates. From a connectivity perspective, the whole IEDN appears as a single switch, which is not STP capable. Connections to different TORs on different zBXs appear as though all the ports were on a single dumb switch, a virtual hub that is not STP capable, nor capable of supporting other switch protocols such as link aggregation.

STP Buds are filtered at each TOR port connecting to the data center network, so RSTP and STP cannot be used for loop resolution in the merged network. A similar IEE protocol-Multiple Spanning Tree Protocol (MSTP) is also ruled out because it is incompatible with the zBX design. MSTP is backward compatible with STP and interacts with STP. Although MSTP provides load balancing to solve the same problems as are being solved with another protocol called PVST, MSTP does so by externalizing STP instead of encapsulating it. It uses the same BPDUs as STP and is therefore filtered by the Bridge Protocol Data Unit (BPDU) filter rules of the TOR. As a result, the zBX cannot participate in this connection to a node using MSTP.

2.2 The basics of high availability

High Availability in an IT infrastructure requires multiple factors, such as the following:

- Hardware and software redundancy

  If one component fails, another component can assume the failing component’s role.

- Conscientious implementation of management techniques to tune subsystems

  Well-tuned or self-tuning subsystems and networks.

- Security to prevent failure of access to required hardware or software

  Introducing security technologies for identification, access control, and so on.

- Rapid recovery or bypass of failed functions

  - Dynamic rerouting around failed components.
  - Hot Standby components such as routers, adapters, switches, and so on.

- Provisioning of target servers to present a single system image, and so on

2.2.1 Basic high availability with physical redundancy for routers and switches

When you want to back up failing components related to a router or a switch, you need to think in terms of first protecting against a physical link failure on the platform and second, protecting against an entire platform failure.

Connecting to a switch

The most basic switch connection to the IEDN is to a single switch in the data center network. For redundancy reasons two links are used to connect this switch to the IEDN.

One link from the data center switch is connected to TOR B in the first zBX, while the second link is connected to TOR B in another zBX, as shown in Figure 2-4 on page 19. This creates a network topology loop because the IEDN can be viewed as a single switch. You can use several methods to resolve this loop:

- If you have a Layer-2/Layer-3 (L2/L3) switch and want to build redundancy with two links on the same switch, you can use a Layer-2 technology that offers high availability for an IP address, such as:
  - IBM System Networking (BNT®) HotLinks
  - Cisco FlexLinks
  - Brocade Protected Link Group
  - Juniper Redundant Trunk Group

  In either case, the two links are backed up by placing one in *Active* mode and the second link in *Standby*. The value of these settings for attaching to the TOR is that they do not rely on STP, which the TOR would not support. These failover technologies minimize disruption to the network by protecting critical links from loss of data and power. With HotLink, FlexLink, Protected Link Group, or Redundant Trunk Group one port in the group acts as the primary or active link, and the other port acts as the secondary or standby link. The active link carries the traffic. If the active link goes down, one of the standby links takes over.

- Another method to resolve the loop is to use an STP version, which works on a per VLAN basis, as opposed to on a link basis (STP/RSTP) or on the basis of a group of VLANs (MSTP). Per VLAN implementation of STP is compatible with the zBX external TOR port implementation, whereas the STP versions that rely on the latter types, such as link and VLAN group, are not.

  Unfortunately, there is no standardized way to implement STP versions that work on a per VLAN basis. Each vendor can choose its own name and even a slightly different way of implementing this type of technology. For example, Cisco uses Per VLAN Spanning Tree (PVST) and PVST+ while other vendors use Simple Loop Prevention Protocol (SLPP) or other similar technology.

  The crucial aspect of such loop resolution protocols is that they work on top of VLANs, encapsulating STP BPDUs, so that they look like regular traffic for the IEDN. The whole IEDN is transparent for such types of traffic. This means that when the data center switch, as depicted in Figure 2-4 on page 19, issues these protocol packets on top of each VLAN, the packets traverse the IEDN and end up back at the very same switch from which they originated. In response to this loop detection, the data center switch would not deactivate an entire link to the IEDN, but rather only a path for an individual VLAN. It may be a different link that is deactivated per VLAN; therefore, a data center switch may block a port for a particular VLAN and then block a different port for an entirely different VLAN.

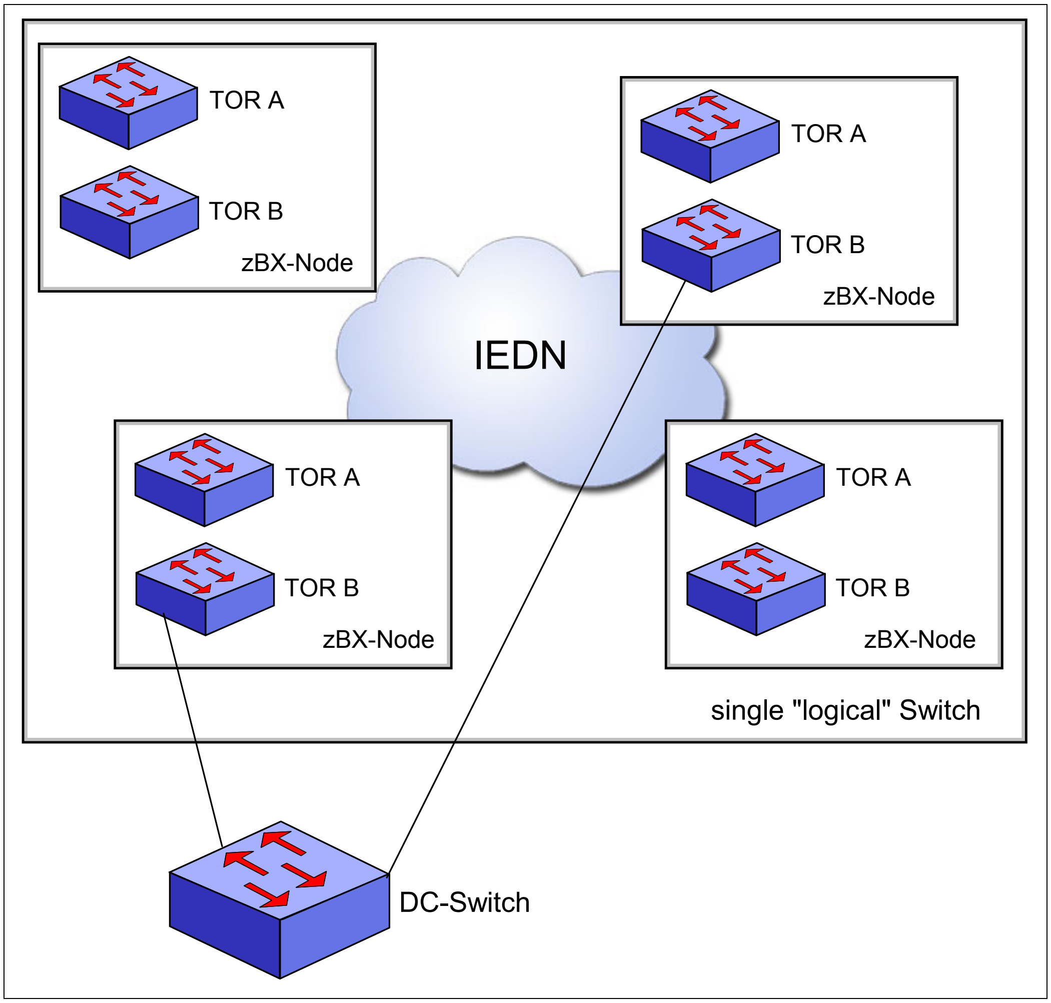

Figure 2-4 IEDN TOR switches in all zBXs simulate a single logical switch

Of course, you can set up more than a single pair of links in a primary and backup configuration using Hotlinks, as long as each pair of links has a distinct set of VLANS, that is, no overlap of VLANs across the multiple pairs is permissible. Alternatively, if VLANS do overlap between the links, you can use PVST/SLPP, allowing for more than a single active 10 Gb connection to the IEDN.

Note: The diagrams that follow show multiple zBXs with their TORs connected together. This is done to illustrate the fact that whether a single zBX or multiple zBXs are deployed, the IEDN connects to the data center network as a single cloud. The network environment and associated issues are not more complex in an environment with eight zBXs and their 16 TOR switches, nor are they any simpler in an environment where a single zBX is deployed with its single pair of TOR switches.

Connecting to a switch stack

The process for connecting to a switch stack is similar to the one for connecting to a single switch with the exception that links from the zBX TORs connect to a switch stack instead of a single data center switch. A switch stack consists of several physically distinct interconnected switches that behave like a single switch. Figure 2-5 shows how the first link goes to the top switch in the stack while the second link goes to the second link in the stack.

