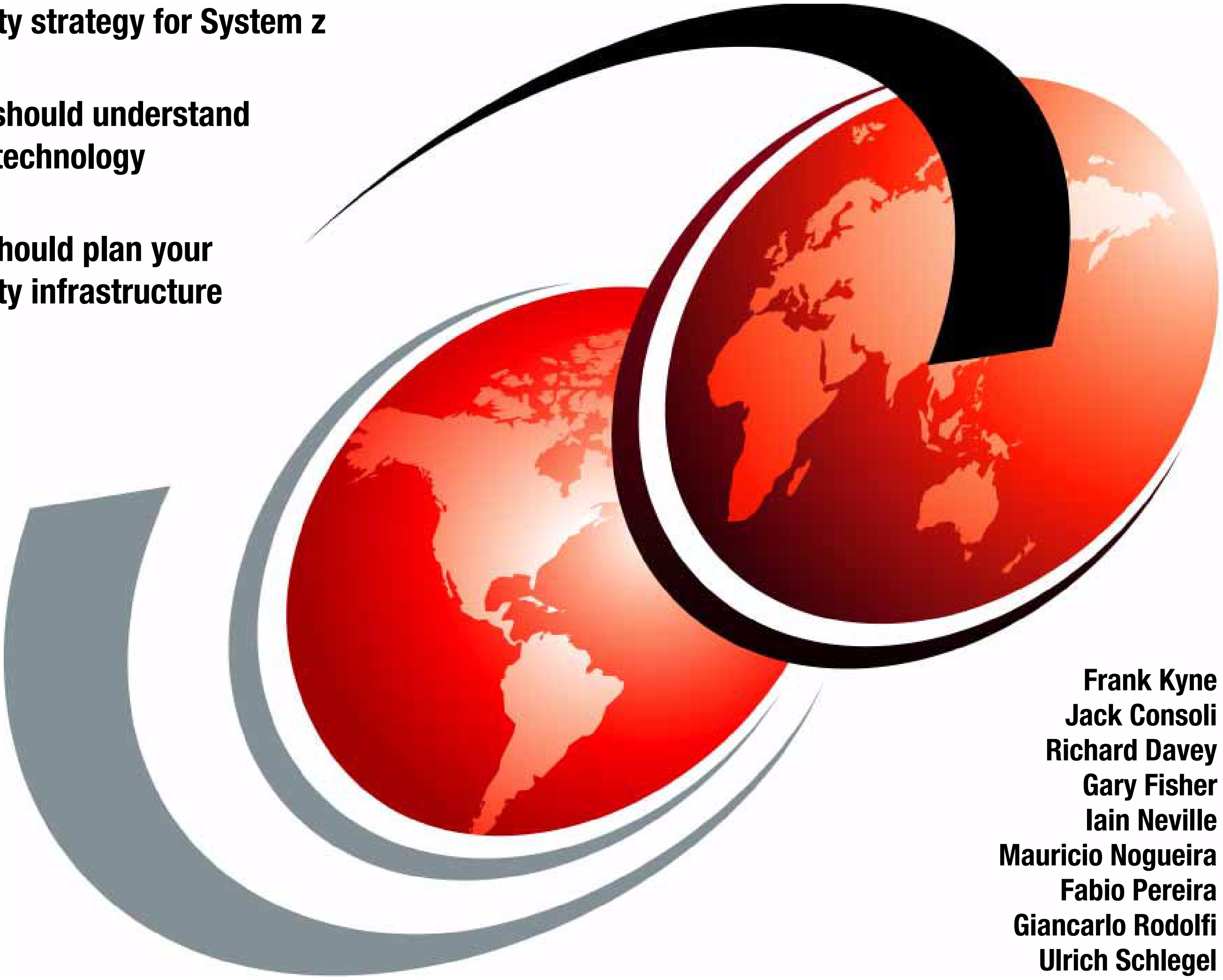
System z End-to-End Extended Distance Guide

Why you should have an end-to-end connectivity strategy for System z

What you should understand about the technology

How you should plan your connectivity infrastructure



Frank Kyne
Jack Consoli
Richard Davey
Gary Fisher
Iain Neville
Mauricio Nogueira
Fabio Pereira
Giancarlo Rodolfi
Ulrich Schlegel

Redbooks

ibm.com/redbooks

International Technical Support Organization

System z End-to-End Extended Distance Guide

March 2014

SG24-8047-00

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

First Edition (March 2014)

Contents

Notices

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

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

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

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

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

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

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

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

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

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

COPYRIGHT LICENSE:

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

Trademarks

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

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

BladeCenter®
CICS®
DB2®
DS8000®
Easy Tier®
ESCON®
eServer™
FICON®
FlashCopy®
GDPS®
Geographically Dispersed Parallel Sysplex™
Global Technology Services®
Guardium®
HyperSwap®
IBM®
IMS™
InfoSphere®
MVS™
Parallel Sysplex®
PR/SM™
RACF®
Redbooks®
Redpapers™
Redbooks (logo) ®
Resource Link®
RMF™
System p®
System Storage®
System x®
System z®
System z10®
System z9®
Tivoli®
VTAM®
WebSphere®
z/OS®
z/VM®
z/VSE®
z10™
z9®
zEnterprise®

The following terms are trademarks of other companies:

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

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

UNIX is a registered trademark of The Open Group in the United States and other countries.

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

Preface

This IBM® Redbooks® publication will help you design and manage an end-to-end, extended distance connectivity architecture for IBM System z®. This solution addresses your requirements now, and positions you to make effective use of new technologies in the future.

Many enterprises implement extended distance connectivity in a silo manner. However, effective extended distance solutions require the involvement of different teams within an organization. Typically there is a network group, a storage group, a systems group, and possibly other teams.

The intent of this publication is to help you design and manage a solution that will provide for *all* of your System z extended distance needs in the most effective and flexible way possible. This book introduces an approach to help plan, optimize, and maintain all of the moving parts of the solution together.

Authors

This book was produced by a team of specialists from around the world working at the International Technical Support Organization (ITSO), Poughkeepsie Center.

Frank Kyne is an Executive IT Specialist at the ITSO, Poughkeepsie Center. He writes extensively and teaches IBM classes worldwide about all aspects of IBM Parallel Sysplex® and high availability. Before joining the ITSO 15 years ago, Frank worked in IBM Ireland as an IBM MVS™ system programmer.

Jack Consoli is a System Engineer based in Connecticut, US, on the IBM original equipment manufacturer (OEM) team at Brocade Communications Systems, Inc. Jack has over 26 years of experience in the development and marketing of enterprise-class switching and storage, business continuity, and disaster recovery products. His roles include driving qualification programs for mainframe storage area network (SAN) solutions. He holds a Bachelor of Science in Electrical Engineering and a Master of Business Administration.

Richard Davey is a mainframe system programmer with the Standard Bank of South Africa. He has 24 years of mainframe experience, 16 of them as a mainframe system programmer. His areas of expertise are in mainframe hardware configuration, IBM Geographically Dispersed Parallel Sysplex™ (IBM GDPS®) extended remote copy (XRC)/Peer-to-Peer Remote Copy (PPRC), and data center site migration.

Gary Fisher has been a programmer for IBM for 33 years. Gary has worked on a wide variety of software and hardware development projects focused on managing computer interconnections for network and data transfer. Gary has received several awards and authored several patents, mostly for work in multi-system processes and automation for connectivity management. Gary received a Bachelor of Science from Buffalo State College, a Master of Science in Computer Science from Rensselaer Polytechnic Institute, is a doctoral candidate at Pace University, and is an adjunct professor at Marist College where he teaches Mainframe Networking. Gary is currently a Mainframe Network consultant based in Poughkeepsie, NY, helping IBM clients create and manage networks to interconnect large and diverse computer installations.

Iain Neville is a Certified Consulting IT Specialist with IBM United Kingdom. He has 24 years of experience in mainframe technical support and consultancy. His areas of expertise include Parallel Sysplex, Server Time Protocol (STP), IBM z/OS®, IBM Fibre Channel connection (FICON®), InfiniBand, and mainframe high availability solutions. Iain's responsibilities include pre-sales mainframe technical consultancy and end-to-end infrastructure design that supports numerous large financial institutions across the UK.

Mauricio Nogueira is a System Programmer at Banco do Brasil, a government bank in Brazil. He has 6 years of experience in mainframe systems, including SAN, data center connectivity, and hardware configuration. He holds a degree in Computer Science from Unimar (Universidade de Marília). His areas of expertise include mainframe hardware configuration, and storage director architecture and implementation.

Fabio Pereira is a Senior System Programmer at Banco do Brasil, a government bank in Brazil. He has 12 years of experience in mainframe systems, including GDPS, Data Facility Storage Management Subsystem (DFSMS), high-end storage systems, and remote copy solutions. He holds a degree in Data Processing, and a graduate degree in High Performance Computing Environments Management from Uniceub (Centro Universitário de Brasília). His areas of expertise include mainframe hardware configuration, GDPS, high-end storage architecture, storage performance for mainframe environments, and remote copy solutions.

Giancarlo Rodolfi is a System z Consultant Technical Sales Specialist in Brazil. He has 28 years of experience in the mainframe field. He has written extensively about the z/OS Communication Server, security, z/OS, and IBM zEnterprise®.

Ulrich Schlegel is Director of Business Development Data Center solutions at ADVA Optical Networking in Munich, Germany. Uli has over 13 years of experience in wavelength-division multiplexing (WDM) technology and optical networking systems. He holds an engineering degree (Dipl.-Ing. Physikalische Technik) from the University of Applied Sciences (TFH) in Berlin, Germany. Uli currently provides consulting to clients on design and implementation of data center networking. He is also driving interoperability programs with various vendors in the data center and SAN space.

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

Connie Beuselinck
Michael Browne
Pasquale (PJ) Catalano
Andrew Crimmins
Casimer DeCusatis
Charles (Hugh) Howard (Retired)
Mark Lewis
Sam Mercier (Retired)
Phil Muller
Ray Newsom
Dennis Ng
Lou Ricci (Retired)
IBM US

Mark Detrick
Ben Hart
David Lytle
Dr. Steven Guendert
Brocade US

Now you can become a published author, too!

This is 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.

Obtain more information about the residency program, browse the residency index, and apply online:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

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

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

  ibm.com/redbooks

- Send your comments in an email:

  redbooks@us.ibm.com

- Mail your comments:

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

Stay connected to IBM Redbooks

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks

- Follow us on Twitter:

  http://twitter.com/ibmredbooks

- Look for us on LinkedIn:

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

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

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

- Stay current on recent Redbooks publications with RSS Feeds:

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

1

Introduction

This chapter provides information about why this IBM Redbooks document was created, and about the background information that you should be familiar with before proceeding with the remainder of this book. Specifically, this chapter covers the following topics:

- Why we wrote this book
- Why you would have multiple data centers
- Common data center models
- The importance of an end-to-end architecture
- Role of the connectivity architecture group
- What needs to be connected
- Connectivity options
- System z qualification and testing programs
- Planning for the future
- Where to go for help
- Layout of this book

1.1 Why we wrote this book

At one time, it was generally larger mainframe clients that had multiple data centers. They tended to be geographically dispersed and not connected to each other, except by Systems Network Architecture (SNA) and Internet Protocol network (Transmission Control Protocol/Internet Protocol, or TCP/IP) connections.

Today, it is routine for System z clients to have multiple data centers. However, now the data centers must increasingly be interconnected for the purposes of continuous availability and disaster recovery (DR). This is in response to the ever-increasing dependence of businesses, governments, and society on IT services. It is also required to meet government regulations related to the availability of those services. Additionally, changes in connection technology mean that it is now possible to do things that previously were not possible or not feasible.

However, in enterprises that connect devices across sites, it is not uncommon to find that the different groups (storage, network, sysplex, DR, and so on) have implemented different strategies, with some common components, some unique components, and no one with overall responsibility for the complete end-to-end solution.

This results in a sub-optimized configuration with a higher overall cost, but reduced flexibility and resilience. Figure 1-1 illustrates a typical configuration with multiple platforms and connectivity options. You can see the challenges faced by someone that is trying to get all of this technology under control.

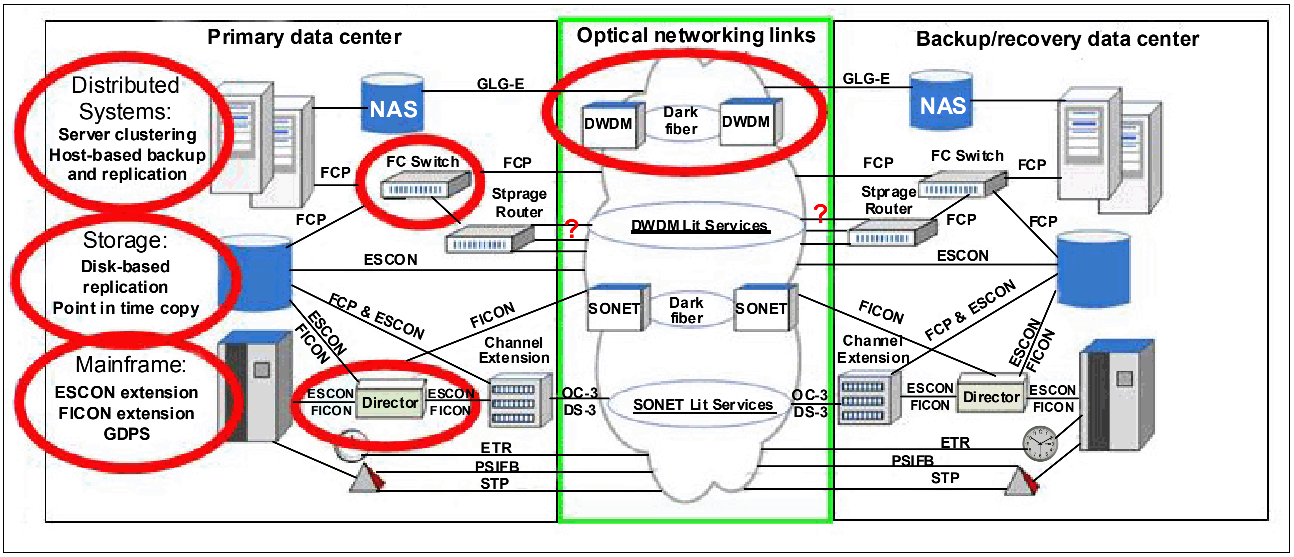


Figure 1-1 Typical multi-data center configuration

The situation is compounded by the fact that the connectivity technology is specialized, and constantly evolving. IBM provides excellent reference documentation for System z connectivity. However, before this document, there was no comprehensive guide to help someone with little or no experience in this area to create an end-to-end connectivity infrastructure that meets your needs now and into the future.

1.1.1 The scope of this book

If you are familiar with a modern storage area network (SAN) environment, you will be aware that many connectivity devices can be used by both mainframe and distributed platforms.

However, if we were to include distributed platforms in the scope of this book, we would have a far larger book, and one where the majority of readers are likely to only be interested in a subset of the contents.

For these reasons, we have limited the scope of this book to the mainframe environment. Where appropriate, we will include information about the relationship between mainframe and distributed connections. However, the *focus* is on System z.

You might be wondering exactly which configurations will be included in this book. If all of your devices are connected directly to the central electronics complex (CEC), it might be obvious that this is not a configuration that will be covered here. As you can imagine, this could potentially be a huge book unless we impose some bounds on the configurations that are included. Therefore, we focus on configurations that have one or both of the following characteristics:

- The distance between the devices and the CECs or other devices that they are connected to is greater than the maximum supported unrepeated distance for that device.
- Dense wavelength division multiplexing (DWDM) is used to connect the two sites.

> **Note:** Strictly speaking, the technology is wavelength division multiplexing (WDM), and coarse wavelength division multiplexing (CWDM) and DWDM are just different ways to use that technology. However, because this book is written in the context of qualified System z distance extension solutions, and all qualified WDMs (at the time of writing) are DWDMs, we use the terms *WDM* and *DWDM* interchangeably in this book. However, if we are specifically speaking about a CWDM device, we will use the term *CWDM*.

Assumptions

Although we acknowledge that there are many enterprises that still use older hardware (for example, parallel channel-attached check sorters), the reality is that those devices will be replaced over time with current technology, and current technology is used for the majority of devices.

For this reason, the examples in this document use technology that is current at the time of writing (for example, FICON Express8S rather than 2 gigabit (Gb) or 4 Gb FICON), while also bearing in mind that the infrastructure that you design should cater for future technology changes. Having said that, where appropriate, we also address the issue of older technology that must coexist with, and be supported by, the connectivity infrastructure.

1.2 Why you would have multiple data centers

There are many reasons why you might have (or plan to have) multiple data centers. This section will touch briefly on the more common ones. For more comprehensive information about data center resiliency, see the IBM Redbooks publications *IBM System Storage Business Continuity: Part 1 Planning Guide*, SG24-6547, and *IBM System Storage Business Continuity: Part 2 Solutions Guide*, SG24-6548.

One reason for having multiple data centers is that you need the ability to have a planned data center outage (in locations where the power supply is constrained or unreliable, for example) without affecting the availability of your applications. In that case, you can spread your production sysplex over the two sites, and use IBM HyperSwap® to move the primary disk back and forth between the sites. This can enable you to shut down either site without affecting application availability.

Such a configuration has the most stringent connectivity requirements, because of the bandwidth that is required between the sites and the broad range of devices that need to be connected to systems in both sites. Because applications will use resources in both sites, there are realistic limits to how far apart the sites can be located.

Another reason for a second site might be to have a DR capability. If all applications will run in just one site, you might want to place the second site at a large distance from the production site to minimize the risk of one event affecting both sites.

Placing the second site a large distance away minimizes the risk. However, it also increases the connectivity cost and increases the amount of data that is in-flight (and therefore must be recreated) in case of a disaster. Such a DR capability can be viewed as a form of insurance, and as with insurance, the less risk you are willing to accept, the higher the cost.

An additional possible reason is that you already have multiple data centers, and you want to derive more value from that investment. So, for example, you might have independent IBM TS7700 Virtualization Engine grids in two sites. To provide more resiliency in case of a site outage, you might decide to reconfigure the grids so that you have a grid that spans both sites, with the ability to keep mirrored copies of critical volumes in both sites.

An increasingly common reason for having more than one data center is to conform to government regulations. The global financial crisis of 2008 resulted in increased recognition of the vital societal role that financial institutions play. That realization has resulted in governments in many countries imposing more stringent DR requirements on businesses in certain industries.

These are just some of many possible reasons for having multiple interconnected System z data centers. Regardless of the business reason for your envisioned extended distance configuration, this book should help you.

1.3 Common data center models

Every System z environment is unique, and the continuous availability and DR requirements vary from enterprise to enterprise. However, broadly speaking, configurations typically fall into one of these categories:

Single data center All System z equipment is in one data center, and is within the supported maximum unrepeated distance from the CECs.

This environment does not have any extended distance connectivity requirements other than traditional SNA and Internet Protocol networks, and therefore is not the focus of this document.

Campus environment There is more than one System z data center, however all devices are within the supported distance from the CECs.

This configuration differs from the single data center in that it provides the possibility to maintain application availability across a data center outage. However, because all of the devices are within the supported distance, there are no extended distance connectivity requirements, and therefore this configuration is not the focus of this document.

Multisite sysplex There is more than one data center, the production sysplex spans more than one site, and the distance between the data centers is greater than the maximum supported unrepeated distance for the devices that need to be connected across the sites.

This category would also include situations where the data centers *are* within the maximum supported distance, but DWDM devices are being used for the connectivity between the sites.

This configuration *does* use extended distance equipment, so it is in the scope of this document. Because the production sysplex spans both sites, the connectivity infrastructure must support both CEC-to-device and disk-to-disk (mirroring) requirements and coupling links. These are used to connect the coupling facilities (CFs) and provide Server Time Protocol (STP) support.

Data mirroring There is more than one data center, and the distance between the data centers is greater than the maximum supported channel distance. However, the sysplex does *not* span both sites.

Host-to-device communication would generally be used in relation to a mirroring technology, reading from, and writing to, devices in the remote site.

This configuration also has extended distance requirements. This configuration is typically used together with some form of asynchronous mirroring technology to provide a DR capability at an alternative location that is independent of any events, natural or man-made, that would cause the primary data center to become unavailable.

Of course, there are variations on all of these configurations. You might use channel extension devices across sites that are only 10 kilometers (km) apart. Or you might use some combination (for example, a production sysplex spanning two sites, combined with disk mirroring to an out-of-region third site). Chapter 4, “Common multisite models” on page 113 goes into greater detail about these configurations. However, the important point is that this document will address connecting sites at both metro and extended distances.

1.3.1 Distances

You might be wondering, “What do they mean by *extended distance*? I have two sites that are *x* km apart: is that considered extended distance?”

Extended distance is quite a subjective term, and the meaning might change depending on the context. So, in this book we use the following terms to broadly describe different ranges of distances:

Extended distance Any distance greater than the maximum supported unrepeated distance for a device.

In the context of this book, extended distance also includes any configuration that is using a DWDM, even if the distance between the device and the CEC is within the maximum supported unrepeated distance.

Metro distance This term is also somewhat subjective. However, in the context of this book, we use it to define a distance that is less than the maximum supported distance for synchronous disk remote copy (which, at the time of writing, is 300 km). Even more imprecisely, you can think of it as any distance that requires some form of repeater, but is less than *out-of-region*.

Note that just because two data centers are within metro distance, that does *not* imply that they must be using synchronous mirroring. Metro distance is a term to describe *distance*, not the mirroring technology that might be used.

Global distance This is a new term, created by the writers of this document. It refers to any distance greater than metro distance.

Based on these definitions, your next question is probably, “Do I need extended distance equipment, or can I just connect all my devices directly to the CEC or an existing switch?” The maximum unrepeated distance that is supported between a CEC and a connected device depends on many things:

- The type of channel card in the CEC
- The device type
- The type of adapter in the device
- The channel technology (for example, the supported distance for a FICON channel depends on the channel bandwidth and whether it is using single-mode (SM) or multimode (MM) fiber)
- Whether a switch or director is being used
- The quality of the connecting link and the number of splices or connections

Information about the maximum supported unrepeated distance for various devices is provided in the IBM Redbooks publication *IBM System z Connectivity Handbook*, SG24-5444. We do not include information in this document for every device, because the maximum supported unrepeated distance varies over time. In addition, the Connectivity Handbook document is updated to reflect new technologies, and duplicating that information in this book would only lead to confusion.

If you determine that extended distance equipment *is* required to support your planned configuration, this document is intended to help you plan for your current and future needs.

Meaning of supported distance

IBM performs testing of all IBM devices that are designed to work with System z. The testing ensures that the interaction with the device conforms to the System z input/output (I/O) architecture. Part of the testing might involve connecting the device at distances beyond those that would be encountered in a normal data center. As a result of that testing, IBM determines the maximum supported distance for that device.

However, just because a device will function correctly at a given distance, does *not* necessarily mean that the service times that can be achieved at that distance would enable you to meet your application response time objectives.

Successfully delivering a business IT service requires a combination of a supported configuration that operates without errors *and* a configuration that delivers the service times that enable all of your applications to meet their service level objectives (SLO). It is the responsibility of your technical leaders to combine the information about what is supported with information about the relationship between distance and service times.

Relationship between distance and performance

When planning a multisite configuration, one of the things that must be considered is the performance effect of the distance between the sites, and the relationship between the workloads running in the connected sites.

If the second site only contains secondary disks and an asynchronous mirroring mechanism is used, the distance between the sites should not make any difference to the performance of the production systems (assuming that there are no bottlenecks in the mirroring infrastructure).

Alternatively, if synchronous mirroring is used between the sites, the distance between the sites *does* have a real effect on the service times experienced by the production systems. And if the production sysplex spans both sites, the effect of the distance on CF response times also needs to be taken into consideration.

The relationship between distance, service times, and application response times is a complex one. Depending on your requirements, your application design, and your database design, a distance that delivers acceptable response times for one enterprise (or even one application) might result in an unacceptably high response time for another.

This topic is covered in detail in 4.1.1, “Difference between continuous availability and disaster recovery” on page 114. However it is important to point out that IBM strongly suggests that any client that is considering implementing a multisite sysplex should perform a benchmark to determine the effect of the planned distance on their production environment prior to making a final decision.

1.3.2 Connectivity solutions

Readers that have been around the mainframe world for a long time might be familiar with the term *channel extender*. This was originally a stand-alone device that typically was used to connect tape drives or printers that were located beyond the maximum supported distance for a device (which was 400 feet with the original bus and tag parallel channel cables).

However, connectivity technology in this area has changed dramatically since those old channel extender days. The channel extension function is now likely to exist on a blade that will be a part of your SAN switch or director. The connectivity devices might support a variety of options for connecting across sites, and those options might cover distances from tens of kilometers up to thousands of kilometers.

Similarly, WDM technology is changing dramatically in response to the massive volumes of data that must be moved across global networks. Depending on the distance between the WDMs and the availability of dedicated fiber, different protocols might be used for connecting the WDMs.

Switches and WDMs provide functions that are unique to each, and they also provide some functions that are common across both devices. Therefore, when looking at your options for connecting two sites, it is not necessarily an “either/or” situation.

Depending on your situation and requirements, the solution might be a switch, or a WDM, or (more likely) a combination of the two. This document is intended to help you identify the most appropriate solution for *your* enterprise (one that will meet your needs today, and provide a framework that will support your needs into the future).

1.4 The importance of an end-to-end architecture

The System z end-to-end connectivity architecture is growing increasingly complex, and existing environments might not be built on current preferred practices or optimized for overall efficiency. This presents clients with a bewildering range of options when upgrading or extending old systems or installing new ones.

Much of this complexity comes from the fact that many data centers are relatively ad hoc in their structure. The System z multisite connectivity architecture has evolved incrementally over the years, and therefore was not engineered as one complete solution, such as a car or aircraft. Silos of different elements have evolved, often owned and managed by different teams or departments within your environment. These might be interconnected, making changes difficult and complex.

Although operationally the end-to-end connectivity infrastructure might be designed, built, managed, and maintained as a series of silos, at a high level it should be viewed as a single, seamless layer. As industry requirements grow, it will become even more important that this infrastructure is dynamic, responsive, scalable, and optimal across all components.

Figure 1-2 shows a high-level view of the components in the end-to-end connectivity infrastructure. As complex as this might appear, it does not even show another key set of components in the end-to-end architecture: the physical cabling infrastructure within the data center.

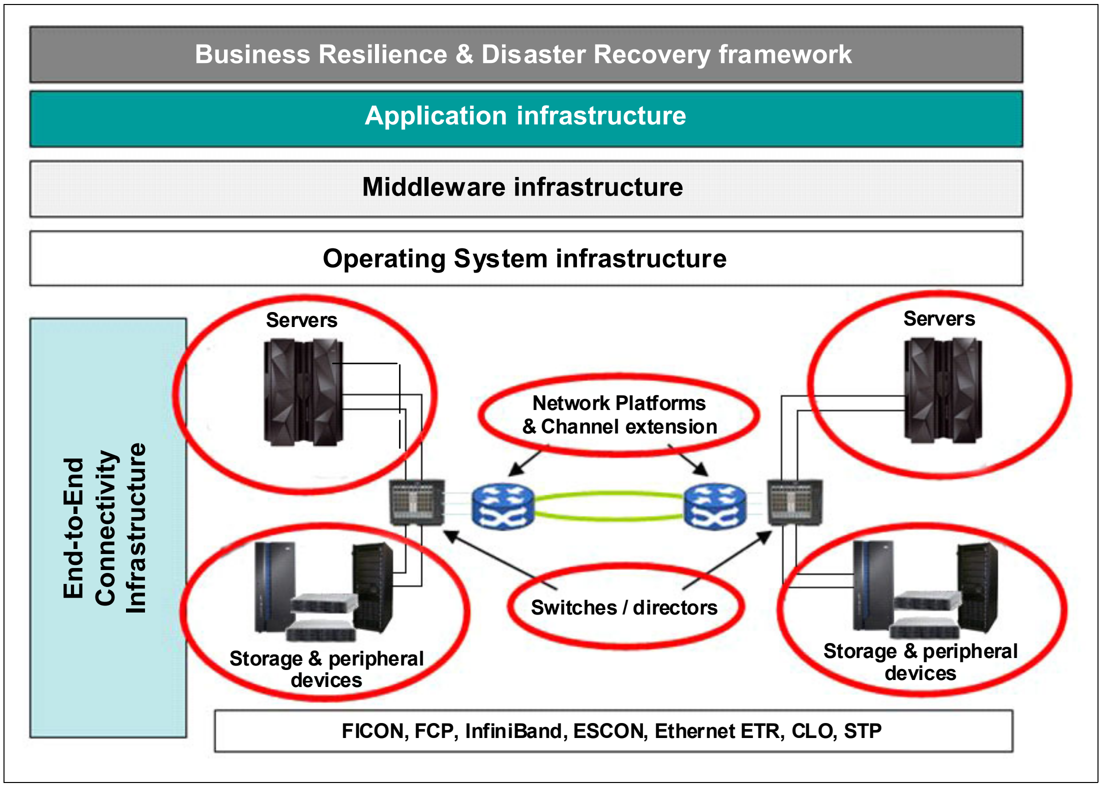

Business Resilience & Disaster Recovery framework
Application infrastructure
Middleware infrastructure
Operating System infrastructure
End-to-End Connectivity Infrastructure
Servers
Servers
Network Platforms & Channel extension
Switches / directors
Storage & peripheral devices
Storage & peripheral devices
FICON, FCP, InfiniBand, ESCON, Ethernet ETR, CLO, STP

Figure 1-2 System z end-to-end connectivity infrastructure

Figure 1-2 illustrates the core requirement for one complete solution, where all the moving parts must follow a single scope with core end-to-end credentials. This book provides detailed information about these credentials and attributes.

There are two key end-to-end connectivity attributes that must play a role in your end-to-end solution:

- End-to-end design

  Provision one complete architecture that includes all of the components under one umbrella, rather than having individual silos. Provide an end-to-end strategy for the complete solution to meet the business demands, and to use future connectivity and information technology (IT) infrastructure advances. Underpin the solution with end-to-end design attributes that include the following components:

  - Functionality
  - Manageability
  - Performance
  - Capacity planning
  - Documentation
  - Use of connectivity technology capabilities, now and in the future

  Without an end-to-end approach, the solution might be fractured or sub-optimal, where some components negate the benefits of others, or end-to-end functionality is not enabled or optimized. Changes to individual components might introduce problems or compromise the end-to-end solution.

- End-to-end interoperability

  Provide confidence in a tested solution. Provide an authoritative reference regarding the interoperability, metrics, tolerances, and components for each interconnected part of the end-to-end infrastructure. This facilitates further end-to-end attributes:

  - Qualified vendor support, with clear, contractual obligations
  - Maintenance capability that provides a usable framework to maintain interoperability and resolve field issues in a timely manner
  - Alignment to rigorously-tested architectural blueprints

  Without these attributes, the solution might be unsupported, or make it difficult to quickly identify the root cause of problems. If it is not formally supported, identifying a resolution to a field problem might be on a best-effort basis across multiple infrastructure vendors.

1.4.1 End-to-end support structure

You must understand the requirements for a complete end-to-end connectivity solution before you can establish a model to support the day-to-day technical and organization requirements. As a core part of the end-to-end design, you must consider the interconnects. These can be considered the piping between the components that make up the end-to-end connectivity solution.

Interconnection between technology silos

A basic portrayal of the inter-connects in a typical end-to-end connectivity solution across two data centers is shown in Figure 1-3.

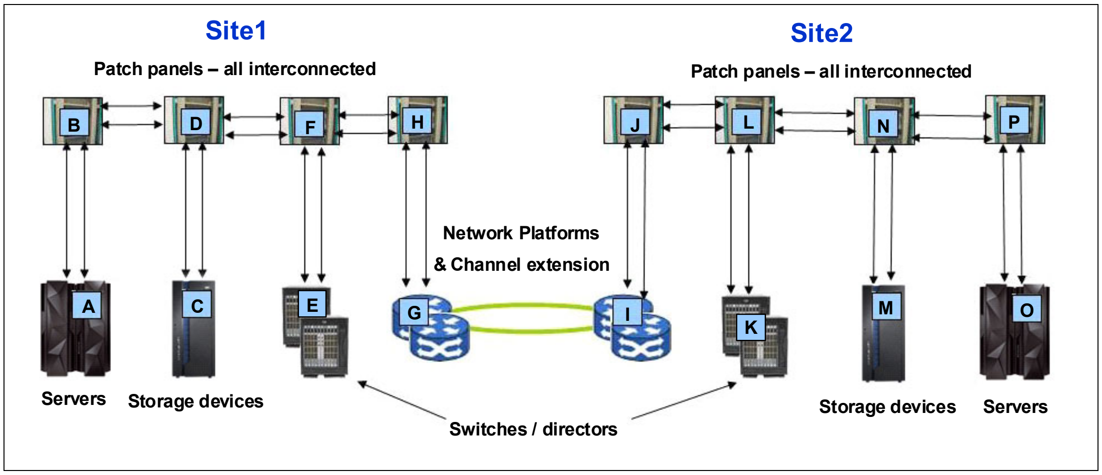


Figure 1-3 End-to-end connectivity inter-connects

Consider the end-to-end connectivity requirements for a FICON connection from a server in Site1 to attach to the storage in Site2. For the purposes of this illustration, we only describe the connectivity in Site1. In Figure 1-3, the components labeled *B*, *D*, *F*, and *H* in Site1, and *J*, *L*, *N*, and *P* in Site2 are all patch panels. The other components are the hardware boxes that are part of the configuration.

Site1 host connection inter-connects to a SAN

The following inter-connects provide Site1 host connection to a SAN:

- A FICON port at inter-connect A is cabled to the patch panel at inter-connect B. (In practice, multiple FICON ports should be connected to multiple patch panels, but we only show one here in the interest of simplicity.)
- The patch panel at inter-connect B is cabled to patch panel at inter-connect F to make the connection to the SAN switch for host connection.
- The host connection to the switch is completed with the inter-connect from patch panel F to the SAN switch at inter-connect E.

Site1 SAN connection inter-connects to a channel extension

The following inter-connects provide Site1 SAN connection to a channel extension:

- The host connection has a link address pointing to the inter-switch link (ISL) on this SAN switch. That link address represents a port on the SAN that is cabled to a different port in the patch panel at inter-connect F.
- The patch panel at inter-connect F is cabled to the patch panel at inter-connect H to make a connection to the DWDM.
- The SAN connection is completed with the inter-connect from patch panel H to the DWDM equipment at inter-connect G.

Site1 channel extension connection inter-connects to Site2

The channel extension inter-connect at patch panel H will be cabled into dedicated fiber (also known as *dark fiber*) or telecommunications equipment (not shown).

At this point, we have described the inter-connect requirements in Site1. The same requirements are now followed to complete the inter-connects in Site2, and to establish the path to the storage controller at M.

1.4.2 Support models

There are different aspects to how the configuration will be supported that must be considered when analyzing the requirements of the end-to-end connectivity solution:

Client support model This describes how your organization internally supports the IT infrastructure. This will typically consist of individual teams providing support within the scope of the particular technologies that they are responsible for.

Vendor support model This describes how the vendors collectively or individually are contracted to provide you with support for their individual components.

These support models might be optimal within each technology silo, but fractured when we look at the requirements for the complete solution and the inter-connects between the different silos. There might be numerous individual support teams working with their own specific vendors in isolation of each other.

To be effective, the support model must address the end-to-end infrastructure as one complete solution, not many individual optimized components. The connectivity solution underpins the IT infrastructure, which in turn supports the organization’s applications and business requirements.

This introduces some important questions that you should be able to answer. For example, consider how you would address a link failure in the end-to-end connectivity solution:

- Is this a supported solution, and has it been certified? If so, which parts are certified?
- Where is the problem in the end-to-end connectivity solution? Where do you look first?
- Who is responsible for addressing the link failure?
- Is there up-to-date, end-to-end documentation in place to support the connectivity infrastructure? Can you easily trace the physical cabling and inter-connect plugging through all optical connections from one end to the other?
- Who has responsibility for each part of the connectivity infrastructure?

End-to-end support models

To be able to comfortably address these queries, an internal support model must first be established. This must represent a single point of control and ownership that enables all departments with responsibilities within the end-to-end connectivity solution to work optimally together.

You should establish a framework with your vendors that defines clear ownership and maintenance processes throughout the connectivity solution. It might be optimal to establish an end-to-end maintenance contract providing a focal point of support across the complete connectivity infrastructure. The previous difficult questions should be addressed when putting such a contract together.

1.5 Role of the connectivity architecture group

A team that focuses on end-to-end solutions will create a less-fractured infrastructure that represents and conforms to a single architectural blueprint. The foundation of this infrastructure is the connectivity layer, where overlap is required across all of the technology disciplines. This connectivity layer underpins all of the other technology silos, so responsibility for, and ownership of, this layer *must be under the control of one team.*

This often represents a requirement for a new role with responsibility for all of the connectivity pieces. This does not necessarily need to be a new team of people. It might be additional responsibilities for an existing role, or even formalizing a role that already exists in a de facto manner. However, the title associated with this role is important, because the other teams must acknowledge who holds responsibility for the connectivity layer and the components within it. The control of all these pieces *must be in one place*, as shown in Figure 1-4.

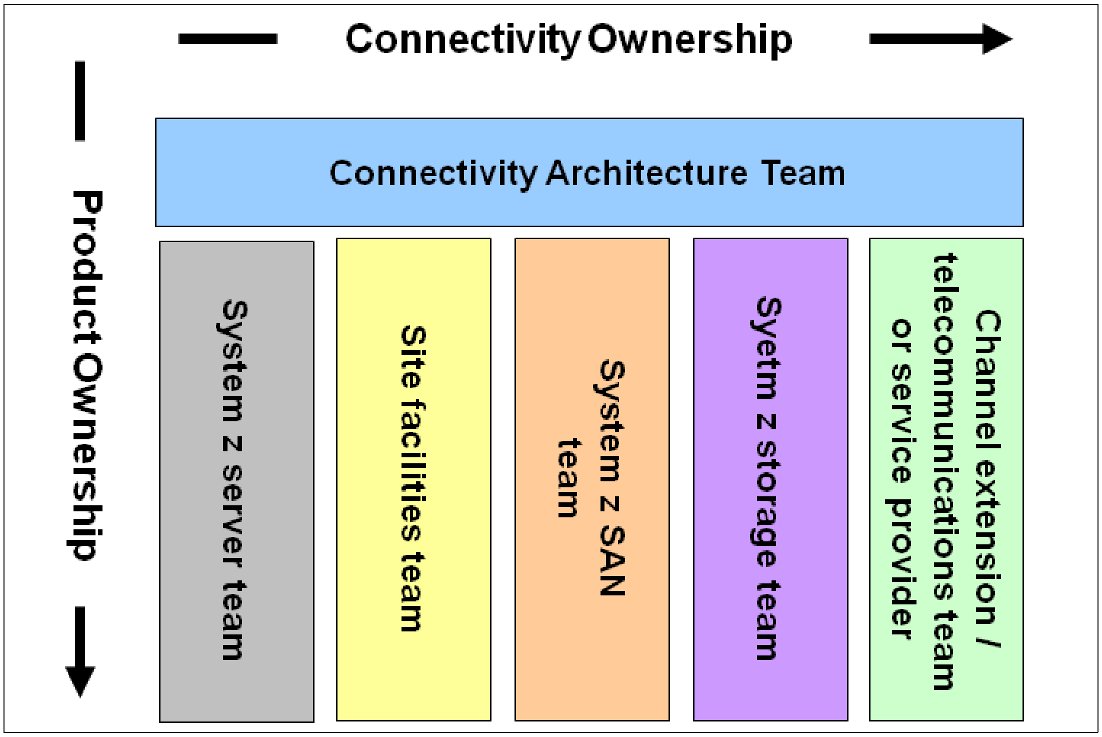


Figure 1-4 Organizational structure for end-to-end control

Figure 1-4 portrays an organizational structure where each individual team still holds responsibility for the optimization of their core technology. However, responsibility for the end-to-end components falls to the connectivity architecture team. This team should be educated, and able to serve as subject matter experts (SMEs) in the intersection points and complex interactions between servers, storage, and networking.

1.5.1 Scope of responsibility

The connectivity architecture team needs to be able to easily establish control of the end-to-end components within the connectivity layer. Each organization will have differing levels of responsibility within each technology silo, and there is likely to be an obvious choice as to which individuals or department take on this responsibility. The staff in this group should have sufficient knowledge of all of the moving parts to deliver an optimized and robust end-to-end connectivity architecture.

However, optimization of the individual silos (for example, storage, server, and physical planning) would still fall to the specialist areas. As an example, control of device adapter assignment might be under the control of the storage team, but the connectivity architect must have clear visibility of the individual port assignments.

Core competencies and responsibilities of the connectivity team are likely to include the following areas:

- Knowledge of the qualification status of the end-to-end components
- Vendor management for qualification, support, and maintenance
- Ownership and control of components in the connectivity layer:
  - System z server:
    - Port assignment on channel adapters (IBM Enterprise Systems Connection (ESCON®), FICON, Open Systems Adapter (OSA), and coupling)
    - Availability mapping of these channel adapters with the channel-path identifier (CHPID) mapping tool
  - SAN switches and directors (host, device, and ISL port assignment)
  - Extended distance equipment (assignment of ports and transponders)
  - Protocol converters (if appropriate):
    - Optica Prizm FICON to ESCON converters
    - Optica Prizm ESCON to Parallel converters
  - Storage (port assignment on device adapters)
  - Cabling:
    - Patching of all inter-connects
    - Adherence to link budget decibels (dB)
  - Optical amplification:
    - Choice of optics (short wave, long wave)
    - Selection of Fibre Channel (FC) providers
    - Ensuring links adhere to the correct transmit and receive levels

1.5.2 Configuration planning

The following list includes typical configuration planning-related responsibilities of the group:

- Input/output definition file (IODF) owning and managing
- Channel-path identifier (CHPID) mapping
- Cabling infrastructure management, including patch panels and trunk cables
- SAN port assignment for host, device, and ISLs
- SAN zoning, flow groups, ISLs, and trunks
- Extended distance equipment planning:
  - SAN ISL or Fibre Channel over IP (FCIP) equipment configuration
  - DWDM configuration
  - Channel allocation, optical amplification, attenuation, and protection

  > **Note:** This might be in the context of ensuring that the service provider provides this detail to the team.

- Port allocation across storage device adapters
- Protocol converters configuration management:
  - Optica FICON to ESCON converter
  - Optica ESCON to Parallel channel converter

1.5.3 Problem determination

An important responsibility of this team is the ownership of the connectivity layer from one end to the other. Control of all inter-connects should be such that a link can be easily traced from end to end.

Connectivity problems for any System z transport protocol become the responsibility of this team. The following list shows examples of such problems:

- FICON, ESCON, or CF link failure or interface control checks (IFCC)
- Peer-to-Peer Remote Copy (PPRC) link failures
- Intraensemble data network (IEDN) connectivity problems between the CEC and the IBM zEnterprise BladeCenter® Extension (zBX)

1.5.4 Performance and capacity planning

The connectivity layer has physical capacity, logical capacity, and performance requirements for which this team should have visibility and control. This team will need to work closely with other teams to understand the effect of link use on the core technology silos, and to ensure that sufficient link capacity is available to support the workload requirements.

These requirements should cover both existing and projected future volumes in the following areas:

- Cabling requirements:
  - Number of trunks and availability of unused channels within them
  - Patch panels and available slots
- Link requirements:
  - ESCON channel use
  - FICON channel use
  - FICON and Fibre Channel port use throughout the link
  - Coupling Facility subchannel and link use
- Network connections and OSA channel use
- Channel extension equipment requirements:
  - Bandwidth through FCIP equipment
  - ISLs on matching DWDM equipment

Important: New requirements driven by capacity or performance needs must be planned in an end-to-end fashion. For example, a change to a FICON card from 4 Gb to 8 Gb will not deliver all expected performance improvements if the other components in the end-to-end link remain at 4 Gb.

Whenever a component in the end-to-end connectivity layer is changed, ask the question: “What does this mean to the other components in the end-to-end configuration?”

1.5.5 Documentation

This team must be responsible for the full set of documentation related to the connectivity infrastructure. This will mean working with the teams that are responsible for any device connected to the infrastructure. It will also mean maintaining an up-to-date knowledge of the latest industry preferred practices, and the latest tools to create and maintain the required documentation.

1.6 What needs to be connected

The possible components in a System z configuration are many and varied. Although the focus of this document is on the components that help you extend the supported distance for a device, you obviously need to take into account which devices must be connected, and how those devices will be used.

The following list describes the types of devices, and functions associated with those devices, that typically exist in a System z configuration:

- CECs, also referred to as central processing units (CPUs) or processors
- CFs
- Sysplex time synchronization (either STP or IBM Sysplex Timer[1])
- Disk
- Tape or virtual tape
- Disk mirroring
- Tape mirroring
- Operating system (OS) consoles
- Hardware Management Consoles (HMCs)
- Printers
- Check sorters
- Card readers
- Channel-to-channel (CTC) adapters
- Remote job entry (RJE) equipment
- Encryption devices
- Network (through OSAs)
- Communications controller (IBM 3745 and similar)
- IBM 2074 Console Support Controller

All of these will be connected to a CEC using some type of channel (parallel, ESCON, FICON[2], or InfiniBand). The connection between devices (for disk mirroring, for example) will depend on the device and what type of connection it supports.

We cover specific planning actions in Chapter 5, "Planning" on page 125. For now, we just want you to start thinking about what System z-connected devices you have, and what their requirements are in terms of cross-site connectivity requirements.

For example, consider that you have a multisite sysplex with production systems running in both sites, and a check sorter in each site. In such a configuration, it might be acceptable to only connect the check sorters to CECs in the same site, because you have production z/OS systems running in both sites.

Alternatively, if all your production systems are in one site, and you have check sorters in your production site and your DR site, you might want to connect the sorter in the DR site to the CECs in the production site, so that it can be used in case the sorter in the production site is unavailable.

Hopefully most of your devices are covered by the IBM qualification process (which is included in 1.8, "System z qualification and testing programs" on page 23). Other devices might not be (typically, older devices that have been replaced by newer technology).

[1] Note that the IBM zEnterprise EC12 (zEC12) is the last generation of System z CEC that will support a mixed-mode timing network, and IBM z10™ was the last generation of System z CEC that supported connection to a Sysplex Timer.

[2] The most recent generation of System z CEC that supported parallel channels was IBM eServer™ zSeries 900 (z900). The last generation of System z CEC that supported ESCON channels was IBM zEnterprise 196 (z196) and IBM zEnterprise 114 (z114).

The move to a multisite topology provides an opportunity to replace those devices with modern equivalents. Support for devices that connect to ESCON or parallel channels might be provided by connecting those devices to an Optica Prizm device. For devices that are not supported by Optica, or are not included in the qualification process, contact the device vendor.

To provide you with more control and flexibility, you probably use patch panels, and possibly physical layer switches, as part of your fiber infrastructure. These help you create and maintain an environment that is easier to manage because you do not need to drag cables all over the place every time you need to make a change. They also provide the benefit of enabling you to change what is connected to what, with a minimal amount of disruption to the cabling infrastructure.

Every time you open a cable, you introduce the risk of dirt getting on the fiber optics. In addition, every time you move a cable, you introduce the risk of damaging the cabling, or creating a kink, knot, or unacceptable angle into the cable. The use of patch panels and physical layer switches enables you to make many changes without having to expose any cables to these risks.

Note: Patch panels are passive devices, and therefore are not included in the IBM qualification process.

Finally, either to provide additional flexibility or extended distance, you might have one or more of the following devices in the configuration:

- Directors or switches
- WDMs

The type of device that you use to provide connectivity over extended distances will depend to a large extent on the distance between the device and the CEC, and the type of devices that you need to connect. Other factors are the type of inter-site connectivity that is available in your country, and cost considerations. In the context of this book (extended distance connectivity), there is no effective minimum distance[3] limitation for any of the connectivity options.

However, there are maximum distance limitations. These maximum distance limitations vary by connected device type, the connectivity device, the firmware level of that device, and other factors. 1.7, “Connectivity options” on page 16 contains information to help you identify the options that are applicable to your environment.

1.7 Connectivity options

The devices that you use to provide your connectivity solution will depend on the distance between the sites, and the types of devices that you need to interconnect.

Determining supported distances

If you have spent time investigating extended distance options, you have probably observed that there are different supported distances for different devices. You may also have wondered where they come from, and why they are not the same for every device.

[3] For 2 gigabits per second (Gbps), 4 Gbps, and 8 Gbps Fibre Channel, the minimum cable length is 2 meters. For 10 Gbps and 16 Gbps, the minimum is 0.5 meters.

We will not go into the technical detail here (that will be covered later in this book), but there are, from a purely technical perspective, two things that determine the maximum supported distance between a pair of devices:

- Strength of the light signal

  The maximum link optical budget must be considered. The optical link budget is the difference between the maximum launch power of the transmitter and the minimum sensitivity of the receiver.

  The stronger the light source, the further the light can reach. If you increase the distance between the transmitter and the receiver, there is a point where the light could not be seen anymore.

  Alternatively, you cannot just boost the strength of the laser because of the damage that that could do to the receiver at the other end of the link. There are also limits on how powerful the laser can be. Furthermore, the vast majority of links are less than 10 km, so it would not be cost-effective for vendors to always provide devices with lasers that are far stronger than the majority of enterprises require.

- Buffer credits

  You might be familiar with the concept of buffer credits. They are described in detail in "How buffer credits work" on page 199. In simple terms, buffer credits are used to control the flow of frames back and forth across a link. The longer the link, the more frames must be in-flight at one time to keep the link fully used.

  However, there are realistic limits to the number of buffers provided. For one thing, each buffer requires physical space on the adapter, so that places limits on the number of buffers you can have. Additionally, there is no benefit to be had from having more buffers than are actually required. If enough buffers are provided to cover the needs of the majority of clients, providing more buffers would increase the cost of the device, with no corresponding benefit for most clients.

The strength of the laser and the number of buffer credits delivered in most devices is sufficient for a majority of installations, especially those with a single data center.

Other optical considerations are the type of fiber used (for example, SM or MM), the quality of the fiber, and the number of connections in the link (every connection results in a reduction in the strength of the signal).

For distances beyond those supported by the standard optics and buffer credits, you have to use devices (such as WDMs and switches) to provide the appropriate optical budget and additional buffer credits. Consideration must also be given to the type of devices you will be connecting. Different devices use different protocols and have different supported distances. Additionally, not all protocols are supported by all switches and DWDMs.

Important: There is no supported way to overcome manufacturers' distance limitations. For buffer credit-based distance limitations, you must use a device (for example, a switch) which provides the additional buffer credits needed for the required distance.

Example of link limitations

Imagine you want to connect your mainframe to a disk subsystem using a standard FICON Express4 adapter, and the distance between the data centers is 25 km.

From a flow control point of view, the FICON Express4 adapter includes 212 buffer credits. There is no need for additional buffer credits because the 212 credits provided are sufficient to support a 4 Gb link at 25 km.

Alternatively, a FICON Express4 adapter includes an SM optical interface that supports a maximum unrepeated distance of 10 km. To extend the signal out to 25 km, you will need some device to increase the signal strength. In this example, you can use a WDM system to overcome the optical budget limitation.

Now, assume that you purchase a new direct access storage device (DASD) subsystem that supports 8 Gb adapters. To get the full benefit from those higher-bandwidth adapters, you decide to also upgrade the CEC channels from FICON Express4 to FICON Express8 adapters.

Similar to the FICON Express4 adapter, the FICON Express8 adapter is limited to a maximum unrepeated distance of 10 km with its optical interface. This is OK because the DWDM can be upgraded to support 8 Gb adapters. However, the FICON Express8 adapter only provides 40 buffer credits per link. Based on the size of your frames, this is not enough to drive the channel at maximum throughput over 25 km.

Although the disk subsystem will operate with fewer than the optimal number of buffers, you will not be able to get the full performance benefit of the 8 Gb adapters. Therefore you have to address the buffer credit-based limitation by using a switch or director in a cascaded director configuration. The switch will provide enough buffer credits at its ISL port to enable you to drive the channel at full usage over 25 km.

This example shows how both switches or directors and DWDMs play a role in a multisite configuration. And, depending on your devices and how you want to use them, you might need both switches *and* DWDMs.

1.7.1 Connecting devices over an extended distance

This section is intended to give you an idea of what makes it possible to connect two or more devices over distance. Because most of the connection options are based on optical fiber, electrical or copper-based connections are not considered in this book.

> **Note:** This section provides a general overview of connection options. It does not provide information about the protocols used or the devices that are connected.

The distances mentioned in this section are theoretical maximum distances. In the case of a real environment, the actual maximum distance depends on the following factors:

- The interface type:
  - MM laser-based transmitter
  - SM laser-based transmitter
  - SM laser-based transmitter using a WDM wavelength
- Protocol-based limitations:
  - Limited buffer credits for Fibre Channel or FICON
  - Throughput considerations
  - Application response time requirements
- Loss and quality figures of the fibers you use, plus the effect of any connections in the link
- Link latency introduced by devices in the signal path such as:
  - WDM devices
  - Ethernet/IP switches or routers
  - Synchronous Optical Network (SONET)/Synchronous Digital Hierarchy (SDH) devices
  - Optical Transport Network (OTN) devices
  - Others

Note: For System z-based connections, all active devices in the signal path are subject to qualification testing. See 1.8, “System z qualification and testing programs” on page 23 or further information about qualified connectivity options.

The generic devices used in the following figures could be any kind of device. For this chapter, it does not matter if those are mainframes, disk, routers, or directors. This is a general overview of how you can connect devices.

Using a direct fiber cable

The most basic is a direct connection from one device to another device. This is shown in Figure 1-5.

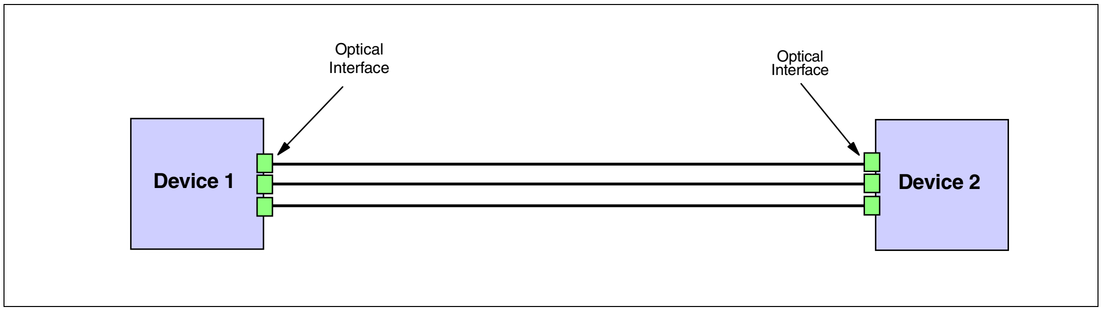


Figure 1-5 Direct optical device connection

In this example, the optical fibers are connected directly to the optical interfaces on both ends of the link. Interfaces using MM lasers can drive a signal up to 860 meters. Interfaces using SM lasers can drive up to 80 km unrepeated. In practice, no current IBM processors or control units (CUs) provide sufficient buffer credits or optical signal strength to work over such a large distance.

Using switches or directors

In practice, anyone connecting channel-attached devices over extended distance is likely to use a pair of SAN directors in a cascaded configuration. The reasons for this are provided at length in Chapter 2, “Storage area network” on page 39. For now we just concentrate on the connectivity options.

The traditional configuration using ISLs is shown in Figure 1-6.

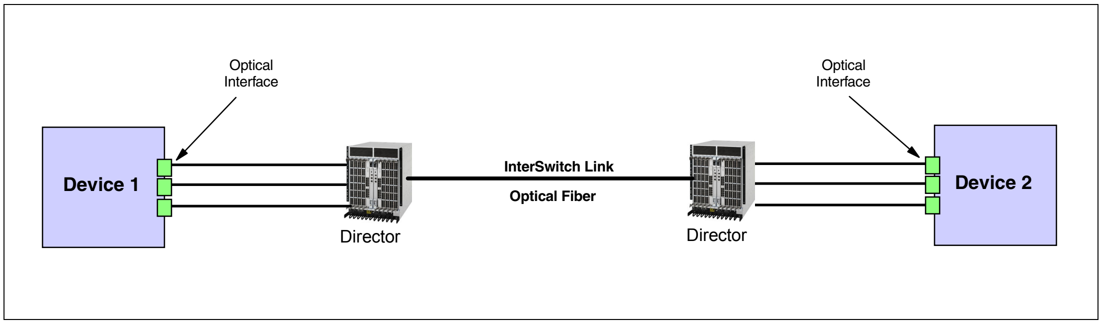


Figure 1-6 Connecting using cascaded directors

The ISL uses FC and requires dedicated fiber connections between the two directors. The maximum distance between the directors will depend on the particular model, and on the type of optics that are used for the ISL connections.

A more recent option is to again use directors, but in this case the directors are connected using FCIP. This configuration would not use dedicated links between the directors, but would instead use an IP network, which could be shared with other devices. Because dedicated fibers are not used, the distance between the directors can be virtually unlimited.

Naturally, however, there are practical limits placed on the distance by the type of devices that will be connected to the directors, and how they will be used. An example of this type of configuration is shown in Figure 1-7. This configuration is also covered in detail in Chapter 2, “Storage area network” on page 39.

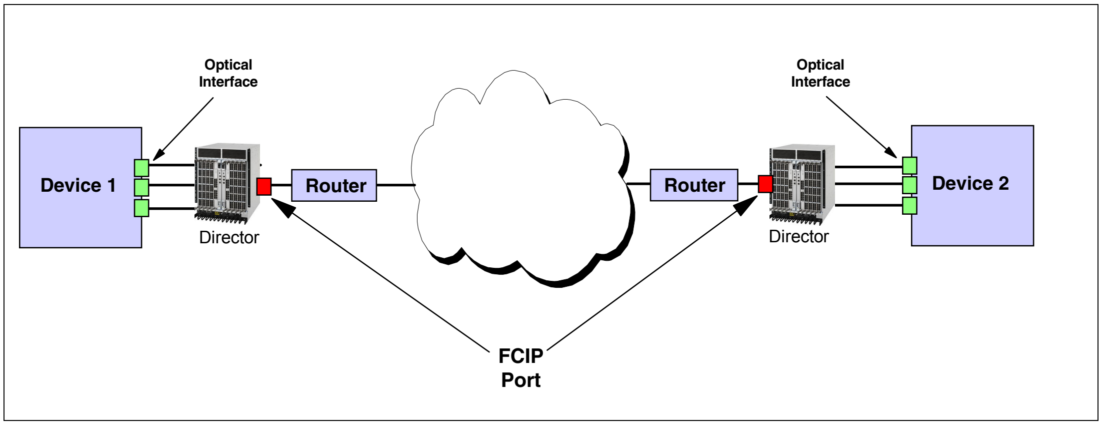


Figure 1-7 Connecting using cascaded directors and FCIP

Using WDMs

In all modern telecommunication networks, WDM is the underlying technology when it comes to transporting large volumes of data over large distances. Therefore, active WDMs are used for ultra-long-haul links, such as transoceanic links. However, active WDMs can also be used to connect data centers that are only a few km apart.

To make the following examples readable and easier to understand, only simple point-to-point networks are shown. These would be using fiber cables that are dedicated to your use (dark fiber). Most modern WDM devices can support all types of network topologies, such as ring or meshed networks, both fixed or reconfigurable. However, typically only point-to-point networks are used for connecting data centers, and only point-to-point topologies are qualified for use with System z.

The simplest way to use WDM technology to connect your devices is to have the devices directly connected to a WDM system, with the WDM systems being connected by optical fiber. This configuration is shown in Figure 1-8.

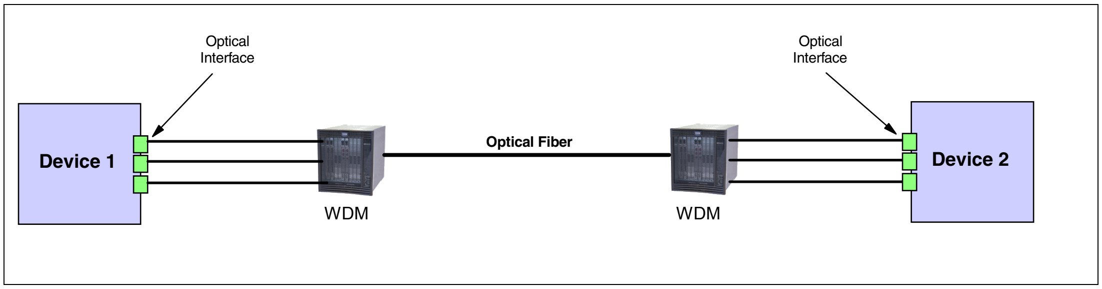


Figure 1-8 Connection using WDM

WDM technology is available in two versions:

- CWDM:
  - Supports a smaller number of lambdas[4]
  - Reaches up to 80 km
- DWDM:
  - Up to 160 WDM lambdas are possible
  - Reaches up to 200 km in a single span (Figure 1-8 on page 20)
  - Reaches up to several thousand kilometers, with multiple spans, using amplification and regeneration sites (Figure 1-9)

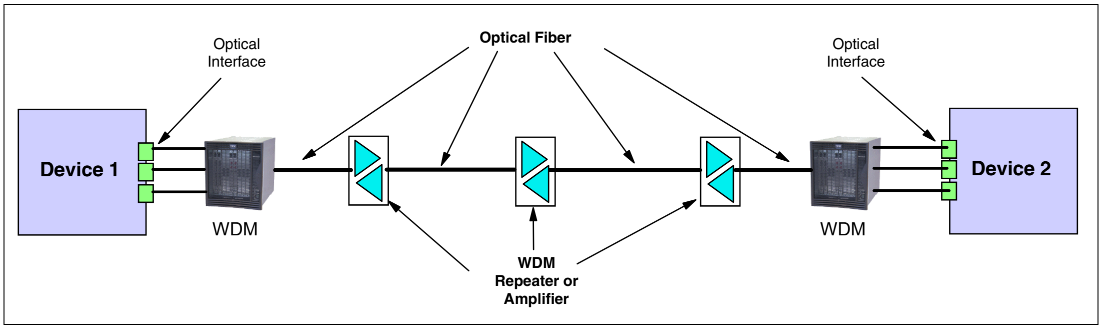


Figure 1-9 Connection using WDM with multiple spans

The requirement to have multiple spans might be a limitation for you, because the fiber has to leave the fiber duct for amplification or regeneration of the signals. At these locations, your data stream could be accessible (an exposure that might not be compatible with your company's data security policies).

Depending on the availability of dedicated optical fiber between the WDM boxes, you might have several options for how this network would be provisioned:

- You own, rent, or buy the optical fiber and the WDM boxes:
  - You can manage, operate, and maintain the WDM boxes yourself.
  - You have the freedom to add, change, or disconnect the services whenever you want.
  - You could outsource operation, maintenance, and management of the configuration.
- The fiber and the WDM boxes are owned by a telecom carrier or service provider:
  - You pay a monthly or yearly fee for the services you use on the WDM, possibly related to the volume of data that is sent across the network.
  - The carrier manages, operates, and maintains the WDM boxes with an agreed-upon service level agreement (SLA).
  - Changes to the network must be ordered and negotiated in advance.

Using SONET/SDH or OTN over WDM

For SONET/SDH-based or OTN networks, the whole transport infrastructure is typically owned and operated by one or more service providers or carriers. The same WDM technology and even the same WDM box as shown in the previous examples can also be used with SONET or OTN networks. However, the carrier will use their chosen hardware platform for offering transport services to you.

[4] A lambda is a specific light wavelength.

An outline of such a network is shown in Figure 1-10.

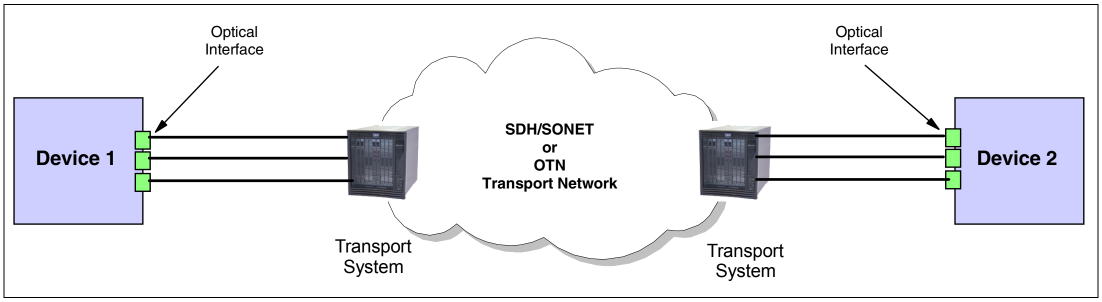


Figure 1-10 Connection using SONET/SDH or OTN-based networks

These networks are typically country-wide or continent-wide. Usually those networks transport your data with a specified delay for backup and protection paths. For security reasons, carrier networks are highly resilient in case of any infrastructure outage. Because these networks are typically designed for carrying network traffic, they might not be ideally suited to transporting storage or cluster protocols.

Important: The use of any WDM network topology other than point-to-point is not qualified by IBM.

For more information about SONET/SDH or OTN enhanced WDM technology, see Chapter 3, “Wavelength division multiplexing” on page 95.

Using WDMs with switches or directors

Another option is to use WDMs in combination with switches or directors. This configuration might be selected because you need to connect devices that are connected to switches, and devices that are not supported by switches (CFs, for example). Or it could be that the optics in the switches are not powerful enough for the distance that you need to connect over.

An example of this configuration is shown in Figure 1-11. In practice, this is one of the most common ways of connecting System z configurations over extended distances.

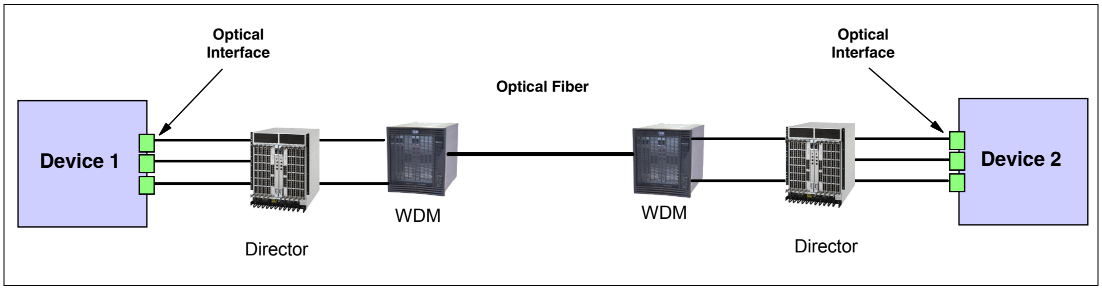


Figure 1-11 Connecting using both directors and WDMs

Using switched or shared networks

As with SONET/SDH or OTN Networks, switched networks, such as Multiprotocol Label Switching (MPLS) networks, are used by carriers or service providers for offering connectivity services to clients. The underlying transport network would also be based on WDM technology, often enhanced by OTN structures.

The internal bandwidth of a switched network is usually shared between several clients, as shown in Figure 1-12.

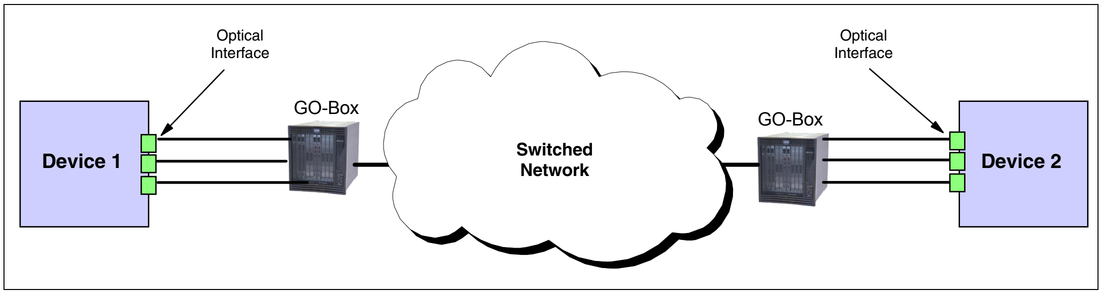


Figure 1-12   Connection using switched networks

The “GO-Box” in Figure 1-12 acts as a protocol converter. The protocol converter translates your protocol into a protocol accepted by the switched network. For example, a direct connection for FC or FICON services is usually not possible. For this kind of configuration, such protocols must first be converted to Ethernet/IP. In most cases, this will be an FCIP gateway or a director with FCIP capabilities.

Switched networks can span the globe. Most of the time you will get an Ethernet/IP service with a committed data rate and a committed availability from your carrier. However, delays in such networks are subject to change, and are usually not predictable.

More information about service provider-based networks is provided in 5.10, “Service provider requirements” on page 175.

Note: Some of those options might not be feasible for all connections (STP or coupling, for example) related to System z environments. However, recall that this section is just meant to be a general technical overview.

1.8 System z qualification and testing programs

IBM has several testing programs that can be used to check the functioning of various IBM and non-IBM equipment. There are different qualification and testing programs:

- System z connectivity testing:
  - DWDM vendor qualification
  - Switches and directors qualification
- The IBM System Storage® interoperability center
- Non-IBM storage vendors
- Testing performed by platforms other than System z

1.8.1 IBM System z connectivity testing

The IBM Vendor Solutions Center (VSC) Lab in Poughkeepsie, New York, US has the facilities and skills to test and qualify extended distance and storage products in various configurations, including the latest and previous generations of IBM hardware. The result of a successful series of tests is a *qualification letter* that describes the precise configuration that was tested, and which protocols were used in the tests.

It is important to understand that *Qualified* has a specific meaning in relation to extended distance and interoperability testing. *Qualified* does not mean *certified, supported,* or *approved*. It means that the product went through a strict and specific process to check the interoperability factors needed on a specific configuration for a specific hardware, software, or solution.

That does not mean that a non-qualified product will not work. But a configuration that has not been tested cannot expect, or have, the same level of confidence and support compared to a qualified configuration.

The qualification process tests IBM protocols and IBM devices with extended distance devices (switches and WDMs). However, it does not include testing of non-IBM intellectual property when used in conjunction with intervening devices. For example, ISL trunking is supported by some switch and WDM vendors. However, it is not included in the IBM DWDM qualification tests because the ISL trunking function belongs to the switch vendor, but that function then relies on support in the DWDM device.

You can imagine the huge number of potential combinations of devices, protocols, switches, WDMs, and functions that would have to be tested. If you want to use a function that is not IBM intellectual property, you should speak to your switch and WDM vendors. Some vendors have their own qualification processes specifically for this purpose, and they are designed to provide an equivalent level of confidence in the configuration that IBM qualification testing provides for IBM intellectual property.

Figure 1-13 shows a sample configuration in the VSC Interoperability lab environment.

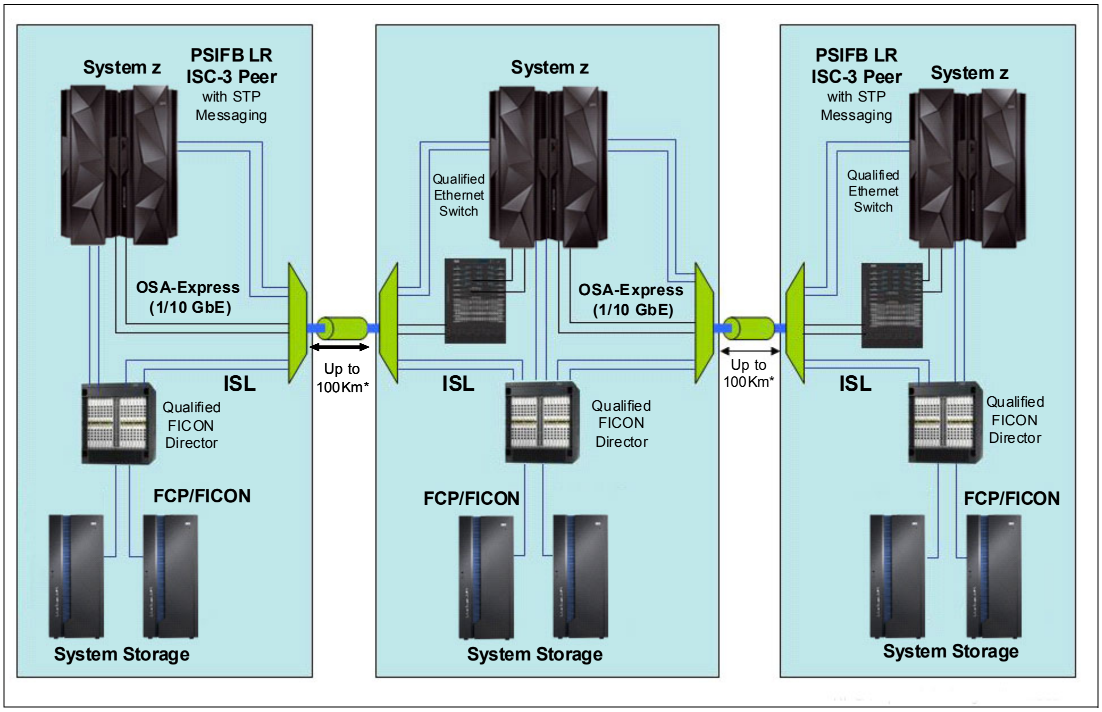


Figure 1-13 VSC Interoperability lab environment

The qualification testing includes a standardized set of tests specifically designed to ensure that the equipment being tested conforms to the standards of the protocols being tested. However, depending on the needs and market specialization of each vendor, not all protocols are necessarily tested. Additionally, the testing for WDMs is not the same as the testing for switches and directors.

The working relationship between the VSC staff and vendors, combined with the facilities of the VSC lab, allow IBM to reproduce any problems that might arise with this equipment in a client's environment.

Components

The following components are used during the qualification process:

- IBM Parallel Sysplex
- IBM System z Servers
- IBM System Storage
- Optical Wavelength Division Multiplexer (WDM)
- IBM System Storage Metro Mirror (PPRC)
- IBM System Storage Global Mirror
- IBM System Storage z/OS Global Mirror (extended remote copy, or XRC)
- IBM Ethernet products

Protocols

Table 1-1 show the protocols that are available for testing in the VSC. Not all protocols are tested and qualified for every device. To determine exactly which protocols were qualified for a given device, review the qualification letter for that device, available on the IBM Resource Link® website.

Table 1-1 Supported protocols

Protocol	Data transfer rates
Enterprise Systems Connection (ESCON)	200 Mbps[a]
Sysplex timer control link oscillator (CLO)	8 Mbps
External time reference (ETR)	8 Mbps
FICON	1 Gbps
FICON Express2	1, 2 Gbps
FICON Express4	1, 2, 4 Gbps
FICON Express8	2, 4, 8 Gbps
FC 100, 200, 400, 800, and 1600 megahertz (MHz)	1, 2, 4, 8, and 16 Gbps
ISL FC 100, 200, 400, 800, 1000, and 1600[b] MHz	1, 2, 4, 8, 10, and 16 Gbps
InterSystem Channel-3 (ISC-3) Compatibility Mode	1 Gbps
ISC-3 Peer Mode	2 Gbps
ISC-3 Peer Mode[c]	1 Gbps
STP (ISC-3 Peer Mode with STP message passing)	2 Gbps
STP (ISC-3 Peer Mode with STP message passing)[b]	1 Gbps
Parallel Sysplex InfiniBand Long Reach (PSIFB LR) 1x IB-single data rate (SDR)	2.5 Gbps
PSIFB LR 1x IB-double data rate (DDR)	5 Gbps
STP (PSIFB LR 1x IB-SDR with STP message passing)	2.5 Gbps

Protocol	Data transfer rates
STP (PSIFB LR 1x IB-DDR with STP message passing)	5 Gbps
Ethernet	1, 10 Gbps

a. Effective channel data rate of an ESCON channel is affected by distance.
b. Intermix of FICON and Fibre Channel Protocol (FCP) traffic on an ISL link is supported. But see 2.9.1, "Mixing FICON and FCP in the same fabric" on page 87 for more information about this configuration.
c. Requires Request for Price Quotation (RPQ) 8P2197. This RPQ provides an ISC-3 Daughter Card that clocks at 1.062 Gbps in peer mode.

The qualification letter for every device that is qualified is available on the Resource Link website. Although some vendors might also make a copy of the letter available on their own website, the superset of all qualified devices is available on Resource Link. If there is no qualification letter for a given device, that device has not yet passed the qualification process.

Important: The IBM VSC lab only qualifies IBM System z servers, IBM storage subsystems, switches, directors, DWDMs, and other vendor's storage devices as requested by that vendor.

The qualification process is always specific. For example, it lists the System z machine type, product model, part number for specific protocols, firmware level, and operating systems level. If anything changes in the configuration and is not covered by the letter (for example, a new machine model for System z, a new part number from the WDM vendor, or a new firmware level), a new qualification process must be completed to obtain a new qualification letter for the new configuration.

WDM vendor qualification for GDPS

The formal name of the WDM qualification program is "System z Qualified Wavelength Division Multiplexer (WDM) products for GDPS solutions". This is partially because such devices are typically part of a GDPS configuration. It is also because many of the early implementers of multisite sysplexes were GDPS clients. However it is important to understand that the qualification process applies to *all* multisite sysplexes, regardless of whether the IBM GDPS solution offering is being used or not.

The VSC documentation sometimes uses the term *GDPS* to refer generically to a sysplex that is spread over more than one site (a geographically dispersed parallel sysplex), and sometimes it uses it to refer specifically to the IBM GDPS solution offering. However, from the perspective of cross-site connectivity, the terms can be used interchangeably. These tested protocols are used in non-GDPS environments as well. For more information about the GDPS solution offering, see the following website:

http://www.ibm.com/systems/z/advantages/gdps/index.html

IBM only supports WDM products qualified by IBM System z for usage in GDPS solutions. To obtain this qualification, WDM vendors obtain licensed IBM patents, knowledge, and intellectual property related to the System z and coupling architecture. This gives them access to IBM protocols and applications used in a GDPS environment (including STP, Sysplex Timer, ISC, IFB, Metro Mirror, Global Mirror, and z/OS Global Mirror).

To qualify a WDM solution, the device being tested must support the following elements:

- At least one type of CF link
- At least one storage channel technology

Qualified vendors normally license this information for an extended period of time, allowing them to subscribe to the latest GDPS architecture changes. These vendor products have been tested and qualified by IBM using the same laboratories and procedures used to test all aspects of a GDPS environment.

This testing includes functionality, recovery, and in some cases performance measurements. Having access to these test facilities enables IBM to configure a fully functional sysplex, and simulate failure and recovery actions that could not easily be tested as part of a working client environment.

The latest list of qualified WDM vendors can be found on the Resource Link website, listed under **Hardware products for servers** on the Library page:

https://www.ibm.com/servers/resourcelink/

The IBM Redbooks website also contains IBM Redpapers™ publications about the qualified WDM devices. These Redpapers publications are based on the qualification letters and contain more detailed information about the process:

http://www.redbooks.ibm.com

Switches and directors qualification

Together with the respective vendors, IBM System z performs connectivity and interoperability testing of FICON and FCP switches and directors to ensure that products adhere to the FICON architecture and FCP architecture. Testing is also performed to ensure that the products support High Performance FICON for System z (zHPF), the intermix of FICON and FCP in the same switch or director, and the cascading of switches or directors.

The link data rates might be 1, 2, or 4 Gbps auto-negotiated using System z FICON Express, FICON Express2, and FICON Express4 channels, and 2, 4, and 8 Gbps auto-negotiated using FICON Express8 channels (exclusive to IBM System z10® and later CECs). ISLs between switches or directors can operate at link data rates of 2, 4, 8, 10, or 16 Gbps, depending on the optics used for the ISL ports.

Connectivity testing is also designed to help ensure switches and directors inter-operate with System z operating systems, such as z/OS, IBM z/VM®, IBM z/VSE®, IBM z/Transaction Processing Facility (z/TPF), and Linux on System z distributions.

Extended distance testing is also performed, and can include unrepeated distances up to 10 km and repeated distances up to 300 km (depending on the product and protocol):

- Intermix. Enables intermixing of FICON and FCP channels within the same physical switch or director. This enables a single director to be shared on a port-by-port basis between FICON-capable servers and devices and FCP-capable servers and devices. Intermix can help increase asset usage. If you have chosen to implement small director footprints which can scale over time, this might be of assistance with that strategy. For more information, see FICON or FCP white papers available from your director vendor.
- Cascaded directors. An ISL between two directors enables the directors to be interconnected, or *cascaded*. Cascading can help minimize the number of channels and cross-site connections, thereby reducing implementation costs for DR and business continuity solutions. System z support for cascaded directors is currently limited to a two-director, single hop configuration.

> **Tip:** Cascading directors can also help to reduce fiber optic cabling infrastructure costs, and can allow for shared links to better use inter-site-connected resources with reduced complexity.

The latest list of qualified switch vendor products can be found on Resource Link at the following website, listed under **Hardware products for servers** on the Library page:

https://www.ibm.com/servers/resourcelink/

Non-qualified WDM and switches vendors

There might be other vendors that do not go through the qualification process with IBM. They might support the same protocols and devices used in the qualification process, but you should consider that such configurations will not have the same level of support from IBM as a qualified solution.

Important: The qualification process is thorough. It provides IBM and the vendors that go through it with a higher level of knowledge and confidence about the solution. It also means that if those products are used in conformity with the qualification letters they will have full support from IBM and from the vendor.

1.8.2 System Storage Interoperation Center

For IBM System Storage, there is a website called System Storage Interoperation Center (SSIC):

http://www.ibm.com/systems/support/storage/ssic/interoperability.wss

This website lists all of the interoperability possibilities between IBM storage products with IBM and selected vendor products. The site enables you to interactively create your own interoperability matrix by selecting from different options:

- Product family, such as IBM System Storage Enterprise Disk
- Host platform, such as IBM System z
- Connection protocol, such as FICON
- Server model
- Operating systems
- Adapter, such as host bus adapter (HBA), converged network adapter (CNA), and so on
- SAN products
- Other options, such as switch module, clustering, multi-pathing, storage controller, and intercluster SAN router

All IBM products listed on that website are either qualified or tested and have IBM support. It also indicates support for interoperability. For example, if you select a specific disk subsystem, all of the other products that are listed, such as switches or servers, will be supported for interoperability.

The independent software vendor (ISV) Solutions Resource Library is also available on the SSIC website:

http://www.ibm.com/systems/storage/solutions/isv/index.html

1.8.3 Non-IBM storage vendors qualification

Non-IBM storage vendors can be either qualified or tested by IBM, or have their own qualification and testing program.

The IBM-qualified vendors are listed on the following website:

http://www.ibm.com/systems/z/advantages/gdps/qualification.html

For vendors not listed on this website, you should contact their sales or service representative to obtain information about their qualification and testing programs. Other vendors also have their own qualification and testing programs.

1.8.4 Other platforms

An end-to-end solution might have other components, such as UNIX, Linux, and Windows servers running application gateways, performing protocol transformation, and acting as a front end to the mainframe. These components also rely on a network, various protocols, and SAN devices to interoperate.

The following list includes components not covered by the IBM Poughkeepsie testing and qualification process:

- Ethernet Switches
- Routers
- IBM servers other than mainframes, such as IBM System p® and IBM System x®
- Non-IBM servers

The lack of specific qualification from those products needs to be considered. This could affect your decision about which devices can be connected to your extended distance solution.

Not all vendors test or qualify their products and end-to-end solutions. One of the reasons that such testing might not be performed is because of the number of options available. If a solution contains products not covered by IBM qualification or tests, it is necessary to contact the specific vendors about their status regarding interoperability at extended distances. The vendor might have internal references or other clients using the same configuration.

Another way to obtain support for these devices might be to negotiate a services contract with the vendor.

1.8.5 Standards for cables and connectors

Cables and connectors are not qualified by IBM. There is already a rigorous set of standards, testing methods, and criteria specified to meet or exceed the standards. A reputable vendor will provide a list of the standards that each of their products can meet. These standards specify cables, connectors (including patch panels), and procedures for communications and power using copper or fiber.

It is important to use a reputable vendor, and equipment that meets or exceeds the requirements. When selecting this equipment, it is important to remember that it will be in place for many years, and can be expensive and disruptive to replace, so consider your current *and* future requirements when installing new cables and connectors.

1.9 Planning for the future

This section provides an overview of the data center networking considerations for the future, based on information available at the time of writing.

Although the majority of this information is specific to the access layer and core data center networking architecture, there is clearly overlap that needs to be considered when designing an end-to-end System z connectivity infrastructure.

Most importantly, the future data center strategy clearly represents a strong requirement for a seamless, dynamic, end-to-end capability, where all components must be integrated and optimized together. The System z end-to-end connectivity model must represent these core values, and integrate with the data center networking strategy where the technology components inter-connect.

1.9.1 The evolution of data center networking

The requirements of the data center network are already stretching the limits and architectures that have evolved within traditional network designs. New requirements are evolving, as shown in Figure 1-14.

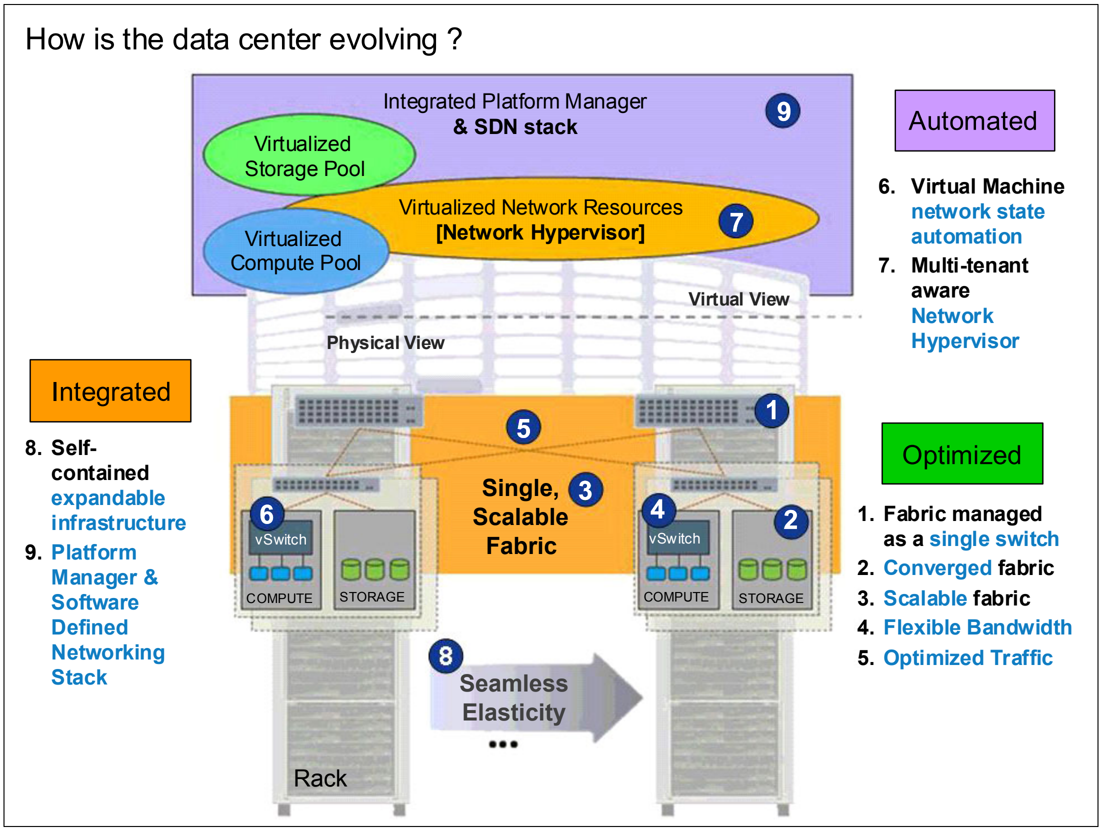


Figure 1-14 The evolution of the data center network

The use of many network tiers, with the associated cumulative latency, can significantly degrade performance. At the same time, proprietary functions and features restrict the ability to meet demands for a dynamic infrastructure. The accelerating pace of innovation demands a flexible data center network with new core credentials:

Optimized System scalability, flexibility, and simplification across a common infrastructure. Provide higher levels of performance, usage, and availability.

Automated Virtualized compute and storage platforms use hybrid computing models. Software-Defined Networking (SDN) virtualizes network resources using open standards, and supports self-configuring capabilities.

Integrated Converged physical infrastructure. Flatter, with a single point of control. Streamlined discovery, management, provisioning, change, and configuration management. Integrated problem resolution and reporting of servers, networking, and storage resources across the enterprise.

1.9.2 Traditional data center network architecture

The traditional data center architecture and compute model is shown in Figure 1-15. Historically, Ethernet was first used to interconnect *stations* (dumb terminals) through repeaters and hubs. Eventually, this evolved into switched topologies for a campus network, which came to form the basis for traditional data center networks.

Conventional Ethernet data center networks are characterized by access, aggregation, services, and core layers, which can have three, four, or more tiers. Data traffic flows North-South from the bottom tier, up through successive tiers as required, and then back down to the bottom tier, providing connectivity between servers. Network management is centered in the switch operating system. Over time, this has come to include a wide range of complex and often vendor-proprietary features and functions.

