A first look into the Inner Workings and Hidden Mechanisms of FICON Performance

- David Lytle, BCAF
- Brocade Communications Inc.

- Tuesday August 9, 2011 – 11am to 12pm

- Session Number - 09368

Legal Disclaimer

- All or some of the products detailed in this presentation may still be under development and certain specifications, including but not limited to, release dates, prices, and product features, may change. The products may not function as intended and a production version of the products may never be released. Even if a production version is released, it may be materially different from the pre-release version discussed in this presentation.
- NOTHING IN THIS PRESENTATION SHALL BE DEEMED TO CREATE A WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, STATUTORY OR OTHERWISE, INCLUDING BUT NOT LIMITED TO, ANY IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, OR NONINFRINGEMENT OF THIRD-PARTY RIGHTS WITH RESPECT TO ANY PRODUCTS AND SERVICES REFERENCED HEREIN.
- Brocade, Fabric OS, File Lifecycle Manager, MyView, and StorageX are registered trademarks and the Brocade B-wing symbol, DCX, and SAN Health are trademarks of Brocade Communications Systems, Inc. or its subsidiaries, in the United States and/or in other countries. All other brands, products, or service names are or may be trademarks or service marks of, and are used to identify, products or services of their respective owners.
- There are slides in this presentation that use IBM graphics.

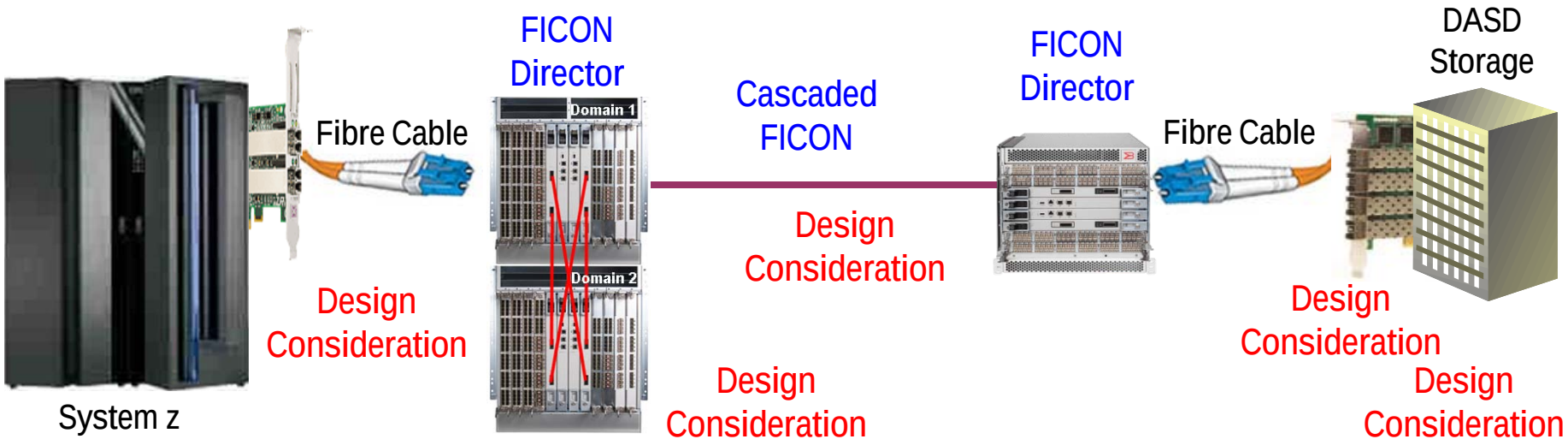
A first look into the Inner Workings and Hidden Mechanisms of FICON Performance

In this session we will discuss some of the architecture and design considerations of a FICON infrastructure.

The session will focus on a number of the design point considerations, from mainframe to storage connection, that affect the way that FICON could perform in your enterprise.

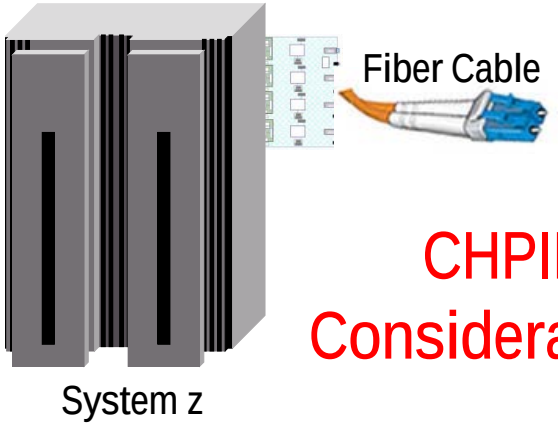
In the 2nd session, my Deeper Dive Into the Inner Workings of FICON performance, I will focus more on FICON Link Congestion; how Buffer Credits are used with FICON; Oversubscription; Slow Draining devices; and RMF reporting of Buffer Credits