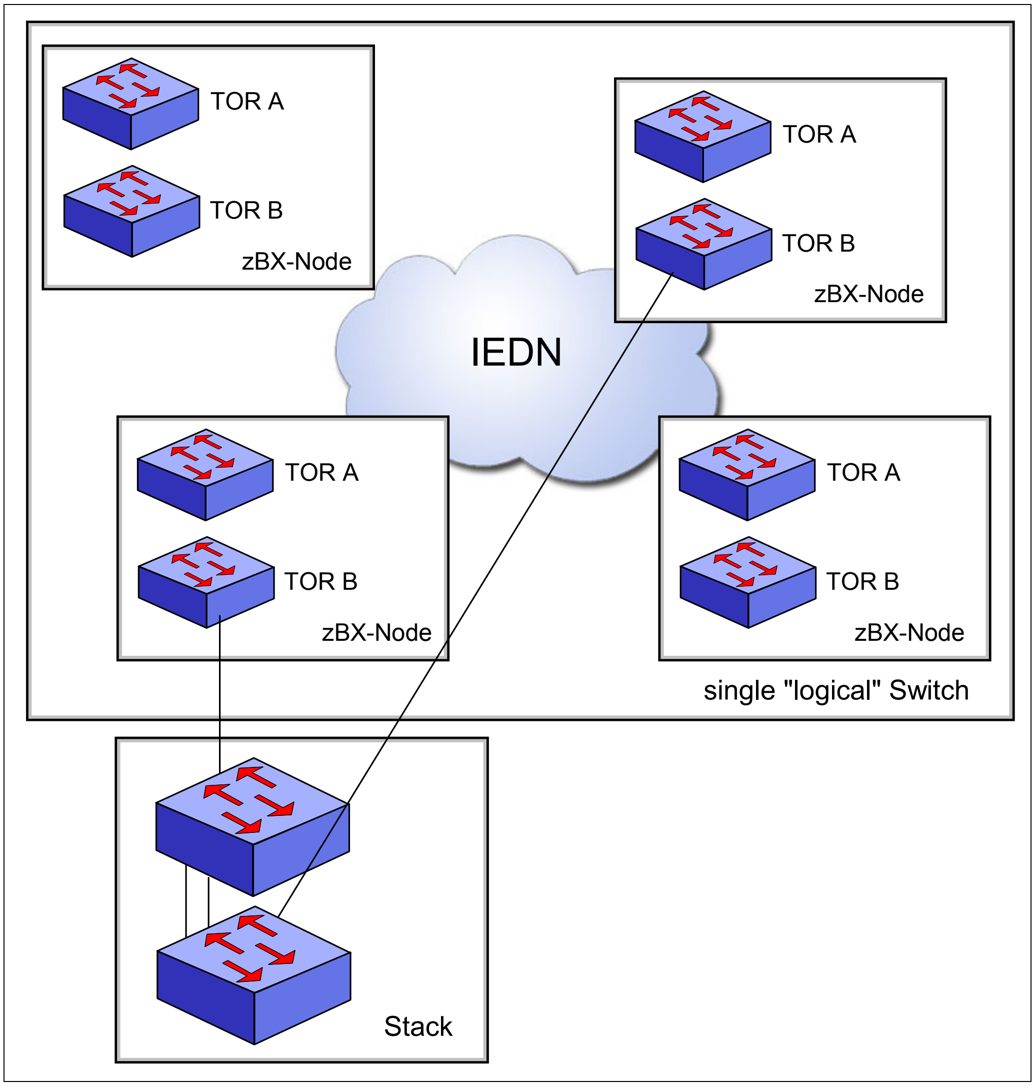

Figure 2-5 zBX logical switch connecting to a switch stack

As in the previous single data center switch example, one link from the switch stack should be connected to one of the TOR switches in one of the zBXs, and a second link should connect to a different TOR (in a different zBX if multiple zBXs are deployed). However, in this example, using two links of two physically distinct switches in the data center switch stack eliminates a single point of failure, and also achieves high availability to protect against link errors as well as data center switch errors.

Connecting to multiple switches

You can also use multiple switches to connect to the IEDN even without using a switch stack, as shown in Figure 2-6.

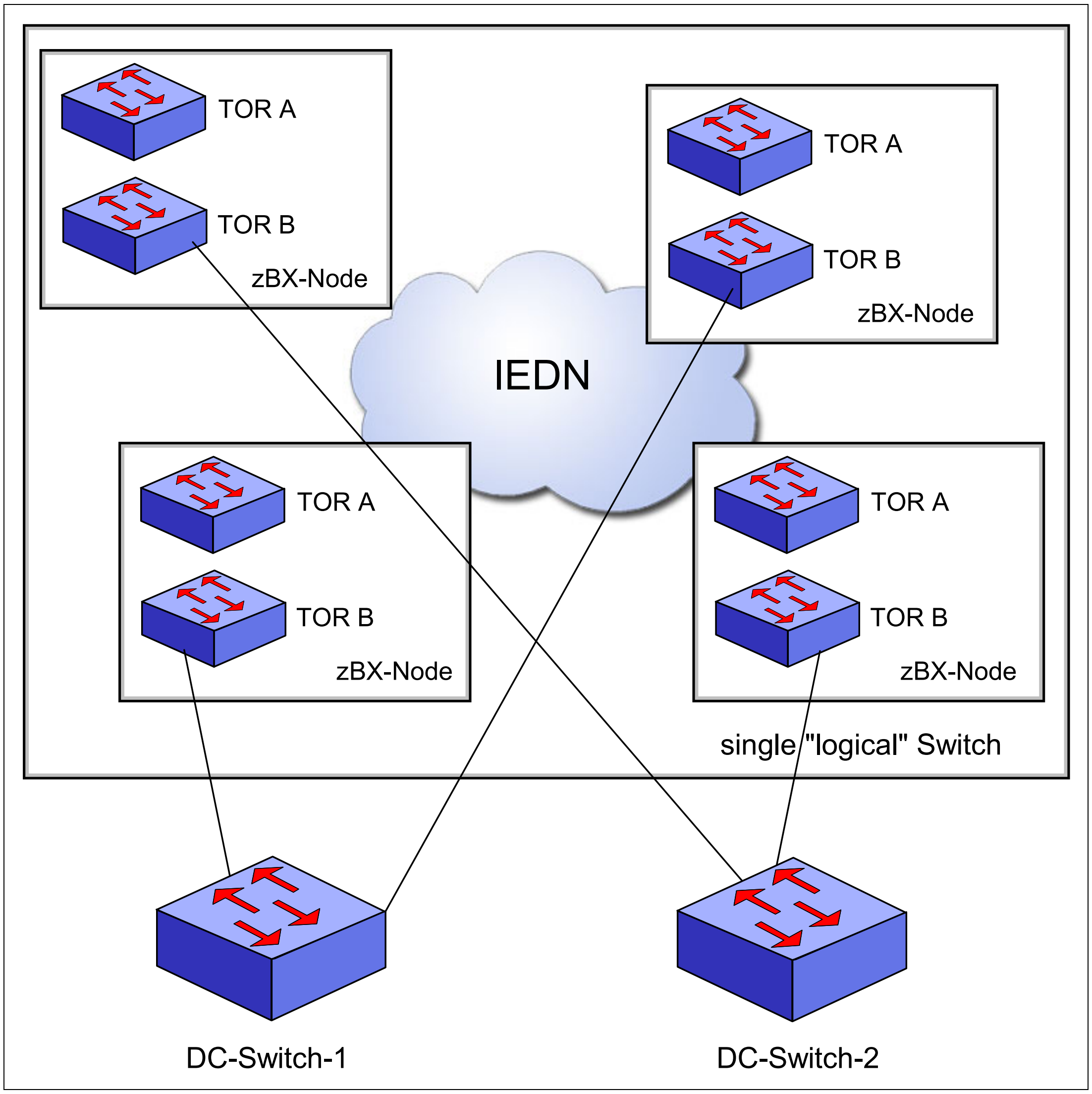

Figure 2-6 zBX logical switch connecting to multiple switches

Because the Hotlinks, or equivalent technology described above works only on a single physical switch or across a switch stack, it cannot be used in this case to deactivate links that would lead to a loop. Because this case operates across distinct physical switches, a per-VLAN loop resolution protocol, such as PVST/SLPP, is the only option in this setup. You can ensure that the link to the IEDN is deactivated instead of the link between the data center switches by carefully selecting link costs.

Connecting to routers: A preferred practice

Figure 2-7 illustrates the recommended solution and the continued best practice. Here the IEDN is connected to a pair of routers. The routers are set up for redundancy using a protocol such as Virtual Router Redundancy Protocol (VRRP) or Hot Standby Router Protocol (HSRP) that provides for dynamic failover capability to the backup router in the event of link or router failures in the primary path. The heartbeat used by the two routers as part of the HSRP/VRRP protocol traverses the IEDN, including potentially multiple zBX nodes (TOR switch pairs).

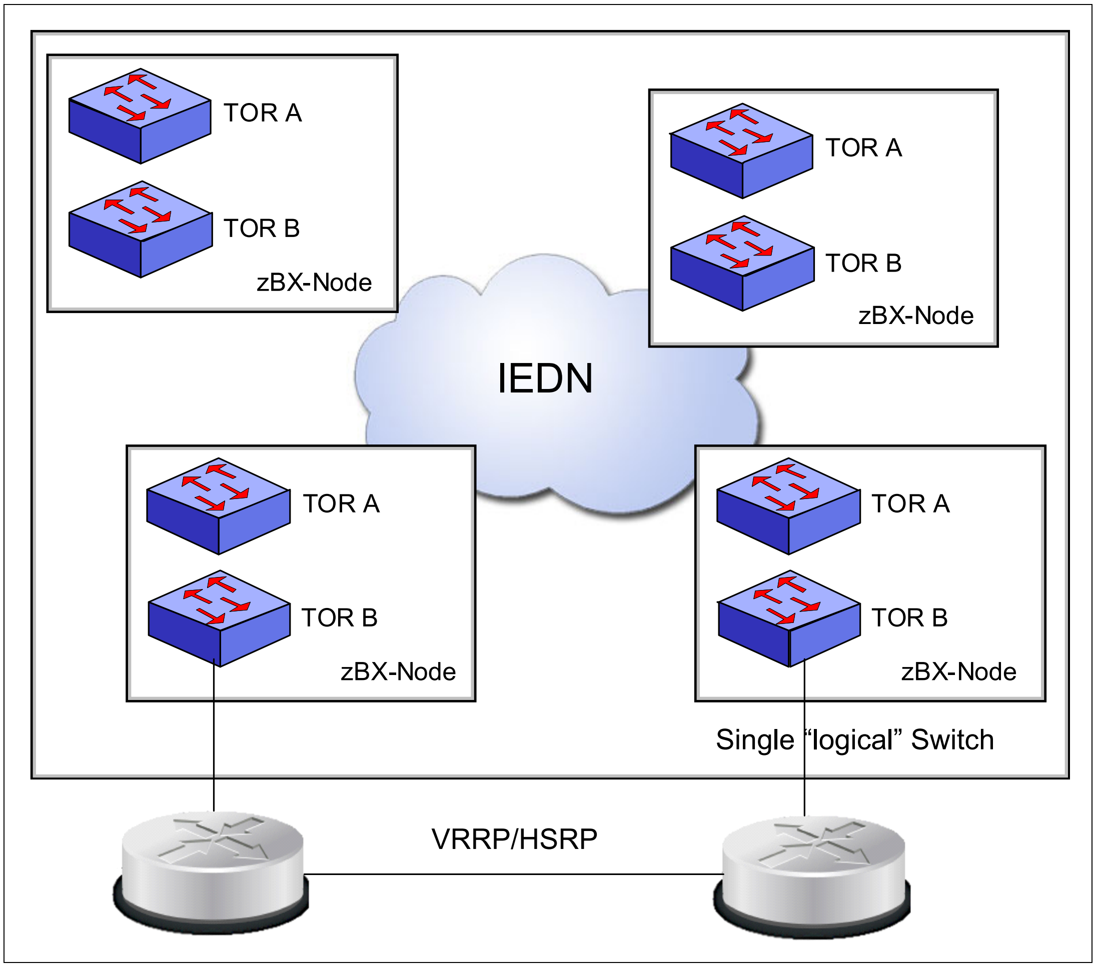

Figure 2-7 zBX logical switch connecting to an external router pair

Connecting to routers with multiple links

You can combine some of the redundancy features described above for greater and more focused failure protection. In Figure 2-8, each router is connected to two TOR switches in the IEDN. The two links from a single router can be combined into an active/backup configuration using Hotlinks technology, with one link from each router being deactivated but available for failover in the case of failure of the active link. This reduces the need for the more extensive router failover action to only be started in the case of a router error.