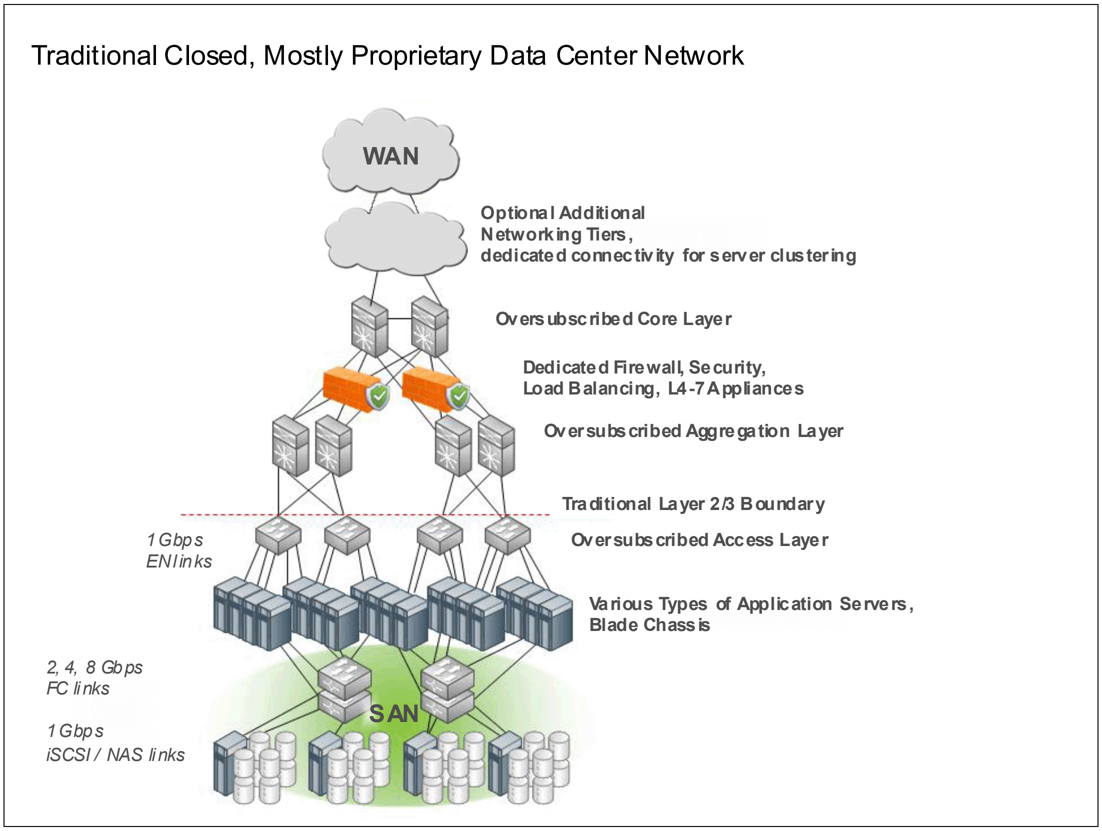

Traditional Closed, Mostly Proprietary Data Center Network
WAN
Optional Additional Networking Tiers, dedicated connectivity for server clustering
Oversubscribed Core Layer
Dedicated Firewall, Security, Load Balancing, L4-7 Appliances
Oversubscribed Aggregation Layer
Traditional Layer 2/3 Boundary
Oversubscribed Access Layer
1 Gbps EN links
Various Types of Application Servers, Blade Chassis
2, 4, 8 Gbps FC links
1 Gbps iSCSI / NAS links
SAN

Figure 1-15 Traditional network design

There are many problems with applying conventional networks to modern data center designs:

- Scalability

  Conventional networks do not scale in a cost-effective or performance-effective manner. Scaling requires adding more tiers to the network, more physical switches, and more physical service appliances.

- Installation and maintenance

  This physical compute model requires both high capital expense and high operating expense. The high capital expense is due to the large number of under-used servers and multiple interconnected networks.

- Dynamic infrastructure

  Modern data centers are undergoing a major transition toward a more dynamic infrastructure. This enables the construction of a flexible information technology (IT) capability that supports the optimal use of information to support business initiatives.

  As part of the dynamic infrastructure, the role of the data center network is also changing in many important ways, causing many clients to re-evaluate their current networking infrastructure. Many new industry standards have emerged and are being implemented industry wide. The accelerating pace of innovation in this area has also led to many new proposals for next generation networks to be implemented within the next few years.

- Fabric convergence

  The industry has slowly begun moving toward the convergence of fabrics, which used to be treated separately. This includes the migration from FC to Fibre Channel over Ethernet (FCoE), and the adoption of Remote Direct Memory Access (RDMA) over Ethernet standards for high performance, low latency clustering.

1.9.3 The future of data center networking

The Open Data Center Interoperable Network (ODIN) is a set of technical briefs that describe preferred practices for developing a flat, virtualized, converged data center network based on open industry standards. ODIN is intended to provide a set of best practices for data center networking to save both capital and operating expense by suggesting ways to achieve scalability, latency, high availability, convergence, security, and energy efficiency without sacrificing adoption of high performance technologies.

Figure 1-16 illustrates many of these attributes.

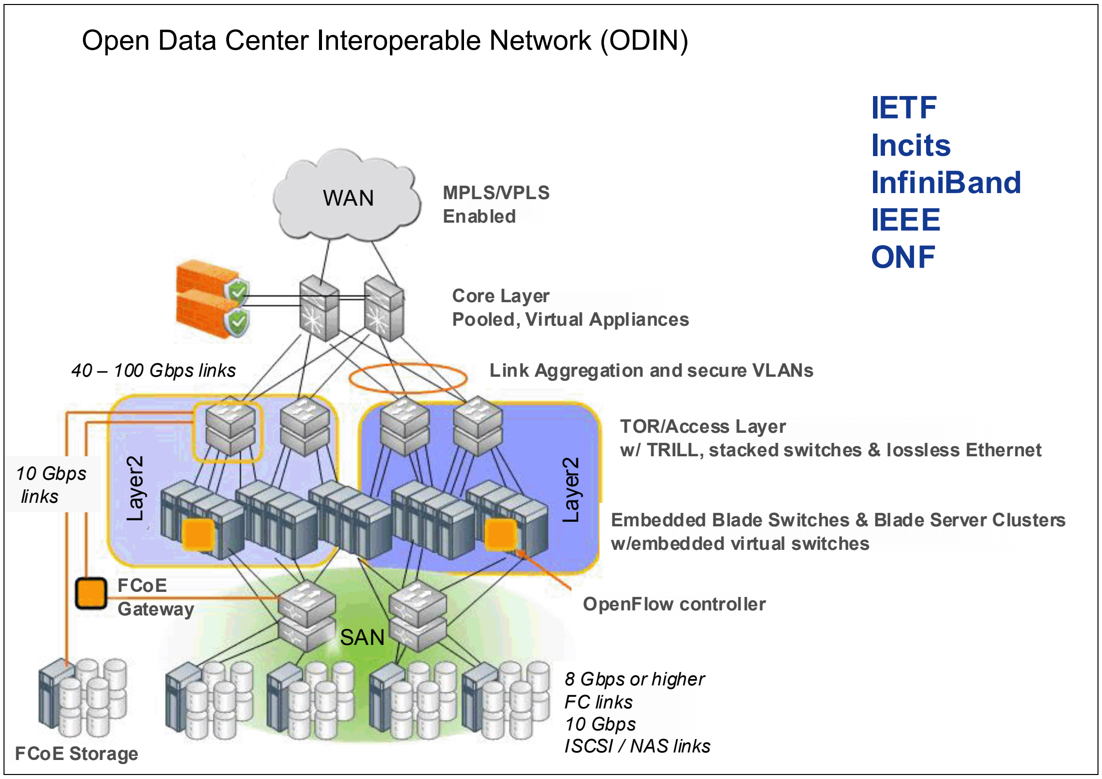

Open Data Center Interoperable Network (ODIN)
IETF
Incits
InfiniBand
IEEE
ONF
WAN
MPLS/VPLS
Enabled
Core Layer
Pooled, Virtual Appliances
40 – 100 Gbps links
Link Aggregation and secure VLANs
TOR/Access Layer
w/ TRILL, stacked switches & lossless Ethernet
10 Gbps
links
Layer2
Layer2
Embedded Blade Switches & Blade Server Clusters
w/embedded virtual switches
FCoE
Gateway
OpenFlow controller
SAN
8 Gbps or higher
FC links
10 Gbps
ISCSI / NAS links
FCoE Storage

Figure 1-16 An ODIN configuration

Some of the key attributes of a next-generation network architecture are described here.

Flattened, converged networks

Classic Ethernet networks are hierarchical, with three, four, or more tiers (such as the access, aggregation, and core switch layers). In order for data to flow between racks of servers and storage, data traffic needs to travel up and down a logical tree structure. This adds latency, and potentially creates congestion on ISLs.

Flattening and converging the network reduces capital expense through the elimination of dedicated storage, cluster and management adapters and their associated switches, and traditional networking tiers.

Virtualized environments

In a data center, virtualization (or logical partitioning) has important implications on the network. All elements of the network can be virtualized, including server hypervisors, network interface cards (NICs), converged network adapters (CNAs), and switches.

Today, there is a trade-off between virtualization and latency, so that applications with very low latency requirements typically do not make use of virtualization. The state of the network and storage attributes must be enabled to move with the virtual machines (VMs). This type of automatic VM migration capability requires coordination and linkages between management tools, hypervisor, server, storage, and network resources.

Scalability

The fundamental property of scalability is defined as the ability to maintain a set of defining characteristics as the network grows in size from small values of *N* to large values of *N*.

Although it is certainly possible to scale any network to very large sizes, this requires a brute-force approach of adding an increasing number of network ports, aggregation switches, and core switches (with the associated latency and performance issues).

Scalability is a prerequisite for achieving better performance and economics in data center networks. Scalability can be facilitated by designing the network with a set of modular hardware and software components. Ideally, this permits increasing or decreasing the network scale while traffic is running (sometimes called *dynamic scalability*).

Network subscription level

Traditionally, enterprise data center networks were designed with enough raw bandwidth to meet peak traffic requirements, which left the networks over-provisioned at lower traffic levels. In many cases, this meant that the network was over-provisioned most of the time. This approach provides an acceptable user experience, but it does not scale in a cost effective manner. New approaches to network subscription levels must be considered based on flatter design, any-to-any connectivity, and bandwidth sharing (or *fairness*).

Latency

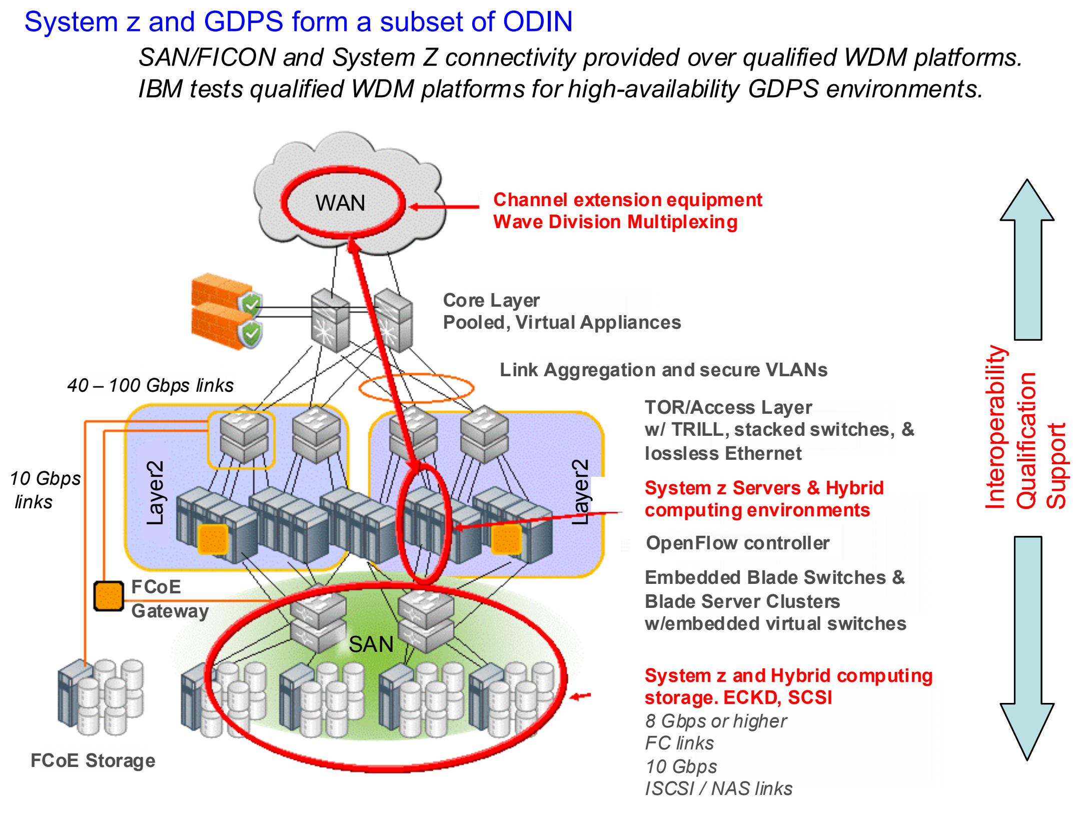
Latency refers to the total end-to-end delay within the network, due to the combination of time of flight and processing delays within the network adapters and switches. New data center networks must adhere to low latency characteristics.

Higher data rates

Anticipating the regular increases in both network equipment capability and application demand for more bandwidth, networks should be designed with future usage considerations, such as the ability to accommodate higher data rates with nondisruptive upgrades and minimal changes to the hardware. Network infrastructures designed for lower data rates might not scale to higher rates without hardware upgrades, making data rate another consideration for cost effective scalability.

1.9.4 System z and ODIN

Now that we have introduced strategic details of how the data center network can evolve, we must consider what this means to the System z end-to-end infrastructure model. Figure 1-17 on page 35 illustrates System z participation in the end-to-end data center architecture, incorporating ODIN practices.



System z and GDPS form a subset of ODIN

SAN/FICON and System Z connectivity provided over qualified WDM platforms.
IBM tests qualified WDM platforms for high-availability GDPS environments.

Figure 1-17 System z and ODIN (future thinking)

The data center network cannot be developed in isolation from the System z end-to-end connectivity infrastructure. There are overlaps and usage opportunities, such as fabric convergence and the extension of hybrid computing technologies.

The approach to maintaining the end-to-end connectivity model will not change. They must still mandate the same levels of support, interoperability, qualification, and control. However, there are extensions to this model that we need to understand and embrace as the wider strategic architecture evolves.

There are many problems and challenges with traditional, evolutionary network design. The network has too many tiers, and the cumulative latency can significantly degrade performance. Furthermore, conventional networks are not optimized for VM migration, east-west traffic patterns, highly virtualized and dynamic servers, or multi-tenant cloud computing applications. All of these problems contribute to high capital and operating expenses.

Modern data center networks should be characterized by two-tier designs, and cost-effective scale-out to thousands of physical ports supporting tens of thousands of VMs, without massive over-subscription.

These networks should be designed to support VM mobility, east-west traffic patterns, automated configuration management from attached servers or network controllers (wiring the underlying physical network only once), arbitrary topologies with switch stacking and link aggregation, and opportunities to use new technologies, such as RDMA networks or FCoE storage.

These networks should be based on best practices from open standards developed by the Institute of Electrical and Electronics Engineers (IEEE), Internet Engineering Task Force (IETF), International Committee for Information Technology Standards (INCITS), InfiniBand Trade Association (IBTA), Open Networking Foundation (ONF), and other groups, which are implemented by multiple vendors to provide lower total cost of ownership (TCO).

In an effort to address these problems, IBM has proposed a reference design based on open industry standards, as outlined in Figure 1-16 on page 33. System z will embrace this architecture, and maintain interoperability and new qualified GDPS architectures as part of this journey, as illustrated in Figure 1-17 on page 35.

More detail, updates to the ODIN documentation, and reference material can be found on the IBM System Networking website:

http://www.ibm.com/systems/networking/solutions/odin.html

1.10 Where to go for help

IBM Global Technology Services® (GTS) has a suite of services to help you assess, design, implement, and manage your end-to-end connectivity infrastructure. These offerings can be tailored to address the following areas:

- Connectivity strategy, assessment, optimization, and integration services
- IT and business solutions to provide expertise, proven methodologies, industry-leading management platforms, and processes
- Strategic partnerships with other industry leaders to help you create an integrated environment that drives business flexibility and growth

Contact your IBM representative for further details.

IBM Lab Services (LBS) is a worldwide team of IBM consultants who can provide the following services:

- Assist with all aspects of solution implementation
- Assess systems for vulnerabilities and mis-configurations:
  - Health checks
  - Fit for Purpose
- Optimize IT infrastructures:
  - Operating system and product configuration and tuning
  - Data center planning, creation, and evaluation
  - Connectivity assessment, planning, installation, and configuration
- Accelerate adoption of new technologies:
  - Jumpstarts
  - Starter kits
  - Production assistance
- Provide skills transfer:
  - Training
  - Classes
  - Seminars and Conferences

Go to http://www.ibm.com/systems/services/labservices or contact your IBM representative for further details.

Useful literature

In addition to this document, there are several other sources of information that can be helpful to you when designing and implementing your end-to-end architecture:

- A list of IBM Redbooks documents that are related to this topic is contained in "Related publications" on page 247.
- For the latest list of WDM qualification letters, log on to Resource Link at the `https//www.ibm.com/servers/resourcelink` website:
  - Go to **Library**, then select **System z Qualified Wavelength Division Multiplexer (WDM) products for GDPS solutions**.
  - Select the vendor that you are interested in. The qualification letters are listed on the right side of the page, under **Download**.
- For the latest list of switch and director qualification letters, log on to Resource Link:
  - Go to **Library**, then select **Switches and directors qualified for IBM System z FICON and FCP channels**.
  - The qualification letters are listed on the right side of the page, under **Download**.

Tip: You can subscribe to these websites to ensure that you are automatically informed should there be any changes to the qualification letters, or if any new devices or firmware levels are qualified.

1.11 Layout of this book

The contents of the remainder of this book are described in this section.

This chapter provides background information, concepts, and an overview of the components within the connectivity architecture. Additionally, you will find organizational suggestions, ways to control the end-to-end architecture, and planning considerations for the future.

Chapter 2, "Storage area network" on page 39**,** provides background information, functional considerations, and concepts associated with the SAN. This chapter includes detail on how to use the SAN over extended distances, and introduces features that make this possible.

Chapter 3, "Wavelength division multiplexing" on page 95**,** provides a background to WDM, including the different types, the components within them, and some of the possible topologies.

Chapter 4, "Common multisite models" on page 113, provides information about common ways that enterprises with multiple System z data centers are configured. The objective of that chapter is to help you understand how other enterprises are using multiple data centers, and briefly provide information about some of the specific connectivity considerations for each configuration.

Chapter 5, "Planning" on page 125, provides all of the information you need to design the extended distance connectivity architecture that addresses the needs of your business today, and positions you to use technology enhancements in the future.

Appendix A, "Performance considerations" on page 179, provides background information about the aspects of system performance that are most likely to be affected in a multisite configuration, and provides information about hardware and software capabilities that can help reduce some of these effects.

1.11.1 Relationship to other documentation

This document represents a point-in-time view. It is not a “living” document, and it will *not* be updated to reflect every future hardware announcement. This document provides high-level information about hardware capabilities. However, its primary purpose is to provide you with a methodology to create, manage, and maintain an extended distance connectivity architecture that is independent of the capabilities of a given device at any particular point in time.

For information about the specific capabilities of a given device, you should always refer to the latest product documentation for that device, because that documentation should be updated to reflect its latest announcements.

For information about System z CEC connectivity capabilities, see the level of *IBM System z Connectivity Handbook*, SG24-5444, that applies to your CEC.

For the latest information about qualified WDMs, see the Resource Link website:

https://www.ibm.com/servers/resourcelink

2

Storage area network

A storage area network (SAN) is a dedicated private network that provides servers with access to storage devices that are physically separated from the server. A SAN enables the storage to be used as though it was local to the server, but also provides the ability to share the storage across many servers. This chapter provides information about SAN topics that are pertinent to extended distance support for System z.

Specifically, this chapter includes the following topics:

- SAN overview
- Channel extension
- SAN support for System z channel protocols
- Considerations for inter-switch link (ISL)-extended SANs
- Considerations for FCIP-extended SANs
- Switch features
- Practical considerations

Important: Because the SAN switching devices are designed and manufactured by vendors other than IBM, this chapter attempts to provide descriptions that are generic across all switches. Although the terminology might vary between vendors, most of the concepts should apply to all vendors.

Some of the concepts are illustrated using a specific switch device's implementation as an example. You should always check with your vendor to ensure that the features and capabilities that you require are available and qualified on the switches you are planning to use, and that you understand the implementation details about the switch that you select.

Also, note that this book represents a point in time. New features might be provided, and existing features might be removed after the publication of this document. Always check with your switch provider for information about the currently-available features.

We suggest that you have a copy of an IBM switch qualification letter available as you read this chapter. The qualification letters are available on the IBM Resource Link website:

`https://www-304.ibm.com/servers/resourcelink/lib03020.nsf/pages/switchesAndDirectorsQualifiedForIbmSystemZRFiconRAndFcpChannels?OpenDocument`

2.1 SAN overview

Before we provide information about the details of how a SAN fits into your extended distance infrastructure and strategy, we will provide a little background about SANs, and list the major features that are relevant within the scope of this book.

Restriction: This is *not* an implementation document. In particular, the complexity of modern switching devices would mean that a dedicated document would be required for each specific switch model.

The intent of this Redbooks publication, and particularly this chapter, is to provide you with sufficient information and terminology to be able to create a high-level design, and to ask the correct questions when working with a SAN specialist. This chapter will not make you a SAN expert, but it should allow you to hold an informed conversation with such a person, and ask the questions that are pertinent to your environment.

In a mainframe environment, storage devices have nearly always been physically separate from the central electronics complex (CEC), and supported connection to multiple CECs. However, in the distributed world, storage was traditionally bundled with the CEC (think about your desktop personal computer (PC) as an example). As distributed systems became more numerous and more powerful, they also needed the ability to connect to a shared storage device from more than one server.

Therefore, the SAN was born. The main difference between a SAN and a traditional mainframe CEC-and-disk configuration is that one or more switches are required to allow multiple CECs to connect to a given storage device in a SAN (in a traditional mainframe environment, CECs would be connected directly to one of several ports on the storage device).

Given that we are talking within the context of a System z environment, a SAN is the set of switches that are used to connect your mainframes to each other, and to the channel-attached devices that they use. In addition, a SAN connects devices to each other, for disk mirroring.

The SAN includes the ports on those switches, but it does *not* include the devices or the processors that are attached to those ports. A SAN might consist of just one switch and its ports, multiple switches, or multiple interconnected switches (called *cascaded switches*), as shown in Figure 2-1 on page 41.

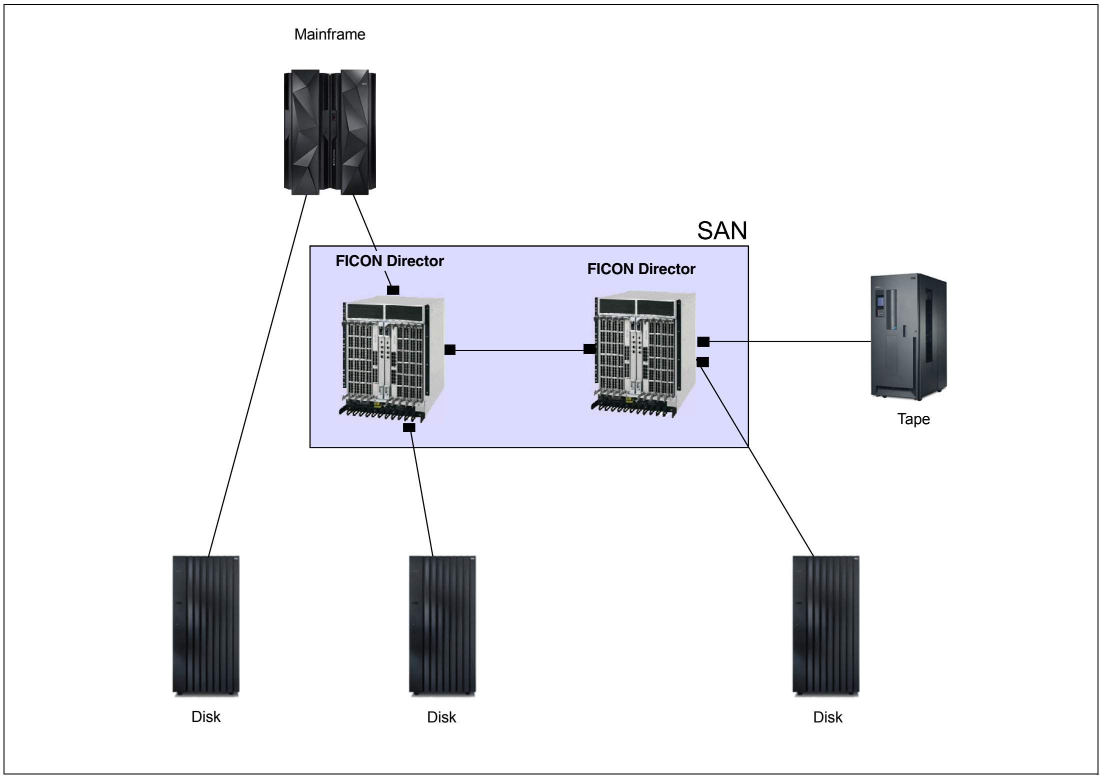


Figure 2-1 Scope of the SAN in a System z environment

A SAN switch in a mainframe environment is generally used to connect devices, such as disks, tape drives, printers, other mainframes, and so on. It does *not* provide connectivity between your mainframe and your corporate network. However, you *can* use Transmission Control Protocol/Internet Protocol (TCP/IP) to connect cascaded switches to each other. Also, note that a switch does *not* support coupling facility (CF) or Server Time Protocol (STP) connectivity.

Depending on the document, the switches and their ports are variously referred to as *switches*, *SAN switches*, or *directors*. In this chapter, those terms are used interchangeably. The set of switches that are connected to each other are called *a SAN* or *SAN fabric*.

Switch versus director: The term *director* refers to a director-class switch. It provides all of the same switching functionality as a switch. When referring to the switching function only, a director might also be referred to as a *switch*.

A switch, however, does not have all of the characteristics of a director. A director is a chassis-based switch that accepts blades with different port counts and functionality. As a result, directors can be used for multiple functions and are more scalable. Also, directors are designed to deliver 5-nines availability, but a switch is typically designed to meet 3-nines availability.

Additionally, as these devices become more powerful and provide more capability and flexibility, they now support virtualization. A physical switch, therefore, might be partitioned into multiple logical switches, just as we have logical partitions (LPARs) in a System z CEC. Furthermore, the connections between the switches might support sharing between logical switches, and the resulting configuration is called a *virtual fabric* or a *virtual storage area network* (VSAN).

2.2 Channel extension

If you have worked with mainframes for some time, you might be familiar with *channel extenders*. These devices (the Ciena CN2000, for example) were connected between the CEC and the control unit (CU). They interpreted the channel protocol, and provided an extension function to allow the CU to be connected at a distance greater than the maximum supported channel distance.

These extenders were completely invisible to the operating system. They were not defined in the hardware configuration definition (HCD), and they had nothing to do with switches. They could optionally be connected to a switch; however, cascaded switch support was limited and not supported at all with some channel extenders.

Today, all Fibre Channel (FC) extension is accomplished using the following methods:

- Directly attaching channels and devices to dense wavelength division multiplexing devices (DWDMs)
- Extending inter-switch links (ISLs) by DWDMs
- Extending ISLs by FCIP

FCIP extension can be accomplished with a blade in a director-class switch, or on a smaller fixed-port switch with built-in FCIP capability. The smaller switch, such as the IBM 2498-R06, can be used as a stand-alone (edge) switch, or can be connected to a larger switch. A stand-alone switch with FCIP capability is often referred to as an *extension switch*.

Users who are familiar with older channel extenders should note that modern extension methods make use of a switch that has a domain ID. The significance of this is that when used as an edge switch, the switch must be defined in HCD, and a two-byte link address must be defined. The first byte of the two-byte link address is the hex equivalent of the edge switch's domain ID. The Control Unit Port function (CUP) is also available for the switch.

2.2.1 FCIP extension switches

There are situations where the most appropriate configuration is two cascaded directors (we use the term *directors* to mean enterprise-level switches designed for maximum resiliency) connected straight to each other. There are other situations where one director connected to a remote extension switch (acting as an edge switch) is the most appropriate. Alternatively, there are situations where you might use both directors *and* extension switches (in which case the extension switches act solely as distance extenders).

Blades with FCIP capabilities within a director inherit the high-availability characteristics of the chassis, and typically have more IP ports and options compared to an extension switch. As a result of the additional IP ports and options, directors can be configured with greater network bandwidth than an extension switch.

An extension switch can be used as an edge switch or as a switch extender. When used as a switch extender, FC ports on the extension switch are connected to a larger switch. This enables a SAN switch without FCIP capability to participate in a SAN fabric extended by FCIP. This approach is useful when the switch does not have FCIP capability, or there are no available slots on the director to accommodate an FCIP blade.

Using an FCIP-capable switch as an edge switch is a cost-effective alternative when the port count of a larger switch is not required. For example, a large data center might have an FCIP blade in a director extended to a remote disaster recovery (DR) or backup site with just an edge switch when there is only a small number of tape or replicated disk interfaces.

Note that in this case the reliability of the switch is usually better than the reliability of the network. Also, because the switch is acting as an edge switch, no local switching is performed, so the lower-availability design (compared to a director) is typically not a concern. Instead, resiliency is obtained through the use of diverse routes, as shown in Figure 2-2.

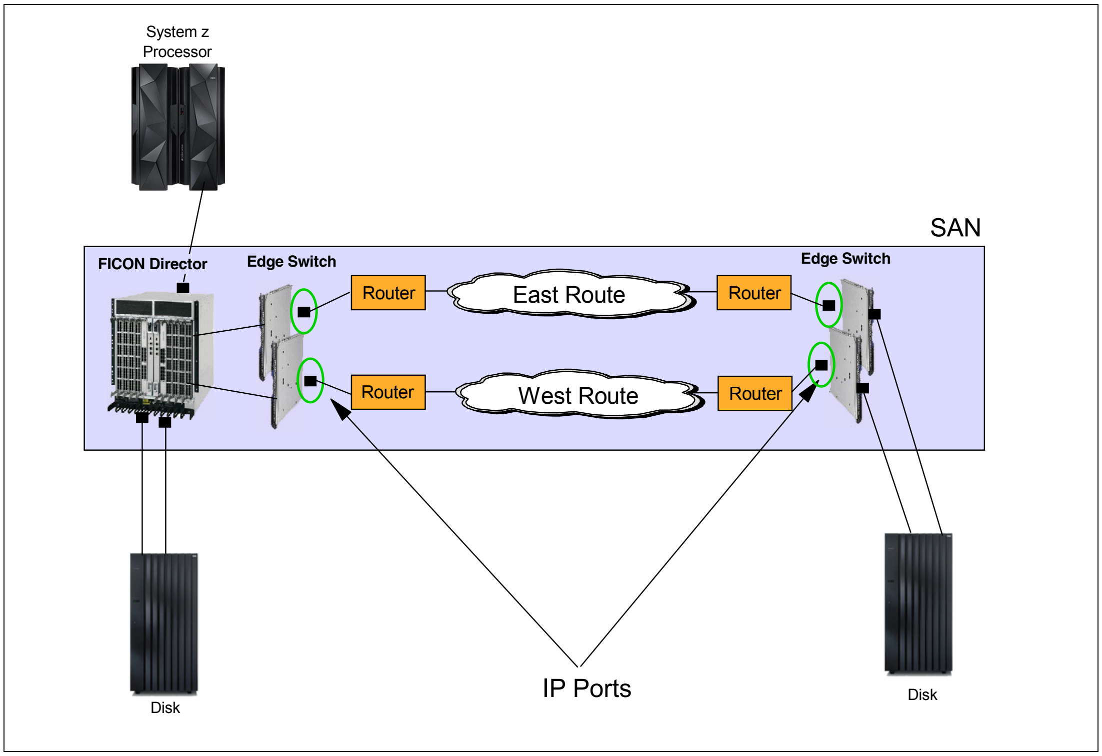


Figure 2-2 Combining director and extension switches

Extension switches support similar types of extended distance inter-switch connections that directors support.

This list includes some of the situations where an extension switch might be used:

- You want to maximize the number of ports in your director that are used to connect channels and CUs. Directors have a finite number of slots. If you want to use FCIP to connect the two sites, the blade that provides the required Ethernet ports on the director would typically support fewer FC ports than a blade that only has FC ports. Moving the Ethernet connectivity to the extension switch, therefore, might enable you to have more FC ports in the director.

  An alternative is to use blades with more ports per blade. This might allow you to provide more FC ports without using all of the slots in the switch. However, it increases the number of ports with a single point of failure. For example, the failure of a 32-port blade would affect 32 ports. If you use a 48-port blade, a failure of that blade would affect 48 ports.

- Site2 does not warrant a director-class switch. If site2 is only used for backup, for example, it might not require the connectivity, capacity, and resiliency provided by a director. In that case, the extension switch can act as the site2 switch, and can connect the site1 director to the CUs in site2.

Note: It is *not* supported for an extension switch to connect to both a switch in the same site *and* to devices in that site. It can either act as an extension device and provide the extended distance support to connect two switches, *or* it can act as a switch and connect to devices in the same site and to the switch in the other site. But it is not supported for it to connect to both devices *and* another switch in the same site.

Alternatively, the use of an extension switch (when acting as an extension device) increases the number of physical boxes that you need to manage, and to provide floor space, power, and cooling for.

Before making a final decision, fully consider your requirements, your strategy, and the costs and capabilities of the various options. For example, a director might support high-bandwidth ports that are not supported on an extension switch.

The capabilities and options provided on these devices are constantly evolving. For more information about determining the most appropriate configuration to meet *your* business requirements, contact your switch vendor.

2.2.2 Switch port types

In a FICON or Fibre Channel Protocol (FCP) switched environment, the ports in the devices (or, to be more accurate, the CUs) and the switches are designated in different ways depending on how the port is being used, as shown in Figure 2-3. We will be referring to these designations as we progress through this chapter, so we briefly describe them.

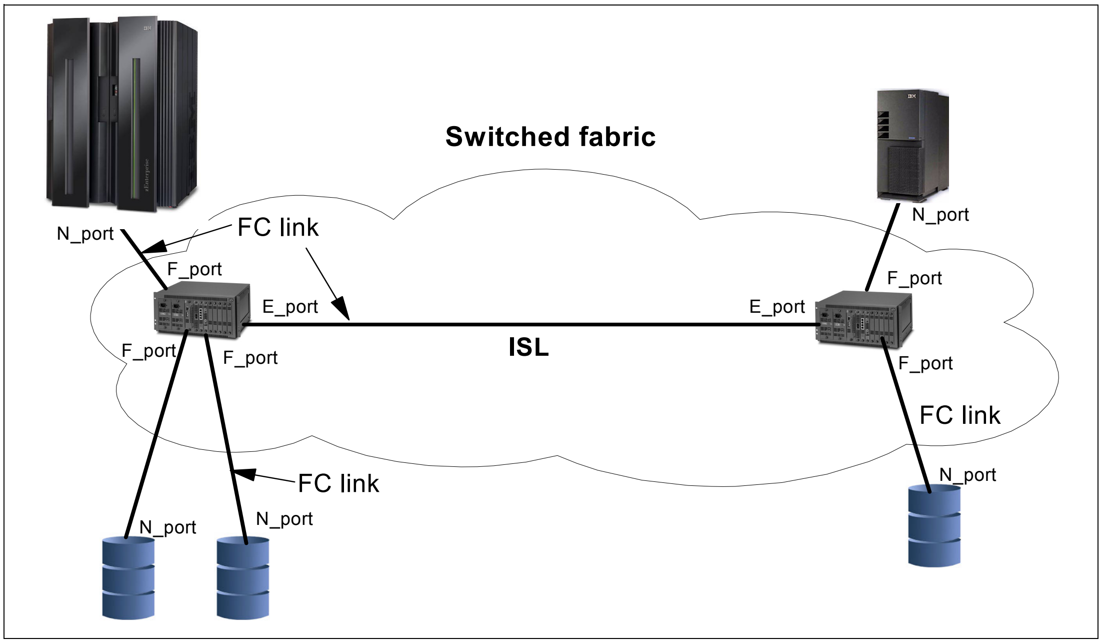


Figure 2-3 Port types in a System z SAN

As shown in Figure 2-3 on page 44, in a System z environment, there are several different designations for FC ports:

- Node ports (N_ports) are on the CEC or the CU that is attached to the switch.
- Fabric ports (F_ports) are the ports on the switch that are used to connect to the CEC or a CU.

- Extension ports (E_ports) are used exclusively for ISL, to connect one switch to another using FC.
- Generic ports (G_ports) are ports that currently are uncabled, but that can operate as either an E_Port or an F_Port.

Additionally, and not shown in Figure 2-3 on page 44, you can use FCIP ports to connect two switches to each other. In this case, the inter-switch communication uses IP rather than FC. As a result, those ports do not have an FC designation. Rather, they are referred to as Virtual E_ports (VE_ports).

To fully understand how to implement and manage these new switches in an extended distance environment, you need a basic knowledge of SAN and the use of two-byte link addresses. If you are not familiar with this, we strongly advise that you read *FICON Planning and Implementation Guide*, SG24-6497, before continuing with this chapter.

2.3 SAN support for System z channel protocols

System z connections to the SAN can use either FICON protocol or FCP[1]. Each port on a FICON Express card can be defined in HCD to run in FICON mode or in FCP mode. Similarly, on a disk subsystem, each port can be defined to run in FICON mode or in FCP mode. FICON mode is used to connect to traditional mainframe devices, and FCP mode is used to connect to Small Computer System Interface (SCSI) devices, and for disk-based mirroring (Metro Mirror and Global Mirror).

A System z server can use FICON mode channels for z/OS, z/VM, z/VSE, Linux for System z, and IBM z/Transaction Processing Facility (z/TPF) partitions. FCP mode channels can be used by z/VM, z/VSE, and Linux for System z partitions. Both modes (FICON and FCP) use the underlying FC architecture. For more information about the relationship between FICON, FCP, and FC, see the “Introduction to FICON” chapter in *FICON Planning and Implementation Guide*, SG24-6497.

2.3.1 Uses of SAN switches in a System z environment

There are several types of traffic that can use a SAN switch in a System z environment:

FICON CEC-to-Device Generally speaking, these devices can be used in the same way at the maximum supported distance that they can be used at local distances (bearing in mind the response time effect of the distance). The qualification letter for each switch indicates the type of devices that are supported on FICON channels and the supported distances (the maximum is 300 km).

FICON CEC-to-Device with FICON Acceleration

A limited subset of applications can be used with FICON-attached devices that are more than 300 km from the CEC. However, this requires a feature in the switch called *FICON Acceleration*.

Although FICON Acceleration is required if you want to use certain devices or applications at distances greater than 300 km, there might also be cases where FICON Acceleration can provide performance benefits at shorter distances.

[1] Because Enterprise Systems Connection (ESCON) channels are not supported on the latest generation of System z processor, and therefore are no longer strategic, we focus on FICON and FCP channels in this chapter.

The applications and device types that are supported by FICON Acceleration vary from one switch to another, and also change over time, so you should work with your switch vendor to determine the applicability of FICON Acceleration to your environment. FICON Acceleration is covered in more detail in 2.6.5, “FICON Acceleration” on page 75.

FCP CEC-to-Device Generally speaking, these devices can be used in the same way at the maximum supported distance that they can be used at local distances. The qualification letter for each switch indicates the type of devices that are supported on FCP channels and the supported distances (the maximum is 300 km).

The use of FCP-connected devices at distances greater than 300 km is not qualified by IBM. There is a subset of applications and devices that might be supported by the vendor over longer distances. If you have a specific requirement to use an FCP-attached device over a distance greater than 300 km, consult your switch vendor.

FCP Disk-to-Disk mirroring IBM disk-based replication products (Metro Mirror and Global Mirror) use FCP to communicate between primary and secondary subsystems. The maximum distance is far greater for asynchronous replication, potentially up to tens of thousands of kilometers, depending on the switch (or the DWDM if a DWDM is being used).

The protocol that is being used by the channel (FICON or FCP) does not matter from a switching perspective. However, switches *do* need to distinguish between FICON and FCP for the following reasons:

- Some features built into the SAN switch intercept the FC frames and interpret the FICON protocol:
  - The CUP is a configurable logical port at link address *nn*.FE. The CUP function intercepts FC frames and routes them to a software function within the switch that responds to certain commands issued from the host. These commands support products, such as IBM Tivoli® System Automation, and provide switch statistics for the FICON Director Activity Report in IBM RMF™.

    The function of the CUP is described in more detail in the section titled “Control Unit Port” in *FICON Planning and Implementation Guide*, SG24-6497.
  - There are differences in how FCP and FICON requests over distances greater than 300 km are handled. This is covered in 2.6.5, “FICON Acceleration” on page 75 and 2.6.6, “FC Fast Write” on page 80.
- FICON and FCP on System z do not support all switch products, and associated firmware levels, due to their strict qualification process.
- In a distributed environment, switches (known as *converged switches*) can be used to provide TCP/IP connectivity as well as storage device connectivity. However, in a System z environment, SAN switches only connect to an IP network to connect two switches. The SAN switch would *not* be used to connect your mainframe to your corporate network.

Both FCP and FICON modes use FC frames to transport their requests. The data field in a FC frame can be up to 2112 bytes long. The first 32 bytes of the data field are reserved for a standard FCP header, and the remaining 2080 bytes are available for data associated with commands, as shown in Figure 2-4.

FCP frames are variable in length, so only the required number of bytes for the payload are added to the frame, along with control and routing bits. The details about the contents of the remaining fields in the frame are not of interest in the context of this book.

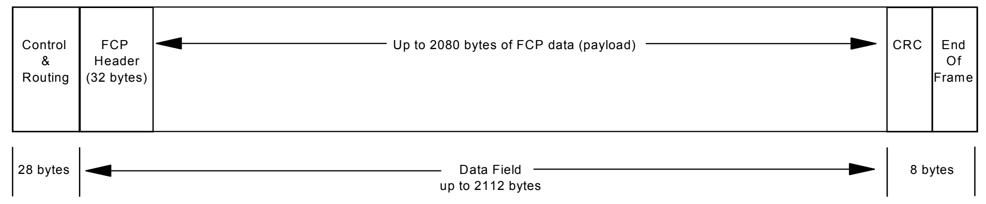


Figure 2-4 FCP frame

The same FCP frame layout is used by FICON, with the exception that a 32-byte FICON header is inserted in the start of the payload area, as shown in Figure 2-5. As a result, the smallest FICON frame is about 100 bytes (32 bytes for the FCP header, plus 32 bytes for the FICON header, plus 36 bytes for the FC header and trailer fields).

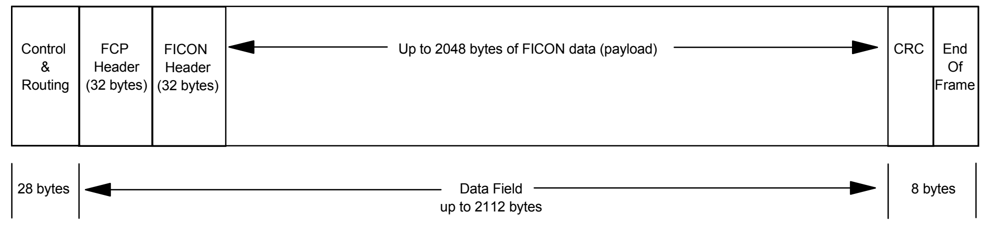


Figure 2-5 FCP frame with FICON header in payload

The variable size of frames and the considerations for the additional FICON header in the data field are relevant when calculating buffer credit requirements. However, the guidelines in this book take these considerations into account, so you do not need to get into this level of detail when planning your environment. For more information about the relationship between frame sizes and buffer credits, see 2.5.2, “Buffer credits” on page 59, and the section about buffer credits in *FICON Planning and Implementation Guide*, SG24-6497.

2.3.2 FC addresses and SAN ports

Every switch port in a SAN fabric has a unique 24-bit address composed of three parts, as described in Table 2-1.

Table 2-1 Constituent parts of the SAN fabric 24-bit address

Domain ID	Bits 23 - 16. First byte of a 2-byte link address. Switch domain ID. Note that domain IDs are specified in decimal on switches, but represented in hex in the link address and in the FC address.
Area	Bits 15 - 8. This is the only byte of the link address when using a single-byte link address, or it is the second byte when using a two-byte address. It is the target port address on the switch.
Arbitrated loop physical address (AL_PA)	Bits 7 - 0. Not used in the link address. Arbitrated loop is a previous protocol that allowed multiple physical CUs to share the same SAN switch port. AL_PA is not supported in a FICON environment, so modern switches configured for a FICON-compatible addressing mode use an AL_PA of `x'00'`. Today, these bits are used for Node Port Identification Virtualization (N-Port ID Virtualization, or NPIV) if that function is enabled for System z FCP channels. Switch vendors might also use some of the bits as an expansion of the area field when the switch port count exceeds 256 ports.

Open systems servers and System z LPARs that access SCSI devices using FCP channels (Linux for System z, z/VM, and z/VSE) reference them by a worldwide name (WWN). A sample z/VM definition of an FCP-attached device is shown in Figure 2-6.

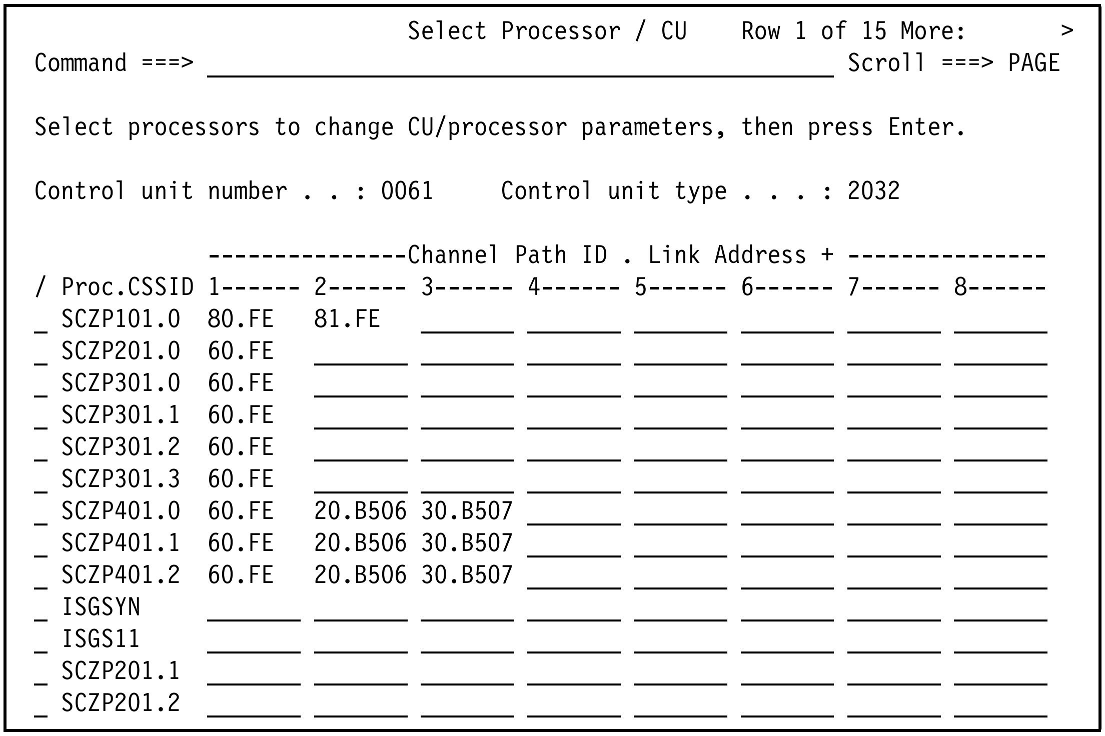
```
/***********************************************
/* SCSI Definition Statements */
/***********************************************
edevice 8000 type fba attr SCSI fcp_dev 6002,
wwpn 5005076300C300AA lun 5010000000000000
edevice 8001 type fba attr SCSI fcp_dev 6003,
wwpn 5005076300C300AA lun 5011000000000000
edevice 8002 type fba attr SCSI fcp_dev 6006,
wwpn 5005076300C300AA lun 5012000000000000
edevice 8003 type fba attr SCSI fcp_dev 6007,
wwpn 5005076300C300AA lun 5013000000000000
```

Figure 2-6 Sample z/VM definitions for an FCP-attached disk device

The SAN switch name server function resolves a WWN to a 24-bit FC address. In FICON, the FC address is determined using information specified on the channel-path identifier (CHPID), and CU definitions in HCD or in the input/output configuration data set (IOCDS).

2.3.3 FICON link addresses and FC addresses

To be able to use the FICON protocol to access a CU that is connected to a switch, the System z host needs to know which switch the CU is connected to, and which port on that switch. The switch and the port are identified by a value known as the FICON link address. The FICON link address is specified in HCD as part of the CU definition, as shown in Figure 2-7.



```
                            Select Processor / CU    Row 1 of 15 More:       >
Command ===> _______________________________________________ Scroll ===> PAGE

Select processors to change CU/processor parameters, then press Enter.

Control unit number . . : 0061     Control unit type . . . : 2032

             ---------------Channel Path ID . Link Address + ---------------
/ Proc.CSSID 1------ 2------ 3------ 4------ 5------ 6------ 7------ 8------
_ SCZP101.0  80.FE   81.FE   _______ _______ _______ _______ _______ _______
_ SCZP201.0  60.FE   _______ _______ _______ _______ _______ _______ _______
_ SCZP301.0  60.FE   _______ _______ _______ _______ _______ _______ _______
_ SCZP301.1  60.FE   _______ _______ _______ _______ _______ _______ _______
_ SCZP301.2  60.FE   _______ _______ _______ _______ _______ _______ _______
_ SCZP301.3  60.FE   _______ _______ _______ _______ _______ _______ _______
_ SCZP401.0  60.FE   20.B506 30.B507 _______ _______ _______ _______ _______
_ SCZP401.1  60.FE   20.B506 30.B507 _______ _______ _______ _______ _______
_ SCZP401.2  60.FE   20.B506 30.B507 _______ _______ _______ _______ _______
_ ISGSYN     _______ _______ _______ _______ _______ _______ _______ _______
_ ISGS11     _______ _______ _______ _______ _______ _______ _______ _______
_ SCZP201.1  _______ _______ _______ _______ _______ _______ _______ _______
_ SCZP201.2  _______ _______ _______ _______ _______ _______ _______ _______
```

Figure 2-7 Defining two-byte link address for CU in HCD

The FICON link address can be specified as a one-byte or two-byte value, as shown in Figure 2-7 (`60.FE` is a one-byte link address, and `20.B506` is a two-byte link address). The FICON channel builds the 24-bit FC address using the FICON link address and other information from HCD.

If a single-byte address is provided, that address represents the port number on the switch. The fact that no switch ID is provided means that the destination port is on the same switch that the channel is connected to, and the switch ID of that switch is obtained from the CHPID definition. Single-byte addressing, therefore, can be used only for CUs that are connected to the local switch, and can never be used for extended distance solutions.

Because the AL_PA byte is always taken from the channel entry link address, the CU must be connected to a port that has the same AL_PA as the channel port. This is why switch vendors must limit the addressing mode used for FICON connections to one that does not use AL_PA bits. Some existing switches used an AL_PA of `x'13'`, but all modern switches configured for FICON-compatible addressing mode use an AL_PA of `x'00'`.

Figure 2-8 shows the relationship between a one-byte link address, a two-byte link address, and the FC address. In this example, the domain ID of the switch the channel is attached to is 181 (x'B5'). The resulting 24-bit FC address, therefore, will be B50600 (B5 is the Domain ID, 06 is the port number, and 00 is the AL_PA), whether you specify a single-byte link address of 06 or a 2-byte link address of B506.

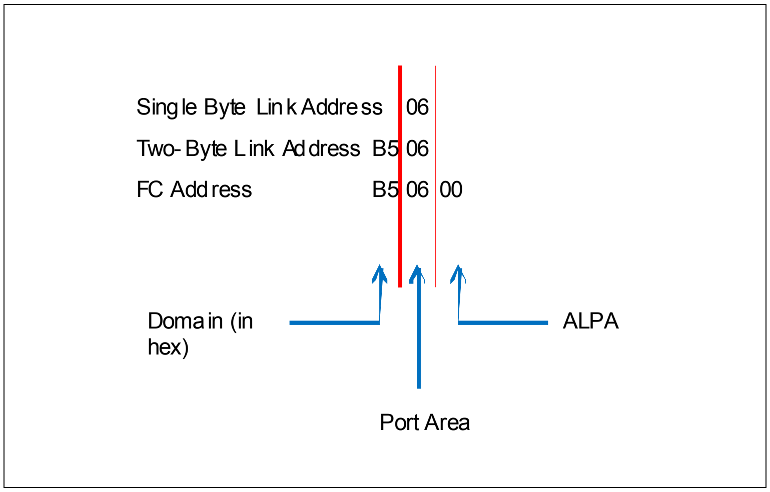


Figure 2-8 FICON Link Address

Two-byte FICON link addresses

Any connection between two switch domains is called a *hop*. At the time of writing, FICON generally permits only one hop, although multiple hops are permitted in certain circumstances where the paths are well-controlled[2] (for example, if you use an extension device). The SAN switch routes frames based on the destination address, so the FICON link address identifies the switch *and* port that the target CU is connected to.

Important: The SWITCH ID on the CHPID statement does not necessarily have to match the switch address when using the 2-byte link address. The Domain ID on the switch or director must match the switch address in input/output configuration program (IOCP)/IOCDS, but the switch address does not have to match the SWITCH ID in the CHPID statement. However, using the same value for both is highly suggested to minimize confusion.

Note that if a two-byte switch address is defined for any CU on a given channel, all link addresses specified for that *channel* must be two-byte addresses, even when accessing CUs that are connected to the local switch. In Figure 2-9 on page 51, the System z host uses a link address of B506 to access the storage CU that is connected to the remote switch (with the Director ID of 181).

Therefore, the link address for the CUP on the local switch (Director ID of 182), which is attached to the same CHPID where link address B506 was defined, must be specified as B6FE and not just FE, and the link address of the storage CU attached to port 06 on that switch would be B606.

[2] *Well controlled* in this context means that the switch configuration is such that the precise route taken for a given input/output (I/O) can be predicted in advance, and will not change as a result of an unplanned configuration change.

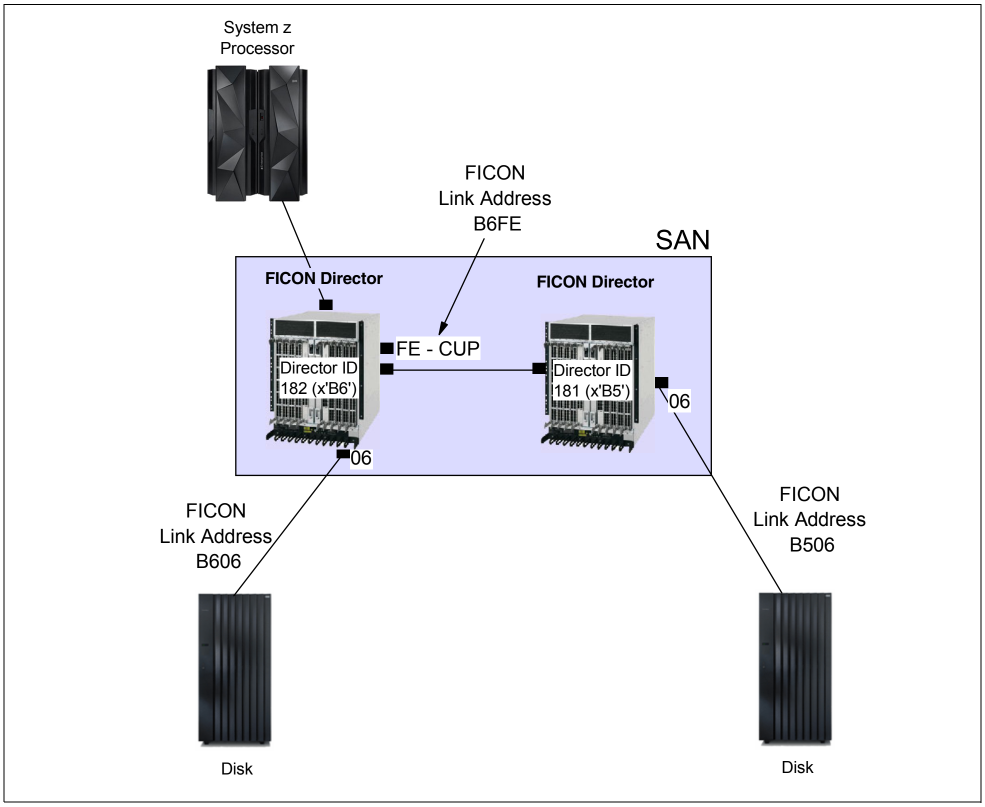


Figure 2-9   Two-byte addressing for extended distance storage

Note that if extension switches were used between the two directors, they would be invisible to the operating system. The two-byte link address identifies the director that the CUs are connected to, not any extension switches that might be used to support the distance between the directors.

2.4 Extending the SAN

To provide greater connectivity, more flexibility, and support for greater distances, multiple SAN switches can be interconnected, as shown in Figure 2-10[3]. Depending on the switch vendor and the specific model, there are several options for connecting the two switches to each other, some of which support an extended distance configuration. In this document we will only provide information about the options that support extended distance.

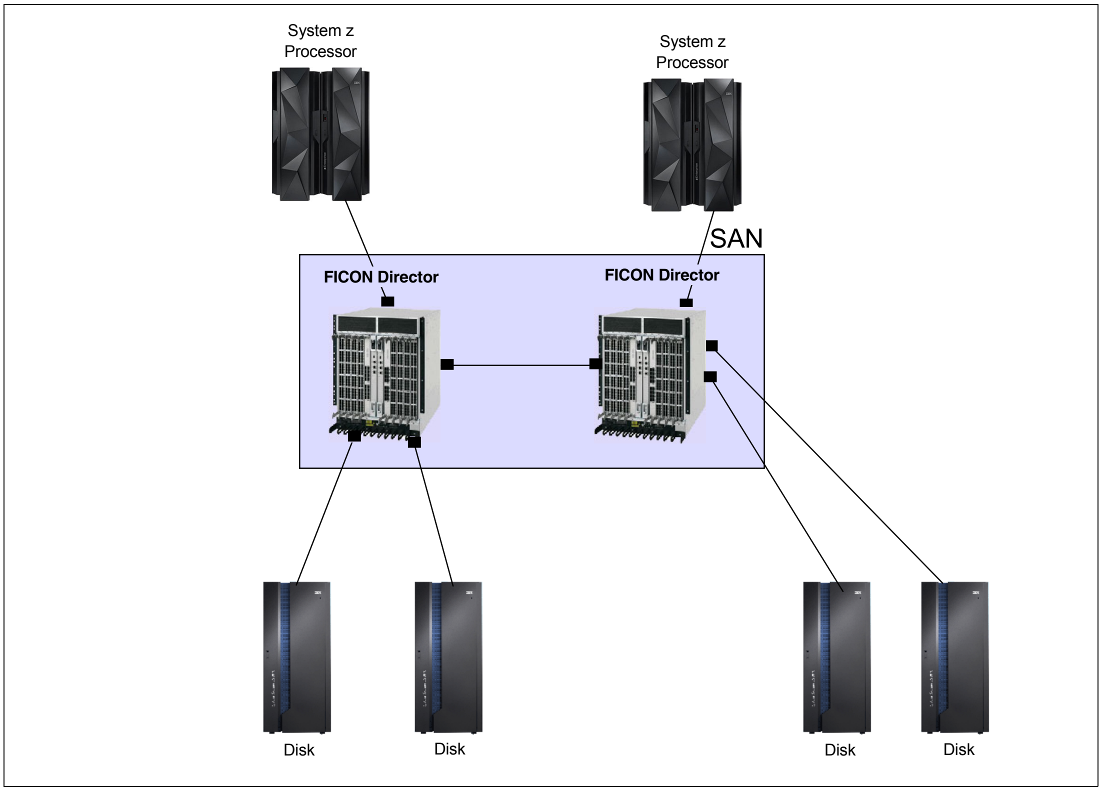


Figure 2-10 Cascaded director configuration

There are three options for SAN extension:

- Direct fiber ISLs
- ISL through DWDMs
- FCIP

The option that is the most appropriate for you will likely depend on the distance between your sites, the availability of dedicated fiber, the performance and availability requirements of your applications, whether you will have DWDMs, and the relative cost of the various options (especially the cost of the connectivity between the two sites). Table 2-2 on page 53 summarizes the different options, and the characteristics of each.

[3] Multiple SAN switches in the same site would typically not be connected to each other. If this were done, only a single fabric is configured, which reduces availability. In a local environment, it would be better to have each CEC connected to each switch, and each switch connected to all of the storage units, but not connected together.

Table 2-2 Extended distance option characteristics

Characteristic (Note 1)	Direct ISL	ISL over DWDM	FCIP
Maximum supported distance	300 km	300 km	Note 2
Requires dedicated fiber between sites	Yes	Yes	No
Supports encryption	Yes (Note 3)	Yes (Note 3)	Yes
Supports compression	Yes (Note 3)	Yes (Note 3)	Yes
Supports IBM z/OS Global Mirror (zGM) remote copy (FICON)	Yes	Yes	Yes
Supports Global Mirror disk mirroring (FCP)	Yes	Yes	Yes (Note 4)
Supports Metro Mirror disk mirroring (FCP)	Yes	Yes	Yes
Supports host-to-CU connections	Yes	Yes	Yes
Supports FICON Acceleration	No (Note 5)	No	Yes (Note 2)
Tape mirroring (Note 6)	No	No	No
Supports use of DWDMs	Not applicable (N/A)	Required	Yes
Supports use of extension switches	Yes	Yes	Yes

Note 1: This table is accurate at the time of writing. Always consult your switch vendor for the latest capabilities and limitations.

Note 2: Limit is 300 km without FICON Acceleration. FICON Acceleration is only *required* for distances greater than 300 km. Maximum distance with FICON Acceleration for applications that support FICON Acceleration is up to 20,000 km, depending on the vendor and the specific switch model.

Note 3: Switch model-dependent. Consult your switch vendor.

Note 4: Maximum supported distance is 20,000 km, depending on the vendor and the specific switch model.

Note 5: If FICON Acceleration is used for distances greater than 300 km, you must remember that ISL does not support such distances. However, a channel that uses FICON Acceleration could be connected to a director, and that director could then be connected to an extension switch using ISLs.

Note 6: Tape mirroring uses IP for the connection between the tape subsystems, and therefore does not connect to a SAN switch.

Performance considerations: Every 1 km of distance added between a CEC and a device will increase service times by about 0.1 millisecond (ms). For FICON-attached devices, the service time increases linearly up to about 120 km. At about 120 km, a phenomenon known as *FICON droop* occurs. Beyond this distance, service times increase non-linearly.

If the I/Os that are going between the sites are synchronous to the execution of a critical transaction or job, their effect on elapsed time must be considered.

If you select the direct fiber ISL or ISL over DWDMs options, the FC ports used to connect the switches to each other must provide sufficient buffer credits for the link speed and the distance.

The switch's ability to provide buffer credits is important because, with 8 Gb and higher speeds, the number of buffer credits provided in the System z channel card[4] is limited to however many are required to support a maximum distance of 10 km.

DWDM vendors no longer provide buffer credits for 8 Gb FC and higher speeds, so the buffer credits required for extended distances must come from the switch.

Assumption: All of the configurations in this document assume that each switch is located within a maximum of 10 km of any CEC that it is directly connected to, and within the supported distance from any CUs that it is directly connected to.

If you select the direct fiber ISL option (that is, you will not be using DWDMs) or the FCIP option without a DWDM, the optics for the ISLs must be powerful enough to support the planned distance. Because TCP/IP flow control is used to control the transport of the FC frames between the switches when using the FCIP option, buffer credits are not required on the IP ports on the switches, meaning that the IP ports will not use any of the switch's buffer credits.

One thing that must be considered is the financial viability of the various options. For example, a DWDM might have the physical ability to drive a signal over a 300 km link. However DWDM links require dedicated fiber[5]. When you consider the number of links that you would require (bearing in mind availability, failover, and capacity requirements), and multiply that by the cost of providing 300 km of dedicated fiber for each link, the cost might be prohibitive.

Therefore, when looking at any extended distance solution, you must consider both the technical feasibility of the configuration that you are considering *and* the financial cost. The combination of these will help you identify the optimal solution.

This section describes several options.

Tip: Regardless of which option you determine to be the most appropriate for your enterprise, one of the inputs to your decision will be the amount of bandwidth that will be required between the switches (remembering to allow for failover). Your switch vendor can help you determine the required number of ISL or Ethernet ports. That information, in turn, will be one of the inputs to your decision about whether you require a DWDM and, if so, how it will be configured.

[4] The number of buffers at each end of a fiber must be sufficient to support the distance, the bandwidth, and the average frame size. If the port at either end of the fiber does not have sufficient buffers, the performance of the link will be gated to that possible with the smaller number of buffers.

[5] The only System z-qualified configuration is to use dedicated fibers between the DWDMs. Other options are possible, but not qualified.

2.4.1 Direct fiber ISLs

Direct fiber ISL connections between SANs are a good option when there are sufficient fiber links available between the locations, as shown in Figure 2-11. Typically, this would be in a campus environment, but you can use direct fiber links whenever enough links are available.

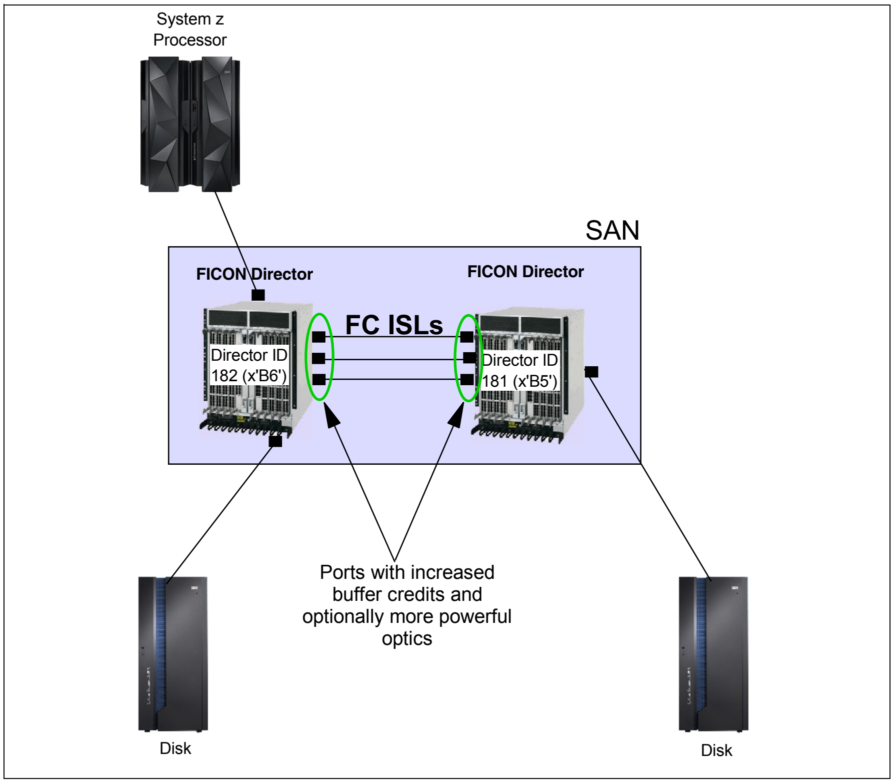


Figure 2-11 Direct fiber SAN extension

The supported distance is dependent on available buffer credits and port optics power, but cannot exceed 300 km. Before you implement this solution, check with your switch provider to confirm that the switch model you plan to use can provide the required number of buffer credits. Additional buffer credits are required on the ports where the ISLs are connected[6], but no additional buffer credits are required on the ports where the System z host or CUs are connected[7].

Also, make sure that the transceivers, generally referred to as *small form-factor pluggables (SFPs)*, have sufficient optical power to meet the link budget requirements required for the distance[8]. In addition, remember that the ports on both ends of a connection must be long wave, or both must be short wave. If none of the qualified SFPs support the distance, and you must use ISLs rather than FCIP, you have to use a DWDM to provide the extended distance support.

[6] The number of buffer credits that are available on the switch, and how they are associated with given ports, varies depending on the particular switching device and the vendor. Speak to your vendor for information about how this is handled in your configuration.

[7] Although it might be technically possible to have a configuration consisting of a CEC, a switch many kilometers away, and a device more kilometers away directly connected to that switch, such a configuration would be highly unusual, and therefore will not be included in this document. We assume that remote devices will be connected using a cascaded switch configuration, and that all devices connected to a switch are a maximum of 10 km from the switch.

[8] The different ports on a blade can use different SFPs, meaning that the ports used for ISLs can be spread across multiple blades for increased resiliency.

2.4.2 ISLs through DWDMs

Because of their ability to multiplex many wavelengths of light into a single fiber, you can use fiber ISL connections through a pair of DWDMs for situations where fiber availability between sites is limited or very expensive, *or* if you need to connect devices that are not supported by SAN switches (InfiniBand (IFB) for CFs and STP, or tape mirroring over IP, for example). A configuration that uses switches and DWDMs is shown in Figure 2-12.

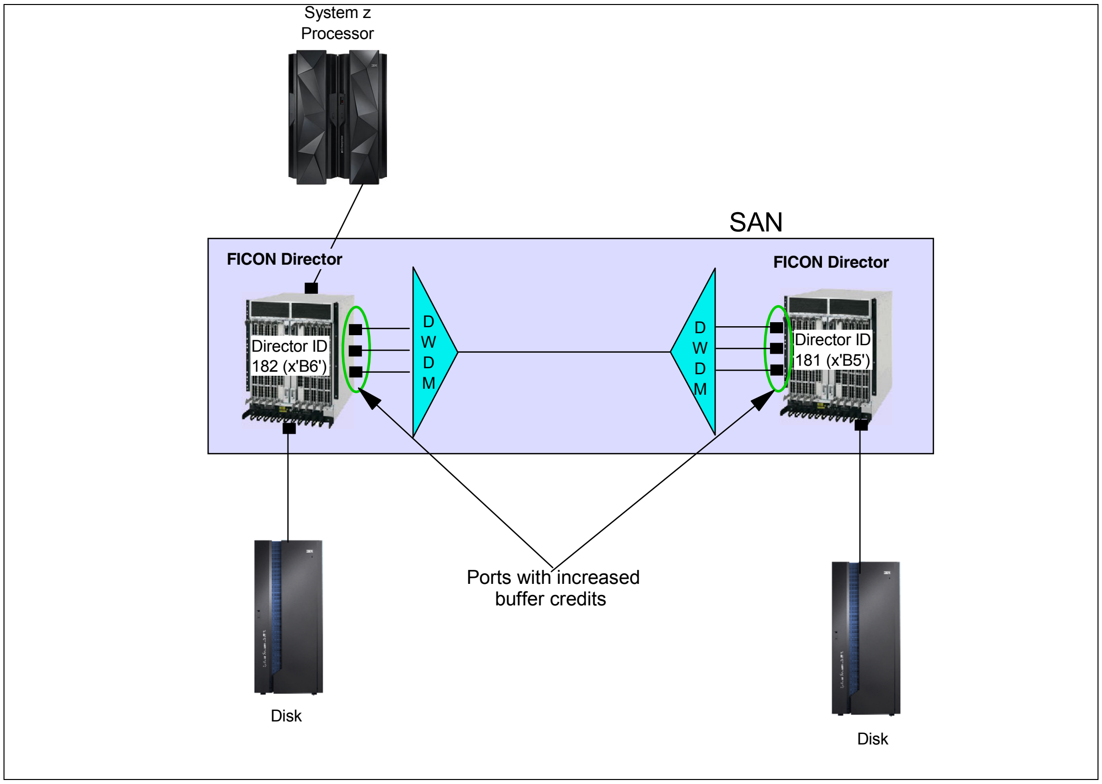


Figure 2-12 SAN extension through DWDM

Typically, this configuration is used for remote sites outside a campus but within 300 km. The actual supported distance depends on the buffer credits available on your switch, the optical power capabilities of the DWDM, and the capabilities of the connected CUs. Before you implement this solution, check with your switch provider to verify that the switch model you are using has sufficient buffer credits. Also ensure that the DWDM has enough optical power to meet the link budget requirements for the distance.

As with direct fiber connections, the only place additional buffer credits are required is in the ports used to connect the two switches. However when using DWDMs, the optical power requirements have moved from the switch to the DWDM.

2.4.3 Fibre Channel over IP (FCIP)

FCIP technology encapsulates FC frames in an IP protocol, allowing those frames to be transported over an IP network. Extending the SAN over an IP network is useful when dedicated fiber between the sites is unavailable, or if the distance exceeds 300 km, as illustrated in Figure 2-13.

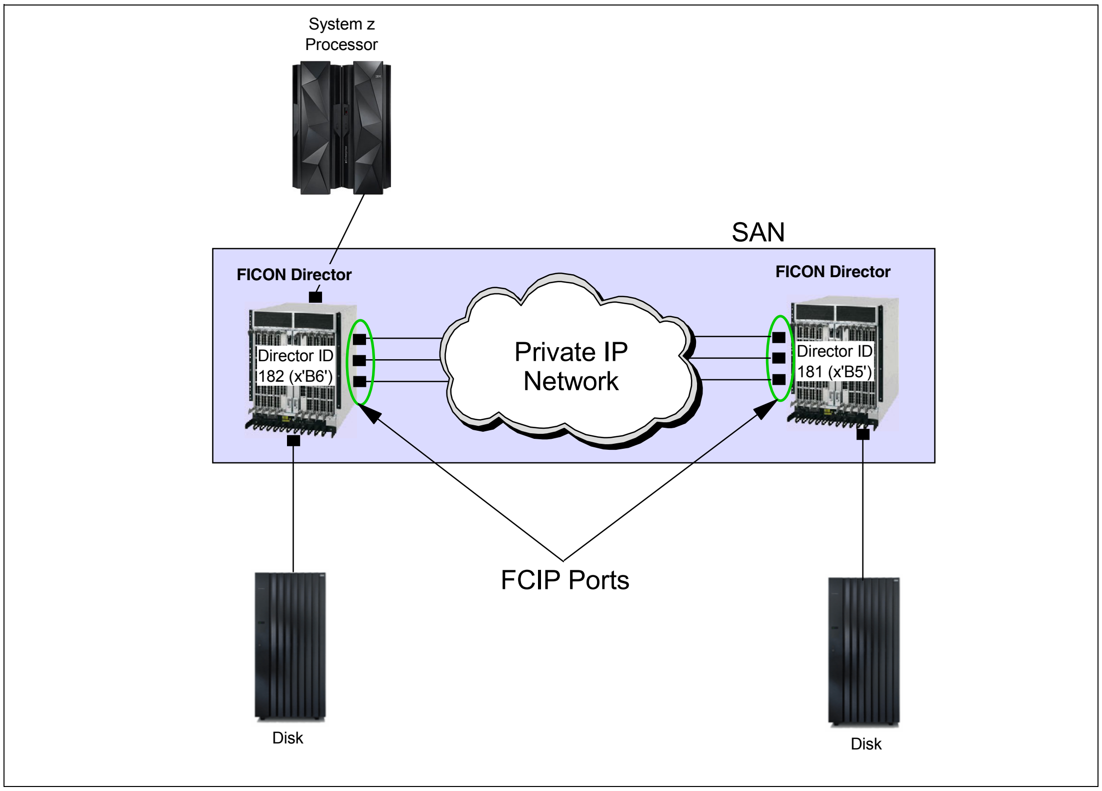


Figure 2-13 SAN Extended over IP

In situations where the distance is less than 300 km, and both dedicated fiber and IP networks are available, you have to choose between using ISLs and FCIP. One of the criteria that must be considered is the performance that will be delivered by either option. The performance difference between an ISL link and an FCIP link over the same distance consists of the following elements:

- Protocol conversion time to convert the FC frames to IP packets and back again
- Time for the packet to traverse the IP network
- Potentially, time spent waiting to reassemble the FC frame if the packets arrive out of order

The first of these (protocol conversion time) is reasonably consistent and predictable. However, the other two are less predictable, and can be affected by other traffic sharing the same network. You should work with your switch vendor and network provider to determine a reliable value for the effect of using FCIP in *your* environment.

The results will vary, but in some cases the response times delivered by FCIP can be close (within fractions of an ms) to those delivered by ISL. For more information about the relative performance of FCIP and ISLs, see the article titled "*When Milliseconds Matter: FCIP and Fibre Channel Performance Over Distance*" on the `http://www.brocade.com` website.

Note that only certain applications (tape read and write, zGM, FCP for disk mirroring, and possibly selected other applications) can be extended beyond 300 km. Also, although the use of an IP network might initially appear to be a less expensive option than dedicated fiber, the stringent quality of service (QoS) requirements when extending a SAN over an IP network, combined with the need for *at least* OC3 links, will increase the price of the IP option.

Although FCIP supports direct connection from the switch to the network, it is also possible to use FCIP in conjunction with a DWDM. One example of where you might do this is if you have a DWDM with under-used Ethernet ports that are used by other IP applications (tape mirroring, for example). Using FCIP and being able to use those DWDM ports might allow you to avoid the cost of adding ISL ports to the DWDMs.

Of course, it is possible to use a combination of more than one of these options. However, if you have a reason to use DWDMs, it is likely that you will route *all* cross-site connections over the DWDMs. Also, if FCIP is not appropriate for some or all of your applications, it is unlikely that you would want to use a combination of ISL links and FCIP links. You should also try to avoid creating a configuration that is more complex than necessary.

2.5 Considerations for inter-switch link (ISL)-extended SANs

Traditionally, the connection between two SAN switches in a mainframe environment was provided by ISL links, so we will cover those first. The considerations are similar regardless of whether DWDMs are used or not.

2.5.1 When ISLs are appropriate

One of the overriding considerations in respect to whether ISLs can be used is the distance between the sites. Regardless of whether the ISLs use direct fiber or if they are routed over DWDMs, the maximum distance supported by ISL links is 300 km. If you need to extend over a larger distance, FCIP is the only supported option.

Another consideration is the availability of dedicated fiber between the sites. Both direct fiber ISL links and ISL links over DWDMs require dedicated fibers. If dedicated fibers are unavailable, FCIP is the only supported option.

Regardless of whether you are using direct fiber or DWDMs, the same number of ISL ports will be required, based on the capacity, performance, and failover requirements of your environment. Your switch must have enough spare FC ports to support that number of ISLs. If the switch is constrained in terms of the number of available FC ports, FCIP might be a viable alternative[9].

To fully use the bandwidth provided by your ISL links, sufficient buffer credits[10] must be available in the switch. If you are unable to provide enough buffer credits to deliver acceptable performance, an alternative is to use FCIP. Because FCIP uses TCP/IP flow control rather than buffer credits, FCIP might be able to provide the required levels of performance. Your vendor can help you identify the required number of buffer credits based on the distance, link speed, average frame sizes, and features used (such as compression).

[9] Depending on the capabilities of your switch and how it is configured, it *might* be possible to add Ethernet ports without decreasing the number of available FC ports.

[10] Smaller frames require more buffer credits to avoid performance issues.

Because ISL links use dedicated fiber over a fixed path, the response times they deliver tend to be predictable and consistent. Alternatively, because of the nature of IP networks, and the likelihood that the path between your sites will be longer when using IP than your own dedicated fiber, response times when using FCIP are likely to be a little longer and less predictable.

If the only I/Os between the sites are asynchronous (Global Mirror or zGM, for example), either ISLs or FCIP should be acceptable. But if the I/Os are synchronous to transaction execution, and your applications have stringent response-time requirements, ISLs might be required. You should work with your switch vendor and network provider to obtain comparative performance information and costs for your configuration.

2.5.2 Buffer credits

In this section, we provide information about guidelines for buffer credit planning when measurements are unavailable.

One buffer credit is required for each in-flight FC frame. Because FC frames vary in length, the number of frames (and, therefore, the number of buffers) required to keep a link full depends on the frame size. Advertised distance support is based on full-size frames. Vendors might adjust buffer credits sizing calculations to allow for occasional frames that are less than full, but the calculations generally assume that average frame size will be 90% of the maximum frame size.

Most open systems’ traffic results in close-to-full average frame sizes. FICON tape, Modified Indirect Data Address Word (MIDAW), and High Performance FICON for System z (zHPF) will also use close-to-full frame sizes. The average frame size for most random access non-zHPF FICON 4 KB block direct access storage device (DASD) traffic tends to be between 800 and 830 bytes.

This is primarily due to a single 35-byte frame with extended count key data that can be as frequent as every other frame when performing random DASD access. Channel-to-channel (CTC) traffic is often less than 100 bytes per frame, and therefore warrants special attention.

As illustrated in Figure 2-11 on page 55 and Figure 2-12 on page 56, additional buffer credits are generally required for the ISL ports, but not the ports where the hosts and CUs are connected. Buffer credits on the ISL ports must be planned based on the applications with the smallest average frame size. This is because applications using smaller frame sizes can use up all the buffer credits available on the E_ports.

Because there is one buffer credit per frame, and frames vary in length, all buffer credits can be used up for small frames before the ISL bandwidth is fully used. This affects the performance of applications that use full frame sizes that are sharing that port. This is one of the reasons that IBM suggests keeping tape traffic on different ISLs than your disk traffic. Your switch vendor will work with you to identify the usage information that is input to the buffer credit calculation.

In addition to providing sufficient buffer credits, you can mitigate some performance effects by using vendor-specific switch features and design approaches. Discuss these options with your switch vendor if you are unable to provide adequate buffer credits.

2.5.3 Identifying the correct number of ISLs

As part of your extended-distance SAN design, you will need to identify the number of ISLs that will be required. The following list describes the inputs to that decision:

- The amount of bandwidth that will be required

  You should work with your switch vendor to identify the traffic that will be using the ISLs:

  - Is DASD mirroring being used, and if so, what are the peak write rates?
  - Will CECs in one site be reading and writing from the DASD in the other site?
  - Will HyperSwap be used to swap the location of the primary and secondary DASD, and if so, how will that affect the volume of DASD traffic between the two sites?
  - Will CECs in one site be reading from and writing to tape in the other side?
  - If the normal mode of operation is that all tape I/Os will be directed to one site, is there a scenario where those I/Os might be directed to tape drives in the other site?
  - What other devices will be connected to the switches, and will they need to be accessed by a CEC in the other site?

  These are just some of the questions that must be answered. You should go through every possible scenario of where your systems and primary and secondary tape and disk might end up, and ensure that your configuration can deliver the required levels of performance in all cases.

- Your availability requirements:

  - One ISL might provide sufficient bandwidth, but that would mean that all of your ISL connectivity is on a single blade. Would that be acceptable?
  - Assuming that you have two failure-isolated paths between the two sites, would you want each switch to be connected to both paths, or just one?

  Ideally you want to spread your connections to each device over more than one port, more than one blade, more than one physical switch, and more than one network. You need to consider single points of failure, as well as the number of paths and switches in the configuration, when determining how many ISLs you want.

- Buffer credits

  Based on the total traffic between the two switches, the buffer credit calculation process should identify how many buffer credits are required to support that traffic. However, each port is assigned a given number of buffer credits, so having a larger number of ISLs means that each port will have fewer buffer credits. Increasing the number of ISL ports (for availability reasons, for example) might increase the number of buffer credits that must be installed in the switch.

You should discuss these, and other, considerations with your switch vendor to identify the optimal number of ISLs, in total and on each switch.

2.5.4 Configuring for availability

When designing the connectivity of your extended distance equipment, consideration should be given to optimizing availability and maintaining a level of separation to limit the effect of errors.

Consider the configuration shown in Figure 2-14. Each switch is connected to both DWDMs. If there is an error on the East Route that results in a path being taken offline, this configuration could potentially result in all paths being taken offline (because each of the four z/OS channels can use either route, and therefore all four are likely to encounter the error).

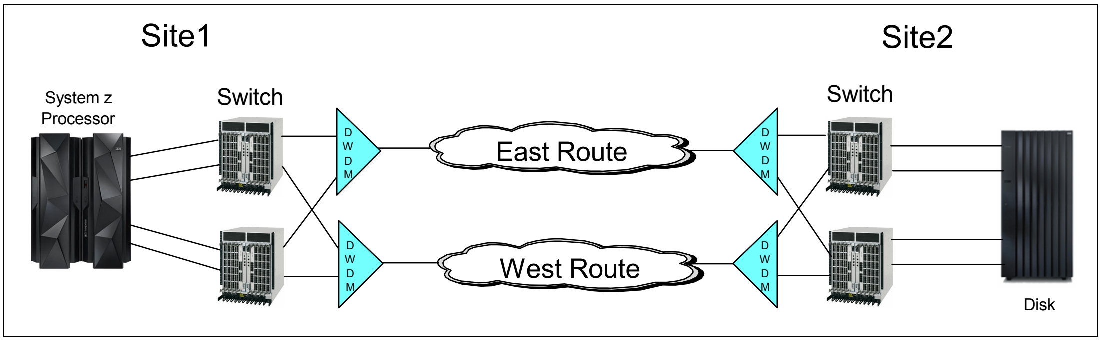


Figure 2-14 Connectivity option 1

A better alternative might be the configuration shown in Figure 2-15.

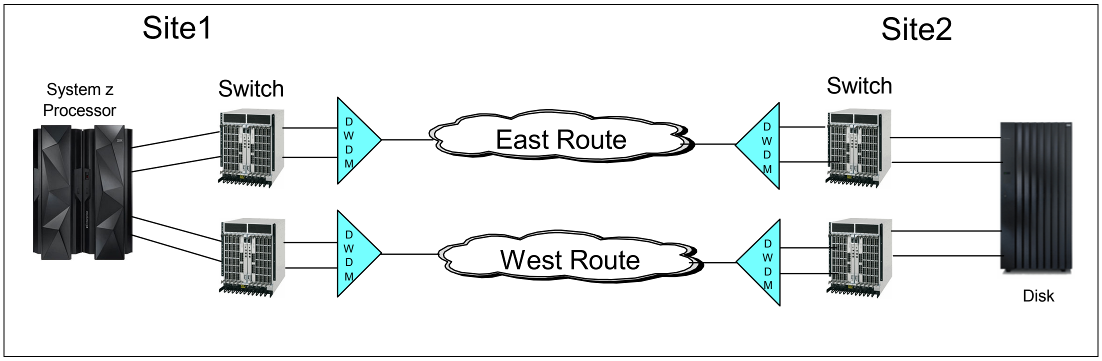


Figure 2-15 Connectivity option 2

In this configuration, an error in either route can only affect two of the channels. Of course, if you have an error on one route *and* a failure of a switch or a DWDM that is connected to the other route, you could lose all access to Site2. However that requires two failures. In the first scenario, a single failure could result in complete loss of access to the disk in Site2.

2.5.5 Optics

When a light signal travels through a fiber optic cable, the signal loses some of its strength (decibels (dB) is the metric used to measure light power loss.). The significant factors that contribute to light power loss are the length of the fiber, the number of splices, and the number of connections.

The amount of light power loss (dB) across a link is known as the *link budget*. All links are rated for a maximum link budget (the sum of the applicable light power loss factors must be less than the link budget) and a maximum distance (exceeding the maximum distance causes undetectable data integrity exposures).

When you connect ISLs to DWDMs, the lowest-power SFP can be used to connect to the DWDM client ports, because the DWDM is typically in the same data center. If there is a large distance between the switch and the DWDM, or a large number of connections in the link, work with your switch or DWDM vendor to identify the appropriate type of SFP to use.

In general, aim to use the lowest-power SFP that will support the required distance. SFPs are designed to work with a certain level of power, and to allow for a loss of signal due to the expected distance. If you connect two SFPs designed for a 40 km distance with a cable that is only 10 meters long, this is likely to result in errors, because the signal will be far stronger than the receiving SFP is expecting.

The distance supported by a given SFP is based on the assumption that no more than 2.0 dB is lost due to connections, and to reasonable losses because of cable attenuation. The two most important considerations for supporting distance are not actually the distance stamped on the SFP, but rather the following measurements:

- Minimum Rx (receive) value (Min Rx)

  The Min Rx is the minimum signal strength that is expected by the SFP receive function.

- Maximum signal loss (link budget)

  The maximum signal loss is the maximum amount of light loss that is acceptable to ensure that the signal strength at the other end of the link will be at least equal to the Min Rx.