End-to-End FICON/FCP Connectivity

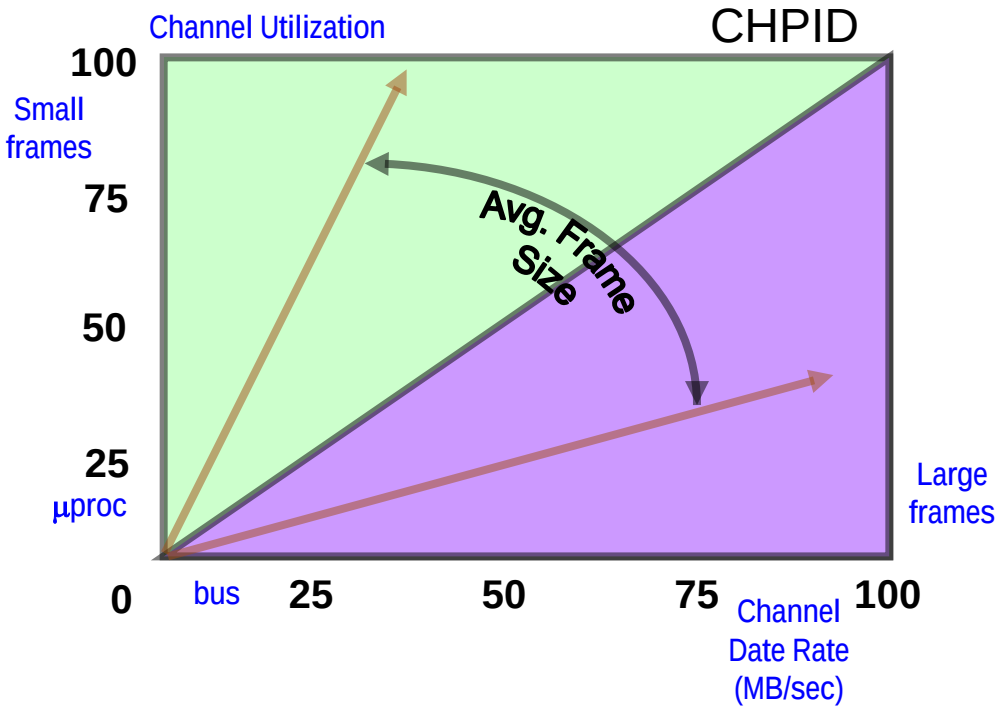


- From End-to-End in a FICON infrastructure there are a series of Design Considerations that you must understand in order to successfully meet your expectations with your FICON fabrics
- This is just a 20,000 foot OVERVIEW!

End-to-End FICON/FCP Connectivity

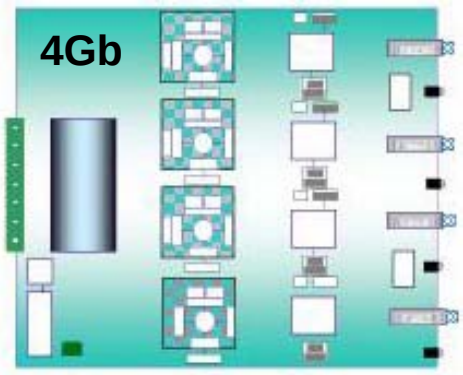


CHPID Considerations



- Channel Microprocessors and PCI Bus
- Average frame size for FICON
- Buffer Credit considerations

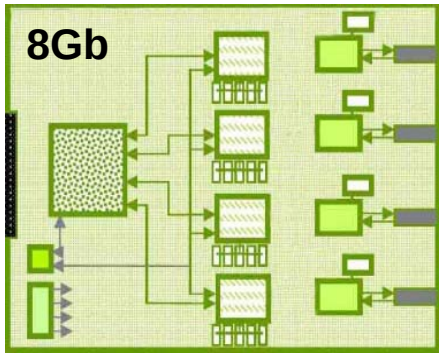
Current Mainframe Channel Cards (Features)



FICON Express4

- z196, z114, z10, z9
- 4 ports per feature
- 4km & 10km LX
- Shortwave (SX)
- 1, 2 or 4 GBps link rate

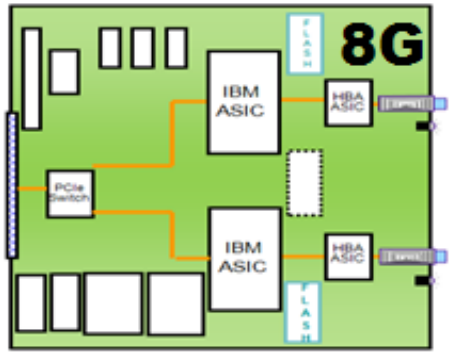
FICON Express4 provides the last native 1Gbps CHPID support



FICON Express8

- z196, z114, z10
- 4 ports per feature
- Longwave (LX) to 10km
- Shortwave (SX)
- 2, 4 or 8 GBps link rate