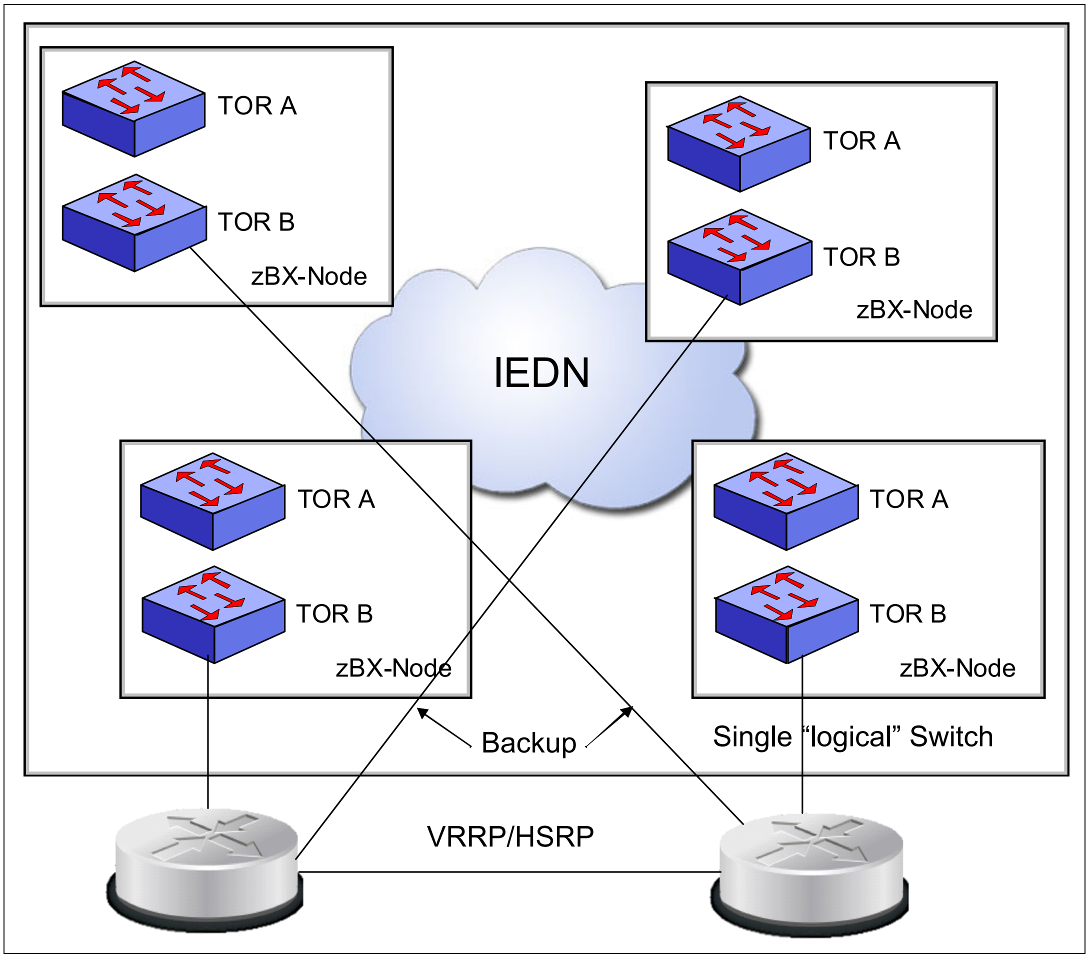

Figure 2-8 zBX logical switch connecting to routers using HotLinks and VRRP/HSRP

If you use a switch that provides some level of Layer-2 and Layer-3 functionality, you can also use the switch stacking function to form a single logical stacked router, as depicted in Figure 2-9. If the stacked router is combined with Hotlinks technology, a router failure would not start HSRP or VRRP, but would rather use the switch stacking recovery mechanisms.

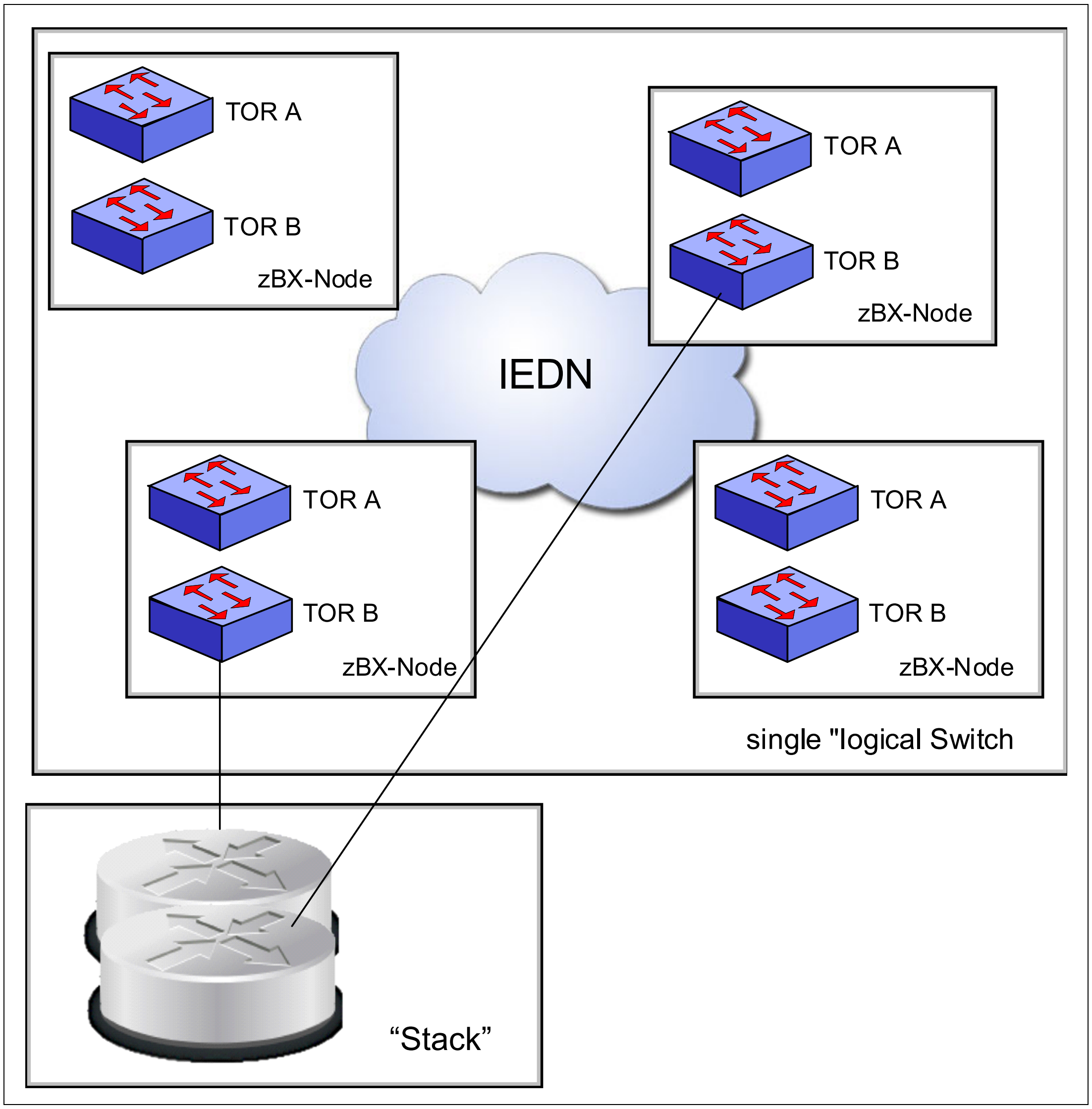

Figure 2-9   zBX logical switch connecting to stacked switches using HotLinks

3

Layer-3 configuration options

Earlier, we discussed supported connectivity options that were previously available between the customer data center core network and the IBM zEnterprise BladeCenter Extension (zBX). The two options described in Figure 2-5 on page 20 through Figure 2-8 on page 23, enforce the isolation of the intraensemble data network (IEDN) VLAN domains from the VLAN domains in the external data network using Layer-3 routing. Both the IEDN endpoint in a logical partition (LPAR) and the endpoint connected to the externally facing ports of the zBX terminate the IEDN Layer-2 network and forward packets to and from non-IEDN destinations on Layer-3, that is, using routing. This connectivity approach makes the zManager the centralized point of control for protecting the IEDN.

In this chapter, we provide more detailed illustrations and descriptions of Layer-3 routed endpoint options using multiple router platforms. These options show you how to adhere to IEDN architecture security constraints while, at the same time, achieve successful connections to the zBX TORs and provide high availability solutions.

> **Note:** These examples are not intended to provide the exclusive set of supported options; many other possible variations can exist. Syntax of sample coding statements uses the platform's newer version of its Industry Standard Command Line Interface (ISCLI) initially available in code level 7.6.1. Differences in syntax and functional capabilities are likely to exist when using other code levels or other vendor platforms. The sample coding statements that are provided are not intended to show all possible configurations, but rather to highlight examples of specific CLI statements.

We have already outlined the rules for connecting the external customer network to the IEDN through the external ports of the zBX, but they are worth repeating here:

- Loop detection and elimination protocols that require the use of IEEE standard Bridge Protocol Data Unit (BPDU) packets to detect and eliminate switching loops are not permitted on the links between customer networking equipment and the zBX top-of-rack (TOR) switches. In other words, Spanning Tree Protocol (STP), Rapid Spanning Tree Protocol (RSTP), Multiple Spanning Tree Protocol (MSTP), and so on, are not permitted.
- Loop detection and elimination protocols that do not rely on IEEE standard BPCU packets, such as Per VLAN Spanning Tree+ (PVST+), Simple Loop Prevention Protocol (SLPP), are permitted.

© Copyright IBM Corp. 2014. All rights reserved.

- The allowable IEDN VLAN range is limited to VLANs 10 - 1030.

We continue now to illustrate and describe configurations for three variations of supported connections that terminate the IEDN at the endpoint and use Layer-3 routing between the external network and the zBX.

3.1 Option 1: Dual IBM RackSwitch G8052s: Routed endpoint with VRRP

You can implement a routed endpoint with either a Layer-3 (routed) definition or a Layer-2 (switched) definition, or both types of definitions depending on the additional capabilities that might be required. Figure 3-1 shows how the G8052 platforms next to the IEDN VLAN and the zBX on the right are routing (R) into the core network on the left using a Layer-2 definition, a virtual routed interface in the switch that presents as a routed endpoint.