If actual optical power measurements are not available when you are planning a new installation, use these guidelines for estimating the amount of light that will be lost:

- Loss of 0.5 dB per connection
- Loss of 0.5 dB per km (1310 nanometer (nm) wavelength)
- Loss of 0.3 dB per km (1550 nm wavelength)

Actual losses should be less than these estimates. However, consider taking steps to address the signal loss if the expected loss for each planned connection is within 20% of the specifications shown in Table 2-3 (speeds measured in gigabits per second, or Gbps).

Table 2-3 Fiber optic cabling for FICON: Maximum distances and link loss budget

	1 Gbps		**2 Gbps**		**4 Gbps**		**8 Gbps**		**16 Gbps**	
Fiber core (light source)[a]	Distance in meters	Link loss budget in dB	Distance in meters	Link loss budget in dB	Distance in meters	Link loss budget in dB	Distance in meters	Link loss budget in dB	Distance in meters	Link loss budget in dB
9 micrometer (µm) single-mode (SM) (10 km long wavelength (LX) laser)	10000	7.8	10000	7.8	10000	7.8	10000	6.4	10000	6.4
9 µm SM (4 km LX laser)	4000	4.8	4000	4.8	4000	4.8	This SFP is not available at these speeds			
50 µm multimode (MM) OM3 (short wavelength (SX) laser)	860	4.62	500	3.31	380	2.88	150	2.04	300	2.6
50 µm MM OM2 (SX laser)	500	3.85	300	2.62	150	2.06	50	1.68	82	2.3
62.5 µm MM OM1 (SX laser)	300	3.0	150	2.1	70	1.78	21	1.58	33	2.4

[a] In an SM jumper cable, the minimum distance between connectors or splices is 4 meters (13.1 ft.)

Table 2-3 on page 62 shows the supported distances and link loss budget for SFPs that are available with System z CECs. The link loss budget is derived from the channel insertion loss budget defined by the FC-PI-4 standard. The FC-PI-4 standard can be viewed on the following website:

http://www.t11.org

Field experience shows that toleration of deviations from the specifications with 2 Gb and 4 Gb links was much greater than with 8 Gb and 16 Gb links. This is an important consideration, and not just for the extended distance ISL connections. You might have 8 Gb ISL links and 4 Gb FICON adapters on the CEC today, but at some point the 4 Gb adapters will be replaced with 8 Gb adapters.

When you upgrade optical speed in any part of the configuration, the preferred practice is a thorough cable cleaning that covers *all* connections, including patch panels. This is because the most common problems during speed upgrades are related to cable hygiene. Furthermore, field experience shows that time spent cleaning all cable connections would be less than the time spent debugging cable issues if a general cable cleaning was not done.

It is good practice to occasionally monitor SFPs to get advance warning of potential failures. There are several attributes that should be monitored:

- Temperature
- Power consumption
- Power of the output signal

Some devices, such as some switches, monitor and report on all three attributes. Others might only report on a subset. Of the three, probably the least important is the power of the output signal. If the SFP power measurements are out of specification, do not immediately conclude that there is a power level problem. What is more important is to monitor the trend, because power degradation over time can indicate an approaching SFP failure.

Low SFP power measurements accompanied by bit errors are a good indicator of power problems. Continued bit errors and low-power measurements after replacing an SFP indicate a connection problem or a kinked cabled.

Cables

The most common type of fiber in most System z environments is long wavelength (LX) single-mode (SM) fiber. One of the advantages of SM fiber is that the cable is the same, regardless of the distance. Multimode (MM) fiber, conversely, uses different types of cable depending on the distance. This makes data center management and cable management significantly more complex.

Another advantage of SM fiber is that the error rates that can be achieved are significantly lower than can be achieved with MM fiber. The FC specification states that the bit error rate (BER) must be less than or equal to 10^{-12}. At 8 Gb, this equates to 28 errors per hour. SM fiber is able to achieve much lower error rates than this. Many SM cable vendors advertise a BER of 10^{-15} or better. At 8 Gb, this equates to less than one bit error per *month*.

In some cases, the DWDM client port is short wavelength (SX) MM, even though all other cable infrastructure is SM. A common mistake in an environment such as this is to forget to order the SX SFPs for the switch. Remember to validate the type of SFPs that are required. Also, when using MM fiber, OM-3 cable is the minimum cable standard, but OM-4 cable is highly suggested.

2.5.6 Managing and prioritizing traffic on ISLs

There are several mechanisms that can be used to influence the performance observed by the different types of traffic using a given ISL.

Virtual fabrics

The best way to isolate traffic from disparate environments (mainframe and distributed, or FICON and FCP, for example) is to put the applications on separate physical switches. If that is not a viable option, the next best alternative is to put the applications on separate virtual fabrics (VSANs).

Virtual fabric capabilities for modern switches or director chassis allow you to create logical switches within a switch or chassis that share the same hardware. All other aspects of a logical switch are independent, and appear and are managed as though they were independent physical switches.

You can configure virtual fabrics to share ISLs, or dedicate ISLs to specific logical switches. Although not all applications support the use of shared ISLs, dedicating ISLs to a specific virtual fabric can result in coarse segregation of traffic. A possible alternative to multiple virtual fabrics is to use some mechanism to limit the ports within a switch that can use a given ISL or set of ISLs. Note that some switches might have complex multi-layer algorithms that determine which ports can use which ISLs.

The best alternative (separate real switches, separate virtual switches, or some other mechanism) will depend on the types of traffic that will be using the switches, your organizational and availability requirements, and the features that are supported by your switches. This is a complex topic, beyond the scope of this document, so you should ensure that you discuss this carefully with your switch vendor.

ISL trunking

ISL trunking is a frame-based trunking algorithm that enables the bandwidth of multiple E_ports on a switch to form a single logical ISL with an effective bandwidth equal to the aggregate bandwidth of all ISLs in the trunk. Frame order is maintained across the logical ISL. ISL trunking maximizes the use of available bandwidth while simplifying ISL management, and therefore it is suggested that you enable ISL trunking where possible.

However, ISL trunking might be restricted to certain ports, and there are stringent signal delay skew requirements. Individual links within an ISL trunk group must be the same length. DWDM equipment adds delay skew, and you should therefore assume that you will not be able to use ISL trunking if you use a DWDM with ISLs. For more information, see 2.5.8, “Using DWDMs with ISL trunks” on page 65. Contact your switch and DWDM vendors to determine if your combination of switch and DWDM support the use of ISL trunks.

2.5.7 Using DWDMs with ISL links

There are several reasons why you might decide that the use of DWDMs with ISL links is appropriate:

- If you are using a DWDM anyway for other types of traffic (for example, IP, distributed systems, voice, coupling links, or STP), you would probably want to use the DWDMs for the inter-switch traffic as well.
- If you buy a connectivity service from a vendor, the service will probably include DWDMs.
- The use of a DWDM can allow you to provide the required connectivity and bandwidth with fewer dedicated fibers, thereby reducing the connectivity cost.

Generally speaking, DWDMs are not visible to the switch. However, there are implementation considerations that might be specific to particular DWDM models or vendors. One example is support for forward error recovery. You should discuss this with both your switch and DWDM vendors.

Another consideration is where is the best place to enable encryption and compression if both your switch and your DWDM support these capabilities. There are several things that should be factored into your decision:

- The cost of this capability on each device.
- The performance effect. Your switch vendor and your DWDM vendor should be able to tell you the elapsed time to encrypt and decrypt, or compress and decompress, each frame.
- Whether the other traffic that is using the DWDM been compressed or encrypted before being sent to the DWDM. There is generally little value in trying to compress data that has already been compressed, or to encrypt data that has already been encrypted.

In general, try to adhere to these guidelines:

- Select one strategy for where you will perform compression, and adhere to that strategy across all devices that will connect to the DWDM. For example, either perform all compression on the DWDM, or all compression before the data is sent to the DWDM.
- Select one strategy for where you will perform encryption, and adhere to that strategy across all devices that will connect to the DWDM. For example, either perform all encryption on the DWDM, or all encryption before the data is sent to the DWDM.

  In addition to there being no value in encrypting a second time, there are situations where encrypting data twice is not recommended. You should check with the vendors that supply your encryption functions to determine if they support their data being encrypted a second time.
- Always perform compression first, and then encryption. Compressing encrypted data results in poor compression ratios.
- Depending on your physical environment and your business needs, there might be an argument for encrypting the data as close to the CPU as possible, to minimize the number of places where data is broadcast in the clear.

2.5.8 Using DWDMs with ISL trunks

Cascaded FCP/FICON directors use ISLs to connect the directors. In certain configurations, ISLs can be grouped or aggregated, typically for performance and reliability. Brocade calls this an *ISL trunk* (frame-based trunking), and Cisco calls this a *PortChannel*. We will generically call this feature *ISL trunking* or just a *trunk*.

Each vendor might implement these trunks in a unique way to provide proprietary features. The vendors' trunked ISLs might contain proprietary frames, proprietary frame formats, or special characters or sequences of characters in the inter-frame gaps.

Often, the difference between a cascaded environment contained in a single data center or campus environment, and one in a metro environment, is the use of a DWDM (often with amplifiers) to carry the ISLs over the extended distance.

The primary concern when attempting to use trunked ISLs with a DWDM is that *the ISL data streams must be unaltered by the DWDM for the proprietary functions to work correctly.* This is sometimes called *bits in, bits out*, to indicate that there is no change to the signals, especially between the cascaded directors.

The challenge with non-symmetric transit times for the ISLs in a trunk is illustrated in Figure 2-16. The scale is *time to arrive* and not *distance traversed per time unit* (which would produce a graph roughly the opposite of this). This diagram shows how the signals, sent at the same time on parallel ISLs, could arrive at the endpoint at different times. The director measures this difference at the time that the trunk is created. The difference is called *skew*.

The director can accommodate a small skew, but an ISL with skew that is too large might be removed from the trunk by the director. An ISL that is carried on circuitry that introduces variable skew will not be detected, because the director does not re-measure the skew. If the variance of the skew becomes too large, the traffic on the trunk could be the cause of interface control checks (IFCCs), or could experience out-of-order frames.

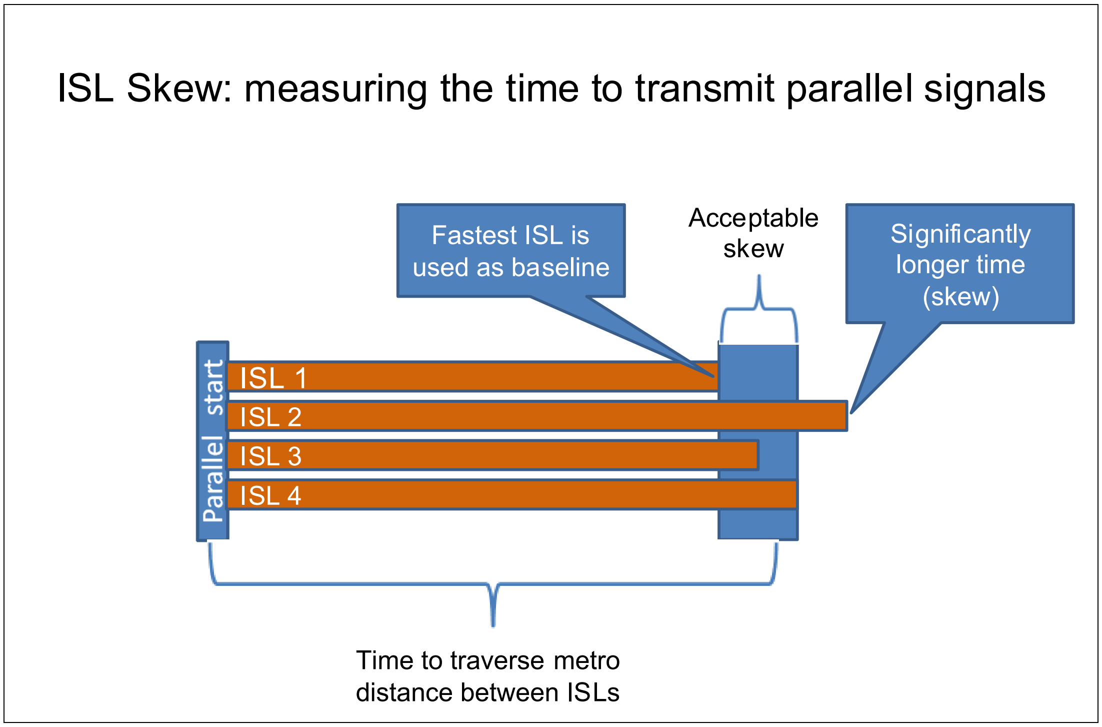

ISL Skew: measuring the time to transmit parallel signals
Fastest ISL is used as baseline
Acceptable skew
Significantly longer time (skew)
Parallel start
ISL 1
ISL 2
ISL 3
ISL 4
Time to traverse metro distance between ISLs

Figure 2-16 ISL transit time skew

It must be noted that the trunks between cascaded directors might appear to work without any issues during testing, because this is often performed with a relatively low I/O load. At that point, only one or two ISLs in a trunk are in use. However, large skew might only be detectable when multiple ISLs in the trunk carry traffic with high I/O loads. Some DWDM features can cause the skew to vary (that is, not be consistent), which can cause out-of-order frames or other issues with the I/O traffic.

Any alteration of the data stream introduced by circuitry or software in the DWDMs might affect the ISLs. The DWDM vendor might alter the data stream for different purposes, as listed in Table 2-4 on page 67. You should check with the DWDM and the FICON director vendors to determine basic ISL compatibility. Some of these features might be implemented in a way that alters the data stream that will not affect a single ISL, but would affect trunking. In general, these DWDM features should not be used on trunked ISLs.

Table 2-4 DWDM Features that could affect trunked ISLs

Feature or characteristic	Purpose	Effect on ISLs
Acceleration	Decrease signals sent through the DWDM to decrease end-to-end latency of the data stream	Alteration of the frames to compress the data stream or Removal of idle or arbitrate primitive signal (ARB) characters that are in the inter-frame gap
Re-framing	Acceleration or To reduce the Fast Etherchannel (FEC) activity and improve the amount of user data content of the data stream	Alteration of the frames or Removal of idle or ARB characters that are in the inter-frame gap
Re-packaging (such as to put into a Synchronous Optical Network (SONET)/ Synchronous Digital Hierarchy (SDH) format)	Use common components from carrier-style (communications service provider, or telco) DWDMs	Alteration of the frames or Removal of idle or ARB characters that are in the inter-frame gap
Re-timing	Adjust delays introduced by time-division multiplexing (TDM) circuits	Byte stuffing (adding or removing Idles and ARB characters in the FCP/FICON data stream)

IBM has experience with DWDMs that could not be used for ISL trunks because of the issues noted, and some experience where DWDMs appeared to support ISL trunking. There are many features on each DWDM and on each FICON director, giving a large number of permutations that would be difficult to test.

For a single example, and definitely not to provide an exhaustive test, we tested a specific configuration with two Brocade FICON Directors whose trunked ISLs were carried on two ADVA FSP 3000s at a distance of 80 km. The test configuration, with significant and varying I/O load, did not find significant increases in IFCCs or out-of-order frames, and the skews between the ISLs in the trunk were within acceptable limits. However, note the following caveats:

- This configuration is *not* qualified by IBM. At the time of writing there are *no* qualified DWDMs for ISL trunking.
- This test does *not* qualify and does *not* endorse this configuration. It was a test to show an example configuration that could be used to achieve trunking between cascaded FICON Directors at a metro distance.

The ADVA FSP 3000 that we used had a specific configuration (shown in Figure 2-17 on page 68). These components introduce a small amount of latency (on the order of 10 -16 ns, according to the DWDM vendor), but do not alter the data stream. In particular, note the following configuration details:

- All the circuitry used is not visible. More importantly, it does not alter the I/O data stream (bits in, bits out).
- All four ISLs in a trunk are on the same card, and every signal has its own wavelength. All ISLs using that card will experience similar latency introduced by the circuitry. If the latency should vary slightly, all ISLs should experience similar variances.

- The multiplexer (mux)/demux filter (the “prism” part of the DWDM) is a passive component that will not vary the latency on any specific wavelength.
- The dispersion compensation unit (DCU) that corrects for the fiber effect on different wavelengths is a passive component that will not vary the latency on any specific wavelength.
- Other components such as amplifiers (Erbium Doped Fibre Amplifiers, or EDFAs) or circuit protectors (Remote Switch Modules, or RSMs) are working on the multiplexed wavelengths, and affect all wavelengths in a highly similar manner.

Figure 2-17 shows the specific ADVA FSP 3000 configuration used in the 80 km test.

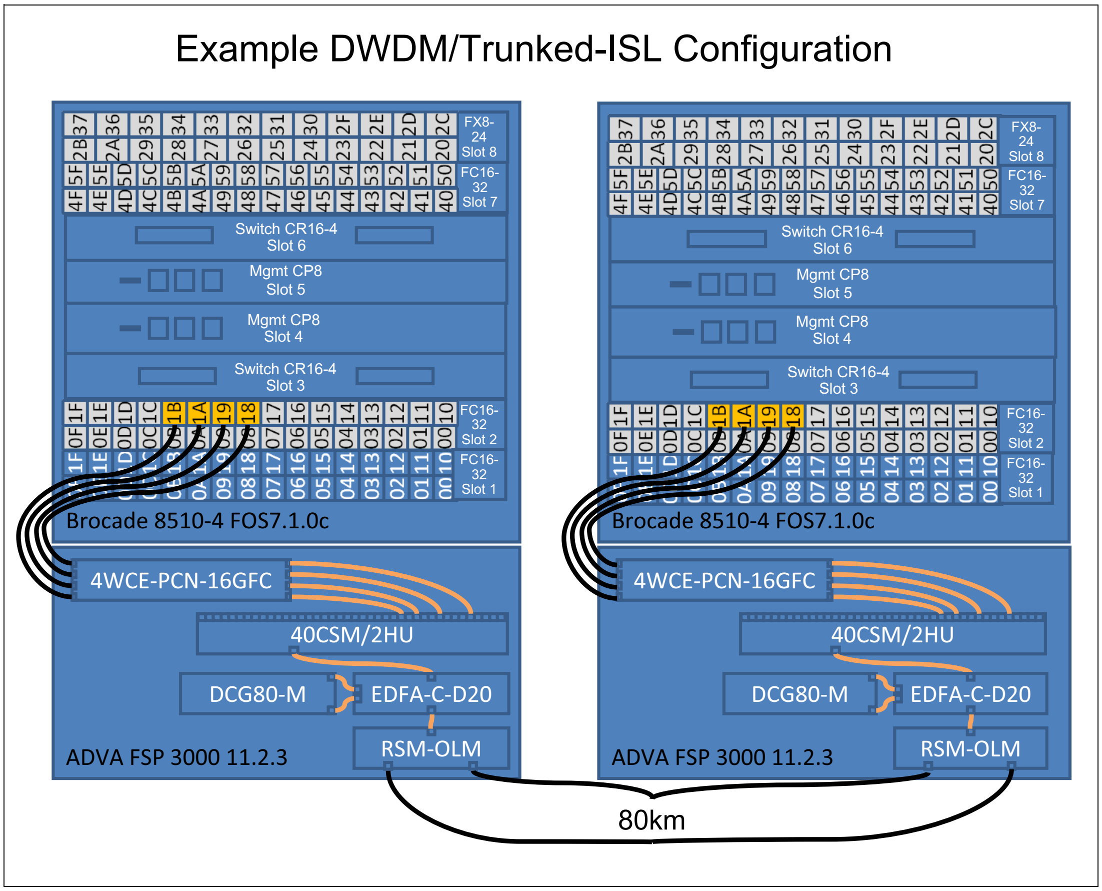


Figure 2-17 Example configuration for trunked ISLs over DWDM

Figure 2-17 shows the path of the four ISLs that were trunked:

- The directors had logical switches (with configured ports highlighted) with the card types and port addresses shown. System z channels and I/O storage adapters were attached at various ports in these logical switches.
- The DWDMs had the card types, two-stage amplifiers with DCU between the stages, and failover circuits as shown.
- The ISLs were all 16 Gbps.

During the testing, the I/O load going through the trunked ISLs was at times not trivial and used multiple ISLs in the trunk. No IFCCs or skew-induced I/O errors were observed during the tests.

For more information about the considerations for using ISL trunks in general, and especially with DWDMs, see the following websites:

- ADVA has a compatibility matrix:

  http://www.advaoptical.com/en/enterprise/advantage-qualification-program-aqp-for-storage-networking.aspx

  The matrix lists host adapters (including IBM System z), storage units, and directors for which ADVA has done their own testing.

- Brocade has a compatibility matrix:

  http://www.brocade.com/downloads/documents/matrices/compatibility-matrix-fos-7x-mx.pdf

  The matrix describes the I/O adapters, storage units, and DWDMs for which they have tested. Note the included caveat, “IMPORTANT: Compatibility with the Brocade Advanced Fabric OS feature Trunking varies by product. Contact your vendor for more detail”.

- Cisco has a table containing valuable information:

  http://www.cisco.com/en/US/docs/switches/datacenter/mds9000/compatibility/matrix/hwswcomp.pdf

  The table describes using DWDM optics in a Cisco 9500 to connect to an existing DWDM. The table does not describe using any particular DWDM or their configurations.

- There are many other DWDM vendors who have their own compatibility documents. Ask your vendor for this information if you are considering combining ISL trunking with a DWDM.

2.6 Considerations for FCIP-extended SANs

If you plan to use TCP/IP to interconnect your switches, there are some considerations that are unique to that environment. A traditional SAN in the data center is interconnected across FC links, using ports referred to as E_ports. An IP SAN is created whenever two or more switches are interconnected with an IP network, using Ethernet rather than FC ports.

2.6.1 QoS and propagation delay

A private IP network is required for FCIP. System z applications, hosts, devices, and switches using FCP have much more stringent timing requirements than are provided by a public Ethernet/IP network. The IP SAN would typically be just one of several users of the private IP network.

Consider the following QoS requirements when planning an IP network (remembering to consider the failover route and ensure that it meets these requirements as well):

- The maximum packet loss cannot exceed 1%.
- The committed data rate must meet the application bandwidth requirements, but must be at least an OC-3 (155 megabits per second, or Mbps) rate.
- The route should be static, which provides a higher QoS and ensures that the maximum acceptable propagation delay is never exceeded.

  A redundant failover path for an occasional outage is not considered a dynamic path.

- Measure the propagation delay to determine the *effective distance*. Equipment in the data path adds delay, and measuring the actual delay provides an extra check that the route being used is actually the planned route. It is ultimately the propagation delay, not the distance, that affects protocol timing. The stated maximum distances are based on, for example, the maximum supported propagation:
  - The speed of light through fiber is approximately 200 meters per microsecond (µs).
  - Example 2-1 shows the propagation delay as reported from a switch utility:

    Example 2-1 Sample report showing propagation delay

    ```
    switch:admin> fcping 20:00:00:00:c9:3f:7c:b8
            Destination:    20:00:00:00:c9:3f:7c:b8

            Pinging 20:00:00:00:c9:3f:7c:b8 [0x370501] with 12 bytes of data:
            received reply from 20:00:00:00:c9:3f:7c:b8:12 bytes time:825 usec
            received reply from 20:00:00:00:c9:3f:7c:b8:12 bytes time:713 usec
            received reply from 20:00:00:00:c9:3f:7c:b8:12 bytes time:714 usec
            received reply from 20:00:00:00:c9:3f:7c:b8:12 bytes time:741 usec
            received reply from 20:00:00:00:c9:3f:7c:b8:12 bytes time:880 usec
            5 frames sent,5 frames received,0 frames rejected,0 frames timeout
            Round-trip min/avg/max = 713/774/880 usec
    ```

    - Based on the highest reported time of 880 µs, that equates to a round-trip distance of (880 µs * 0.2 km/µs) = 176 km.
    - Assuming that the outbound and the return paths are approximately the same length, this indicates that the two sites are roughly 88 km apart.
- Another consideration is that when you are using ISL links you will have your own dedicated fiber, and you will generally know the precise length and propagation delay for each route between your sites. Therefore, that delay tends to be consistent. However, when using an IP network, you might not be able to control the exact route that will be used for every packet, so it is normal for the propagation delay to be more erratic than when using ISL links.

  If you will be using FCIP for FICON traffic and are not using FICON Acceleration, you should work with your network provider to ensure that the maximum propagation delay will not be more than the delay that would be experienced over a dedicated 300 km fiber (3000 µs).

2.6.2 FCIP circuits, tunnels, and trunks

An FCIP circuit is a logical connection between a pair of Ethernet ports on two switches. The number of Ethernet ports that you need on each switch will depend on the volume of traffic between the switches, and the type of traffic that will be using the switch. For high availability, you will want to have at least two Ethernet ports on each switch. The circuit can include all or a portion of the available bandwidth of an interface. Multiple FCIP circuits can share a given Ethernet port.

A tunnel is a collection of one or more FCIP circuits. For better availability, it is best to connect the circuits to different Ethernet ports. If the circuits in the tunnel are connected in this manner, the tunnel is called a trunk. Because an Ethernet port can be shared between multiple IP circuits, that means that a port can also be shared between multiple trunks.

Similar to an ISL trunk, an FCIP trunk is a logical aggregation of multiple IP circuits that form a single logical FCIP link, while maintaining in-order frame delivery. Unlike ISL trunks, there are no restrictions regarding the difference in latency between FCIP links. However, the effective latency of the trunk will be that of the circuit with the highest latency.

Because FC frames must be delivered in order, frames transmitted on the shorter FCIP link must wait for delivery of frames transmitted on the longer FCIP link to be ordered properly. An application that cannot keep the highest-latency FCIP link full will not be able to keep the shorter link full when trunked with higher latency links.

You will recall that we said previously that E_ports are used for ISL links. When using FCIP, Ethernet ports are used rather than FC E_ports. Rather than real ISL links, virtual ISL links are used. And rather than E_ports, VE_ports are used. A tunnel is used to connect a VE_port in one switch to a corresponding VE_port in the other switch, as shown in Figure 2-18.

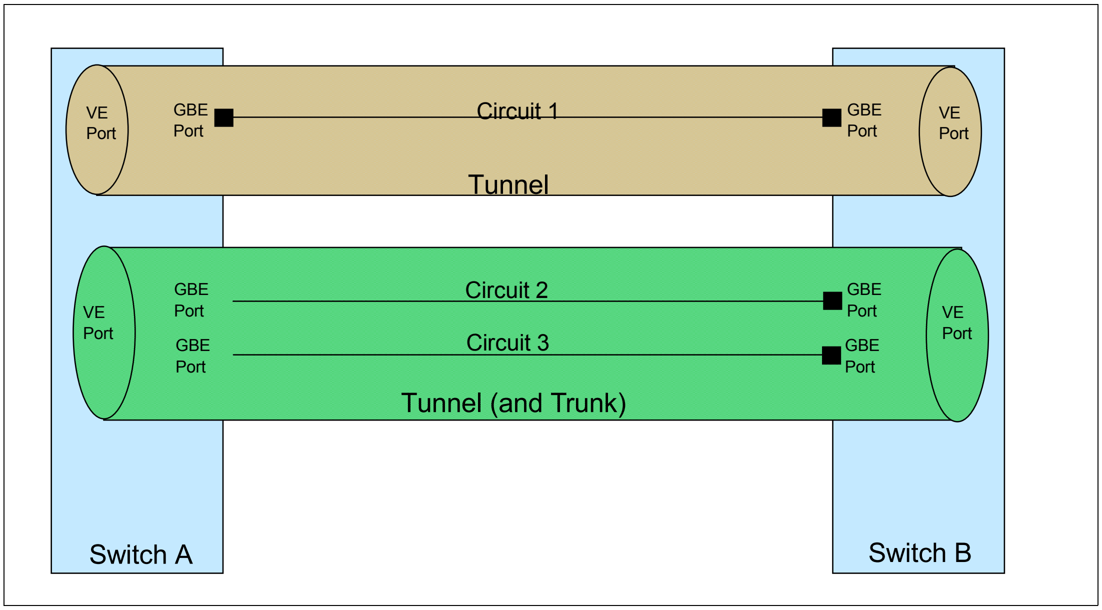


Figure 2-18 Relationship between FCIP circuits, tunnels, and trunks

If you use two 1-gigabit Ethernet (GbE) circuits, one with a 10 ms latency and the other with 50 ms of latency, the FCIP Tunnel will present a 2 gigabit (Gb) connection with 50 ms of latency to the fabric. The tunnel will run fully used if the application data could have 12.5 MB of data outstanding at one time ((2 Gbps/8) *.05 = 12.5 MB).

If the application cannot drive that much outstanding data, the FCIP tunnel will not be able to saturate the circuits. With 2:1 compression, the application would have to have 25 MB of outstanding data to fully use the circuits.

Trunks allow for flexible handling of network failover, because they use multiple physical links. Using trunks with Traffic Isolation (TI) zones and QoS provides control over how and when bandwidth is used, in both fault-tolerant and non-fault-tolerant network scenarios.

The requirements to form an IP trunk are not as restrictive as ISL trunks. However, detailed feature requirements and configuration are beyond the scope of this book. Contact your switch vendor for details.

2.6.3 IP planning and design

FCIP consists of encapsulating FC frames in IP packets. The IP protocol adds an additional 98 bytes. With most traffic, the average payload size is close to the maximum, so the effect of the additional 98 bytes is minimal. The added IP activity is significant only with applications that use small frame sizes (CTCs, for example).

The details of how FC frames are mapped to IP packets varies from switch model to switch model. Contact your vendor for more information, and for information about how performance is affected by the relationship between *your* average FC frame sizes and the methodology that the switch uses.

Typically, there are two distinct network routes for the IP traffic, as shown in Figure 2-19. Unlike fiber ISL trunking, IP trunks can use both routes, and therefore can have grossly different latencies. This enables designers to plan for redundancy in the network, making recovery transparent to the application, and enables the SAN to recover from network faults.

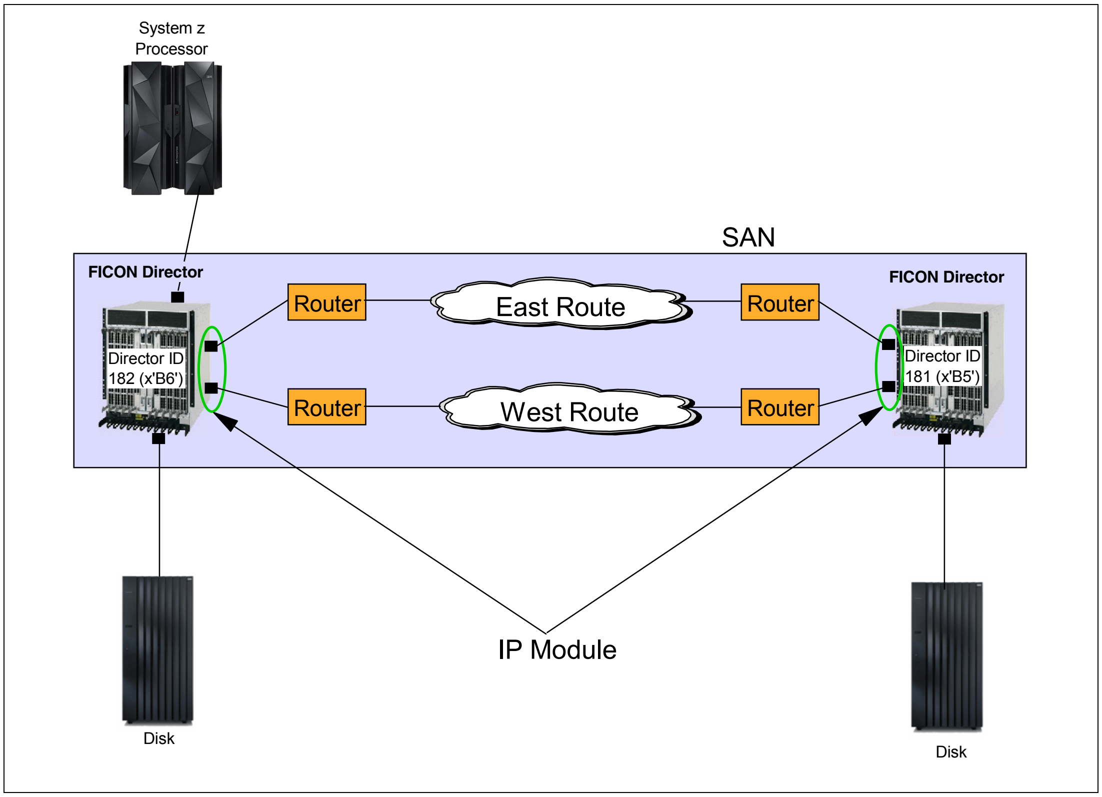


Figure 2-19 Redundant networks

If you plan to use FICON Acceleration (described in 2.6.5, “FICON Acceleration” on page 75), we suggest that you create a single tunnel that spans multiple circuits (physical network interfaces) and (if possible) multiple networks. The result should be that a bandwidth reduction would be the only effect of a network outage.

This occurs because the switch can use one of the other interfaces in the tunnel, meaning that no change would have occurred from the channel path perspective except that the bandwidth is reduced. If only one physical network path is associated with the tunnel, a single network failure could result in losing the entire channel path.

Although FICON Acceleration has stringent requirements governing failover in case of a single network outage, if you are *not* using FICON Acceleration, you have more flexibility for failover from a single network outage. However, the preferred practice is to have the SAN IP networking component manage the network, because the IP component recovers lost frames, but the FC component does not.

FC frames in flight on that network might be dropped if the SAN is left to reroute traffic in a network outage. The number of dropped frames will not exceed the number that FCP can recover, but there is no value for the attached CUs to recover frames. FICON channels use FCP to transport and recover frames, but frame loss on a FICON channel is detected and reported as an IFCC due to out-of-order frames.

Because the outage is detected in a different layer in the SAN, the report originates from a different place in the SAN, but the net result is that the SAN reports the network failure. Because the SAN reports the outage in either case, most designers set the SAN IP networking component to manage network outages.

2.6.4 Adaptive rate limiting

Adaptive Rate Limiting (ARL) is a feature on the switch that automatically and dynamically maximizes available bandwidth by adjusting for network congestion. The use of ARL is a preferred practice if the WAN is shared by two or more FCIP circuits, or if the IP network is shared with other traffic.

Each FCIP circuit is configured with a minimum and maximum bandwidth (BW) value. The minimum bandwidth provides a specified level of bandwidth. The sum of all the minimum bandwidth values for each circuit cannot exceed the total bandwidth available. The maximum bandwidth setting enables you to cap the bandwidth that can be used by a circuit, while still allowing the circuit to use up to the maximum value when the bandwidth is available.

ARL gradually increases the bandwidth allocated to a circuit until congestion is detected. When minor congestion is detected, the rate begins to adjust downward. When massive congestion is detected, the rate immediately adjusts down to the minimum rate, as shown in Figure 2-20.

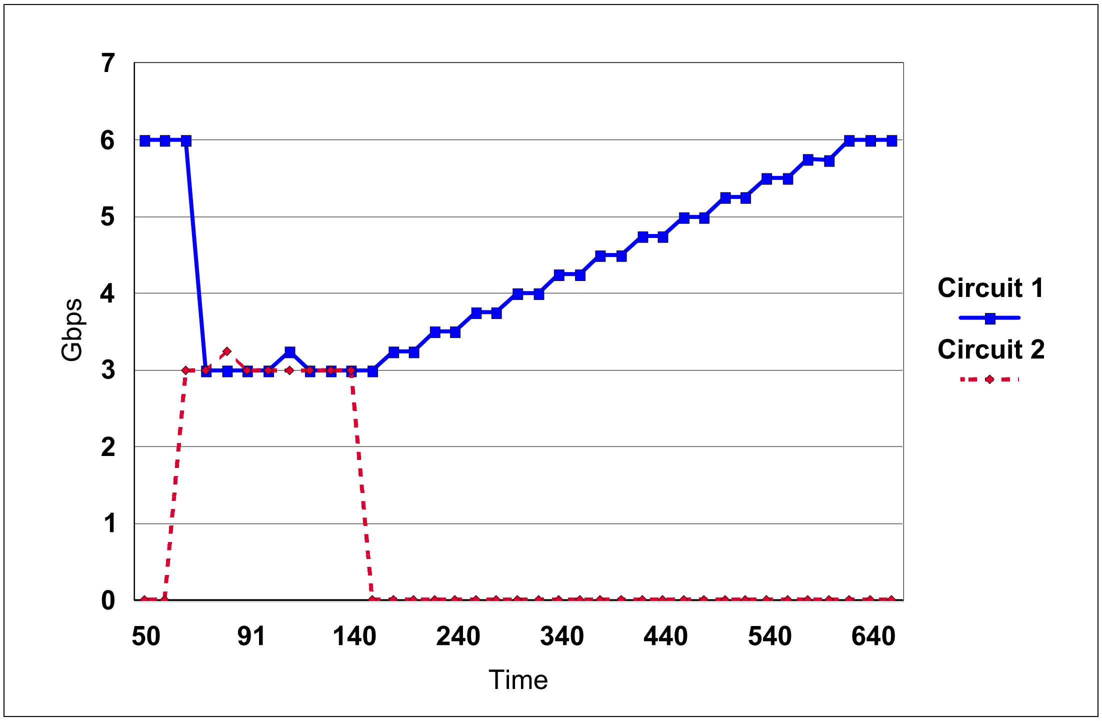


Figure 2-20 Adaptive Rate Limiting example with traffic that suddenly increases and decreases

In Figure 2-20, the total bandwidth available is 6 Gbps. Each circuit has been defined with a minimum bandwidth of 3 Gbps. Circuit 1 has been running for some time while Circuit 2 has been idle, resulting in Circuit 1 being allocated more than its minimum bandwidth. Circuit 2 then has some data to pass and begins to transmit at its minimum configured bandwidth of 3 Gbps. Circuit 1 then encounters heavy congestion and reduces its bandwidth usage to its minimum configured bandwidth of 3 Gbps.

Circuits 1 and 2 attempt to take additional bandwidth but always run into congestion so a sustained rate greater than 3 Gbps is never achieved by either circuit until Circuit 2 stops sending data. After Circuit 2 stops sending data, Circuit 1 begins incrementally increasing its bandwidth until the maximum available bandwidth or the maximum configured bandwidth is again reached.

In Figure 2-21 on page 75, the circuits are configured as they were in the previous example. In this case, however, demand for bandwidth on Circuit 1 remains high, while demand for bandwidth on Circuit 2 fluctuates, as shown in Figure 2-21 on page 75.

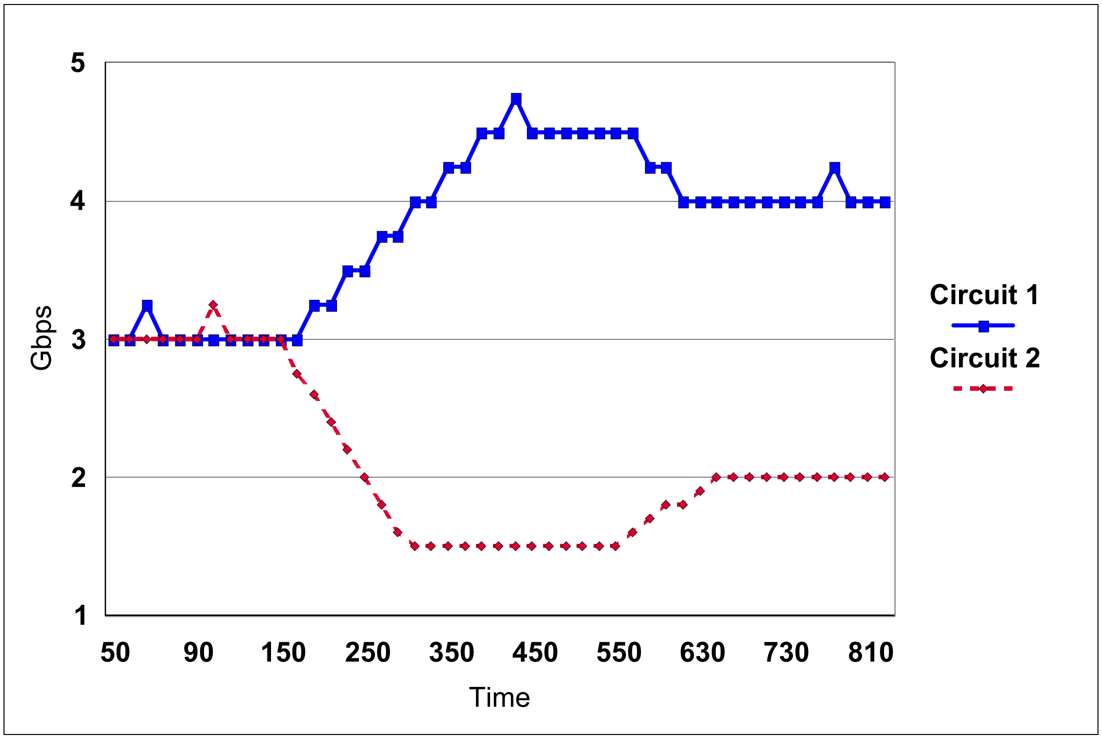


Figure 2-21 Adaptive Rate Limiting example with fluctuating traffic volumes

As demand for bandwidth on Circuit 2 decreases, Circuit 1 has been running for some time before the bandwidth demand for Circuit 2 decreases. As soon as Circuit 1 attempts to use additional bandwidth without running into congestion, it incrementally increases use of additional bandwidth until congestion is encountered. Circuit 1 then backs off and periodically tries an incremental increase in bandwidth again.

This example shows how the use of ARL enables Circuit 1 to use more than its minimum bandwidth. However, the amount of bandwidth available to Circuit 1 is determined by the amount of bandwidth needed by Circuit 2 (which is always less than its specified minimum, meaning that it will always be given what it needs).

2.6.5 FICON Acceleration

Getting help: FICON Acceleration is a complex topic, and the devices, applications, and distances that it supports are constantly changing. This section provides an overview of how it works. However, clients are urged to consult their switch vendor, or enlist the services of the IBM Lab Services (LBS) team to perform detailed planning about the potential role of FICON Acceleration in your configuration.

The FICON flow control protocol requires that acknowledgements are returned after a certain number of commands and data is sent. As the distance between two sites increases, the time for data and commands to reach the remote site, and for the acknowledgments to return, increases.

Buffer credits and optical power capabilities on FICON Express8 and Express8S channels are limited to 10 km. Greater distances are possible when the channel is connected to switches with additional buffer credits and more powerful optics, or when a switch is used in conjunction with a DWDM with the required optics.

However, when the total distance exceeds about 120 km, the protocol starts to experience "FICON droop". FICON droop is the performance degradation as the time for handshakes to traverse the link increases compared to the time required to transmit commands and data.

FICON droop is particularly noticeable for FICON disk I/Os. The nature of modern tape processing (large sequential data transfers) or printers (low throughput compared to disk or tape) means that FICON droop does not affect those I/Os to the same extent until larger distances are reached.

Generally speaking, disks that have high levels of I/Os and that are required to deliver short, consistent, response times would be placed at limited distances from the CEC. Although larger distances are supported, the applications that can tolerate the increased response times that result from large distances become more limited.

One application that might require both long distances *and* high throughput rates is zGM. For that application, Extended Distance FICON (described in 2.7.3, "Extended Distance FICON" on page 83) is available. Extended Distance FICON provides relief from FICON droop for zGM for distances between 120 km and 300 km.

300 km is the maximum qualified repeated distance for FICON-attached devices. The use of FICON-attached devices at distances greater than 300 km requires a feature known as FICON Acceleration in the switches. As stated previously, only selected applications are supported by FICON Acceleration.

Because FCIP is the only inter-switch connectivity option that supports distances greater than 300 km, FICON Acceleration would typically be used on switches that are inter-connected using FCIP. However, there might be cases where the performance benefits of FICON Acceleration might be attractive, even at distances less than 300 km.

How FICON Acceleration works

FICON Acceleration is a licensed switch feature that provides near-distance performance for channels and CUs that are separated by up to thousands of kilometers. This is done by emulating certain aspects of the channel and CU protocol. Emulation, therefore, is limited to certain devices that stream data:

- zGM (formerly extended remote copy, or XRC)
- Tape
- Terradata
- Selected printers

Emulation monitors the exchange of channel command words (CCWs) between the channel and CU. When a CCW pattern for transporting bulk data streams is recognized, the emulation software in the switch that is in the same site as the CEC provides CU acknowledgements back to the channel, or initiates read commands to the CU, depending on the data flow direction. The switch at the CU site builds the CCW strings that are sent to the CU. This supports a continuous data stream in the link between the two sites.

From the channel perspective, data that is still in-flight appears to have been written to the CU. From the CU perspective, the command to read data that was created by the switch appears to have come from the channel.

FICON Acceleration for tape

The functions provided by FICON Acceleration vary based on the target device type, and also on the type of request that is being processed.

Tape write

All tape CUs have a cache that contains data that has been written but not yet committed to media. Certain channel commands require the data in the cache to be written to the media before completing. When the CU receives these commands, it responds with a status that instructs the channel to suspend further processing until the CU completes the operation.

If you are using FICON Acceleration, the FICON Acceleration function in the switch responds to the channel in the same way that the CU would. It then completes writing all data previously accepted from the channel to the CU. After all data in flight has been written to the CU, the last command is sent to the CU. After the CU completes the operation and provides final status to the switch, the switch fabric sends that final status to the switch where the channel is connected, and presents that status to the channel, as shown in Figure 2-22.

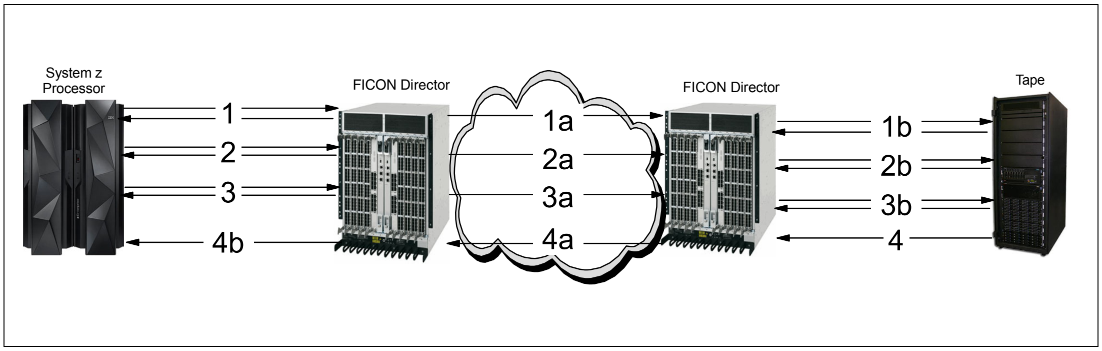


Figure 2-22 FICON tape write acceleration

Figure 2-22 helps illustrate how tape write processing is handled when using FICON Acceleration:

- Items 1, 2, and 3 represent write I/Os being sent to the CEC-site director, and the corresponding Channel End (CE) and Device End (DE) statuses being sent back to the channel for each I/O. Note that the CE and DE are likely to be sent to the channel before the data is written to the CU in the remote site.
- Items 1a, 2a, and 3a represent the data being sent from the CEC-site director to the CU-site director.
- Items 1b, 2b, and 3b represent the data being written to the CU. At the end of each I/O, the CU returns CE and DE to the CU-site director. These CEs and DEs are *not* sent back to the channel, because the CEs and DEs for those I/Os were already returned to the channel by the CEC-site director.
- After the last I/O has been sent to the CU and written to the media, the ending status DE is sent to the CU-site director, as shown by item 4.
- The CU-site director forwards the ending status DE to the CEC-site director (item 4a).
- Upon receiving the ending status DE, the CEC-site director sends the final device status (item 4b) to the channel.

Tape read ahead

When data is being read from tape, the FICON Acceleration function posts reads to the CU when it determines that the host is attempting to perform sequential reads. Anticipating that the host will read to the end of tape or the next tape mark, the director that the tape unit is connected to generates commands to read tape data and to send it to the local site. This process maximizes the use of the available bandwidth between sites, as shown in Figure 2-23.

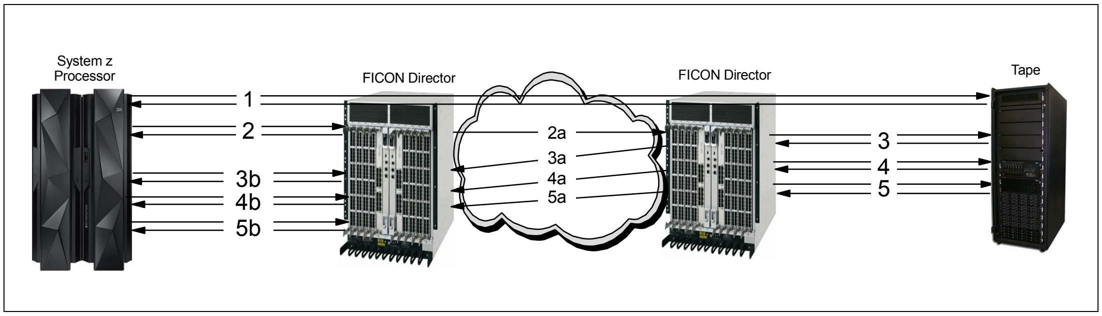


Figure 2-23 FICON tape read acceleration

Figure 2-23 shows an example of a tape read-ahead configuration:

- Item 1 represents a read I/Os to the tape CU. The FICON Acceleration feature monitors all I/Os, looking for indications that sequential reads are being sent for a device.
- Item 2 represents a read I/O to the tape in the CU site. The FICON Acceleration intercepts the read request and sends back a response indicating that the channel should try the read request again.
- Item 2a represents a command being sent to the director in the CU site. The command instructs the director in the CU site to send multiple read requests to the CU.
- The director in the CU site accepts the command and sends read requests (items 3, 4, and 5) to the CU. In response to each read I/O, data is returned to the CU-site director.
- As the data for each I/O is received at the director, it is forwarded back to the CEC-site director (items 3a, 4a, and 5a), where it is held in a cache until it is retrieved by the host.
- Items 3b, 4b, and 5b represent read I/O requests being sent from the channel to the CEC-site director. The requested data is retrieved from the cache and immediately returned to the channel.

Error recovery and other tape commands

Consider a normal tape write operation. When the channel sends a block of data to the tape drive, there will be some data in the channel that is on the way to the CU, and some data that is cached in the CU, waiting to be written to tape. When all of the data has been received by the CU, it returns Device End (DE) and Channel End (CE) responses to the channel (even though the data might not have actually been written to tape yet). When the channel sees the DE and CE, it knows that it can start sending the next block of data to the CU.

When using FICON Acceleration, the director accepts all of the data from the channel and immediately returns DE and CE to the channel. At that instant, some of the data might have reached the CU. Some will be in the remote director, waiting to be sent to the CU. Some will be in flight, between the two directors, and some might be in the CEC-site director, waiting to be sent to the remote director.

Because we do not have to wait for all of the data to reach the CU before DE and CE are issued, the channel can start writing the second block of data much faster. From a recovery perspective, the important point is that there are various caches in the path between the channel and the CU interface, and the data that is part of one I/O request might exist in several of those caches at any instant during the I/O.

Now, in a tape device or CU that is not using FICON Acceleration, consider what happens if there is an error before all of the data has been saved on the tape media. The tape CU will send a response to the channel, indicating that it has encountered an error. That will cause the channel to send a command to the tape CU, instructing it to return any data that has not yet been saved. It can then use that data to redrive the request.

If an error occurs with a CU that *is* using FICON Acceleration, the error will be returned to the director, and the director will forward it to the channel. In this case, however, the data that must be recovered is the data that was in the CU cache, the data that was in-flight, and the data that was in the cache of either director. The director will return the *total* amount of data (in-flight data plus data in the CU) to the channel.

The emulation software intercepts the commands that are intended to read the CU cache, and instead returns any in-flight data to the channel. This continues until all in-flight data has been returned to the channel. After all in-flight data has been returned to the channel, subsequent commands to read cached data are sent directly to the CU, and data that was in the CU cache will be returned.

That brings us to an important point in the design of your director configuration when using FICON Acceleration. To be able to return the correct information to the channel about data that has not yet been written to the tape media (that is, data that is in-flight or in the director or CU cache), *it is vital that every I/O to the CU from that channel takes the same path through the director configuration*. You should work with your vendor to ensure that your configuration adheres to this requirement.

FICON Acceleration for z/OS Global Mirror

FICON Acceleration for zGM works similarly to tape read ahead. The reads generated by the System Data Mover function are emulated by continuously reading modified data using the Read Record Set (RRS) command and sending that data to the site running the System Data Mover function.

While zGM is reading updates from the primary disk, the emulation software posts Read Track Set (RTS) commands, keeping the network bandwidth fully used. This keeps the available bandwidth completely in use when there are modified records to read, and maximizes the rate at which data is read from the primary disk.

Using this RRS/RTS approach, the primary and secondary disk can be virtually any distance apart, as shown in Figure 2-24.

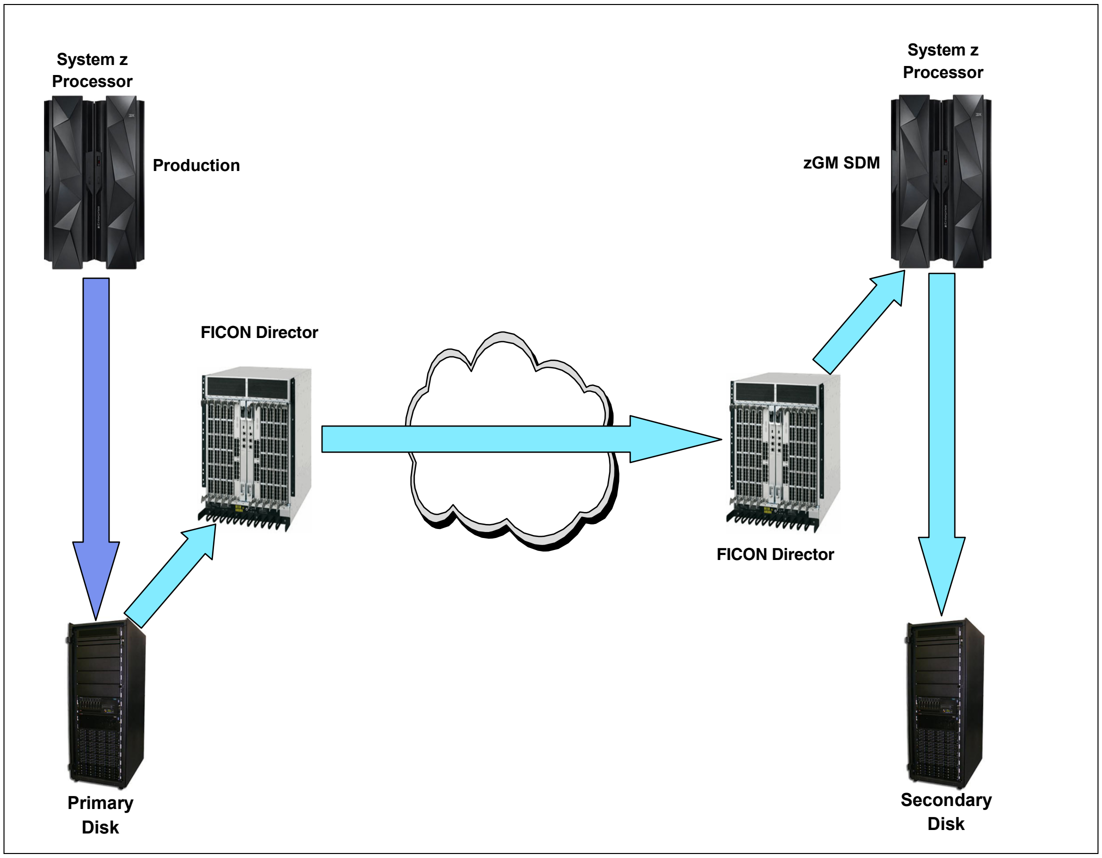


Figure 2-24 FICON Acceleration for z/OS Global Mirror

Returning home after a disaster

At one time, channel extenders that were used for large distance cross-site connectivity would have to be configured specifically for the site that the extender was in. For example, if the production disk were in site1, and the zGM system was in site2 together with the secondary disk, the extenders in site1 would be configured differently than those in site2. As a result, if you swapped the roles of the two sites (meaning that you are now mirroring from site2 to site1), the extenders would need to be reconfigured.

With FICON Acceleration, the feature is installed on switches in both sites. If you need to reverse the direction of mirroring, there is no need to reconfigure the switch or the FICON Acceleration feature because of that swap. There might be *other* reasons why you would need to alter the configurations; however, FICON Acceleration would not be one of them.

2.6.6 FC Fast Write

FC Fast Write was designed to address an inefficiency in the FCP protocol that affected performance at larger distances.

Fast Write mitigates the latency effects of FCP write operations when the initiator and target are not at the same location. Fast Write enables the entire data segment of an FCP write to be transported across the long-distance link between the initiator and the target without the inefficiencies of waiting for transfer ready (FCR_XFER_RDY) commands to travel back and forth across the high-latency link.

By not having to wait for potentially numerous roundtrip handshake messages, SAN switches can expedite FCP write operations and improve throughput by up to 200%.

The behavior of an I/O to an FCP device when Fast Write is *not* used is shown in Figure 2-25.

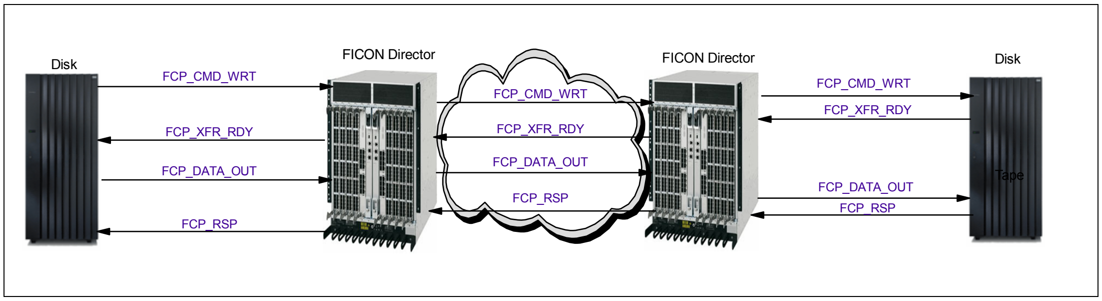


Figure 2-25 FCP transfers without Fast Write

In Figure 2-25, you can see that the data write (FCP_DATA_OUT) does not start until the FCP_CMD_WRT has been sent all the way to the target device, and the FCP_XFR_RDY is received back from the device. FCP_XFR_RDY is a transfer-ready protocol that was designed for older devices with a small or non-existent cache.

Today, most target devices have a large, efficient cache, so there is rarely a case when an information unit (IU) cannot write to the target. If the target is operating slowly, the IU is held in the switch until the target is ready, and normal FC buffer credit flow control slows the transfer of additional data from the initiator. The Fast Write function causes the switch to return FCP_XFR_RDY to the initiator of the I/O without having to wait to receive it from the target device. This is shown in Figure 2-26.

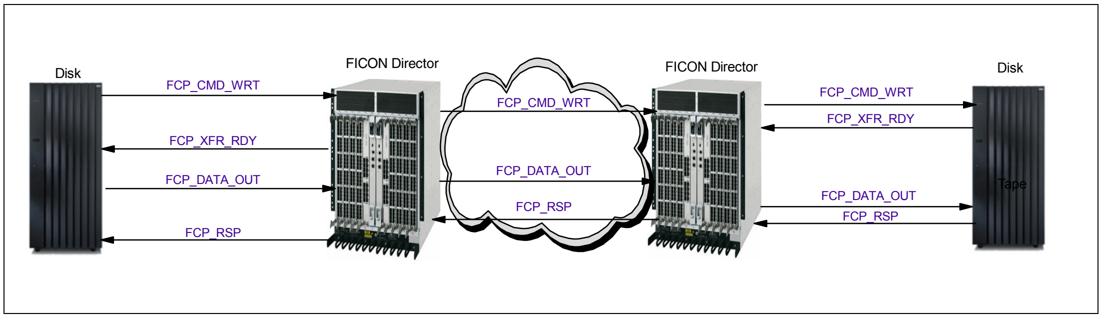


Figure 2-26 FCP write transfers with Fast Write

The extra transfer ready handshake is not required with reads because the host initiator does not request the data until it is ready to receive it.

In practice, the FCP processing for most modern devices has been enhanced to address this inefficiency without having to use FC Fast Write. You should consult with your switch and storage vendors to determine if this capability is required and available for your configuration.

2.6.7 Monitoring and managing FCIP ports

Ethernet ports on a switch are not visible to the CUP, so any applications that use the CUP to monitor or manage a port will not be able to work with the Ethernet ports. Consult your switch vendor for information about alternative tools that you can use to manage those ports.

2.7 Switch features

There are some common switch features that you might want to enable.

2.7.1 Compression

One of the optional licensed features on most switches is compression. Compressing and decompressing the data adds some latency. Alternatively, transmission times might be reduced because less data needs to be transmitted to the other site. Typically, the reduced transmission times more than compensate for the added time to compress and decompress the data. However, for applications with stringent response time requirements, you should carefully consider the latency effect of using compression.

Switch vendors offer a variety of compression algorithms. The different algorithms can deliver different levels of compression, but also have differing latencies. The details of the algorithms are beyond the scope of this book, but you should contact your switch vendor regarding compression options if a significant percentage of the inter-switch traffic is synchronous to the execution of a transaction or batch job.

For asynchronous applications (asynchronous mirroring, for example), an automatic compression detection mode is nearly always the best option. The automatic mode determines the best compression algorithm to use for the network. Although the automatic mode is generally used for synchronous applications as well, we encourage SAN architects to discuss the applications and network resources with their switch vendor.

Note that you should ensure that data compression takes place before encryption if you are using both functions. If compression and encryption are enabled on the switch, the switch performs compression before it encrypts data. However, if the data is already encrypted before being compressed, enabling compression is unlikely to significantly reduce the size of the data, and therefore will not improve the effective bandwidth.

Additionally, there is probably no point in trying to compress the data twice. If you are using an extension switch together with a director, enable compression in one place, but *not* in both.

Also, if you are using compression with ISL connections, bear in mind that compressing the data has the effect of reducing the average frame size, which will have an effect on the number of buffer credits required to keep that link fully used.

2.7.2 Encryption

If you are transmitting data outside your data center, it is likely that you will want to encrypt the data. Depending on whether you are using ISLs or FCIP, and whether you are using DWDMs or not, you might have several alternatives available in relation to where the encryption is performed.

The following list includes some factors that you should consider when deciding where the data will be encrypted:

- The latency effect of encrypting the data. For example, the latency might be different on a DWDM than it is on your switch.
- The financial cost of the encryption feature. Encryption is generally a chargeable licensed feature, and the price will probably be different on different platforms.
- You should not encrypt the same piece of data more than once.

- You should ensure that all data is encrypted. For example, if several different types of traffic are transmitted through your DWDMs, and one or more does not provide an encryption capability, it might be simpler to perform the encryption for all traffic types in the DWDM.
- Depending on your corporate requirements and the physical attributes of your data center, there might be value in encrypting the data as close to the CEC as possible.

There is no single correct answer for all enterprises. You need to evaluate the considerations that apply to your enterprise, work with your vendors and security colleagues, and identify the configuration that is most appropriate for you.

2.7.3 Extended Distance FICON

Extended Distance FICON is a feature on disk CUs. As such, it is strictly speaking not a switch feature. However, it is designed to be used when the distance between the CU and the CEC is between 120 km and 300 km, and those distances are not achievable unless you use a switch. Extended Distance FICON works with both ISLs and FCIP, so we have included it in this section.

Extended Distance FICON is designed to address the phenomenon known as FICON droop. It is specifically designed to deliver the improved throughput that is required if you use zGM at extended distances. Extended Distance FICON does not play any role in disk I/Os other than those for zGM, and it is not available on other types of CU (tape or printer, for example).

A deterioration in performance can occur before buffer credits are all used. FICON uses strings of CCWs. An IU represents one or more buffer credits required for a CCW. By default, each CCW string can have a maximum of 16 IUs outstanding, regardless of how many buffer credits are available. When this maximum is met, an acknowledgement is required before additional IUs, and therefore CCWs, can be sent.

Extended Distance FICON (also known as *Persistent IU Pacing*) enables the maximum number of outstanding IUs to be increased up to 256. Extended Distance FICON improves zGM performance within the maximum supported distance of 300 km, but the feature is not supported without an Request for Price Quotation (RPQ) at distances greater than 300 km. For distances greater than 300 km, FICON Acceleration should be used.

For more information about Extended Distance FICON, see the section about Extended Distance FICON in *FICON Planning and Implementation Guide*, SG24-6497.

2.8 Redundant network and switch topologies

The only way to achieve the levels of availability required by modern enterprises is to eliminate as many single points of failure as possible. However, there is a cost associated with providing redundancy. Just as different enterprises have different availability objectives, similarly there are different levels of redundancy that you can provide in your switch and network configuration.

This section provides information about how network-related failures are handled, and provides some common examples of different levels of redundancy in the network and hardware configuration.

2.8.1 Effect of a network failure

The fabric shortest path first (FSPF) algorithm routes frames through a SAN using the path with the lowest cost. The algorithm uses a cost that is calculated for each route. The factors that are used in arriving at the route cost, and the weighting that is assigned to those factors, can vary from one switch to another. Additionally, a particular switch might have additional, more sophisticated, routing algorithms.

Because it is vital for you to understand how the various routes will be used, you should work closely with your switch vendor to understand the algorithms that are used for *your* configuration. Note that any DWDM equipment that might be part of the configuration is not visible to this algorithm.

SAN designers sometimes overlook the effects of a network fault when configuring a SAN for extended distance solutions. In Figure 2-27, the shortest path between the host and the storage is from Switch A to Switch B. If the link between Switch A and Switch B fails, the new fabric path would be Switch A <-> Switch C <-> Switch B. This path adds distance, but it is also a configuration that is currently not supported by System z.

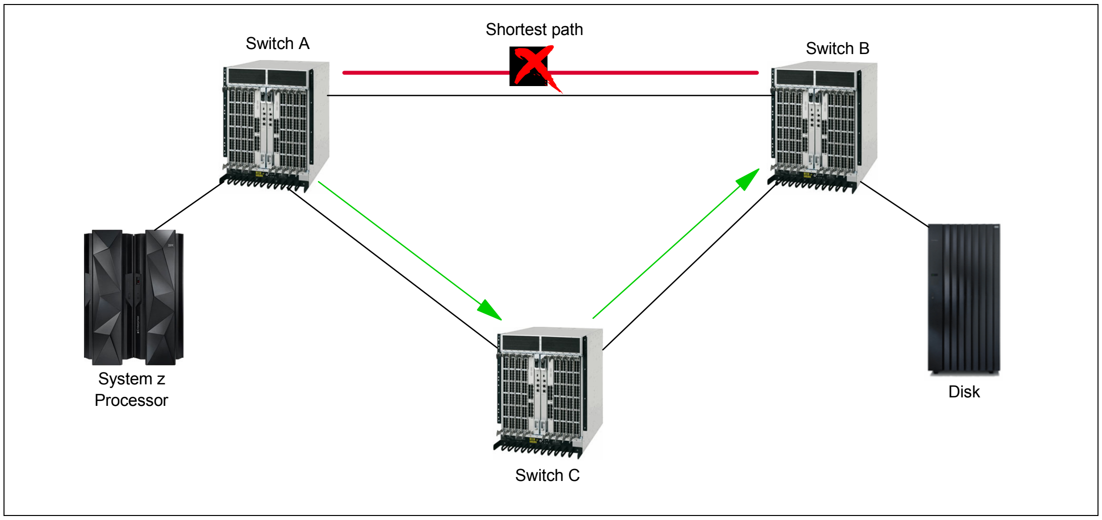


Figure 2-27 Switched shortest paths

You can build a triangular fabric, but you need to avoid configurations that normally conform to the System z restrictions, but that could violate the restriction in failure scenarios. Consult your switch vendor whenever you implement designs that involve (or could involve) more than one SAN hop.

A good redundant network plan considers both the available bandwidth with the backup network, and the effect of reduced bandwidth. Bandwidth is typically reduced to half during an outage in active-active network topologies.

2.8.2 Calculating paths for FCIP

The algorithm that calculates the most desirable path in an FCIP network is not as straightforward as that for a pure FC network. FCIP link costs take additional factors into account. As a result, two FCIP links with the same number of hops can each have different link costs.

Depending on the switch design, the FSPF algorithm can potentially override designs that are intended to force traffic down certain paths (FICON Acceleration, for example). Contact your switch vendor for mechanisms that can ensure that the algorithm will work as designed. Include this issue in your planning process, and be sure to discuss it with your switch vendor before implementation.

2.8.3 FCIP routed network

The preferred practice for providing a high-availability, multisite configuration is to accomplish the following goals:

- Build network redundancy into the network. Provide two, failure-isolated networks, and connect each switch to both.
- Build redundancy for hardware faults into the fabric design, for example by having at least two switches in each site.

Routers will recover and re-route frames in a network failure if an alternative route is available. Trunked IP interfaces in a switch also recover and re-route frames, but the capabilities to do so might be limited to the same hardware, and not span different hardware modules. Putting redundant networks on the same switch hardware modules only protects you from a network fault; it does not protect you from cable issues, patch mistakes, or hardware faults.

Fully redundant solutions

A continuously resilient, fully redundant solution is required in many System z environments. A fully redundant solution, as shown in Figure 2-28, includes the following components:

- Channel paths from the processor to the storage
- Switching hardware
- Cable infrastructure
- Networks

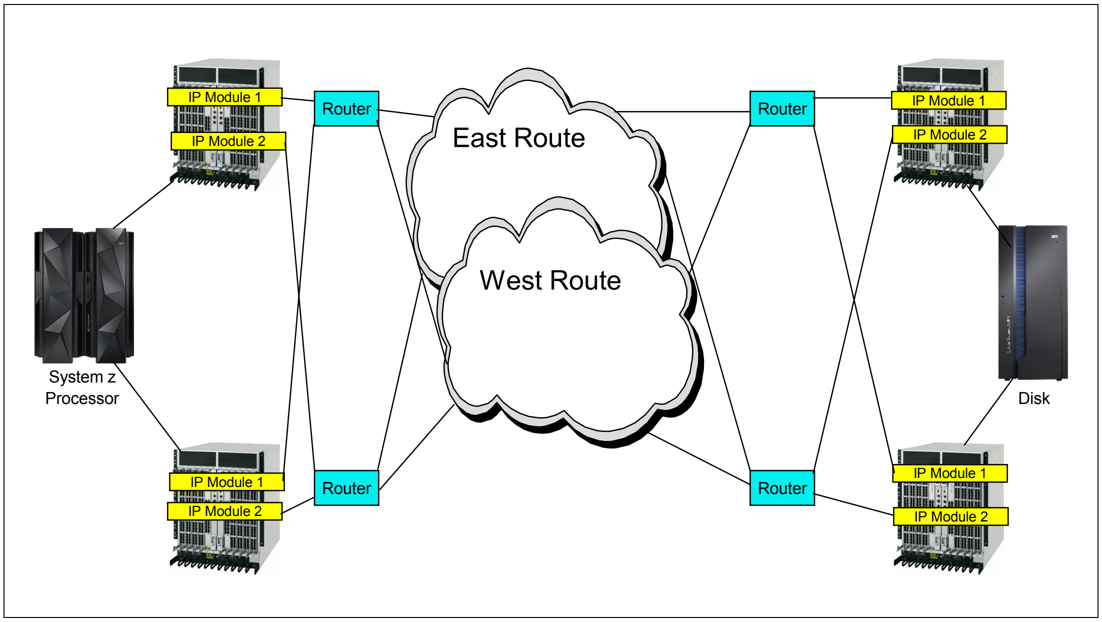


Figure 2-28 Fully redundant routed FCIP solution

Limited redundancy

Some applications, batch-tape backup for example, might not require continuous availability. In a component failure scenario, hardware costs can be reduced if failed components can be identified and replaced quickly enough to avoid missing availability targets. The network is the most likely component to fail, so it does not make sense to have redundancy elsewhere if you do not have redundant networks.

A typical configuration includes a redundant network with single points of failure in the components that are in the data center. In Figure 2-29, a failed optic, cable, channel port, switch port, or storage port will result in a complete path failure. However, either network could fail and data will continue to flow uninterrupted between the host and the storage device. In the case of a component failure, system operators can move workload to avoid the failed component, or dispatch service personnel to replace the failed components.

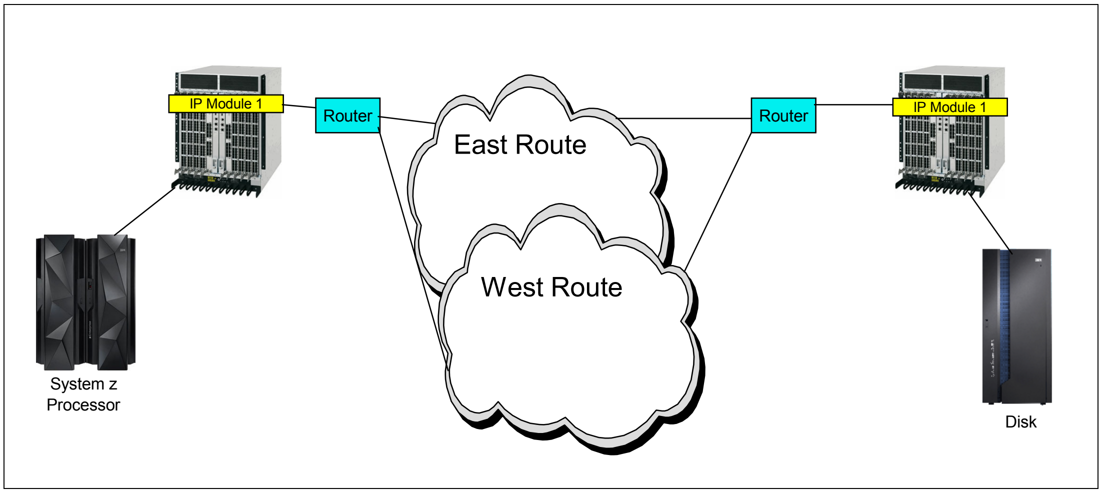


Figure 2-29 Network redundancy only

2.8.4 FCIP with DWDM

FCIP with DWDM is similar to FCIP over a routed network, except that frames in flight that are lost as a result of a failed component are not recovered in the network. Instead, dropped frames are recovered by the channel or host bus adapter (HBA). Upper protocols might get involved for severe failures, but usually frames are recovered using standard FCP. A FICON channel will report an out-of-order frame to z/OS, which will report it as an IFCC.

In a fully redundant high availability solution, as shown in Figure 2-30 on page 87, the error is usually recovered without further incident.

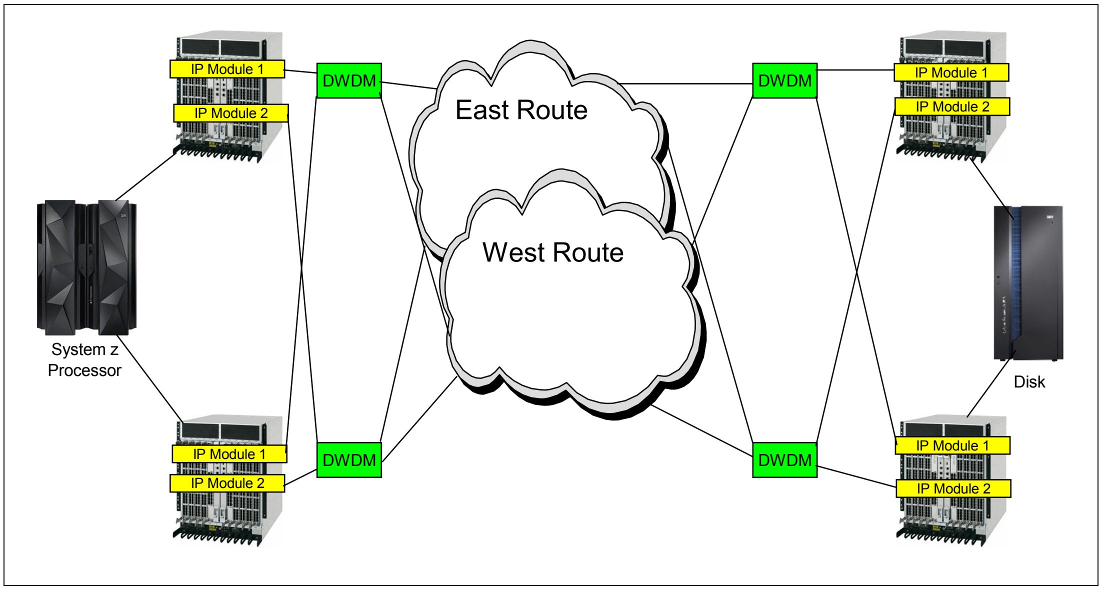


Figure 2-30 Fully redundant DWDM FCIP solution

Some people prefer this approach, because the SAN is aware of the problem and reports the condition through the SAN management software. Another reason is that SAN is typically managed by storage administrators who are immediately aware of network issues. Alternatively, when recovery is done in the network, the network teams must inform the storage team that the network was compromised.

2.9 Practical considerations

Information technology (IT) professionals will know that there are usually many ways to do a particular thing, but experience shows us that some are more effective than others. Similarly, just because something is supported by a vendor does not necessarily mean that it is a good idea to do that in your particular configuration. In this section we provide advice based on the experiences of the authors of this document.

2.9.1 Mixing FICON and FCP in the same fabric

The IBM qualification letters indicate whether a given switch configuration is qualified to have FICON and FCP in the same fabric. However, just because it might be qualified to have FCP and FICON in the same fabric, that does not necessarily mean that you should always configure to run in that way. The suggested preferred practice is to segregate FICON and FCP into their own logical fabrics. If the reason you are considering putting FICON and FCP in the same fabric is to share the ISL bandwidth, see 2.9.2, “Sharing ISLs” on page 89.

There are several considerations when mixing FICON and FCP in the same fabric:

- Placing FCP and FICON in the same fabric means that they must be set up consistently. For example, FCP prefers exchange-based routing, but at the time of writing, FICON does not support exchange-based routing.
- Switch firmware updates for FCP are delivered much more frequently than levels that are qualified for System z. If FICON and FCP share the same fabric, you might have to make a decision between not installing an update that you need for FCP, or installing the update and then running your System z traffic on a non-qualified firmware level.

- Distributed systems typically use zoning much more intensively than mainframe systems. This is largely because distributed systems use dynamic discovery (meaning that you have to use zoning to limit what each system can see), but mainframe systems only use devices that are explicitly defined in the IOCDS.
- Mainframe support groups tend to cable the infrastructure, including planning for growth, and then not touch it (on the basis that every change is a potential opportunity for a problem). Distributed systems, conversely, tend to be much more dynamic, with a far higher level of plugging and unplugging of cables.
- Is the FCP traffic associated with distributed systems, or Linux on System z, or disk mirroring? You might have preferences for the type of traffic that you would be willing to share the fabric with your FICON traffic.
- Who is responsible for managing the FCP and FICON infrastructures?

  The processes and management approaches used by the staff that support distributed systems are often different than those used by the mainframe group. Even in the mainframe group, there might be a different set of staff that manage the FICON infrastructure than manage the FCP infrastructure.

  Also, potentially, even a different set that manage the FCP infrastructure used for disk mirroring. If you intermix FICON and FCP in the same fabric, the support groups need to be cognizant both of their own processes and management, and those of the other groups sharing the fabric.
- The traditional model for controlling a System z I/O configuration is that each system will only touch devices that are defined in the I/O gen for that system. In the distributed world, the model is that every server will perform dynamic discovery to see what devices it has physical connectivity to.

  However, it goes beyond that. Some distributed systems will both look to see what devices they can connect to, and also write to all of those devices. If one of those systems were to write on a System z disk, that disk would then be unusable by System z. Segregating those systems from your System z fabric is one of the safest ways of ensuring that this does not happen.

For these reasons, it is preferable to separate FICON and FCP into different fabrics if possible. If this is not feasible, try to separate the mainframe infrastructure into a different fabric than the distributed systems. When FICON traffic and FCP traffic share the same physical switch, the traffic can be segregated by placing different traffic in separate logical fabrics, or by separating the traffic into different zones. The preferred practice is to place FICON and FCP in separate logical fabrics.

However, there is a drawback to configuring in this manner. Placing the FICON and FCP traffic in different logical fabrics will limit your ability to use RMF to analyze the performance of traffic on the ports being used by FCP. It will also not be possible to use other System z tools that use the CUP feature to obtain information about the FCP ports. For example, disk mirroring uses FCP protocol even though the data being replicated is from FICON channels. The traffic that is being reported on is z/OS data, even though the mirroring ports use FCP.

If you have an absolute requirement to use the CUP to gather information about the FCP ports, you have a few choices:

- Place the FCP traffic in a separate logical fabric from the FICON traffic, and define one FICON port in the FCP fabric to provide access to a CUP in that fabric.
- Place the FCP ports in the same logical fabric as the FICON ports, but keep the FCP and FICON ports in separate zones. In this case, ensure that you place the CUP port in the FICON zone.

2.9.2 Sharing ISLs

ISLs can be shared between different logical switches within the same fabric. This is done by creating a logical switch that is used just for transporting data between other logical switches in the same chassis. This logical switch is referred to as a *base switch*. Logical switches can be defined to use the base switch, or to have dedicated ISLs, but not both.

For the purposes of this material, an ISL between base switches is referred to as an *extended inter-link switch (XISL)*, and an ISL connected directly between two logical switches that cannot be shared with any other logical fabric is referred to as a *direct inter-link switch (DISL)*, as shown in Figure 2-31.

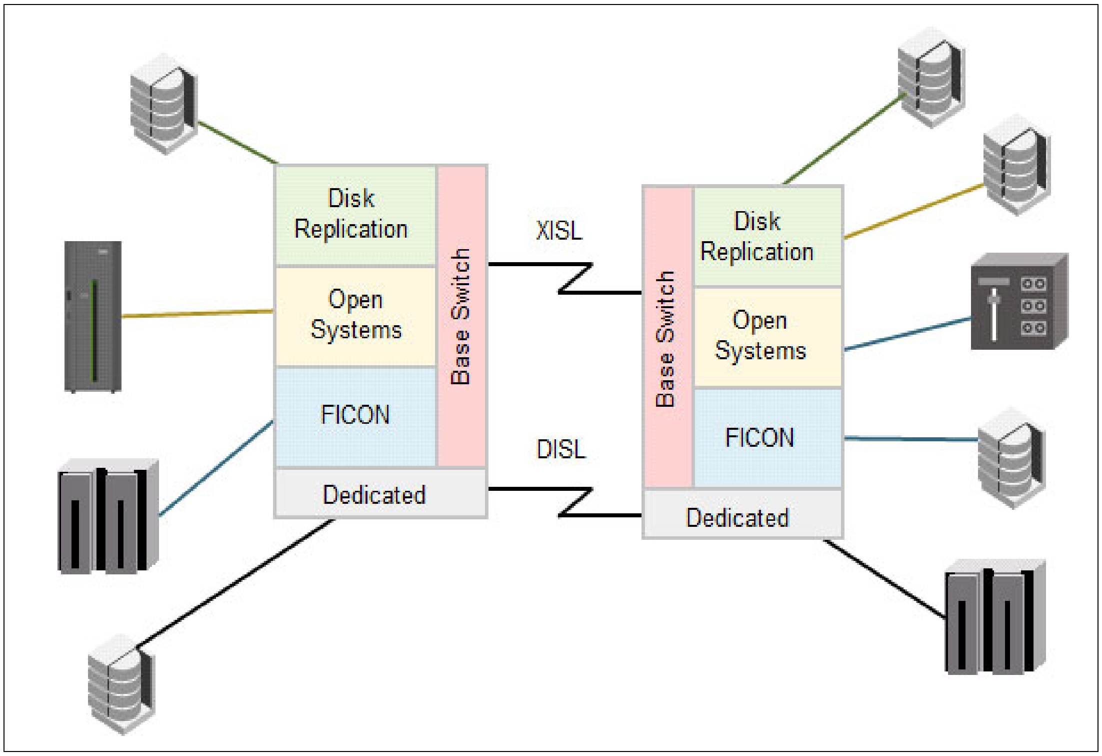


Figure 2-31 Sharing ISLs

Routing method

All logical switches defined to allow XISLs will adopt the same routing policy as the base switch. Device-Based Routing (DBR) is highly suggested. See 2.9.3, “Routing considerations” on page 90 for additional important routing considerations.

Control unit port

XISLs do not have a FICON address, and therefore cannot be managed through a CUP. Statistics for the XISL ports are not returned to RMF for the FICON Director Activity Report.

Supported ports in base switch

Only FC E_ports (ISLs and intercluster links, or ICLs) are supported in the base switch. All F-Ports (server and storage connections) must be in other logical switches. In the case of Brocade switches, IP ports (VE_ports) on the FX8-24, FX8-24E, and 7800 are not supported in the base switch; however, a physical GbE or extended GbE (XGbE) interface can be put in the default switch with circuits defined in different logical switches, with the minimum and maximums set to effect the same result.

Traffic Isolation zones

TI zones are supported, but only with failover enabled. TI Zones with failover disabled are not supported on an XISL.

Bandwidth sharing considerations

There are no restrictions to how many logical switches can share the XISLs. Traffic from FICON partitions and FCP partitions can share the same XISLs. Careful attention should be paid to the bandwidth requirements and device types that will be sharing the same bandwidth.

Slow-drain FCP devices using the same XISL that a FICON partition is using, as well as over-subscription, can cause performance problems and possibly FICON protocol timeouts. When using QoS zones to remedy this potential issue, System z connections should be in the highest QoS zone.

2.9.3 Routing considerations

Switch routing is a complex topic, and optimizing it requires the involvement of a switch specialist. This section just points out some recent enhancements in this area that should be considered when configuring the switches.

Routing considerations with XISLs

When using XISLs (an ISL in a base switch), the effective routing policy for all logical switches configured to allow XISL use is the routing policy of the base switch, regardless of how the routing policy is defined. The preferred practice is to explicitly set the routing policy in all logical switches configured to allow XISL use to have the same routing policy as the base switch - this avoids confusion about which routing policy is actually being used.

The suggested routing policy for the base switch whenever FICON traffic will use the base switch for transport is Device-Based Routing (DBR). DBR is supported for both FCP and FICON traffic.

Device-Based Routing

Previously, only Port-Based Routing (PBR) was supported for FICON. PBR establishes a route for every possible path through the fabric whenever a port logs in. With DBR, routes are only established as needed.

The routing hash algorithm attempts to balance routes across available ISL trunks based on the count of routes required. With PBR, all possible routes for each port pair are determined at login time. This results in an evenly distributed number of port paths (source ID (SID)/destination ID (DID) pairs) across the ISLs.

Because channel paths are not always cascaded, not all port paths require a route across ISLs. As a result, paths requiring ISL routes might end up more heavily loaded on some ISL trunks while other ISL trunks end up with little cascaded traffic. These routes remain static until there is a fabric change.

With DBR, routes are determined when a route is required. As with PBR, the route remains static after it is determined; however, because the routing algorithm is only applied to port paths that require a route, better ISL load balancing is achieved.

Example of how routes can end up unbalanced with PBR

In the (over-simplified) configuration shown in Figure 2-32, CHPID 10 logs in first, and has a path defined to device D01A. A route is established to D01A (red line). Because PBR establishes all routes at login time, a route is also established to device D03B (gray line).

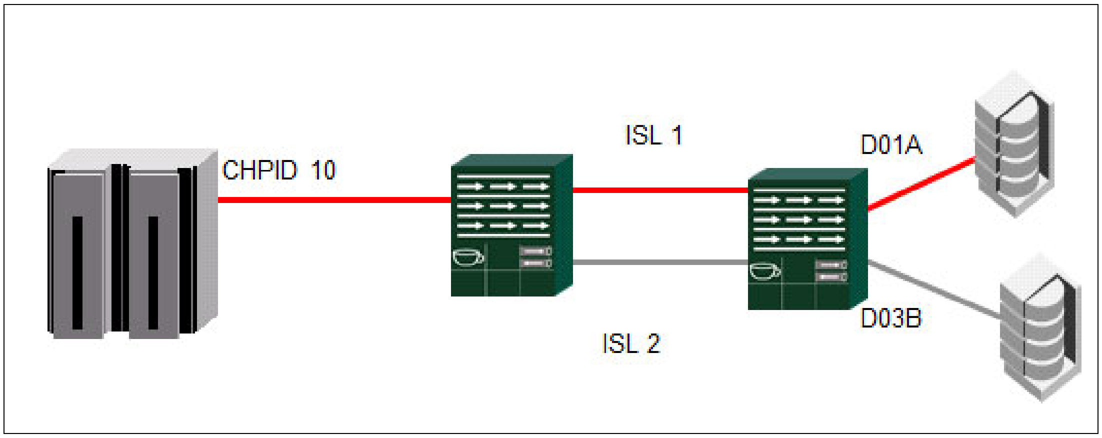


Figure 2-32 Paths from CHPID 10 to device D01A using PBR

Some time later, CHPID 20 logs in and has a path defined to device D03B (blue line) as shown in Figure 2-33. Other routes that are not needed have been established, such that when the route for D01A (grey line) is determined it chooses ISL 2 and the route for D03B is put on ISL 1.