FICON buffer credits have become very limited per CHPID



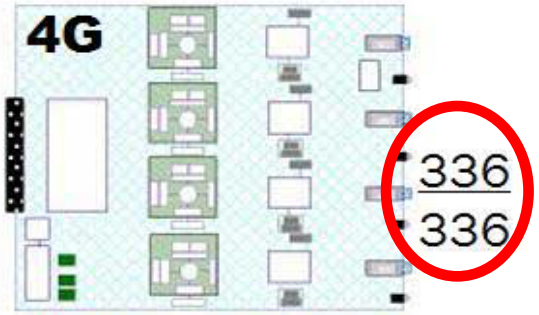
FICON Express8S

- z196, z114
- 2 ports per feature
- Longwave (LX) to 10km
- Shortwave (SX)
- 2, 4 or 8 GBps link rate

Reduced Ports per feature
...BUT...
Better Performance

One or more IBM Graphics are used above

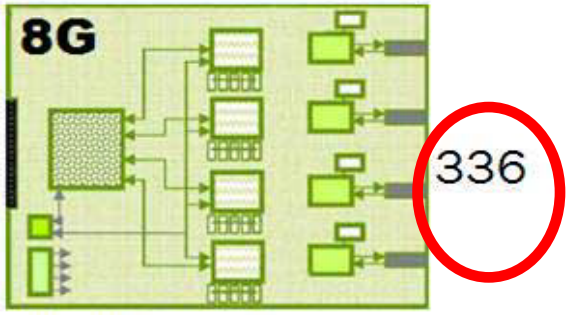
Mainframe Channel Cards



FICON Express4 – 4 ports
400MBps+400MBps=800MBps

FICON Express4

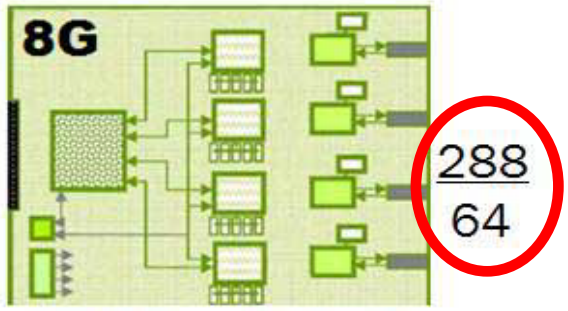
- z10, z9
- 1, 2 or 4 GBps link rate
- **Cannot Perform at 4Gbps!**
- Standard FICON Mode:
  <= 350MBps Full Duplex
  out of 800 MBps
- zHPF FICON Mode:
  <= 520MBps Full Duplex
  out of 800 MBps
- 200 Buffer Credits per port
  - Out to 50km
    assuming 1K frames



FICON Express8 – 4 ports
800MBps+800MBps=1600MBps

FICON Express8

- z10
- 2, 4 or 8 GBps link rate
- **Cannot Perform at 8Gbps!**
- Standard FICON Mode:
  <= 620 MBps Full Duplex
  out of 1600 MBps
- zHPF FICON Mode:
  <=770 MBps Full Duplex
  out of 1600 MBps
- **40 Buffer Credits per port**
  - Out to 5km
    assuming 1K frames



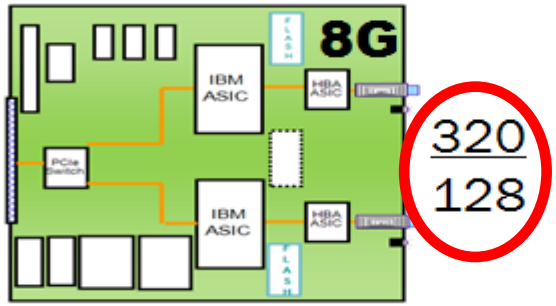
FICON Express8 – 4 ports
800MBps+800MBps=1600MBps

FICON Express8

- z196, z114
- 2, 4 or 8 GBps link rate
- **Cannot Perform at 8Gbps!**
- Standard FICON Mode:
  <= 620 MBps Full Duplex
  out of 1600 MBps
- zHPF FICON Mode:
  <=770 MBps Full Duplex
  out of 1600 MBps
- **40 Buffer Credits per port**
  - Out to 5km
    assuming 1K frames

One or more IBM Graphics are used above

Mainframe Channel Cards



FICON Express8S – 2 ports
800MBps+800MBps=1600MBps

FICON Express8S

- z196, z114
- 2, 4 or 8 GBps link rate
- **zHPF Performs at 8Gbps!**
- Standard FICON Mode:
  <= 620MBps Full Duplex
  out of 1600 MBps
- zHPF FICON Mode:
  <=1600 MBps Full Duplex
  out of 1600 MBps
- **40 Buffer Credits per port**
  - Out to 5km
    assuming 1K frames