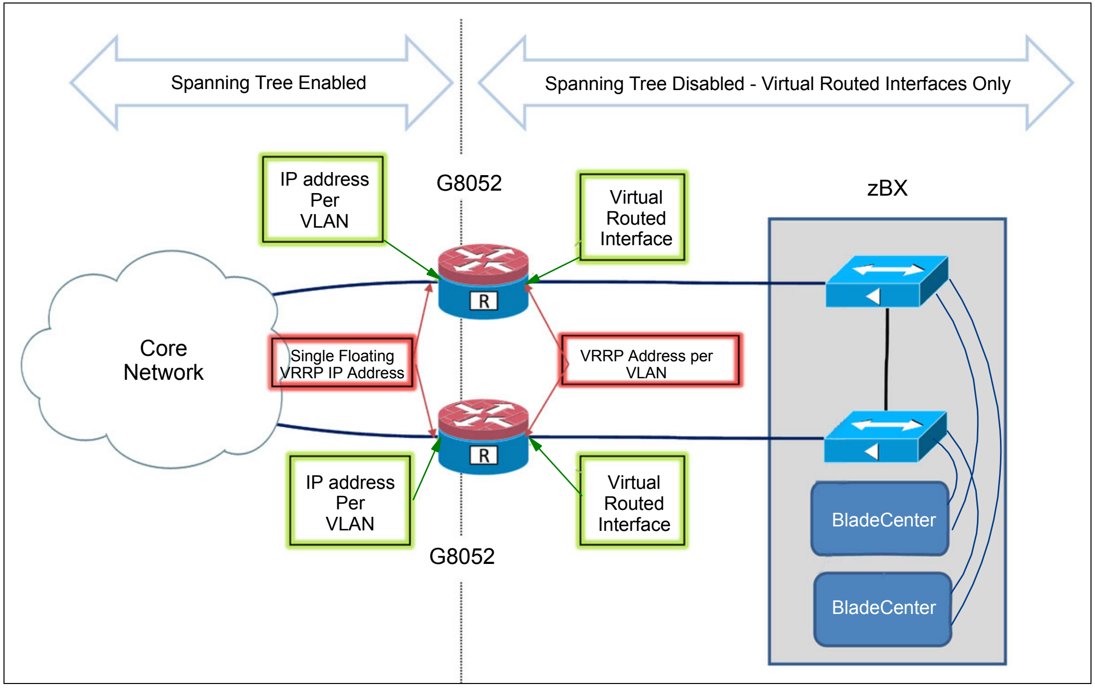

Figure 3-1 Routed endpoint with redundancy through VRRP on the IBM G8052

For high availability, Figure 3-1 shows that Virtual Router Redundancy Protocol (VRRP) provides failover from one routed endpoint to another in the event of a router failure. VRRP is deployed in the core part of the network on the left and in the network attaching to the zBX on the right.

Using VRRP, the two physical routers are grouped to provide the appearance of a single logical router to end-hosts, enabling higher availability without requiring configuration on the end-hosts of dynamic routing or router discovery protocols. One G8052 is the VRRP master for each VLAN and provides the primary/active path to the zBX; the second G8052 serves as the backup router that can take over in the event of a failure of the primary router.

In the core part of the network on the left in Figure 3-1, a single floating IP address is assigned to the virtual router, with ownership of this IP address always assigned to the interface of the active/master physical router. On the right side of the graphic, separate VRRP addresses for each of the IEDN VLANs are assigned to the interface of the active/master physical router.

3.1.1 Base router configuration

Table 3-1 shows the command-line interface (CLI) commands for the base router configuration for this option. These commands identify that the newer version of the ISCLI will be used on a pair of G8052 platforms, and assigns host names to each of them. It also globally enables STP on the platform.

Table 3-1 CLI commands for basic router configuration

DMZ-IBM-G8052-06	DMZ-IBM-G8052-07

`version "7.6.1"`
`switch-type "IBM Networking Operating System RackSwitch G8052"`
`iscli-new`
`!`
`hostname "DMZ-IBM-G8052-06"`
`!`
`spanning-tree loopguard`
	`version "7.6.1"`
	`switch-type "IBM Networking Operating System RackSwitch G8052"`
	`iscli-new`
	`!`
	`hostname "DMZ-IBM-G8052-07"`
	`!`
	`spanning-tree loopguard`

We now examine the left side of Figure 3-1 on page 26, or the core network, more closely.

3.1.2 Connectivity to the core data center network

In Figure 3-2 on page 28, note that the blue arrow at the top indicates that this configuration is Spanning Tree friendly in that it is compatible with enabling STP on core network connections. Enabling STP on behalf of the DMZ enables the automated detection and resolution of network loops.

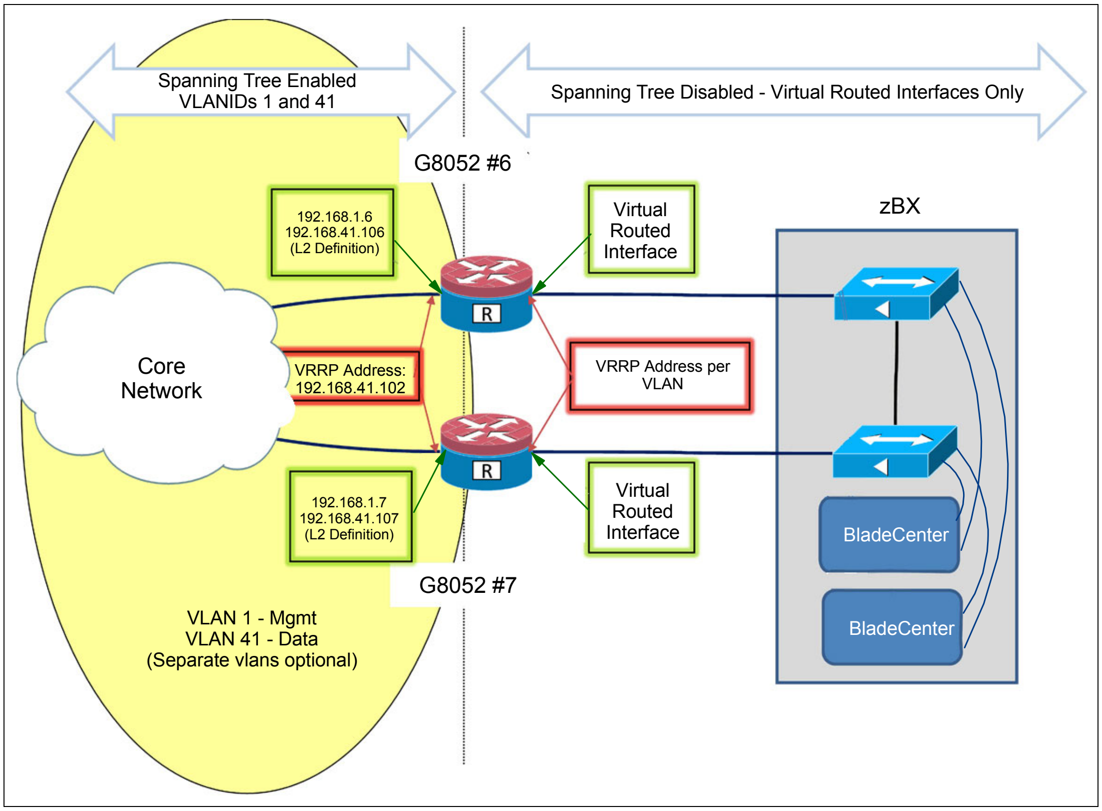

Figure 3-2 Connectivity to core network, routed endpoint with redundancy through VRRP on the IBM G8052

We used Per VLAN Rapid Spanning Tree (PVRST) in the core network and extended to the attached G8052s. Any STP variation supported in your data center switches and the platform that attaches to the zBX can be used on the core network side when using this connectivity option. No special changes are required in any of the core routers if you add the G8052/zBX configuration to a previously existing core network because they route all zBX traffic to the VRRP floating address (192.168.41.102). Although Figure 3-2 shows only a single connection between the core network and each G8052, you can add additional redundancy, for example, when using multiple connections, aggregating multiple connections through Link Aggregation Control Protocol (LACP), and so on. In this example, each G8052 has a management IP address in management VLAN 1 (which would be an optional part of the configuration), and each G8052 also has a router IP address in data VLAN 41, the two of which are merged into a single floating virtual router address that is pointed to by routing tables within the core network.

3.1.3 Core network interface configuration

Table 3-2 on page 29 shows sample ISCLI commands for configuring the uplink from each router platform to the core network, which in this example is embodied in an IBM G8264 platform, for both the data and management VLANs.

Table 3-2 ISCLI commands for configuring uplink from router to core network

DMZ-IBM-G8052-06	DMZ-IBM-G8052-07

`vlan 1`
`name "VLAN_1_DMZ_192.168.1.0-63"`
`!`
`vlan 41`
`name "VLAN_41_DMZ_192.168.41.0-255"`
`!`
`spanning-tree stp 1 vlan 1`
`spanning-tree stp 41 vlan 41`
`!`
`interface port XGE4`
`description "'TO DMZ-IBM-G8264-51 RACK 71' - VLANs"`
`switchport mode trunk`
`switchport trunk allowed vlan 1,41`
`spanning-tree guard loop`
`exit`
	`vlan 1`
	`name "VLAN_1_DMZ_192.168.1.0-63"`
	`!`
	`vlan 41`
	`name "VLAN_41_DMZ_192.168.41.0-255"`
	`!`
	`spanning-tree stp 1 vlan 1`
	`spanning-tree stp 41 vlan 41`
	`!`
	`interface port XGE4`
	`description "'TO DMZ-BNT-G8264-51 RACK 71' - VLANs"`
	`switchport mode trunk`
	`switchport trunk allowed vlan 1,41`
	`spanning-tree guard loop`
	`exit`

3.1.4 Core network IP configuration

Table 3-3 shows sample ISCLI commands for configuring IP addresses in the two router platforms. An IP address is configured in VLAN 1 to receive management network traffic, and a separate IP address in VLAN 41 for receiving data network traffic destined for hosts inside the zBX. The default VLAN for an interface is VLAN 1, so it is not necessary to specify VLAN 1 explicitly when defining the interface for the management traffic.

The default gateway statement in Table 3-3 indicates to the G8052 platforms the router address to which they should direct traffic received on VLAN 41 that is destined to hosts on another VLAN.