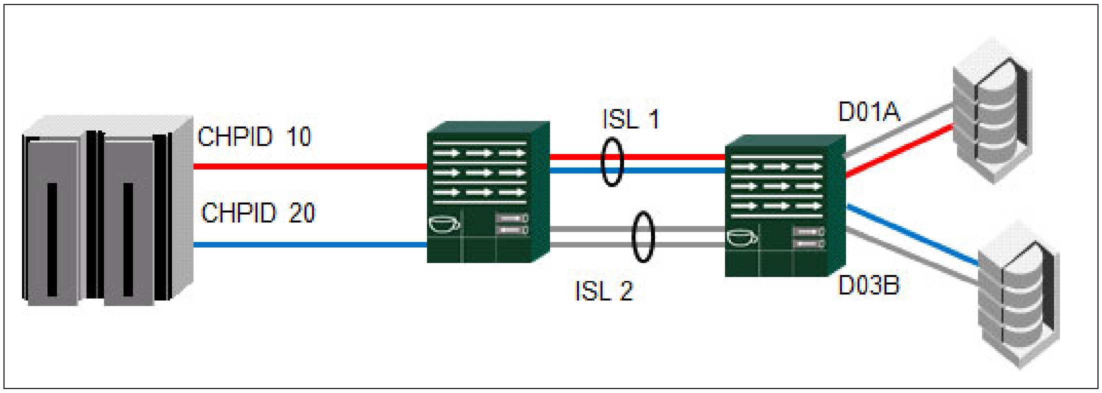


Figure 2-33 Paths from CHPID 20 to device D03B using PBR

Although routes are balanced across all available ISLs, both paths with data are routed to ISL 1 and the unused routes are assigned to ISL 2.

Example of how DBR rectifies this potential problem

Following the previous example but with DBR, Figure 2-34 shows CHPID 10 logging in first. However because DBR is being used, the route to device D01A is not established until that device is varied online (red line). No other cascaded routes are required, so no other paths are established.

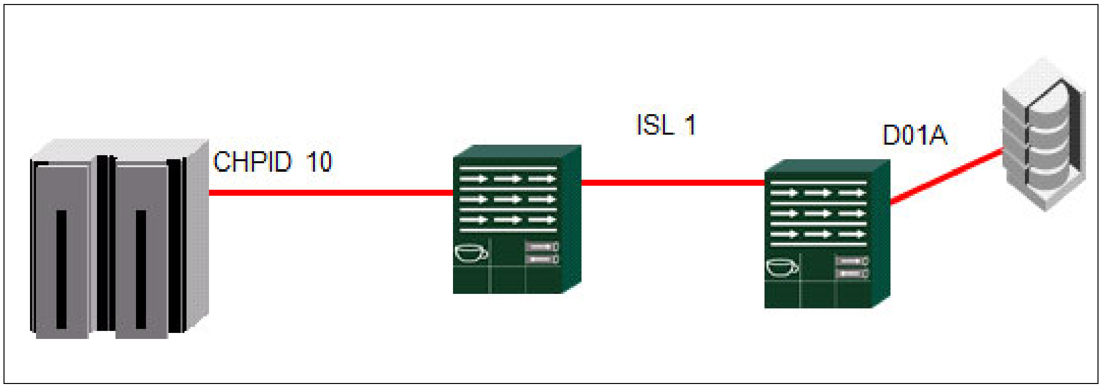


Figure 2-34 Path from CHPID 10 to device D01A using DBR

In Figure 2-35, CHPID 20 then logs in. But again, the route to D03B is not established until that device is varied online. When D03B is varied online, no other routes are established across the ISLs because routes are only established as needed, and no other paths required routes across the ISLs. In this case, because ISL_1 is being used by the path to device D01A, ISL_2 is used because it is the next ISL.

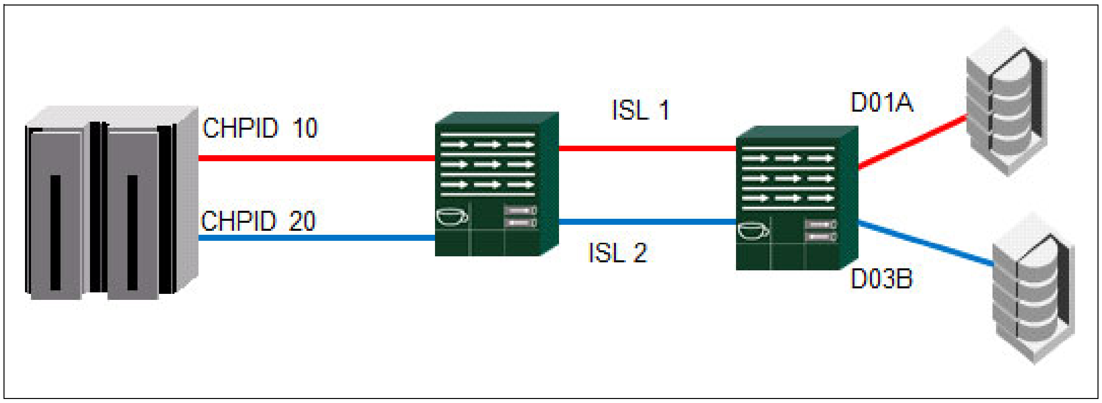


Figure 2-35 Path from CHPID 20 to device D03BA using DBR

Because routes are only established as needed, routes are balanced across the ISLs, and the cascaded paths are balanced across the ISLs as well.

2.9.4 Plan ahead

There might be activities over the life of a switch that require a planned outage of part of the switch, or potentially the entire physical switch. Such outages can be disruptive, both because of the bandwidth that is lost when a switch goes down, and because of the volume of manual intervention that might be required to take all of the paths offline in advance, and then online again after the change has been completed.

However, careful advance planning and ongoing management can help you minimize the number and effect of such outages. The specific activities that require outages will vary by vendor, by switch device, and potentially even depending on how the switch is configured. You should work closely with your vendor during the planning stages for the switch to ensure that you configure it to minimize potential planned outages in the future.

The following list describes some examples of the type of activity that might require an outage:

- Changing a switch from being configured as a single switch to being configured as one or more virtual switches.
- Upgrading to a new firmware level. For example, it might be possible to upgrade from one level to the next one non-disruptively, but if you want to upgrade across several levels, that might require an outage.
- Adding capacity to the switch.

The specifics will vary from one switch device to another. The important thing is to identify the changes that require a planned outage and, wherever possible, configure the switch in such a manner that those outages are not required.

2.9.5 Plan for firmware upgrades

Depending on your vendor, specific switch model, and how it is configured, there will be different considerations when the time comes to upgrade the switch firmware:

- Ensure that the target firmware level has been qualified with your mix of devices, connection types, and features.
- Work with your vendor to determine whether all switches in the fabric should be upgraded at the same time, or individually. The normal mainframe practice is to upgrade one device at a time, so that you can validate the new code, and minimize the number of things that are changed at the same time (to facilitate problem determination should any problems arise). However, there might be a requirement that both switches have to be on the same firmware level.
- Consider whether the upgrade requires a planned outage. If it does, perhaps it would be better to upgrade both switches at the same time, rather than having to work around two planned outages.

Your switch vendor will be the best one to help you identify the preferred practice for your specific environment and configuration.

2.9.6 Miscellaneous

To protect yourself from the risk of a non-System z server being accidentally connected to a System z device, you might consider disabling any ports that are not currently in use.

If your switch is shared between System z and distributed systems, it is likely that there will be a higher-than normal level of cabling changes within the switch, due to the dynamic nature of distributed systems. The small amount of additional time that will be required to enable the port when you want to use it will be far less than the time that would be required to restore a device that is accidentally damaged from some other system.

To ensure that you have optimum availability, every device should be connected to the System z CEC by at least two CHPIDs, and those CHPIDs should be connected to different blades and different switches.

To minimize the chance of bottlenecks within the configuration, the speed of the links between the CECs and the switches should be at least as high as the links between the switches and the connected CUs.

For example, having 8 Gb links from the CECs to the switches, and 4 Gb links from the switches to the attached CUs would be considered to be a balanced configuration. However, having 8 Gb links between the CUs and the switch, and 4 Gb between the switch and the CEC would not be considered balanced, and would be more likely to encounter bottlenecks.

When looking at the volume of data that will be going to the devices, and the speeds of the various links, bear in mind features of the devices that can affect the volume of data. For example, some tape units support compression, meaning that the write rate of data onto the tape media would be less than the rate at which the (decompressed) data is being sent to the tape CU.

It is generally considered to be a preferred practice to separate different types of traffic onto different ISLs. For example, tape traffic should be kept separate from FICON disk traffic. And FICON disk traffic should be kept separate from FCP disk mirroring traffic. These different types of traffic display different types of behavior (frame sizes, burstiness of the I/Os, and so on), and therefore should ideally be kept separate from each other.

Generally speaking, distance has a larger effect on the performance of applications that use small frames than it does on applications using large frames. So, for example, tape processing, which tends to use very large blocksizes, is less affected by distance than applications that rely on CTC adapters, which tend to have very small frames.

Therefore, adding a given distance might have a different effect on different applications, depending on their I/O profile. It also means that you need to be more cognizant of having sufficient buffer credits for applications with small frame sizes than you would be for one with large frame sizes.

2.10 Further information

For more detailed information about SANs and how to use SANs in a System z environment, see the following IBM Redbooks documents:

- *FICON Planning and Implementation Guide*, SG24-6497
- *IBM SAN Survival Guide*, SG24-6143

3

Wavelength division multiplexing

This chapter describes the situations where you might have a good reason to use wavelength division multiplexing (WDM). It also provides a description of WDM technology, and defines the terms that you will encounter when involved in the implementation of WDM in an end-to-end connectivity solution.

IBM does not offer a WDM device. However, IBM works closely with multiple vendors to test, qualify, and provide extended distance solutions. This chapter attempts to provide descriptions that are generic across all vendors. Although the terminology might vary between vendors, most of the concepts should apply to all vendors.

You should always check with your vendor to ensure that the features and capabilities you require are available and qualified on the WDMs that you are planning to use. The IBM WDM qualification letters are available on the IBM Resource Link website:

https://www-304.ibm.com/servers/resourcelink/lib03020.nsf/pages/systemzQualifiedWdmProductsForGdpsSolutions?

3.1 WDM description and functionality

WDM is a mature technology that enables multiple independent signals to be transported over a single fiber or a pair of optical fibers[1]. WDM mixes different wavelengths, or lambdas (λ), of light. The different wavelengths are typically represented as different colors when WDMs are shown in a diagram. The different wavelengths remain independent of each other while traveling within the optical fiber. As a result, each lambda can be used to transport its own signal, independently from the others. Figure 3-1 shows the basic concept behind WDM.

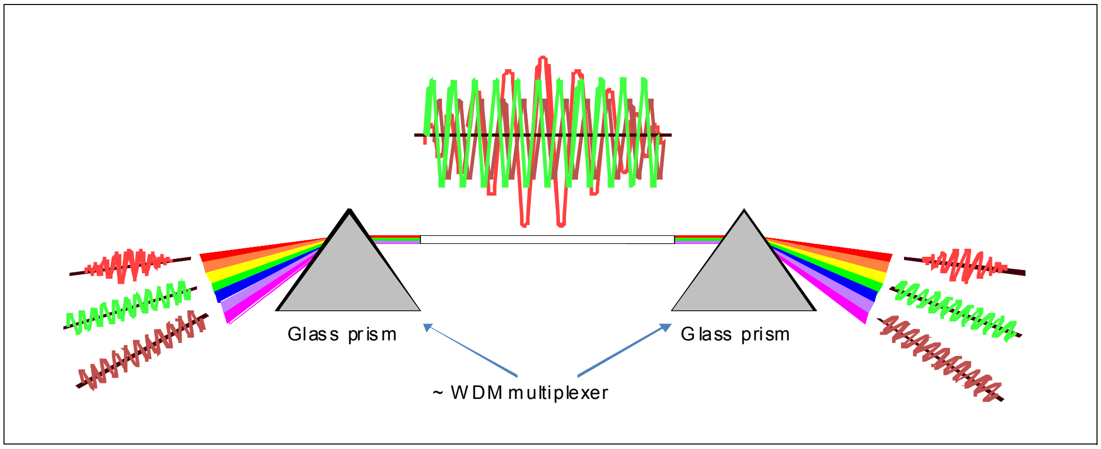


Figure 3-1 Fundamental concepts of WDM technology

With the availability of improved optical amplifiers, WDM vendors have the ability to amplify all WDM lambdas as one combined signal within the fiber. This capability provided a commercial breakthrough for WDM-based transport systems.

3.2 Benefits of WDM technology

When planning your end-to-end connectivity solution, one of the decisions you need to make is whether WDMs should be a part of that solution. The following list includes some of the most important questions to help you determine the requirements:

- Are my fiber connections greater than the maximum supported unrepeated distance for the optical links that I need to connect?

  For example, if your fibers are longer than 10 km, they are beyond the maximum unrepeated distance of long wavelength (LX) FICON and Fibre Channel Protocol (FCP), Open Systems Adapter (OSA), and 1X Parallel Sysplex InfiniBand (PSIFB).

- Do I have a constraint (technical or financial) on the number of fiber pairs between my sites?
- Do I have a variety of connectivity types (coupling links, OSA, FICON and FCP channels, inter-switch links (ISLs), and so on) that I need to connect across my sites?
- Do I have a cross-site connectivity requirement for devices that are not supported by a storage area network (SAN) switch or a network switch?
- Is it possible to get dedicated fiber connections between my sites? IBM or your communications service provider (telco) should be able to help you get an answer to this question.

[1] Most WDM vendors currently support only dual fiber (a single pair of fibers for the WDM link).

If your answer to any of these questions is *yes*, there might be a compelling reason to use WDM technology as a key component in your end-to-end connectivity architecture.

The following list describes other reasons why you might benefit from WDM technology:

- Dedicated fiber is usually charged on a per-fiber basis. A WDM might help you reduce costs by enabling you to obtain the bandwidth you require with fewer fibers.
- A WDM could help you improve availability by providing the ability to connect to fibers that are connected over failure-isolated routes, and automatically switch from one route to the other in the event of a failure or planned outage on one of the routes.
- A WDM can provide a defined termination and handover point between the SAN/local area network (LAN) and the wide area network (WAN).
- A WDM provides a centralized management point to control all of your cross-site connections.
- A WDM can provide continuous fiber inspection for loss of light and changes in the length of the route between the WDMs.
- A WDM provides alarm correlation through advanced management systems, enabling better troubleshooting capabilities.
- A WDM might provide an encryption capability, meaning that encryption for *all* inter-site traffic could be handled in one place.
- If you purchase a cross-site connectivity solution from a telecommunications vendor, it is likely that your solution will include WDMs. However, even though the WDMs will be owned by that vendor, it is still critical that you ensure that the proposed models will meet your System z requirements, and that the DWDM has been qualified by IBM. The latest System z DWDM qualification letters can be found on the IBM Resource Link website.

These and other features might be offered by current WDM offerings. The specific features that are available will vary by WDM vendor and your implementation needs.

The rest of this chapter provides further details to help you understand WDM technology and terminology in a little more detail.

3.3 Terminology used with WDMs

The following list provides a description of some common WDM terminology that is subsequently used in this chapter:

Link capacities Information technology (IT) staff are used to talking about data rates, which are generally referred to in bytes per second (Bps). Network staff generally talk about capacity in terms of bits per second (bps).

Electrical circuits handle information using 8 bits for each byte of information. Commonly available capacities are T1, which provides 1.544 megabits per second (Mbps), and T3, which provides 45 Mbps.

Optical circuits handle the same information using 10 bits for each byte of information. The common designations for optical circuits are OC1 (50 Mbps), OC3 (155 Mbps), OC48 (2.4 gigabits per second, or Gbps), OC12 (9.6 Gbps), and OC768 (40 Gbps).

So, for example, a 1G Ethernet electrical connection would need a 10 Gbps optical connection. And a 10 GBps circuit would require a 100 Gbps optical connection.

Lambda A lambda (λ) is a WDM wavelength (physicists use the greek symbol lambda for a wavelength). In a WDM, a lambda is used to transport data. One lambda could be used to transport a single 10 Mb signal, or could be used to transport a 100 Gb data stream, depending on the technology used by the WDM.

Transponder A transponder is a module within the WDM. It has two types of ports:

A *client port* that is connected to a device (a SAN switch, for example)

A *network port* that passes a signal to another component within the WDM.

Generally speaking, a transponder will have an equal number of client and network ports.

The transponder converts the signal coming from the connected device into a signal using a WDM lambda that is launched at the transponder's network port. It is common for a WDM to have more than one transponder.

See Figure 3-3 on page 100 for a diagram showing the role of the transponder in the WDM.

Muxponder A WDM muxponder is a module within the WDM that combines the following elements:

Multiple client ports.

An electrical multiplexer that combines the incoming client signals. The electrical multiplexing is done using a technology called time-division multiplexing (TDM).

A transponder to convert the output from the electrical multiplexer into a lambda.

A muxponder has more client ports than network ports. A WDM can contain both transponders and muxponders. But a given client device would only be connected to one or the other.

See Figure 3-3 on page 100 for a diagram showing the role of the muxponder in the WDM.

Multiplexer (Or Demultiplexers) are used to combine different WDM lambdas onto a single fiber. For example, up to 160 lambdas could be multiplexed into a single optical signal before traveling down the optical fiber. At the receiving end, the demultiplexer separates out the combined signal into single lambdas, each on a separate fiber.

CWDM Coarse wavelength division multiplexing. Usually 8 to 16 lambdas can be used, with up to 10 Gbps per lambda. The maximum distance is about 60 kilometers (km), because fiber amplifiers cannot be used. CWDM is for smaller implementations given these limitations. None of the IBM-qualified WDM solutions use CWDM technology.

DWDM Dense wavelength division multiplexing. This technology supports in the region of 40 to 160 lambdas, with up to 100 Gbps per lambda. DWDM is the underlying technology for nationwide backbones and transatlantic data transmission, to interconnection between remote data centers. Additionally, all IBM-qualified WDM solutions are based on DWDM technology.

SONET and SDH Synchronous Optical Networking (SONET) and Synchronous Digital Hierarchy (SDH) are standardized multiplexing protocols that transfer multiple digital bit streams over optical fiber using lasers or highly coherent light from light-emitting diodes (LEDs). At low transmission rates data can also be transferred across an electrical interface.

The method was developed to replace the Plesiochronous Digital Hierarchy (PDH) system for transporting large amounts of telephone calls and data traffic over the same fiber without synchronization problems.

System z implementations require dedicated fiber for synchronous cross-site communications. For asynchronous applications (such as IBM z/OS Global Mirror (zGM) or Global Mirror), the use of SONET or SDH might be acceptable.

However, the response time and bandwidth that are delivered by such a configuration are not directly under your control, and are not as predictable as when you use dedicated fiber. Therefore, the use of any topology other than point-to-point dedicated fiber is *not* qualified by IBM.

OTN The Optical Transport Network (OTN) was created with the intention of combining the benefits of SONET/SDH technology with the bandwidth expansion capabilities offered by WDM technology. OTN also enhances the support, administration, maintenance, and provisioning of SONET/SDH in WDM networks.

3.4 Types of WDM

There are two types of WDM. The most commonly used is DWDM. The other is CWDM.

The difference from a commercial perspective is that a CWDM is slightly cheaper, but offers a limited number of lambdas. Also, it is not possible to amplify all CWDM lambdas using an optical amplifier, therefore it has limited distance capabilities.

At the time of writing, DWDMs typically support up to 160 lambdas, and distances up to several thousand kilometers[2].

[2] The maximum distance qualified for use with System z is 300 km.

Figure 3-2 shows the difference between CWDM lambdas, which have a wide 20 nanometer (nm) spacing between lambdas, and DWDM lambdas, which have a narrow 0.4 nm spacing between the wavelengths. At the time of writing, all WDMs qualified by IBM for use with System z are DWDMs.

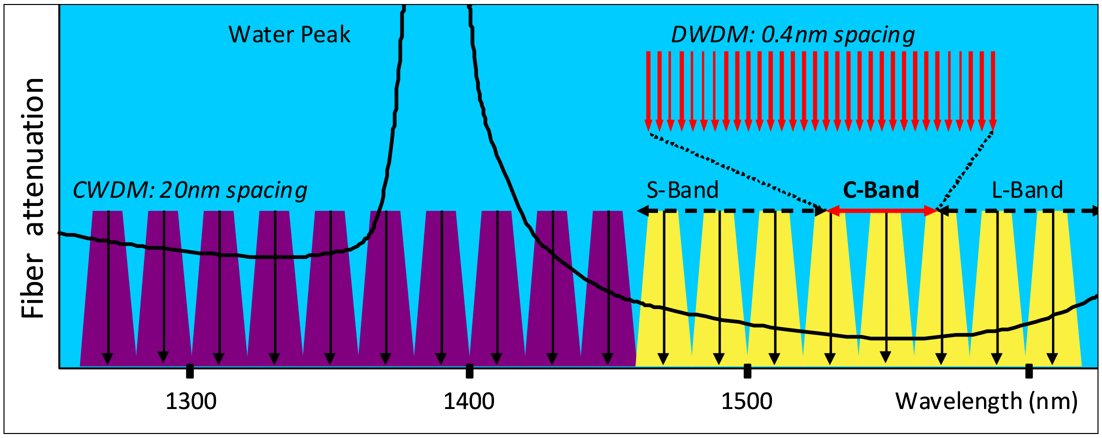


Figure 3-2 CWDM and DWDM wavelength spacing

3.5 WDM systems

A WDM system will usually have one or more chassis containing all of the modules. Typically, WDM systems consist of the following components:

- Transponder or muxponder modules
- Optical multiplexer/demultiplexer units
- Amplifiers, dispersion compensating modules, fiber switches, and other needed functions
- A control unit for external communication with the WDM system to enable management of the WDM and the WDM network

Figure 3-3 illustrates how these different modules are used in the overall WDM solution.

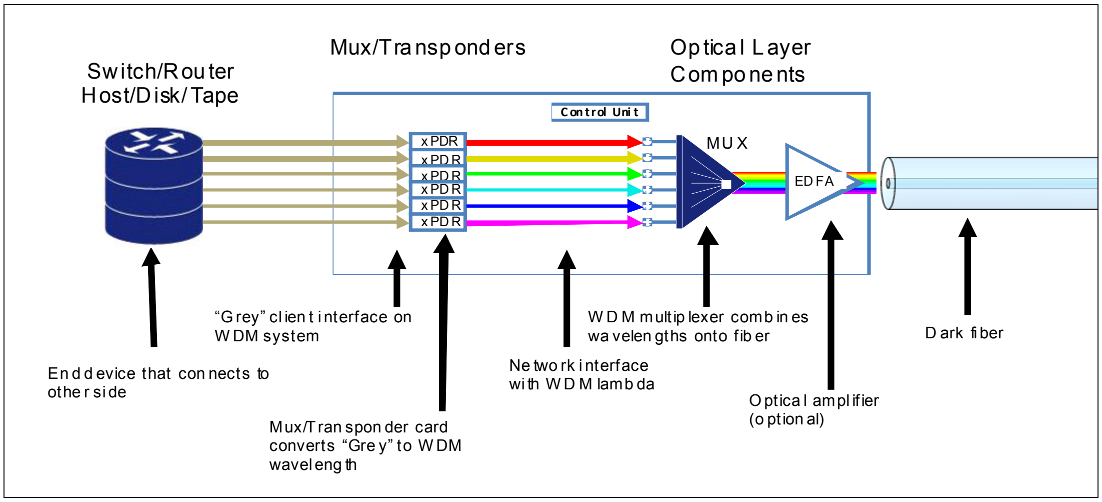

Switch/Router
Host/Disk/Tape
Mux/Transponders
Optical Layer
Components
Control Unit
xPDR
xPDR
xPDR
xPDR
xPDR
xPDR
MUX
EDFA
"Grey" client interface on WDM system
End device that connects to other side
Mux/Transponder card converts "Grey" to WDM wavelength
Network interface with WDM lambda
WDM multiplexer combines wavelengths onto fiber
Optical amplifier (optional)
Dark fiber

Figure 3-3 The building blocks of a WDM system

The xPDR in Figure 3-3 represents either a muxponder or a transponder.

3.5.1 Dark fiber

The optical fiber used in a WDM network is single mode (SM). The following list includes some of the associated characteristics of an SM fiber:

- The speed of light in a fiber is reduced to about 200,000 km/sec, so it takes about 5 microseconds (µs) to travel 1 km within a fiber.
- The attenuation of a fiber is a function of the wavelength that you transmit. This is shown in Figure 3-4. The preferred wavelengths for WDMs are the areas (known as *optical windows*) where the fiber attenuation is low:
  - The first optical window is used to transport short wavelength (SX) signals at 850 nm.
  - The second optical window at 1310 nm is used by LX signals.
  - The third optical window is between 1500 nm and 1610 nm.
- From the three available optical windows, the second and third are used by CWDMs.
- Only the 3rd optical window is used by DWDM.

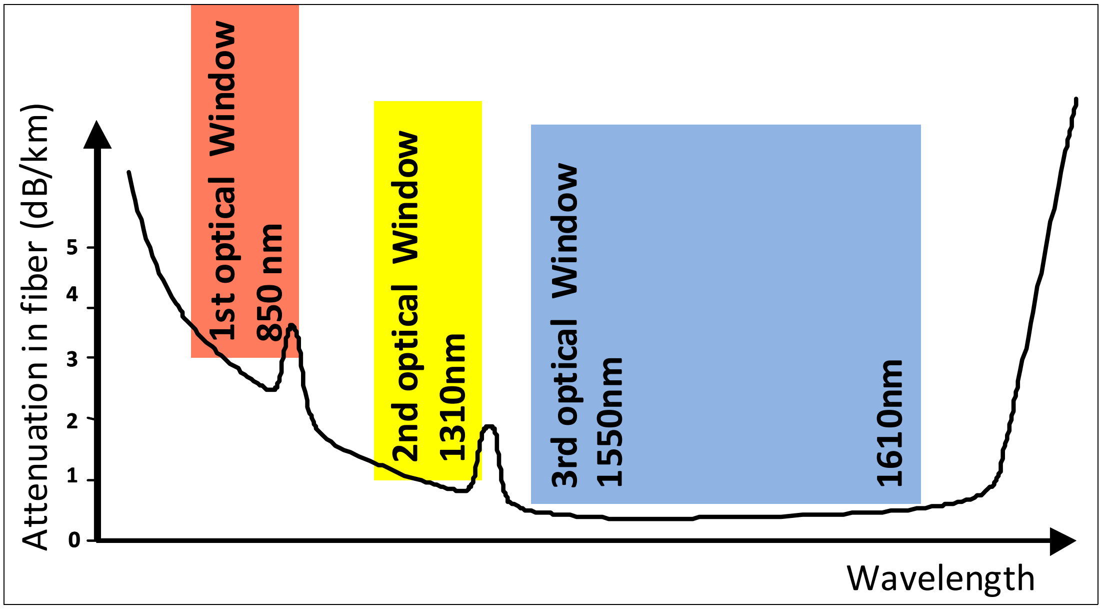


Figure 3-4 Attenuation on a fiber and the optical windows used

3.5.2 The control unit

The control unit (CU) of a WDM device provides full provisioning operation and maintenance functions and services. An Ethernet port and, typically, a serial port are used for external connectivity. Usually, the controller provides a web-based or Java-based graphical user interface (GUI), the possibility for Telnet or Secure Shell (SSH) access, and services for Simple Network Management Protocol (SNMP) access.

3.5.3 The optical multiplexer/demultiplexer

The optical multiplexer is a passive optical module to multiplex (combine) different lambdas into a sum signal. The port count for such multiplexers or demultiplexers are typically between 4 and 40 lambdas.

A demultiplexer is the counterpart of the optical multiplexer at the receiving end of the link. It separates all the different lambdas from the sum signal.

The building blocks for a fixed multiplexer/demultiplexer unit are shown in Figure 3-5.

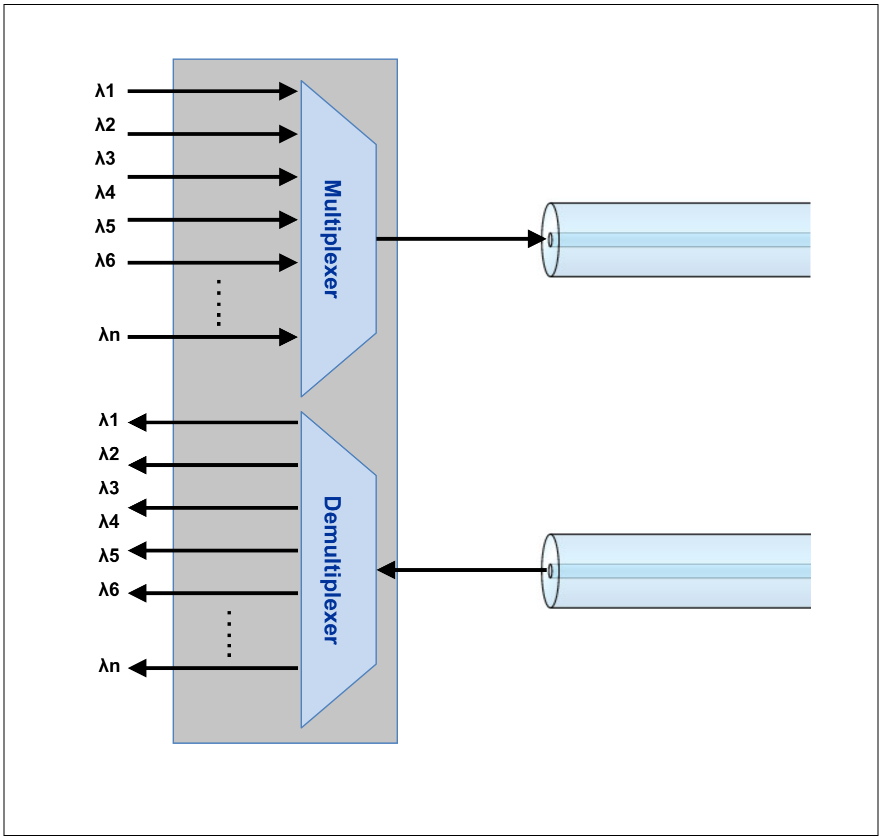


Figure 3-5 Sample of a multiplexer/demultiplexer module

3.5.4 The transponder and muxponder modules

Transponders and muxponders are the active parts of a WDM system.

A transponder is a module with one or more sets of ports. Each set contains a client port (connected to a device, or an ISL, or an InfiniBand (IFB) link, for example), and a network port (that is connected to the WDM multiplexer). The transponder converts the signal coming from a device attached to its client port into a signal using a WDM lambda that is transmitted from the network port, as shown in Figure 3-6. If there is more than one client port, there will be an equal number of WDM network ports on that module.

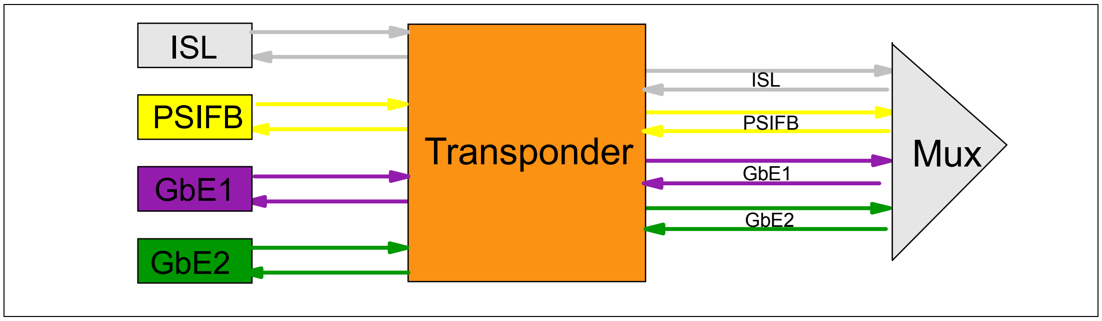


Figure 3-6 WDM transponder

A muxponder is a combination of an electrical multiplexer and a WDM transponder. Therefore a muxponder has two or more client ports, but only one WDM-based network port. The electrical multiplexing is done using TDM. Within a muxponder, several lower data-rate signals are multiplexed into one serial bitstream, which has to be at a higher data rate that is equal to the sum of all of the multiplexed streams. The elementary concept of TDM is shown in Figure 3-7.

The signals and protocols that feed into the client ports of those modules could be all kinds of signals, such as Ethernet, Fibre Channel (FC), gigabit Ethernet (GbE), IFB, or others. Typically, the data rates range from 8 Mbps up to 16 Gbps. For different data rates and different signal types, you might need different transponder or muxponder modules.

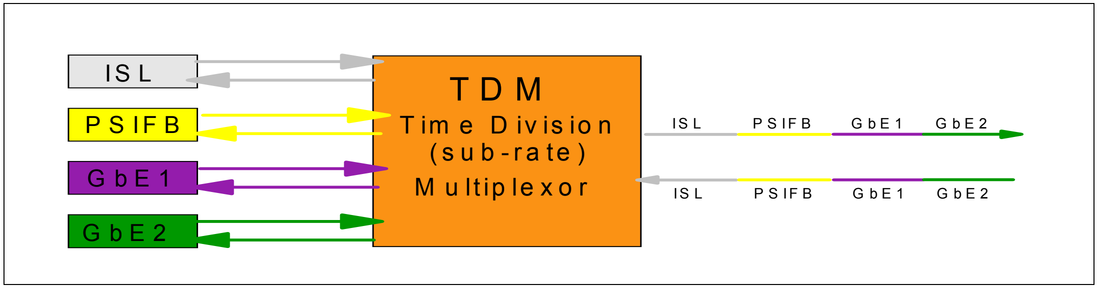


Figure 3-7 Basic TDM process

Mux/transponders for data center applications

There are two distinct markets for DWDM technology: data center extension and telco providers. Transponders and muxponders for data centers are usually much simpler than their counterparts for the telco/Internet service provider (ISP) market. For data center modules, the WDM manufacturers have the freedom to use nonstandard transport and TDM techniques. This is possible because most of the networks for data center applications are private networks using a dedicated fiber infrastructure.

Transponders

Transponders for data center applications are the fastest WDM modules. They act like a media converter, converting the incoming signal from the client device to a WDM signal with exactly the same data rate. A diagram of a dual-lane transponder is shown on the left in Figure 3-8 on page 104.

Muxponders

Muxponders for data center applications are usually based on standard components, because it is complicated to design bespoke, reliable, TDM technology. However, some manufacturers do offer this kind of device. The benefit of having a proprietary TDM scheme is that it could be developed and optimized for special latency requirements, and for handling nonstandard signals.

An outline of such a muxponder is shown on the right in Figure 3-8.

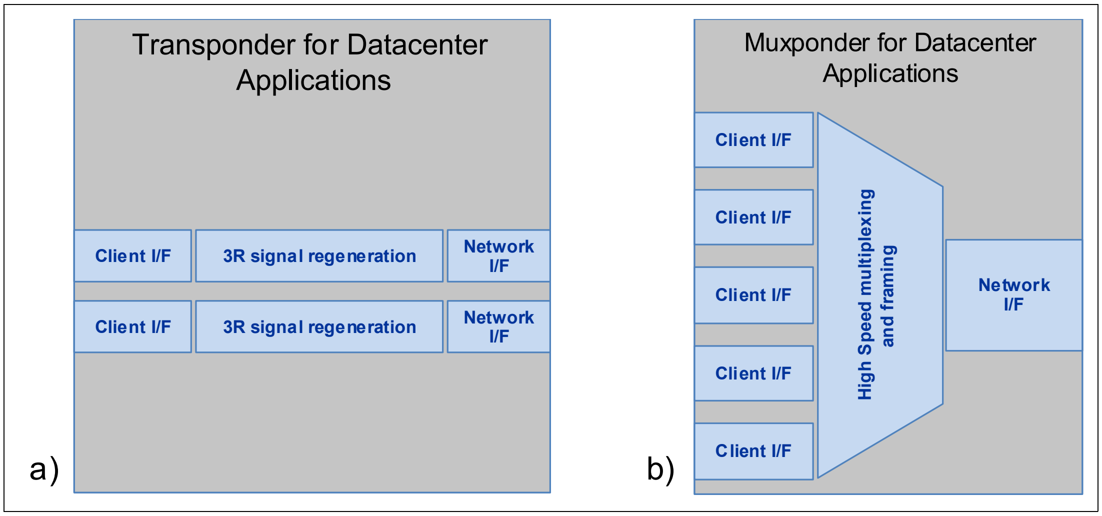


Figure 3-8   Transponder and Muxponder for data center applications

Transponders and muxponders for data center applications are specifically designed to connect data centers over metro distances (<100 km) using private networks. They usually offer smaller and more stable latency figures as a result.

The latency of a modern data center transponder could be as low as 5 ns per module, which roughly equates to the latency of one meter (m) of fiber. For muxponders, the latency figures might vary for different client signals. Additionally, data center transponders and muxponders offer support for nonstandard signals and features.

Mux/transponders for telco/ISP applications

Important: For WDM latency figures, and protocol or feature support for devices that you want to connect, check with your respective vendors. Vendor-specific features might not be supported by all WDM devices or modules.

It is important to be aware that connecting data centers is only one use of WDM technology. In fact, the majority of WDM devices are actually used for reasons other than connecting data centers. Therefore, in this section we briefly describe some of the terminology that you are likely to encounter when dealing with WDMs.

Because many WDM manufacturers serve multiple markets, most of them develop WDM transport modules for multiple purposes. Those WDM devices could be used to build country- or continent-spanning backbone infrastructures. Therefore support for technologies from SONET/SDH and other technologies, such as Optical Transport Hierarchy (OTH), is mandatory for those WDM modules.

The International Telecommunication Union (ITU) and especially the ITU Telecommunications Standardization Sector (ITU-T) defined an optical transport layer for WDM transport systems. This standard is OTN (also known in the industry as G.709, which indicates the ITU-T internal naming).

Transponders

Telco/ISP transponders take the client signal and map it directly into a G.709 container, or by using SONET/SDH encapsulation before encapsulating it into OTN structures.

A simplified telco/ISP transponder is shown on the left in Figure 3-9.

Muxponders

Muxponders need to time-division multiplex the signals. For incoming FC or Ethernet signals, a standardized TDM scheme is used. This scheme is known as the generic framing procedure (GFP). The output from GFP multiplexing is SONET/SDH containers. They could be further encapsulated into standardized OTN containers. There are also techniques for a direct encapsulation of non-SONET/SDH signals directly into G.709. Finally, GFP could feed directly into G.709.

The simplified building blocks for a telco/ISP muxponder are shown on the right of Figure 3-9.

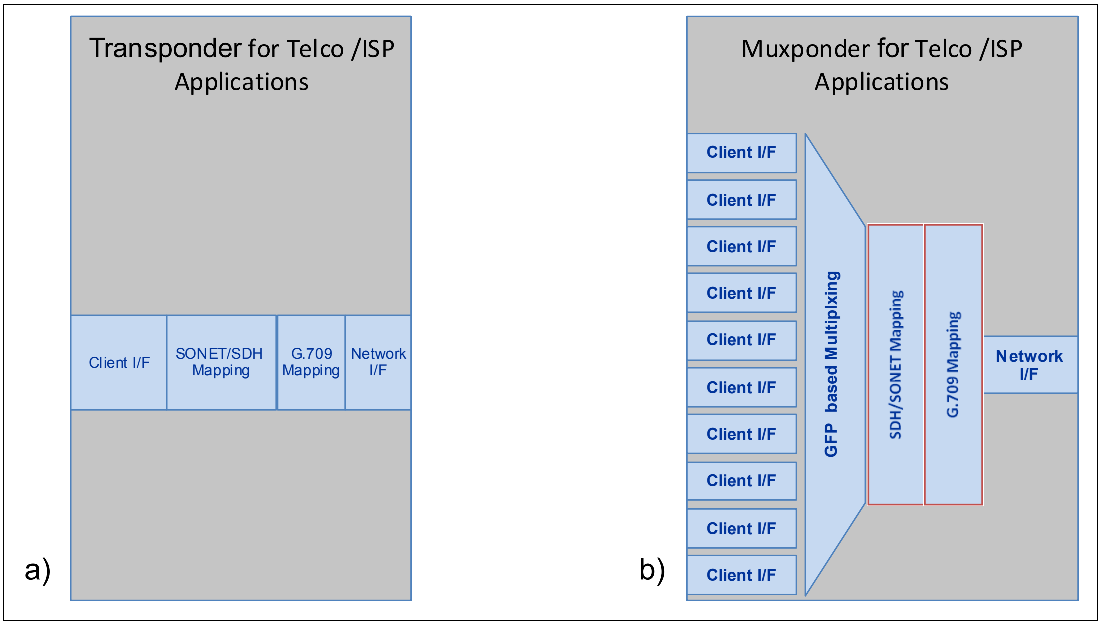


Figure 3-9 Transponder and muxponder for telco/ISP applications

These types of modules offer rich functionality, and can support distances of up to several thousand km using optical amplification. They are used to build large infrastructure networks, and could also be used to feed into third-party equipment. This is possible because SONET/SDH and OTN signals are highly standardized. For example, a muxponder with a SONET network interface from Company A could be connected to an existing SONET network that is based on devices from Company B.

The drawback of this kind of module is that all of those standardized procedures were not generally designed to meet low and stable latency criteria. Also, signals that do not conform entirely to the standard might lead to unexpected complications.

Therefore, if you have stringent latency requirements (for example, coupling facility (CF) response times in the µs range), or you must support special features and protocols (such as System z and FICON), it is better if you use modules that are designed specifically for data center applications. Obviously, all qualified DWDMs use components that are aimed specifically at the data center market.

3.6 Signal quality degradation

When a signal travels over a fiber, some physical effects can affect the quality of the signal. The effect is absorption, which weakens the optical power of the signal. This effect can be compensated for by using optical amplifiers. Another effect, called *dispersion*, also affects the signal quality. There can also be other effects related to the fiber quality and installation. However, the WDM devices are designed to handle and mask those effects.

3.6.1 Optical amplifiers

In a lab environment using WDM modules and simple optical filter structures, a WDM can drive a signal up to 100 km without additional amplification. However, when using real fiber infrastructures and optical WDM filter structures, the result is often much higher attenuation or smaller distances. As a result, the use of optical amplifiers is common for distances greater than 50 km. Preamplifiers are used to amplify weak signals over long-distance fiber. For distances beyond 100 km, additional special-purpose optical amplifiers might be necessary.

3.6.2 Dispersion compensation units

Depending on the distance between the WDMs, dispersion can cause signal distortion. Dispersion compensation (using dispersion compensation units, or DCUs) is usually required for links greater than 100 km. The amount of dispersion-based distortion that a WDM system can handle will vary depending on the signal data rate and WDM supplier.

In general, the support for handling distortion is built into the WDM, and you do not need to worry about it. However there is one aspect of dispersion that should be noted in the context of this publication.

There are three different types of dispersion-compensating modules available:

- Fiber-based dispersion compensating modules:
  - These are based on a special spool of dispersion compensating fiber.
  - The use of these spools can add significantly to the latency of your transport system (50 µs and more).
- Fiber Bragg Grating (FBG)-based dispersion compensating modules:
  - These are special optical devices to negate the dispersion effect.
  - The latency of these devices is usually less than 50 ns.
- Electronic dispersion compensation modules:
  - These can be a separate unit, or in some cases they are built into a transponder or muxponder.

If you require very low latency for your connectivity, and you have distances where DCU is necessary, you should check with your supplier to determine which types of DCU they support. Not all DCUs are qualified for use with System z, so check with the vendor to ensure that the DCU they propose using *is* qualified.

3.7 WDM topologies and protection schemes

Most WDM transport platforms offer a variety of topologies and protection schemes. The options and capabilities will vary from one device to another, and from vendor to vendor. Consult your WDM supplier for information regarding the available and qualified options for the WDMs that you are considering using.

3.7.1 Topologies

The most common WDM topologies are:

- Point-to-point topology
- Ring topology
- Meshed topology

All of these topologies can be designed with fixed-WDM filter structures, or with reconfigurable filters. Ring and meshed topologies (shown in Figure 3-10) are not qualified for System z connectivity solutions.

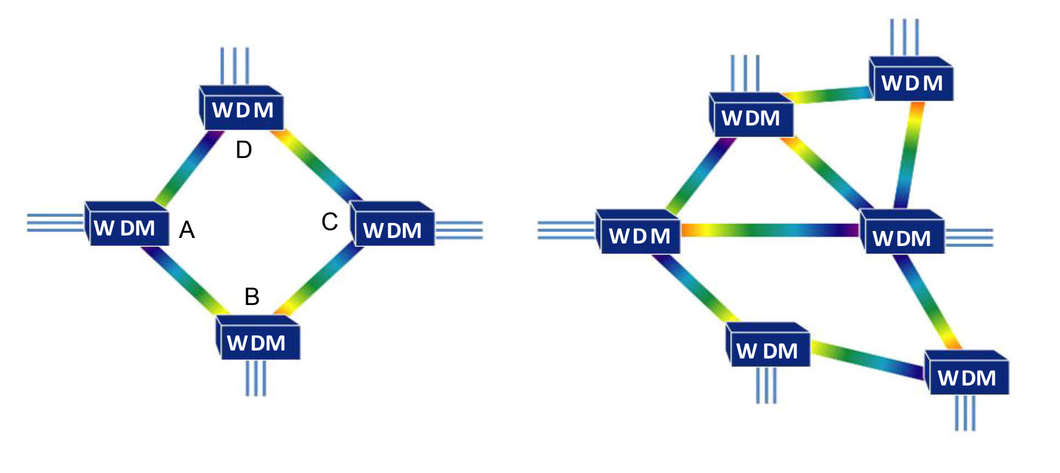


Figure 3-10 Ring and meshed WDM topologies

System z multisite connectivity solutions use a point-to-point topology, as shown in Figure 3-11, so that is the type that we focus on in this document. This shows a configuration where a pair of devices (a switch and a System z central electronics complex (CEC), for example) has three connections to each of the two WDM boxes in its respective site. This configuration will act as the base for the information about network availability.

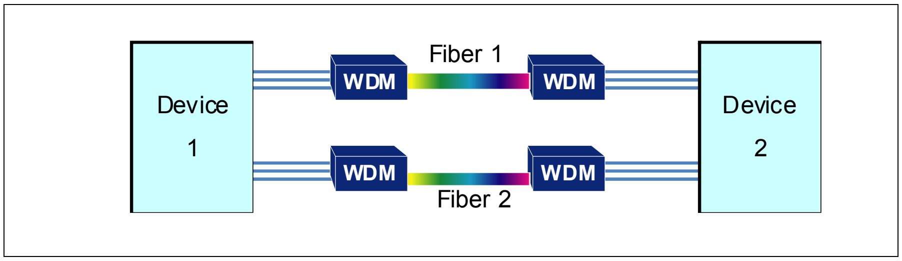


Figure 3-11 Typical dual Point-to-Point-based network layout for data centers

3.7.2 Protection schemes

In WDM terminology, the ability to continue operation across a link failure is known as *protection*. There are typically two protection approaches:

- If a link is lost, the error is presented back to the devices that are using the WDM, so the attached client device can recover from the failure by trying the operation again, or by routing the traffic to a different path. This is known as *client-based protection*, because the recovery is handled by the devices that are using the WDM, and behaves as though the WDM did not exist.
- The WDM equipment is responsible for failing over to a backup link and partially masking the failure from the connected devices. This is known as *WDM-based protection*.

Additionally, those two protection schemes can be combined.

Client-based protection

A client-based protection configuration using dual point-to-point systems is shown in Figure 3-11 on page 107. In this scenario, the client devices determine which path to choose. This assumes that the configuration is designed to provide multiple paths between every pair of devices, and that the devices include logic to recognize and recover from link failures. For most System z links, this is the mandatory protection mode for high-availability solutions.

The preferred practice for any System z configuration is that there should be no single points of failure in the configuration, *especially* any part of the configuration that is outside the protected data center environment. System z is designed to automatically recover from any connectivity failure, but that requires two things:

- At least one alternative path must be available to fail over to.
- The operating system (z/OS for example) and the connected devices must have visibility to the link failure, so that they can recover any requests that might have been in transit over that link at the time of the failure.

If a WDM or cross-site fiber fails, connectivity continues to be available through the other WDM *when both paths are configured to have no single points of failure*.

WDM-based protection

Different WDM vendors offer different protection schemes. One example of this is Reconfigurable Optical Add/Drop Multiplexer (ROADM)-based restoration.

Fiber switching modules are another option. This option is commonly known as *trunk protection* or *optical switch protection*. Most WDM suppliers offer this solution for mainframe implementations. The outline for this is shown in Figure 3-12.

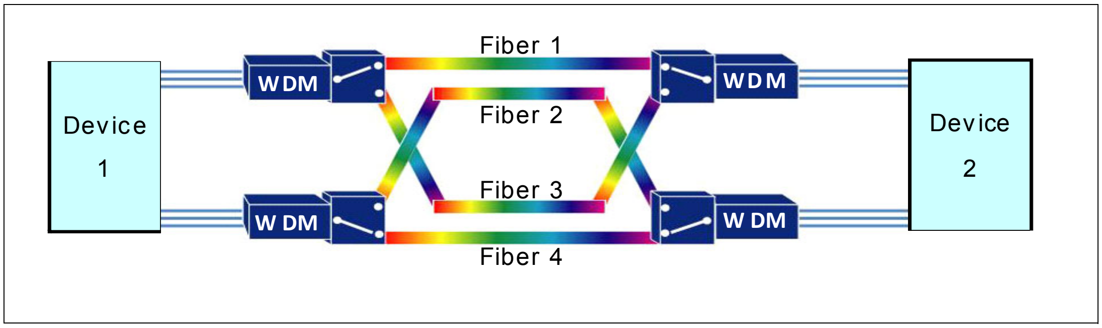


Figure 3-12 WDM-based protection using optical switches

In Figure 3-12 on page 108, four pairs of inter-site fibers are required. The fiber switch is a module within the WDM, and is connected to the WDM trunk where all lambdas are present. In the example, fiber 1 and fiber 2 use one route, and fiber 3 and fiber 4 share a different one. In normal operation, the upper pair of WDMs in the figure runs on fiber 1, and the lower pair of WDMs use fiber 4 as the active connection. Fiber 2 and fiber 3 are only used in failure scenarios.

If a break occurs in the fiber duct containing fiber 1 and fiber 2, the fiber switch of the upper system flips all of its traffic to fiber 3, while the lower WDM systems continue to run on fiber 4. This is shown in Figure 3-13. This switchover will normally take less than 50 ms.

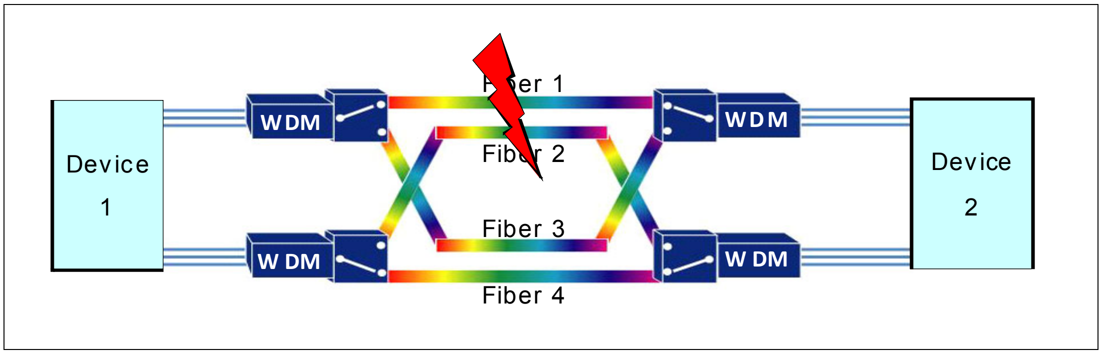


Figure 3-13 WDM protection switching using a fiber switch

WDM-based protection can be combined with client-based protection for added resilience. Client-based protection ensures that the connected devices are aware of the failure, and will drive normal link-error recovery processing. WDM-based protection automatically brings the backup path online so that both WDMs continue to operate.

In a System z environment, Client-based protection *must* be enabled, so that the System z devices can start their normal link recovery processing. If WDM-based protection is also enabled, the loss of one fiber or one route causes the WDM to automatically switch to the backup route. From the perspective of the System z devices, the event appears as a temporary loss of signal, and they will recover in a similar manner to how they would recover if you unplugged, then quickly re-plugged a fiber that was in use.

The advantage of using WDM-based protection is that you do not lose any bandwidth between the two sites. If you only use client-based protection, and one fiber or one route goes down, you would then be operating with just one fiber, meaning that you have lost half of your inter-site bandwidth. The combination of client-based and WDM-based protection means that, even in the event of a fiber or route outage, all services remain operational. Even double failures, such as a WDM and fiber failure, can be catered for.

However, cost and availability of the long-distance fiber must be taken into consideration. Additionally, remember that both client and WDM recovery are running in parallel, so it is impossible to predict which one will complete first, meaning that the recovery actions following a link failure cannot be reliably predicted in advance. For this reason, it is imperative that automation is put in place to ensure that the System z operations and technical personal are aware that there has been a failure, even if the system recovered automatically.

Important: During WDM failover switching, the connected devices must be informed about the event. The optimal approach is to force a loss of light/signal at the ports of the client devices. *Masking such events can lead to unstable environments.*

3.8 Connecting WDM into the end-to-end architecture

Figure 3-14 illustrates how a WDM system is established as part of the end-to-end connectivity solution.

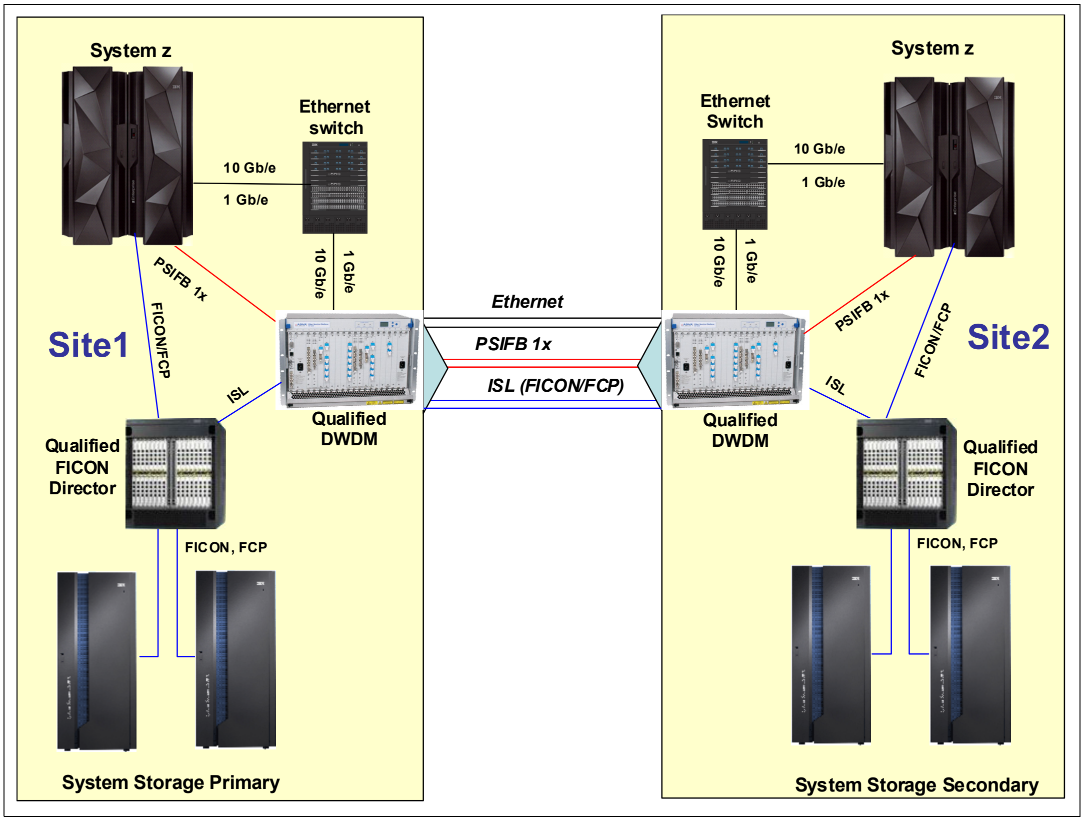


Figure 3-14 WDM systems in an end-to-end connectivity solution

Figure 3-14 shows the core technology components in a sample end-to-end connectivity solution across two data centers. The physical connectivity is not shown. In this example, the WDM is used to extend Ethernet (OSA), PSIFB 1X, and FICON/FCP using ISLs:

- Coupling links are connected directly to the WDM boxes in each site.
- FICON and FCP links are typically connected using a director or switch. FICON and FCP connectivity is usually extended using the ISLs, which are transported using the WDM infrastructure.
- Connections like OSA can also be transported through the WDM using an Ethernet switch or router.

3.9 Additional WDM capabilities

Obviously, your first priority will be to identify candidate WDMs that support all of the different types of traffic that you need to move between your sites. Depending on your configuration, this might mean that not every qualified WDM can meet your requirements. For example, some qualified WDMs do not support 10 Gb ISLs.

Having identified the set of WDMs that meet your needs, you should then look at optional features that might provide additional value for you. Because the WDMs are developed and manufactured by different vendors, different WDMs will offer different options. Some options might be attractive to you, and others might not be.

One example is an encryption capability. Because the WDM acts as the focal point for all of your traffic before it exits your data center, performing encryption in the WDM might be an attractive way to be sure that no unencrypted data is transmitted outside your premises. Having one device that performs all encryption might also be easier to manage (from a key management perspective) than if each of the devices connected to the WDM performs its own encryption (with its own set of keys that need to be managed).

This is just one example. The optional features will vary from one vendor to another, and some features might be available from several vendors. Nevertheless, we encourage you to include an investigation of these features as part of your evaluation process.

3.10 Selecting a WDM

Although WDMs might deliver cost savings compared to a configuration with similar connectivity and capabilities that does not use WDMs, they are still a significant investment. The following brief list includes some of the things that you should consider when selecting the most appropriate vendor and model for your enterprise:

- Does the precise configuration that you are considering conform to the IBM qualification letter for that device?

  The qualification letters are detailed about precisely what devices, connectivity types, and features have been qualified and should be studied carefully.

- Does the WDM support all the types of connectivity that you need to use across your sites?
- How much latency does the WDM add to each request?

  At large distances, additional latency might constitute a small part of the overall total, because the time it takes the light signal to traverse the distance adds so much latency. However, over shorter distances, additional latency becomes more important. For example, a direct fiber connection over 10 km would add 100 ms to the response time for a request sent over that link. If the WDM takes 50 µs to process each request, that is the equivalent of adding another 5 km between the sites.

- How discernible is the WDM? That is, does the WDM place restrictions on protocols and features that you can use on connected devices? If it does have restrictions, would those restrictions affect you?
- What management capabilities or tools does the WDM provide? Does it feed into your existing system management tools?
- Does the WDM support both client-based and WDM-based protection? Can both types of protection be used concurrently?
- Does the WDM provide any error correction capability?

  Particularly in configurations where a large number of temporary errors might be experienced, you do not want all of them being presented as an interface control check (IFCC) back to the connected devices.

- If you are interested in a feature that is supported on more than one platform (encryption, for example), obtain and compare the financial and performance costs of each alternative.
- If you buy the connectivity from a telco/ISP, confirm that the WDM they propose using is a qualified one, and that it supports all of your traffic types and any features that you might be using on your SAN switches.

These are just some of the things that you should consider when evaluating different WDMs. This is a complex and constantly-evolving technology, and the WDM you select will be an integral part of your end-to-end connectivity architecture. It is important to select devices that will support your needs today, and support your strategy for the future.

3.11 WDM connectivity preferred practices

The following list describes some high-level preferred practices for WDM connectivity:

- Have at least two completely failure-independent WDMs in each site (separate power supplies, separate locations, separate cooling, separate entry points, and so on).
- Balance all of your cross-site connectivity across those WDMs.
- All devices should be connected to both WDMs.
- The long-distance fibers must run on different fiber routes.
- The long-distance fibers should never be within 10 meters of each other at any point between the two sites, and they should never cross.
- Where possible, combine client-based protection with WDM-based protection.
- If you connect your data centers using a telco/ISP, make sure that the previous points are taken into consideration in the solution that they are proposing.

Tip: Not all WDM vendors will require a dual point-to-point configuration for their equipment. However, for high-availability solutions, it is the most appropriate choice.

4

Common multisite models

This chapter provides information about the most common configurations for connecting IBM System z components over more than one site. We use the IBM Geographically Dispersed Parallel Sysplex (GDPS) offerings to illustrate the most common configurations and their capabilities. This does not mean that GDPS *must* be used in an extended distance configuration.

However, there are GDPS offerings that are aligned with the most common extended distance configurations, so they provide a convenient model. Additionally, if you want more detailed information about any of the models, see the IBM Redbooks publication *GDPS Family - An Introduction to Concepts and Capabilities*, SG24-6374.

The primary objective of this chapter is to show the direction of the industry, and how enterprises are solving configuration issues, so that you can examine models that most closely reflect your planned configuration. The detailed information that you will need to develop your extended distance connectivity architecture is provided in Chapter 5, “Planning” on page 125.

4.1 Considerations for extended distance models

Most enterprises configure System z data centers that are separated by extended distances along the lines of one of three models:

- Two data centers within metro distance of each other, using synchronous disk mirroring
- Two data centers at larger distances using asynchronous disk mirroring
- Three data centers, with two connected at metro distance using synchronous mirroring, and the third at a greater distance using asynchronous disk mirroring

Each model has variations, depending on the constraints imposed by the location of the data centers, and the availability and disaster recovery requirements of the business.

However, before we examine different models, we need to level-set by briefly describing and contrasting some relevant concepts.

4.1.1 Difference between continuous availability and disaster recovery

The terms *continuous availability* and *disaster recovery* are often used interchangeably. Both deal with information technology (IT) operations. However they are not the same, and can even place competing requirements on those responsible for delivering the service.

Continuous availability generally means that the applications should be available with minimal planned or unplanned outages. Any event that could threaten the availability of the application should be planned for in advance, and a mechanism identified to ameliorate the risk. For example, if loss of access to a device would affect the availability of the application, two failure-independent paths to the device should be configured.

Disaster recovery readiness means that after a disaster, the applications can be recovered within a certain time (called the recovery time objective (RTO)) and within a certain loss of data (called the recovery point objective (RPO)). Obviously, the ideal RTO would be “immediately” and the ideal RPO would be “with no loss of data”. However there are financial and practical limitations that must be considered.

For example, consider that you have two data centers. And because you are using synchronous mirroring, you expect that all your data is secure. Due to some event, you suddenly lose connectivity to the secondary disks.

If the cause of the loss of connectivity is that someone accidentally removed the wrong cable, this is not a real disaster scenario. You can quickly re-establish connectivity and resynchronize the disks, and all processing continues as normal. In this case, to achieve your continuous availability objective you would obviously want to keep the production systems up and running during this event.

But what if the loss of connectivity was caused by a fire in the computer room, and the first device to be affected is the dense wavelength division multiplexing (DWDM) that is used to connect the two sites? The initial symptoms will appear to be the same: A loss of connectivity between the sites. To meet your continuous availability objective, and not realizing the true cause of the connectivity failure, you keep your production systems up and running.

However, you are now applying updates to the primary disks that are not being mirrored. Five minutes later, the fire spreads and destroys one of your primary disk subsystems. You now have five minutes’ worth of data updates that were applied to your primary disks but not to the secondary disks and that you will somehow need to recover.

In the latter scenario, the preferred choice would have been to stop the production systems at the instant that you lost the ability to mirror updates to the second site. However, if you took that same action in the first scenario, you would have caused an unnecessary outage to the production applications (this is often referred to as a *false freeze* situation, because the trigger for stopping the mirror is not a true disaster).

This is an example of how disaster recovery and continuous availability can sometimes appear to be incompatible. Using products such as IBM GDPS will help you manage your environment to your selected strategy. *Which* strategy you choose is really a business rather than an IT decision. The business needs to decide which is more important to it: The highest possible levels of availability, or ensuring that no data is lost in a disaster.

4.1.2 Relationship between distance and continuous availability

Generally, when clients request continuous availability, what they really want is continuous, or near-continuous, *application* availability. The clients might not know (or care) what system they are using: Their requirement is that their *application* is available when they request it.

The critical prerequisite for near-continuous application availability is that there are no single points of failure in the components that service the application. Therefore, taking this concept to its logical conclusion, the application must be running in two sites, so that if one site becomes unavailable, the application is still available in the other site. In System z, this would be achieved by running the application in both sites, with the systems in both sites sharing the same set of primary disks and the coupling facility (CF) structures that enable data sharing[1].

In a single-site configuration, disk response times of 1 millisecond (ms) or less, and CF response times of single-digit µs, are common. In a multisite configuration, the speed of light can be a significant part of the observed response time, because the round-trip time for light in a fiber is about 10 microseconds (µs) per kilometer (km).

Moving the CF and primary disk 10 km from the IBM z/OS central electronics complex (CEC) would increase CF and disk response times by about 100 µs. Because normal disk response times are in the 1000 µs range, the effect of this distance on disk response times (10%) is relatively less than the effect on CF response times (1000%).

This phenomenon effectively limits the distance between the sites in a multisite sysplex configuration, particularly when data sharing is being used across both sites, to less than the maximum distance that the underlying technology supports. If all systems accessing the primary disk and CFs are in the same site and only the secondary disk and tape are remote, larger distances might be acceptable.

The effect of distance on your transaction response times, and the level of effect that is acceptable, will vary from enterprise to enterprise, and even from application to application. For this reason, IBM strongly suggests that any client considering implementing a multisite sysplex should perform a benchmark of the proposed configuration before making a final decision.

An increasingly common configuration is a three-site configuration. If you only have two sites, you might not want to place them close together, in case a single event affects both data centers. Alternatively, placing them a larger distance apart might limit or even eliminate your ability to use functions such as HyperSwap or cross-site data sharing.

[1] The GDPS/active-active Sites solution is intended to provide the ability to effectively share data across sites that are separated by hundreds or thousands of kilometers.

However, if you have a third, out-of-region, site, you can place the first two sites closer together (thereby reducing the performance effect of distance), while still having the third site to be able to recover from a regional event that affects the first two data centers.

4.1.3 Industry terms for site roles

There are several terms used to describe configurations that are spread over more than one site. To avoid confusion, we describe the terms, and clarify how they are used in this book:

Active/active A multisite sysplex configuration, where applications are running in more than one site.

This term originated in the distributed world, where there was traditionally only one application per system.

In a mainframe configuration, there would typically be many applications running in each z/OS instance. As a result, an active/active sysplex might contain applications that are spread over both sites, with both instances updating the same database. Or you might have one set of applications running in one site, and a different set of applications running in the other site. You might also have a combination of these.

Active/standby A multisite sysplex where all of the applications only run in one site at a given time.

MultiSite Workload (MSW) This term tends to be used in relation to GDPS/Peer-to-Peer Remote Copy (PPRC) configurations. A MSW configuration is a multisite sysplex with a GDPS controlling system in at least one of the sites, and production applications running in both sites.

SingleSite Workload (SSW) Another GDPS/PPRC term. A SSW is a multisite sysplex, with a GDPS controlling system running in the same site as the secondary disks, and all applications running in the other site.

Site1 In GDPS terms, Site1 is the site that normally contains the primary disks. For simplicity, we will use the same terminology in this book.

Site2 In GDPS, Site2 is the site that normally contains the secondary disks, so we will use that term in this book as well.

Site3 Given that we call the first two sites Site1 and Site2, we will call the third, out-of-region, data center Site3 when talking about three-site configurations.

Important: It is important not to confuse a multisite sysplex active/active configuration with the GDPS offering called GDPS/Active-Active Sites. The GDPS/Active-Active Sites solution uses software data replication techniques to send database updates from one site to another site. The two sites would typically be located beyond metro distance from each other.

Because of its unique requirement to have a controlling system in the same site as the secondary disks, the GDPS documentation will use the terms MSW and SSW when referring to a GDPS/PPRC multisite sysplex configuration. If we are referring to the GDPS/Active-Active Sites solution, we will refer to it by that name.

In *this* document, because we do not want to infer that GDPS *must* be used in a multisite sysplex, we will use active/active to refer to a multisite sysplex with production running in both sites, and active/standby to refer to a multisite sysplex with production only running in one site.

For more information about the GDPS offerings, see the IBM Redbooks publication, *GDPS Family - An Introduction to Concepts and Capabilities*, SG24-6374.

4.1.4 Considerations for distributed systems

There are few, if any, mainframe-only installations left in the world. The current IT model is that applications will be deployed on multiple platforms. To bring more control and increased efficiencies to the distributed systems, it is becoming increasingly common to find that data for the distributed systems is in the same disk subsystems as the System z data. Additionally, the complexity of some modern applications means that a single unit of work might update data on System z, a UNIX system, and X-86 based systems.

Even if such applications are not yet in use in your enterprise, it is likely that they will be in the foreseeable future. And even if multiple platforms are not affected by a single transaction, there is no denying that there is an ever-closer relationship between the data on System z and on other platforms.

Apart from the efficiencies of scale and the management advantages, the relationship between the data on the different platforms makes it attractive to store all of the data in one set of disk subsystems, and to manage the data (at least from a disaster recovery perspective) in a similar manner.

If the relationship between the data on the different platforms is acknowledged in your enterprise, thought needs to be given to the location of the primary data for each platform. Some enterprises have found that placing the primary data for all platforms in the same disk subsystem causes performance issues.

The expedient resolution to this situation problem is to place the primary copy of distributed data in the same subsystem as the secondary copy of the System z data and vice versa. This might be effective at addressing the performance issues.

However, consideration needs to be given to the issues related to mirroring in opposite directions. For example, consider what would happen if some failure disrupts the ability to mirror between the two sites. If the systems continue running, you would have the situation where each site will contain data for one platform that is frozen at the point in time of the failure, while the data for the other platform continues to be updated.

If the cause of the mirroring failure transpires to be a disaster event, you now have two sets of data at different points in time. This is, at best, inconvenient. At worse, it might represent a data integrity issue. For this reason, we suggest that serious consideration should be given before a decision to mirror in two directions is made. Additionally, the owners of the data should fully comprehend the issues and agree to your intended solution.

4.1.5 Performance considerations

The speed of light seems like an academic topic to most people. However, when considering multiple data centers, the time it takes light to travel over relatively small distances suddenly becomes real. At the time of writing, 5 µs would be considered to be a “good” CF service time. If you increase the distance between the z/OS CEC and the CF CEC by 5 km, that will add 50 µs to the CF service time (an 11-fold increase).

However, it is not even as simple as that, because increasing service times can have difficult or even impossible-to-predict secondary effects. For example, if CF service times increase, that could increase transaction elapsed times, meaning that they remain resident in storage for longer, increasing use of real storage, and giving the transaction manager more concurrent transactions that it has to manage.

This is an important topic when considering how far apart you can place your sites, especially if you hope to span a single sysplex over both sites. However, it is too lengthy and complex to include in this chapter, so we have placed it in an appendix of its own, Appendix A, “Performance considerations” on page 179. We strongly suggest that you read that appendix before making a final decision on how far apart you want your sites to be located.

4.2 Consider your objective

Before we provide information about the most common configurations, it is vital that your enterprise is completely clear about its reasons for considering a multisite configuration. In particular, the application owners must clearly state their requirements. And the technicians must ensure that the application owners understand the implications and limitations of the solution they are proposing.

For example, if you have a history of site outages affecting application availability, the application owners might see multisite as the solution to that issue. If that is the case, they must understand that there is a relationship between performance, distance, and protecting your data. Placing the two sites close together will minimize the performance effect (and the connectivity infrastructure cost), but it exposes your data to the risk of a single event destroying or incapacitating both sites.

If the priority is to ensure that no single event could ever wipe out all their data, it is likely that the target distance between the sites will be measured in tens, hundreds, or possibly even thousands of kilometers. In that case, the application owners must understand that switching from one site will be disruptive and is likely to take several hours to complete.

Just as the technicians might not understand all of the subtle nuances of the business, similarly, it must be remembered that the application owners are probably not familiar with disk and CF response times, and the relative effect of distance on them. Therefore, before any significant investments are made, it is vital that all parties completely understand the benefits and limitations of the proposed configuration.

Having agreed on the objective of the project and the requirements of the business, we hope that the following descriptions of some common multisite models will be helpful.

4.3 Metro distance active/active configuration

A metro distance active/active configuration has production systems in both sites. To achieve the availability benefits of two sites, all critical applications will use sysplex data sharing and will run in both sites, as shown in Figure 4-1. The disk subsystems will support the ability to non-disruptively switch between primary and secondary, and each site will be configured with all of the hardware necessary to support all applications in the event that the other site is unavailable.

Because the applications will be running in both sites, the systems in both sites require access to all of the devices necessary to run those applications, and to be in the same sysplex as the systems in the other site. The most obvious example is the primary disks. But, they will also need to be in the same Coordinated Timing Network (CTN), and each system in the sysplex requires access to the CFs in both sites.

From a disk channel bandwidth perspective, systems in both sites need access to the primary disk. Therefore, the systems in Site2 need *at least* as much channel capacity between them and the primary disks as the systems in Site1 have.

Additionally, to fully use the flexibility provided by HyperSwap, the Site1 systems will require sufficient channel capacity to the Site2 disks to provide service required levels. And, of course, you need connectivity between the primary and secondary disk for mirroring purposes, and you need to be configured to enable mirroring in either direction.

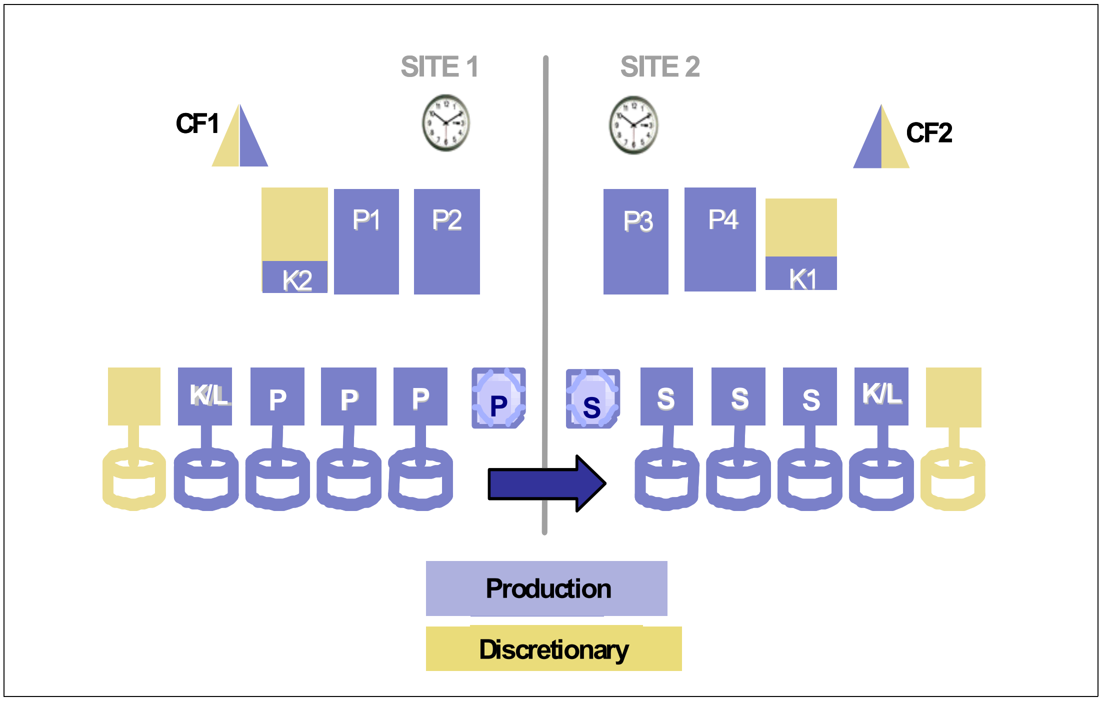


Figure 4-1 Typical GDPS/PPRC active/active configuration

The active/active configuration has stringent connectivity requirements for the following reasons:

- Systems must be able to run in either site.
- Systems in both sites must be in the same sysplex and the same CTN.
- Systems should be able to use devices in either site.

- Because of the latency effect of the distance, cross-site input/output (I/O) and CF request response times will be greater than local response times. The increased response time increases unit control block (UCB) and CF subchannel usage. This increased usage must be taken into account when sizing the channel and CF link bandwidth between the sites.
- Because the links between the sites run through public property, there is a heightened risk of damage. Therefore it is vital to ensure that there is no single point of failure between the links connecting the two sites.

The GDPS offering that is most commonly used in this configuration is GDPS/PPRC. For more information, see *GDPS Family - An Introduction to Concepts and Capabilities*, SG24-6374.

If you are using z/VM or Linux on System z, there might be some additional considerations. Both z/VM and Linux on System z support fixed-block architecture (FBA)-format disks, as well as HyperSwap support. Therefore, you must plan to run those devices in either site if you are including Linux systems in the scope of GDPS control. And, as with the z/OS disks, those devices must be accessible to the systems in the other site.

4.3.1 Connectivity considerations

The active/active configuration has additional stringent connectivity requirements:

- The diversity of devices that must be accessible to systems in the other site is likely to be higher than in the other configurations.
- If you use protocol converters to access some devices, such as parallel channel-attached printers or check sorters, you must ensure that those converters function with any equipment used for connectivity between the sites.
- Because of the performance implications of distance, the distance between the sites is likely to be limited to some tens of kilometers.

A qualified wavelength division multiplexing (WDM) is required in this configuration, because you must provide CF and Server Time Protocol (STP) access in addition to Fibre Channel connection (FICON) and mirroring connectivity between the two sites.

4.4 Metro distance active/standby configuration

An active/standby configuration is similar to an active/active configuration. An active/standby environment also consists of a multisite sysplex, however, it differs from an active/active configuration in that all production applications normally run in just one site, with the other site containing hot standby systems and possibly a GDPS controlling system. If using GDPS/PPRC in an active/standby configuration, the controlling system will normally run in the site containing the secondary disk volumes.

The multisite sysplex must be configured with redundant hardware so that the production applications can run in either site, using only the devices in that site, as shown in Figure 4-2 on page 121. The bandwidth that is required between the sites depends on the following factors:

- The bandwidth required for mirroring.
- Whether you require the ability to run systems in one site, with the primary disk in the other site.

- Whether there will be times, for example, during a planned switch from one site to the other, when you run production applications in both sites at the same time. If so, cross-site coupling link capacity must be sufficient to handle the expected volume of CF requests.

Figure 4-2 shows a typical GDPS/PPRC active/standby workload configuration. The logical partitions (LPARs) in blue (P1, P2, P3, and K1) are in the production sysplex, as are the CFs CF1 and CF2. The primary disks are all in Site1, with the secondaries in Site2.

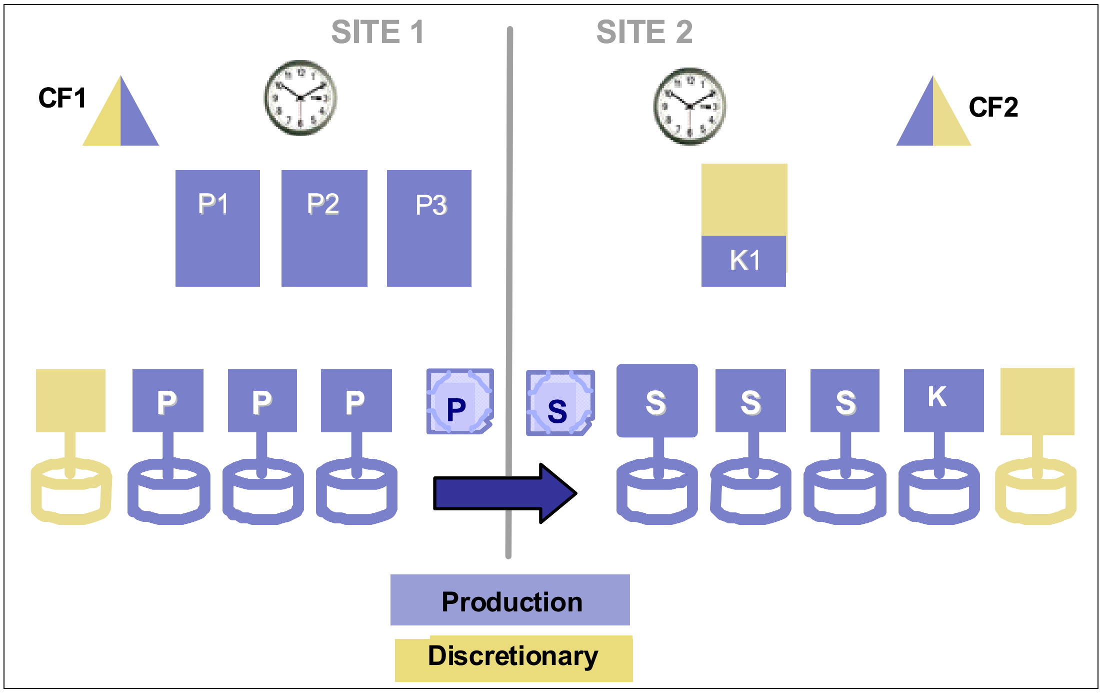


Figure 4-2 GDPS/PPRC active/standby configuration

In this example, all of the production systems are running in Site1, with only the GDPS controlling system (K1) running in Site2. Notice that system K1's disks, those marked K, are also in Site2. The unlabeled boxes represent work that can be displaced, such as development or test systems.

4.4.1 Active/standby connectivity considerations

The connectivity considerations for an active/standby configuration are essentially the same as those for an active/active configuration. The devices that need to be connected are probably the same, depending on how much flexibility you need to run the systems in one site, but using devices in the other site.

The more flexibility you require, the more comprehensive the connectivity requirements. However, the active/standby configuration does require a cross-site sysplex, which means that your connectivity infrastructure must support coupling links and STP.

4.5 Extended distance disaster recovery configuration

An extended distance disaster recovery configuration is simpler from the perspective that all production systems run in just one site, and there are no cross-site sysplex or STP considerations.

In this configuration, the production workload will only be run in the second site in case of a disaster being declared at the normal production site. Because of the extended distance, some form of asynchronous mirroring would be used. This precludes the use of HyperSwap, meaning that regular planned site swaps are unlikely. A sample configuration is shown in Figure 4-3.

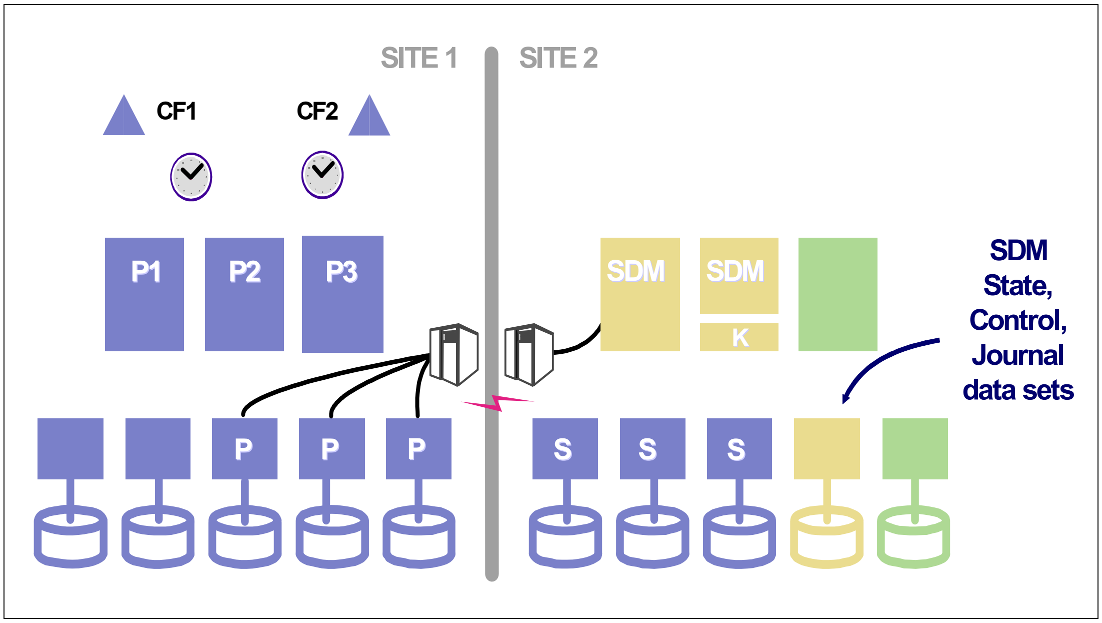


Figure 4-3 Sample extended distance disaster recovery configuration

The hardware that will be required in the remote site will depend to some extent on the type of mirroring that you use. If you use IBM zSeries Global Mirror (zGM, previously known as extended remote copy (XRC)), the remote site must contain one or more z/OS systems that drive the mirroring function. If you use Global Mirror, a remote system is not required for Global Mirror. However, you will need CECs in that site to perform disaster recovery tests, and for use in the case of a real disaster.

There was a time when channel extension devices that were used with asynchronous mirroring would be configured differently at the remote site than at the production site. However, this is no longer the case. Nevertheless, when performing your connectivity planning, you should include the scenario where you will have to mirror in the opposite direction (perhaps in the case of returning to normal after a prolonged but temporary outage of the primary site).

Apart from disk mirroring, the only other “applications” that are likely to be connected across the two sites are tape mirroring, and writing to remote tapes using FICON Acceleration. The specific requirements will depend on your tape devices, the mirroring technology they use (is it based on Transmission Control Protocol/Internet Protocol (TCP/IP), for example), and whether you will use FICON Acceleration. If you *do* plan to use FICON Acceleration, that feature must be installed on the switches at both sites.

4.6 Three-site configuration

An increasingly common configuration is one consisting of three sites. Two of the sites would be located relatively close together (tens of kilometers or less), thereby minimizing the performance effect, enabling cross-site data sharing (and the near-continuous availability associated with that) and regular planned site switches.

The third site would be located at a distance such that no natural or man-made disaster would affect all three sites. The third site would be viewed as insurance. Having that insurance enables you to place the two proximate sites closer together than you would if they held the only two copies of your corporate data.

A sample three-site configuration is shown in Figure 4-4.

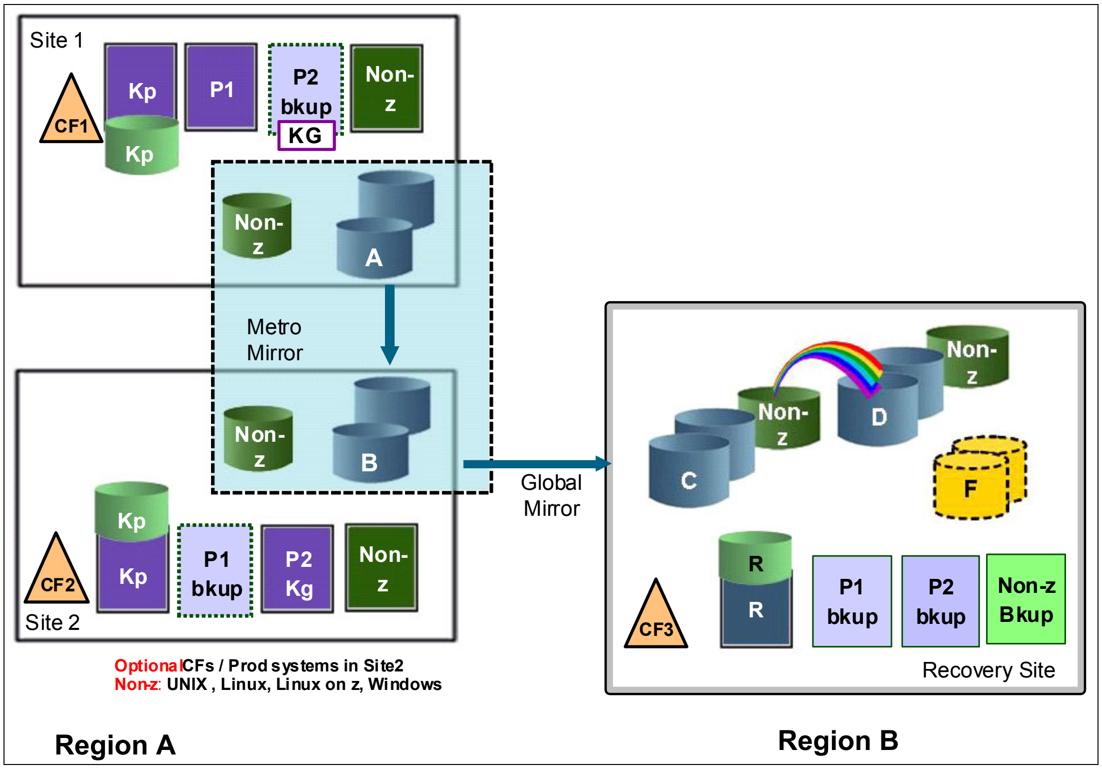


Figure 4-4 Sample three-site configuration

The IBM three-site GDPS solutions support both zGM and Global Mirror forms of asynchronous mirroring. The zGM solution supports only System z data. When using this solution, the same disk acts as primary for the synchronous mirroring *and* the zGM asynchronous mirroring.