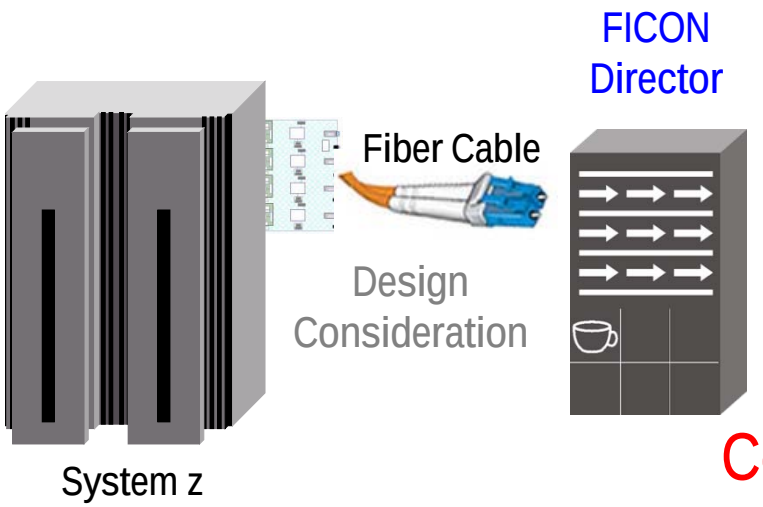
- **FICON Express8S (I call it Speedy):**
  - New IBM ASIC which supports...
  - PCIe 8 GBps host bus in a new...
  - PCIe I/O drawer
  - Increased start I/Os over FICON Express8
  - Improved throughout for zHPF and FCP
  - Increased port granularity – 2 CHPIDs/FX8S
  - Introduction of a Hardware Data Router

- The new Hardware Data Router supports the zHPF and FCP protocols providing path length reduction and increased throughput

- 2 CHPIDs/FX8S versus the 4 CHPIDs/FX8 helps facilitate purchasing the right number of ports to help satisfy your application requirements and to better optimize your infrastructure for redundancy

One or more IBM Graphics are used above

FICON/FCP Switching Devices



FICON Director Chassis Considerations

- Point-to-Point versus switched FICON connectivity
- Redundant fabrics to position for five-9s of availability
- Multimode cables and short wave SFP limitations

Switched-FICON is a Best Practice for System z

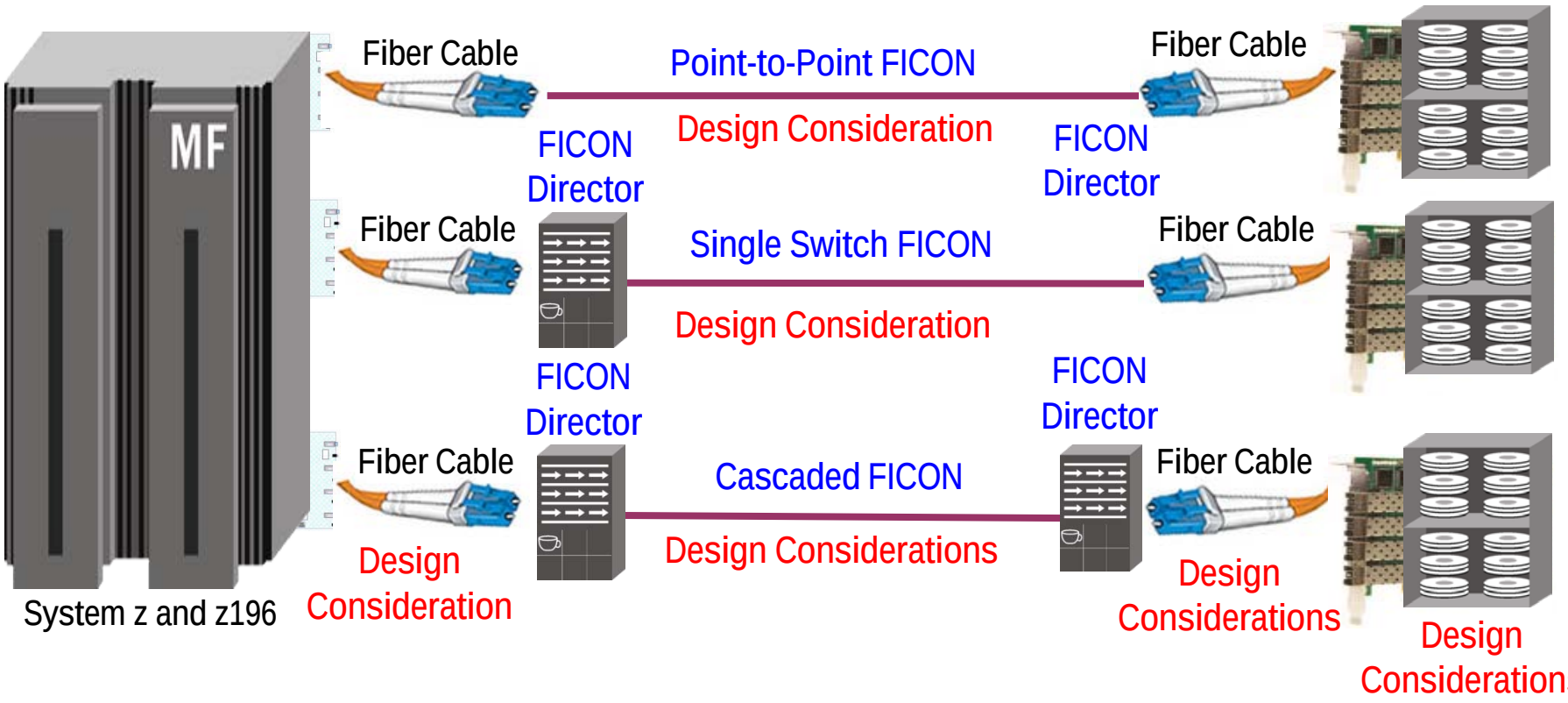
- Architected and deployed correctly, Brocade FICON switching devices do not cause performance problems in a local data center nor across very long distances
  - Cut-through frame routing and very low frame latency times
- In fact, use of Brocade switched-FICON and Brocade FCIP long distance connectivity solutions can even enhance DASD replication performance and long distance tape operations effectiveness and performance
  - XRC emulation and Tape Read and Write Pipelining (tape emulation)
- Switched-FICON is the only way to efficiently and effectively support Linux on System z connectivity
  - Makes use of Node_Port ID Virtualization (NPIV) channel virtualization
- Switched-FICON is the only way to really take advantage of the full value of the System z I/O subsystem
  - Let's see why….