Table 3-3 ISCLI commands for configuring IP addresses on routers

DMZ-IBM-G8052-06	DMZ-IBM-G8052-07

`! Management Interface`
`interface ip 1`
`ip address 192.168.1.6 255.255.254.0`
`enable`
`exit`
`!`
`! Data Interface`
`interface ip 41`
`ip address 192.168.41.106 255.255.255.0`
`vlan 41`
`enable`
`exit`
`!`
`!`
`! Default Gateway on VLAN 41 (not us)`
`ip gateway 41 address 192.168.41.100`
`ip gateway 41 enable`
	`! Management Interface`
	`interface ip 1`
	`ip address 192.168.1.7 255.255.254.0`
	`enable`
	`exit`
	`!`
	`! Data Interface`
	`interface ip 41`
	`ip address 192.168.41.107 255.255.255.0`
	`vlan 41`
	`enable`
	`exit`
	`!`

3.1.5 Connectivity to the zBX

Consider the connectivity of the router platforms to the zBX as in the yellow oval in Figure 3-3. The arrow at the top of the figure shows that spanning tree is not permitted for this connectivity, which is in line with the connectivity rules we outlined earlier.

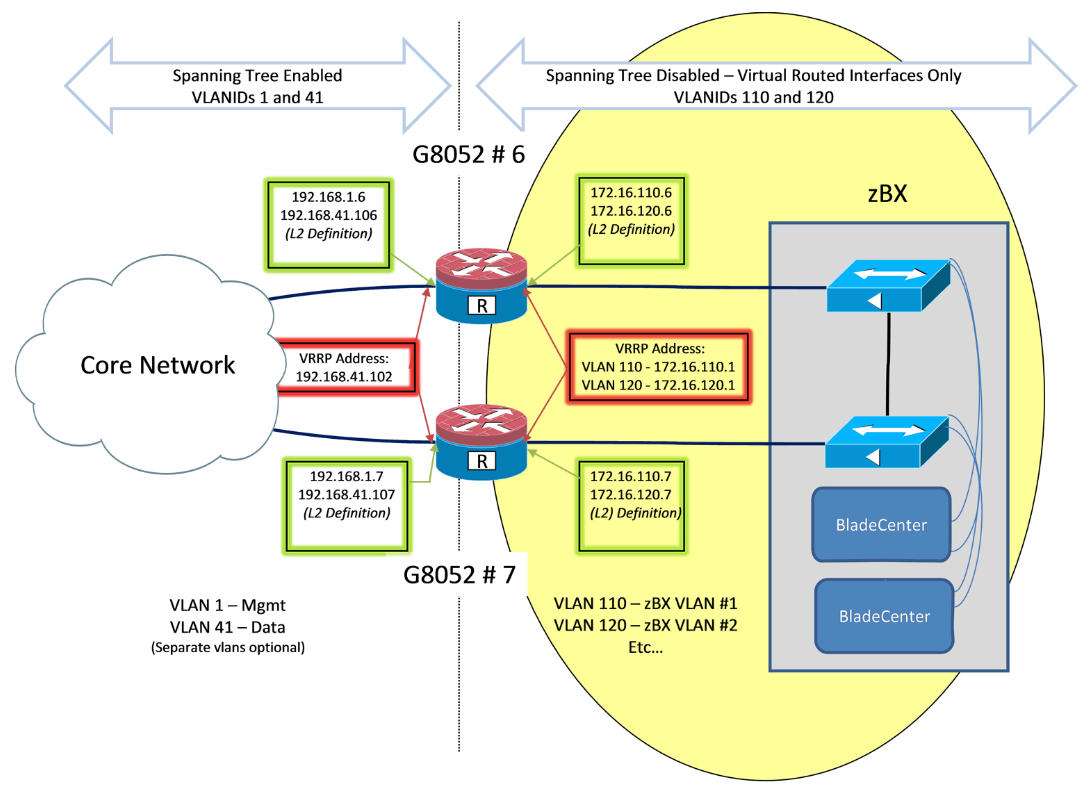

Figure 3-3 Connectivity to zBX, routed endpoint with redundancy through VRRP on the IBM G8052

The connection between the routers and the zBX can be in access mode or in trunk mode with multiple VLANs being trunked on each link (VLAN ID 110, VLAN ID 120, and so on). If the router ports depicted in each G8052 are operating in trunk mode, each router has a separate router IP address in each zBX VLAN ID. However, a single floating VRRP address is shared by the two G8052s for each zBX VLAN, and this IP address serves as the VLAN's default route.

3.1.6 zBX interface configuration

Table 3-4 on page 31 shows ISCLI statements that define the interfaces on the G8052s that link to the zBX TORs. The ports are configured in trunk mode with allowable VLAN IDs 110 and 120. STP is disabled on these VLANs, in recognition of the rule stated earlier regarding the prohibition of BPDUs being exchanged with the zBX TORs.

Table 3-4   ISCLI commands for configuring G8052s to link to zBX TORs

DMZ-IBM-G8052-06	DMZ-IBM-G8052-07
! Link to zBX TORs ! vlan 110 name "VRRP TEST 110" ! vlan 120 name "VRRP TEST 120" ! ! Disable Spanning Tree on zBX VLANS ! spanning-tree stp 110 vlan 110 spanning-tree stp 120 vlan 120 no spanning-tree stp 110 enable no spanning-tree stp 120 enable ! interface port XGE1 description "'Cable #2 TO zBX 003 TOR SWITCH Port 31' - VLANs" switchport mode trunk switchport trunk allowed vlan 110,120 vlan dot1q tag native switchport trunk native vlan 110 exit	! Link to zBX TORs ! vlan 110 name "VRRP TEST 110" ! vlan 120 name "VRRP TEST 120" ! ! Disable Spanning Tree on zBX VLANS ! spanning-tree stp 110 vlan 110 spanning-tree stp 120 vlan 120 no spanning-tree stp 110 enable no spanning-tree stp 120 enable ! interface port XGE1 description "'Cable #1 TO zBX TOR SWITCH Port 31' - VLANs" switchport mode trunk switchport trunk allowed vlan 110,120 vlan dot1q tag native switchport trunk native vlan 110 exit

3.1.7 zBX IP configuration

Table 3-5 shows ISCLI statements that define the IP addresses for the two VLAN interfaces to the zBX. These IP addresses are used by hosts inside the zBX as their next hop router address when accessing hosts within the core network.

Table 3-5   ISCLI statements for defining two VLAN interfaces to zBX

DMZ-IBM-G8052-06	DMZ-IBM-G8052-07
! Guest VLAN 110 interface ip 110 ip address 172.16.110.6 255.255.255.0 vlan 110 enable exit ! ! Guest VLAN 120 interface ip 120 ip address 172.16.120.6 255.255.255.0 vlan 120 enable exit	! Guest VLAN 110 interface ip 110 ip address 172.16.110.7 255.255.255.0 vlan 110 enable exit ! ! Guest VLAN 120 interface ip 120 ip address 172.16.120.7 255.255.255.0 vlan 120 enable exit

3.1.8 VRRP configuration

Table 3-6 shows the ISCLI commands for configuring VRRP across the pair of G8052s, presenting a single virtual router interface to the connection to the core network VLAN 41. Both router definitions are identical except for differing priorities. Generally, the router with the higher priority is the master, so in this case G8052-07 becomes the master and G8052-06 becomes the backup. The first advertised parameter ensures a faster (subsecond) recovery from link outages. The track interfaces parameter ensures that the switch with the most active router interfaces becomes the master. This ensures that a link failure on the zBX side, which decreases the active routed interface count for that router, reduces the priority of the switch with the failed link, and causes a corresponding failover on the core. This ensures that the core network sends traffic only to the switch that has a functional link to the zBX.

Table 3-6 VRRP configuration for interfaces to core network

DMZ-IBM-G8052-06	DMZ-IBM-G8052-07

`router vrrp`
`enable`
`!`
`virtual-router 41 virtual-router-id 41`
`virtual-router 41 interface 41`
`virtual-router 41 priority 106`
`virtual-router 41 address 192.168.41.102`
`virtual-router 41 enable`
`virtual-router 41 timers advertise 20`
`virtual-router 41 fast-advertise`
`virtual-router 41 track interfaces`
`!`
	`router vrrp`
	`enable`
	`!`
	`virtual-router 41 virtual-router-id 41`
	`virtual-router 41 interface 41`
	`virtual-router 41 priority 107`
	`virtual-router 41 address 192.168.41.102`
	`virtual-router 41 enable`
	`virtual-router 41 timers advertise 20`
	`virtual-router 41 fast-advertise`
	`virtual-router 41 track interfaces`
	`!`

3.1.9 VRRP configuration for connecting to the zBX

Table 3-7 on page 33 shows the ISCLI commands for configuring the VRRP protocol across the pair of G8052s for their interfaces that connect to the zBX on VLANs 110 and 120.

Table 3-7 VRRP configuration for connecting to the zBX

DMZ-IBM-G8052-06	DMZ-IBM-G8052-07
router vrrp enable ! virtual-router 110 virtual-router-id 110 virtual-router 110 interface 110 virtual-router 110 priority 106 virtual-router 110 address 172.16.110.1 virtual-router 110 enable virtual-router 110 timers advertise 20 virtual-router 110 fast-advertise virtual-router 110 track interfaces ! virtual-router 120 virtual-router-id 120 virtual-router 120 interface 120 virtual-router 120 priority 106 virtual-router 120 address 172.16.120.1 virtual-router 120 enable virtual-router 120 timers advertise 20 virtual-router 120 fast-advertise virtual-router 120 track interfaces !	router vrrp enable ! virtual-router 110 virtual-router-id 110 virtual-router 110 interface 110 virtual-router 110 priority 107 virtual-router 110 address 172.16.110.1 virtual-router 110 enable virtual-router 110 timers advertise 20 virtual-router 110 fast-advertise virtual-router 110 track interfaces ! virtual-router 120 virtual-router-id 120 virtual-router 120 interface 120 virtual-router 120 priority 107 virtual-router 120 address 172.16.120.1 virtual-router 120 enable virtual-router 120 timers advertise 20 virtual-router 120 fast-advertise virtual-router 120 track interfaces !

Verification

Table 3-8 outlines the verification process for the VRRP to zBX connection.