The GDPS three-site solution that uses Global Mirror supports both System z and distributed data. When using Global Mirror, the secondary disk for the synchronous mirroring configuration acts as the primary disk for the Global Mirror configuration.

4.7 GDPS/Active-Active

The GDPS/Active-Active configuration enables systems in two sites to run the same application, and to update the same set of data. Distance between the sites is unlimited, because this configuration is based on software data replication rather than on sysplex data sharing.

In addition to the critical applications that you use GDPS/Active-Active for, you probably have other applications that do not support GDPS/Active-Active, or that do not require that level of availability. For those applications, GDPS/Active-Active supports cooperation with GDPS/PPRC or GDPS/Metro Global Mirror (MGM). The distance considerations for those applications are the same as for GDPS/PPRC or GDPS/MGM.

4.8 Selecting the model that is appropriate for you

The requirements of your enterprise provide you with a “wish list” of what the business would want to have if there were no financial or technology constraints. As a technician, you then need to balance that with what the technology can support now and in the future, adding industry directions and government regulations into the mix.

You also need to take into account your current environment. If you have two data centers that are 30 km apart today, do you consolidate all production into one site and treat the second site as a disaster recovery site? Or do you eliminate one of the sites and replace it with another closer one that would be more suited to cross-site data sharing, or a more distant one that would provide more isolation from a regional disaster?

Perhaps you decide on two proximate data centers, with a strategy to eventually implement a distant third site, and implement the GDPS/Active-Active Sites solution encompassing all three sites and providing a local and remote near-continuous availability capability.

Hopefully, the information provided this far in this document will be valuable in assisting you to narrow down the list of potential options.

Chapter 5, “Planning” on page 125 contains information to help you move forward with planning for your extended distance configuration.

5

Planning

This chapter provides information to help you create an end-to-end architecture that will support your current and future business requirements as they relate to a multisite interconnected configuration. Specifically, we will provide information about the following areas:

- Creating your connectivity architecture group
- Identifying your objectives
- Documenting your System z configuration
- Creating a balanced end-to-end configuration
- Security considerations
- IBM qualification for extended distance devices
- Selecting your extended distance equipment
- Benchmarking your proposed configuration
- Service provider requirements
- Physical connectivity considerations

5.1 Creating your connectivity architecture group

The section 1.5, “Role of the connectivity architecture group” on page 12, provided information about the need for a group that will own the end-to-end connectivity architecture. If such a group does not already exist in your organization, it should be created now, at the start of your project.

The creation of this team does not necessarily mean that more people must be hired by your organization. It might just mean a realignment of existing responsibilities or reporting structures. However, it is important that this team is recognized by all involved parties as the owner of the architecture. Therefore, the team should be in place before any discussions about identifying requirements and building the architecture commence.

It is likely that the groups responsible for the individual components within the architecture (network, storage, switches/directors, and so on) will possess deeper technical skills within each of their respective areas than the members of the connectivity architecture group. However, the ownership of the overall architecture is logically within the IBM System z department for the following reasons:

- The applications that run on System z depend on all of the components within the architecture, so the System z group has a vested interest in ensuring that the total end-to-end infrastructure works as efficiently as possible and is supportive of future growth.
- System z typically has the highest availability objectives of all of the IT platforms, so it should have the final say over any changes or plans that might affect the ability to meet those objectives.
- System z has more stringent requirements on the connectivity infrastructure. Not all dense wavelength division multiplexing (DWDM) and director firmware levels are qualified to work with System z. But levels that are qualified for System z should work with distributed systems. Therefore, it makes sense that the architecture owners should be in the System z group.

Not having a group responsible for the overall end-to-end architecture is akin to trying to build the world’s best car by combining the engine from the most powerful car, the brakes from the fastest-stopping car, the suspension from the best-handling car, the gearbox from the smoothest car, and the interior from the most comfortable car, and expecting the result to be the best car in the world.

In practice, a normal family car would probably be more effective overall than a car built from these fantastic components, which were not designed to work with each other. Similarly, your end-to-end connectivity infrastructure will be the most effective if it adheres to a well-designed and managed end-to-end architecture.

Information about the suggested scope and responsibilities of the connectivity architecture group is provided in 1.4, “The importance of an end-to-end architecture” on page 7 and 1.5, “Role of the connectivity architecture group” on page 12.

Tip: To avoid wasting valuable time later in the process, it would be wise for the members of the connectivity architecture group to spend a little time now getting familiar with the IBM GDPS qualification program. Information about the program is available in the following places:

- "System z qualification and testing programs" on page 23
- IBM Redpapers publications related to the qualification letters for various DWDM devices. These are available on the IBM Redbooks website:

  http://www.redbooks.ibm.com

5.2 Identifying your objectives

One of the first responsibilities of the connectivity architecture group should be to identify and agree upon what your business is trying to achieve by having multiple data centers. As you progress through this chapter, you might encounter limitations that would stop you from being able to achieve all of your objectives. By clearly documenting your objectives, you create a checklist that helps highlight any relevant restrictions.

For example, is your objective purely disaster recovery? If so, what are your recovery point objectives (RPOs) and recovery time objectives (RTOs)? If your enterprise absolutely must have a zero data loss capability, that means that you *must* have a synchronous mirroring solution. And that in turn places realistic limits on the distance between the sites.

Alternatively, if your primary data center is in an area that is prone to natural disasters (earthquakes, flooding, ice storms, and so on), you have a requirement for an out-of-region disaster recovery (DR) capability. That requirement effectively forces you to use an asynchronous mirroring solution, which means that zero data loss is not an option.

Maybe your company has both a zero data loss requirement *and* an out-of-region DR requirement. To meet these conflicting requirements, the only effective solution is likely to be a three-site configuration, with one site within metro distance of your primary site and using synchronous mirroring, and the third site a large distance away (possibly even on another continent) and using asynchronous mirroring.

If having three sites seems to be overkill, consider that some enterprises are investigating *four*-site System z solutions at the time of writing of this book. Given the ever-increasing reliance of business, government, and society on information technology (IT), DR and continuous availability requirements are likely to become ever-more stringent. Enterprises that are considering a four-site solution have determined that the risk of having just a single site after a regional disaster is unacceptable, therefore their interest in a fourth site.

Perhaps a review of your outage history (both planned and unplanned) and your risk profile indicate that in addition to a DR capability, you also have a requirement for near-continuous availability for your critical applications. The first step toward near-continuous availability for any application is to remove all possible single points of failure. In a System z environment, that means that the applications should support Parallel Sysplex data sharing and dynamic transaction routing.

The next logical step is to span the sysplex across multiple sites. This gives you the ability to maintain application availability across planned outages of the transaction manager, database manager, operating system, and central electronics complex (CEC), in addition to planned *site* outages (assuming that you implement HyperSwap or a similar capability).

If your objective is to stretch your data sharing sysplex across multiple sites, that imposes realistic limits on the distance between the sites. Three-site configurations are especially well suited to multisite data sharing, because the existence of the third site gives you the ability to recover from a regional disaster that would incapacitate the two metro distance data centers. Therefore, you can place those two sites closer together than you might want to if you only had two sites.

If you want to have a multisite sysplex, how do you plan to distribute work across the sysplex members? Will one site be the primary site, where all the staff are located, all the applications normally run, and all the peripherals (printers, check sorters, tape drives) are located? Or will the applications run in both sites, with the sites being mirror images of each other, providing the ability to run all critical applications in either site if necessary?

Perhaps your most important requirements are the highest possible levels of availability, combined with minimal performance effect (meaning a limited distance between the sites), but you cannot cost-justify a third site. In that case, you might not want to "place all your eggs in one basket", so you might enter an agreement that enables you to place just disk or tape in an out-of-region data center.

The lack of a running z/OS image in the remote location means that IBM z/OS Global Mirror (zGM, previously known as extended remote copy (XRC)) is not an option, so some other asynchronous mirroring mechanism, such as Global Mirror, must be used.

There are also non-technical requirements that must be considered. Are there government regulations for your industry that specify a minimum distance between the sites? Are there government regulations that prohibit keeping a copy of client data outside the country? Are there regulations governing the maximum amount of application downtime in a year? Do your corporate insurance or your auditors specify requirements for distance, continuous availability capability, or regular DR testing?

These are just some examples. The range of requirements for System z clients is as wide as the types of enterprises that rely on System z. If you have not already reviewed it, Chapter 4, "Common multisite models" on page 113 contains information that might give you more ideas about what capabilities your company might want to use.

It is vital that you have a clear understanding of the needs of your business, as well as any financial limitations that must be taken into consideration when looking at the technical options for meeting your companies objectives. The lack of a clear mission statement might mean that you create a more elaborate (and more expensive) solution than your business actually requires. Conversely, you might develop an architecture that is not capable of meeting the business needs.

As part of the process of documenting the requirements of the business, you should start building a picture of the responsibilities of each data center. Will it only ever run production systems in the event of a regional disaster? Will there be operations staff in that location, meaning that it might be used for roles such as bulk printing or check sorting?

When the normal and disaster-state role of each data center becomes clear, this will make it easier to identify the devices that will need to be in each data center, *and* the connectivity requirements of those devices.

While gathering the objectives of the business, do not forget to include the requirements of the IT department. Perhaps one of their primary objectives is to be sure that the solution you deploy will be supported. The easiest way to do this is to ensure that it uses a configuration (switches or directors, and DWDMs) that has been qualified by IBM. This ensures support from IBM and the vendor if problems are subsequently encountered.

The IT department will also require a solution that can be effectively managed. The management aspects of an end-to-end architecture are covered in 1.5, “Role of the connectivity architecture group” on page 12.

5.3 Documenting your System z configuration

Having identified the objectives for your project, the next step is to create an inventory of the devices required to support your System z applications, and to investigate how your proposed configuration would affect those devices.

You should expect to encounter some inhibitors to creating an extended distance configuration that meets *all* of your objectives. Some of these inhibitors might be “show-stoppers”, although it might be possible to address or circumvent others. The important thing is to identify them as early in the planning process as possible, so that you can adjust your plans. Perhaps you might even have to go back and re-visit and re-prioritize your objectives.

The following list includes some concepts that you should consider in the context of designing your multisite configuration:

- Create an inventory of the devices in each data center that must be connected to the other data center:
  - Processors (CECs)
  - Switches:
    - Storage (Fibre Channel connection (FICON) Directors)
    - Network (switches/bridges, routers)
    - Physical layer (optics, cable standards, connector types, path panels)
  - Storage:
    - Disk
    - Tape
  - Others:
    - Printers
    - Check sorters
    - Encryption devices
    - Special devices
  - “Wish list” items (for example, additional resources that you want to include, but that are not *required*)
- Identify and document the maximum supported distance for each of the devices in your configuration. This list will help you identify what is possible, considering the devices in your inventory.
- Distance between data centers. If you have multiple existing data centers and want to use one or more of them, the distance between them needs to be factored into your planning.
- Available connectivity options:
  - Private fiber
  - Leased fiber (dark fiber)
  - Leased connections over a shared network

- Time available to implement:
  - Immediate or short term
  - Tactical
  - Strategic
- Financial limitations.

5.3.1 Creating your device inventory

Before you can decide whether your objectives are achievable, you must create an inventory of the devices that connect to System z and that currently exist in each or either site. Presumably all of those devices are used by one or more System z applications.

The inventory should contain a list of all of those devices, which site they are in, and which CECs and switches they are currently connected to. Table 5-1 shows the start of a spreadsheet that you could create to contain the information that you will need as you proceed through this chapter.

In a subsequent step, you will identify the devices required to support each critical application, and how that affects the enterprise's requirements in terms of which applications are required in which (or both) sites. That in turn will help you identify the connectivity requirements for each device.

Table 5-1 Device inventory (subset of attributes)

Vendor	Type	Model	Serial #	Port / Speed	Port / Type	Port / Connect or Type	Location
IBM	2827	H43	12345	PCHID 1 / 8 Gbps	MM	LC Duplex	Room 1 Grid 2D
IBM	2827	H43	12345	PCHID 2 / 8 Gbps	MM	LC Duplex	Room 1 Grid 2D
IBM	2827	H43	12345	AID 10 PSIFB 12X	MM	MPO	Room 1 Grid 2D

Depending on the size of your System z environment, the industry that you operate in, and the services that you provide, compiling your inventory might be a simple exercise, or it could be complex, requiring input from many groups.

The best place to start is by looking at a list of all of the devices that are currently connected in your System z environment. This information might come from several sources:

- One of the tools that process an input/output definition file (IODF) file:
  - Hardware configuration definition (HCD)
  - Hardware Configuration Manager (HCM)

    HCM is a PC-based client/server interface to HCD that combines the logical and physical aspects of hardware configuration management. HCM displays an interactive configuration diagram that enables you to maintain both the logical connectivity data in IODF, and the physical configuration information.

Note: This exercise will be much easier if you use a standard that a given physical device will be referenced by the same device number in any system that uses it. If you have a situation where a given device number could be associated with a different physical device depending on which system you look at, that makes this exercise more complex.

It would be even easier if you have a device numbering convention that indicates the location of each device. For example, all devices with an odd number in the first digit of the device number are in Site1, and all devices with an even number are in Site2.

If you have just one site today, you should think about such a convention when deciding which devices will be in the second site, and their addressing.

- Various home-grown graphical tools or spreadsheets that you might use to document your hardware configuration.
- System, IBM VTAM®, and Transmission Control Protocol/Internet Protocol (TCP/IP) commands to display the defined and online devices that are accessible to each system.

  You might find that there are many devices defined in your IODF or network definitions that no longer exist. This is an opportune time to remove such definitions.
- Hardware Management Console (HMC) and Support Element (SE) displays that show information (node descriptor) about the device that is connected to every channel.
- Switch functions that provide information (device type, manufacturer, and serial number) about every device connected to every port on the switch.
- Computer room floor plans.
- Perform a physical audit of all mainframe-connected devices.

This should help you produce the inventory of currently-installed devices. The information for each device should include its physical location, which CECs or switches it is connected to, and the speed and type of its interfaces (for example, 8 gigabits per second (Gbps) and single-mode (SM) fiber).

Combining that list with information about your target configuration will help you identify devices that will be required but that are not currently installed. For example, if you want each site to be able to handle your full printing workload, you might need to purchase additional printers for the second site.

Central electronic complexes

CEC capacity and how the CECs will be used is one of the most important factors to consider in planning an end-to-end solution. Specifically, the following attributes of each CEC should be included in your inventory information:

- Processing capacity (millions of service units, or MSU)
- Special purpose processors (System z Integrated Information Processors (zIIPs), System z Application Assist Processors (zAAPs), Integrated Facilities for Linux (IFLs), and so on)
- Memory
- Special features (crypto cards, flash memory, IBM zEnterprise Data Compression (zEDC), Capacity Upgrade on Demand (CUoD) capability, and so on)
- Connectivity capacity and bandwidth

In addition to compiling this information for each CEC, you need to aggregate it for each site. For example, does each site have sufficient IFL capacity to take over the entire Linux workload if one site is down?

Also, if you will be using Fibre Channel Protocol (FCP) channels with z/VM or Linux, do not forget to ensure that the switches and disk subsystems are configured appropriately. For example, your secondary disk subsystems might normally only use FCP ports for mirroring connections. You must ensure that those subsystems are configured to support any z/VM or Linux workload that could potentially run in that site, while at the same time using FCP ports for mirroring back to the disks Site1.

Capacity Backup

The type of site that a CEC is in will determine the amount of configured and configurable capacity that is required on that CEC. Capacity Backup (CBU) can be used to provide cost-effective configurable capacity that is only used for DR testing or in case of a real disaster.

For example, if the site is purely a DR site, little or no central processor (CP) capacity will normally be required. If zGM is being used for asynchronous mirroring, zIIPs can be a cost-effective means of providing the capacity required for the System Data Movers (SDMs). CBU would then be used to provide most of the CPs and other processor types that are required to run the production workload.

CBU enables you to temporarily and concurrently activate additional CPs, Internal Coupling Facilities (ICFs), IFLs, zAAPs, zIIPs, and system assist processors (SAPs). Note that you cannot use the capacity of a CBU upgrade to provide additional capacity for normal workload peaks.

If you will have a SingleSite Workload (SSW) configuration, it is likely that the primary site will have more configured capacity than the backup site. The backup site often has more capacity installed than is being used, depending on which types of work you plan to run in the second site. For example, if production systems span four large IBM zEnterprise EC12 (zEC12)-type CECs, and your development workload running in the backup site requires just one CEC, three CECs in the backup site will normally be idle.

To effectively use multisite data sharing, you should spread test and production sysplexes over both sites and over both sets of CECs. In this case, both sites are likely to have more capacity installed than is actually in use in normal operation. If an event requires all production work to be moved to either site, CBU would be used to enable the additional capacity on the CECs in that site.

Also, regardless of which configuration you run, remember that whenever you upgrade a production CEC, you should make a corresponding change to one of the CECs in the other site. Similarly, if you add features such as zEDC to a production CEC, the CEC that will act as its backup should also be equipped with that feature.

Coupling links

If you currently have a single site configuration and are planning on moving to some form of a multisite sysplex configuration, consideration must be given to the following concepts:

- The CECs in each site will access one or more Coupling Facilities (CFs) in the same site, and one of more CFs in the other site.

  Therefore, each CEC *and* each CF should be configured with InfiniBand (IFB) 12X links (which support a maximum of 150 meters (m)) for connections to the local CECs, and IFB 1X links (which support up to 200 km, depending on the type of DWDM used) for connection to the remote CECs.

- As covered in “CF subchannel considerations” on page 188, increasing the distance between a z/OS system and a connected CF will increase both z/OS subchannel and CEC link buffer usage.

  IFB 1X links, when used on z196 or later CECs, provide relief by supporting up to 32 subchannels and link buffers per channel-path identifier (CHPID). The IFB ability to assign multiple CHPIDs to a single IFB port also provides flexibility and relief. Work with your IBM sales representative to ensure that your CECs will have sufficient IFB links installed to deliver acceptable subchannel usage.

- The performance effect of System-Managed CF Structure Duplexing increases as the distance between the two CFs increases. If you are using this capability today, you should re-assess whether you want to continue doing so in your target multisite sysplex configuration.

  The number of duplexed requests might have an effect on the amount of cross-site IFB bandwidth that you need to provide. Alternatively, you might decide to configure a second CF in each site, and perform the duplexing across two CFs in the same site.

- You also need to consider the need to avoid single points of failure in the cross-site coupling connectivity. Depending on your workload, it is conceivable that a single IFB link might provide adequate bandwidth. However, if that link is lost, all cross-site coupling connectivity (including Server Time Protocol (STP) connectivity) would be lost, resulting in a system, site, or potentially even a sysplex outage.

Important: It is difficult to over-state the importance of ensuring that there are *no* single points of failure in the connectivity between your sites. It is difficult, if not impossible, for a system to differentiate between a connectivity failure and a failure of the component at the other end of a link.

If you are 100% positive that there are no single points of failure in your cross-site connectivity infrastructure, it is safe for your systems to assume that if they can no longer communicate with the devices in the other site, then that site has suffered an outage.

The appropriate response to a connectivity failure would be different than the correct response to a site failure. By providing a configuration where no single event can remove all paths between your sites, your recovery processes can be designed to react to a total loss of communication as a site failure.

Remember that it is not sufficient to provide a configuration with zero physical points of failure. You must also ensure that all of your connectivity (CF links, FICON channels, mirroring links, and so on) is distributed across all of the available inter-site connections, so that the loss of one path does not remove all connections to a given device.

For information about the capabilities, limitations, and distance considerations for IFB links, see the IBM Redbooks publication *Implementing and Managing InfiniBand Coupling Links on System z*, SG24-7539. For information about coupling link type support on the various System z CEC types, see the IBM Redbooks Technical Guide for the relevant CEC.

Note: System z10 was the last System z server generation to support Integrated Cluster Bus-4 (ICB-4) connections. It was also the last System z server generation to support connection to a Sysplex Timer.

IBM zEnterprise 196 (z196) and IBM zEnterprise 114 (z114) are the last System z server generation to offer ordering of InterSystem Channel-3 (ISC-3) connections, and zEC12 is the last generation to support ISC-3 links, even if they are carried forward from a previous generation.

Server Time Protocol

The connectivity considerations for STP are essentially the same for a multisite Coordinated Timing Network (CTN) as for one that is contained in one site. Every CEC should have at least two failure-isolated links to the Preferred Time Server (PTS), and two failure-isolated links to the Backup Time Server (BTS). In addition, the Arbiter must be connected to both the PTS and the BTS.

Although this provides sufficient connectivity for normal operations, you should consider adding coupling or timing-only links between the CECs that do not currently have a special STP role. This gives you the ability to move the PTS, BTS, or Arbiter roles to another CEC during planned outages, and maintain the same level of STP functionality and resilience. You should also plan for the situation where only one site is operating. In that case, you want to have at least a PTS *and* a BTS, and ideally an Arbiter.

In addition to the normal guidelines for configuring STP, there are additional STP considerations that are unique to a multisite sysplex configuration.

The first one is that *you* must ensure that any DWDMs that are part of the connectivity path used by STP must be qualified for STP. STP has no knowledge of the type of DWDMs that are being used. Consider a configuration where each site has two DWDMs that have been qualified for STP, and two that have not. If you have coupling or timing-only links routed through all four DWDMs, there is no way for STP to know if it is currently using a DWDM that is not qualified.

This could potentially result in a data integrity issue. The preferred solution is that *all* of your DWDMs would be qualified for any use that you might make of them. If you are forced to use a configuration where only a subset of DWDMs is qualified for STP, you must ensure that you only configure coupling or timing-only links through the qualified DWDMs[1].

The other consideration relates to failing over from one STP link to another. When STP starts using a coupling link, it communicates with the CEC at the other end of that link and determines the length of the link. This information is then input to STP processing, to ensure that all CECs are synchronized to within the allowable time difference between CECs in a given CTN.

If something were to happen that suddenly changed the length of the link, *and STP was not aware of that change*, there is a risk that, for a small period of time, the synchronization between the CECs might not be within the acceptable tolerances. For this reason, it is critical that any DWDMs that could be used by STP are configured to use client-based protection. This ensures that any link failures are reflected back to STP and it can then handle switching over to a different coupling or timing link.

[1] The same considerations apply to Sysplex Timer connections. If you are still using Sysplex Timers and only a subset of DWDMs are qualified for use with Sysplex Timers, ensure that the timing links are only routed through those DWDMs.

Sysplex Timer (9037)

Because IBM z10 was the last System z server to support connection to a Sysplex Timer, and zEC12 is the last generation to support a mixed CTN, we will not include extension of Sysplex Timer in this document. If you are still using Sysplex Timers, we suggest that you upgrade to an STP-only configuration before implementing a multisite timing network.

Disk

Disks are obviously a critical part of your configuration. Apart from the obvious fact that they contain the programs and data for your applications, they play a pivotal role in the performance your applications will deliver, and they are at the heart of your mirroring configuration. This section provides information about several attributes and perspectives of your disk configuration.

Disk performance

Because applications are becoming increasingly data-intensive, the volume of data that is read and written (and therefore the number of I/Os that are performed) plays a large part in the response time that your applications' users will observe.

Disk performance is a complex topic, especially in an extended distance configuration. "Disk-related considerations" on page 193 provides information about disk performance in detail, particularly the components of disk response time that are sensitive to distance.

Much of the information in that section is more relevant to the post-implementation time frame, where you have some performance data to analyze. In this time frame, you are also in a position to determine the effect of enabling capabilities, such as Modified Indirect Data Address Word (MIDAW) and High Performance FICON for System z (zHPF).

In this section, we just want to remind you that disk performance is something that you must plan for and must be conscious of. However, for now, our focus is on identifying the aspects of your disk configuration that must be considered when you are designing your end-to-end connectivity architecture.

As covered in "Disk-related considerations" on page 193, extending the distance between the host and the disk subsystem, or between the primary and secondary disk subsystems, can increase usage of the channels, switch ports, switch-to-subsystem links, disk subsystem adapters, and potentially components within the disk subsystem.

In general, increasing the distance between the primary and secondary subsystems will have the following effects:

- Read operations from the systems in the same site as the primary disks would be unaffected.
- Read operations from the systems in the remote site would be elongated by roughly 10 µs/km between the primary disks and the z/OS system.
- Write operations from the systems in the same site as the primary disk subsystem would experience longer response times, to the tune of roughly 10 µs/km between the two sites.
- Write operations from the systems in the remote site would have their response times increased by 10 µs/km between the system and the primary disks *plus* 10 µs/km between the primary and secondary disks (a total of 20 µs/km between the sites).

We said "in general" because you should allow for how increased usage of the components in the disk configuration will affect response times. The best way to estimate how this will affect you in advance is to use a tool such as IntelliMagic's Direction (formerly called Disk Magic) and Vision (formerly called RMF Magic) modeling and analysis tools, or to work with your disk vendor.

You might find that additional links or ports will be required. Or you might be able to achieve acceptable performance by reassigning the links and ports that will be used by the various systems.

Host-to-disk operations

Host-to-disk operations are the regular input/outputs (I/Os), issued by applications in Site1 or Site2. There are several technologies that you can use to minimize the effect of the additional distance on the response time of application I/Os:

- PAV and HyperPAV

  Every disk device is represented by a subchannel and an associated unit control block (UCB). Information about the I/O operation is stored in the UCB, meaning that each UCB can only support one I/O at time. If the elapsed time of the I/O is increased because of the increased distance, the UCB will be busier for longer, meaning that subsequent I/Os will have to queue for the previous I/O to complete.

  However, if you had two UCBs for the device, you could initiate a second I/O before the first one completes, eliminating the wait for the first UCB to become available. This is accomplished by using parallel access volumes (PAVs). PAVs allow a single physical disk device to be represented by more than one UCB, enabling multiple I/Os to be in process to the same device at the same time. Because longer distances drive up UCB usage, the use of PAVs can significantly reduce UCB contention in a multisite configuration.

  The original PAV capability was a part of z/OS. However, the more powerful HyperPAV capability might require a chargeable feature in your disk subsystem, so this should be borne in mind when performing your financial planning.

  You can use IntelliMagic's Direction and Vision modeling and analysis tools to estimate the appropriate number of PAV or HyperPav aliases required for your solution. HyperPAV is enabled on a z/OS system using the `IECIOSxx` member of parmlib. For more information about PAV and HyperPAV, see "PAV and HyperPAV" on page 207.

- MIDAW

  The number of interactions between two disk subsystems, or between the CEC and the primary disk subsystem, has an increasingly negative effect as the distance between the two boxes increases, because the elapsed time for every interaction is affected by the distance. Anything that you can do to reduce the number of interactions between the boxes will improve performance.

  One option is to use larger block sizes and more efficient buffering. This enables the same amount of data to be processed with fewer I/O interaction between the channel and the control unit (CU).

  You can also use the MIDAW facility. This was introduced with IBM z9®, and is a method for gathering or scattering data into and from non-contiguous storage locations during the execution of an I/O operation by the channel.

  Channel command words (CCWs) with data chaining were used before MIDAW, but this technique adversely affected FICON channels' performance because of the handshake between the FICON channel and the CU. With MIDAW, you can gather and scatter data from and to the CEC memory using only one CCW. Without MIDAW, for each data scattered in the CEC memory, you need one CCW to read the data from memory and write it to the disk subsystem memory.

  MIDAW is part of z/OS and the System z family of CECs, and is not a separately chargeable option. However, not all disk subsystems support MIDAW, so you should verify with your disk vendor that your subsystem works with MIDAW.

  MIDAW is enabled in the `IECIOSxx` member of parmlib. For more information about MIDAW, see "MIDAW" on page 208.

- The zHPF protocol

  The zHPF protocol is a data transfer protocol that is optionally employed for accessing data from IBM DS8000® storage and other subsystems.

  The zHPF reduces FICON channel activity by using features in the FICON channel, the z/OS operating system, and the CU. These features combine to reduce the number of information units (IUs), and therefore the number of handshakes between the channel and the CU. This normally results in more efficient use of the FICON channel.

  With zHPF, FICON architecture has been streamlined by removing significant system activity in the disk subsystem and the FICON channel. A command block is created to chain commands into significantly fewer IUs. The system activity required to convert individual commands into FICON format is reduced, because multiple System z I/O commands are packaged together and passed directly over the fiber optic link. One single zHPF command block replaces a series of FICON CCWs.

  The zHPF feature builds more efficient CCWs (called *transport control words*, or TCWs), moving more data than a single CCW. When you enable zHPF in your environment, the FC frame payload tends to be bigger than when you are using native FICON, decreasing the number of interactions between the FICON channel and the control unit.

  zHPF is a chargeable feature for FICON Express2, FICON Express4, FICON Express8 and FICON Express8S FICON cards in a z10 system or later CEC. The disk subsystem will also require zHPF support, and that might be a chargeable feature.

  zHPF is enabled on a z/OS system using the IECIOSxx member of parmlib. For more information about zHPF, see the IBM zHPF frequently asked questions (FAQ), available on the following website:

  `http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/FQ127122`

- Extended Distance High Performance FICON

  The zHPF protocol included an additional interaction between the host and the CU at the start of every write I/O. This interaction is not noticeable at short distances. However, at larger distances it can adversely affect response times.

  To improve write performance on zHPF channels at extended distances, the zHPF protocol was modified to remove that additional exchange. This enhancement is included on the IBM DS8700 at the 5.15 driver level, and on the DS8800 with the 6.1 driver level.

  There is no charge for this feature, and it is enabled automatically when the appropriate level of code is available in the disk subsystem.

- FICON dynamic channel-path management

  FICON dynamic channel-path management (FICON DCM) was delivered with z/OS1.11 and IBM z10 CECs. FICON DCM provides the ability for the system to dynamically configure channel paths on and offline to managed CUs under the direction of the MVS Workload Manager (WLM).

  FICON DCM is intended to enable you to configure a control with fewer permanent channels, while still having the ability to add more paths at times when the load on that control unit increases. When DCM chooses a managed channel path to add to a control unit, it takes into consideration the usage of the target port.

  That is, it prefers ports with lower usage levels. After a managed channel path has been assigned, DCM can move the managed channel from one port on the switch to another if it detects high port usage for the previously assigned port.

  Further, FICON DCM uses its knowledge of the internals of the disk subsystems and switches to create connectivity that provides the optimal availability. FICON DCM is included in z/OS and System z servers. It is not a chargeable feature.

For more information about FICON DCM, see the chapter about FICON DCM in *z/OS Version 1 Release12 Implementation*, SG24-7853, and the white paper titled "FICON DCM for System Programmers", available on the following website:

http://www.ibm.com/support/docview.wss?uid=tss1wp101544&aid=1

- IBM System Storage Easy Tier®

  Easy Tier is an optional and no initial charge feature supported on the later IBM DS8000 disk subsystems that offers the following enhanced capabilities:

  - Automated hot spot management and data relocation
  - Auto-rebalancing
  - Manual volume rebalancing and volume migration
  - Rank depopulation
  - Merging of extent pools
  - Thin provisioning support

  IBM System Storage DS8000 Easy Tier is designed to automate data placement throughout the storage system disks pool. It enables the system (automatically and without disruption to applications) to relocate data (at the extent level) across up to three drive tiers. The process is fully automated. Easy Tier also automatically rebalances extents among ranks within the same tier, removing workload skew between ranks, even within homogeneous and single-tier extent pools.

  Easy Tier has two operating modes. Easy Tier manual mode enables a set of manually initiated actions to relocate data among the storage system resources dynamically without any disruption to the host operations.

  In Easy Tier automatic mode, the disk subsystem manages any combination of solid-state drives (SSDs), Fibre Channel (FC) enterprise disks, and Serial Advanced Technology Attachment (SATA) nearline disks. Easy Tier also provides an auto-rebalance capability that adjusts the system to continuously maintain high performance by balancing the load on the ranks within a given tier in an extent pool.

  SSDs help to improve data transfer rate (input/output operations per second, or IOPS) and response times. The use of SSDs, combined with Easy Tier in a DS8000 disk subsystem, can help to maintain critical data used over extended distances to be optimally placed within the disk subsystem.

  Note that Copy Services (Metro Mirror, Global Mirror, and IBM FlashCopy®) are supported with Easy Tier, and Copy Services do not affect the operations of Easy Tier in either Automatic or Manual mode.

  When using Easy Tier in Automatic mode in a Metro Mirror, Global Copy, or combined environment, the workload monitoring on the primary and secondary disk subsystems might differ. Easy Tier on the primary disk sees the normal read and write workload, while in the secondary disk, the workload consists only of writes.

  Therefore, the optimized extent distribution on the primary disk might differ considerably from the secondary disk. The optimized extent reallocation based on workload in the primary is not sent to the secondary.

  In a DR situation, fail over to the secondary disk will mean that the extent distribution of the volumes on the secondary disk subsystem is not optimized to match the primary disk subsystem workload. Easy Tier would need to relearn the production I/O profile, which could take between 12 and 48 hours.

- The z/OS I/O-related parameters

  In addition to controlling the enablement of the features listed here, there are also I/O-related functions in z/OS that must be considered when you make any change that will significantly affect disk response times.

The IECIOSxx member of parmlib enables you to control the following aspects of z/OS I/O management:

- Missing Interrupt Handler

  z/OS monitors the elapsed time of every I/O. No matter how fast or slow a device is, every successful I/O must end after some amount of time. If no response is received from an I/O, that is an indication of a component or connectivity failure that the system must somehow address.

  The amount of time that the system will allow for an I/O to complete is controlled using the **MIHTIME** parameter in the IECIOSxx member. You can specify times for specific classes of devices (such as disk and tape) or for specific devices based on the device address or range of addresses.

  Be sure that you read the entire description of the IECIOSxx member in *z/OS MVS Initialization and Tuning Reference*, SA22-7592. That description contains valuable information about the I/O monitoring and error recovery functions provided by z/OS. It also explains the functions associated with the various keywords in that member, and provides information about which settings can be changed dynamically and which require an initial program load (IPL) to change.

  You should also work with your disk vendor to determine the appropriate setting for their devices, bearing in the mind the distance between your data centers and whether mirroring is being used (and if it is, what type of mirroring).

- I/O Timing

  The I/O timing facility abnormally ends the following I/O requests that have exceeded the I/O timing limits specified for a device:

  - Queued requests waiting for execution
  - Start pending requests
  - Active requests

  The I/O timing facility can be enabled to trigger a HyperSwap. Therefore, it is especially important that the thresholds used by the I/O timing facility are appropriate for the distance between your data centers, to ensure that I/Os are not erroneously ended or that HyperSwaps are not triggered unnecessarily.

- I/O Priority Management

  WLM I/O priority queuing is used to control non-paging direct access storage device (DASD) I/O requests that are queued because the device is busy. You can optionally have the system manage I/O priorities in the sysplex based on service class goals. The default for I/O priority management is NO, which sets I/O priorities equal to dispatching priorities. If you specify YES, workload management sets I/O priorities in the sysplex based on goals. The I/O priority that is set by WLM is used to influence the following factors:

  - The prioritization of I/O requests that are queued on the UCB for that device
  - The use of dynamic PAVs
  - Queuing of I/Os in the channel subsystem (a feature of Intelligent Resource Director)
  - Queuing within the DS8700 or DS8800 disk subsystem if the (optional and chargeable) DS8000 I/O Priority Manager feature is installed

  WLM I/O Priority Management is enabled using the Service Coefficient/Service Definition Options panels in the WLM policy.

When planning your end-to-end configuration, we strongly suggest that you plan for the exploitation of these host technologies if you are not already using them.

Disk-to-disk operations

The other aspect of disk operations that is central to your planning is disk mirroring. After all, the minimal requirement for nearly every multisite configuration is to provide an offsite copy of critical data. However, the considerations are different for disk-based mirroring technologies (Metro Mirror and Global Mirror) than for host-based mirroring (zGM). And they are different for synchronous mirroring (Metro Mirror) than for asynchronous mirroring (Global Mirror and zGM).

The specific tuning and capacity planning considerations are specific to the type of mirroring you use, and will vary from one disk vendor to another. You should work with your vendor to understand the implications for your configuration, and the metrics that should be monitored to provide a warning of potential delays or problems.

The following suggestions are a partial list, but for the full set that is appropriate to your configuration, consult your disk vendor:

- For Global Mirror or zGM, ensure that the primary disk subsystem has sufficient cache to hold all updates until they can be mirrored.
- For zGM, insufficient primary subsystem cache or insufficient bandwidth between the sites can result in device blocking or device pacing, both of which can have a negative effect on application writes to the primary disk subsystem. For more information, see the section about application workload pacing in *z/OS DFSMS Advanced Copy Services*, SC35-0428.
- When using Metro Mirror and HyperSwap, remember to plan for sufficient host channels, switch ports, and disk subsystem interfaces to fully support HyperSwap, meaning that systems could be running in either or both sites with the primary disks in Site1 or in Site2.
- When using Metro Mirror, configure CECs, switches, and disk subsystems to support mirroring in either direction.
- Strongly consider only mirroring in one direction at a time, for the reasons described in 4.1.4, “Considerations for distributed systems” on page 117.
- For any type of mirroring, provide FlashCopy or a similar function in the disk subsystems in both sites. This will give you the ability to create a consistent set of disks before resynchronizing mirroring after a stoppage.
- When using zGM over an extended distance, the Extended Distance FICON feature described in 2.7.3, “Extended Distance FICON” on page 83 can be used to provide improved performance.

  Extended Distance FICON is an enhancement to the industry-standard FICON architecture (FC-SB-3) that can help avoid degradation of performance at extended distances by implementing a new protocol for persistent IU pacing. CUs that use the enhancement to the architecture can increase the pacing count (the number of IUs allowed to be in flight from channel to CU).

  Extended Distance FICON can allow the channel to remember the last pacing update for use on subsequent operations, to help avoid degradation of performance at the start of each new operation. Improved IU pacing can help to improve the usage of the link (for example, it can help keep a 4 Gbps link fully used at 50 km) and support increased distance between servers and CUs.

  Extended Distance FICON can reduce the need for channel extenders in DS8000 series two-site and three-site zGM configurations by allowing an increased number of read commands to be in flight simultaneously.

  It can drastically reduce the total cost of ownership (TCO) of two-site and three-site zGM configurations, and give clients the choice of selecting lower-cost channel extenders built on frame-forwarding technology. The Extended Distance FICON capability is available on the DS8000 series at no additional charge.

Technology extensions for disk

Follow configuration guidelines to ensure the best possible configuration and performance when you order and configure a disk subsystem for use in an extended-distance environment. Each disk vendor will have different configuration recommendations, but supported technologies and performance should remain the primary focus. Create a standard set of requirements and apply them to all disk subsystems in the environment, irrespective of vendor.

Examples of these requirements could be the amount of cache in a subsystem, and the number of usable channel connections. Will the disk subsystem need to support Global Mirror, Metro Mirror, Incremental Resync, zHPF, and other functions? Keep upgrade scenarios in mind for configuration purposes. This can include configuring each disk subsystem with the same amount of base and alias devices on each logical control unit (LCU), to reduce future change outages to make additional storage addressable.

Host adapter allocation on the disk subsystem should be carefully analyzed depending on extended distance requirements. For example, mixing PPRC and zGM with primary host systems connectivity on the same host adapter is not recommended.

Tape

The latest tape technology from IBM supports FICON extended distance connections up to 250 km using a mixture of cascaded FICON directors and supported DWDM equipment. Depending on the application, and the need for a constant tape backup, you can choose to run a single tape library or multiple tape libraries in a Grid configuration.

Some enterprises run with the tape library in the remote location but connected to the primary site through FICON directors and DWDM equipment. With this configuration, the tape infrastructure is already in the DR site if a disaster occurs in the primary site.

Other enterprises with a grid configuration (two or more connected tape subsystems) run the tape infrastructure in `RUN` mode. Therefore, a copy of the tape exists in both locations before the batch job step ends.

For the IBM TS7700, there are two different modes: *immediate* and *deferred*. In immediate mode the distance between the systems will a job time performance. In deferred mode, you need to determine how far out of synchronization is acceptable. For more information about the remote copy options available for TS7700, see the IBM Redbooks document *IBM Virtualization Engine TS7700 with R3.0*, SG24-8122.

The remote copy options provided by the IBM TS7700 are constantly evolving. You should work with your IBM tape specialist to identify the option that best meets your requirements.

For older tape equipment, for example, stand-alone 3490 or 3590 tape drives, you can either extend the channel connectivity if the tape subsystem can support it, or ensure that sufficient tape resources exist in the remote location. For vendors other than IBM, check with the vendor for supported connectivity distances and recommendations.

Consoles

Few installations use real z/OS system consoles any more. Operators generally monitor an automated operations product interface rather than using a real console. Other options for providing a console interface are to use the z/OS HMC System Console or to use a Systems Network Architecture (SNA) console. Additionally, zEC12 and z/OS V2 provide the ability to have a real 3270-style console on the HMC.

Nevertheless, there are situations where a real console is required to control a stand-alone dump, perhaps, or for use with the Disabled Consoles Communication Facility. For this reason, you should ensure that your remote data center provides the same console capabilities as the primary site, and that the consoles are generated and defined to the production systems. The CECs in both sites should be defined to all of your HMCs so that all CECs can be controlled from either site.

Other devices

The following list includes some examples of other devices that might be connected to your z/OS systems:

- Printers
- Check sorters
- Card readers
- Remote job entry (RJE) equipment
- Network (through Open Systems Adapters, or OSAs)
- Communications controller (3745 and similar)
- IBM 2074 Console Support Controller

You should ensure that your plan addresses how each of these would be handled in the case of a loss of either site. Also, some of these devices are obsolete and no longer supported, so you should take this opportunity to replace those devices with current, supported equivalents. Apart from anything else, you need to think about how you would replace those devices if they were lost as part of a site failure. It is better to replace them in a planned manner, than to have to replace them in an emergency when it might not be possible to obtain such devices.

Obsolete devices and technology

Most System z environments now use FICON technology to replace the older ESCON architecture. ESCON is now only used for older technologies, such as tape drives and printers. If you have not upgraded from ESCON, we strongly suggest that you do so.

For older devices that only support ESCON, there is converter technology available that will allow older ESCON devices to run over FICON infrastructure. Remember that the current zEnterprise server, zEC12, no longer supports ESCON. The last generation of System z servers to support ESCON was z196 and z114.

Similarly, FICON Bridge (FICON converter, or FCV) channels are no longer supported as of z196 and z114. In addition, parallel channels (also known as bus-and-tag channels) have not been supported since the z900 generation of CECs.

If you are using devices attached to parallel channels, ESCON channels, or FCV channels, you should be planning to replace those devices with more modern equivalents. If you are unable to replace the devices with ones that support FICON or some other modern attachment, Optica Prizm devices can be used to provide connectivity to FICON channels. For more information about the use of ESCON and parallel channel devices in a FICON environment, see *IBM System z Connectivity Handbook*, SG24-5444.

If you still have older 2 Gbps FICON channels, try to use them for less bandwidth-sensitive devices, such as tape drives, printers, consoles, and protocol converters.

If you are still using Sysplex Timers (device type 9037), you should have a plan for upgrading from them to STP before moving to a multisite sysplex configuration. Sysplex Timer has unique connectivity and support requirements. The last CEC to support direct connection to a Sysplex Timer was z10. There is little value in providing a cross-site connection capability for a device that is non-strategic, and where the connection adapters would not be re-used by some other device.

If you are still using an IBM 3745 Communications Controller, you should be planning to replace it with a newer function, such as the Communications Controller for Linux (CCL). For more information about CCL, see *IBM Communication Controller for Linux on System z V1.2.1 Implementation Guide*, SG24-7223.

5.3.2 Supported distances by device type

All devices, no matter how they connect to System z, have a maximum unrepeated distance over which the device can be connected to the CEC without the use of extended distance equipment. The supported distance will vary by device type. In particular, just because one device supports a given distance does *not* mean that another device will also be supported at that distance.

All devices are connected to System z using channels, coupling links, or some form of adapters. The maximum distance supported by the channel and by the device is determined by the performance requirements, the bandwidth supported by the device, and the communication protocol that the device uses. Table 5-2 shows the maximum supported *unrepeated* distances and link budgets for the various System z channels and other types of connections.

Table 5-2 Fiber optic connections: unrepeated distances

Feature type	Fiber type	Link data rate	Maximum distance[1]	Link budget dB
ESCON (SBCON)	MM 62.5 µm	200 Mbps	2 km 3 km	8
	MM 50 µm		2 km	8
ETR (for Sysplex Timer)	MM 62.5 µm	8 Mbps	3 km	8
	MM 50 µm		2 km	8
FICON long wavelength (LX)	SM 9 micrometers (µm)	1 Gbps	10 km	7.8
		1 Gbps	4 km	4.8
		2 Gbps	10 km	7.8
		2 Gbps	4 km	4.8
		4 Gbps	10 km	7.8
		4 Gbps	4 km	4.8
		8 Gbps	10 km	6.4

Feature type	Fiber type	Link data rate	Maximum distance[1]	Link budget dB
FICON short wavelength (SX)	MM 62.5 µm (OM1)	1 Gbps	300 m	3.00
	MM 50 µm (OM2)		500 m	3.85
	MM 50 µm (OM3)		860 m	4.62
	MM 62.5 µm (OM1)	2 Gbps	150 m	2.10
	MM 50 µm (OM2)		300 m	2.62
	MM 50 µm (OM3)		500 m	3.31
	MM 62.5 µm (OM1)	4 Gbps	70 m	1.78
	MM 50 µm (OM2)		150 m	2.06
	MM 50 µm (OM3)		380 m	2.88
	MM 62.5 µm (OM1)	8 Gbps	21 m	1.58
	MM 50 µm (OM2)		50 m	1.68
	MM 50 µm (OM3)		150 m	2.04
HCA2-O LR (1X IFB) HCA3-O LR (1X IFB)	SM 9 µm	2.5 Gbps 5.0 Gbps[2]	10 km	5.66
HCA2-O (12X IFB) HCA3-O[3] (12X IFB or IFB3)	MM 50 µm (OM3)[4]	3 GBps 6 GBps	150 m	2.06
ISC3 Peer Mode	SM 9 µm	1 Gbps	10 km	7
		2 Gbps		
Gigabit Ethernet LX	SM 9 µm	1 Gbps	5 km	4.6
	MM 62.5 µm		550 m	2.4
	MM 50 µm		550 m	2.4
Gigabit Ethernet SX	MM 62.5 µm	1 Gbps	275 m	2.6
	MM 50 µm		550 m	3.6
10 Gigabit Ethernet LR[5]	SM 9 µm	10 Gbps	10 km	6
10 Gigabit Ethernet SR	MM 62.5 µm	10 Gbps	33 m	2.5
	MM 50 µm		82 m	2.3
			300 m	2.6

[1] Some features might support larger distances when using Request Price Quotations (RPQs)

[2] Auto-negotiated, depending on DWDM equipment.

[3] HCA2-O supports 3 gigabytes per second (GBps) when connecting to HCA1-O on IBM System z9®. HCA3-O does not support connection to HCA1-O.

[4] Requires Multi-fiber Push-On (MPO) connectors.

[5] Does not include OSA-Express2 10 gigabit Ethernet (GbE) Long Reach (LR).

Table 5-3 contains more detailed information about the various types of coupling links.

Table 5-3 Coupling link unrepeated distance and link data rate support

Coupling link type	Maximum unrepeated distance	Link data rate
ICP (Internal)	N/A	Memory-to-Memory
ISC3 Peer mode at 2 Gbps	10 km 12 km[1]	2 Gbps
ISC3 Peer mode at 1 Gbps	10 km 20 km[2]	1 Gbps
ICB-4[3]	7 m, 10 m	2 GBps
PSIFB 12X	150 m	6 GBps or 3 GBps
PSIFB 1X	10 km	5 Gbps or 2.5 Gbps

1. Requires RPQ 8P2263 (System z Extended Distance) or RPQ 8P2340 for the z10 BC
2. Requires RPQ 8P2197. RPQ provides an ISC-3 daughter card that clocks at 1 Gbps in peer mode. This enables the ISC-3 peer mode link to have an unrepeated distance extended up to 20 km. Under certain conditions, RPQ 8P2263 (or RPQ 8P2340 for the z10 BC) might be required in conjunction with RPQ 8P2197. Check with your IBM representative.
3. ICB-4 uses unique 10m copper cables. The maximum distance between each server's raised-floor entry cutout is 7m due to cable routing.

5.3.3 Existing data center considerations

If you already have multiple data centers, you might want to use that investment to help you deliver the capabilities that your company requires, but with a smaller up-front investment.

If the data centers are within metro distance of each other, you need to determine the following information:

- If it is possible to get dedicated fiber between the two sites
- The cable distance for each path between the two sites

  Remember that you need at least two, completely failure-isolated, paths between the sites. It is common to source the fiber from two suppliers, in an attempt to ensure that there are no single points of failure. However, it is not unknown for vendors to share a cable duct, so you should make it clear to your suppliers that failure isolation is an absolute requirement.

  You also need to be careful to determine the precise cable length of each path. In particular, you need to be sure that all paths are within the maximum supported distance. It is not unusual to find differences of 10s of kilometers in the length of two paths between a pair of data centers.

Alternatively, if you are looking for a location to use as an out-of-region disaster recovery site, you might find that your company already has remote data centers that are currently only used for distributed systems. In most modern data centers, the System z equipment only occupies a small percentage of the total floor space. The largest users of power, cooling, and floor space are typically the distributed systems. As a result, it might be possible to add the required System z equipment with only a relatively small effect on the environmentals.

The important point is that you should factor your existing corporate data centers into your calculations, even if they currently do not contain System z equipment.

5.3.4 Available connectivity options

One of the things that you might have no control over is the type of connectivity that you can have between the sites. In some countries, dedicated fiber is not available, regardless of how much you are willing to pay for it.

A more common scenario is that you can lease dedicated fiber, or avail yourself of a managed offering that includes the fiber, possibly the DWDMs that connect to the fiber, and management of the entire connectivity infrastructure. You would need to discuss your requirements with the service providers in your country to determine the available options and the cost of each option.

Another option is that you would pay a provider to lay your own dedicated fiber between your sites. The cost of this option might be prohibitive, depending on the distance, and the amount of disruption that would be caused by the cable-laying operation.

The final option is that you could use a shared network. This might initially appear to be the least expensive option. However, the stringent quality of service (QoS) requirements for a System z configuration are likely to bring the cost closer to the cost of a dedicated fiber solution. Also, remember that coupling links require dedicated fiber between the sites.

The type and cost of connectivity that is available to you is critical in determining the type of service that you can provide. If the only option is to use a shared network, it will not be possible to have a multisite sysplex configuration. If you need to pay for the amount of data that is sent over the connections, financial constraints might limit you to having all work running in the primary site and only use the links for mirroring traffic.

The important point is that you should determine what connectivity options are available to you. The options might preclude the configuration that you were planning for. If the required type of connectivity is *available*, you need to determine its cost.

5.3.5 Time limitations

The timing of the configuration project often determines what you can and cannot do. If you need synchronous off-site storage replication within 60 days, you have to implement one option very quickly, without much detailed study and not much purchasing leverage. If you have two years for the same project, you can be more flexible, create a phased implementation plan, choose the best sites, collect multiple bids, and evaluate different solution offerings.

5.3.6 Financial considerations

The amount of funding that is available to implement your end-to-end connectivity infrastructure is the most important and possibly the most difficult issue to address. The funding can be approached from one of several aspects:

- Government regulations in some countries dictate a maximum amount of application downtime per year. They also require that you have at least two sites, and that you prove every quarter that you can successfully fail over from one site to the other.

  In this situation, the financial question comes down to whether your enterprise wants to continue to operate in that country. If they do, the question becomes “how do we meet (and hopefully exceed) the government requirements, and how much will that cost?”.

- If the budget has already been determined, you must decide if the required objectives can be achieved using the infrastructure that can be put in place based on that level of investment.
- Identify the business value of improved availability and disaster readiness.

  For example, assume that your business loses an average of $10000.00 for each minute the online trading system is not available, and having two sites is expected to reduce the amount of downtime for that application by 10 hours a year. So the benefit of having two sites to the trading system is 10 hours/year * 60 min./hour * $10000.00 cost/min., or $6,000,000.00 per year.

The work you have completed so far in this chapter should help you determine the cost for the proposed solution. The following list includes the likely major cost items:

- The cost of the physical links between the sites.

  Depending on your favored solution, this could be the cost of procuring and laying the cables, or it might be the cost of paying for a managed solution.
- DWDMs to connect the sites. You will need a minimum of two devices in each site to avoid single points of failure.
- Directors or director upgrades to add inter-switch link (ISL) ports, Fibre Channel over IP (FCIP) support, buffer credits, and so on.
- Any additional CEC capacity that might be required in the second (or third) site. The use of CBU should be investigated to help you minimize costs.

  You might be able to reduce costs by providing fewer CECs in the second site, with the understanding that less capacity might be required for test and development systems in case of a real disaster.
- Any additional storage devices, such as secondary disks, additional tape subsystems, and so on.

  Similar to the CECs, you might be able to control costs by providing secondary devices that have (for example) larger or slower hard disk drives (HDDs), on the basis that reduced levels of performance might be acceptable in case of a real disaster.
- Additional features on the CECs, switches, or disk subsystems that optimize performance over longer distances (compression, for example).
- Other peripheral devices, such as additional printers, network devices, encryption devices, and so on.
- Do not forget to budget for the time of the technical staff that will be required to take on such a complex project.
- You might also want to budget for the use of outside experts, for example people with DWDM expertise.

5.4 Creating a balanced end-to-end configuration

Anyone with experience with performance tuning will know that good, and consistent, performance is the result of creating and maintaining a balanced configuration. The System z server design architecture, shown in Figure 5-1, balances all of the core resources necessary for the workload:

- Uni-processor (or engine) size
- N-way effect (number of available engines)
- Memory size
- I/O back-end and front-end speed and throughput

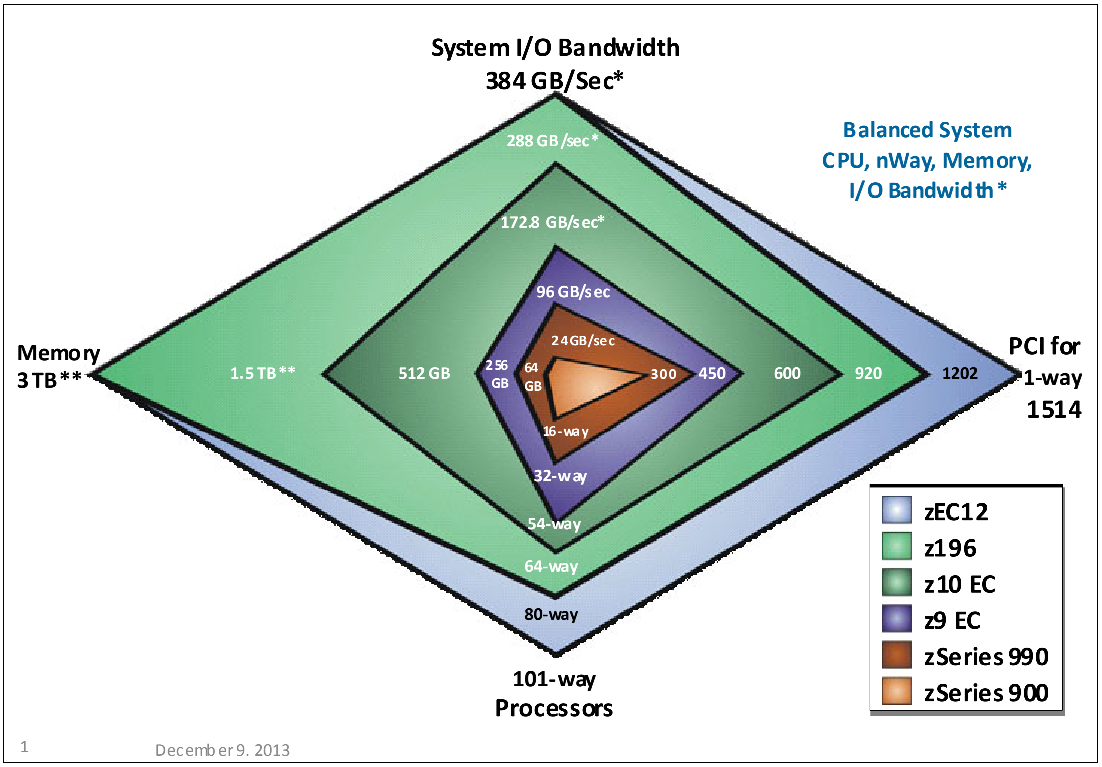


Figure 5-1 System z balanced system design

An architecture that creates or exposes an imbalance could effectively create a weakness in the overall design, where one of these core requirements becomes a performance bottleneck. The end-to-end connectivity infrastructure needs a similar balanced design approach for optimal performance and functionality throughout the infrastructure.

Core weaknesses might be introduced, resulting in individual, back-level, components negating the benefits of investments that refresh other parts of the end-to-end solution. Figure 5-2 on page 149 shows link and network bandwidth components in the end-to-end connectivity infrastructure.

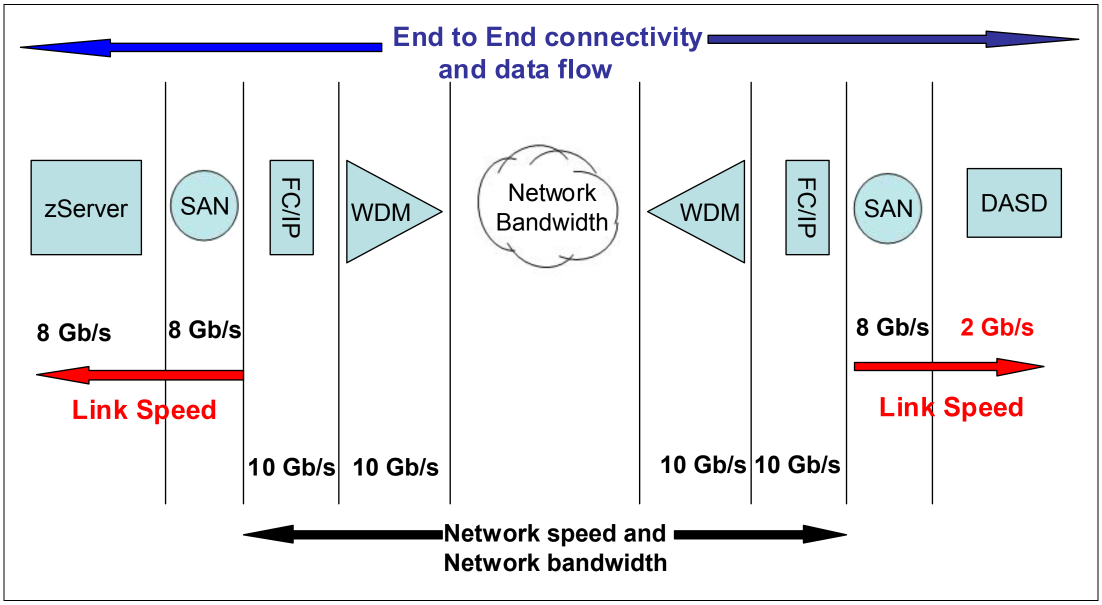


Figure 5-2 Balanced connectivity design considerations

You need to consider the complete end-to-end configuration when you change components in the connectivity design. For example, in Figure 5-2:

- The server-to-director link is optimal, with matching 8 Gbps optics.
- A 10 Gbps network pipe from the local director all the way through to the remote director. In this example, the network bandwidth is assumed to provide sufficient throughput to deliver optimal performance from the disk subsystem.
- Director-to-disk link is *sub*-optimal. This link will auto-negotiate to the weakest point of 2 Gbps at the disk subsystem device adapter, which might cause performance bottlenecks back to the server.

Maintaining a balanced configuration requires an investment of both time and money to ensure that all connectivity components function correctly and are optimally aligned. However, the performance and functionality might also be satisfactory in an imbalanced (albeit unstressed) configuration. In this case, there must be an assessment to validate the effect of this compromise in the end-to-end architecture.

Every time you change or upgrade a component, you need to ask “What does this change mean to my end-to-end connectivity infrastructure?” Consider the following guidelines in answering the question:

Buffer credits An increase in link speed will require buffer credits to keep the link full, so check that each point in the connectivity infrastructure has sufficient credits.

Cabling Adding links will require new cables. Spare capacity needs to be available in the trunk when trunking is being used. Cable types must be optimal to match the features being deployed; links might be compromised where sub-optimal cables are used.

Technology enablement New features might require changes in other areas of the connectivity infrastructure. For example, a new server using zHPF will require enabling functions in the storage subsystem. Part of the process for enabling the new function should include checking for any related requirements throughout the end-to-end infrastructure.

Auto-negotiation Links will auto-negotiate to the weakest point in the infrastructure. Whenever you upgrade components, review the end-to-end link connectivity for gaps or mismatches, and ensure that port settings are optimally set for end to-end link performance where necessary (for example, on switches and directors).

Qualification Check changes to the end-to-end infrastructure against qualification letters, where appropriate, to ensure that any components that you plan to use are qualified.

5.5 Security considerations

The data center is usually a highly secure area with limited access. However, connecting data centers over a distance can pose security risks, because the data has to leave the secure area of your building. The problem is shown in Figure 5-3.

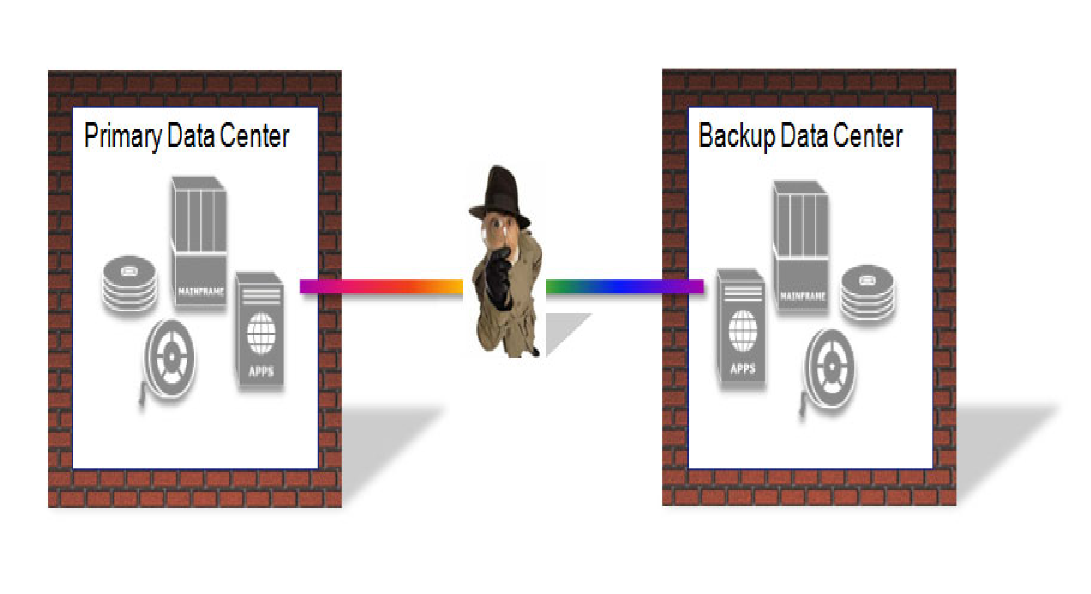


Figure 5-3 Secure data centers but insecure transmission

Risks to your data while it is being transported include:

- Tapping of the optical fiber with a tapping clamp
- Accessing the signals at optical amplifier stations
- Having a built-in optical splitter in a part of the fiber that is outside your control

Tapping the fiber can be detected because tapping devices add attenuation to the fiber that can be detected. Some DWDM manufacturers offer special features to generate an alarm if there are changes in fiber attenuation. If the fiber tapping device is connected into the fibers before the DWDM is connected, there is no way to determine its presence.

If you connect data centers, the preferred practice for dark fiber-based networks is to use fibers that consist of a single (unsplit) fiber all the way through from one data center to the other. However, if your data centers are more than 100 km apart, you will have to amplify the signals in the middle, as shown in Figure 5-4.

To amplify or regenerate signals, the whole fiber has to be routed through an amplification regenerating station, where a WDM device is used for amplification or regenerating. This could be done in a building or some other container near the fiber route. Those amplification stations must be secured in the same way as your data centers, because of the potential access to your data that they provide. If your data is transported by a service provider, your agreement should ensure the security of the fibers that will carry your data.

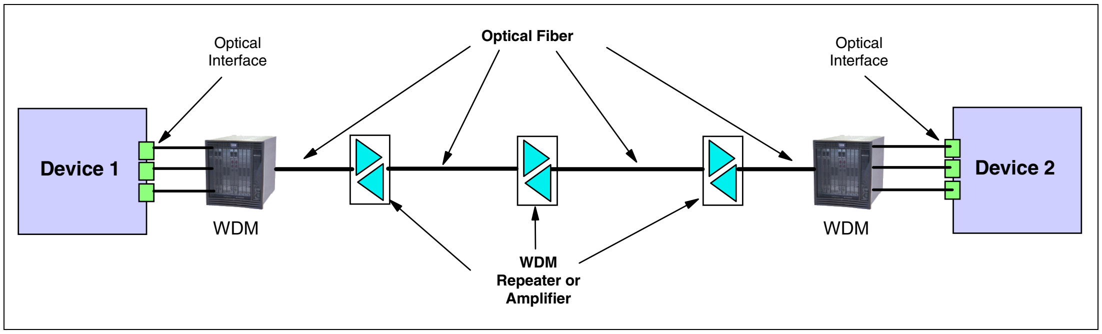


Figure 5-4 Long fiber span with inline amplification

However, considering the increasing public and governmental interest in data privacy, it is likely that transmitting data in the clear outside your premises will at some point be deemed to be unacceptable. Therefore, when designing your end-to-end infrastructure, you should plan on encrypting the data that will be transmitted between the two sites.

There are numerous options for encrypting your data, and new options are constantly being added. The following sections provide information about some of the options available at the time of writing, however you should work with your corporate data privacy department and security experts to identify the most appropriate option to meet your needs.

Host-based data encryption

There are several IBM products that provide the ability to encrypt data in z/OS before it is transmitted on the channel. IBM InfoSphere® Guardium® Data Encryption for IBM DB2® and IBM IMS™ Databases provides encryption for DB2 for IBM z/OS and IMS data systems. It uses IBM System z cryptographic hardware to protect sensitive data at the DB2 row level and IMS segment level. For more information about encryption for DB2, see the IBM Redbooks document *Security Functions of IBM DB2 10 for z/OS*, SG24-7959.

For sequential disk and tape-based data sets, the Encryption Facility for z/OS provides the ability to encrypt and decrypt the data in z/OS, meaning that the data is not sent in clear to the channel. For more information about the Encryption Facility for z/OS, see the IBM Redbooks document *Encryption Facility for z/OS Version 1.10*, SG24-7318.

However, both of these solutions are aimed at specific data types. They do not encrypt data sent to the CF, nor do they encrypt *all* of the data sent to the channel.

Device-level encryption

Both the IBM TS7700 Virtualization Engine and the DS8000 disk subsystem support device-level encryption. That is, they provide the ability to encrypt data before it is written to the media, and decrypt it before it is sent back to the host.

Both of these methods secure your data storage media. However, they do not address the issue of data being transmitted in the clear between the two sites.

In-flight encryption

The other option is to use in-flight encryption. That is, the data is encrypted after it leaves the CEC, but before it is transmitted outside the data center, and then decrypted when it reaches the other data center. There are several options for this type of encryption:

- A stand alone encryption system that is hooked in before the signal leaves the data center. You need one system for each signal that you want to encrypt.
- Some directors provide an encryption ability, so that all data transmitted on an ISL or FCIP port is encrypted.

  Work with your switch vendor to understand the performance effect (increased latency) on each I/O of encrypting and decrypting each packet.

  This is an attractive option because *all of the data that is transmitted by the director will be encrypted*, regardless of whether it is destined for a tape, disk, or any other type of device. It is also independent of data types (whether the data is going to a sequential data set, Virtual Storage Access Method (VSAM), DB2, IMS, or otherwise makes no difference).
- The other option is to perform encryption and decryption in the DWDM.

  As with director-based encryption, performing encryption in the DWDM is likely to add some amount of latency to all requests that are passed between the two sites. Alternatively, DWDM-based encryption has the advantage that *all* of the data passed between the DWDMs will be encrypted, including data on coupling links.

> **Important:** Encryption devices or encryption-enhanced devices might add latency to your link. Consider this especially for links whose flow control is based on buffer credits.

Regardless of where the encryption is performed, there will be a financial cost for the device or the software product that performs the encryption. To determine the optimal solution for your environment, you need to consider financial cost, performance effect, the type of encryption that is supported, and whether you have a need for encryption of data that is not transmitted between the sites (on local disk subsystems, for example).

Also, remember that you should not encrypt data that has already been encrypted. For example, if data is encrypted before it leaves the CEC, it should not be encrypted again before it leaves the building.

5.6 IBM qualification for extended distance devices

We now provide information about how to use the IBM qualification process to ensure that your planned environment will be fully supported.

The following list includes components in an end-to-end solution:

- System z CEC
- Storage area network (SAN) Switches and Directors
- Wavelength division multiplexers (WDMs)
- IBM storage
- Non-IBM storage
- Other platforms
- Network service providers

You can find machine documentation, education material, qualification status of switches and WDMs, fixes, tools, and other information about System z on the Resource Link website. Use the following link to access the site:

https://www.ibm.com/servers/resourcelink/

5.6.1 System z CEC

It is valuable to monitor the driver exception letters for System z CECs, because they contain up-to-date information about interoperability considerations for System z CECs.

Driver exception letter

The Driver exception letters are available on the Resource Link website:

1. Select the **Fixes** link on the left side of the home page.
2. On the resulting page, select **Exception Letters** in the Hardware section. The resulting window lists all of the IBM System z CEC types.
3. Select the ones that correspond to your CECs. You will then be presented with a list of the Driver levels that are available for that CEC.

There are two types of exception letter for each machine type and driver level: A Service letter and a Client letter. Both letters are intended for the IBM System z service representative and for the client. The letters contain information that is specific to that driver level:

- Functional exceptions
- Code restrictions
- General information
- Specific service information

5.6.2 WDM

The qualification letters for every qualified WDM are available on the Resource Link website. To access the letters, follow these steps:

1. Select **Library** on the home page (on the left side of the page)
2. On the resulting page, in the Hardware products for servers section, click **System z Qualified Wavelength Division Multiplexer (WDM) products for GDPS solutions**.
3. The resulting page lists the vendors that have qualified WDM solutions. Select the vendors that you are interested in.
4. The resulting page lists the models and firmware levels that are qualified. The qualification letters are listed on the right side of the page. Select the model and firmware level that you are interested in.

Unfortunately, there is no single page that shows all qualified WDMs and the specific features that are qualified on each one in a matrix format. You need to retrieve and read the qualification letter for each model that you are interested in.

The beginning of a typical DWDM qualification letter is shown in Example 5-1. This shows the specific model and firmware level that the letter relates to.

Example 5-1 Qualified product and software version

```
IBM GDPS and Server Time Protocol (STP) Application Qualification support for the
ADVA FSP3000* Dense Wavelength Division Multiplexer (DWDM) Platform running
software release 11.2.3
```

The next section, shown in Example 5-2, shows the list of System z servers that are qualified for use with this DWDM.

Example 5-2 Qualified System z servers

```
IBM Parallel Sysplex and Geographically Dispersed Parallel Sysplex(GDPS), IBM
zEnterprise EC12 (zEC12), IBM zEnterprise BC12 (zBC12), IBM zEnterprise 196
(z196), IBM zEnterprise 114 (z114), IBM zEnterprise BladeCenter Extension (zBX),
IBM System z10 (z10 EC, z10 BC), and IBM System z9 (z9 EC, z9 BC) environments.
```

However, you need to continue reading. Just because a server is listed in the section shown in Example 5-2 does not necessarily mean that all protocols and channel types are supported on every server. The next section down, shown in Example 5-3, lists the supported solutions and protocols that are included in this qualification.

Example 5-3 Qualified solutions and protocols

```
GDPS / Peer-to-Peer Remote Copy (PPRC) (Metro Mirror) using the following
protocols:
- High Performance FICON for System z (zHPF) & FICON for Storage Access
- FCP for disk mirroring
- 1x InfiniBand (1x IFB) or ISC-3** peer mode for exchanging Server Time Protocol
(STP) messages to provide synchronization of servers
- ISC-3 for coupling facility (CF) messaging
- GDPS / Extended Remote Copy (XRC) (z/OS Global Mirror) using zHPF & FICON for
asynchronous remote copy
- zBX intraensemble data network (IEDN) over 10 Gigabit Ethernet (10 GbE)
```

Coupling links: If a qualification letter mentions ISC-3 or 1x-IFB, both STP and CF messaging are supported.

The next letter section contains a table showing the specific part numbers with a detailed list of the protocols, speeds, and the distance that each was qualified with. An extract from the table is shown in Table 5-4.

Table 5-4 Qualified part number, protocols, and distance

Adva P/N descriptor	Description	Supported protocols	Supported distance
5TCE	5-port 10G TDM module: 2:1 5G InfiniBand (1x IFB DDR) 4:1 ISC-3 Peer Mode 3:1 4G FCP/ISL 1:1 8G FCP/ISL 1:1 10G ISL 1:1 10 GbE	1x IFB 5 Gbps (DDR), ISC-3 Peer Mode, 4, 8 Gbps FCP1/ ISL, 10 Gbps ISL, 10 GbE	100 km

Notice that each vendor has a specific supported configuration, and the letters can be different. For a complete sample letter, see Appendix B, “Sample qualification letters” on page 209.

DWDM qualification letters provide detailed information. Review each letter carefully to ensure that the protocols that you require and are considering for the future are qualified. For example, some WDM devices are qualified for FCP connection for mirroring, but not for CEC-to-storage subsystem connection. Ensure that the device you select supports the protocols that you use today, as well as ones that you expect to use in the future.

Even if a specific module or transponder (the terminology is vendor-dependent) supports a specific distance, this does not mean that your application will be able to perform as expected at that distance. For each environment, the performance requirements for the applications will dictate the maximum achievable distance. Testing is the only way to determine the maximum distance for *your* configuration.

“It depends”: We often get questions from clients asking “Over what distance can I perform active/active sysplex data sharing?” Unfortunately, the only correct answer is “it depends”. The effect of distance on application performance will vary from one configuration to another, from one application to another, and even from one transaction to another.

IBM and the qualified vendors can tell you the maximum distance that is *supported* for a given configuration. But the only way to determine the performance effect is to benchmark that configuration. Your company is the only one that can determine if the resulting response times are acceptable to you. For example, in one enterprise, a doubling of response times might be acceptable, but for another, such an increase might be completely unacceptable.

5.6.3 SAN switches and directors

The qualification status for switches and directors is also found in qualification letters on the Resource Link website. To access the letters, follow these steps:

1. Select **Library** on the home page (on the left side of the page).
2. On the resulting page, in the Hardware products for servers section, click **Switches and directors qualified for IBM System z FICON and FCP channels**.
3. The resulting page contains generic information about the switch qualification process, and the protocols and devices that are included in the testing. The qualification letters are all listed down the right side of the page. Select the model and firmware level that you are interested in.

Switch qualification letters are more detailed and complex than the corresponding WDM letters. The switch letter has tables that cross-reference each other. Table 5-5 shows a summary of sections in a switch qualification letter.

Table 5-5 Switch qualification letter summary

Table number/name	Description
1) Switches and Directors	Contains all of the switch and director models supported with specific code levels, specific SFP optics, and speeds.
2) System z servers	The supported System z servers with specific driver levels.

Table number/name	Description
3) System z functions and features	Specific protocols (such as FICON and FCP) and features (such as zHPF). Also contain information about specific operating systems.
4) I/O devices	Specific storage controllers and protocol converters.
5) Maximum supported distance for non-repeated and non-amplified switch/director optics	Lists all the different SFP and speeds and the maximum distance supported. This table shows only LX SFPs but SX SFPs are also supported.
6) Supported distance extension	Lists the supported extended distances and the supported protocols and technologies.
7) Supported software	Lists all supported operating systems and products.
8) FCIP configuration	Lists all FCIP configurations supported by this letter.

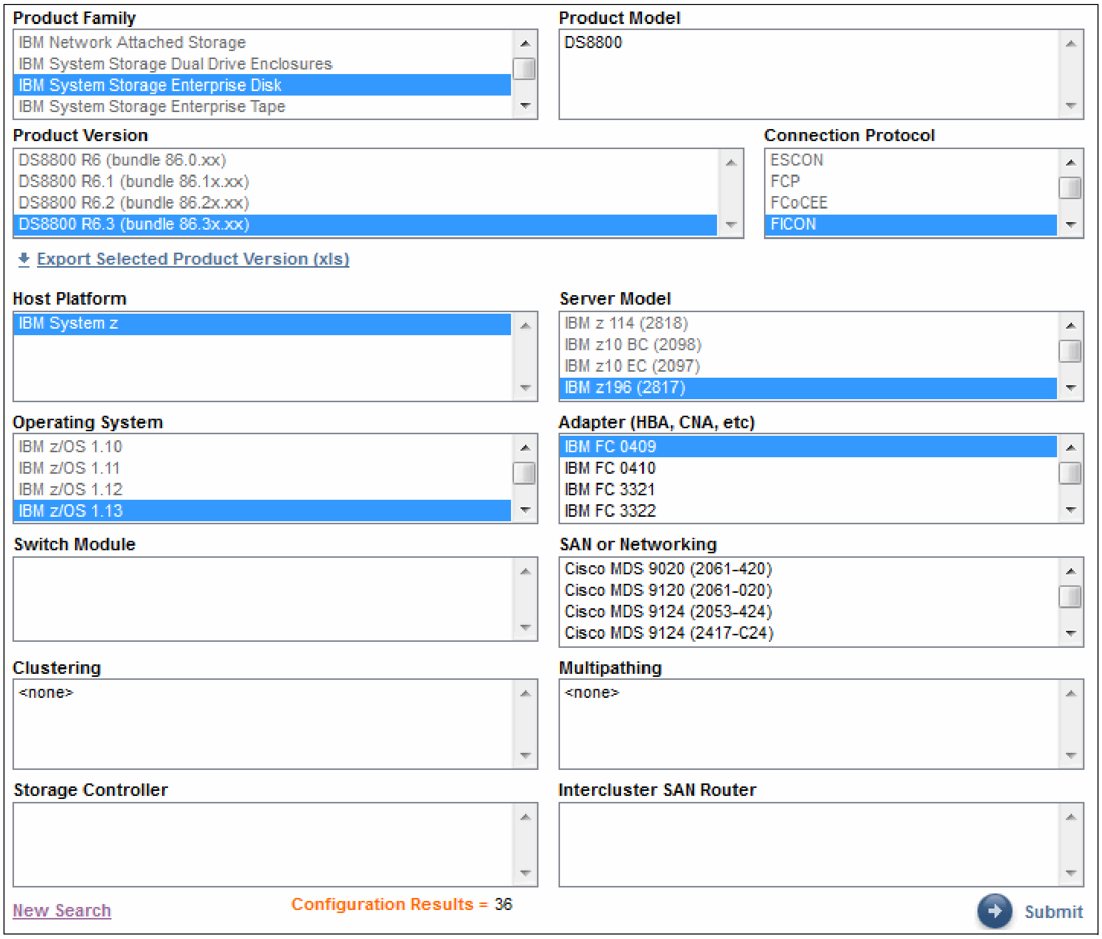
See Appendix B, “Sample qualification letters” on page 209 for a complete sample switch qualification letter.

5.6.4 IBM storage

The IBM System Storage Interoperation Center (SSIC) website (`http://www.ibm.com/systems/support/storage/ssic/interoperability.wss`) is a powerful interactive tool for building and checking specific storage-related configurations. You select a component for a specific solution, and the tool provides a list of the interoperable components.

After you narrow the options down to fewer than 100 results, as shown in Figure 5-5 on page 157, you can submit your selections and receive a report of the supported components (for example, for a specific storage controller running a specific code level).

You can also export the results to a spreadsheet that will have one workbook for each platform and protocol. For example, if you select a DS8800 for System z, the tool will show a link for the *Copy Services Fibre Channel Extension Support Matrix* on the *IBM System z FICON* workbook.



Product Family
IBM Network Attached Storage
IBM System Storage Dual Drive Enclosures
IBM System Storage Enterprise Disk
IBM System Storage Enterprise Tape

Product Model
DS8800

Product Version
DS8800 R6 (bundle 86.0.xx)
DS8800 R6.1 (bundle 86.1x.xx)
DS8800 R6.2 (bundle 86.2x.xx)
DS8800 R6.3 (bundle 86.3x.xx)

Connection Protocol
ESCON
FCP
FCoCEE
FICON

Host Platform
IBM System z

Server Model
IBM z 114 (2818)
IBM z10 BC (2098)
IBM z10 EC (2097)
IBM z196 (2817)

Operating System
IBM z/OS 1.10
IBM z/OS 1.11
IBM z/OS 1.12
IBM z/OS 1.13

Adapter (HBA, CNA, etc)
IBM FC 0409
IBM FC 0410
IBM FC 3321
IBM FC 3322

Switch Module

SAN or Networking
Cisco MDS 9020 (2061-420)
Cisco MDS 9120 (2061-020)
Cisco MDS 9124 (2053-424)
Cisco MDS 9124 (2417-C24)

Clustering
<none>

Multipathing
<none>

Storage Controller

Intercluster SAN Router

Configuration Results = 36

Figure 5-5 SSIC component selection sample

An example of an SSIC report generated by the selections in Figure 5-5 is included in “System Storage Interoperation Center” on page 219.

Example 5-4 shows a sample report generated after you submit your selections.

Example 5-4 Example of components selection

```
Configuration                  Name
Product Family:                IBM System Storage Enterprise Disk
Product Version:               DS8800 R6.3 (bundle 86.3x.xx)
Connectivity:                  FICON
Host Platform:                 IBM System z
Server Model:                  IBM z196 (2817)
Operating System:              IBM z/OS 1.13
Adapter (HBA, CNA, and so on): IBM FC 0409
Product Model:                 DS8800

Result SAN or Networking
0001 Cisco MDS 9020 (2061-420)
Show details | Hide details
```

A list of supported SAN products, switches, and directors is shown after the selected components. Click **Show Details** to view detailed information about that device.

5.7 Physical connectivity considerations

To ensure a problem-free implementation and reliable service, the connectivity architecture group should consider the following questions:

- What devices need to be connected, and what speed are the connections (for example, 2 Gbps, 4 Gbps, 8 Gbps, and so on)?
- What will be the cable lengths?
- What are my link budgets, and are my links within specifications?
- Do I require attenuation at any point in my end-to-end infrastructure?
- Do I require SX or LX transceivers on my servers?
- Do the transceivers match at both ends of each physical connection?
- Are my cables the correct type, and do I require trunking?

5.7.1 Optical data flow

Figure 5-6 shows the data flow from the server through the physical and optical infrastructure, to the transceiver on the next component in the end-to-end connectivity infrastructure. This is a generic description, and is intended to highlight all of the physical connectivity items that you must consider when planning and implementing your end-to-end architecture:

1. Data flows from the server through the feature on the server (specific to the architecture/protocol) to the transceiver, which will use an light-emitting diode (LED) or a laser (SX or LX).

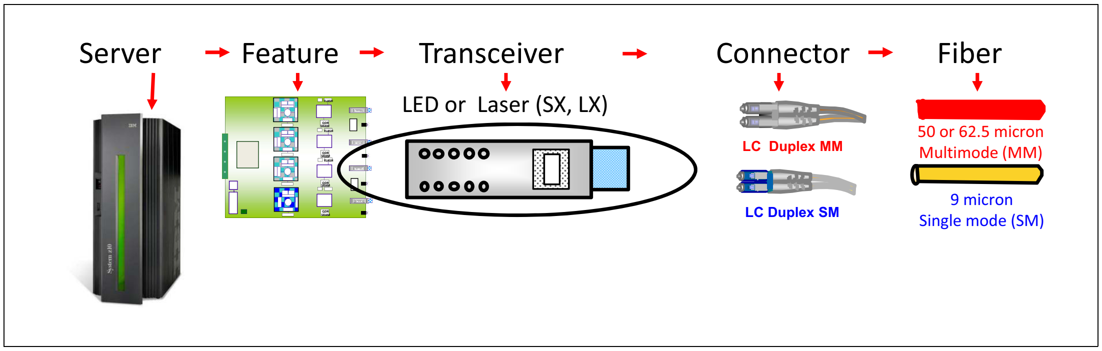


Figure 5-6 Data flow through optical and physical infrastructure

2. Flow continues through the cable connector. Depending on the type of link, this can be one of the following connectors:

   – ESCON Duplex
   – Mechanical Transfer Registered Jack (MTRJ)
   – SC Duplex
   – LC Duplex

   The default connector for all FICON channels is LC Duplex unless you explicitly specify otherwise.

   You can find more information about the various types of connectors in the section about connector types for fiber cables in *IBM System z Connectivity Handbook*, SG24-5444.

3. The flow continues to the fiber optic cable. The cable might be physically attached through multiple patched connections (see Figure 5-7). These connections will be one of the following types:
   - Multimode fiber (MM), which could be either 50 or 62.5 micron. The numbers 50 and 62.5 microns describe the diameter of a cable's core. Core diameter influences fiber optic cable performance. When MM fiber is used, the light source can vary depending upon the optical transceiver (transmits and receives light) on the feature, which can be either a light-emitting diode (LED) or an SX laser.
   - Single-mode fiber (SM), which is 9 micron. When SM fiber is used, the light source is an LX laser optical transceiver.

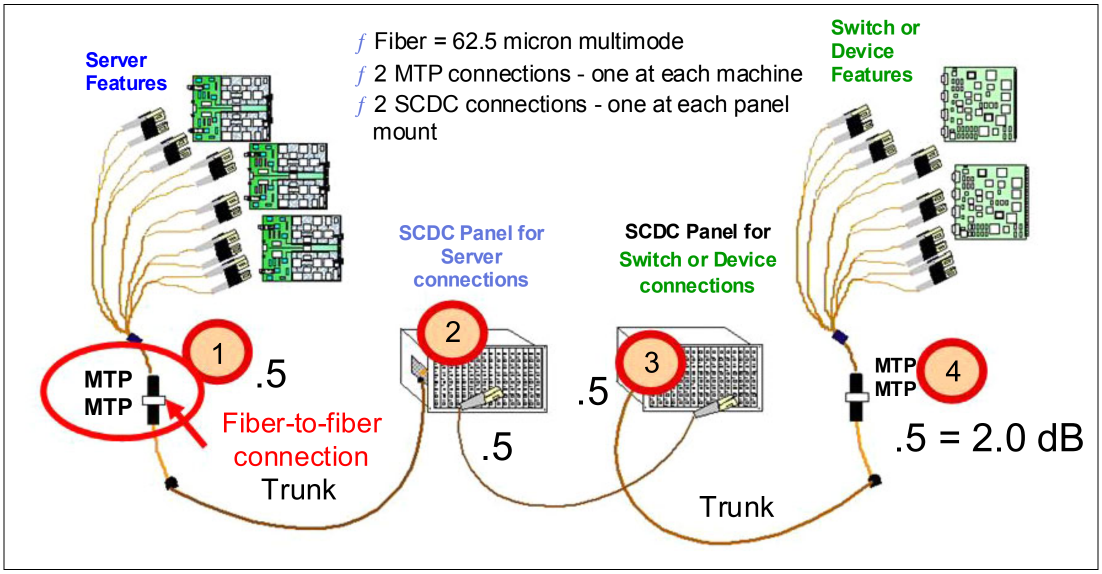

Server Features
ƒ Fiber = 62.5 micron multimode
ƒ 2 MTP connections - one at each machine
ƒ 2 SCDC connections - one at each panel mount
Switch or Device Features
SCDC Panel for Server connections
SCDC Panel for Switch or Device connections
1
2
3
4
MTP
MTP
.5
Fiber-to-fiber connection
Trunk
.5
.5
MTP
MTP
.5 = 2.0 dB
Trunk

Figure 5-7 Multiple connections and the effect on link budget

4. The data flow continues to the next transceiver in the end-to-end connectivity solution. This transceiver must match the type coming from the server (LX or SX). This will be attached to a feature on the next device:
   - The device adapter on the target device
   - The port on a SAN switch
   - The transponder on extended distance equipment (DWDM or other)

For detailed planning considerations, see *Planning for Fiber Optic Links (ESCON, FICON, InfiniBand, Coupling Links, and Open System Adapters)*, GA23-0367. Also, for preferred practice information regarding planning for and managing your cabling infrastructure, consult the documents contained on the IBM *Cabling Considerations in Storage Area Networks* support website:

http://www.ibm.com/support/docview.wss?uid=ssg1S1004299

5.7.2 Physical layer switching

Physical layer switches for computer communication have been used for a long time. For example, the IBM 2914 bus-and-tag crosspoint switch was available in the 1970s. A physical layer switch is often called an electronic patch panel, because a microprocessor controls which ports talk to each other, and the microprocessor is controlled from a local or remote console, command-line interface (CLI), or graphical user interface (GUI). A physical layer switch enables you to perform the following functions:

- Change which cables are connected to each other without having to open and reconnect any cable connections.