Recent z/OS and System z Functionality

Some of the new z/OS and/or System z functionality will REQUIRE that a customer deploy switched –FICON:

- **FICON Express8 CHPID buffer credits**: Only 40 BCs per FICON Express8 CHPID limits long distance direct connectivity to <=10km. Use up to 1,300 port buffer credits on FICON switching devices for longer distances.
- **FICON Dynamic Channel Management**: Ability to dynamically add and remove channel resources at Workload Manager discretion can be accomplished only in switched-FICON environments.
- **zDAC**: Simplified configuration of FICON connected disk and tape through z/OS FICON Discovery and Auto Configuration (zDAC) capability of switched-FICON fabrics.
- **NPIV**: Excellent for Linux on the Mainframe, Node_Port ID Virtualization allows many FCP I/O users to interleave I/O across a single physical but virtualized channel path which minimizes the number of total channel paths
  - There is additional functionality that switched-FICON provides and we will discuss that on the following slides

End-to-End FICON Connectivity

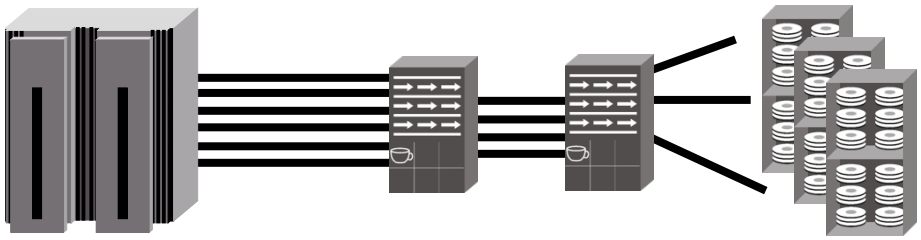


- These are the typical ways that FICON is deployed for an enterprise.
  - Long wave ports (Single Mode cables) can go from 4-100km
  - Short wave ports (Multimode cables) can go from 50-500 meters

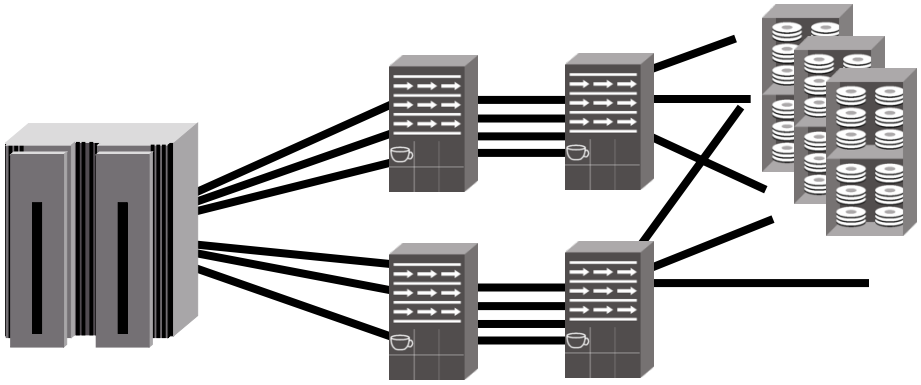
Native FICON with Simple Cascading (FC)

- Uses FICON switching devices
- Single fabrics provide no more than four-9s of availability – if a switching devices fails (a very rare occurrence) it could take down all connectivity [1]
- Redundant fabrics might provide five-9s of availability – a fabric failure would not take down all connectivity – but, loss of bandwidth is another consideration to create five-9s environments