Table 3-8 VRRP to zBX connection verification

DMZ-IBM-G8052-06	DMZ-IBM-G8052-07
DMZ-BNT-G8052-06> show ip information VRRP information: 41: vrid 41, 192.168.41.102, if 41, renter, prio 114, backup 110: vrid 110, 172.16.110.1, if 110, renter, prio 114, backup 120: vrid 120, 172.16.120.1, if 120, renter, prio 114, backup	DMZ-BNT-G8052-07>show ip vrrp information VRRP information: 41: vrid 41, 192.168.41.102, if 41, renter, prio 115, master 110: vrid 110, 172.16.110.1, if 110, renter, prio 115, master 120: vrid 120, 172.16.120.1, if 120, renter, prio 115, master

3.2 Option 2: Dual Cisco routers: Routed endpoint with HSRP

This configuration option offers a variation on Option 1: Dual IBM networking division G8052s with a routed endpoint using VRRP. However, this option has two distinctions: First, Cisco routers are used instead of IBM routers; and second, Hot Standby Router Protocol (HSRP) is used instead of VRRP. HSRP is the redundancy protocol that provides failover in the case of the failure of a router or of the single link between the router and the zBX TOR.

3.2.1 Topology

Because this example is similar to Option one, it illustrates only a setup diagram, which is shown in Figure 3-4, along with sample configuration commands.

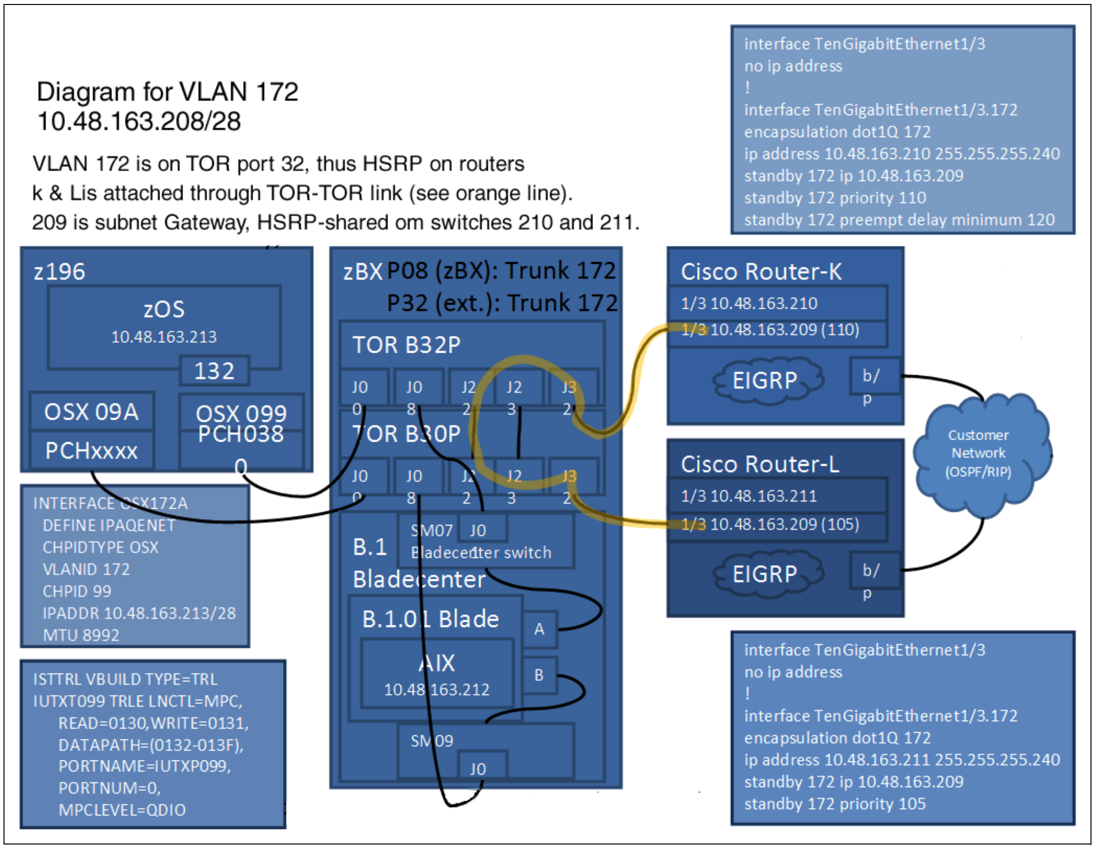

Diagram for VLAN 172
10.48.163.208/28

VLAN 172 is on TOR port 32, thus HSRP on routers
k & Lis attached through TOR-TOR link (see orange line).
209 is subnet Gateway, HSRP-shared om switches 210 and 211.

```
interface TenGigabitEthernet1/3
no ip address
!
interface TenGigabitEthernet1/3.172
encapsulation dot1Q 172
ip address 10.48.163.210 255.255.255.240
standby 172 ip 10.48.163.209
standby 172 priority 110
standby 172 preempt delay minimum 120
```

```
INTERFACE OSX172A
  DEFINE IPAQENET
  CHPIDTYPE OSX
  VLANID 172
  CHPID 99
  IPADDR 10.48.163.213/28
  MTU 8992
```

```
ISTTRL VBUILD TYPE=TRL
IUTXT099 TRLE LNCTL=MPC,
    READ=0130,WRITE=0131,
    DATAPATH=(0132-013F),
    PORTNAME=IUTXP099,
    PORTNUM=0,
    MPCLEVEL=QDIO
```

```
interface TenGigabitEthernet1/3
no ip address
!
interface TenGigabitEthernet1/3.172
encapsulation dot1Q 172
ip address 10.48.163.211 255.255.255.240
standby 172 ip 10.48.163.209
standby 172 priority 105
```

Figure 3-4 Cisco router: Routed endpoint with redundancy through HSRP

In this example, IP address 10.48.163.209 is the shared gateway address for both routers, although each router has their own IP address on the configured subnet/VLAN. Router-K has a higher priority (110 versus 105), so it will be the primary router and Router-L will be the backup. Using HSRP provides for redundant paths to the subnet gateway. The blades inside the zBX have a path to the external network, with one active and one standby TOR connection to a pair of primary and secondary HSRP switches, respectively. As described earlier, the “heartbeat” used by the HSRP routers flows through the TOR switches (through the IEDN), so that the HSRP routers see each other through the TOR-TOR links. A failure of either the active TOR or the primary HSRP router will cause a failover to the secondary (backup) HSRP router, and the backup TOR will gain access to the gateway address on the new primary HSRP router (after it has quickly taken over ownership of the gateway address).

3.2.2 Configuration

Table 3-9 illustrates the Cisco CLI commands for configuring HSRP across a pair of Cisco router interfaces that connect to the zBX on VLAN 172.

Table 3-9 Cisco CLI commands for configuring HSRP across Cisco routers to zBX on VLAN 172

Cisco Router-K	Cisco Router-L

`!`
`interface TenGigabitEthernet1/3`
`no ip address`
`!`
`interface TenGigabitEthernet1/3.172`
`encapsulation dot1Q 172`
`ip address 10.48.163.210 255.255.255.240`
`standby 172 ip 10.48.163.209`
`standby 172 priority 110`
`standby 172 preempt delay minimum`
	`!`
	`interface TenGigabitEthernet1/3`
	`no ip address`
	`!`
	`interface TenGigabitEthernet1/3.172`
	`encapsulation dot1Q 172`
	`ip address 10.48.163.211 255.255.255.240`
	`standby 172 ip 10.48.163.209`
	`standby 172 priority 105`
	`!`

3.2.3 Verification

This verification process, which is shown in Table 3-10, indicates that HSRP is active and Cisco Router-K is primary and is enabled for take back (as indicated by preemption), while Cisco Router-L is backup. HSRP is active on both routers.

Table 3-10 Verification for Cisco routers connecting to zBX through HSRP

Cisco Router-K	Cisco Router-L
`Router-K#show stand ten1/3.172` ► TenGigabitEthernet1/3.172 - Group 172 ► Local state is Active, priority 110, may preempt ► Preemption delayed for at least 120 secs ► Hellotime 3 sec, holdtime 10 sec ► Next hello sent in 1.950 ► Virtual IP address is 10.48.163.209 configured ► Active router is local ► Standby router is 10.48.163.211 expires in 7.512 ► Virtual mac address is 0000.0c07.acac ► 5 state changes, last state change 2w0d ► IP redundancy name is "hsrp-Te1/3.172-172" (default)	`Router-L#show stand ten1/3.172` ► TenGigabitEthernet1/3.172 - Group 172 ► Local state is Standby, priority 105 ► Hellotime 3 sec, holdtime 10 sec ► Next hello sent in 0.732 ► Virtual IP address is 10.48.163.209 configured ► Active router is 10.48.163.210, priority 110 expires in 9.420 ► Standby router is local ► 7 state changes, last state change 1w6d ► IP redundancy name is "hsrp-Te1/3.172-172" (default)

3.3 Option 3: Single IBM Networking Division G8052: A routed identity with HotLinks

We now describe a third configuration and another variation where the connection endpoint terminates the IEDN Layer-2 network and forwards packets to and from other non-IEDN destinations on Layer-3, in other words, using routing.

3.3.1 Description

In contrast to the previous example where VRRP was used to provide redundancy across multiple connections to the zBX, in this option, redundancy is provided with HotLinks in a single attached platform. HotLinks provides the ability for multiple links emanating from the same platform to provide redundant connections to an attached host platform. One link is established as the active/primary link, and the other as the backup link. A single floating IP address is assigned to the virtual routed interface, and network traffic destined for this floating IP address always flows over the link that is currently active. In the event of a failure of the active link, the backup link becomes the new active link and assumes ownership of the associated IP address.

It is of course possible, and even recommended, to use two router platforms where each has two links to the zBX, one connected to each of the TOR switches in the zBX. In this case, you would configure VRRP as in Option 1 to provide for failover in the case of an entire router platform failure, and HotLinks as in Option 3 to provide failover in the case of a single link failure.

3.3.2 Topology

Figure 3-5 outlines Option 3: Routed endpoint using HotLinks.