  This is an important attribute, because every time a cable is opened, there is a chance that dirt will be introduced into the connection, causing problems when the link is re-enabled, or (even worse) causing a problem at some time in the future.

- Make changes quickly.
- Make connections, even unplanned ones, easily.
- Easily return to a previous configuration that worked if you find a problem in a new configuration.

Figure 5-8 shows a physical layer switch connecting the processor at the upper left (1) with a storage device at the lower left (A), and a processor in the upper right (2) with a storage device in the lower right (B). The connections are made in the physical layer switch shown in the center of the diagram.

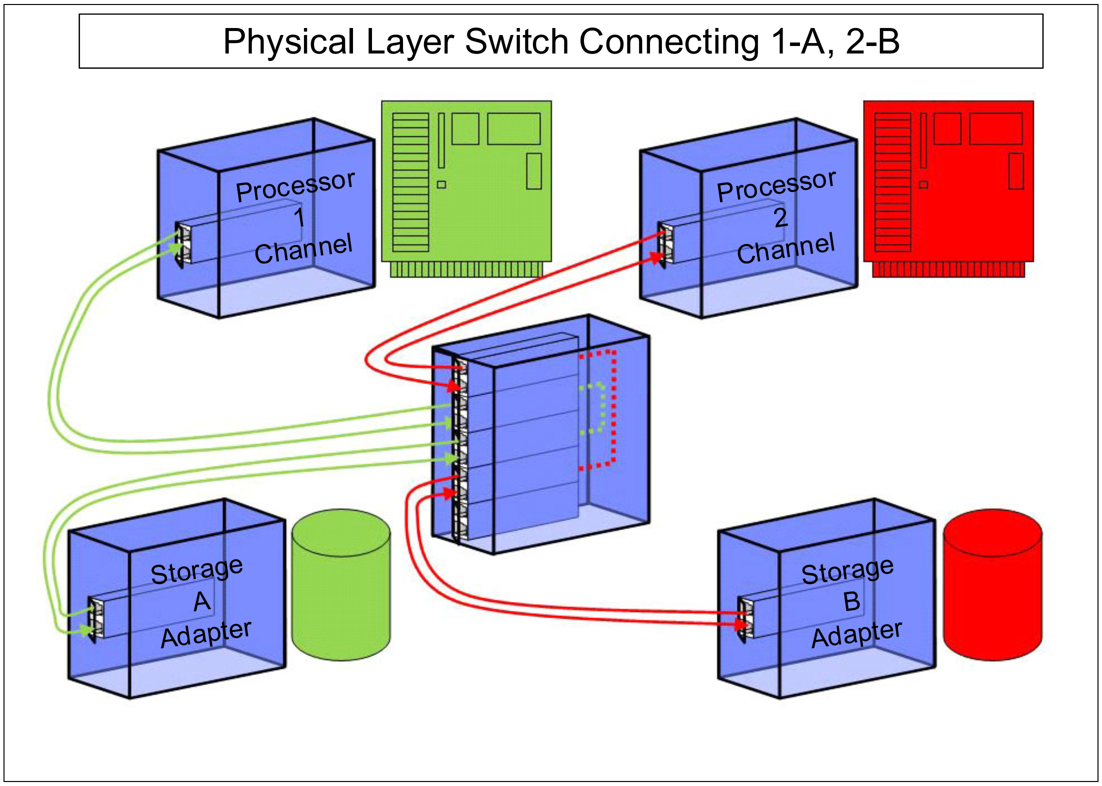


Figure 5-8 Physical Layer Switch Connecting 1-A, 2-B

Figure 5-9 shows how the physical layer switch can change the configuration without having to move any cables. You can see that the cables are all still plugged into the same ports, but the upper-left processor (1) is now connected to the lower-right storage device (B).

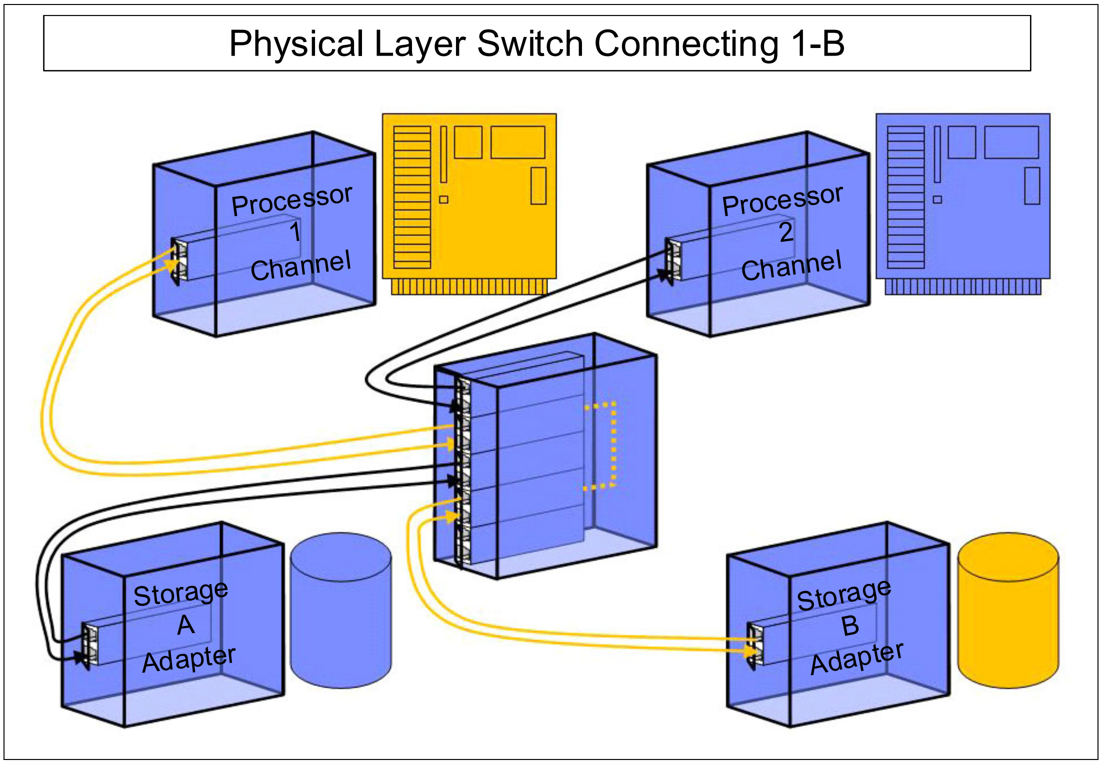


Figure 5-9 Physical Layer Switch Connecting 1-B

Even though this change is disruptive to the I/O paths, it can be completed quickly using a CLI or a GUI from a remote terminal. Also, because there is no physical recabling, there is no need to verify the physical connection.

For more information about the physical layer and how it relates to the various protocols and devices that use it, see Appendix C, “Physical layer information” on page 221.

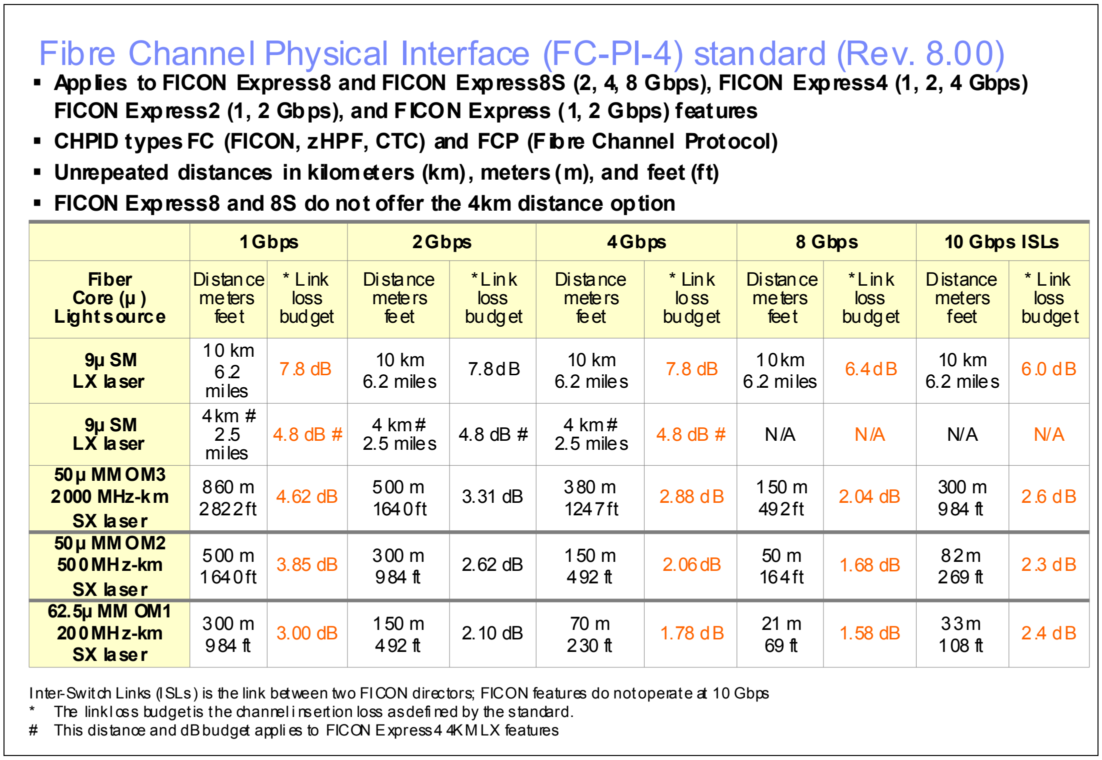
5.7.3 Link specifications and considerations

A link is a physical connection over a transmission medium (a fiber, for example) used between an optical transmitter and an optical receiver. The maximum allowable link loss, or link budget, is the maximum amount of link attenuation, or loss of light (expressed in decibels (dB)), that can occur without causing a possible failure condition or bit errors. The unrepeated distance and link budget both decrease as the link data rate increases when using MM fiber.

Physical cabling considerations

The link budget is derived by combining the channel insertion loss budget with the deallocated link margin budget. For details about link specifications and budgets, see *S/390 Fiber Optic Link (ESCON, FICON, Coupling Links and OSA) Maintenance Information,* SY27-2597.

Figure 5-10 shows the current link specifications for System z FICON and FCP channel features. This figure reflects the budget difference between various features, and also the difference between SM and MM features. For example, FICON Express8 and Express8S features have a tolerance of 6.4 dB for SM and 1.58 dB for MM.



Fibre Channel Physical Interface (FC-PI-4) standard (Rev. 8.00)

- **Applies to FICON Express8 and FICON Express8S (2, 4, 8 Gbps), FICON Express4 (1, 2, 4 Gbps) FICON Express2 (1, 2 Gbps), and FICON Express (1, 2 Gbps) features**
- **CHPID types FC (FICON, zHPF, CTC) and FCP (Fibre Channel Protocol)**
- **Unrepeated distances in kilometers (km), meters (m), and feet (ft)**
- **FICON Express8 and 8S do not offer the 4km distance option**

	1 Gbps		**2 Gbps**		**4 Gbps**		**8 Gbps**		**10 Gbps ISLs**	
Fiber Core (µ) Light source	Distance meters feet	* Link loss budget	Distance meters feet	* Link loss budget	Distance meters feet	* Link loss budget	Distance meters feet	* Link loss budget	Distance meters feet	* Link loss budget
9µ SM LX laser	10 km 6.2 miles	7.8 dB	10 km 6.2 miles	7.8 dB	10 km 6.2 miles	7.8 dB	10 km 6.2 miles	6.4 dB	10 km 6.2 miles	6.0 dB
9µ SM LX laser	4 km # 2.5 miles	4.8 dB #	4 km # 2.5 miles	4.8 dB #	4 km # 2.5 miles	4.8 dB #	N/A	N/A	N/A	N/A
50µ MM OM3 2000 MHz-km SX laser	860 m 2822 ft	4.62 dB	500 m 1640 ft	3.31 dB	380 m 1247 ft	2.88 dB	150 m 492 ft	2.04 dB	300 m 984 ft	2.6 dB
50µ MM OM2 500 MHz-km SX laser	500 m 1640 ft	3.85 dB	300 m 984 ft	2.62 dB	150 m 492 ft	2.06 dB	50 m 164 ft	1.68 dB	82 m 269 ft	2.3 dB
62.5µ MM OM1 200 MHz-km SX laser	300 m 984 ft	3.00 dB	150 m 492 ft	2.10 dB	70 m 230 ft	1.78 dB	21 m 69 ft	1.58 dB	33 m 108 ft	2.4 dB

Inter-Switch Links (ISLs) is the link between two FICON directors; FICON features do not operate at 10 Gbps

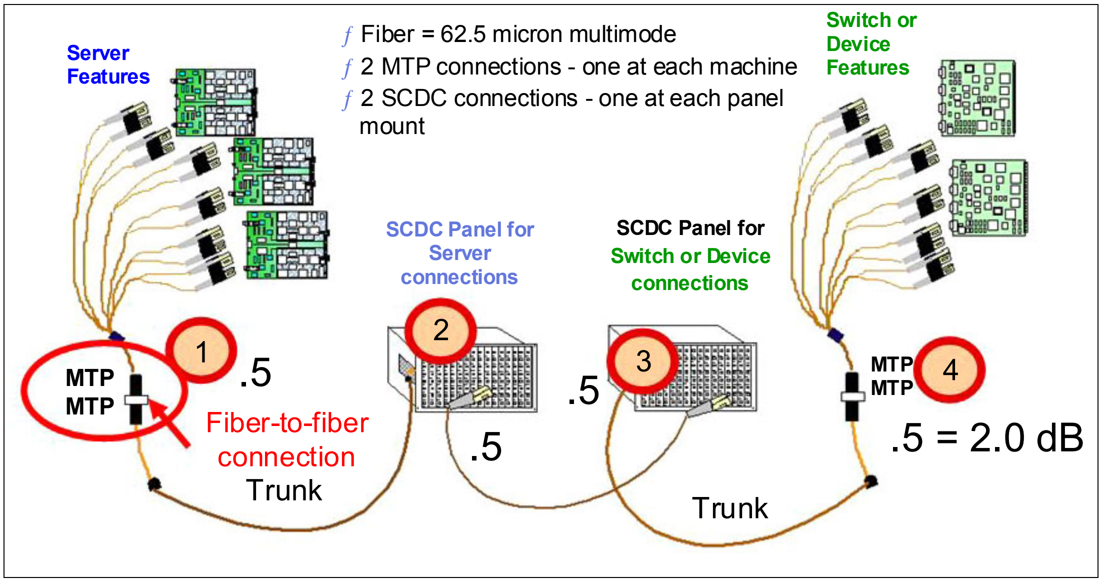
* The link loss budget is the channel insertion loss as defined by the standard.

# This distance and dB budget applies to FICON Express4 4KM LX features

Figure 5-10 FICON and FC link budgets

There can be multiple physical connections in the end-to-end cable connection to join each pair of optical connections. These can be a combination of patch panel and Multi-fiber Termination Push-on (MTP) connections.

The effect of multiple connections is shown in Figure 5-11. Each connection can have an effect of up to 0.5 dB on the available link budget. In a 62.5 micron MM cabling for an 8 Gbps FICON connection, the link loss budget of 1.58 Db would have been exceeded, resulting in link errors or interface control checks.



Server Features
ƒ Fiber = 62.5 micron multimode
ƒ 2 MTP connections - one at each machine
ƒ 2 SCDC connections - one at each panel mount
Switch or Device Features
SCDC Panel for Server connections
SCDC Panel for Switch or Device connections
1
.5
MTP
MTP
Fiber-to-fiber connection
Trunk
2
.5
.5
3
MTP
MTP
4
.5 = 2.0 dB
Trunk

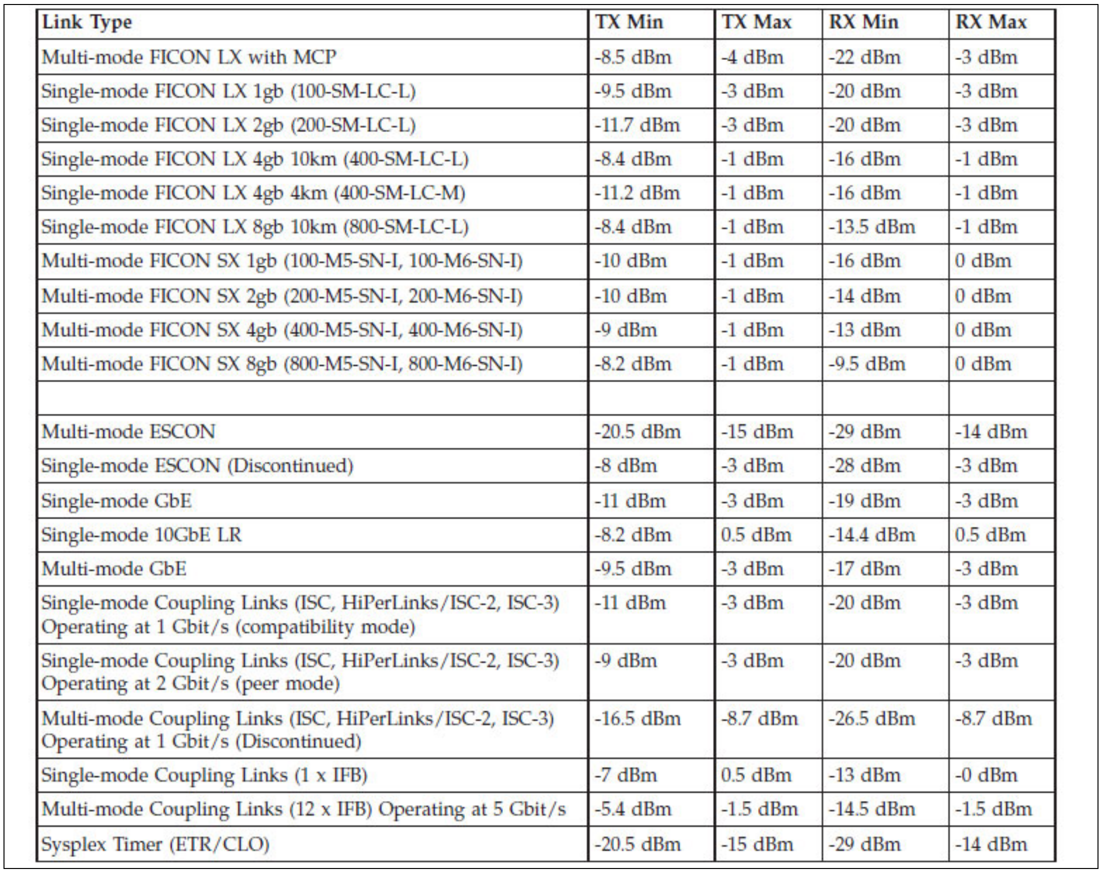
Figure 5-11 Effect of multiple connection on link budget

Attenuation considerations

Transmit and receive signals that are outside the link specification can cause the link to not initialize or come online, possibly due to the light signal being too strong at either the transmitter or receiver. This might be a result of an optical transceiver in the end-to-end configuration providing too much amplification. An example is when channel extension equipment uses extended distance optics but is connected at a short distance.

Fiber Optic Attenuators are components installed in the fiber optic transmission system to reduce the power of the optical signal. An attenuator is effectively the opposite of an amplifier, though the two work by different methods. An amplifier provides gain, but an attenuator provides loss. This reduces the amplitude or power of a signal without appreciably distorting its waveform. You can use different attenuator levels to bring the link back into specification when the light is too strong.

Figure 5-12 shows the various link *transmit* and *receive* specifications. For more details, see *S/390 Fiber Optic Link (ESCON, FICON, Coupling Links and OSA) Maintenance Information,* SY27-2597.



Link Type	TX Min	TX Max	RX Min	RX Max
Multi-mode FICON LX with MCP	-8.5 dBm	-4 dBm	-22 dBm	-3 dBm
Single-mode FICON LX 1gb (100-SM-LC-L)	-9.5 dBm	-3 dBm	-20 dBm	-3 dBm
Single-mode FICON LX 2gb (200-SM-LC-L)	-11.7 dBm	-3 dBm	-20 dBm	-3 dBm
Single-mode FICON LX 4gb 10km (400-SM-LC-L)	-8.4 dBm	-1 dBm	-16 dBm	-1 dBm
Single-mode FICON LX 4gb 4km (400-SM-LC-M)	-11.2 dBm	-1 dBm	-16 dBm	-1 dBm
Single-mode FICON LX 8gb 10km (800-SM-LC-L)	-8.4 dBm	-1 dBm	-13.5 dBm	-1 dBm
Multi-mode FICON SX 1gb (100-M5-SN-I, 100-M6-SN-I)	-10 dBm	-1 dBm	-16 dBm	0 dBm
Multi-mode FICON SX 2gb (200-M5-SN-I, 200-M6-SN-I)	-10 dBm	-1 dBm	-14 dBm	0 dBm
Multi-mode FICON SX 4gb (400-M5-SN-I, 400-M6-SN-I)	-9 dBm	-1 dBm	-13 dBm	0 dBm
Multi-mode FICON SX 8gb (800-M5-SN-I, 800-M6-SN-I)	-8.2 dBm	-1 dBm	-9.5 dBm	0 dBm
Multi-mode ESCON	-20.5 dBm	-15 dBm	-29 dBm	-14 dBm
Single-mode ESCON (Discontinued)	-8 dBm	-3 dBm	-28 dBm	-3 dBm
Single-mode GbE	-11 dBm	-3 dBm	-19 dBm	-3 dBm
Single-mode 10GbE LR	-8.2 dBm	0.5 dBm	-14.4 dBm	0.5 dBm
Multi-mode GbE	-9.5 dBm	-3 dBm	-17 dBm	-3 dBm
Single-mode Coupling Links (ISC, HiPerLinks/ISC-2, ISC-3) Operating at 1 Gbit/s (compatibility mode)	-11 dBm	-3 dBm	-20 dBm	-3 dBm
Single-mode Coupling Links (ISC, HiPerLinks/ISC-2, ISC-3) Operating at 2 Gbit/s (peer mode)	-9 dBm	-3 dBm	-20 dBm	-3 dBm
Multi-mode Coupling Links (ISC, HiPerLinks/ISC-2, ISC-3) Operating at 1 Gbit/s (Discontinued)	-16.5 dBm	-8.7 dBm	-26.5 dBm	-8.7 dBm
Single-mode Coupling Links (1 x IFB)	-7 dBm	0.5 dBm	-13 dBm	-0 dBm
Multi-mode Coupling Links (12 x IFB) Operating at 5 Gbit/s	-5.4 dBm	-1.5 dBm	-14.5 dBm	-1.5 dBm
Sysplex Timer (ETR/CLO)	-20.5 dBm	-15 dBm	-29 dBm	-14 dBm

Figure 5-12 Minimum and maximum transmit and receive levels

5.7.4 Considerations for fiber routes with different lengths

Generally speaking, all active paths between a pair of sites should be similar in speed and latency. In a single data center, latency due to different cable lengths is usually irrelevant compared to disk access times or network latency. However, latency due to distance can become a larger factor as the distance between sites increases.

For availability reasons, you should always have at least two, completely failure-isolated, fiber connections between your sites. Further, each path should provide sufficient capacity to handle your entire workload in the event of the other path being unavailable. You must have a second path to allow for unforeseen incidents (someone digging up your cable, for example), and also to allow for planned outages for maintenance, upgrades, and so on.

Unless the distance between your sites is very short, it is likely that there will be a difference of at least hundreds of meters, if not multiple kilometers, in the length of the two paths. Depending on how large the difference is, and how sensitive to the latency introduced by the distance your applications are, you might have to perform some planning to cater for the different latencies of the two paths.

Light travels through fiber at about 200,000 km/sec with a delay of 5 µs/km. It is common to use 10 µs/km as a gauge, because this is the cumulative round-trip-time over fiber for each send-receive transfer.

This 10 µs/km can become a major consideration in connecting remote data centers, potentially making the difference between whether CF requests are handled synchronously or asynchronously.

We assume that you have already indicated to your fiber providers that you want the shortest possible paths between the two sites. If you have done that and are still concerned about the latency difference between the two paths, there are some actions that you can take. To verify the latency of each path, most directors and DWDMs provide facilities to measure the time to send a test signal up and down the link.

Additionally, the DWDMs might have the ability to raise an alert if the length of the link changes at any time. Remember that fiber providers might have the ability to temporarily re-route a link if they need to perform maintenance or make repairs, so this ability to detect changes in the link length is an important capability.

There is nothing further that you can do to give the longer path the same latency as the shorter path. You cannot make light travel faster. You *could* add fiber to increase the length of the shorter path to make it equal to the longer path. However we believe that most installations would not want to do that, and in most cases, it should not be necessary.

For most IT departments, consistency is more important than achieving the lowest possible response time for *some* transactions. And, if you think about it, most transactions and certainly most batch jobs consist of multiple I/O requests and multiple CF requests. So your objective should be to create a configuration where I/O and CF requests are balanced across the available paths, so that each job and each transaction uses both the shorter and the longer path.

For availability reasons, we assume that you already plan to distribute the paths from every host to every remote device across the available directors and DWDMs. Further, in cases where a single link can be used by multiple LPARs, you need to ensure both that the physical links are distributed across the inter-site paths, and that the logical paths used by each LPAR are distributed across the available paths.

Beyond that, balancing the load across the paths requires an understanding of FICON and coupling link path selection algorithms.

You should use tools, such as the Resource Management Facility (RMF) Channel Path Activity report, plus any reports that might be provided by your DWDMs, to understand the distribution of traffic across the available paths. Note that the RMF CF subchannel activity report only reports accumulated activity across the full set of CHPIDs that are used to connect a CF. It does not provide a breakdown of usage at the individual CHPID level[1].

Also, in relation to CF activity, remember that multiple sysplexes might be using the same IFB link, but the RMF CF report only reports the activity for one sysplex. To get a view of the activity at the physical link level, you need to combine the reports for all sysplexes that share the links.

STP considerations

To ensure that all CECs in the CTN remain within the required time synchronization of each other, STP takes into account the length of the link that is used to connect the Current Time Server (CTS) to every other CEC in the CTN. If the length were to change without STP being aware of the change, there is the potential of a data integrity exposure.

[1] If you are operating on a zEC12 or later, the RMF Channel Path Details section in the RMF CF Subchannel Activity report shows the length (to the closest tenth of a kilometer) of the path used by each coupling CHPID.

For this reason, you *must* use “client-based protection” for DWDMs that are used for System z STP connections. This is covered in more detail in 3.7.2, “Protection schemes” on page 108. Client-based protection means that if there is any event that would cause the DWDM to switch from one path to the alternative one, the connected devices will be notified of the change.

This enables them to perform normal channel recovery. In the case of STP, STP will select another path. Additionally, as a safeguard, STP checks the length of each link on a regular basis.

5.7.5 Connectivity infrastructure preferred practices

As with many things in IT, there are a small number of situations where there is only one correct way to do something. But there are many situations with a huge gray area, where something will not fail if you do it a particular way, but it will not perform as efficiently or as reliably as though you did it a different way.

These are the situations where *preferred practices* are valuable. Preferred practices are suggestions based on experiences in a particular area. If the preferred practice is not adhered to, the product or service might still function. However, we believe that the best way to manage or set up or perform a function is described in a related preferred practice. This section contains some preferred practices related to extended distance configurations, based on experiences and observations in this area.

Cable maintenance and cleaning

One of the most time-consuming and frustrating problems to track down can be dirt in a fiber optic connection. Dirt that is invisible to the eye can cause intermittent faults and bit errors. Figure 5-13 shows two fiber optic cables, viewed under a microscope. Both appeared to be clean to the naked eye, however when magnified you can see the amount of dirt on the fiber on the left. How many times have you seen someone “cleaning” a fiber cable by rubbing it on their shirt?

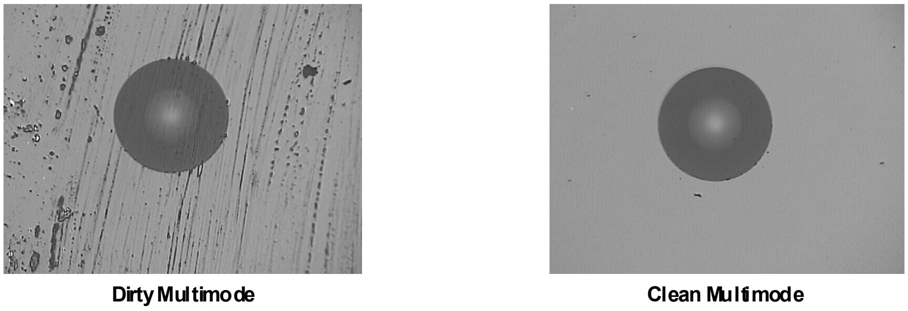


Figure 5-13 Cable cleanliness

The problem with dirty fiber optics increases in line with the speed of the adapter the cable is connected to. Many clients found that fiber optic infrastructures that had worked with no problems with 2 Gbps or 4 Gbps adapters suddenly started experiencing problems when the adapter was upgraded to 8 Gbps.

Even worse, the vibrations of someone walking past can cause a piece of dirt to move into the path of the light, meaning that you can suddenly start experiencing problems with a link that has not been opened for months.

The best way to avoid these issues is to purchase a fiber optic cleaning device and clean both the cable end and the receptacle every time you open a connection. The extra few seconds it will take to clean the cable are insignificant compared to the hours of time and service disruption that can be caused by having to track down a dirty connection some time in the future.

For more information about maintaining and cleaning your fiber optic links, see the following website:

http://www.ibm.com/support/docview.wss?uid=ssg1S1004299&aid=1

Make all switch domain IDs unique

Although the only technical requirement for switch domain IDs is that they be unique within a *fabric*, making all switch domain IDs unique in an *environment* will make it easier to identify switches. This is especially useful when evaluating FICON link addresses, because the first byte of the two-byte link address is the hex equivalent of the domain ID (the Switch Address that you define in HCD must be the hex equivalent of the Domain ID). Using unique domain IDs means that the switch can immediately be determined from just the link address.

Although domain IDs are configured on switches in decimal, it appears in the FC address as hex. The first byte of the link address matches the first byte of the FC address.

Domain IDs must be in the range 1 - 239. In very large environments where the switch count exceeds 239, try breaking the switches into smaller logical environments.

Make the switch ID in HCD the hex equivalent of the switch domain ID

For FICON, the switch ID that you specify in HCD is effectively a comment. Setting the switch ID to the hex equivalent to the domain ID (and therefore equal to the Switch Address) will make it easier to match the switch ID to the switch.

Include the hex equivalent of the domain ID in the switch name

Consistent with the advice “Make All Domain IDs Unique” and “Make the Switch ID in HCD the hex equivalent of the switch domain ID”, including the hex equivalent of the domain ID in the switch name that you define on the switch will make it easier to associate switches with link addresses.

Always use two-byte link addresses

Previous FICON switches required a license to support the security features required for two-byte addressing. As a result, many clients used single-byte addressing when two-byte addressing was not required. Today, there is no additional license for this capability. Always using two-byte link addressing avoids any confusion about which switch a control unit is connected to.

Keep tape and disk on separate ISLs

Because of their different use profiles, it is best to keep tape and disk traffic on separate ISLs.

Plan for future changes

Depending on the capabilities of your switch, there are some changes that might be disruptive. For example, if you want to implement logical switches, the first time you switch to using a logical switch might require an outage, but adding more logical switches in the future might be nondisruptive.

Work with your vendor to identify all configuration changes that require an outage, and attempt to configure the box now in a manner that will avoid planned outages in the future.

Disable unused switch ports

Remember that z/OS will only use devices that are defined to it. Even if it has physical connectivity to a device that it does not know about, it will not use that device.

Not all operating systems behave in this manner. The norm for distributed platforms is to perform automatic discovery. That is, they will go out and look to see what devices they have connectivity to. Worse than that, some of them might even go and write on every device they can access.

To protect your z/OS devices from accidentally being accessed by some other system, it is suggested to disable any switch ports that are not currently in use. This might marginally lengthen the process of bringing a new link into use, but it ensures that if someone connects a cable to the wrong port by accident, your devices will not be exposed.

Zoning

Unless you need to use smaller, more granular, zones for security reasons, create one large zone with all FICON ports in it. Port zoning, not worldwide name (WWN) zoning, should be used for FICON. Zones should never have a mix of WWN zone members and port zone members. If it is necessary to mix FICON and FCP on the same fabric, a zone configuration can have a mix of WWN zones and port zones, but each zone should have either WWN *or* port members but not both.

The CUP, FE, is a logical address on the switch that is reachable even when it is not in a zone, so there is no need to put the CUP in a zone. The CUP can manage and gather RMF 74-7 records for all ports regardless of how zoning is configured.

Because only node ports matter in zones, there is no need to include extension ports (E_ports) in zones. Furthermore, a registered state change notification (RSCN) is sent whenever a port changes state. An RSCN requires a frame, but RSCN frames are discarded by FICON fabric ports (F_ports). Not putting E_ports in a zone eliminates sending these frames to ports where they are not needed.

All FCP channels for Linux on System z and z/VM should use WWN zoning. Only include the channel WWNs and the WWNs for the targets the channel needs access to. When Node Port Identification Virtualization (NPIV) is in use, it is the virtual WWN associated with the virtual server and not the base channel WWN that belongs in the zone. The least significant byte of the base channel WWN when NPIV is in use is 00.

Disk and some tape mirroring use FCP even if the data is from z/OS or IBM z/Transaction Processing Facility (z/TPF), so WWN zoning should be used for all mirroring ports regardless of what operating system is associated with the data.

Segregate disparate environments into different logical fabrics

Today's SAN fabric switches can be virtualized into multiple logical switches. ISLs can still be shared, but this capability might require a software license. Consult your switch vendor before planning and configuring different virtual fabrics to share ISLs.

Disk or tape mirroring uses FCP whether mirroring FICON or open systems data. It is usually best to put the ports used for mirroring in their own logical fabric; however, some clients using RMF to analyze performance prefer to put these ports in the same logical fabric as the FICON traffic so as not to have to expend a CHPID just for this purpose. If the mirroring ports will be put in the same logical fabric as the FICON ports, the FICON and FCP port for mirroring should be in separate zones.

Additional information

For more information about preferred practices for your extended distance configuration, consult your director and DWDM vendors. They will have extensive experience in this area and can offer valuable guidance.

Also, see the following IBM Redbooks documents:

- *Fabric Resiliency Best Practices*, REDP-4722
- *FICON Planning and Implementation Guide*, SG24-6497
- *IBM System z Connectivity Handbook*, SG24-5444
- *IBM SAN Survival Guide*, SG24-6143
- *I/O Configuration Using z/OS HCD and HCM*, SG24-7804

5.8 Selecting your extended distance equipment

There are many connectivity options available, ranging from installing your own dark fiber, to using a managed solution from a service provider. When you were compiling your device inventory, one of the pieces of information you should have included was the number and type of interfaces on each device.

This information will help you identify the number, speed, and type of interfaces on your switches. It will also help you estimate the required bandwidth between the sites. For example, “I need ten 1 Gbps Ethernet, twenty 8 Gbps FICON, two 10 Gbps Ethernet, two 5 Gbps InfiniBand” and so on.

With this information in hand, you need to consider the following things when discussing your options with vendors:

- Is the configuration that they are proposing qualified? The qualification letters are detailed and specific, so work with the vendor to ensure that not just the devices and the firmware levels, but also all feature codes, conform to the qualification letter.
- If your configuration will consist of both directors/switches and WDMs, ensure that the devices have been qualified to work with each other. If purchasing a WDM, ensure that it fully supports any proprietary protocols that might be used by your directors.
- Ensure that the proposed devices support all of the features that you might require, such as compression, encryption, virtual switches, CUP, FICON Acceleration, and so on.
- Obtain mean time between failures (MTBF) values from the vendor, and factor that information into your decision. Remember that director-class devices are designed to deliver higher levels of availability than switches.
- Get information about the latency that will be injected by the device, plus any additional latency that might result from the use of compression, encryption, and so on.
- Understand the ability of the device to support dynamic changes. What types of changes require an outage? Can firmware or hardware features be added dynamically? Do firmware upgrades require an outage? Do the devices in both sites need to be at the same firmware level (meaning that both must be down at the same time for an upgrade)?
- Ensure that the recovery processing that the device uses is appropriate for a System z configuration. For example, failing over from one ISL to a different one without the host being made aware of the failover is not supported.
- Does the switch/director have sufficient buffer credits for your current and planned configuration? For example, if you are using 4 Gbps interfaces today, your buffer credit requirement will increase when you move to 8 Gbps ports.

- What management tools are supported? Does the device support an interface that provides performance information? Configuration information? Availability information? Is it possible to inspect the error logs, particularly those related to any link failures? Does the device provide diagnostic tools? Is the operator interface intuitive? Does it communicate with industry-standard system management tools using Simple Network Management Protocol (SNMP)?
- Is the device compatible with your current SAN configuration? Is it extensible, so that it can support future configuration changes (for example, moving from 8 Gbps to 16 Gbps connections at some time)?
- Does the device have spare capacity? You do not want to purchase a device that does not provide room for growth without an expensive upgrade or replacement.
- What redundancy does the device provide? Or, looking at it from the other perspective, are there any single points of failure in the design of the device?
- What are the power, cooling, and floor space requirements?
- Does the vendor have experience and reference sites that are similar to your proposed configuration?
- What are the maintenance costs?
- What are the acquisition costs? Do not forget to include the costs of all features plus any management software.

5.9 Benchmarking your proposed configuration

No matter how you approach this, implementing a second or third data center is a significant investment that you will have to live with for many years. Therefore, before making the final decision to proceed, especially if your plans include a multisite sysplex, it is critical that you benchmark your proposed configuration to ensure that it can deliver the required levels of performance and throughput.

Increasing the distance between a CEC and the devices it is connected to will affect the service time for requests to those devices. The relationship between distance, service times, and batch and transaction response times is a complex one.

The complexity is compounded by the secondary effects of increased service times. Without going into all the detail here, consider the following scenario:

- An existing sysplex is ported to a multisite configuration, with half the systems, half the CFs, and half the workload moved to a new data center 20 km from the existing one.
- Because of the greater distance, log writes during a transaction (T1) take longer. Because the transaction cannot release serialization on the resources it was using and complete until the log records have been saved to disk, all locks held by the transaction are held for longer.
- Another transaction (T2) that will require those resources starts running and obtains serialization on some other resources. Transaction T2 then tries to serialize the resource held by transaction T1, but has to wait until T1 releases its locks.
- Transaction T3 starts running, but is immediately delayed because it cannot serialize a resource that T2 is holding.

In this scenario, transaction T3 cannot run until transaction T2 completes and releases its locks. And transaction T2 cannot complete until transaction T1 completes and releases *its* locks. So, the increased elapsed time for T1 has a knock-on effect on T2 and T3.

But consider a similar scenario, except that this time T2 is not run. Because T3 requires resources that T2 would have used, but not the resources that T1 was using, T3 would not be delayed in this case.

You can see that the elapsed time for each transaction is affected, both by the change in service times for I/Os and CF requests, and by the relationship between the transactions and the arrival pattern of transactions.

Appendix A, “Performance considerations” on page 179 provides valuable information to help you quantify the *direct* effect of increased distances on service time, and things that you can do to minimize that effect.

Unfortunately, due to the complexity of the interactions and relationships between your work, there are no tools and no methods to identify the *secondary* effects of the increased service times. It is for this reason that IBM strongly suggests that every enterprise that is considering implementing a multisite sysplex configuration should perform a benchmark before making a final decision or committing significant funding.

In an ideal world, you would be able to benchmark any change that significantly changes your configuration. However, in practice, the urgency of performing a benchmark depends on the likelihood that the proposed change will affect your ability to meet your service level objectives. In the case of a multisite configuration, the following list shows the order of likelihood of the configuration causing performance issues (from most to least likely):

1. Multisite sysplex with work spread evenly across both sites
2. Multisite sysplex with all work running in the same site as the primary disk
3. Synchronous mirroring to a metro-distance disaster recovery site
4. Asynchronous mirroring to remote disaster recovery site

The first configuration is most likely to experience the largest performance effect because of the following facts:

- *Every* access (both reads and writes) to the primary disk from the remote site will incur a service time effect.
- Every access to a CF structure in the other site will incur a service time effect.

5.9.1 Benchmark planning

A critical consideration in performing a benchmark is that the workload volume and mix used in the benchmark accurately reflect the production environment. Ideally, you would also have the ability to measure workloads that reflect your planned future workload growth.

It is possible that the secondary effects of increased service times (increased contention) will have a larger effect than the increased service times themselves. If you perform a measurement with a lower-than-normal volume of transactions, the likelihood of contention between transactions is decreased, meaning that the value of the benchmark is also reduced. Just because your workloads do not display contention today does *not* mean that they will not encounter contention when service times increase.

You must consider the following items when planning a benchmark:

- Do you have a testing environment that is capable of driving production levels of both online and batch work? Consider the following questions:
  - Does the test environment contain the same mix and ratio of transactions as your production environment?
  - Do you have the ability to adjust the transaction rate?

- Performing a benchmark is a time-consuming exercise, so you must make sure that you drive the maximum value from each set of measurements:
  - You must have a thoroughly documented base case. Do you know the normal transaction rates and response times, I/O rates and response times, CF request rates and response times?
  - You must have identified all the metrics and tools that you require to report, investigate, and understand the results of each run.
  - You must have processes in place to collect and archive the required performance data.
- Identify and address any anomalies in a measurement:
  - Know what results you expect to get, so that any divergence can be quickly investigated and addressed if appropriate.
  - Be able to reference your current performance information quickly, so that you can immediately identify any changes:
    - Has the transaction rate decreased? If so why?
    - What part of the transaction (I/Os, CF requests, waiting for a resource, and so on) is taking longer?

    Without your normal performance information easily accessible, it is too easy to go through a series of performance measurements, and only discover later, after the benchmark is over, that some configuration error skewed the results.
- Do you have an existing multisite configuration?

  If so, can you incrementally modify your current infrastructure to make it look like your proposed configuration, so that you can gradually move to the target environment and be easily able to back out if the performance proves to be unacceptable?

5.9.2 Benchmark options

The following list includes available options for performing an extended distance benchmark:

- Depending on your workload, your requirements, and your ability to create a production level of work, it is possible that one of the IBM benchmark centers might be able to provide a configuration for your benchmark. You should discuss your requirements with your IBM representative. If an IBM benchmark center *can* provide the required configuration, bring a copy of your production environment to the benchmarking center, along with the tools and products necessary to generate realistic production levels of work.

  Because of the huge array of sources of z/OS transactions (your own staff, automated teller machines (ATMs), connected enterprises, Internet transactions, distributed platforms, and so on), gathering all of the resources required to create peak production levels of workload could be a challenge.
- Insert the extended distance equipment into your production configuration.

  The advantage here is that benchmark results reflect precisely what you should expect in your production environment if you implement that configuration.

Keep these items in mind as you select a benchmark approach:

- Document your benchmark objectives in detail in advance. Are you solely interested in obtaining performance data? Or do you also want to validate that all of your devices will function correctly with the planned equipment?
- Configure the equipment for the benchmark (for example, switches and DWDMs), in exactly the same way that you would configure them in your planned extended distance solution. For example, if you will use encryption in your solution, ensure that encryption is enabled for the benchmark so that the performance effects are included in the results that you observe.
- Match the device topology in each site, how devices will be connected to switches and WDMs, to that of the eventual solution.
- Obtain different lengths of fiber that can be easily daisy-chained together to create different distances between the sites. Many clients perform measurements of distances that are larger than they plan to implement to determine the point at which performance becomes unacceptable.
- Include recovery testing in the benchmark, especially recovery related to losing one of the inter-DWDM links.

If you opt for installing the benchmark equipment in your production environment, be aware that this method can be disruptive and time-consuming. Schedule the changes during a planned outage window to minimize these issues.

If you decide to change the length of inter-site links, consult with your WDM vendor to determine how much disruption to expect. Also work with your IBM representative to determine how your connected devices would react if the WDM changes from one path to another when you open one of the paths to inject the additional distance.

The DWDM might have the ability to switch from one path to another without the host being aware of the switch. Although this might appear attractive from the perspective of making configuration changes less noticeable, it is not the suggested way to configure DWDM connections with System z.

Changing the distance between the DWDMs without making the CEC aware of the change could cause unexpected results. For this reason, regardless of the capability of the WDM, we suggest that any time you change a cross-site path, do it in a manner that is reflected to the host as a loss of signal.

Director and WDM latency

In planning end-to-end solutions for data centers, most enterprises focus on distance to estimate the effect of that solution in application response time. But another important aspect to consider is the latency caused by equipment, such as Directors and WDM.

Latency injected by directors can vary from 700 nanoseconds (ns) to 2.5 microseconds (µs). If both input and output ports are served by the same ASIC, the latency will be, on average, 700 ns. If the input and output ports are served by different application-specific integrated circuits (ASICs), the latency will be, on average, 2.5 µs.

When you compare the response times of your applications, most of them are likely to be in the multi-milliseconds (ms) range, the equipment latency will not be a concern in a correctly functioning SAN. However, it can become an issue if you are dealing with sensitive applications, or in a SAN where bottlenecks can occur.

Latency injected by the DWDM is more complicated, because it depends on whether you are using a transponder or a muxponder. Latency on such equipment can vary from as little as 10 ns up to more than 100 µs, both per link, or 20 ns and 200 µs per round trip. Your vendor should be able to provide information about the latency injected by their device, and actions that you can take to minimize the latency.

5.9.3 Obtaining the equipment

Work with IBM staff and your vendors to identify the configuration that will be used for the benchmark if you choose an IBM benchmarking center. If you plan to perform the benchmark in your own site, consult with your vendors to see if you can rent or borrow the required equipment.

You will need fiber to perform the benchmark. The IBM interoperability lab uses “fiber suitcases” that contain long spools of minimally shielded fiber. The suitcases can be daisy-chained together to create varying distances. These devices are fragile and easily damaged in transit, so allow sufficient time to locate and obtain them. Your WDM vendor can help locate suppliers in your area.

Electronic devices are available that will inject a delay into a fiber link. However, these are not qualified for use with links that carry coupling link signals. They might not provide the same results as actual fiber, because their physical characteristics are different.

5.9.4 Interpreting the results

As mentioned in 5.9.1, “Benchmark planning” on page 171, it is vital that you perform advance planning for how you are going to determine and analyze the results of your benchmark. There is a plethora of performance metrics available from RMF and the various subsystem monitors. One of the challenges is in identifying the subset of those that you need.

You want one set of metrics that give you high-level information, such as transaction rates, average response times, service times, CPU usage, and so on. These are vital to help you quickly identify if everything is performing in line with your expectations. If some metric is not in the range that you expect, you need a second set of metrics.

The second set includes the metrics and tools that you need to drill down to understand the components of transaction and batch elapsed times. This will likely include IBM CICS® and DB2 System Management Facilities (SMF) records. It is important that you go through the exercise of performing an analysis before the benchmark starts so that you are prepared in the following ways:

- You are familiar with the tools, how to use them, the critical fields that you should be looking at, and the precise meaning of each field.
- You already have a library of information about your current environment that acts as the basis for comparison. Knowing that a transaction spends 17 ms waiting for I/Os to complete in the benchmark environment is meaningless unless you know how long it normally takes.
- You identify in advance all of the data that you need to collect, and all of the options that must be turned on to enable the creation and collection of that data. Having to rerun measurements because the required performance data is not available for the first run is expensive, time-consuming, and frustrating.

For information about the metrics and tools that we used to perform a comparison of various distances in a multisite sysplex environment, see the chapter about CICS/DB2 workload and metrics in *Considerations for Multisite Sysplex Data Sharing*, SG24-7263. That chapter provides detailed information about which record types to collect, and identifies the key fields to help you analyze the results.

5.10 Service provider requirements

Many large organizations rely on business partners, outside vendors, and independent service providers to implement and maintain their connectivity infrastructure. This makes good business sense in many cases; however, outsourcing part of your service delivery environment can complicate the management and maintenance of your environment, especially when you expand the environment into an extended distance solution.

Cost constraints mean that only a subset of System z environments are likely to have a completely dedicated end-to-end infrastructure. It makes business sense to share the cost of some components across different platforms, or potentially even different enterprises. The network and DWDM components are the most commonly shared infrastructure components.

There is no single solution for providing and managing an extended distance configuration, because every environment is unique. You need to examine and fully document your current environment and your requirements to arrive at the configuration that is ideal for your enterprise.

You might already use service providers to manage parts of your System z environment. If so, examine each individual component and gather the service level agreements (SLAs) for each service provider to ensure that each can accommodate your extended distance solution. You might need to contract with additional service providers if your existing vendors do not provide extended distance services. Both you and the service providers must understand the additional complexity that can be introduced in an extended-distance solution.

Reasons to use service providers

Some organizations use service providers for those services that are not part of their core business strategy, or that are not cost-effective for them to manage themselves. One example might be data center cabling, a time-consuming function with intermittent periods of high activity. A service provider might be able to provide the cabling needs, including manufacturing, installing, and testing, at a lower cost or with greater expertise.

Another example is to use the service provider as a network provider. The cost of installing and maintaining a dark fiber network might not be affordable for your organization. Apart from the cost, some countries do not allow private enterprises to install their own networks across public property.

Using a service provider assures that you have the technical resources to provide a highly stable network service. Also, because provisioning and managing the network infrastructure is the service provider’s core business, it enables you to focus on the System z aspects of the extended distance solution. To provide failure isolation of the network, the devices, and support, it is not uncommon to use more than one service provider.

5.10.1 Service provider environments

Using a service provider might offer benefits and give you access to expertise that you might not otherwise have. However, it can also change the way your environment must be managed, because service providers usually operate within specific bounds. Their view of the service they are responsible for is often restricted to everything between two specific connection points.

In an extended distance System z environment, it is likely that you will use multiple service providers to provide the required level of failure isolation. However, this can also complicate management, coordination, and problem resolution because three parties might need to be involved: You, and the two service providers.

One thing that you need to be cognizant of is that the service providers might have an existing agreement with one or a specific set of preferred WDM and switch vendors. However, they *must* be made to understand that it will be a contractual requirement to provide equipment that is qualified by the IBM qualification process.

Similarly, because of the stringent operating tolerances of System z, the vendors must understand the importance of providing accurate information about the length of each path between your two sites. If one path is longer than the supported distance, this can result in difficult-to-detect problems and potential data integrity exposures.

Be sure to set clear expectations with a service provider in terms of clear reporting structures, problem resolution responsibilities, and interactions with other service providers. In System z environments, problem resolution can be unnecessarily prolonged if each service provider only looks at the components that they are responsible for.

For example, errors on the storage SAN can be mistaken for faulty cables or channels, when they actually could be a network link running in a degraded mode. Similarly, channel errors can be mistaken for network issues. Remember that a problem in one part of the System z environment can quickly affect operations in others. Speedy and accurate identification of the problem requires that all parties work together effectively.

The Connectivity Architecture Group should own the relationship with all infrastructure service providers, because they are the one team that is most likely to be dealing with all of the groups involved in managing the end-to-end infrastructure. Each individual department (storage, SAN, performance, sysplex, and so on) should maintain the responsibility for managing and monitoring their part of the solution, but all should liaise and work closely with the Connectivity Architecture Group.

Depending on the services offered by the service provider, your in-house skills, and your view of where responsibilities should be located, you might be able to select specific components from the service providers and provide others yourself. For example, you might use a service provider for the cross-site connectivity, but use your own technical resources to configure the ISL definitions that connect to that network. Work closely with the service providers in your area to determine what services they offer, and choose the ones that best meet your needs and budget.

5.10.2 Client and network service provider responsibilities

The network is always the most critical component in a System z end-to-end extended distance solution. When you choose a network service provider, ensure that their service uses qualified DWDM equipment, including qualified firmware levels. This might mean running older levels of firmware, because the most recent levels of code might not yet be qualified.

The following list includes some items that you might want to include in the contract with the service provider:

- The service provider cannot change firmware levels without your prior agreement.
- The service provider cannot use a firmware level that is not qualified.
- The service provider cannot make a change that increases the distance of the links between the DWDMs.
- State clearly who is responsible for monitoring the status of the link, and the processes for notification in case of errors or problems.
- The service provider must stipulate that there are no Single Points Of Failure in the configuration (if you are using two service providers, they are responsible for ensuring that they are not sharing any network infrastructure).
- The monitoring information and tools that will be made available to you.
- The maximum amount of time that a link can be down for.
- The duration and frequency of maintenance windows.

5.10.3 Service provider monitoring

Monitoring is critical in an extended distance System z solution. A failure in one component can cause a sysplex-wide outage, or affect critical applications or systems. Make sure that the network server provider you select offers a robust monitoring solution that includes a notification process, some form of application programming interface (API) for System Automation, and a problem diagnosis, tracking, and management function.

Service providers can use vendor equipment to send notifications to a monitoring tool. The higher the level of automation and monitoring included in an extended-distance solution, the faster potential problems can be detected and corrected.

Capacity and usage statistics are also critical components. Your network service provider might offer a “black box” solution that does not give you access to tools to investigate capacity and usage statistics, or to view information that warns you of potential bottlenecks or under used areas.

In such cases, the network service provider should supply regular feedback regarding capacity statistics. A better solution is for the network service provider to offer you an interface that enables you to access the information yourself.

5.10.4 Client and service provider partnership

To ensure success, the partnership between you and the service provider must be mutually beneficial; failures by either party can result in costly, avoidable, outages.

On your side, you must inform the vendor of any planned workload or configuration changes that would change the load on the extended distance infrastructure. Outages can result when resources are overused (for example, by adding more or faster disk storage to the infrastructure without first verifying that the network can handle the increased capacity).

Conversely, if the network service provider does not inform you of infrastructure changes (applying code to DWDMs to fix a problem experienced by other users of the DWDM, for example), the configuration might no longer be qualified for System z.

It is vital that both parties understand that System z has unique and stringent requirements that do not exist on the other platforms that the supplier might be more familiar with. IBM, WDM, and switch vendors would not make such significant investments in the qualification process unless experience proved it to be necessary.

A

Performance considerations

This appendix provides information about the relationship between distance and performance that is a critical factor when creating an end-to-end architecture that will support your current and future business requirements. Specifically, we provide information about the following areas:

- Performance and distance
- Relationship between the sites
- Physical and logical connectivity
- General considerations
- Coupling facility-related considerations
- Disk-related considerations

Performance and distance

For many years, it has been known that one of the best ways to improve performance for online transactions and batch jobs was to avoid performing input/outputs (I/Os). The data-in-memory exercises of the 1980s were intended to help address the disparity between processor speeds and disk speeds by keeping more data in memory, thereby reducing the number of times the program had to wait for an I/O to complete.

Conversely, anything that you do that *increases* service times when you go outside the processor will adversely affect transaction and batch job performance. Obviously, increasing the distance between the central electronics complex (CEC) and the devices it is using will increase service times. To understand how this increase will affect your applications, you need to consider two key questions:

- Are the trips outside the processor synchronous to transaction or batch job execution?

  For example, asynchronous mirroring I/Os occur after the I/O initiated by the application program has completed, so they do not affect transaction response times. Alternatively, requests to a coupling facility (CF) lock structure to obtain serialization before making a database update *do* happen inside the execution of the transaction, so any increases in the service time of those requests have the potential to increase the overall transaction response time.

- For I/Os and CF requests that *are* synchronous to transaction execution, how large will the increase in service time be, in comparison to current service times?

  For example, if the service times will increase by 5%, that might not have a noticeable effect on response times. If the service times increase by 50%, or 500%, that is more likely to have a noticeable effect.

This appendix helps you understand the relationship between distance and performance, and includes some of the things that you can do to help offset some of the performance and capacity effects of distance.

Relationship between the sites

The first question (whether the trips outside the CEC are synchronous to transaction execution) is largely related to the relationship between the sites. If the second site's primary role is as a remote disaster recovery (DR) site, and asynchronous mirroring is going to be used, the mirroring I/Os *should* not have an effect on transaction or batch job response times[1].

The performance of the asynchronous mirroring infrastructure is important:

- Bottlenecks or slowdowns in the mirroring process can delay writes from the production system to the primary disk subsystem.
- If the mirroring I/Os are delayed, the secondary disks will fall further behind the primary disks, meaning that more data would be lost in case of a disaster.

The specific actions to take to address asynchronous mirroring performance problems are specific to the replication technique being used. See the product documentation for information about addressing performance issues. The remainder of this appendix will focus on configurations based on synchronous mirroring, because application performance is more likely to be affected in those configurations.

[1] Assuming that there are no bottlenecks in the mirroring infrastructure.

Physical and logical connectivity

To assess and address the potential for performance issues, we need to look at the aspects of the configuration and understand how they contribute to application performance. Figure A-1 contains a simplified representation of the physical connectivity in a multisite sysplex configuration.

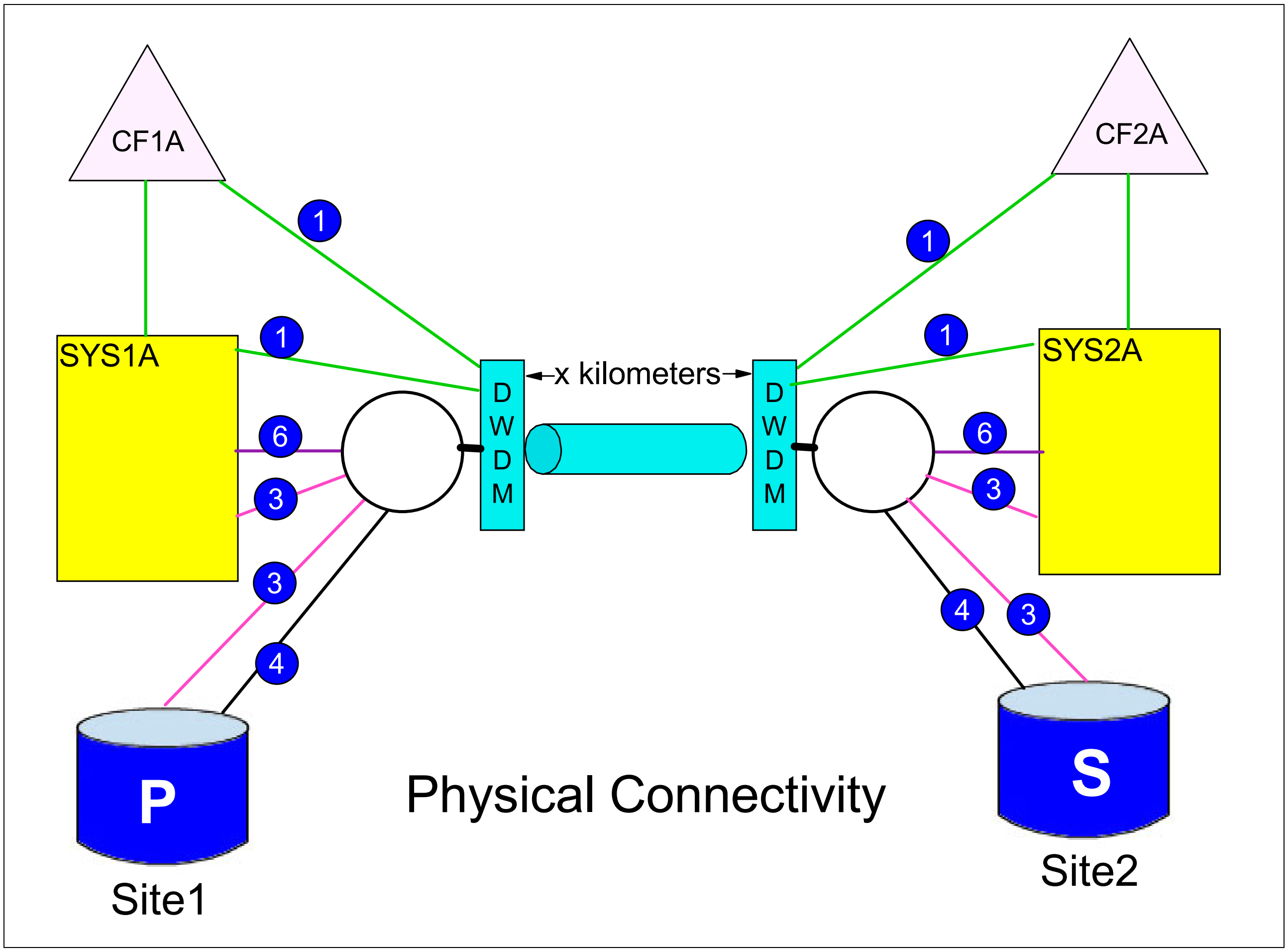


Figure A-1 Simplified physical connectivity for multisite sysplex

Figure A-1 is only intended to show the types of connectivity that you would need between the two sites. Obviously a real configuration should have multiple instances of all critical components (for example, multiple switches in each site, multiple dense wavelength division multiplexers (DWDMs), multiple links from each device to the switches or DWDMs, and so on). Figure A-1 uses the following conventions:

- The connections labeled *1* are 1X InfiniBand (IFB) Coupling Links.
- The connections labeled *3* are Fibre Channel connection (FICON) channels used for host-to-disk I/Os.
- The connections labeled *4* are Fibre Channel Protocol (FCP) links used for synchronous disk mirroring.
- The connections labeled *6* are FICON channels used for channel-to-channel (CTC) links between the systems in the two sites.

To understand which I/Os (which includes CF requests) are going to be affected by the distance between the sites, we also need to look at the logical configuration. This is shown in Figure A-2.

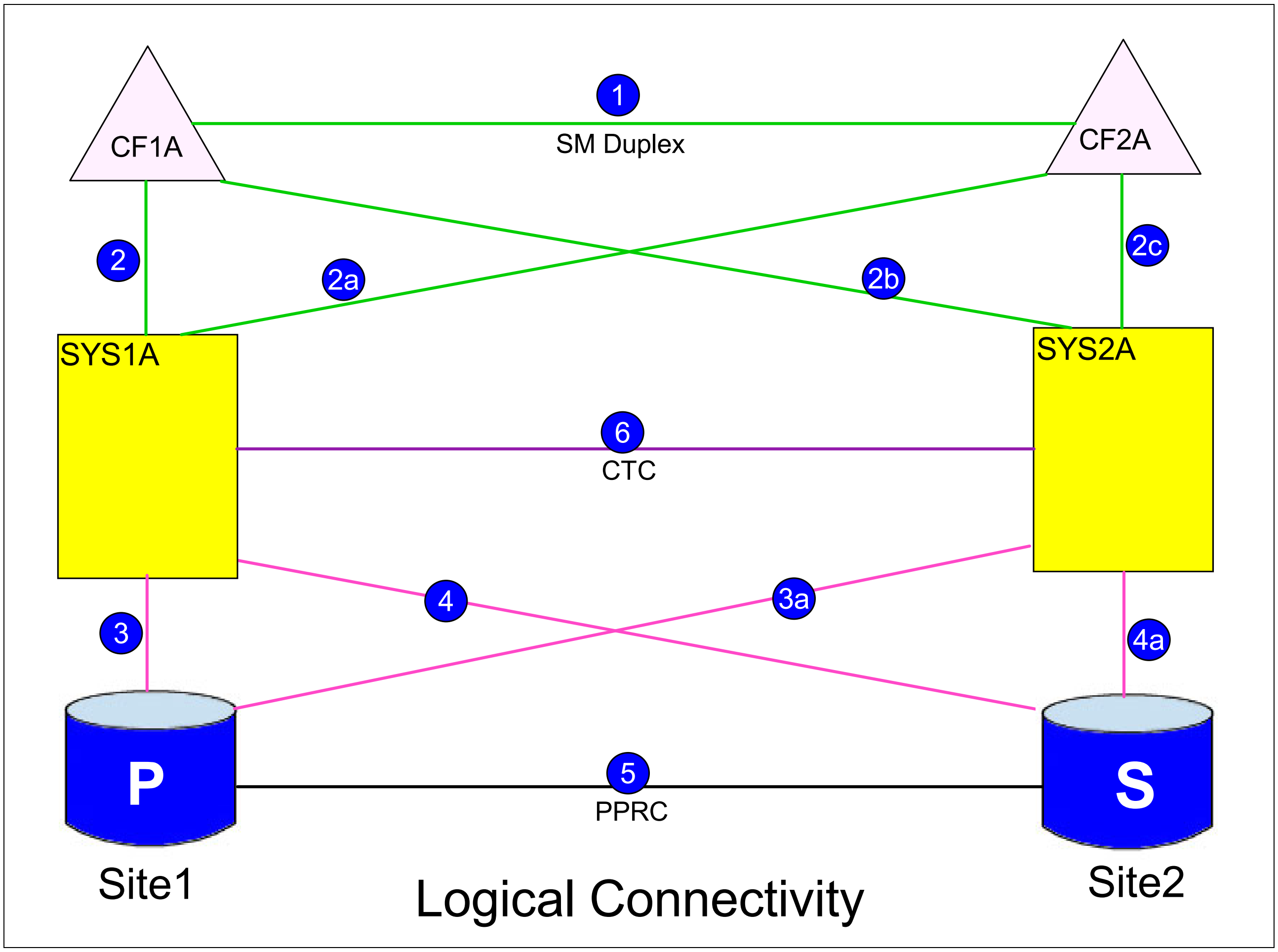


Figure A-2 Multisite sysplex logical connectivity

In summary, we have several types of activity using the connections shown in Figure A-2. These include host-to-CF requests, CF-to-CF requests (for System-Managed Duplexed structures), host-to-disk I/Os, synchronous mirroring I/Os, and CTC I/Os. Figure A-2 shows the following labeled connections:

- The connections labeled *1* are CF-to-CF links used for System-Managed Duplexing.
- The connections labeled *2* are used for CF requests from SYS1A to CF1A.

  Requests sent from SYS1A to CF1A generally should not be affected by the distance between the two sites, with the following exceptions:

  – Requests to a structure that is duplexed using System-Managed Duplexing will experience longer service times. System-Managed Duplexing will be covered in more detail in “System-Managed Duplexing” on page 189.

  – Writes to a cache structure in CF1A that involve driving a cross invalidate to a system in Site2 will experience longer service times. The write to the cache structure does not complete until the cross-invalidate signal has been received by all other systems that have an in-storage copy of that data. If the target system is in the other site, the service time for the write to the local CF will include the time to send the cross invalidate to the other site.

As a result, writes from a database manager to a local cache structure will tend to see elongated service times when other members of the data sharing group move further away from the CF.

- The connections labeled 2a are used for CF requests from SYS1A to the CF in Site2 (CF2A). CF requests from SYS1A to any structure in CF2A will experience longer service times. The service time will be roughly whatever the service time would have been if the structure had been in CF1A, plus 10 microseconds per kilometer (µs/km) between SYS1A and CF2A.

  Additionally, if the service time for a synchronous request from SYS1A to CF2A would exceed the sync/async threshold (a minimum of 26 µs for a zEC12 at the time of writing), the request will be converted to an asynchronous request, which will add some variable amount of time to the service time.

  If the request is to a System-Managed Duplexed structure, the service time would be likely to increase by an additional 20 to 30 µs/km between SYS1A and CF2A.

  If the request is an update to a cache structure, and a system in Site1 has an in-storage copy of the data that has been updated, the write to CF2A will be elongated by the time it takes to perform the cross invalidate to the systems in Site1.

- The connections labeled 2b are used for requests from SYS2A in Site2 to CF1A in Site2. The performance considerations for those requests would be the same as those sent across connection 2a from SYS1A.
- The connections labeled 2c are used for requests from SYS2A to CF2A. Similarly to requests sent over connection 2 by SYS1A, requests sent on connection 2c would not be affected by the distance between the sites, except for requests to System-Managed Duplexed structures, or write requests to a cache structure that result in a cross invalidate being sent to a system in Site1.

> **Important:** Not every CF request will be affected by the distance between the CFs. If the workload is split evenly across the two CFs and the two systems, roughly half the CF requests will be affected, because half of the CF requests from each system will go to the local CF, and half will go to the remote CF. Of course, if the CF and system workloads are not split evenly across the two sites, the number of requests that go to the local and remote CF would be adjusted accordingly.

- The connections labeled 3 are used for read and write disk I/Os from SYS1A to the primary disk.

  The service time of the read I/Os is unaffected by the distance between the sites. However, the service time of write I/Os will increase by the time it takes to mirror the writes to the secondary disks in Site2.

- The connections labeled 3a are used for read and write disk I/Os from SYS2A to the primary disk in Site1.

  The service time of read I/Os will increase by roughly 10 µs/km between the sites. The service time of write I/Os will increase by 20 µs/km, because the write has to be sent across to the primary disk in Site1, then mirrored back to Site2.

  The connections labeled 4 and 4a would be used if a HyperSwap is performed, to swap the primary disks to Site2. If your plans include the ability to perform HyperSwaps, you must ensure that there is sufficient channel bandwidth between SYS1A and the disk in Site2, and between SYS2A and the disk in Site2.

- The connections labeled 5 are the Peer-to-Peer Remote Copy (PPRC) links. The performance effect of enabling PPRC is typically less than 0.5 milliseconds (ms), plus 10 µs/km.

- The connections labeled 6 are used for CTC I/Os between the sites.

  The reason that we separated these out is that CTC I/Os tend to be small, meaning that they are particularly sensitive to long distances. Also, as mentioned in 2.9.6, “Miscellaneous” on page 93, because of their small frame size, CTCs tend to use a large number of buffer credits, so you must ensure that the switches provide enough buffer credits to keep the link fully used.

Having briefly gone over the different types of requests that pass between the systems and the connected devices, and how they are affected by the distance between the sites, we will now go into a bit more detail about the considerations for the different types of requests, and actions that you can take to minimize the effect. Before that, however, we will provide information about general performance considerations for a multisite configuration.

General considerations

At this point, you can see that the performance effect of running a multisite configuration is related to the type and number of requests sent between the sites, and the distance between the sites.

If all of the workload runs in the same site as the primary disk, and the sysplex does not span the two sites, the only performance effect will be on write I/Os. Your IBM storage specialist has access to tools that can help you estimate the response time that you would expect to observe if you add a given distance between your primary and secondary disks. Information about offsetting some of the effects of the longer service times is provided in “Disk-related considerations” on page 193.

Capacity considerations

Although most of the performance-related attention will naturally be focused on response times and throughput, changes that increase I/O and CF request service times can also have an effect on used capacity:

- Because of the increased distance between the CEC and the disk subsystems and CFs, I/Os and CF requests take longer to complete.

  Therefore, the subchannels and control blocks associated with the requests will be busy for longer. You cannot easily decrease the service times, but there are capabilities that you can use (described in “CF subchannel considerations” on page 188 and “Related disk and z/OS features” on page 207) to alleviate the subchannel contention that could result from the longer service times.
- The increased service times cause transactions and batch programs to run longer and hold resources longer.
- Because the transactions and batch programs take longer to complete, there is an increased chance of contention with other work, which increases the elapsed time for the work that is delayed by the contention.
- The resulting longer-running transactions increase the number of concurrently active transactions that the transaction manager has to manage. This increases central processing unit (CPU) usage and memory occupancy by the transaction manager.
- For the transactions that encounter contention, the database manager has additional work to do (more CPU and more cross-system coupling facility (XCF) signaling) to manage the contention. Depending on the amount of contention, this can result in increased resource consumption by the database manager.

One of the key metrics to monitor when performing a benchmark, or after moving to a multisite configuration, is the amount of contention, particularly contention for database locks. If there is a significant increase in the level of contention, monitor the average CPU consumption per transaction and adjust the capacity available to the system as appropriate.

Workload split

Another important consideration is how the work will be spread across the sites. If all the work runs in one site (SingleSite Workload, or SSW), the effect would be the response time effect of mirroring disk writes, plus possibly duplexing CF structures across the two sites.

Alternatively, if works run in both sites, the effect is much more variable. This is because there are so many ways to configure your workload:

- You could run all applications in both sites, with an objective to have half the work running in each site.
- You could run all transactions and batch jobs that perform updates in one site, and all the read-only work in the other site.
- You could run one set of applications in one site, and a different set in the other site.
- You could aim for a loose 50/50 split, but with selected jobs or transactions running in just one site.
- You could use the range of workload routing tools (Sysplex Distributor, Virtual Telecommunications Access Method (VTAM) Generic Resources, Workload Manager (WLM) Scheduling Environments, and so on) to gradually adjust the percentage of transactions run in each site.

Perhaps the important point is that you have flexibility in where your work runs and how it is split across the sites. Both sites will require sufficient capacity to run all work in case either site is unavailable, so it should be possible to dynamically adjust the balance of work across the sites, based on your experience with service times, contention, transaction response times, and batch job elapsed times.

If you find it necessary to fine-tune what work runs where, you might find it advantageous to run update-intensive batch jobs in the same site as the database lock manager and the primary disks.

Using data in memory techniques

Because every trip outside the processor has the potential to slow a transaction or batch job, anything that you can do to avoid the number of times something has to be retrieved from a device or CF is a good thing. There are several techniques that should be investigated. Note that these make sense in a single-site configuration, too, it is just that the benefits are likely to be even more pronounced in a multisite configuration.

System-managed buffering

Virtual Storage Access Method (VSAM) can use system-managed buffering (SMB) to determine the number of buffers and the type of buffer management to use for VSAM data sets. This is done with an objective to minimize the number of I/Os required to retrieve the required data from the data set.

To indicate that VSAM is to use SMB, use one of the following options:

- Specify the `ACCBIAS` subparameter of the job control language (JCL) data definition (DD) statement **`AMP`** parameter, and an appropriate value for record access bias.
- Specify Record Access Bias in the data class, and an application processing option in the access method control block (ACB).

To be eligible for SMB, the data set must have the following characteristics:

- Storage management subsystem (SMS)-managed
- Defined as an extended format VSAM data set (`DSNTYPE=EXT`) in the data class

For more information about SMB, see the following publications:

- *z/OS DFSMS Using Data Sets*, SC26-7410
- *z/OS V1R3 and V1R5 DFSMS Technical Guide*, SG24-6979

System-determined blocksize

When the distance between the systems and the disk increases, you quickly get to the point that it takes more time to send the signal up and down the channel than it does to retrieve the data from the device. So you can see that it is important to retrieve the required data in as few I/Os as possible.

One way to achieve that for sequential data sets is through the use of efficient blocksizes. It is amazing how many sequential data sets still use a blocksize of 3200 bytes. It is far more efficient (and also a better use of disk space) to use half-track blocking.

Rather than having to go through the exercise of calculating the optimum blocksize for a given logical record length, you can (and should) use a blocksize of 0. This indicates that you want the use a system-determined blocksize. In other words, you want the system to calculate the optimum blocksize for the data set.

To avoid the onerous task of having to go through and fix all of your JCL, you can use the SMS Data Class `Forced System Determined Blocksize` attribute to indicate that the data set should use a System-determined blocksize, even if `BLKSIZE` is specified by the user.

For more information about the use of System-determined blocksize, see the following publication:

- *z/OS DFSMS Implementing System-Managed Storage*, SC26-7407

LLA and VLF

The MVS library lookaside (LLA) and virtual lookaside facility (VLF) both provide the ability to buffer objects in storage, avoiding the need to perform I/Os to retrieve this information.

LLA is designed to buffer executable modules in storage, enabling faster program loading, and avoiding the time to load the program from disk.

VLF provides a generalized service to store frequently used and infrequently updated objects in memory. IBM users of VLF include LLA, the catalog address space (for buffering catalog records), IBM RACF®, and Time Sharing Option Extensions (TSO/E).

For more information about LLA and VLF, see the following publications:

- *System z Mean Time to Recovery Best Practices*, SG24-7816
- *z/OS MVS Initialization and Tuning Guide*, SA22-7591
- *z/OS MVS Initialization and Tuning Reference*, SA22-7592

Sort utilities

A mainstay of batch processing is still sorting data into the required sequence. Sort programs can use huge amounts of memory, and if poorly tuned, they can also generate large numbers of I/Os. You should review your batch jobs and the performance information provided by your sort program to ensure that it is operating as efficiently as possible.

In a multisite configuration, it might make sense to make additional memory available to the sort program if that would result in a reduced number of disk I/Os (especially sort work I/Os, which would end up getting mirrored to the remote secondary disks).

Use of compression

Big data has reignited the dialog about the use of compression on the mainframe. There is no doubt that compressing data sets can reduce the cost of storage devices to contain all that data. On the other hand, there is a CPU cost in compressing and decompressing the data.

The Data Facility Storage Management Subsystem (DFSMS) provides the ability to compress sequential and VSAM data sets. DB2 also supports the ability to compress its databases. Similarly, IMS supports data compression for its databases.

Most installations have policies that determine at what point in the life of a data set it should be compressed. If it is compressed while it is still being frequently accessed, the decompression cost must be borne multiple times. So, there is a balance to be achieved between saving on disk and tape storage, against the z/OS CPU resources to compress and decompress the data.

Implementing a multisite configuration might cause that balance to be revisited. Fewer I/Os are required to read and write a compressed file (meaning fewer traversals of the distance between the two sites). Therefore, the performance benefits might make it viable to compress the data set earlier in its life.

Additionally, the announcement of the zEnterprise Data Compression (zEDC) capability on IBM zEnterprise EC12 (zEC12) and IBM zEnterprise BC12 (zBC12) will significantly affect the cost equation for data compression, adding to the argument for compressing data earlier in its life.

For more information about zEDC, see the following publication:

- *IBM zEnterprise EC12 Technical Guide*, SG24-8049

Well-balanced switch configuration

It is important to work with your switch vendor to ensure that you provision a well-balanced configuration, bearing in mind the speeds of all adapters connected to the switches, the distances involved, and the buffer credit requirements.

However, this should not be a once-off exercise. Modern System z environments are dynamic, with more capacity constantly being added, and older devices being replaced with newer ones with different capabilities (for example, replacing a CEC with FICON 4 Gb adapters with one with FICON Express 8S adapters). You should have an ongoing relationship with your switch vendor to ensure that the storage area network (SAN) continues to be able to deliver the required levels of performance and availability.

Coupling facility-related considerations

The role of CFs in the performance of your multisite configuration depends mostly on how extensively you use the CF's capabilities. If the CF workload is small and limited to resource sharing, it is likely that the service times delivered by the CF will not have a significant effect on your overall levels of performance.

However, if you are serious enough about availability to be investing in a multisite configuration, it is reasonable to assume that you *do* use CF capabilities, such as lock and cache structures, to enable data and queue sharing. This section contains information to help you optimize CF performance and capacity in a multisite configuration.

CF subchannel considerations

Every coupling link channel-path identifier (CHPID) has either 7 or 32 subchannels. The number of subchannels is specified when the CHPID is defined in the hardware configuration definition (HCD). When a CF request is started, it obtains a subchannel. The subchannel is not released until the CF response is processed by XCF. Therefore, the amount of time that the subchannel is busy for each request is roughly equal to the service time of that request.

Because the service time increases when the CF moves further away from the z/OS CEC, the subchannel usage will also increase. Additionally, if the request is converted to an asynchronous request because of the service time, that will also increase the service time, increasing subchannel usage even more.

As an example, assume that the CF is located beside the z/OS CEC, and delivers 10 µs service times. The CHPID would have 7 subchannels associated with it, so 50,000 requests a second would result in subchannel usage of

```
50,000 x 10
--------------- = 7.14%
1,000,000 x 7
```

If the CF is moved 10 km from the z/OS CEC, the service time would increase to roughly 10 (original service time) + 100 (10 µs x 10 km) + 40 (allow for synch to async conversion), or 150 µs. Inserting that new service time into the described calculation, we get

```
50,000 x 150
--------------- = 107%
1,000,000 x 7
```

The IBM guideline for CF subchannel usage is that it should not exceed 30%, so you can see that the usage would now far exceed the suggested maximum. Fortunately, HCA3 1X IFB adapters added the ability to have 32 subchannels for each coupling link CHPID. In this case, the CHPID definition in HCD could be changed to specify that 32 subchannels should be defined, rather than the original 7. This changes the calculation as follows:

```
50,000 x 150
--------------- = 23.4%
1,000,000 x 32
```

This is still high, but it is less than the suggested maximum.

Suggestion: Coupling link CHPIDs for local CFs should be defined with 7 subchannels, even for 1X links. Coupling link CHPIDs for remote CFs should be defined with 32 subchannels.

System-Managed Duplexing

System-Managed Duplexing can be used to provide a recovery capability to CF structures, as shown in Figure A-3. It is generally used by programs that do not support user-managed rebuild. However, there are some components (such as independent resource lock manager (IRLM)) that use it to recover from a failure affecting both its lock structure *and* one or more connectors.

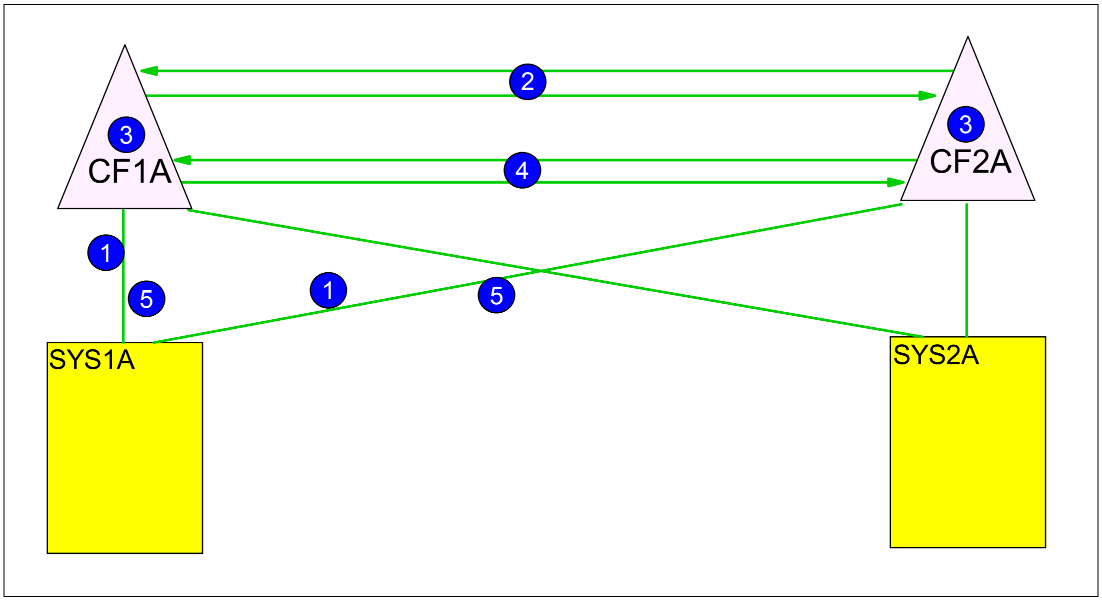


Figure A-3 System-Managed Duplexing interactions

Figure A-3 illustrates the steps in a request to a System-Managed Duplexed structure:

1. The CF request is sent to both instances of the structure (in CF1A and CF2A).
2. Each CF sends a Ready-To-Execute (RTE) signal to its peer CF, indicating that it is ready to start processing the request. When the CF has both sent its RTE signal *and* received the corresponding signal from its peer CF, it can start running the request.
3. The request is run in each CF.
4. When each CF finishes processing the request, it sends a Ready-To-Complete (RTC) signal to its peer, informing it that it has completed processing and is ready to release the resources associated with that request.
5. When the CF has sent its RTC signal and received the corresponding RTC from its peer, it sends its response back to SYS1A.

For a simplex request to a structure in the remote site, the distance only affects the request once. But for a request to a System-Managed Duplexed structure, *every* interaction between the peer CFs will also be affected by the distance. Every enterprise needs to make a determination about the value of System-Managed Duplexing. If you use it for some structures and it is the only way to provide recovery for those structures, you have three options:

- Duplex across the sites and accept the resulting service time.
- Set up a second CF in each site (possibly as an internal CF in the same CEC as the z/OS systems) and duplex between the two CFs in the same site.
- Investigate and document the process for recovering from a loss of that structure, then revert the structure to simplex mode.

Additionally, coupling facility control code (CFCC) Level 16, available on IBM z10, added support for an expedited version of the System-Managed Duplexing protocol that might reduce service times for structures that are duplexed over peer CFs that are remote from each other. The modified protocol is enabled using the **SETXCF FUNCTIONS,ENABLE=DUPLEXCF16** command. You can also enable it by updating your COUPLExx member of parmlib, as shown in Example A-1.

Example: A-1 Enabling DUPLEXCF16 in COUPLExx member

```
COUPLE SYSPLEX(&SYSPLEX.)
       PCOUPLE(SYS1.XCF.CDS05)
       ACOUPLE(SYS1.XCF.CDS06)
       OPNOTIFY(+3)
       CLEANUP(15)
       MAXMSG(2000)

FUNCTIONS ENABLE(DUPLEXCF16)
```

Tip: To optimize the System-Managed Duplexing protocol for a multisite sysplex configuration, enable DUPLEXCF16 on all systems in the sysplex.

Duplexed DB2 group buffer pools

DB2 is able to recover from the loss of a CF containing one or more group buffer pool (GBP) structures. It achieves this by reading back through its logs, identifying all updated pages that had been written to the GBP but not yet cast out to disk, and updating the database with those changed pages.

Although this is effective, recovery can take a long time. To expedite recovery from a CF outage (and also speed up the process of moving a GBP from one CF to another), DB2 supports User-Managed Duplexing for its GBP structures. User-Managed Duplexing is similar to System-Managed Duplexing in that it enables rapid recovery from a CF failure. However, the mechanics of how it works are completely different from System-Managed Duplexing.

As the name infers, User-Managed Duplexing is managed by the structure user (DB2 in this case). DB2 is aware that a GBP structure is duplexed and is responsible for updating the secondary GBP structure with changed pages. To achieve this in as efficient a manner as possible, DB2 sends an asynchronous write request to the secondary GBP. While that is running, it sends a synchronous request to the primary GBP.

This enables the two updates to run in parallel. It also avoids the interactions between the CFs that are an integral part of the System-Managed Duplex process, and as a result, duplexed GBPs are not as sensitive to the distance between the CFs as System-Managed Duplexing is.

With both System-Managed and User-Managed Duplex structures, read requests are always sent to the primary structure instance, and castout processing is always done from the primary instance. Updated database pages are sent to both structures. Based on this behavior, it might be valuable to try to configure DB2 so that the castout owner of heavily updated pagesets is in the same site as the primary disks.

The z/OS heuristic algorithm

CF requests can be handled by XCF in one of two ways:

- The request is sent to the CF, and XCF spins, waiting for the response to come back from the CF. This delivers the best possible service time because there is no delay between when the response arrives back in the z/OS CEC and when XCF starts processing it. If the service time is short, this is also the most efficient way to handle the request from the perspective of the amount of z/OS CPU expended on handling the request. This type of request is known as a synchronous request.
- The request is sent to the CF and XCF then passes control to the MVS Dispatcher. The Dispatcher determines what task should run next and passes control to that task. At some point, control is passed back to the MVS Dispatcher. The Dispatcher checks to see if the response from the CF has arrived back yet. If it has, control is returned to XCF and it then processes the response.

If the request took a long time to process, handling it as an asynchronous request is more efficient, because rather than spinning, the CPU was actually performing some work. However, switching tasks four times also uses some amount of CPU. For example, suppose that this time is 26 µs on a zEC12.

If the synchronous request would have taken longer than 26 µs to run, less z/OS CPU is used by treating it as an asynchronous request. Alternatively, if the synchronous request would have completed in 5 µs, then treating it as an asynchronous request would have used an additional 21 µs of z/OS CPU time.

In an effort to minimize the z/OS CPU usage of CF requests, z/OS 1.2 introduced a heuristic algorithm that controls whether a synchronous request gets converted to an asynchronous one, based on its knowledge of the CPU cost of handling an asynchronous request and the expected service time for the request, derived from recent service times for that type of request to that CF.

If the expected service time is greater than the cost of an asynchronous request, the request will be converted to an asynchronous one. If the expected service time is less than the cost of an asynchronous request, the request will be left as a synchronous request.

So, presume that the average service time for a given structure is about 10 µs in a local CF. That means that the z/OS CPU time used by that request was 10 µs. Assuming that the z/OS CPU cost of an asynchronous request would be 26 µs, the heuristic algorithm leaves this as a synchronous request, and the z/OS CPU cost is 10 µs. If that structure were moved to a CF 10 km from z/OS, the service time would increase to about 110 µs. Because 110 µs is greater than the asynchronous cost of 26 µs, the request is converted to an asynchronous one.

In this example, the heuristic algorithm saved 84 µs of z/OS CPU time by converting the request to an asynchronous one. However, the request still used 16 µs *more* z/OS CPU time than if the structure had been in the local CF.

Depending on the distribution of work across the two sites, structures across the two CFs, and the location of specific structures in relation to the programs that will use them, moving to a multisite sysplex might result in an increase in the overall z/OS CPU usage used to drive CF requests. Judicious positioning of structures might help you reduce this cost to some extent.

Shared queues

In an ideal Parallel Sysplex, any work can run on any system in the sysplex. Further, you should be able to stop any component in the sysplex, and all of the work in the sysplex should continue running with minimal manual intervention.

Key enablers of that capability are shared queue structures in the CF. These allow work to be added from any system in the sysplex, and to be retrieved and processed by any system in the sysplex. The two primary users of this capability at the time of writing are IBM WebSphere® MQ Series and IMS Transaction Manager.