Switched-FICON and a Cascaded FICON Fabric

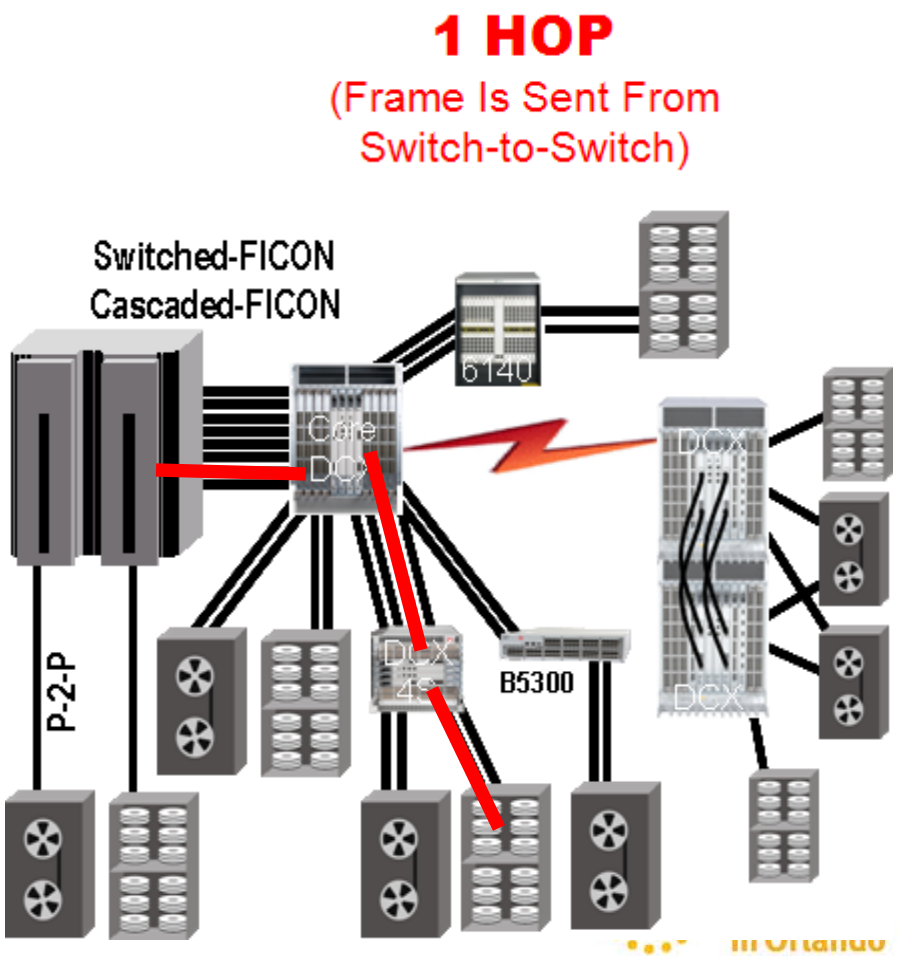


Redundant Switched-FICON and Cascaded FICON Fabrics

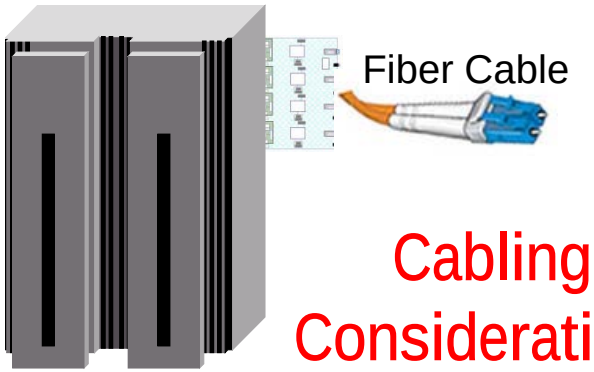


Native FICON with Cascading

- Scalable FICON benefits.
- Cascaded FICON allows:
  - Scalability of resources
  - Ease of growth and change
  - Multiple protocols
  - Support for dynamic connectivity to a local or remote environment
- Notice that there can be several switches/Directors attached to a core Director but there can only be 1 hop (switch to switch) between a CHPID and a storage port