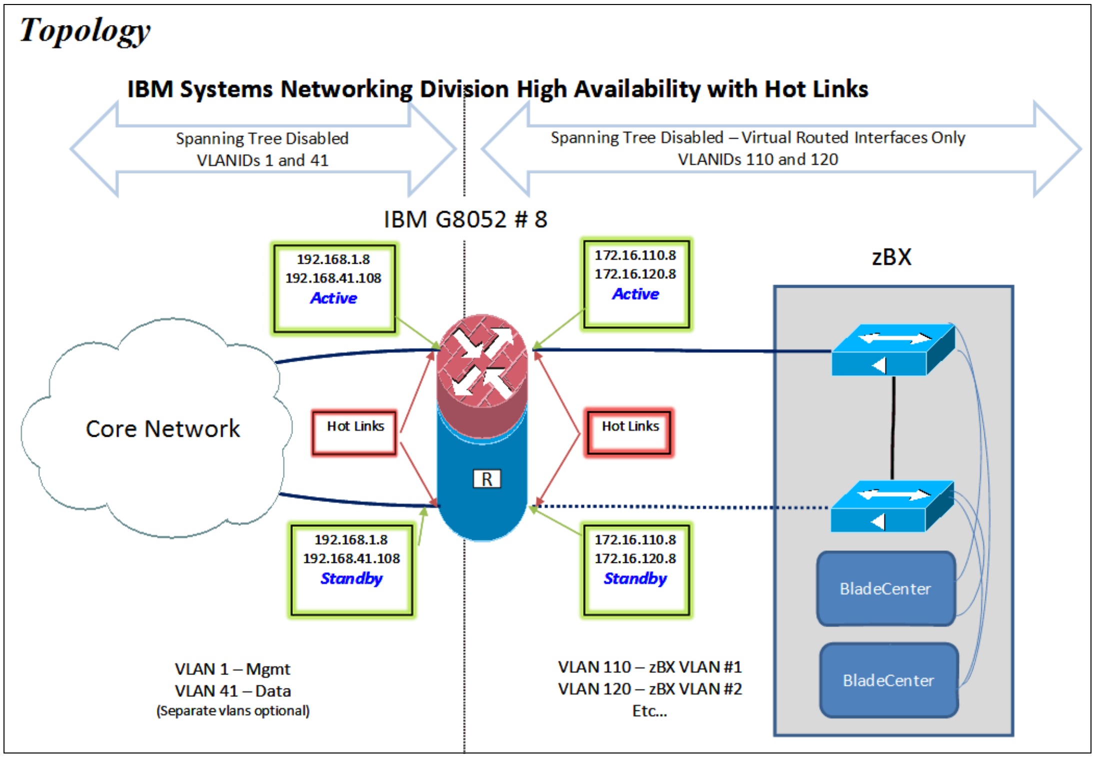

Figure 3-5 Routed endpoint with redundancy through HotLinks on IBM G8052

4

Layer-2 configuration options

IBM recently announced that it would support connectivity to the zEnterprise BladeCenter Extension (zBX) using switched endpoints, that is, connections at Layer-2. This support provides an easier migration path for moving existing data center workloads to the zBX environment to take advantage of System z governance and zManager's unique workload management functions.

Earlier, we described how the zBX and the intraensemble data network (IEDN) are configured for high availability, and how some Layer-2 connectivity options are not compatible with that configuration. We noted that you need to adhere to certain rules to successfully connect to the zBX external ports. We repeat these important rules here:

- You cannot use any Layer-2 protocol that attempts to use Bridge Protocol Data Units (BDPUs) to communicate with the top-of-rack (TOR) switches in the zBX. However, you can use protocols that encapsulate BDPUs in a frame sent to the externally facing TOR ports if their VLAN tagging is valid. This means that the VLAN tagged encapsulated frames are authorized to traverse the IEDN TORs.
- The allowable IEDN VLAN range is limited to VLANs 10 - 1030.

In this chapter, we describe two options that use Layer-2 switching to forward frames between the external network and the zBX, namely:

1. Merging the IEDN VLAN domain with the external network VLAN domain while eliminating BPDUs between the node next to the zBX's external ports. The merge is implemented using an IBM Systems Networking Division G8052.
2. Merging the IEDN VLAN domain with the external network VLAN domain while implementing Per VLAN Spanning Tree (PVST) on a pair of Cisco devices.

> **Note:** These examples are not intended to provide the exclusive set of supported options, and many other possible variations may exist. Syntax of sample coding statements uses the platform's newer version of its Industry Standard Command Line Interface (ISCLI) initially available in code level 7.6.1. Differences in syntax and functional capabilities are likely to exist when using other code levels or other vendor platforms. The sample coding statements that are provided are not intended to show all possible configurations, but rather to highlight examples of specific CLI statements.

© Copyright IBM Corp. 2014. All rights reserved.

4.1 Option 1: IBM Systems Networking Division G8052 with switched endpoint using HotLinks

This option is an example of how you can implement the newly supported attachment of the data center network to the zBX using Layer-2 switching to forward frames between data center hosts and hosts that are inside the zBX. Figure 4-1 illustrates this configuration. You can see that the data center and IEDN VLAN IDs merge, as frames are simply forwarded at Layer 2 through the switching platform. The switching platform does not route, and thus does not have its own IP address that would be necessary for participating in routing.

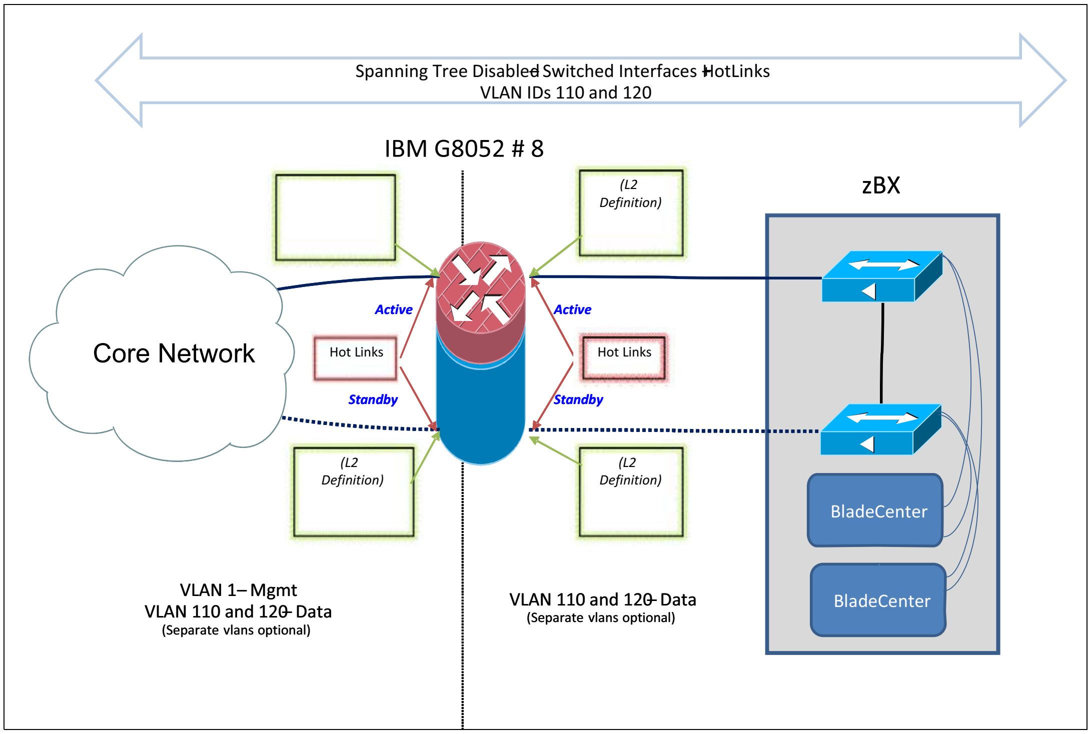

Figure 4-1 Switched endpoint with HotLinks for failover on the IBM G8052

> **Note:** STP is not compatible with attaching to zBX. Using IBM ISCLI 7.6.1, it is not possible to globally enable STP while disabling it on a per switchport basis, so in this example it has been globally disabled. This might be possible if you are using other vendor Layer-2 technologies similar to HotLinks; it might also be possible in a subsequent level of IBM ISCLI.

Tables in the following sections provide sample IBM ISCLI level 7.6.1 coding statements that were used to configure this option, a switched connection to the zBX with redundancy through HotLinks on a single IBM G8052 as the switching platform. In this example, the base configuration would be similar to 3.1, “Option 1: Dual IBM RackSwitch G8052s: Routed endpoint with VRRP” on page 26, except that the core and zBX Interface statements differ, along with the core and zBX IP configurations.

4.1.1 Connectivity to the core data center network

This connectivity involves an uplink from each router platform to the core network, which is embodied here in an IBM G8264 platform, for both of the data VLANs 110 and 120. VLAN 1 is included because it would likely be used for the management interface.

Configuration

Table 4-1 shows sample ISCLI commands for configuring this option.

Table 4-1 ISCLI commands to configure the uplink from each router to the core network

DMZ-IBM-G8052-08

```
vlan 110
        name "VLAN_1_DMZ_192.168.110.0-255"
!
vlan 120
        name "VLAN_120_DMZ_192.168.120.0-255"
!
spanning-tree stp 110 vlan 110
spanning-tree stp 120 vlan 120
!
!
interface port XGE3
        description "'LACP TRUNK 1 of 2  - Core ""
        switchport mode trunk
        switchport trunk allowed vlan 1,110,120
        spanning-tree guard loop
        exit
!
interface port XGE4
        description "'LACP TRUNK 2 of 2 - Core"
        switchport mode trunk
        switchport trunk allowed vlan 1,110,120
        spanning-tree guard loop
        exit
```

4.1.2 Core network IP configuration

Table 4-2 displays sample ISCLI commands for configuring IP addresses in the switch. An IP address is configured in VLAN 1 to receive management network traffic.

Table 4-2 Sample ISCLI commands for configuring IP address in switch

DMZ-IBM-G8052-08

```
! Management Interface for VLAN 1
interface ip 1
        ip address 192.168.1.8 255.255.254.0
        vlan 1
        enable
        exit
!

!
! Default Gateway on VLAN 1  (not us)
ip gateway 1  address 192.168.1.1
ip gateway 1  enable
```

4.1.3 Connectivity to the zBX interface

Table 4-3 displays ISCLI statements for defining G8052 interfaces that link to the zBX TORs. The ports are configured in trunk mode with allowable VLAN IDs 110 and 120. STP is disabled globally in the switch, a requirement to enable HotLinks. However, after the HotLinks functionality is enabled, one interface becomes the master and the second interface becomes the backup.

Table 4-3 Sample ISCLI statements for configuring zBX interfaces

DMZ-IBM-G8052-08

```
!
spanning-tree mode disabled
!
!
interface port XGE1
        description "'LAYER 2 HOTLINK TRUNK 1 TO zBX TOR SWITCH' - VLANs"
        switchport mode trunk
        switchport trunk allowed vlan 110,120
                exit
!
interface port XGE2
        description "'LAYER 2 HOTLINK TRUNK 2 TO zBX TOR SWITCH' - VLANs"
        switchport mode trunk
        switchport trunk allowed vlan 110,120
                exit
hotlinks trigger 1 master port XGE1
hotlinks trigger 1 backup port XGE2
hotlinks trigger 1 enable
!
hotlinks enable
```

4.2 Option 2: Cisco switching equipment using PVST next to zBX

Per VLAN Spanning Tree (PVST) is the default protocol for resolving loops on Cisco switches. Unlike the IEEE STP and RSTP protocol suite, the loops are resolved on a per VLAN basis instead of a per cable/link basis. Cisco uses vendor-specific PDUs and protocols for loop detection and elimination. For switches that do not have awareness of PVST, the PDUs appear as regular multicast PDUs and are transferred like regular multicast packets. In essence, the whole non-PVST aware cloud looks like a single hub or STP/PVST transparent switch.