In addition to making the work accessible to all systems in the sysplex, the use of shared queues also provides a workload distribution mechanism. However IMS and WebSphere MQ work slightly differently in this regard.

Originally, both IMS and WebSphere MQ would inform every interested connector in the sysplex about the arrival of a new piece of work. Every system would be informed at roughly the same time, and whichever system responded first (typically the one with the most spare capacity) would get that message. In a configuration where all connectors are local to the CF, they would all have a similar chance of being the first one to respond to the CF.

However, if some sites are remote to the CF, they would be at a disadvantage, and therefore would be likely to get a smaller share of the workload. And the larger the distance between the connector and the CF, the smaller its chance of being the first to respond.

To address this situation for IMS, the sublist notification delay mechanism was introduced. It enables you to control how long the CF will wait after informing one connector about the existence of a new message before it informs the other connectors. Specifying a very small value will tend to make the system behave in a similar manner to the original design. Specifying a larger value will tend to make it work in more of a round-robin fashion, with each connector being more likely to get its fair share of the messages.

The delay is controlled, at the individual structure level, by the `SUBNOTIFYDELAY` keyword on the structure definition in the coupling facility resource management (CRFM) policy. The value specifies the number of µs to delay before telling the other connectors, and can be between 1 and 1000000. The value can be changed dynamically by updating and starting the CFRM policy, so you can make small adjustments to identify a value that provides the results that you want.

At the time of writing, WebSphere MQ Shared Queues still work in the original way, meaning that, all other things being equal, systems in the same site as a shared queue will tend to process more messages from that structure than a system that is remote to the structure.

Coupling Thin Interrupts

Before zEC12 GA2 (or zBC12 GA1), there was no interrupt mechanism for coupling links. CFCC discovers new requests or signals from a peer by constantly polling its links, looking for new work. On the z/OS end, the system assist processor (SAP) monitors for incoming CF responses and notifications and turns on a flag in the hardware system area (HSA) indicating the arrival of a response to an asynchronous request, or a list transition notification signal. Some time later, the MVS Dispatcher will detect the flag and pass control to XCF to process it.

In the case of both z/OS and a CF logical partition (LPAR), because there was no interrupt, IBM PR/SM™ would be unaware that the LPAR had some work to do, and therefore would not expedite dispatching it.

The zEC12 GA2 supports a new mechanism called Coupling Thin Interrupts. Coupling Thin Interrupts behave in the following ways:

- Do generate an interrupt, meaning that PR/SM is now aware that there is some work waiting for the LPAR to get dispatched.
- Change the way responses from asynchronous requests are handled in z/OS. Rather than waiting for the MVS Dispatcher to check HSA looking for the flag, the coupling adapter will generate an interrupt. This will interrupt the current task in z/OS and immediately pass control to XCF. This should have the effect of reducing the service times for asynchronous CF requests, especially at times when z/OS is not so busy.

Coupling Thin Interrupts are enabled automatically in z/OS when z/OS V2, or z/OS V1.12 or V1.13 with the required authorized program analysis reports (APARs), are run on a zEC12 GA2 or a zBC12. On the CF end, Coupling Thin Interrupts must be explicitly enabled using the **DYNDISP THIN** command on the CF console.

Coupling Thin Interrupts are not specifically intended for a multisite sysplex configuration, however they should result in shorter and more consistent asynchronous service times. And given that you are likely to have more asynchronous requests in a multisite sysplex, they will benefit that configuration.

Disk-related considerations

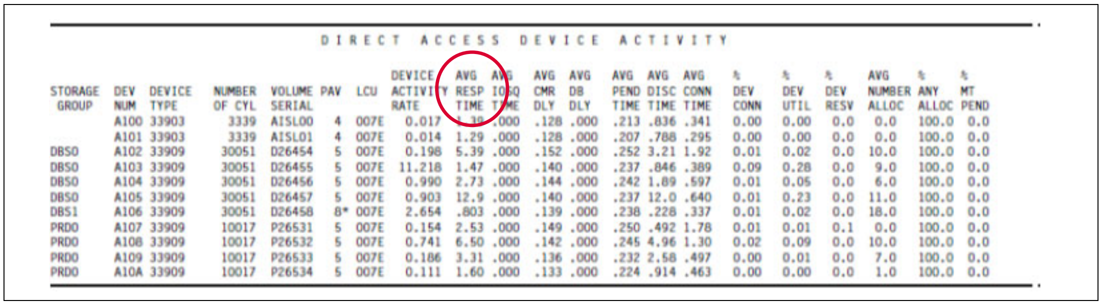
If you recall, in “Physical and logical connectivity” on page 181 we noted that the majority of interactions outside the CEC that are likely to affect transaction and batch performance are for CF requests and disk I/Os. This section provides information about the disk component of the configuration.

The components of disk response times

Disk I/O response times represent how long it takes to read data from, or write data to, a disk subsystem. More precisely, the response time reported by Resource Management Facility (RMF) or other performance monitors is the average of all I/Os to that volume during the interval being reported on. Disk response time consists of several components:

- Input/output supervisor queue (IOSQ) time
- Pend time
- Disconnect time
- Connect time

Figure A-4 shows a sample RMF Direct Access Device Activity (DASD) report. The field AVG RESP TIME is the sum of the four listed components.



```
                                              D I R E C T   A C C E S S   D E V I C E   A C T I V I T Y

                                                     DEVICE   AVG  AVG   AVG  AVG   AVG  AVG  AVG    %      %      %    AVG     %     %
STORAGE  DEV  DEVICE    NUMBER  VOLUME PAV   LCU    ACTIVITY  RESP IOSQ  CMR  DB    PEND DISC CONN   DEV    DEV    DEV  NUMBER ANY    MT
 GROUP   NUM  TYPE      OF CYL  SERIAL              RATE      TIME TIME  DLY  DLY   TIME TIME TIME   CONN   UTIL   RESV ALLOC  ALLOC  PEND
         A100 33903       3339  AISL00  4    007E    0.017    1.39 .000  .128 .000  .213 .836 .341   0.00   0.00    0.0   0.0  100.0  0.0
         A101 33903       3339  AISL01  4    007E    0.014    1.29 .000  .128 .000  .207 .788 .295   0.00   0.00    0.0   0.0  100.0  0.0
DBS0     A102 33909      30051  D26454  5    007E    0.198    5.39 .000  .152 .000  .252 3.21 1.92   0.01   0.02    0.0  10.0  100.0  0.0
DBS0     A103 33909      30051  D26455  5    007E   11.218    1.47 .000  .140 .000  .237 .846 .389   0.09   0.28    0.0   9.0  100.0  0.0
DBS0     A104 33909      30051  D26456  5    007E    0.990    2.73 .000  .144 .000  .242 1.89 .597   0.01   0.05    0.0   6.0  100.0  0.0
DBS0     A105 33909      30051  D26457  5    007E    0.903    12.9 .000  .140 .000  .237 12.0 .640   0.01   0.23    0.0  11.0  100.0  0.0
DBS1     A106 33909      30051  D26458  8*   007E    2.654    .803 .000  .139 .000  .238 .228 .337   0.01   0.02    0.0  18.0  100.0  0.0
PRD0     A107 33909      10017  P26531  5    007E    0.154    2.53 .000  .149 .000  .250 .492 1.78   0.01   0.01    0.1   0.0  100.0  0.0
PRD0     A108 33909      10017  P26532  5    007E    0.741    6.50 .000  .142 .000  .245 4.96 1.30   0.02   0.09    0.0  10.0  100.0  0.0
PRD0     A109 33909      10017  P26533  5    007E    0.186    3.31 .000  .136 .000  .232 2.58 .497   0.00   0.01    0.0   7.0  100.0  0.0
PRD0     A10A 33909      10017  P26534  5    007E    0.111    1.60 .000  .133 .000  .224 .914 .463   0.00   0.00    0.0   1.0  100.0  0.0
```

Figure A-4 RMF Device Activity report showing average total response time for each DASD device

IOSQ time

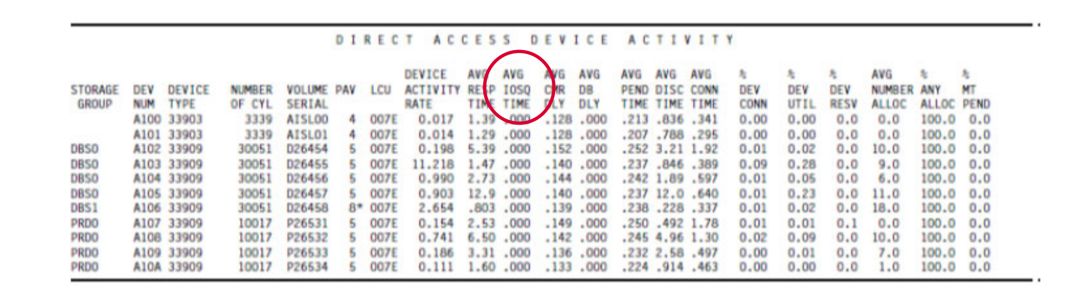
IOSQ time is the time that an I/O request is queued in the LPAR by z/OS. It is the time that an I/O request waits in a queue, possibly because the unit control block (UCB) is busy, waiting to issue the start subchannel (SSCH) instruction. IOSQ time occurs when the I/O supervisor (IOS) tries to start an I/O operation to a device, but the UCB representing the device is being used by a previous I/O request issued by another application in the same LPAR.

High IOSQ time can be caused by elongation of some other response time component. Increasing the distance between the CEC and the disk subsystems, or between the primary and secondary disk subsystems, increases the response time, meaning that the UCB associated with the device will be busy for a longer time. Expect higher IOSQ times when you increase the distance between your data centers.

Another potential source of high IOSQ time is IBM z/OS Global Mirror (zGM), formerly extended remote copy (XRC). XRC can cause increased IOSQ time when the device blocking function is used. Device blocking instructs IOS to inject a delay into every I/O request to prevent the disk subsystem from being overwhelmed when XRC is unable to keep up with the rate at which writes are being sent to mirrored volumes.

Apart from taking actions to reduce the length of the other components of the response time, the only other way to address high IOSQ time is to add more UCBs for each device (or, at least, for the busy devices). Parallel access volumes (PAVs) and HyperPAV can be used to provide alias UCBs. For more information about PAV and HyperPAV, see “PAV and HyperPAV” on page 207.

To identify the current IOSQ time, see the RMF Direct Access Device Activity report. You can also use IntelliMagic’s Direction (formerly called *Disk Magic*) and Vision (formerly called *RMF Magic*) modeling and analysis tools to predict how changing the distance between the CEC and the device, and between the primary and secondary disks (you can get more information about this product at http://www.intellimagic.net). Figure A-5 shows the `AVG IOSQ TIME` field of the RMF Direct Access Device Activity report.



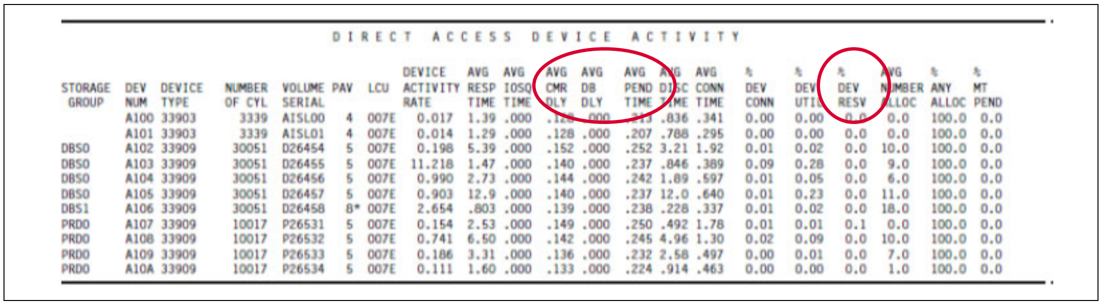
```
                                                 D I R E C T   A C C E S S   D E V I C E   A C T I V I T Y

                                                   DEVICE    AVG  AVG  AVG  AVG    AVG  AVG  AVG    %      %     %     AVG    %     %
STORAGE  DEV  DEVICE    NUMBER  VOLUME PAV  LCU   ACTIVITY  RESP IOSQ CMR  DB     PEND DISC CONN   DEV    DEV   DEV  NUMBER  ANY   MT
 GROUP   NUM  TYPE      OF CYL  SERIAL               RATE   TIME TIME DLY  DLY    TIME TIME TIME   CONN   UTIL  RESV  ALLOC  ALLOC PEND
         A100 33903       3339  AISL00   4  007E    0.017   1.39 .000 .128 .000   .213 .836 .341   0.00   0.00   0.0    0.0  100.0  0.0
         A101 33903       3339  AISL01   4  007E    0.014   1.29 .000 .128 .000   .207 .788 .295   0.00   0.00   0.0    0.0  100.0  0.0
DBS0     A102 33909      30051  D26454   5  007E    0.198   5.39 .000 .152 .000   .252 3.21 1.92   0.01   0.02   0.0   10.0  100.0  0.0
DBS0     A103 33909      30051  D26455   5  007E   11.218   1.47 .000 .140 .000   .237 .846 .389   0.09   0.28   0.0    9.0  100.0  0.0
DBS0     A104 33909      30051  D26456   5  007E    0.990   2.73 .000 .144 .000   .242 1.89 .597   0.01   0.05   0.0    6.0  100.0  0.0
DBS0     A105 33909      30051  D26457   5  007E    0.903   12.9 .000 .140 .000   .237 12.0 .640   0.01   0.23   0.0   11.0  100.0  0.0
DBS1     A106 33909      30051  D26458   8* 007E    2.654   .803 .000 .139 .000   .238 .228 .337   0.01   0.02   0.0   18.0  100.0  0.0
PRD0     A107 33909      10017  P26531   5  007E    0.154   2.53 .000 .149 .000   .250 .492 1.78   0.01   0.01   0.1    0.0  100.0  0.0
PRD0     A108 33909      10017  P26532   5  007E    0.741   6.50 .000 .142 .000   .245 4.96 1.30   0.02   0.09   0.0   10.0  100.0  0.0
PRD0     A109 33909      10017  P26533   5  007E    0.186   3.31 .000 .136 .000   .232 2.58 .497   0.00   0.01   0.0    7.0  100.0  0.0
PRD0     A10A 33909      10017  P26534   5  007E    0.111   1.60 .000 .133 .000   .224 .914 .463   0.00   0.00   0.0    1.0  100.0  0.0
```

Figure A-5 RMF Direct Access Device Activity report showing IOSQ time

Pending time

Pending time is the time between when the SSCH command is issued, and when the disk subsystem accepts the command. You can see the AVG PEND TIME and some of its components in the RMF Direct Access Device Activity report, as shown in Figure A-6 on page 195. Note that PEND time will never be zero, because there is always some amount of time for SAP processing (PEND time includes queuing time in the specific SAP initiative queue).



```
                                    D I R E C T   A C C E S S   D E V I C E   A C T I V I T Y

                                                      DEVICE   AVG  AVG   AVG  AVG   AVG  AVG  AVG    %      %     %    AVG     %     %
STORAGE  DEV  DEVICE   NUMBER  VOLUME PAV  LCU   ACTIVITY RESP IOSQ  CMR  DB    PEND DISC CONN   DEV    DEV   DEV  NUMBER  ANY   MT
 GROUP   NUM  TYPE     OF CYL  SERIAL            RATE     TIME TIME  DLY  DLY   TIME TIME TIME   CONN   UTIL  RESV ALLOC  ALLOC PEND
         A100 33903      3339  AISL00   4  007E   0.017   1.39 .000  .128 .000  .213 .836 .341   0.00   0.00   0.0   0.0  100.0  0.0
         A101 33903      3339  AISL01   4  007E   0.014   1.29 .000  .128 .000  .207 .788 .295   0.00   0.00   0.0   0.0  100.0  0.0
DBS0     A102 33909     30051  D26454   5  007E   0.198   5.39 .000  .152 .000  .252 3.21 1.92   0.01   0.02   0.0  10.0  100.0  0.0
DBS0     A103 33909     30051  D26455   5  007E  11.218   1.47 .000  .140 .000  .237 .846 .389   0.09   0.28   0.0   9.0  100.0  0.0
DBS0     A104 33909     30051  D26456   5  007E   0.990   2.73 .000  .144 .000  .242 1.89 .597   0.01   0.05   0.0   6.0  100.0  0.0
DBS0     A105 33909     30051  D26457   5  007E   0.903   12.9 .000  .140 .000  .237 12.0 .640   0.01   0.23   0.0  11.0  100.0  0.0
DBS1     A106 33909     30051  D26458   8* 007E   2.654   .803 .000  .139 .000  .238 .228 .337   0.01   0.02   0.0  18.0  100.0  0.0
PRD0     A107 33909     10017  P26531   5  007E   0.154   2.53 .000  .149 .000  .250 .492 1.78   0.01   0.01   0.1   0.0  100.0  0.0
PRD0     A108 33909     10017  P26532   5  007E   0.741   6.50 .000  .142 .000  .245 4.96 1.30   0.02   0.09   0.0  10.0  100.0  0.0
PRD0     A109 33909     10017  P26533   5  007E   0.186   3.31 .000  .136 .000  .232 2.58 .497   0.00   0.01   0.0   7.0  100.0  0.0
PRD0     A10A 33909     10017  P26534   5  007E   0.111   1.60 .000  .133 .000  .224 .914 .463   0.00   0.00   0.0   1.0  100.0  0.0
```

Figure A-6 RMF Direct Access Device Activity report showing PEND time-related fields

High PEND time can be caused by the following factors:

- High FICON Director port or disk subsystem host adapter usage:
  - High FICON Director or disk subsystem FICON port usage can be caused by a high activity rate on those ports.

    High FICON Director or disk subsystem FICON port usage can be due to multiple FICON channels from different CECs using the same port on the FICON Director to communicate to the disk subsystem host adapter. In this case, the FICON channel usage as seen from each host might be low, but the sum of the usages of the channels that share the same port can be significant.

    You can use the RMF FICON Director Activity Report to see the usage of each port on the director. For more information about the RMF FICON Director Activity report, see “RMF FICON Director Activity report” on page 205.
  - Some disk subsystems’ host adapters keep track of each active I/O operation running inside them. This monitoring increases the usage of the host adapter, consuming some cycles of its application-specific integrated circuits (ASICs).
  - Increasing the distance between the primary and secondary disk subsystems also increases the service time of active I/Os, resulting in higher host adapter usage and higher PEND time.
- Delays:
  - SAP-busy queuing time. SAP usage should not be higher than 40% to avoid delays. All SAP-related usage and delays are reported in the RMF I/O Queuing Activity report. This time is usually 100 µs or less for a non-overloaded SAP.
  - Command response (CMR) delay. CMR time is the time it takes for the channel to process the first few channel command words (CCWs) in the channel program and send commands to the disk subsystem. This time does not end until the disk subsystem sends a CMR to the channel.

    It is shown in the RMF Device Activity report and the I/O Queuing Activity report. CMR indicates how busy the FICON card is in the disk subsystem. This delay can be higher at long distances because of higher disk subsystems’ host adapter usage.
  - Device busy delay. This can be caused by an extent conflict, because of a read or write operation to an extent that is currently being updated by another I/O request. Device busy delay might also be caused by use of the Reserve macro.

    However, in modern systems Reserves are generally converted to SYSTEMS scope enquiry characters (ENQs). This component of the PEND time is not directly related to your end-to-end solution or distance.

PEND time is caused by anything that prevents the starting of the I/O dialog between the FICON channel and the control unit (CU), after the SSCH has been issued.

For example, imagine a scenario with a connection between a FICON channel and a CU where the CU has fewer buffer credits than the FICON channel. If the control unit cannot process the frames, and does not release the buffers, it will not send the R_RDY back to the FICON channel. Without the R_RDY, the FICON channel will eventually decrement the buffer credit count until it reaches 0, and it will stop transmitting frames to the CU. (For more information about buffer credits, see “Buffer credits” on page 199.)

The time that the channel cannot send frames to the CU is not reported separately by RMF, but it *is* indirectly counted in the PEND time delay. PEND time delay might indicate buffer credits starvation in your environment. Non-zero values in the AVG FRAME PACING field in the RMF Ficon Director Activity report are an indication of a buffer credit shortage. Buffer credits can be an issue at long distances.

Delays introduced by equipment in the connection between the CEC and the disk subsystem (for example, directors and wavelength division multiplexers (WDMs)) might also be counted in the PEND time. In this case, the SSCH has already been issued, but the dialog between the channel and CU did not yet begin.

If remote CECs are attached in the primary disks using WDMs and directors, you might see higher CMR and PEND times, injected by this hardware, and the delay will be higher due to the distance between CEC and the disks. The longer the distance, the more time it takes for the frames to reach the destination.

Summary: In a multisite sysplex, you would expect to see the PEND time increase on the system that is remote to the primary disks. If the increase is more than 10 µs/km between the disk subsystems, or if the system that is local to the primary disk subsystem also experiences an increase in PEND time, some other bottleneck is causing a delay and should be investigated.

Disconnect time

DISC time is when an I/O has already started (SSCH and CMR issued) but the FICON channel and the disk subsystem are not currently exchanging data or control information. The following list includes some causes of high disconnect time:

- A read cache miss

  This occurs when the requested block is not in the disk subsystem cache. The disk subsystem will disconnect from the channel until the requested data has been retrieved from the hard drive.

- High hard disk drive usage

  The average response time of Fibre Channel (FC) hard disk drives (HDDs) or serial-attached Small Computer System Interface (SCSI), or SAS, HDDs is about 10 ms. If the usage of the HDD exceeds 50%, the response time might be twice or even four times that (when usage exceeds 75%). You can check the HDD response time in the RMF ESS Rank Statistics report.

- Internal disk subsystem delays

  Some DISC time delay might be related to back-end queuing in the disk subsystem. In disk subsystem architecture, back-end is the internal physical components, processors and buses, that are responsible for managing and communicating with the HDD.

- Synchronous remote copy

  DISC time contains both the time that it takes for data to travel from the primary to the secondary disk subsystem, and the time the secondary disk subsystem takes to acknowledge receipt of the data to the primary disk subsystem. Note that only write operations experience this delay.

  Consider a write I/O operation to a disk subsystem where that disk subsystem is not using any type of remote copy. The write is recorded in the disk subsystem's cache and the disk subsystem posts an I/O completion to the host that initiated the I/O. No delay is experienced.

  Now consider a write I/O operation to a disk subsystem where that disk subsystem is in a synchronous replication environment. This operation will also be recorded in the primary disk subsystem's cache.

  The subsystem then disconnects from the channel and the data is sent to the secondary disk. The data is recorded in the secondary disk's cache and an acknowledgement signal is sent back to the primary disk, indicating that the write completed. The primary disk then reconnects to the channel and posts the I/O completion to the host.

  Figure A-7 illustrates the sequence of a write I/O when using synchronous remote copy.

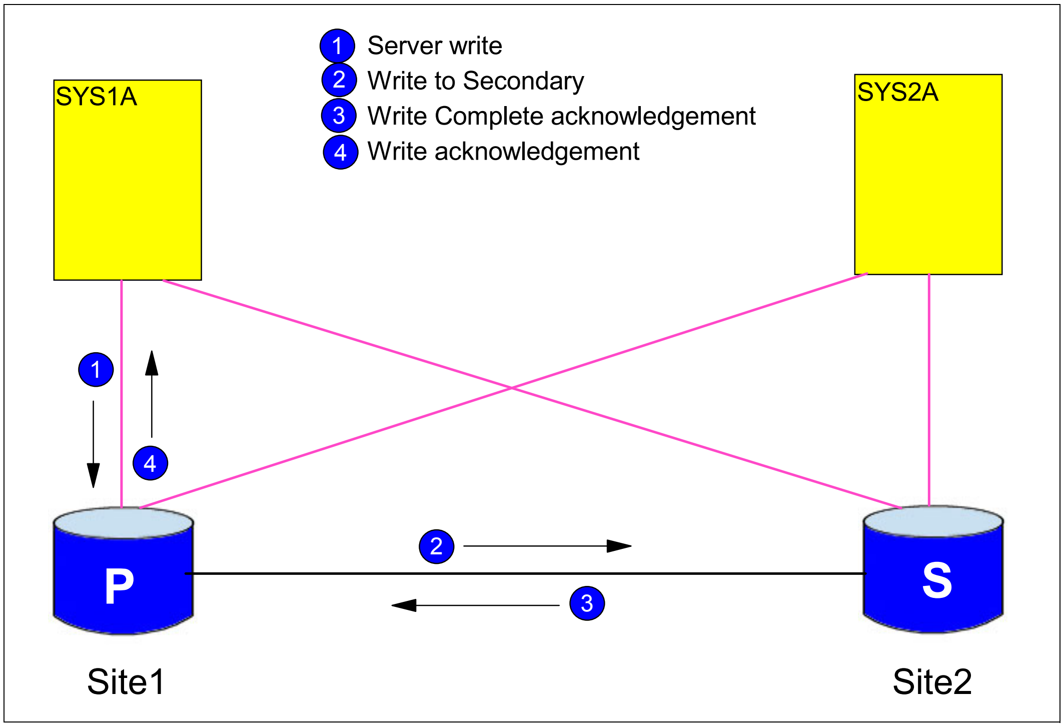


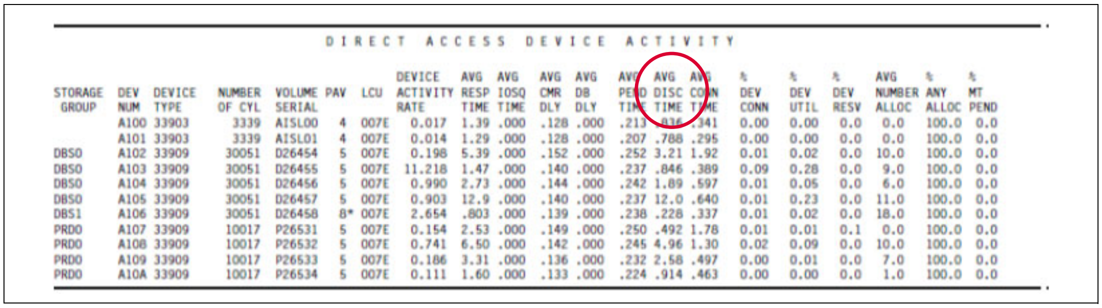
Figure A-7 Synchronous sequence

Increasing the distance between primary and secondary disks will inject some delay in your I/O operation. Actually, just turning on the remote copy relationship adds about 0.1 ms to the response time, assuming that the secondary disk is physically next to the primary disk.

The round trip time for light in a fiber is about 10 μs/km, so moving your secondary disk 10 kilometers from the primary disk would add a further 100 μs to the I/O response times.

DISC time will typically be a major component of disk response times in an extended distance configuration. The longer the distance between your primary and secondary disks, the higher the time that data takes to travel between them. There might be a performance effect on your applications because they cannot proceed until the write to the secondary device completes.

If the increase in DISC time exceeds 10 µs/km between the primary and secondary disk subsystems, it is important to determine the cause of the additional delay. You can use the RMF Direct Access Device Activity report to identify the DISC time for each device, as shown in Figure A-8.



```
                                         D I R E C T   A C C E S S   D E V I C E   A C T I V I T Y

                                                      DEVICE   AVG  AVG   AVG  AVG   AVG  AVG  AVG   %      %     %     AVG    %     %
STORAGE  DEV  DEVICE   NUMBER  VOLUME PAV  LCU  ACTIVITY RESP IOSQ  CMR  DB    PEND DISC CONN  DEV    DEV   DEV   NUMBER ANY   MT
 GROUP   NUM  TYPE     OF CYL  SERIAL               RATE   TIME TIME  DLY  DLY   TIME TIME TIME  CONN   UTIL  RESV  ALLOC  ALLOC PEND
         A100 33903      3339  AISL00  4   007E   0.017  1.39 .000  .128 .000  .213 .836 .341  0.00   0.00   0.0    0.0  100.0  0.0
         A101 33903      3339  AISL01  4   007E   0.014  1.29 .000  .128 .000  .207 .788 .295  0.00   0.00   0.0    0.0  100.0  0.0
DBS0     A102 33909     30051  D26454  5   007E   0.198  5.39 .000  .152 .000  .252 3.21 1.92  0.01   0.02   0.0   10.0  100.0  0.0
DBS0     A103 33909     30051  D26455  5   007E  11.218  1.47 .000  .140 .000  .237 .846 .389  0.09   0.28   0.0    9.0  100.0  0.0
DBS0     A104 33909     30051  D26456  5   007E   0.990  2.73 .000  .144 .000  .242 1.89 .597  0.01   0.05   0.0    6.0  100.0  0.0
DBS0     A105 33909     30051  D26457  5   007E   0.903  12.9 .000  .140 .000  .237 12.0 .640  0.01   0.23   0.0   11.0  100.0  0.0
DBS1     A106 33909     30051  D26458  8*  007E   2.654  .803 .000  .139 .000  .238 .228 .337  0.01   0.02   0.0   18.0  100.0  0.0
PRD0     A107 33909     10017  P26531  5   007E   0.154  2.53 .000  .149 .000  .250 .492 1.78  0.01   0.01   0.1    0.0  100.0  0.0
PRD0     A108 33909     10017  P26532  5   007E   0.741  6.50 .000  .142 .000  .245 4.96 1.30  0.02   0.09   0.0   10.0  100.0  0.0
PRD0     A109 33909     10017  P26533  5   007E   0.186  3.31 .000  .136 .000  .232 2.58 .497  0.00   0.01   0.0    7.0  100.0  0.0
PRD0     A10A 33909     10017  P26534  5   007E   0.111  1.60 .000  .133 .000  .224 .914 .463  0.00   0.00   0.0    1.0  100.0  0.0
```

Figure A-8 RMF Device Activity report showing DISC time

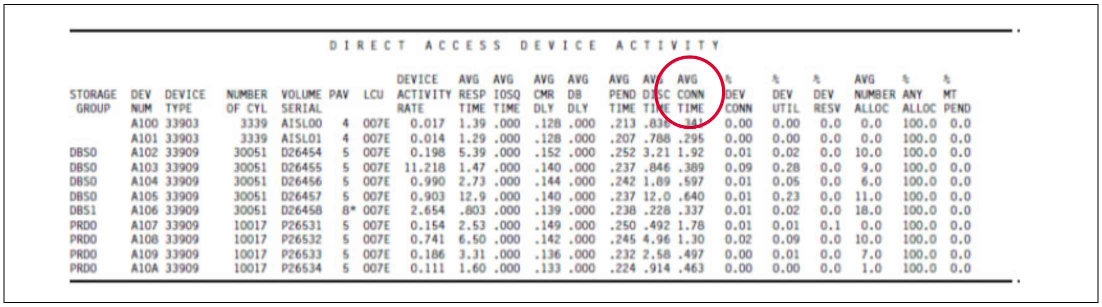
If the DISC time is higher than you expect, the RMF Cache Activity report and the RMF ESS Rank Statistics report can help determine the cause of the delay. The points you need to consider are the average disconnect time, cache hit ratio, and HDD response time. The IntelliMagic Vision product is able to break out the DISC time and show each component separately.

Summary: In a multisite sysplex, you would expect to see the DISC time increase on all systems that write to mirrored disks. If the increase is more than 10 µs/km between the disk subsystems, some other bottleneck is causing a delay and should be investigated.

Connect time

Connect time is the time when the host channel is transferring data from or to the disk subsystem cache, or exchanging protocol control information about an I/O operation. When there is a high level of usage of resources, such as the host channel, director port, or disk subsystem port, significant time can be spent in contention and management, rather than transferring data.

CONN time is influenced by the amount of data being transferred, the speed and the load of disk host adapter, the channel speed and usage, the director load, and remote copy activities. AVG CONN TIME is included in the RMF Direct Access Device Activity report, as shown in Figure A-9.



```
                                         D I R E C T   A C C E S S   D E V I C E   A C T I V I T Y

                                                      DEVICE   AVG  AVG   AVG  AVG   AVG  AVG  AVG   %      %     %     AVG    %     %
STORAGE  DEV  DEVICE   NUMBER  VOLUME PAV  LCU  ACTIVITY RESP IOSQ  CMR  DB    PEND DISC CONN  DEV    DEV   DEV   NUMBER ANY   MT
 GROUP   NUM  TYPE     OF CYL  SERIAL               RATE   TIME TIME  DLY  DLY   TIME TIME TIME  CONN   UTIL  RESV  ALLOC  ALLOC PEND
         A100 33903      3339  AISL00  4   007E   0.017  1.39 .000  .128 .000  .213 .836 .341  0.00   0.00   0.0    0.0  100.0  0.0
         A101 33903      3339  AISL01  4   007E   0.014  1.29 .000  .128 .000  .207 .788 .295  0.00   0.00   0.0    0.0  100.0  0.0
DBS0     A102 33909     30051  D26454  5   007E   0.198  5.39 .000  .152 .000  .252 3.21 1.92  0.01   0.02   0.0   10.0  100.0  0.0
DBS0     A103 33909     30051  D26455  5   007E  11.218  1.47 .000  .140 .000  .237 .846 .389  0.09   0.28   0.0    9.0  100.0  0.0
DBS0     A104 33909     30051  D26456  5   007E   0.990  2.73 .000  .144 .000  .242 1.89 .597  0.01   0.05   0.0    6.0  100.0  0.0
DBS0     A105 33909     30051  D26457  5   007E   0.903  12.9 .000  .140 .000  .237 12.0 .640  0.01   0.23   0.0   11.0  100.0  0.0
DBS1     A106 33909     30051  D26458  8*  007E   2.654  .803 .000  .139 .000  .238 .228 .337  0.01   0.02   0.0   18.0  100.0  0.0
PRD0     A107 33909     10017  P26531  5   007E   0.154  2.53 .000  .149 .000  .250 .492 1.78  0.01   0.01   0.1    0.0  100.0  0.0
PRD0     A108 33909     10017  P26532  5   007E   0.741  6.50 .000  .142 .000  .245 4.96 1.30  0.02   0.09   0.0   10.0  100.0  0.0
PRD0     A109 33909     10017  P26533  5   007E   0.186  3.31 .000  .136 .000  .232 2.58 .497  0.00   0.01   0.0    7.0  100.0  0.0
PRD0     A10A 33909     10017  P26534  5   007E   0.111  1.60 .000  .133 .000  .224 .914 .463  0.00   0.00   0.0    1.0  100.0  0.0
```

Figure A-9 RMF Device Activity report showing CONN time

CONN time can be increased if the channel usage at the host exceeds 50%. In FICON channels, the data being transmitted is divided into frames. When the channel is busy with multiple I/O requests, the frames from one I/O request can be multiplexed with the frames of other I/Os, therefore elongating the time that it takes to transfer all frames that belong to that I/O. The total time, including the transmission time of the other multiplexed frames, is counted as CONN time. This phenomenon is called *FICON elongation*.

The usage of the components in the I/O path tend to be higher as the distance increases. This results in higher response times, which in turn means that the I/O takes longer to finish. Also, as the distance between the CEC and disks increases, more frames will be in flight in the cable. More unfinished I/Os and more frames in flight mean more multiplexing and overhead in the channel, elongating the CONN time and increasing the channel usage.

Decreasing the number of I/Os (SSCHs), even if you transfer the same total amount of data, will consistently decrease total connect time because fewer I/O operations means less overhead.

Modified Indirect Data Address Word (MIDAW) and High Performance FICON for System z (zHPF), features implemented by z/OS and the disk subsystem, improve the efficiency of the I/O, transferring more data per I/O and reducing resource usage. We suggest using these features if you plan an extended end-to-end solution. More information about zHPF and MIDAW is provided later in this Appendix.

Buffer credits

Buffer credits provide a flow control mechanism for FICON channels. In a configuration where all components are within a single data center and relatively close to each other, it is likely that you are completely unaware of the existence of buffer credits or the effect that they have on performance.

However, as the distance between the CEC and the attached disk or tape subsystem increases, buffer credits are likely to play an increasingly important role in the levels of performance and efficiency that you can drive from a given configuration. To help you identify a buffer credit shortage situation, and to provide you with the information you need to perform effective buffer credit planning, this section provides detailed information about how buffer credits work.

We believe this will be valuable information for anyone responsible for performance in an extended distance configuration, and also for the individuals responsible for planning and implementing the extended distance equipment.

How buffer credits work

Buffer credits are used to control the flow of frames between two adjacent ports in a data path. Normally buffer credits are only an issue for extended distances. The number of buffer credits that are supported by the hardware determines the distance that the two nodes can be apart and still maintain the supported link data rate.

Buffer credit flow control occurs at the link level:

- Between two directly attached node ports (N_ports), CHPID *and* CU
- Between an N_Port (CHPID *or* CU) and a fabric port (F_port) or SAN Port
- Between two extension ports (E_ports), inter-switch links (ISLs)

Buffer-to-buffer credits are the mechanism of flow control implemented by the FCP architecture.

FICON ports exchange communication parameters when they connect to each other during link initialization. The key parameter for the purposes of this book is the number of buffer credits. Each port tells the other port how many buffers it has to hold frames received from the other port (these are called *receive buffers*). Each receive buffer is equal to one buffer credit.

So, if a port has 10 receive buffers to store frames, it tells the transmitting port on the other end of the cable that it has 10 buffer credits. Note that each port can have different values. This information is stored in a counter in each port. Figure A-10 illustrates a configuration with one channel, one switch, and one CU. Note that the number of buffer credits for each port can be different.

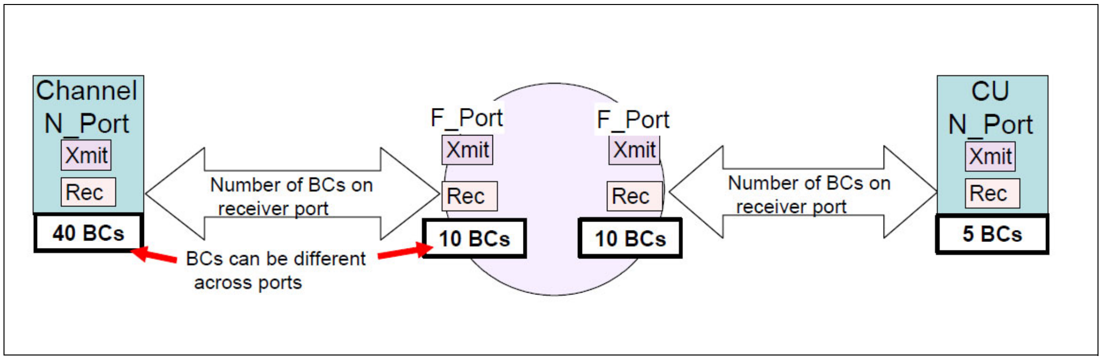

Channel N_Port
Xmit
Rec
40 BCs
Number of BCs on receiver port
BCs can be different across ports
F_Port
Xmit
Rec
10 BCs
F_Port
Xmit
Rec
10 BCs
Number of BCs on receiver port
CU N_Port
Xmit
Rec
5 BCs

Figure A-10 Buffer credit illustration

ESCON also relied on a buffer credit mechanism to control the flow of frames between two ports. However, we do not include ESCON use of buffer credits because ESCON channels are not supported on current CEC technology.

If a port has been told that the port on the other end of the cable has 10 buffer credits, that port can only transmit 10 frames before it needs an acknowledgement from the receiving port. A FICON Express card has a fixed number of buffer credits per CHPID.

FICON Express8S and FICON Express8 features have 40 buffer credits, equivalent to about 80 kilobytes (KB) of buffers at the receiver. Therefore, the FICON Express8S and FICON Express8 receivers can store up to 40 frames of data at any one time. Other FICON features support other numbers of buffer credits:

- FICON Express4 with 212 buffer credits
- FICON Express2 with 107 buffer credits
- FICON Express with 64 buffer credits

FICON directors typically have a pool of buffer credits, and can assign differing numbers of buffer credits to different ports as required. The total number of buffer credits available depends on the ASIC the ports are attached on. Storage devices, especially DASD, usually have a fixed number of buffer credits per port.

Note: For more details about the specific number of available buffer credits, see the product documentation for your disk subsystems and Switch/Directors, or contact the vendor.

When a port transmits a frame, it decrements its buffer credits count. For example, if a port has a buffer credit count of 10 and 1 frame is transmitted, the buffer credit count is decremented to 9. If another frame is transmitted, the count is decremented to 8, and so on until the buffer credit count reaches zero. At that point, all transmissions to the remote port will stop. To transmit more frames, the transmitting port must receive an acknowledgment (FC Receiver-Ready (R_RDY)) from the receiving port.

Figure A-11 shows that 3 frames have been sent from the channel to the switch. Because the switch has 10 buffer credits, the initial buffer counter in the channel was set to 10. Similarly, the buffer counter for the outgoing port on the switch is set to 5 because that is the number of buffers supported by the control unit. The buffer counter in the channel has been decremented three times and is now set to 7. However, because none of the frames have reached the switch yet, the buffer counter in the switch is still set to 10.

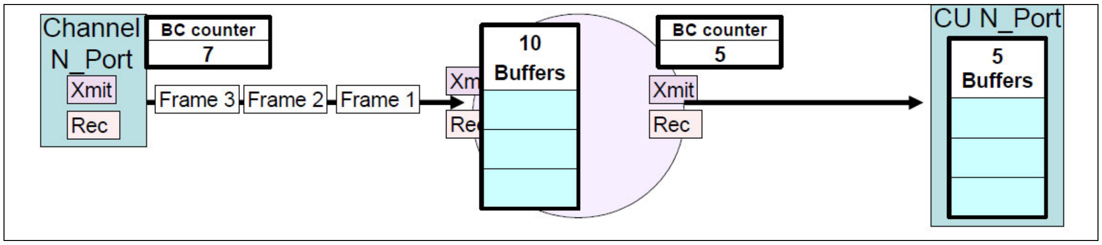


Figure A-11 Buffer credits counters

Each frame is placed into a buffer as it is received on the receiving port. If a second frame is received, it is placed into a second buffer. As the port processes and moves frames to the next step, it releases the buffer and sends an R_RDY acknowledgment to the transmitting port. When the transmitting port receives the R_RDY, it increments the buffer credit count by one.

In Figure A-12, Frame 1 has been received by the switch and been forwarded to the port that is connected to the CU. The buffer count for that port has been decremented to 4. On the port that is connected to the channel, the R_RDY has been sent to the channel (because the frame was forwarded to the other switch port), and the buffer count is now 10 again (because Frame 2 has not arrived yet). On the channel, the buffer count has been increased to 8 because the R_RDY was received from the switch.

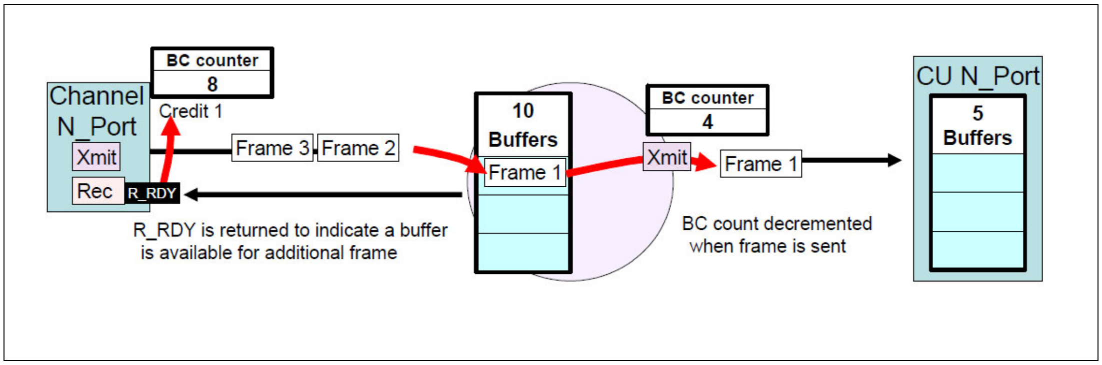


Figure A-12 Buffer credit R_RDY

Buffer credits only play a direct role in the communication between a pair of ports. The buffer count is decremented and incremented based on the interaction between the ports at each end of the cable. In this example, the number of buffer credits in the switch transmitting port and the CU do not directly affect the number of buffers being used for interactions between the channel and the switch.

However, they have a strong indirect affect on the overall throughput that can be achieved from a given configuration. If the port cannot process the frames for any reason, and therefore does not release the buffers, it will not send R_RDY back to the transmitting port. Without the R_RDY, the transmitting port will eventually decrement the buffer credit count until it reaches 0, and will stop transmitting frames.

With this flow control mechanism, the transmitting port can never send more frames than there are available buffers on the receiving port. This is another important reason why your connectivity planning must consider the complete end-to-end path.

Note that an N_Port (CHPID or CU) always tells the port that it is directly connected to how many buffer credits it has. In the case of a switched connection between a CHIPD and a CU, the flow control always happens between one of these end nodes and the switches or director's port. A different mechanism, called End-to-End credits, is used to control the communication between the two end nodes; that is, between the channel and the CU.

A buffer credit is a frame count and not a data size. A 1-byte frame uses one buffer credit and a 2000-byte frame also uses one buffer credit. Two primary factors contribute to the buffer credit count reaching zero:

- Congestion
- Distance

> **Remember:** If the ISL between the switches uses Fibre Channel over IP (FCIP), the control of the frame flow between the switches is based on Transmission Control Protocol/Internet Protocol (TCP/IP), so buffer credits will *not* be used for that part of the transfer. The specifics of TCP/IP flow control is beyond the scope of this book. However, the fact that the transfer of frames when using FCIP does not rely on buffer credits is one of the reasons why FCIP supports much larger distances than FCP.

Congestion

Network or SAN congestion can affect ports upstream. Consider a switch that has four ingress FICON ports, all using the same single ISL to get to the next switch. If all of the ports on this switch are 8 gigabits per second (Gbps) ports, there is potential for 32 Gbps of ingress data needing to use a single 8 Gbps ISL. Although this is not a good SAN design, this does illustrate a real congestion point in the SAN.

The switch will prevent the ingress ports from overwhelming the ISL port, meaning that the ingress ports will eventually start to fill up their receive buffers. This will cause them to stop sending R_RDYs back to the transmitting ports. If this continues, the buffer credits counters will decrement to zero, and the ports that are transmitting to the switch will stop sending frames.

Distance

As the length of the link between two ports increases, the number of buffer credits needs to increase to fully use the link. This is because the longer the link, the more frames can be in flight between the transmitting port and the receiving port. The number of buffer credits needed to fully use the link depends on several factors:

- Actual circuit distance between nodes
- Link speed
- Frame size

Actual circuit distance between the nodes

The FICON frame size is 2048 bytes, including the 32 header bytes. Considering 8/10 encoding, the maximum FICON frame is 20,480 bits. At a link speed of 4 Gbps, a full frame is a little bit more than 1 km long. This is based on the time it takes to move the frame, bit by bit, onto the circuit. When the last bit is being moved into the cable, the first bit of the frame has traveled 1 km down the link. If the link is 10 km long, you can have 10 frames in flight.

Now consider an 8 Gbps link. Because the link speed is twice as fast as the previous example, it takes half as much time to move the frame onto the circuit, meaning that a full frame is only 0.5 km long. As a result, you can have 20 frames in flight on a 10 km link. The longer the distance between the two links, and the faster the link, the more buffer credits are needed to fully use the bandwidth in a cable.

Table A-1 illustrates the relationship between link bandwidth and the number of buffer credits that are required to keep the link fully used, based on a full frame size of 2048 bytes.

Table A-1 Number of in-flight frames

Link speed	Length of each full frame (km)	Number of full frames in a 10 km cable	Number of full frames in a 20 km cable
4 Gbps	1	10	20
8 Gbps	0.5	20	40

Link speed

As the link speed increases, the number of buffer credits needs to increase to maintain efficient link usage. Generally, you need twice as many buffer credits, plus 20%, as the number of frames that can be in-flight on the link. This supports the sum of the following times:

- Time for the frames to reach the other end of the link
- Time for the receiving port to process the frames
- The return time for the R_RDY to come back to the transmitting port

You can use the following formula to calculate the number of buffer credits needed:

```
# buffer credits needed = 1 + Link Speed in Gbps + (Distance in km / AVG Frame Size)
```

Table A-2 shows the relationship between distance, link speed, and buffer credits necessary to maintain a similar link usage at 10 km and 20 km, assuming a 2048-byte frame size.

Table A-2 Number of buffer credits required at 10 km and 20 km

	10 km		20 km	
Link speed	Frames in flight	Buffer credits	Frames in flight	Buffer credits
4 Gbps	10	20	20	40
8 Gbps	20	40	40	80

> **Tip:** Plan for double the number of buffer credits when connected to a SAN fabric with ISL compression enabled.

Frame size

FICON frames have a maximum size of 2048 bytes, of which 32 bytes is the header. One frame can carry anything from simple control information up to a full frame of application data. Therefore, frames can range in size from 64 bytes to 2048 bytes. Typical z/OS environments do not sustain one pattern for the transfer of data. Normally, reads and writes are mixed and the data block sizes vary.

The z/OS disk-intensive workloads rarely produce full frames. For a 4 KB transfer (online applications), the *average* frame size is 819 bytes. For a 27 KB transfer (batch applications), the average frame size is 1502 bytes. You can use the RMF FICON Director Activity report to determine the average frame size for your environment (see the field AVG FRAME SIZE in the report, as shown in Figure A-13 on page 205).

Note that the buffer credit numbers shown in Table A-2 on page 203 were based on a full frame size of 2048 bytes. As the frame size decreases, more buffer credits are needed to support more frames in flight, and to fully use the available bandwidth.

Table A-3 shows the relationship between frame size, link speed, and buffer credits at a distance of 20 km with the link 100% used, for frame sizes of 819 and 1502 bytes.

Table A-3 Required number of buffer credits by size of frame and link speed at 20 km

Frame Size (bytes)	**% Smaller than a full frame**	**Number of Buffer Credits required for a 4 Gbps link**	**Number of Buffer Credits required for a 8 Gbps link**
819	60%	100	199
1502	28%	56	111

Note that Table A-2 on page 203 shows that you need 80 buffer credits with 8 Gbps links, a distance of 20 km and full frame size of 2048 bytes. However, Table A-3 shows that you would need 200 buffer credits at the same link speed and distance when using a frame size of 819 bytes, which is the average of many online applications. Therefore, 150% more buffer credits are needed to support the frames in flight and to keep the link full.

You can use MIDAW and zHPF to increase the size of FICON frames. MIDAW and zHPF are designed to improve I/O efficiency and resource usage. MIDAW is designed to decrease channel, fabric, and CU overhead by reducing the number of CCW's and frames required to process a given I/O. MIDAW provides significant performance benefits, especially when processing extended format data sets with FICON channels.

The zHPF feature is available on FICON Express2, FICON Express4, FICON Express8, and FICON Express8S FICON cards in a z10 system or later CEC. The zHPF feature builds more efficient CCWs (called transport command words or TCWs) that transfer more data with a single CCW or TCW.

Therefore, when you enable the zHPF feature in your environment, the FC frame payload tends to be larger than when you are using native FICON, decreasing the number of interactions between the FICON channel and the CU, and decreasing the number of frames (and therefore buffer credits) that are required to transfer a given (large) amount of data.

Frame pacing

A frame pacing delay occurs after all buffer credits for a port are exhausted. These delays generally result in longer FICON connect times or longer PEND times that show up on the volumes attached to these links. You can use the RMF FICON Director Activity report to determine if your environment is experiencing frame pacing delay. See “RMF FICON Director Activity report” on page 205 for further details.

Look for a non-zero value in the field AVG FRAME PACING in this report. The field reports the average time (in µs) that a frame had to wait before it could be transmitted. The larger the number in this column, the larger the performance problem.

RMF FICON Director Activity report

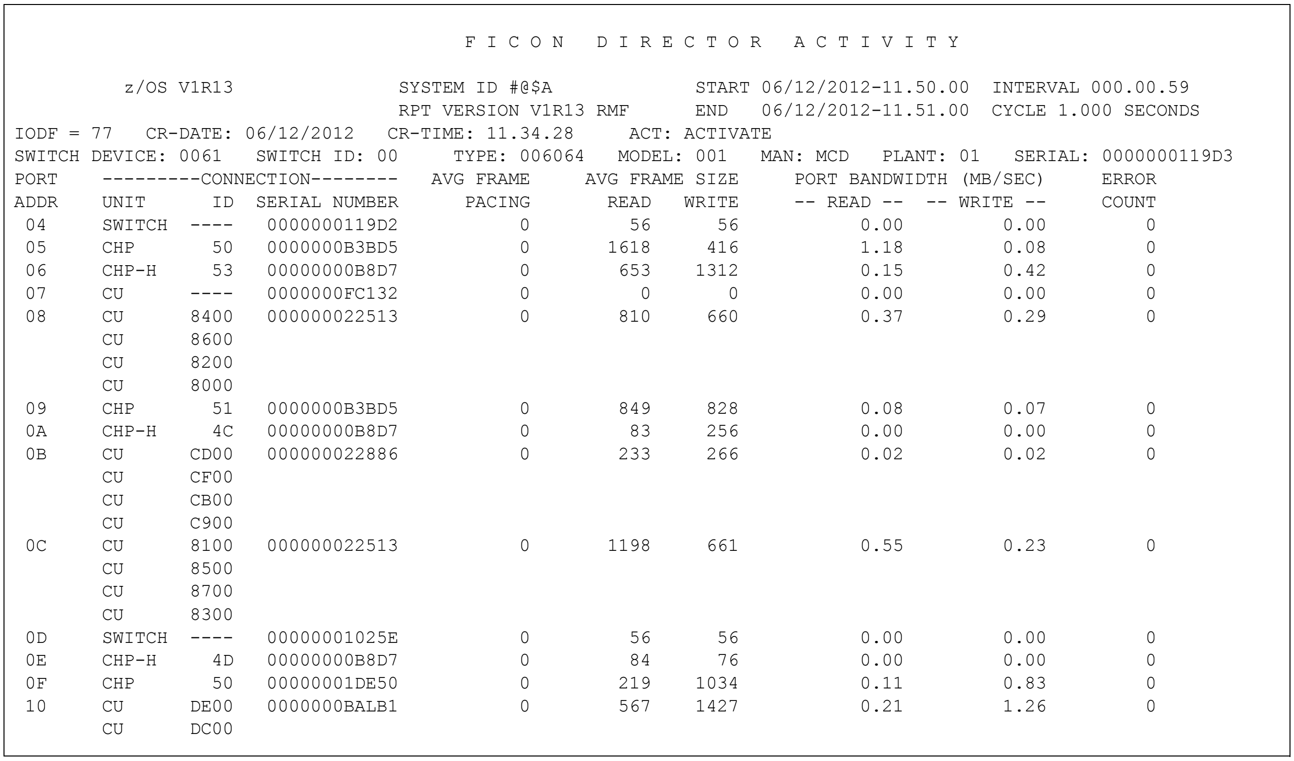
The FICON Director Activity report provides configuration information for connected FICON directors, and details about the connectivity and activity of each director port. The measurements provided for a port include the I/Os for the system on which the report is run, in addition to *all* I/O that is directed through this port, regardless of which LPAR requests the I/O.

FICON **STATS=NO** should be specified in the IECIOSxx member of all but one or maybe two systems in the sysplex. This parameter controls the collection of switch statistics. The statistics contain information for all I/Os processed by that switch, for all systems, so it is not necessary to collect this information on every system in the sysplex. In fact, doing so might cause I/O contention issues. The default is **STATS=YES**, so you need to explicitly turn it off to avoid every system collecting this information.

Additionally, the FCD option in the ERBRMFxx member of parmlib must be enabled for RMF to receive the information. As with the FICON **STATS** parameter, in a sysplex environment you should enable this option in only one system. The data will be recorded in the RMF type 74 subtype 7 System Management Facilities (SMF) records.

You need to have the CUP installed and enabled on your FICON directors, and the device associated with the CUP should be online to whichever system will collect the RMF data. Even if the director is shared between two or more sysplexes, you need to enable this feature in only one sysplex.

The report is created using the RMF Postprocessor, specifying **REPORTS(FCD(*option*))**, where *option* is the device number of the CUP. Figure A-13 shows a sample report for one director and its ports.



```
                                         F I C O N   D I R E C T O R   A C T I V I T Y

          z/OS V1R13               SYSTEM ID #@$A             START 06/12/2012-11.50.00  INTERVAL 000.00.59
                                   RPT VERSION V1R13 RMF      END   06/12/2012-11.51.00  CYCLE 1.000 SECONDS
IODF = 77   CR-DATE: 06/12/2012   CR-TIME: 11.34.28     ACT: ACTIVATE
SWITCH DEVICE: 0061   SWITCH ID: 00     TYPE: 006064   MODEL: 001   MAN: MCD   PLANT: 01   SERIAL: 0000000119D3
PORT    ---------CONNECTION--------   AVG FRAME     AVG FRAME SIZE     PORT BANDWIDTH (MB/SEC)     ERROR
ADDR    UNIT      ID  SERIAL NUMBER      PACING       READ   WRITE     -- READ --  -- WRITE --     COUNT
 04     SWITCH  ----   0000000119D2           0         56      56           0.00         0.00         0
 05     CHP       50   0000000B3BD5           0       1618     416           1.18         0.08         0
 06     CHP-H     53   00000000B8D7           0        653    1312           0.15         0.42         0
 07     CU      ----   0000000FC132           0          0       0           0.00         0.00         0
 08     CU      8400   000000022513           0        810     660           0.37         0.29         0
        CU      8600
        CU      8200
        CU      8000
 09     CHP       51   0000000B3BD5           0        849     828           0.08         0.07         0
 0A     CHP-H     4C   00000000B8D7           0         83     256           0.00         0.00         0
 0B     CU      CD00   000000022886           0        233     266           0.02         0.02         0
        CU      CF00
        CU      CB00
        CU      C900
 0C     CU      8100   000000022513           0       1198     661           0.55         0.23         0
        CU      8500
        CU      8700
        CU      8300
 0D     SWITCH  ----   00000001025E           0         56      56           0.00         0.00         0
 0E     CHP-H     4D   00000000B8D7           0         84      76           0.00         0.00         0
 0F     CHP       50   00000001DE50           0        219    1034           0.11         0.83         0
 10     CU      DE00   0000000BALB1           0        567    1427           0.21         1.26         0
        CU      DC00
```

Figure A-13 Sample RMF FICON Director report

The following list describes the important fields in the report:

CONNECTION

Provides information about the device connected to each port.

The UNIT field contains one of the following connection attributes:

CHP denotes a channel path.
CHP-H denotes a channel path of the system that requested this report.
CU denotes that the port is connected to a CU.
SWITCH denotes a switch.

AVG FRAME PACING

The average time (in µs) that a frame had to wait before it could be transmitted.

AVG FRAME SIZE READ/WRITE

The average frame size (in bytes) used to transmit and receive data during this interval.

PORT BANDWIDTH READ/WRITE

The rate (in megabytes per second (MBps)) of data transmitted and received during the interval.

ERROR COUNT

The number of errors that were encountered during the interval.

> **Important:** The focus of your analysis should be the ISLs, which are identified as a CONNECTION UNIT of SWITCH in the report.

Buffer credit considerations

The buffer credit represents the number of receive buffers supported by a port for receiving frames. The minimum number of buffer credits is one. The number of buffer credits typically doers not affect performance until high data rates are attempted over long distances. If there are insufficient buffer credits, there might be a hard limit on the data rate that can be sustained.

The number of buffer credits available for each port on the FICON director is implementation-dependent. The optimal amount of buffer credits is determined by the distance, the processing time at the receiving port, the link data rate, and the frame size.

Consider these four implications when planning buffer credit allocation:

- Ports do not negotiate buffer credits down to the lowest common value. A receiver advertises buffer credits to the linked transmitter.
- The exhaustion of buffer credits at any point between an initiator and a target limits the performance of the entire path.
- For write-intensive applications using an ISL (tape and disk replication), the buffer credit value advertised by the E_port on the target-cascaded FICON director is the major factor that limits performance.
- For read-intensive applications using an ISL (regular transactions), the buffer credit value advertised by the E_port on the local FICON director is the major factor that limits performance.

Buffer credits are a concern mainly for extended distances. However, a poorly designed configuration can use all available buffer credits and affect performance. For example, assume you have a FICON 8 Gbps channel attached to two different CUs running at lower link rates of 4 Gbps and 2 Gbps.

Depending on the traffic pattern, it is possible for the lower-speed device to use all of the available buffer credits of the 8 Gbps link, stopping more packets from being sent between the server and FICON director, resulting in the average link usage going down.

This scenario is also applicable in the case where the FICON 8 Gbps channel is attached to the CU through an ISL. The slower devices will use all of the buffer credits available in the ISLs, and the faster devices will suffer buffer credit starvation.

Also, if you have applications that use small FC frames and share an ISL, it is possible that this application uses all of the buffer credits. For example, CTC traffic usually uses very small block sizes, commonly 80 bytes, creating small FC frames on the fiber. If you mix this application with disks and tapes, which typically use larger blocks and frames, these devices might suffer buffer credits starvation.

This is because the small CTC FC frames use all of the buffer credits available in the ISLs. For more information about frame sizing effects on buffer credits, see “Frame size” on page 203.

Do not mix small-frame applications with large-frame applications, such as disks and tapes, in the same ISL. Isolate CTC traffic whenever possible.

Important: FICON and FCP traffic must not share the same ISLs. Use Traffic Isolation (TI) zones to direct traffic to different ISLs. Consult your director vendor representative for more information.

Related disk and z/OS features

There are several recent enhancements to z/OS, System z I/O architecture, and disk subsystem support that are particularly relevant to an extended distance configuration. This section describes these features.

PAV and HyperPAV

As covered in “IOSQ time” on page 194, the increased disk response times in an extended distance configuration can cause higher UCB usage, resulting in high IOSQ times.

As an example, we will use a disk subsystem that is 20 meters (m) from a CEC with an I/O response time of 1000 µs. Therefore, the UCB associated with the disk device will be busy for 1000 µs while the physical device will be busy for some smaller amount of time (say 700 µs).

If you move that disk subsystem 50 km from the CEC, the additional distance increases response time by 500 µs. Therefore, the UCB associated with the device will also be busier for 500 µs longer. However, the device will still only be busy for 700 µs. So the limit on how many I/Os can be handled by the device is actually the UCB, rather than the device.

However, if you had two UCBs for the device, you could initiate a second I/O before the first one completes, eliminating the wait for the UCB to become free. This is accomplished by using PAVs. PAVs allow a single physical disk device to be represented by more than one UCB, enabling multiple I/Os to be in process to the same device at the same time.

Because longer distances drive up UCB usage, using PAVs is beneficial when there is a considerable distance between the system and the primary disk devices, or between the primary and secondary disk devices. As you plan an extended solution, you should strongly consider using PAV or HyperPav if you are not already using this capability. You can use IntelliMagic’s Direction and Vision modelling and analysis tools to predict the number of PAV or HyperPav aliases required for your configuration.

MIDAW

The number of interactions between two disk subsystems, or between the CEC and the primary disk subsystem, has a greater negative effect as the distance between the two devices increases. Anything you can do to reduce the number of I/Os between the devices will improve performance.

One option is to use larger block sizes, where many records can be sent and retrieved with a single block. This enables the same number of bytes to be read or written with fewer interactions between the channel and the control unit.

The MIDAW facility is a method for gathering or scattering data into and from non-contiguous storage locations during the execution of an I/O operation by the channel. CCWs with data chaining were used before MIDAW, but this technique adversely affected FICON channel performance because of the handshake between the FICON channel and the CU.

With MIDAW, you can gather and scatter data from and to the CEC memory using only one CCW. Without MIDAW, for each data scattered in the CEC memory, you need one CCW to read the data from memory and write it to the disk subsystem memory.

The zHPF function

The zHPF function is available on FICON Express2, FICON Express4, FICON Express8, and FICON Express8S FICON cards in a z10 system or later CEC.

The zHPF function reduces FICON channel overhead by using features in the FICON channel, the z/OS operating system, and the CU. These features combine to reduce the number of Information Units (IUs), and therefore the number of handshakes between the channel and the CU. This results in more efficient use of the FICON channel, especially in an extended distance configuration.

With zHPF, the FICON architecture has been streamlined by removing significant resource usage to the disk subsystem and the microprocessor within the FICON channel. A command block is created to chain commands into significantly fewer IUs. The overhead required to convert individual commands into FICON format is removed, because multiple System z I/O commands are packaged together and passed directly over the fiber optic link. One single zHPF command block replaces a series of FICON CCWs.

The zHPF feature builds more efficient CCWs, moving more data than a single CCW or TCW. When you enable zHPF in your environment, the FC frame payload tends to be bigger than when you are using native FICON, decreasing the number of interactions between the FICON channel and the CU.

B

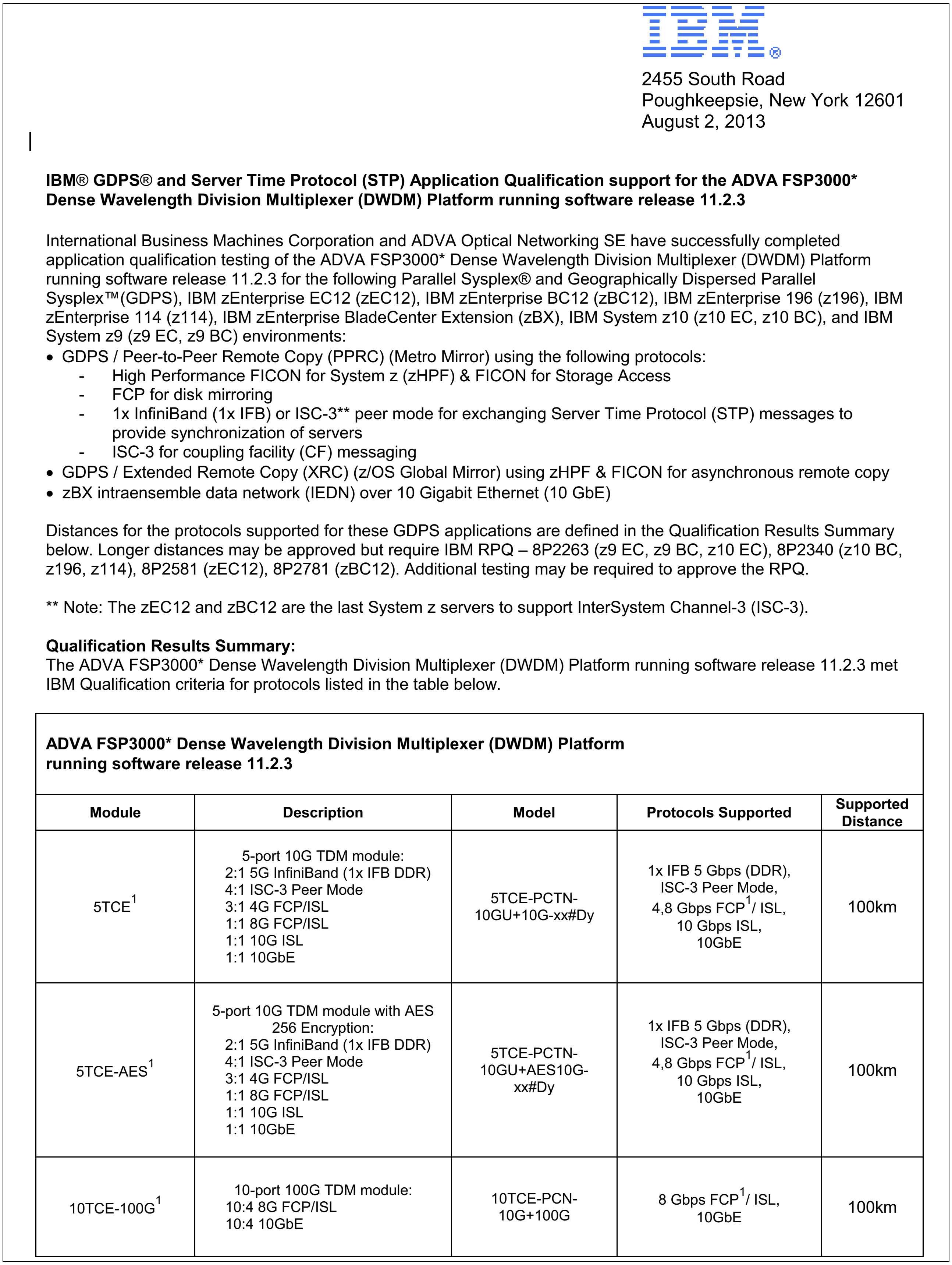
Sample qualification letters

This appendix provides sample qualification letters for a wavelength division multiplexer (WDM) and a switch. It also shows a sample report from the IBM System Storage Interoperation Center (SSIC) website.

Note that you should always get the latest qualification letters from the Resource Link website to ensure that you have the most up-to-date information.

Sample WDM qualification letter

Figure B-1 on page 211, Figure B-2 on page 212, and Figure B-3 on page 213 contain a sample qualification letter for a WDM device. They are provided here to illustrate the information about WDM qualification letter contents in 5.6, “IBM qualification for extended distance devices” on page 152.



2455 South Road
Poughkeepsie, New York 12601
August 2, 2013

IBM® GDPS® and Server Time Protocol (STP) Application Qualification support for the ADVA FSP3000* Dense Wavelength Division Multiplexer (DWDM) Platform running software release 11.2.3

International Business Machines Corporation and ADVA Optical Networking SE have successfully completed application qualification testing of the ADVA FSP3000* Dense Wavelength Division Multiplexer (DWDM) Platform running software release 11.2.3 for the following Parallel Sysplex® and Geographically Dispersed Parallel Sysplex™(GDPS), IBM zEnterprise EC12 (zEC12), IBM zEnterprise BC12 (zBC12), IBM zEnterprise 196 (z196), IBM zEnterprise 114 (z114), IBM zEnterprise BladeCenter Extension (zBX), IBM System z10 (z10 EC, z10 BC), and IBM System z9 (z9 EC, z9 BC) environments:

- GDPS / Peer-to-Peer Remote Copy (PPRC) (Metro Mirror) using the following protocols:
  - High Performance FICON for System z (zHPF) & FICON for Storage Access
  - FCP for disk mirroring
  - 1x InfiniBand (1x IFB) or ISC-3** peer mode for exchanging Server Time Protocol (STP) messages to provide synchronization of servers
  - ISC-3 for coupling facility (CF) messaging
- GDPS / Extended Remote Copy (XRC) (z/OS Global Mirror) using zHPF & FICON for asynchronous remote copy
- zBX intraensemble data network (IEDN) over 10 Gigabit Ethernet (10 GbE)

Distances for the protocols supported for these GDPS applications are defined in the Qualification Results Summary below. Longer distances may be approved but require IBM RPQ – 8P2263 (z9 EC, z9 BC, z10 EC), 8P2340 (z10 BC, z196, z114), 8P2581 (zEC12), 8P2781 (zBC12). Additional testing may be required to approve the RPQ.

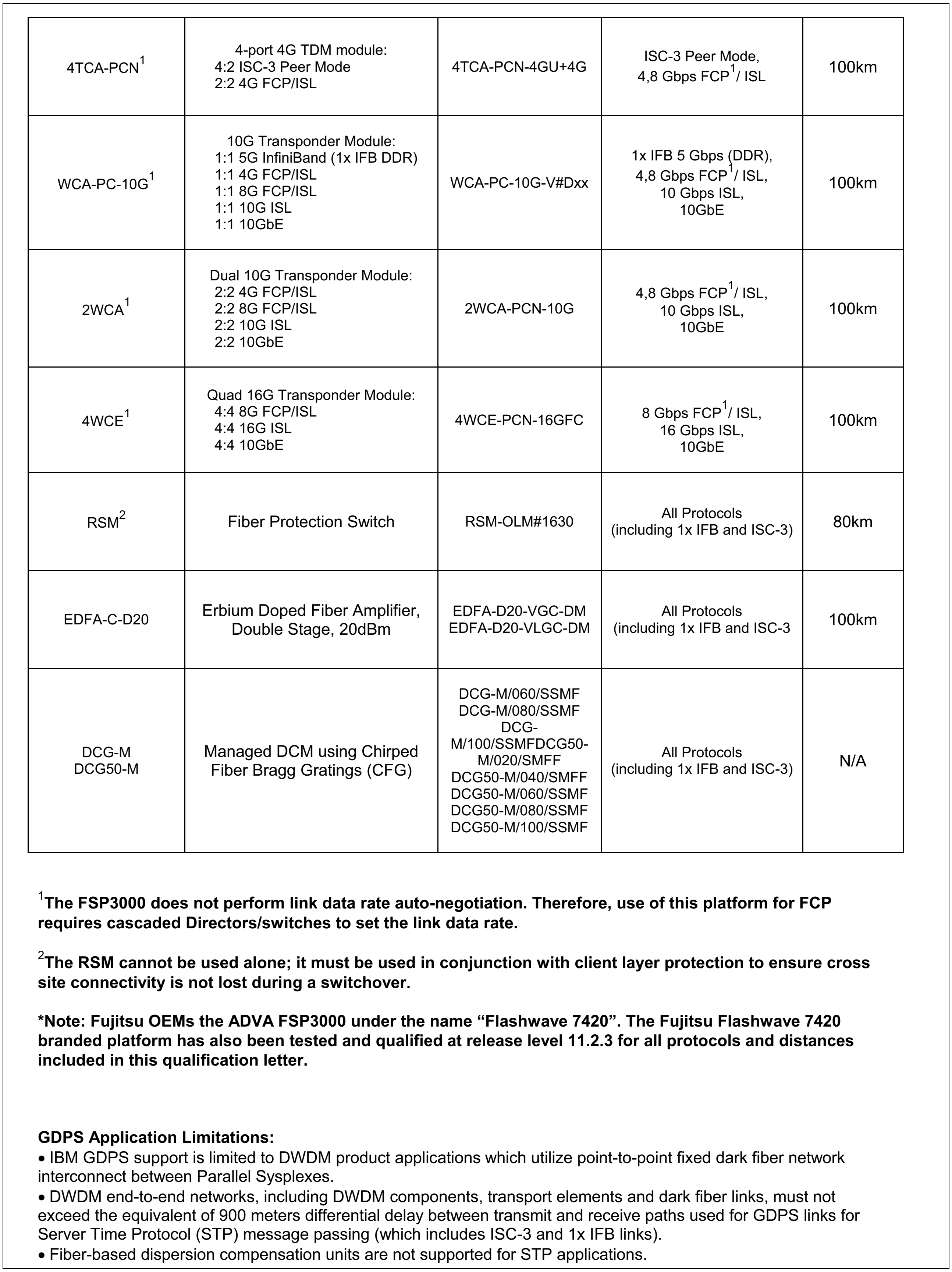
** Note: The zEC12 and zBC12 are the last System z servers to support InterSystem Channel-3 (ISC-3).

Qualification Results Summary:
The ADVA FSP3000* Dense Wavelength Division Multiplexer (DWDM) Platform running software release 11.2.3 met IBM Qualification criteria for protocols listed in the table below.

ADVA FSP3000* Dense Wavelength Division Multiplexer (DWDM) Platform running software release 11.2.3

Module	Description	Model	Protocols Supported	Supported Distance
5TCE[1]	5-port 10G TDM module: 2:1 5G InfiniBand (1x IFB DDR) 4:1 ISC-3 Peer Mode 3:1 4G FCP/ISL 1:1 8G FCP/ISL 1:1 10G ISL 1:1 10GbE	5TCE-PCTN-10GU+10G-xx#Dy	1x IFB 5 Gbps (DDR), ISC-3 Peer Mode, 4,8 Gbps FCP[1]/ ISL, 10 Gbps ISL, 10GbE	100km
5TCE-AES[1]	5-port 10G TDM module with AES 256 Encryption: 2:1 5G InfiniBand (1x IFB DDR) 4:1 ISC-3 Peer Mode 3:1 4G FCP/ISL 1:1 8G FCP/ISL 1:1 10G ISL 1:1 10GbE	5TCE-PCTN-10GU+AES10G-xx#Dy	1x IFB 5 Gbps (DDR), ISC-3 Peer Mode, 4,8 Gbps FCP[1]/ ISL, 10 Gbps ISL, 10GbE	100km
10TCE-100G[1]	10-port 100G TDM module: 10:4 8G FCP/ISL 10:4 10GbE	10TCE-PCN-10G+100G	8 Gbps FCP[1]/ ISL, 10GbE	100km

Figure B-1 Sample WDM qualification letter (page 1 of 3)



4TCA-PCN[1]	4-port 4G TDM module: 4:2 ISC-3 Peer Mode 2:2 4G FCP/ISL	4TCA-PCN-4GU+4G	ISC-3 Peer Mode, 4,8 Gbps FCP[1]/ ISL	100km
WCA-PC-10G[1]	10G Transponder Module: 1:1 5G InfiniBand (1x IFB DDR) 1:1 4G FCP/ISL 1:1 8G FCP/ISL 1:1 10G ISL 1:1 10GbE	WCA-PC-10G-V#Dxx	1x IFB 5 Gbps (DDR), 4,8 Gbps FCP[1]/ ISL, 10 Gbps ISL, 10GbE	100km
2WCA[1]	Dual 10G Transponder Module: 2:2 4G FCP/ISL 2:2 8G FCP/ISL 2:2 10G ISL 2:2 10GbE	2WCA-PCN-10G	4,8 Gbps FCP[1]/ ISL, 10 Gbps ISL, 10GbE	100km
4WCE[1]	Quad 16G Transponder Module: 4:4 8G FCP/ISL 4:4 16G ISL 4:4 10GbE	4WCE-PCN-16GFC	8 Gbps FCP[1]/ ISL, 16 Gbps ISL, 10GbE	100km
RSM[2]	Fiber Protection Switch	RSM-OLM#1630	All Protocols (including 1x IFB and ISC-3)	80km
EDFA-C-D20	Erbium Doped Fiber Amplifier, Double Stage, 20dBm	EDFA-D20-VGC-DM EDFA-D20-VLGC-DM	All Protocols (including 1x IFB and ISC-3	100km
DCG-M DCG50-M	Managed DCM using Chirped Fiber Bragg Gratings (CFG)	DCG-M/060/SSMF DCG-M/080/SSMF DCG-M/100/SSMFDCG50-M/020/SMFF DCG50-M/040/SMFF DCG50-M/060/SSMF DCG50-M/080/SSMF DCG50-M/100/SSMF	All Protocols (including 1x IFB and ISC-3)	N/A

[1] The FSP3000 does not perform link data rate auto-negotiation. Therefore, use of this platform for FCP requires cascaded Directors/switches to set the link data rate.

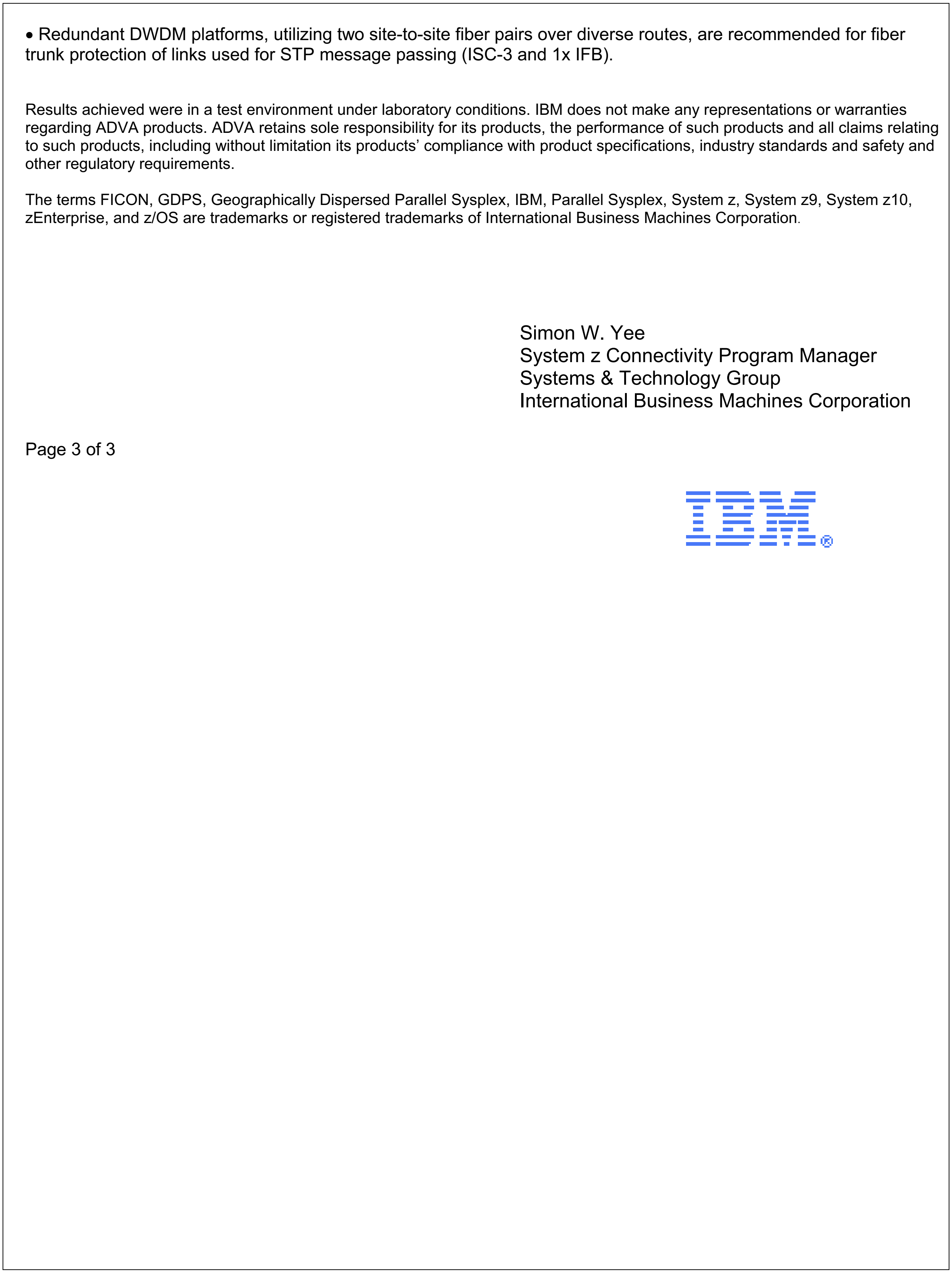
[2] The RSM cannot be used alone; it must be used in conjunction with client layer protection to ensure cross site connectivity is not lost during a switchover.

***Note: Fujitsu OEMs the ADVA FSP3000 under the name “Flashwave 7420”. The Fujitsu Flashwave 7420 branded platform has also been tested and qualified at release level 11.2.3 for all protocols and distances included in this qualification letter.**

GDPS Application Limitations:

- IBM GDPS support is limited to DWDM product applications which utilize point-to-point fixed dark fiber network interconnect between Parallel Sysplexes.
- DWDM end-to-end networks, including DWDM components, transport elements and dark fiber links, must not exceed the equivalent of 900 meters differential delay between transmit and receive paths used for GDPS links for Server Time Protocol (STP) message passing (which includes ISC-3 and 1x IFB links).
- Fiber-based dispersion compensation units are not supported for STP applications.

Figure B-2 Sample WDM qualification letter (page 2 of 3)



- Redundant DWDM platforms, utilizing two site-to-site fiber pairs over diverse routes, are recommended for fiber trunk protection of links used for STP message passing (ISC-3 and 1x IFB).

Results achieved were in a test environment under laboratory conditions. IBM does not make any representations or warranties regarding ADVA products. ADVA retains sole responsibility for its products, the performance of such products and all claims relating to such products, including without limitation its products' compliance with product specifications, industry standards and safety and other regulatory requirements.

The terms FICON, GDPS, Geographically Dispersed Parallel Sysplex, IBM, Parallel Sysplex, System z, System z9, System z10, zEnterprise, and z/OS are trademarks or registered trademarks of International Business Machines Corporation.

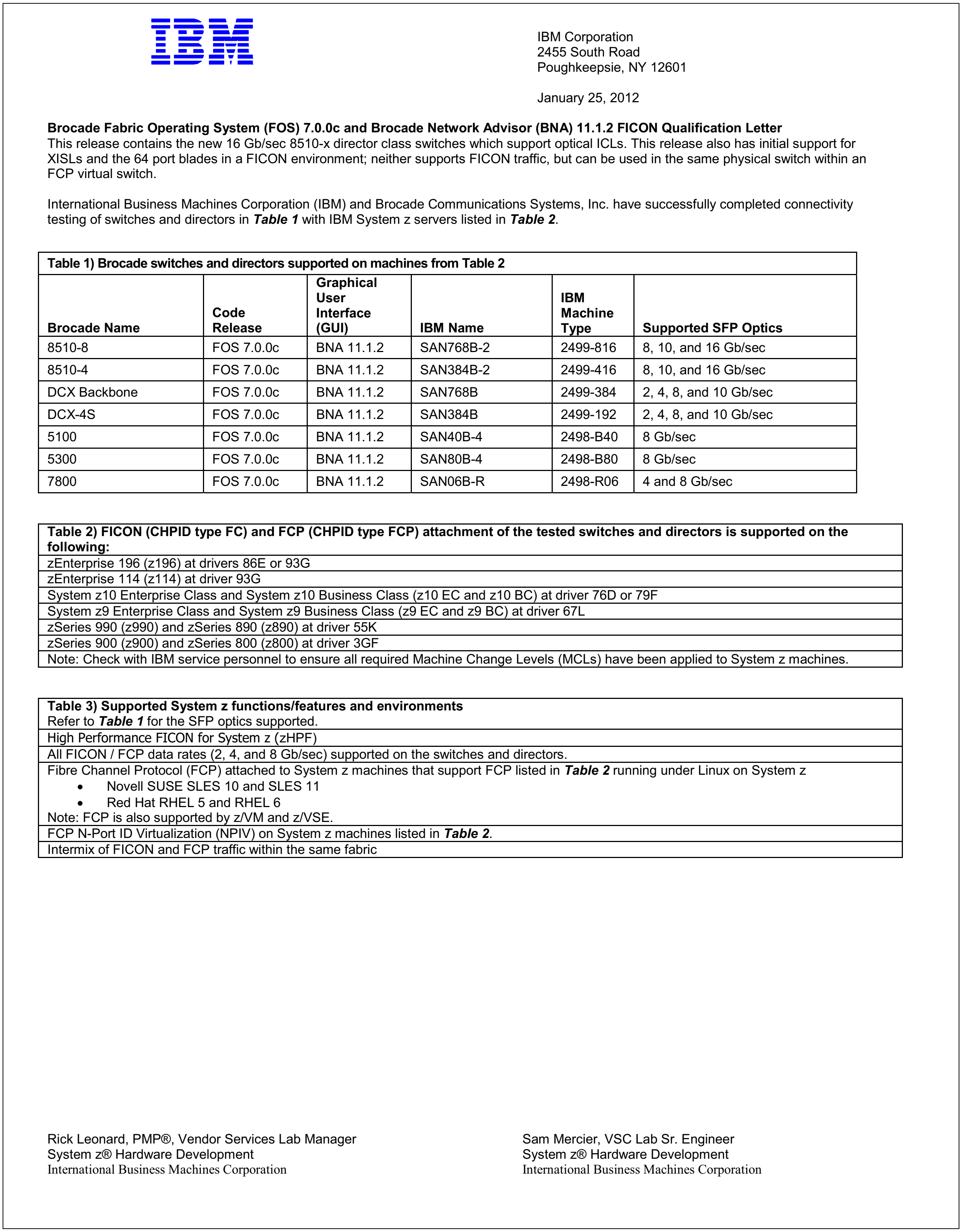
Simon W. Yee
System z Connectivity Program Manager
Systems & Technology Group
International Business Machines Corporation

Page 3 of 3

Figure B-3 Sample WDM qualification letter (page 3 of 3)

Sample switch qualification letter

Figure B-4 on page 215, Figure B-5 on page 216, Figure B-6 on page 217, and Figure B-7 on page 218 contain a sample qualification letter for a switch device. They are provided here to illustrate the information about switch qualification letter contents in 5.6, “IBM qualification for extended distance devices” on page 152.



IBM Corporation
2455 South Road
Poughkeepsie, NY 12601

January 25, 2012

Brocade Fabric Operating System (FOS) 7.0.0c and Brocade Network Advisor (BNA) 11.1.2 FICON Qualification Letter
This release contains the new 16 Gb/sec 8510-x director class switches which support optical ICLs. This release also has initial support for XISLs and the 64 port blades in a FICON environment; neither supports FICON traffic, but can be used in the same physical switch within an FCP virtual switch.

International Business Machines Corporation (IBM) and Brocade Communications Systems, Inc. have successfully completed connectivity testing of switches and directors in ***Table 1*** with IBM System z servers listed in ***Table 2***.