Multi-mode cable distance limitations



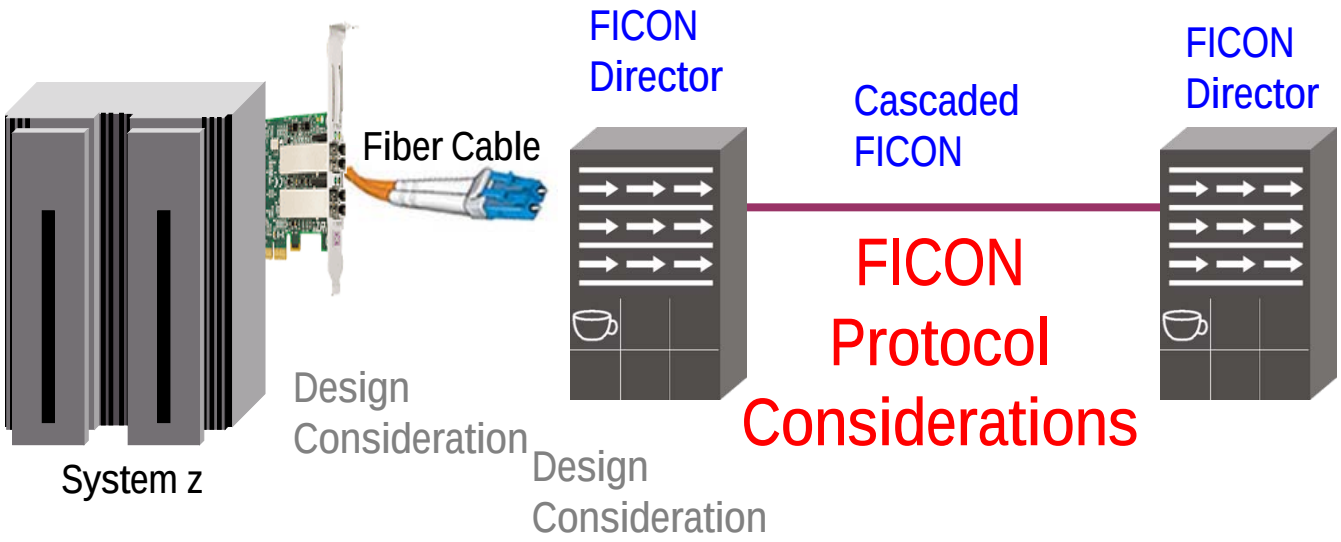
Cabling Considerations

- Long wave single mode (SM) still works well
  - 1/2/4/8/10 Gbps out to 10km with SM
- *Short wave multi-mode might be limiting!*
- 4G optics auto-negotiate back to 1G and 2G
- 8G optics auto-negotiate back to 2G and 4G
  - 1G storage connectivity requires 4G SFPs
- 16G optics will auto-negotiate back to 4G and 8G
  - 2G storage connectivity will require 8G SFPs

Distance with Multi-Mode Cables (feet/meters)

Protocol (FC)	Encoding	Line Rate (Gb/sec)	OM1-62.5m (200mHz) Multi-Mode	OM2-50m (500mHz) Multi-Mode	OM3-50m (2000mHz) Multi-Mode	OM4-50m (4700mHz) Multi-Mode
1G	8b10b	1.0625	984/300	1640/500	2822/860	~
2G	8b10b	2.125	492/150	984/300	1640/500	~
4G	8b10b	4.25	230/70	492/150	1247/380	1312/400
8G	**8b10b**	**8.5**	**69/21**	**164/50**	**492/150**	**656/200**
10G	64b66b	10.53	108/33	269/82	~984/300	~984/300
16G	64b66b	14.025	34.5/10.5	82/25	328/100	427/130

End-to-End FICON/FCP Connectivity



- **With 8b10b, ~ 25% overhead per full frame on FICON links**
- **With 64b66b, ~ 3% overhead per full frame on FICON links**

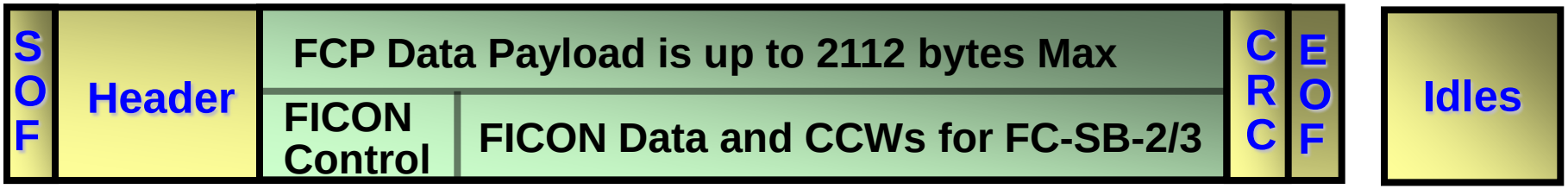
Can use 2 / 4 / 8G and/or 10G for ISL traffic today

The FICON Protocol uses 8b10b data encoding for most link rates – but there is 25% frame payload overhead associated with it

Newer 64b66b data encoding (10G and 16G) is also in use and is more performance oriented (only 3% data payload overhead)