Figure 4-2 on page 41 illustrates this scenario as it relates to the zBX environment. The client data center network is running Cisco intra-switch links (ISLs) and the Cisco PVST+ protocol for loop resolution. However, the links to the zBX are using standard 802.1q encapsulated packets. The PVST+ packets transmitted by the Cisco device are traversing the IEDN like regular multicast packets. The PVST packets emitted by the data center switch 1 (DC-Switch 1) are entering the IEDN at the lower left zBX node-TOR B in the figure. The packets traverse the whole IEDN remember that the IEDN TOR pairs are interconnected, and leave the IEDN at the lower right zBX node: TOR B. They finally arrive at the DC-Switch 2 on the right side of the figure.

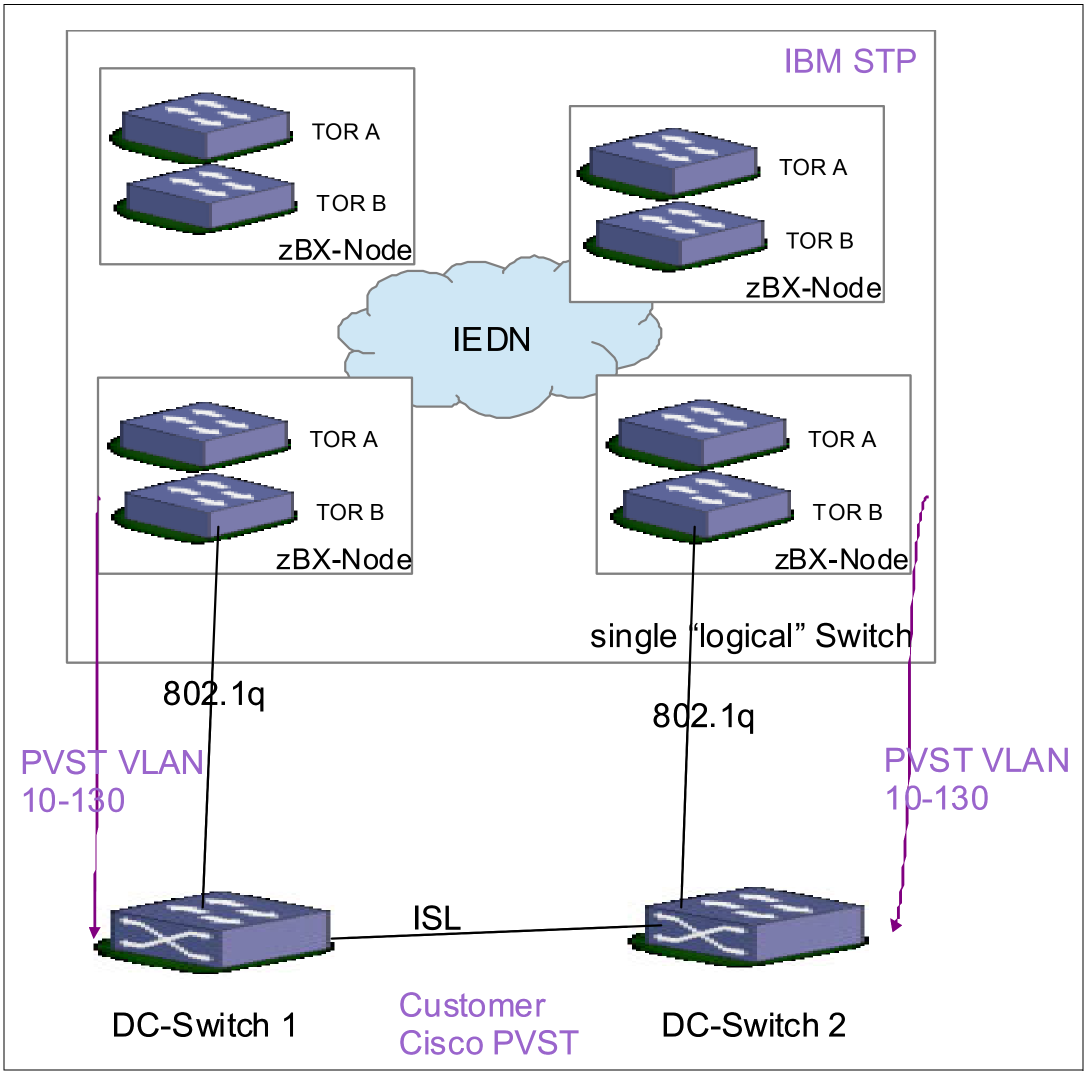

Figure 4-2 Configuration for Cisco switches using PVST

The DC-Switch 2 is now able to detect a loop and shutdown a link. Based on path weights, one of the links to the IEDN is shut down, for example the one from the DC switch 2 to the IEDN. This loop determination and resolution is done on a per-VLAN basis, so a different link can be chosen to be shut down for each of the configured VLANs.

Cisco has special handling for VLAN ID 1. You can refer to the Cisco documentation for more details. This VLAN is also used for compatibility mode with STP and RSTP, and so on. By avoiding the use of this VLAN ID, there is no interference with the IEDN STP topology.

In this example, the use of Cisco switches with PVST enabled allows the entire IEDN to look like a single switch to the Cisco data center domain, with the IEDN infrastructure transparent to PVST as the Cisco data center domain uses this protocol to detect and resolve network loops.

You can use the configuration steps below for this example. The Cisco default settings are such that PVST is enabled, but we recommend that you verify that PVST is turned on and the settings are as wanted. We recommend the following configuration steps on the Cisco switch:

```
# configure terminal
# spanning-tree mode rapid-pvst
# spanning-tree vlan 170
```

The last step enables PVST on VLAN 170, and it should similarly be enabled for each deployed VLAN. As noted earlier, PVST is enabled by default for each created VLAN on Cisco devices, so this step is generally optional. To verify that PVST is enabled, you can issue the following optional commands:

```
# configure terminal
# show spanning-tree vlan 170
```

5

Summary

This paper provided a basic review of how to connect to the IBM zEnterprise BladeCenter Extension (zBX) using both Layer-3 (routing) and Layer-2 (switching) technologies. The support for attaching external platforms to the zBX using switched endpoints, that is, connecting at Layer-2, applies to all existing zBX models across all supported zEnterprise ensemble configurations.

It reviewed the intraensemble data network (IEDN) design and described how and why it was intended to be a private, isolated, and flat Layer-2 network that is exclusively managed and secured by Unified Resource Manager with zEnterprise firmware, along with the value propositions such a configuration provides. It also described how to achieve IEDN high availability in the zBX, including its reliance on the highly controlled use of STP. You need to understand the definitions and functions available on the platform that is adjacent to the zBX as you design the connectivity between an external platform and the zBX. Do not rely on a design that insists that the zBX and the adjacent platform exchange switching messages (Bridge Protocol Data Units (BPDUs)).

It described how you can use multiple technologies other than STP/RSTP/MSTP that are widely available across networking switch vendors to provide high availability Layer-2 connections to the zBX. It also illustrated how to use these technologies with Layer-3 failover protocols to provide high availability when routed interfaces in Layer-2/Layer-3 platforms are used as part of the solution. The connectivity examples provided in this document provide a set of configurations that IBM has tested in either the Washington Systems Center Lab or in conjunction with one of the early adopter clients. These examples are not intended to describe the only supported configurations for connection at Layer-2. Your configuration must abide by the rules outlined in this paper. The VLAN support within the Unified Resource Manager and zBX has not changed, so that the current VLAN range limit of VLANs 10 - 1030 remains in effect. VLANs outside of this range are not supported for traffic entering the zBX.

You can continue to maintain separate VLAN domains inside and outside of the IEDN even when you connect to the zBX using Layer-2 technologies. We recommend that you allow for this level of isolation where possible in order to realize the security provisions for IEDN VLANs provided by zEnterprise, although blending of Layer-2 VLAN domains is not prohibited. In fact, a routed connection remains our recommended best practice because it preserves all of the value propositions of IEDN design and avoids introducing complications of multiple managers of a single Layer-2 network.

© Copyright IBM Corp. 2014. All rights reserved.

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this paper.

IBM Redbooks

The following IBM Redbooks publications provide additional information about the topic in this document. Note that some publications referenced in this list might be available in softcopy only.

- *IBM zEnterprise System Technical Introduction, SG24-8050*
- *IBM zEnterprise EC12 Technical Guide, SG24-8049*
- *IBM zEnterprise BC12 Technical Guide, SG24-8138*
- *IBM System z Connectivity Handbook, SG24-5444*
- *Building an Ensemble using IBM zEnterprise Unified Resource Manager, SG24-7921*
- *IBM zEnterprise EC12 Configuration Setup, SG24-8034*

You can search for, view, download, or order these documents and other Redbooks, Redpapers, Web Docs, draft and additional materials, at the following website:

ibm.com/redbooks

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

© Copyright IBM Corp. 2014. All rights reserved.

IBM zEnterprise BladeCenter Extension

Network Connectivity Options

Redpaper™

Learn about zBX connectivity options

Use technologies from multiple vendors for zBX connections

See how zBX achieves high availability

This IBM Redpaper publication describes the configuration of the networking equipment that attaches to the IBM zEnterprise BladeCenter Extension (zBX), which allows communication with the applications that reside on the intraensemble data network (IEDN). In most cases, the IEDN remains a closed Layer-2 network to maintain a highly available and secure environment that IBM can support. Therefore, when connecting to the IEDN, Layer-3 routed connectivity is still the preferred method. However, now the zBX top-of-rack (TOR) switches support Layer-2 switched connections that can provide an easier migration path when moving data center workloads to the zBX environment.

This paper includes a brief introduction to the IEDN architecture and configuration and how these types of connections work. It also introduces the zBX architecture and explains the implications that network connections can have on the redundancy and high availability setup for this system. Finally, this paper provides concrete examples for connecting the IEDN and external data network through zBX for both Layer-3 routed and Layer-2 switched connection configuration options.

This paper is intended for network architects and network administrators who are responsible for designing and implementing zBX network configurations. It is assumed that you have a basic background in IBM zEnterprise and network concepts.

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks

REDP-5036-00