Table 1) Brocade switches and directors supported on machines from Table 2

Brocade Name	Code Release	Graphical User Interface (GUI)	IBM Name	IBM Machine Type	Supported SFP Optics
8510-8	FOS 7.0.0c	BNA 11.1.2	SAN768B-2	2499-816	8, 10, and 16 Gb/sec
8510-4	FOS 7.0.0c	BNA 11.1.2	SAN384B-2	2499-416	8, 10, and 16 Gb/sec
DCX Backbone	FOS 7.0.0c	BNA 11.1.2	SAN768B	2499-384	2, 4, 8, and 10 Gb/sec
DCX-4S	FOS 7.0.0c	BNA 11.1.2	SAN384B	2499-192	2, 4, 8, and 10 Gb/sec
5100	FOS 7.0.0c	BNA 11.1.2	SAN40B-4	2498-B40	8 Gb/sec
5300	FOS 7.0.0c	BNA 11.1.2	SAN80B-4	2498-B80	8 Gb/sec
7800	FOS 7.0.0c	BNA 11.1.2	SAN06B-R	2498-R06	4 and 8 Gb/sec

Table 2) FICON (CHPID type FC) and FCP (CHPID type FCP) attachment of the tested switches and directors is supported on the following:
zEnterprise 196 (z196) at drivers 86E or 93G
zEnterprise 114 (z114) at driver 93G
System z10 Enterprise Class and System z10 Business Class (z10 EC and z10 BC) at driver 76D or 79F
System z9 Enterprise Class and System z9 Business Class (z9 EC and z9 BC) at driver 67L
zSeries 990 (z990) and zSeries 890 (z890) at driver 55K
zSeries 900 (z900) and zSeries 800 (z800) at driver 3GF
Note: Check with IBM service personnel to ensure all required Machine Change Levels (MCLs) have been applied to System z machines.

Table 3) Supported System z functions/features and environments Refer to ***Table 1*** for the SFP optics supported.
High Performance FICON for System z (zHPF)
All FICON / FCP data rates (2, 4, and 8 Gb/sec) supported on the switches and directors.
Fibre Channel Protocol (FCP) attached to System z machines that support FCP listed in ***Table 2*** running under Linux on System z • Novell SUSE SLES 10 and SLES 11 • Red Hat RHEL 5 and RHEL 6 Note: FCP is also supported by z/VM and z/VSE.
FCP N-Port ID Virtualization (NPIV) on System z machines listed in ***Table 2***.
Intermix of FICON and FCP traffic within the same fabric

Rick Leonard, PMP®, Vendor Services Lab Manager
System z® Hardware Development
International Business Machines Corporation

Sam Mercier, VSC Lab Sr. Engineer
System z® Hardware Development
International Business Machines Corporation

Figure B-4 Sample switch qualification letter (page 1 of 4)

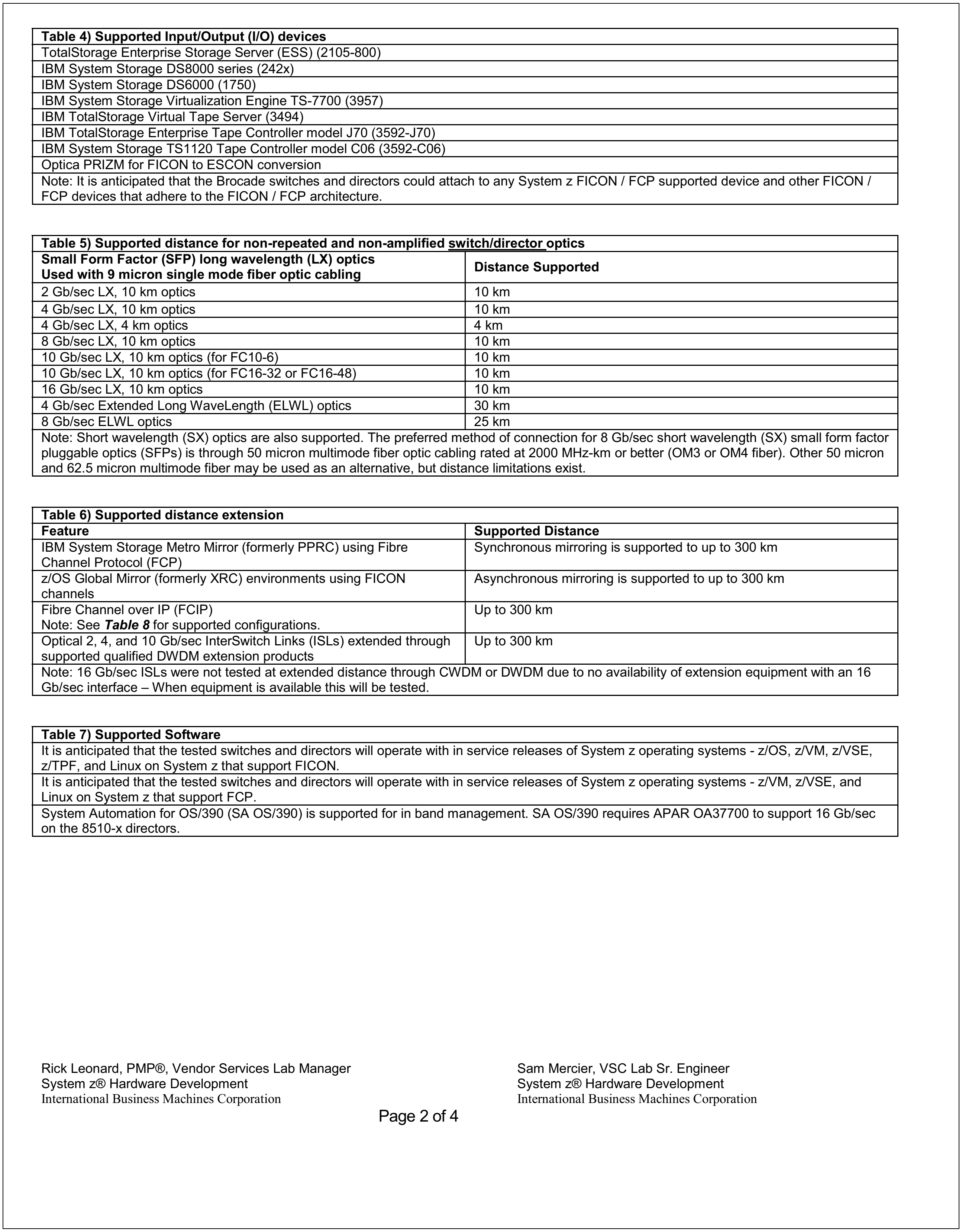


Table 4) Supported Input/Output (I/O) devices
TotalStorage Enterprise Storage Server (ESS) (2105-800)
IBM System Storage DS8000 series (242x)
IBM System Storage DS6000 (1750)
IBM System Storage Virtualization Engine TS-7700 (3957)
IBM TotalStorage Virtual Tape Server (3494)
IBM TotalStorage Enterprise Tape Controller model J70 (3592-J70)
IBM System Storage TS1120 Tape Controller model C06 (3592-C06)
Optica PRIZM for FICON to ESCON conversion
Note: It is anticipated that the Brocade switches and directors could attach to any System z FICON / FCP supported device and other FICON / FCP devices that adhere to the FICON / FCP architecture.

Table 5) Supported distance for non-repeated and non-amplified switch/director optics	
Small Form Factor (SFP) long wavelength (LX) optics **Used with 9 micron single mode fiber optic cabling**	**Distance Supported**
2 Gb/sec LX, 10 km optics	10 km
4 Gb/sec LX, 10 km optics	10 km
4 Gb/sec LX, 4 km optics	4 km
8 Gb/sec LX, 10 km optics	10 km
10 Gb/sec LX, 10 km optics (for FC10-6)	10 km
10 Gb/sec LX, 10 km optics (for FC16-32 or FC16-48)	10 km
16 Gb/sec LX, 10 km optics	10 km
4 Gb/sec Extended Long WaveLength (ELWL) optics	30 km
8 Gb/sec ELWL optics	25 km
Note: Short wavelength (SX) optics are also supported. The preferred method of connection for 8 Gb/sec short wavelength (SX) small form factor pluggable optics (SFPs) is through 50 micron multimode fiber optic cabling rated at 2000 MHz-km or better (OM3 or OM4 fiber). Other 50 micron and 62.5 micron multimode fiber may be used as an alternative, but distance limitations exist.	

Table 6) Supported distance extension	
Feature	**Supported Distance**
IBM System Storage Metro Mirror (formerly PPRC) using Fibre Channel Protocol (FCP)	Synchronous mirroring is supported to up to 300 km
z/OS Global Mirror (formerly XRC) environments using FICON channels	Asynchronous mirroring is supported to up to 300 km
Fibre Channel over IP (FCIP) Note: See ***Table 8*** for supported configurations.	Up to 300 km
Optical 2, 4, and 10 Gb/sec InterSwitch Links (ISLs) extended through supported qualified DWDM extension products	Up to 300 km
Note: 16 Gb/sec ISLs were not tested at extended distance through CWDM or DWDM due to no availability of extension equipment with an 16 Gb/sec interface – When equipment is available this will be tested.	

Table 7) Supported Software
It is anticipated that the tested switches and directors will operate with in service releases of System z operating systems - z/OS, z/VM, z/VSE, z/TPF, and Linux on System z that support FICON.
It is anticipated that the tested switches and directors will operate with in service releases of System z operating systems - z/VM, z/VSE, and Linux on System z that support FCP.
System Automation for OS/390 (SA OS/390) is supported for in band management. SA OS/390 requires APAR OA37700 to support 16 Gb/sec on the 8510-x directors.

Rick Leonard, PMP®, Vendor Services Lab Manager
System z® Hardware Development
International Business Machines Corporation

Sam Mercier, VSC Lab Sr. Engineer
System z® Hardware Development
International Business Machines Corporation

Page 2 of 4

Figure B-5 Sample switch qualification letter (page 2 of 4)

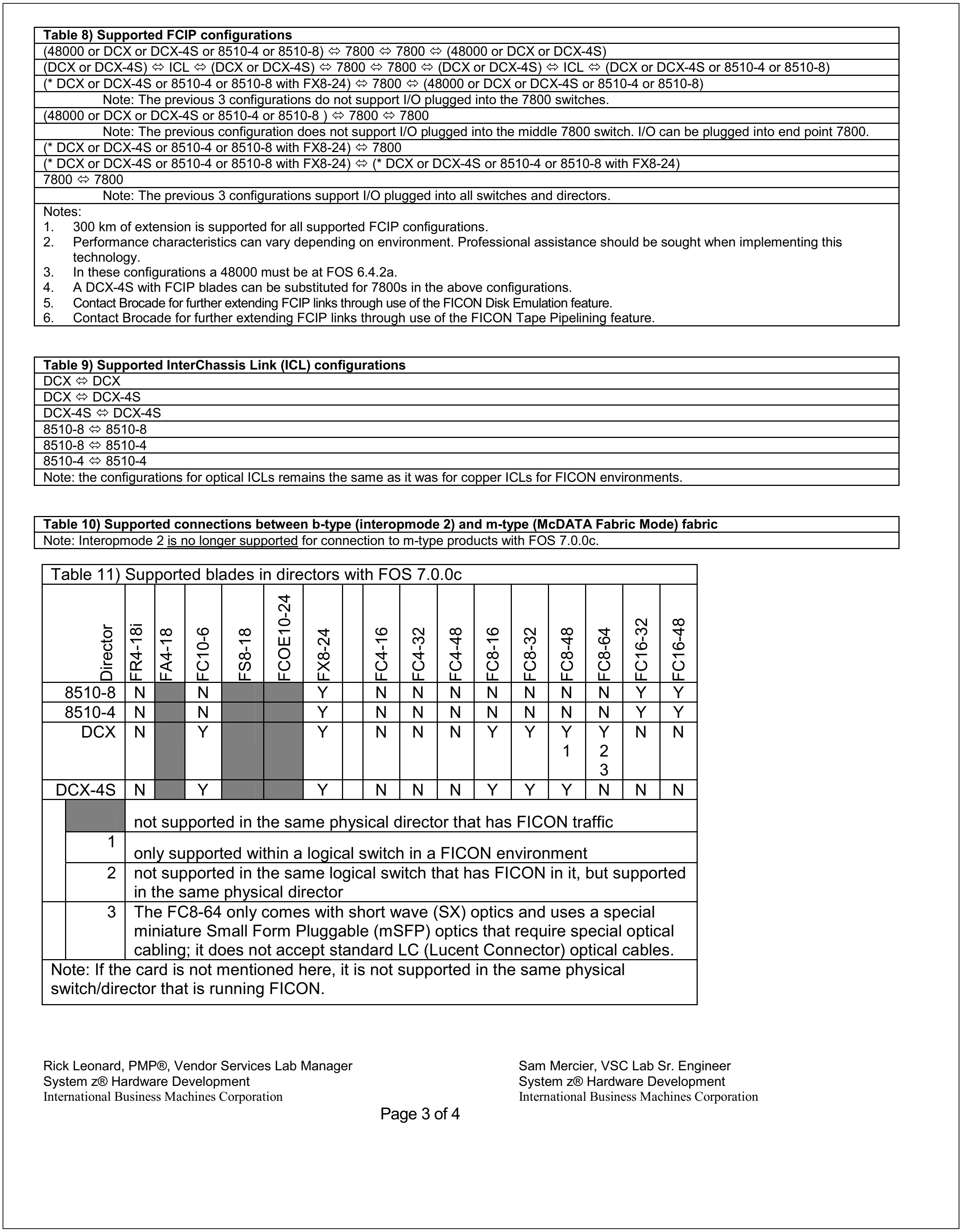


Table 8) Supported FCIP configurations
(48000 or DCX or DCX-4S or 8510-4 or 8510-8) ⇔ 7800 ⇔ 7800 ⇔ (48000 or DCX or DCX-4S)
(DCX or DCX-4S) ⇔ ICL ⇔ (DCX or DCX-4S) ⇔ 7800 ⇔ 7800 ⇔ (DCX or DCX-4S) ⇔ ICL ⇔ (DCX or DCX-4S or 8510-4 or 8510-8)
(* DCX or DCX-4S or 8510-4 or 8510-8 with FX8-24) ⇔ 7800 ⇔ (48000 or DCX or DCX-4S or 8510-4 or 8510-8)
Note: The previous 3 configurations do not support I/O plugged into the 7800 switches.
(48000 or DCX or DCX-4S or 8510-4 or 8510-8) ⇔ 7800 ⇔ 7800
Note: The previous configuration does not support I/O plugged into the middle 7800 switch. I/O can be plugged into end point 7800.
(* DCX or DCX-4S or 8510-4 or 8510-8 with FX8-24) ⇔ 7800
(* DCX or DCX-4S or 8510-4 or 8510-8 with FX8-24) ⇔ (* DCX or DCX-4S or 8510-4 or 8510-8 with FX8-24)
7800 ⇔ 7800
Note: The previous 3 configurations support I/O plugged into all switches and directors.
Notes: 1. 300 km of extension is supported for all supported FCIP configurations. 2. Performance characteristics can vary depending on environment. Professional assistance should be sought when implementing this technology. 3. In these configurations a 48000 must be at FOS 6.4.2a. 4. A DCX-4S with FCIP blades can be substituted for 7800s in the above configurations. 5. Contact Brocade for further extending FCIP links through use of the FICON Disk Emulation feature. 6. Contact Brocade for further extending FCIP links through use of the FICON Tape Pipelining feature.

Table 9) Supported InterChassis Link (ICL) configurations
DCX ⇔ DCX
DCX ⇔ DCX-4S
DCX-4S ⇔ DCX-4S
8510-8 ⇔ 8510-8
8510-8 ⇔ 8510-4
8510-4 ⇔ 8510-4
Note: the configurations for optical ICLs remains the same as it was for copper ICLs for FICON environments.

Table 10) Supported connections between b-type (interopmode 2) and m-type (McDATA Fabric Mode) fabric
Note: Interopmode 2 is no longer supported for connection to m-type products with FOS 7.0.0c.

Table 11) Supported blades in directors with FOS 7.0.0c

Director	FR4-18i	FA4-18	FC10-6	FS8-18	FCOE10-24	FX8-24		FC4-16	FC4-32	FC4-48	FC8-16	FC8-32	FC8-48	FC8-64	FC16-32	FC16-48
8510-8	N	▨	N	▨	▨	Y		N	N	N	N	N	N	N	Y	Y
8510-4	N	▨	N	▨	▨	Y		N	N	N	N	N	N	N	Y	Y
DCX	N	▨	Y	▨	▨	Y		N	N	N	Y	Y	Y 1	Y 2 3	N	N
DCX-4S	N	▨	Y	▨	▨	Y		N	N	N	Y	Y	Y	N	N	N

▨	not supported in the same physical director that has FICON traffic
1	only supported within a logical switch in a FICON environment
2	not supported in the same logical switch that has FICON in it, but supported in the same physical director
3	The FC8-64 only comes with short wave (SX) optics and uses a special miniature Small Form Pluggable (mSFP) optics that require special optical cabling; it does not accept standard LC (Lucent Connector) optical cables.

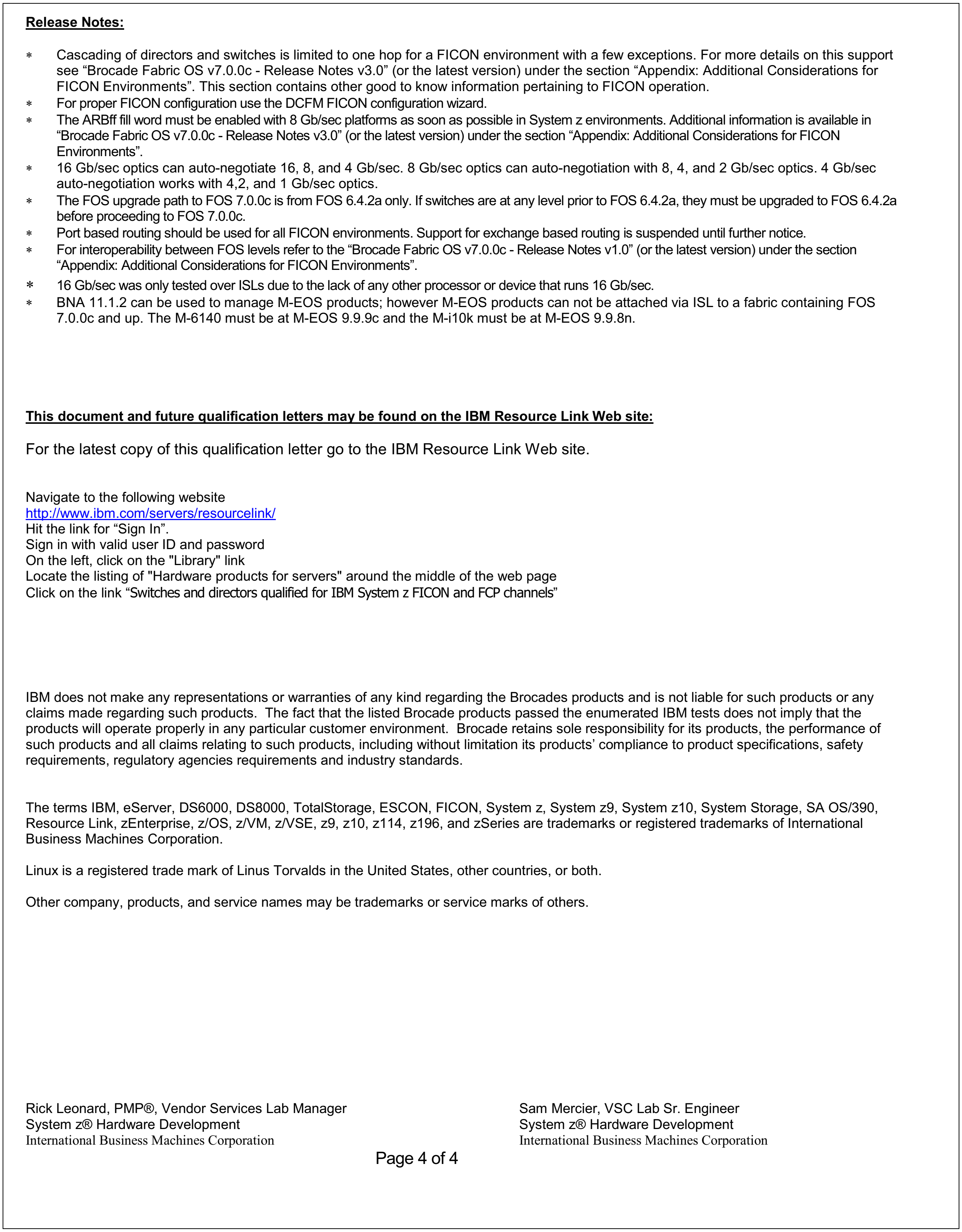
Note: If the card is not mentioned here, it is not supported in the same physical switch/director that is running FICON.

Rick Leonard, PMP®, Vendor Services Lab Manager
System z® Hardware Development
International Business Machines Corporation

Sam Mercier, VSC Lab Sr. Engineer
System z® Hardware Development
International Business Machines Corporation

Page 3 of 4

Figure B-6 Sample switch qualification letter (page 3 of 4)



Release Notes:

- Cascading of directors and switches is limited to one hop for a FICON environment with a few exceptions. For more details on this support see “Brocade Fabric OS v7.0.0c - Release Notes v3.0” (or the latest version) under the section “Appendix: Additional Considerations for FICON Environments”. This section contains other good to know information pertaining to FICON operation.
- For proper FICON configuration use the DCFM FICON configuration wizard.
- The ARBff fill word must be enabled with 8 Gb/sec platforms as soon as possible in System z environments. Additional information is available in “Brocade Fabric OS v7.0.0c - Release Notes v3.0” (or the latest version) under the section “Appendix: Additional Considerations for FICON Environments”.
- 16 Gb/sec optics can auto-negotiate 16, 8, and 4 Gb/sec. 8 Gb/sec optics can auto-negotiation with 8, 4, and 2 Gb/sec optics. 4 Gb/sec auto-negotiation works with 4,2, and 1 Gb/sec optics.
- The FOS upgrade path to FOS 7.0.0c is from FOS 6.4.2a only. If switches are at any level prior to FOS 6.4.2a, they must be upgraded to FOS 6.4.2a before proceeding to FOS 7.0.0c.
- Port based routing should be used for all FICON environments. Support for exchange based routing is suspended until further notice.
- For interoperability between FOS levels refer to the “Brocade Fabric OS v7.0.0c - Release Notes v1.0” (or the latest version) under the section “Appendix: Additional Considerations for FICON Environments”.
- 16 Gb/sec was only tested over ISLs due to the lack of any other processor or device that runs 16 Gb/sec.
- BNA 11.1.2 can be used to manage M-EOS products; however M-EOS products can not be attached via ISL to a fabric containing FOS 7.0.0c and up. The M-6140 must be at M-EOS 9.9.9c and the M-i10k must be at M-EOS 9.9.8n.

This document and future qualification letters may be found on the IBM Resource Link Web site:

For the latest copy of this qualification letter go to the IBM Resource Link Web site.

Navigate to the following website
http://www.ibm.com/servers/resourcelink/
Hit the link for “Sign In”.
Sign in with valid user ID and password
On the left, click on the "Library" link
Locate the listing of "Hardware products for servers" around the middle of the web page
Click on the link “Switches and directors qualified for IBM System z FICON and FCP channels”

IBM does not make any representations or warranties of any kind regarding the Brocades products and is not liable for such products or any claims made regarding such products. The fact that the listed Brocade products passed the enumerated IBM tests does not imply that the products will operate properly in any particular customer environment. Brocade retains sole responsibility for its products, the performance of such products and all claims relating to such products, including without limitation its products’ compliance to product specifications, safety requirements, regulatory agencies requirements and industry standards.

The terms IBM, eServer, DS6000, DS8000, TotalStorage, ESCON, FICON, System z, System z9, System z10, System Storage, SA OS/390, Resource Link, zEnterprise, z/OS, z/VM, z/VSE, z9, z10, z114, z196, and zSeries are trademarks or registered trademarks of International Business Machines Corporation.

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

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

Rick Leonard, PMP®, Vendor Services Lab Manager
System z® Hardware Development
International Business Machines Corporation

Sam Mercier, VSC Lab Sr. Engineer
System z® Hardware Development
International Business Machines Corporation

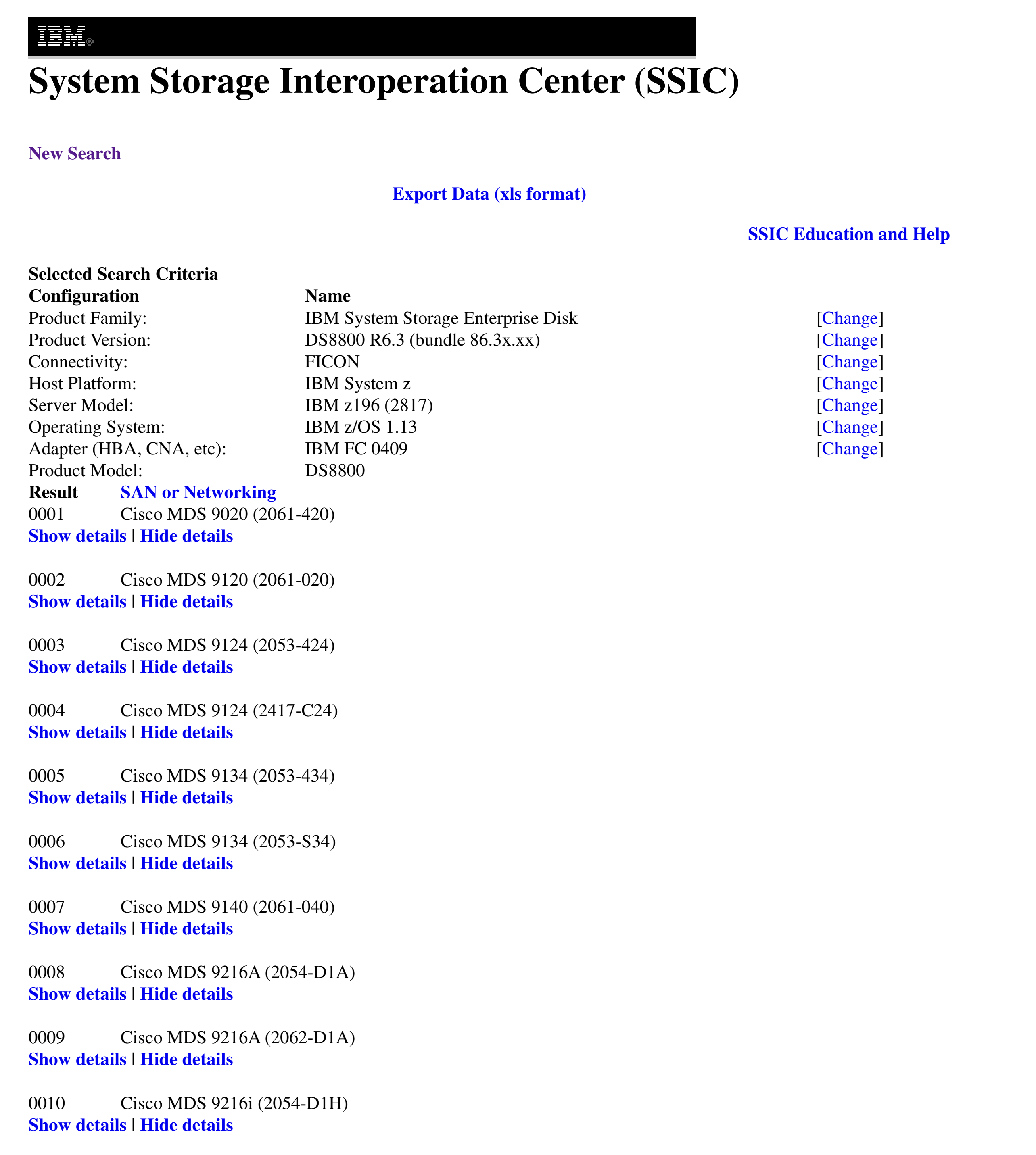
Page 4 of 4

Figure B-7 Sample switch qualification letter (page 4 of 4)

System Storage Interoperation Center

The SSIC website helps you identify combinations of storage devices, central electronics complexes (CECs), features, functions, channel types, and switches and directors that are supported to work together. An excerpt from a sample report is shown in Figure B-8. The website is located at the following address:

http://www.ibm.com/systems/support/storage/ssic/interoperability.wss



System Storage Interoperation Center (SSIC)

New Search

Export Data (xls format)

SSIC Education and Help

Selected Search Criteria

Configuration	**Name**	
Product Family:	IBM System Storage Enterprise Disk	[Change]
Product Version:	DS8800 R6.3 (bundle 86.3x.xx)	[Change]
Connectivity:	FICON	[Change]
Host Platform:	IBM System z	[Change]
Server Model:	IBM z196 (2817)	[Change]
Operating System:	IBM z/OS 1.13	[Change]
Adapter (HBA, CNA, etc):	IBM FC 0409	[Change]
Product Model:	DS8800	

Result **SAN or Networking**

0001 Cisco MDS 9020 (2061-420)
Show details | Hide details

0002 Cisco MDS 9120 (2061-020)
Show details | Hide details

0003 Cisco MDS 9124 (2053-424)
Show details | Hide details

0004 Cisco MDS 9124 (2417-C24)
Show details | Hide details

0005 Cisco MDS 9134 (2053-434)
Show details | Hide details

0006 Cisco MDS 9134 (2053-S34)
Show details | Hide details

0007 Cisco MDS 9140 (2061-040)
Show details | Hide details

0008 Cisco MDS 9216A (2054-D1A)
Show details | Hide details

0009 Cisco MDS 9216A (2062-D1A)
Show details | Hide details

0010 Cisco MDS 9216i (2054-D1H)
Show details | Hide details

Figure B-8 Sample SSIC report showing the supported switches for a given DS8000 configuration

C

Physical layer information

This appendix provides information about the planning and management considerations for the physical layer.

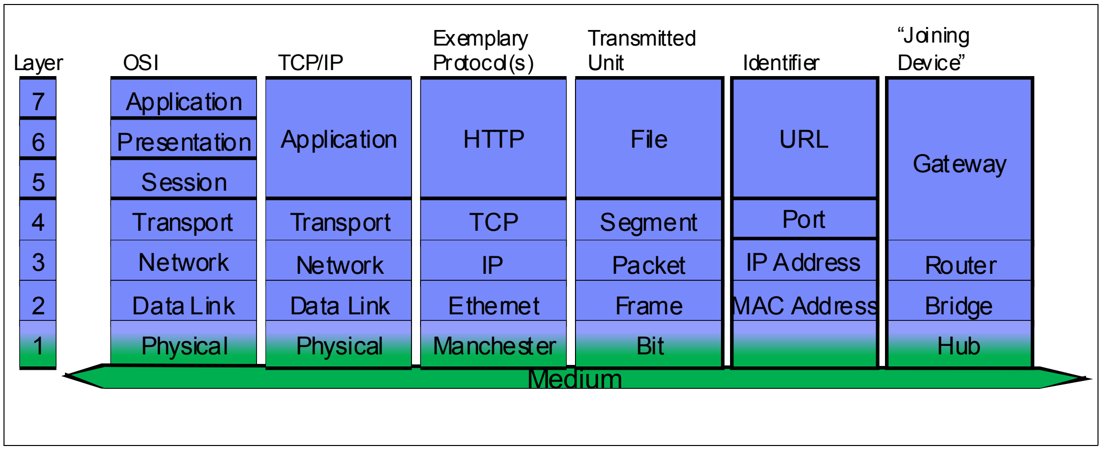
About physical layer switches

The physical layer is the first, or number one, layer in standard protocol stacks, as shown in Figure C-1. The open systems interconnection (OSI) model on the left is often used as a reference protocol stack, and has seven well-defined layers. To the right of the OSI is the Transmission Control Protocol/Internet Protocol (TCP/IP) protocol stack with 5 layers, driven more by pragmatism than architectural purism.

Well-known examples of protocols that run in each layer are to the right of the TCP/IP stack. Next are examples of materials transmitted by the protocol, followed by the identifiers of objects in that layer. To the far right are devices that connect sections of the protocol.

The IP transmits packets to IP addresses through routers and runs in the Internet Protocol network layer. Ethernet transmits frames to Media Access Control (MAC) addresses through hubs or bridges, or Layer 2 switches, and runs in the TCP/IP data link layer. The Ethernet physical interface transceiver (PHY) uses Manchester encoding for the layer that transmits bits, where hubs (simple repeaters) can extend or join different physical sections of a local area network (LAN).

Figure C-1 illustrates that only bits are transmitted at the physical layer, with no examination or concern for addressing or protocols. It also shows that there are some physical layer switches (for example, robotic switches) that are not even concerned with bits, because a signal of any format is switched to different destinations, but the signal is never intercepted.



Layer	OSI	TCP/IP	Exemplary Protocol(s)	Transmitted Unit	Identifier	“Joining Device”
7	Application	Application	HTTP	File	URL	Gateway
6	Presentation					
5	Session					
4	Transport	Transport	TCP	Segment	Port	
3	Network	Network	IP	Packet	IP Address	Router
2	Data Link	Data Link	Ethernet	Frame	MAC Address	Bridge
1	Physical	Physical	Manchester	Bit		Hub

Figure C-1 Communications protocol stacks and related information

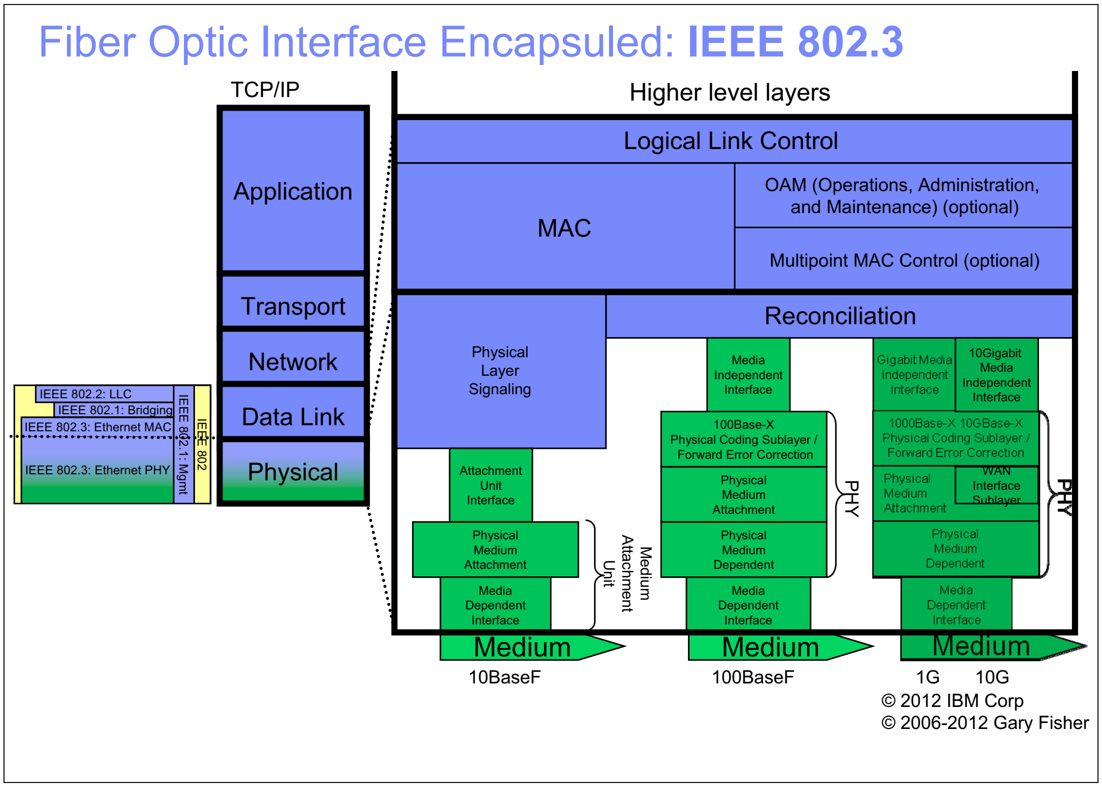
The physical medium is attached to layer 1, the physical layer. This could be copper, fiber, or wireless; the difference is hidden from upper layers. The data link layer protocol does not change for a change in transmission medium. In fact, the physical layer is often split into two: The upper part that interacts with the data link layer, and the lower part that must affect the change of medium.

Figure C-2 on page 223 shows the protection that a protocol stack offers other layers. The following list includes some of these details:

- The standardized version of Ethernet is Institute of Electrical and Electronics Engineers (IEEE) 802, or Carrier Sense Multiple Access with Collision Detect (CSMA/CD).
- IEEE 802.2 describes the interface to the network layer, such as IP.

- IEEE 802.3 describes the MAC functions, which are architecturally the lower part of the data link layer. IEEE 802.3 also describes the physical layer, or PHY, which is split into an upper interface to the data link layer, and a lower interface which must change for each physical media.
- Ethernet Physical Layer and different speeds over fiber shows 4 different PHY standard specifications that provide support for 10 megabits per second (Mbps), 100 Mbps, 1000 Mbps (1 gigabits per second (Gbps)), and 10 Gbps over fiber.

  Note that changes of the physical media are standardized and can be changed quickly without affecting any other portion of the system. For fiber optic communication, you can do this by changing the small form-factor pluggable (SFP) optical transceiver; the attaching devices and the fiber do not change.



Fiber Optic Interface Encapsuled: **IEEE 802.3**
TCP/IP
Application
Transport
Network
Data Link
Physical
IEEE 802.2: LLC
IEEE 802.1: Bridging
IEEE 802.3: Ethernet MAC
IEEE 802.3: Ethernet PHY
IEEE 802.1: Mgmt
IEEE 802
Higher level layers
Logical Link Control
MAC
OAM (Operations, Administration, and Maintenance) (optional)
Multipoint MAC Control (optional)
Reconciliation
Physical Layer Signaling
Attachment Unit Interface
Physical Medium Attachment
Media Dependent Interface
Medium Attachment Unit
Medium
10BaseF
Media Independent Interface
100Base-X Physical Coding Sublayer / Forward Error Correction
Physical Medium Attachment
Physical Medium Dependent
Media Dependent Interface
PHY
Medium
100BaseF
Gigabit Media Independent Interface
10Gigabit Media Independent Interface
1000Base-X 10GBase-X Physical Coding Sublayer / Forward Error Correction
Physical Medium Attachment
WAN Interface Sublayer
Physical Medium Dependent
Media Dependent Interface
PHY
Medium
1G
10G
© 2012 IBM Corp
© 2006-2012 Gary Fisher

Figure C-2 Physical medium changes encapsulated in physical layer of protocol stack

Optical to electrical to optical conversion and optical forms

In optical to electrical to optical conversion (OEO), the optical signal on a fiber is received and converted into an electrical version that is passed through the switch to the next transceiver, which converts it back into an optical signal sent to the destination. OEO switches amplify a source signal.

You can use OEO switches as media conversions, for example, converting multimode (MM) to single-mode (SM). These switches can also convert from electrical signals, for example, from an RJ-45 SFP to optical signals that make them electrical to electro-optical (EEO).

OEO switches introduce some delay, typically a few nanoseconds (ns), during the conversion. Typically, OEO switches do not perform any packet or frame analysis, because they operate solely in the physical layer. Bits come in and are sent back out.

OEO switches have optical transceivers and connection backplanes. The transceivers must be capable of handling the speed of the signal that they receive, and so are technology-dependent. The connection backplane is also technology-dependent. That is, it is not enough to remove a 4 Gbps transceiver from an OEO switch port and plug in an 8 Gbps transceiver in its place, the OEO switch must also be able to pass 8 Gbps traffic on its backplane. Newer optical technologies use polarity and phase modulation for high speeds.

The goal of an optical switch is to remove the need for the OEO process. Light comes in one port and is sent out to any of the other ports on the switch. Optical switches come in many formats, from mechanical (robotic) to some esoteric technologies.

A mechanical switch is essentially a robotic telephone switchboard operator in a sterile, environmentally controlled environment. These switches introduce no latency, have little loss in the connection, and can work with MM and SM fiber, even in the same switch. They also pass the light even in the event of a power loss.

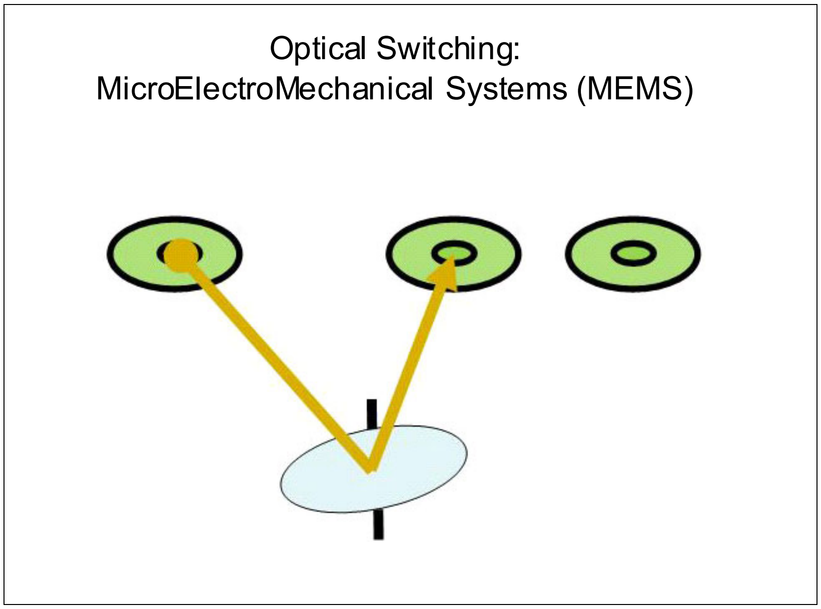
However, their switching time is long compared to other electronic patch panel options, usually by many seconds per new connection. The connections are usually serialized, which means a major configuration change could take many minutes. Also, there is a large but finite number of reconfigurations before physical wear-and-tear will affect the connections.

MicroElectroMechanical Semiconductors

MicroElectroMechanical Semiconductors (MEMS) are specialized microchips with adjustable mirrors that bounce light from the incoming port to one of several outgoing ports, as shown in Figure C-3 on page 225 and Figure C-4 on page 225. There are several forms of MEMS and some inventive configurations: Some with a single mirror per light path, others with multiple mirrors in an array, or adjustable prisms.

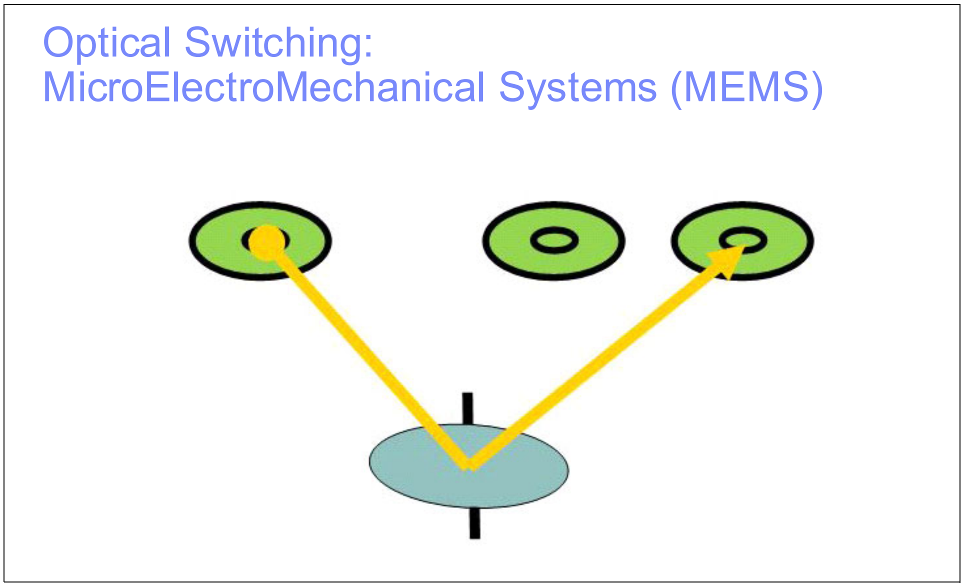
Wave Steering switches use electronically controlled arms to point fibers toward each other. Both of these switches introduce no latency, and have negligible polarity or phase modulation, so they can handle the newer high-speed connections. They typically switch to new connections in milliseconds (ms). A drawback is that they work only with SM fiber due to the spray of MM fiber, although there are some wave steering switches for small numbers of MM fiber, and some new attempts at MEMS for MM fiber.

The problem with these switches is the relatively large loss of optical power because the light is passed through air. The problem is not the air, but the extreme accuracy required to send light to a 9 micron hole in a remote fiber end. Such devices tend to have some optical power fluctuation, because they constantly self-correct to maintain the appropriate direction. Therefore, they do not keep a connection if they lose power.



Optical Switching:
MicroElectroMechanical Systems (MEMS)

Figure C-3 MEMS connecting left to middle



Optical Switching:
MicroElectroMechanical Systems (MEMS)

Figure C-4 MEMS connecting left to right

Table C-1 summarizes the physical layer switch characteristics covered in this section. It does not show other considerations, such as port count, power usage, and price per port. A red box around the cell indicates that the value is less desirable than the others, but you have to determine whether that factor is important for your purposes.

Table C-1 Physical layer switch characteristics (box indicates undesirable)

Technology	Switch MM fiber	Switch SM fiber	Switching through backplane	Latency through switch	Switching speed	Power outage affects
OEO	Yes	Yes	Yes	Yes	Fast	Yes
Robotic	Yes	Yes	No	No	SLOW	No
Optical	No	Yes	No	No	Fast	Yes
Wave steering	Limited	Yes	No	No	Fast	Yes

D

Fiber cabling services

This appendix describes the IBM fiber cabling services options being offered by IBM Global Technology Services (GTS) to clients.

The following topics are covered:

- Fiber cabling services options:
  - System z fiber cabling services
  - Enterprise fiber cabling services
- Fiber transport system (FTS) components

Fiber cabling services options

When integrating a System z server into a data center, an IBM Installation Planning Representative (IPR) provides planning assistance to clients for equipment power, cooling, and the physical placement of the server.

However, the fiber optic cable planning and connecting of the System z server channels to I/O equipment, coupling facilities (CFs), networks, and other servers is the client's responsibility, both for new server installations and server upgrades.

Tip: Clients, especially those with complex system integration requirements, can request connectivity assistance from IBM. See the next section for details.

Clients with the resources and personnel to plan and implement their own connectivity, or those with less complex system configurations, can consult the following manuals to help them determine and order the required fiber optic cabling:

- *IBM zEnterprise 196 Installation Manual for Physical Planning*, GC28-6897
- *System z10 EC Installation Manual for Physical Planning*, GC28-6865
- *System z9 Installation Manual for Physical Planning*, GC28-6844
- *Planning for Fiber Optic Links*, GA23-0367

These manuals are available on the IBM Resource Link website:

http://www.ibm.com/servers/resourcelink

IBM Site and Facilities Services

IBM Site and Facilities Services has a comprehensive set of scalable solutions to address IBM cabling requirements, from product-level to enterprise-level for small, medium, and large enterprises:

- IBM Facilities Cabling Services: Fiber transport system
- IBM IT Facilities Assessment, Design, and Construction Services: Optimized Airflow Assessment for Cabling

Planning and installation services for individual fiber optic cable connections are available. An assessment and planning for IBM FTS trunking components can also be performed.

These services are designed to be right-sized for your products or the end-to-end enterprise, and to take into consideration the requirements for all of the protocols and media types supported on the IBM zEnterprise EC12 (zEC12), IBM zEnterprise BC12 (zBC12), IBM zEnterprise 196 (z196), IBM zEnterprise 114 (z114), System z10, and zSeries (for example, ESCON, FICON, Coupling Links, and Open Systems Adapter (OSA)-Express), whether the focus is the data center, the storage area network (SAN), the local area network (LAN), or the end-to-end enterprise.

IBM Site and Facilities Services are designed to deliver convenient, packaged services to help reduce the complexity of planning, ordering, and installing fiber optic cables. The appropriate fiber cabling is selected based on the product requirements and the installed fiber plant.

The following list describes how the services are packaged:

- Under IBM Facilities Cabling Services there is the option to provide IBM FTS trunking commodities (fiber optic trunk cables, fiber harnesses, and panel-mount boxes) for connecting to the zEnterprise central electronics complexes (CECs). IBM can reduce the cable clutter and cable bulk under the floor. An analysis of the channel configuration and any existing fiber optic cabling is performed to determine the required FTS trunking commodities.

  IBM can also help organize the entire enterprise. This option includes enterprise planning, new cables, fiber optic trunking commodities, installation, and documentation.

- Under IBM IT Facilities Assessment, Design, and Construction Services there is the Optimized Airflow Assessment for Cabling option to provide you with a comprehensive review of your existing data center cabling infrastructure. This service provides an expert analysis of the overall cabling design required to help improve data center airflow for optimized cooling, and to facilitate operational efficiency through simplified change management.

FTS connectivity

The need for data center cabling implementation arises from the following scenarios:

- Establishing a new data center
- Upgrading an existing data center by replacing the cabling
- Adding new equipment to an existing data center

There are two choices when implementing fiber optic cabling. The first uses discrete fiber optic jumper cables (Figure D-1).

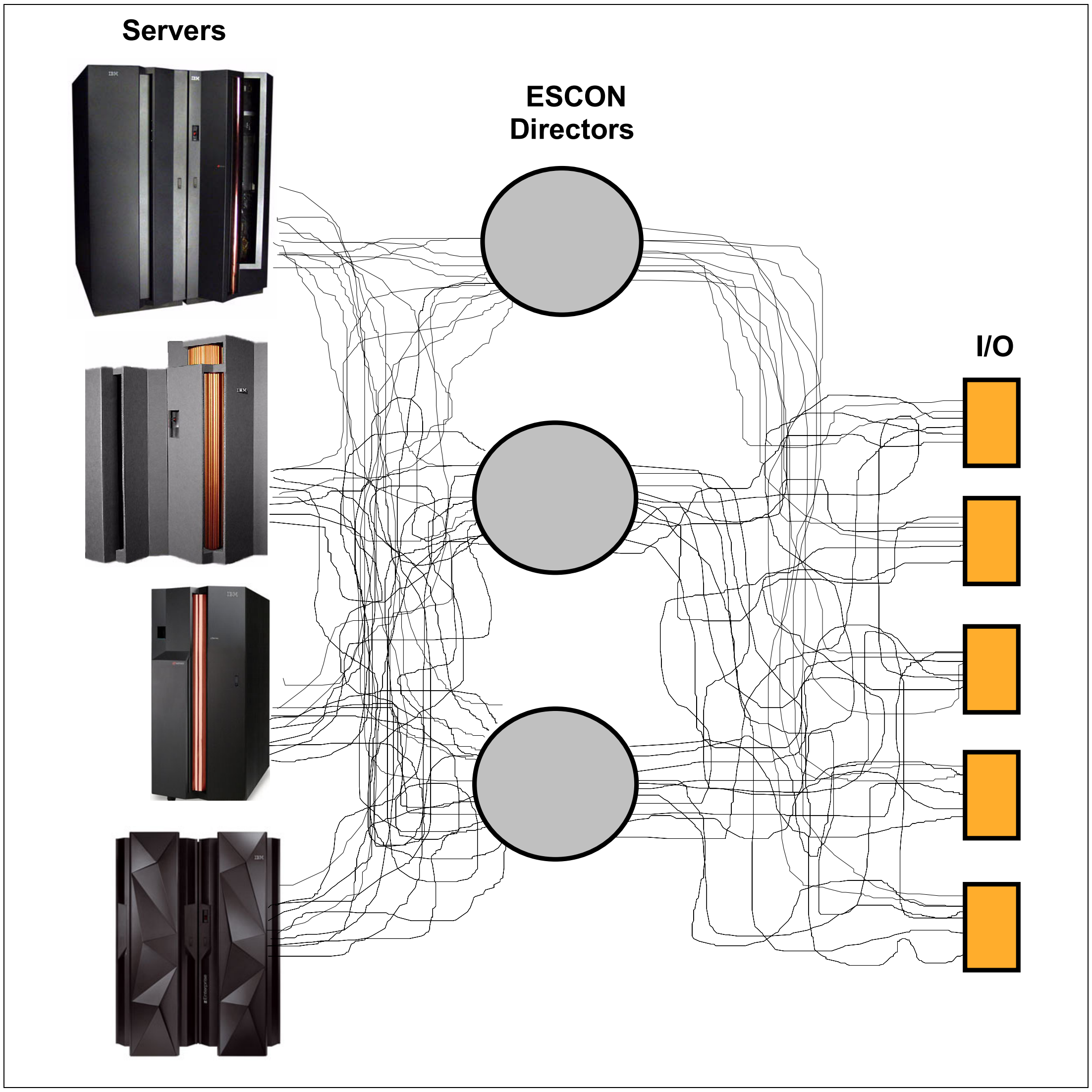


Figure D-1 Unstructured cable environment

Each jumper cable connects one server port directly to another to form a link (for example, one server channel-path identifier (CHPID) port to one ESCON Director port). In today's data centers, with the huge number of fiber optic cables and their diversity, an underfloor cabling system can soon get out of control and show deficiencies:

- Unknown cable routing
- No cable documentation system
- Unpredictable effect of moves, adds, and changes
- Unknown risk presented at every underfloor activity

The second choice for ESCON fiber optic cabling is a structured trunking system (Figure D-2).

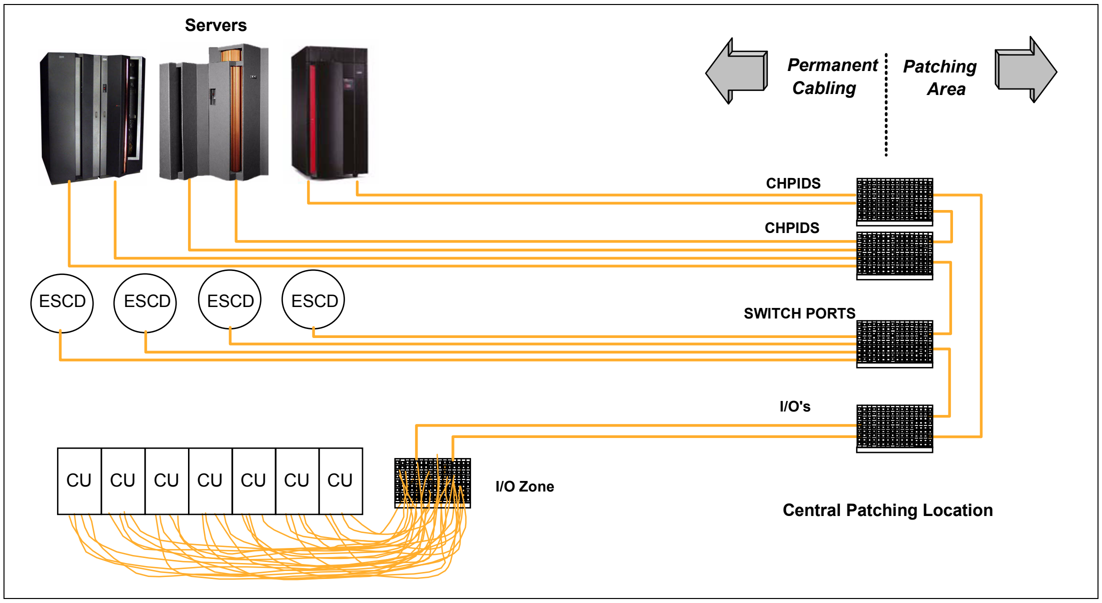


Figure D-2   FTS in an ESCON environment

A structured fiber optic trunking system greatly reduces the number of discrete jumper cables running under the raised floor.

FTS offers complete end-to-end connectivity of the fiber optic environment plant for all System z server link applications.

The fiber optic trunk cables connect the server ports to the back of patch panels that are in the central patching location (CPL). The CPL usually is made up of cabinets or racks that hold the patch panels. The fronts of the patch panels have individual ports that now represent the server ports. Connections between two server ports can be done quickly and easily by running a short jumper cable to connect the two patch panel ports.

FTS benefits

The most obvious benefit of the structured trunking system is the large reduction in the number of fiber optic cables under the raised floor. The smaller number of cables makes documenting what cables go where much easier. Better documentation means that tracing a fiber optic link is much easier during problem determination, and when planning for future growth.

A less apparent and often overlooked benefit of a structured system is its ability to make future data center growth implementation much easier. With a structured system installed, channels, ESCON Director ports, and control unit I/O ports are connected by fiber optic trunk cables to patch panels in the CPL. All of the connections between the equipment are made with short jumper cables between the patch panels in the CPL.

When new equipment arrives, it is connected to patch panels in the CPL as well. Then the actual reconfiguration takes place in the CPL by moving short jumper cables between the patch panels, not by moving long jumper cables under the raised floor.

Also, none of the change activity is done near the active equipment, unlike the case with the discrete jumper cable solution. Future equipment additions and changes can be done in the same manner, and are not affected by the amount of equipment already installed on the floor.

Table D-1 shows the advantages of a structured cabling system over a non-structured cabling system.

Table D-1 Benefits of the structured cabling system

Non-structured cabling	Structured cabling
Unknown cabling routing.	Known cable pathways.
No cable documentation system.	Defined cable documentation.
Unpredictable effect of moves, adds, and changes.	Reliable outcome of moves, adds, and changes.
Every underfloor activity is an unknown risk.	Underfloor activity can be planned to minimize risk.

FTS is a long-term connectivity solution that provides an organized network of cabling options for future equipment reconfigurations and additions. Changes can be performed with minimal floor disruptions, and are accomplished by rearranging short jumpers at the panel-mount boxes in the CPL. Each FTS design is unique in that it is based on physical room characteristics and equipment configurations and placement preferences.

FTS provides the following benefits:

- Tested and approved connectivity:
  - FTS components are IBM-tested, approved, and sold under the IBM logo.
  - FTS trunk-mounting kits are designed and tested with System z servers and devices to prevent any effect on equipment operation or serviceability.
  - Multimode (MM) and single-mode (SM) fiber solutions are available.
- Elimination of long jumper cables, making underfloor areas less congested and easier to manage by using direct-attach trunking.
- Cost reductions and productivity gains:
  - Faster, less disruptive installation and removal of equipment
  - Easier reconfiguration of equipment
  - More efficient use of underfloor space (potential savings in air conditioning requirements)
- Improved cable management with a reduced risk of damage to fiber cables for the following environments:
  - ESCON
  - FICON
  - Parallel Sysplex
  - OSA
  - Open system architectures (Fibre Channel (FC) and Gigabit Ethernet (GbE))
- Growth flexibility:
  - Expandable
  - Relocatable
  - A long-term solution

These benefits and potential cost savings must be assessed for each data center and its particular environment.

The FTS structured cabling has two implementation solutions. Both use direct-attach trunking at the server to connect to the CPL. The difference is the type of panel-mount boxes used at the CPL.

Modular panel-mount connectivity

The first FTS implementation solution is called modular panel-mount connectivity. This uses Multi-fiber Termination Push-on (MTP)-to-MTP trunk cables to connect to the CPL (Figure D-3).

The modular panel-mount boxes are designed to allow you to change the port type in the panel-mount box by unplugging the trunk and changing the modular inserts in the panel-mount box.

Modular panel-mount connectivity provides the following benefits:

- MTP-to-MTP trunks connect quickly to the panel-mount box.
- The connector type in front of the panel mount can be easily changed.
- Panel-mount capacity can be customized.

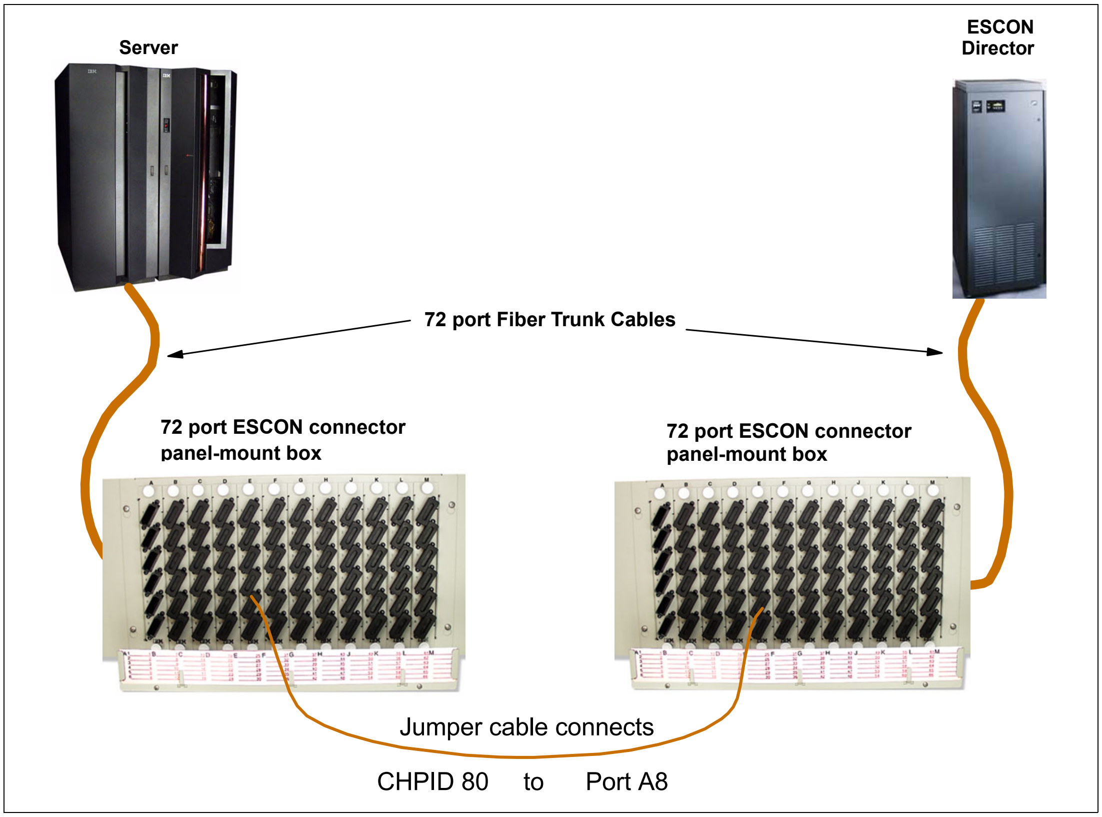

Server
ESCON Director
72 port Fiber Trunk Cables
72 port ESCON connector panel-mount box
72 port ESCON connector panel-mount box
Jumper cable connects
CHPID 80 to Port A8

Figure D-3 FTS solution 1: Modular panel-mount connectivity

Single connector-dual contact connectivity

The second FTS implementation solution is called single connector-dual contact (SC-DC) connectivity. It uses MTP-to-SC-DC trunk cables to connect to the CPL (Figure D-4).

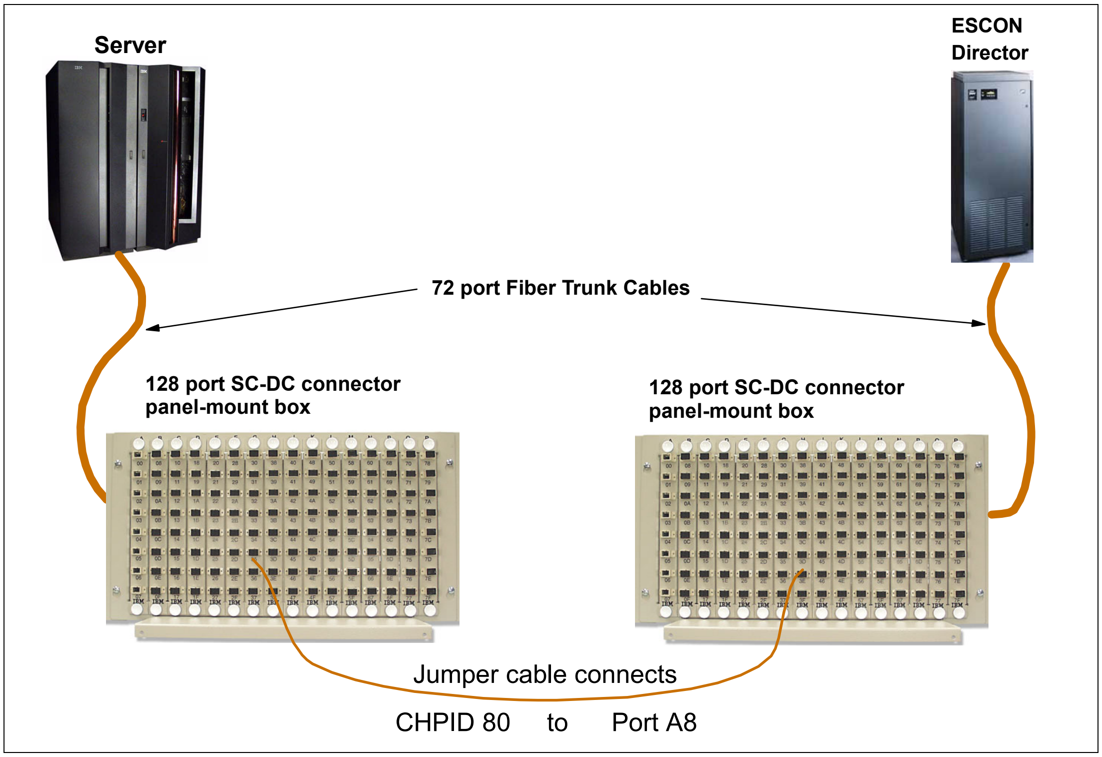

Server
ESCON Director
72 port Fiber Trunk Cables
128 port SC-DC connector panel-mount box
128 port SC-DC connector panel-mount box
Jumper cable connects
CHPID 80 to Port A8

Figure D-4 FTS solution 2: SC-DC connectivity

The SC-DC small form factor optical connector is the standard for FTS connectivity, eliminating the need for multiple connector types at patch panels and for patch cables. By standardizing on a connector with a history of reliability, the different optical fiber connectors used in equipment do not create a management problem when a system configuration must change.

The SC-DC supports termination of both sizes of MM fiber and SM fiber. As a result, CPL patch panel connections and patch cables are easily managed and differentiated by their color. This provides the following benefits:

- Server port order at the panel-mount is independent of the harness plugging at the server.
- Panel-mount ports come factory-labeled with the server port addresses.
- There is a single connector type at the CPL.

Table D-2 provides a comparison between the two types of FTS solutions.

Table D-2 FTS connectivity solutions comparison

Benefits	Jumper cables	Modular panel-mount connectivity solution	SC-DC connectivity solution
Organized, manageable underfloor cabling	No	Yes	Yes
All server ports in one central location	No	Yes	Yes
System reconfigurations done without lifting floor tiles or opening server covers	No	Yes	Yes
Server ports arranged in sequential order and labeled at the panel-mount boxes	No	No	Yes
Space allocated for future server growth	No	No	Yes

Fiber Quick Connect

Fiber Quick Connect (FQC) is an option in the eConfig configuration tool when ordering one of the following options:

- A new build zEnterprise CEC
- A new build System z10 server
- An upgrade of a System z

The FQC features are for factory installation of IBM FTS fiber optic harnesses for connection to all ESCON and FICON long wavelength (LX) on zEnterprise CECs and System z10. FTS fiber optic harnesses enable connection to FTS direct-attach fiber optic trunk cables.

FQC, when coupled with the FTS products from IBM GTS, delivers a solution to reduce the amount of time required for onsite installation and set up of cabling, to minimize disruptions and to isolate the activity from the active system as much as possible. FQC facilitates adds, moves, and changes of ESCON MM fiber optic cables and FICON LX SM fiber optic cables in the data center, and reduces fiber optic cable installation time.

Enterprise fiber cabling services provide the direct-attach trunk harnesses, patch panels, and CPL hardware, in addition to the planning and installation required to complete the total structured connectivity solution.

CPL planning and layout is done before arrival of the server onsite, and documentation is provided showing the channel layout and how the direct attach harnesses are plugged.

FQC supports all of the ESCON channels and FICON LX channels in all of the zEnterprise CECs and System z10. FQC cannot be ordered for selected channels and cages within the server. For example, FQC for ESCON is based on a quick connect/disconnect trunking strategy using the 12-fiber MTP connector, so it is able to transport six ESCON CHPIDs.

For example, six trunks, each with 72 fiber optic pairs (twelve MTP connectors), can displace up to 420 fiber optic cables, the maximum quantity of ESCON channels supported in one input/output (I/O) cage on a z196 or System z10. This significantly reduces ESCON cable bulk. The MTP connector enables FQC trunk cables to be installed, and later disconnected and relocated quickly. The MTP connector also enables FQC to bring its fiber trunk cables directly under the covers of the System z servers and ESCON Directors.

FQC direct-attach harness

FICON LX direct-attach harnesses have one MTP connector that breaks out to six LC-Duplex connectors.

ESCON direct-attach harnesses have one MTP connector that breaks out to six MT-RJ connectors (for 16-port ESCON features) or six ESCON Duplex connectors (for earlier 4-port ESCON features).

The direct-attach harnesses connect the trunk cable to the individual server ports, and accommodate the different industry-standard optical connectors used on System z servers (Figure D-5).

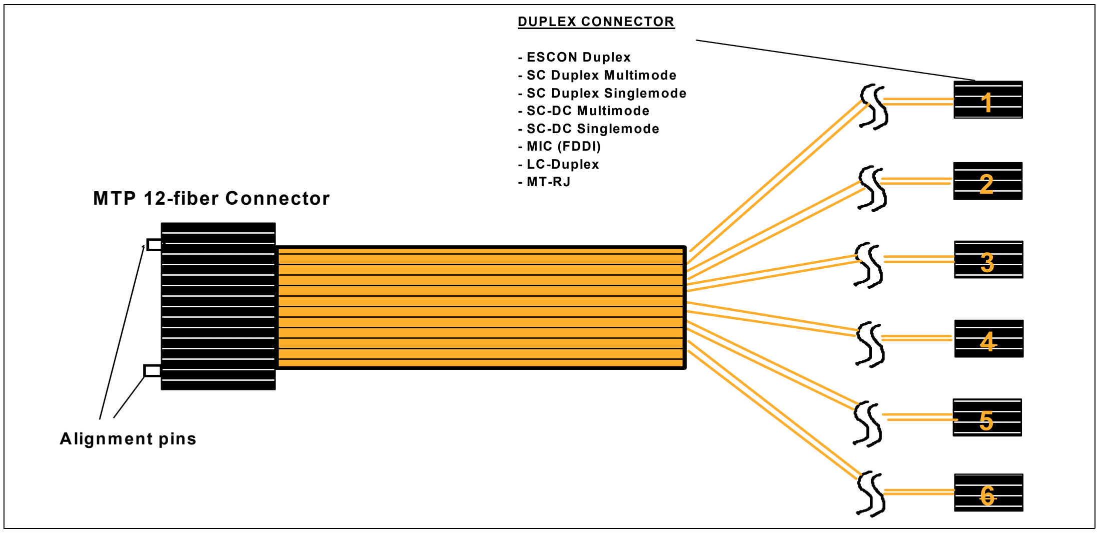


Figure D-5 FQC direct-attach harness schematic

The number of harnesses required depends on the number of channel ports in the server.

MTP coupler brackets

The MTP coupler brackets provide the mechanical support to connect the harness MTP connectors to the trunk cable MTP connectors. They are directly mounted on the server frames.

Direct attach trunk cable

The maximum density of fiber trunk cables is 144 fibers (72 ESCON channels), which use 12 MTP connectors on the server trunk cable end. The trunk cables run under the raised floor and connect the harnesses in the server to the panel-mount box in the CPL.

Fiber trunking is available with both overhead I/O exit and underfloor cable exit.

Central Patch Location panel-mount box

The CPL panel-mount box is the connection point for the System z server. This isolates the active server from future system moves, adds, and changes.

> **Restriction:** Although FTS provides cabling solutions for the different industry-standard optical connectors used on System z servers, the FQC feature, which is the FTS factory-installed direct attach trunking system, is only available for ESCON-attached channels.

Summary

Enterprise fiber cabling services provide a structured FTS fiber cabling system that combines planning, installation, and service. It consists of fiber trunk cables, direct-attach harnesses, and a variety of main distribution frame (MDF) panel mounts. The installation might also include a fiber conveyance system, which is an underfloor tray system that protects the fiber optic cables.

Each enterprise fiber cabling services design is unique in that it is based on physical room characteristics and equipment placement preferences. It is a long-term *connectivity solution* that provides an organized network of cabling options for future equipment reconfigurations and additions:

- Secure configuration environment.
- Under-floor cable control. Long jumper cables can be eliminated, making underfloor areas less congested and easier to manage.
- Efficient cable management:
  - Uniform cable documentation
  - Structured cable routing
  - Simplified system configuration
  - Server ports in one central distribution facility
  - CHPIDs and device ports in order at patch panels
  - Factory-installed harnesses and tailgates to ensure consistent routing, allowing quick connection of fiber trunks
  - Space reserved for future growth
- Changes can be performed with minimal floor disruptions and are accomplished by rearranging short jumpers at the panel-mount boxes in the MDF.

In summary, Enterprise fiber cabling services solutions can provide the following benefits:

- Complement the ESCON architecture by providing easier copper-to-fiber migration.
- Cost reductions and productivity gains:
  - Faster, less disruptive install and removal of equipment
  - Easier reconfiguration of equipment
  - More efficient use of under-floor space (potential savings in air conditioning requirements)
- Improved cable management with a reduced risk of damage to fiber optic cables for the following environments:
  - ESCON
  - FICON
  - Parallel Sysplex
  - OSA
- Growth flexibility:
  - Expandable
  - Relocatable
  - A long-term solution

These benefits and potential cost savings need to be assessed for each data center and its particular environment.

References

For more information about the IBM Networking Services fiber cabling services offered by IBM GTS and related topics, see:

- The IBM Resource Link website:

  http://www.ibm.com/servers/resourcelink

- *Fiber Transport Services Direct Attach Planning,* GA22-7234
- *Installing the Direct Attach Trunking System in zSeries 900 Servers*, GA27-4247
- *ESCON I/O Interface Physical Layer Document,* SA23-0394
- *Coupling Facility Channel I/O Interface Physical Layer,* SA23-0395
- *Fiber Channel Connection for S/390 I/O Interface Physical Layer,* SA24-7172
- *Planning for Fiber Optic Links,* GA23-0367
- *Fiber Optic Link (ESCON, FICON, Coupling Links and OSA) Maintenance Information,* SY27-2597

E

Fiber optic cables

This appendix describes the physical attributes of optical fiber technologies supported on System z servers.

The following topics are covered:

- Fiber description
- Fiber connector types
- Mode Conditioning Patch (MCP) cables
- Conversion cables

Description

Fiber optic cables use light for data transmission, rather than electrical current on copper cables. Fiber optic cables have many advantages. For example, they are many times lighter and have substantially reduced bulk, no pins, a smaller and more reliable connector, reduced loss and distortion, and are free from signal skew or the effects of electro-magnetic interference.

Figure E-1 illustrates the two types of optical fiber used in a data center environment in conjunction with System z servers:

- Multimode (MM)
- Single-mode (SM)

The difference between them is the way that light travels along the fiber. MM has multiple light paths, but SM has only one light path.

Each fiber type consists of several parts:

- The core can be 50 or 62.5 micrometers (µm) in diameter for MM, or 9 µm in diameter for SM.
- The cladding that surrounds the core is 125 µm in diameter.
- The outer coating is 250 µm in diameter.

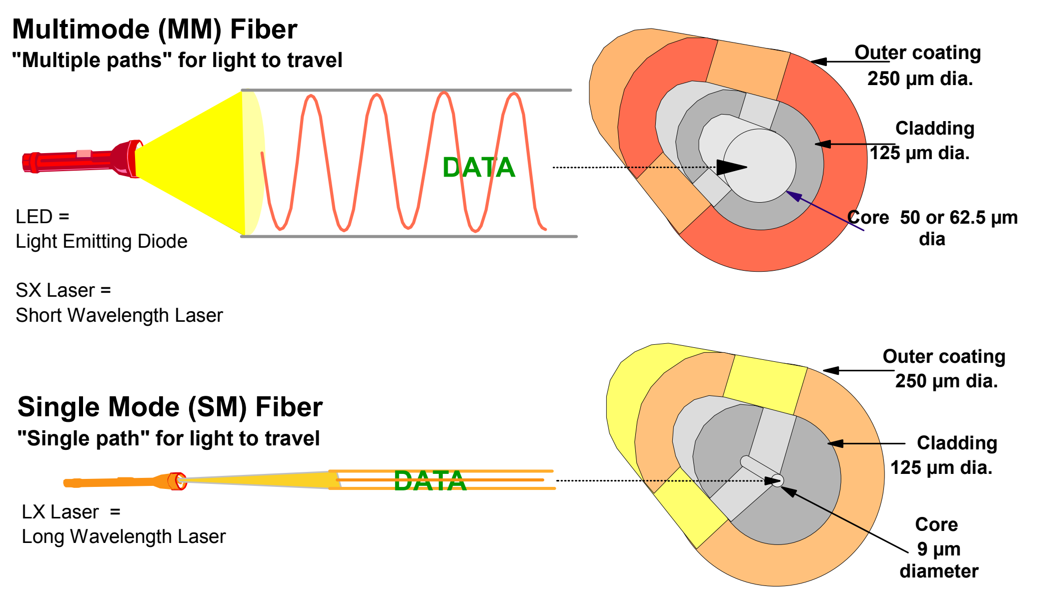


Figure E-1 Fiber optic cable types

Connector types for fiber cables

For all optical links the connector type is LC Duplex, except the following links:

- ESCON has an MT-RJ type connector.
- 12x InfiniBand (IFB) has a Multi-fiber Push-On (MPO) connector.

Figure E-2 shows the most common fiber cable connectors used in data center environments.

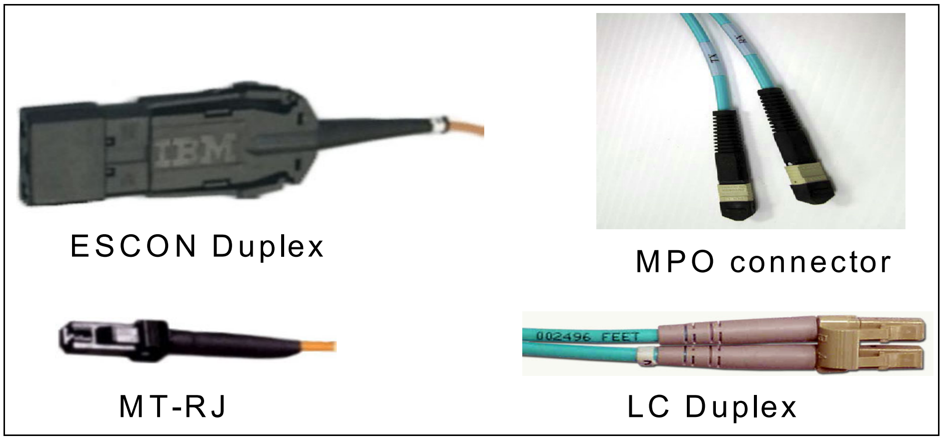


Figure E-2 Most common connectors used for optical cables

Mode Conditioning Patch cables

MCP cables allow for already installed MM fiber cables to be reused for long wavelength (LX) links. The MCP cables are two meters (m) long and have a link loss budget of 5.0 decibels (dB). The MCPs can be used for the following fiber optic links (Table E-1 on page 242):

- InterSystem Channel-3 (ISC-3)
- FICON Express LX
- FICON Express2 LX
- FICON Express4 LX
- Open Systems Adapter (OSA)-Express gigabit Ethernet (GbE) LX
- OSA-Express2 GbE LX
- OSA-Express3 GbE LX
- OSA-Express4S GbE LX

These links have an LX transceiver designed to be used with 9 µm SM fiber cables. MCP cables allow the LX optical signals to be carried over MM fiber cables. The ISC-3 feature can only use MCP cabling to operate in compatibility mode (1 gigabits per second (Gbps)), because 2 Gbps bit rate is not supported with MCP cables. FICON Express4 LX, FICON Express2 LX, and FICON Express LX links only support a link speed of 1 Gbps when MCP cables are used.

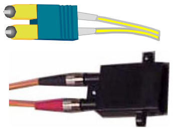
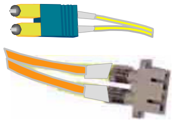
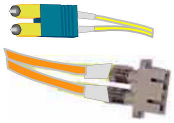
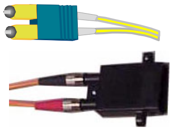
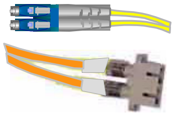
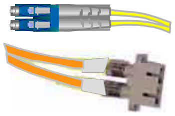
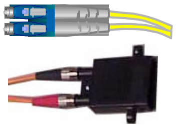
Be aware of the distance reductions when MCP cables are used (Table E-1 on page 242).

Important: One MCP cable must be plugged into the LX transceiver at each end of the link.

Fiber optic MCP cables are not orderable as product feature codes for zEnterprise central electronics complexes (CECs) and System z10.

Fiber optic MCP cables (Table E-1) can be ordered using the IBM Networking Services fiber cabling services options. See Appendix D, “Fiber cabling services” on page 227.

Table E-1 MPC cables

MCP cable description	MCP cable connector/receptacle description	MCP cable connector/receptacle illustration
9 µm SM to 50 µm MM	SC Duplex Connector to ESCON Duplex Receptacle	
9 µm SM to 50 µm MM	SC Duplex Connector to SC Duplex Receptacle	
9 µm SM to 62.5 µm MM	SC Duplex Connector to SC Duplex Receptacle	
9 µm SM to 62.5 µm MM	SC Duplex Connector to ESCON Duplex Receptacle	
ISC-3 compatibility 9 µm SM to 50 µm MM	LC Duplex Connector to SC Duplex Receptacle	
9 µm SM to 62.5 µm MM	LC Duplex Connector to SC Duplex Receptacle	
9 µm SM to 62.5 µm MM	LC Duplex Connector to ESCON Duplex Receptacle	

InfiniBand cables

An OM3 50/125 µm (2000 MHz-km @ 850 nm) MM fiber optic cable with MPO connectors is used for 12x IFB-double data rate (DDR) connections. The cable is an IFB Trade Association (IBTA) industry-standard cable. It contains one pair of fibers per lane for the 12x IFB-DDR connection (resulting in 24 duplex fibers).

The sender and receiver connectors are clearly marked with either transmitter (TX) or receiver (RX), and the connectors are keyed (Figure E-3).

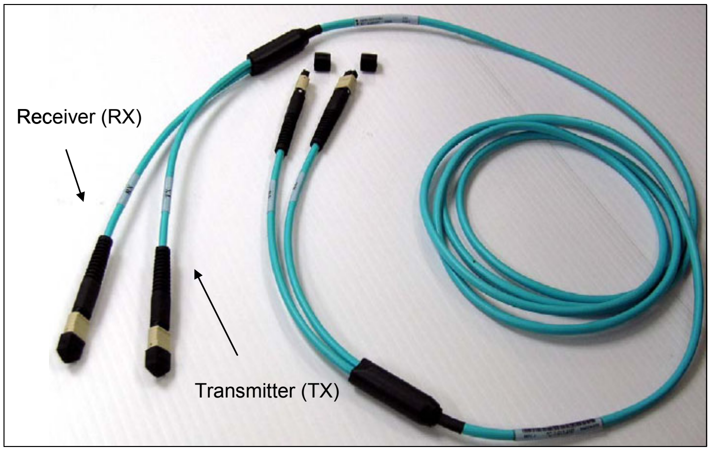


Figure E-3 OM3 50/125 µm MM fiber cable with MPO connectors

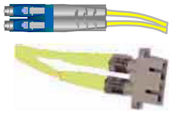
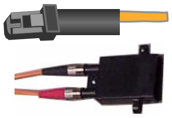
The following standard cable lengths are available from IBM Global Technology Services (GTS), or individual length can be ordered:

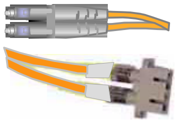
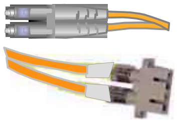
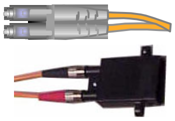
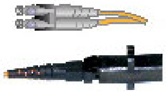
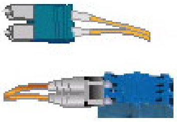
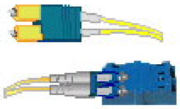
- 10 m (32.8 feet)
- 13 m (42.7 feet)
- 15 m (49.2 feet)
- 20 m (65.6 feet)
- 40 m (131.2 feet)
- 80 m (262.4 feet)
- 120 m (393.7 feet)
- 150 m (492.1 feet)

Conversion kits

Conversion kits allow for the reuse of already installed cables that are the same fiber optic mode but have different connectors from the ones required (Table E-2).

Table E-2 Conversion kit cables

Conversion kit cable description	Conversion kit cable connector/receptacle description	Conversion kit cable connector/receptacle illustration
9 µm SM	LC Duplex Connector to SC Duplex Receptacle	
62.5 µm MM	MT-RJ Connector to ESCON Duplex Receptacle	

Conversion kit cable description	Conversion kit cable connector/receptacle description	Conversion kit cable connector/receptacle illustration
50 µm MM	LC Duplex Connector to SC Duplex Receptacle	
62.5 µm MM	LC Duplex Connector to SC Duplex Receptacle	
62.5 µm MM	LC Duplex Connector to ESCON Duplex Receptacle	
62.5 µm MM	LC Duplex Connector to MT-RJ Connector with Coupler	
62.5 µm MM	SC Duplex Connector to LC Duplex Connector with Coupler	
9 µm SM	SC Duplex Connector to LC Duplex Connector with Coupler	

Restriction: Fiber optic conversion kits are not orderable as product feature codes for zEnterprise and System z10.

Fiber optic conversion kits can be ordered using the IBM Networking Services fiber cabling services options. Each conversion kit contains one cable. See Appendix D, “Fiber cabling services” on page 227.

References

The following publications contain information related to the topics in this appendix:

- *ESCON I/O Interface Physical Layer Document,* SA23-0394
- *Fiber Channel Connection for S/390 I/O Interface Physical Layer,* SA24-7172
- *Coupling Facility Channel I/O Interface Physical Layer,* SA23-0395
- *Planning for Fiber Optic Links,* GA23-0367
- *S/390 Fiber Optic Link (ESCON, FICON, Coupling Links and OSA) Maintenance Information,* SY27-2597
- *Fiber Transport Services Direct Attach Planning,* GA22-7234

Related publications

The publications listed in this section are considered particularly suitable to provide more detailed information about the topics covered in this book.

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.

- *Considerations for Multisite Sysplex Data Sharing*, SG24-7263
- *FICON Native Implementation and Reference Guide*, SG24-6266
- *FICON Planning and Implementation Guide*, SG24-6497
- *IBM System z Connectivity Handbook*, SG24-5444
- *IBM System z Qualified WDM: Adva FSP 2000 at Release Level 6.2*, REDP-3903
- *IBM System z Qualified WDM: Adva FSP 3000 Release Level 9.3*, REDP-4479
- *IBM System z Qualified WDM: Ciena CN 4200 at Release Level 6.0.1 and 6.0.2*, REDP-3907
- *IBM System z Qualified WDM: Ciena ActivSpan 5200 and 5100 Release 11.11*, REDP-4720
- *IBM System z Qualified WDM: Cisco ONS 15454 MSTP Release 8.5 and 9.0*, REDP-4395
- *IBM System z Qualified WDM: Cisco ONS 15454 MSTP Release 9.2*, REDP-4719
- *IBM System z Qualified WDM: Huawei OptiX OSN 6800 and OSN 3800 Release V100R004*, REDP-4478
- *IBM System z Qualified WDM: Nortel Optical Metro 5200 at Release Level 10.0*, REDP-3904
- *IBM System z Qualified WDM: Tellabs 7100 OTS Release FP5.1.1.f4*, REDP-4624
- *Implementing and Managing InfiniBand Coupling Links on System z*, SG24-7539
- *IBM SAN Survival Guide*, SG24-6143

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

ibm.com/redbooks

Other publications

This publication is also relevant as a further information source:

- *Fiber Optic Cleaning Procedure*, SY27-2604

Online resources

These websites are also relevant as further information sources:

- IBM Resource Link

  https://www.ibm.com/servers/resourcelink

- IBM zEnterprise 196 I/O Performance Version 1

  ftp://public.dhe.ibm.com/common/ssi/ecm/en/zsw03169usen/ZSW03169USEN.PDF

- IBM zEnterprise 196 and IBM zEnterprise 114 I/O and FICON Express8S Channel Performance Version 2

  ftp://public.dhe.ibm.com/common/ssi/ecm/en/zsw03196usen/ZSW03196USEN.PDF

- FICON Extended Distance Solution (FEDS)

  http://www.ibm.com/systems/resources/servers_eserver_zseries_library_techpapers_pdf_gm130092.pdf

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Technology Services

ibm.com/services

System z End-to-End Extended Distance Guide

(0.5” spine)
0.475”<->0.873”
250 <-> 459 pages

System z End-to-End Extended Distance Guide

Why you should have an end-to-end connectivity strategy for System z

What you should understand about the technology

How you should plan your connectivity infrastructure

This IBM Redbooks publication will help you design and manage an end-to-end, extended distance connectivity architecture for IBM System z. This solution addresses your requirements now, and positions you to make effective use of new technologies in the future.

Many enterprises implement extended distance connectivity in a silo manner. However, effective extended distance solutions require the involvement of different teams within an organization. Typically there is a network group, a storage group, a systems group, and possibly other teams.

The intent of this publication is to help you design and manage a solution that will provide for all of your System z extended distance needs in the most effective and flexible way possible. This book introduces an approach to help plan, optimize, and maintain all of the moving parts of the solution together.

SG24-8047-00 ISBN 0738439045