MIDAW & zHPF make very good use of 8G FICON switch links

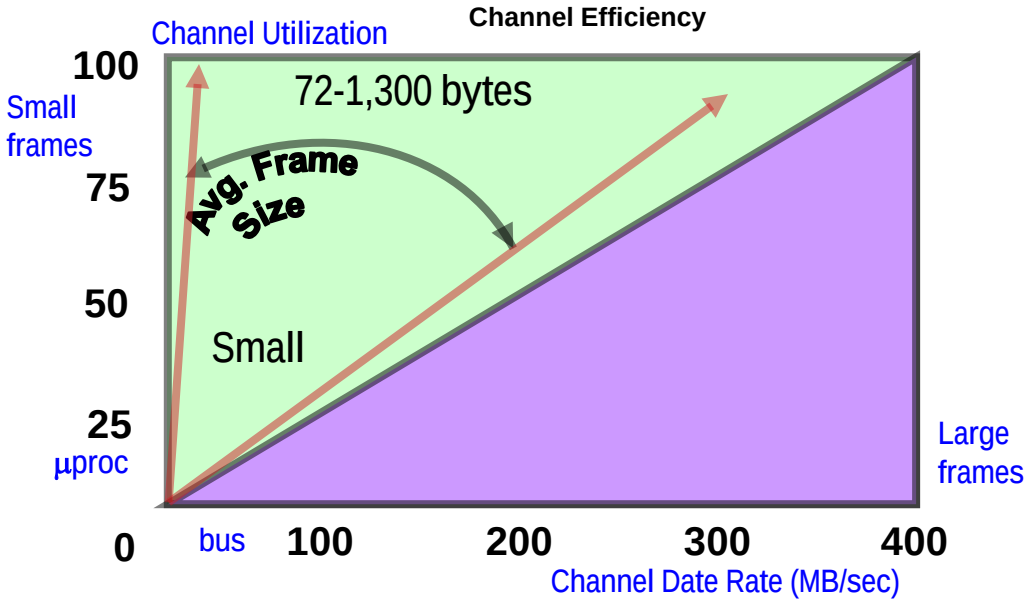
FICON FC-SB-2/3 – Channel Efficiency



64 bytes <===FICON: 2048 bytes Max =====>

<=== 2112 bytes without FICON Control ===>

<==Except for 1st frame, 2084 bytes Max out of 2148 possible==>



FC-SB-2/3

FC-SB-2/3 FICON tends to have an average frame size of between 72 and 1400 bytes

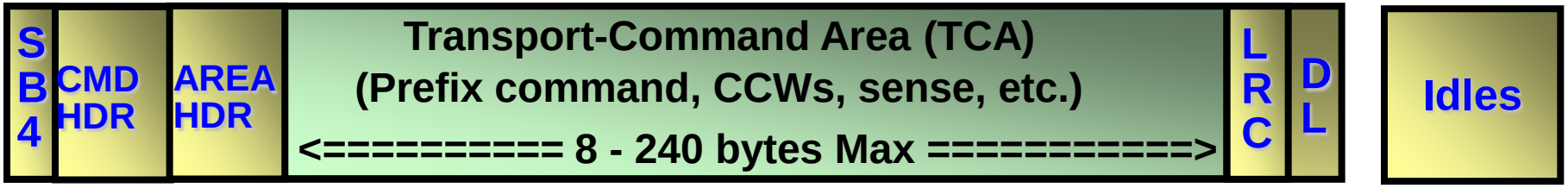
FC-SB-2/3 is used for all FICON BSAM, QSAM and EXCP datasets

High Performance FICON (zHPF)

- Available since October 2008
  - Partly z/OS IOS code and partly DASD control unit code
  - Available on specific IBM, HDS and EMC DASD units
- zHPF is a performance, reliability, availability and serviceability (RAS) enhancement of the z/Architecture and the FICON channel architecture
- It is implemented exclusively in System z10, z196 and z114
- Exploitation of zHPF by the FICON channel, the z/OS operating system, and the DASD control unit is designed to help reduce the FICON channel overhead
  - This is achieved through protocol simplification and a reduced number of information units (IUs) processed, resulting in more efficient use of the channel

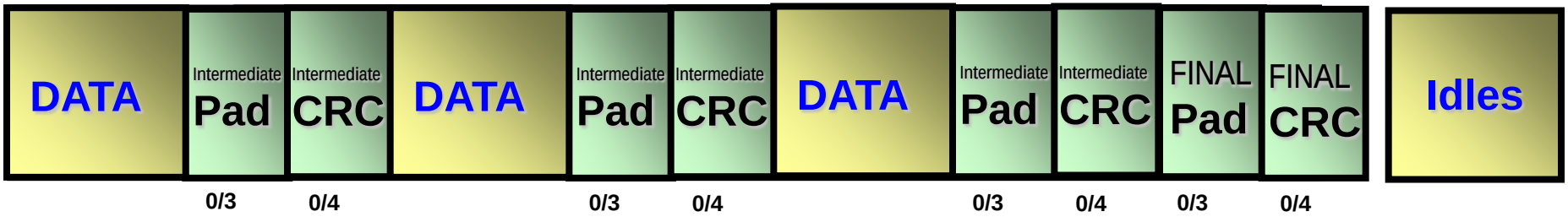
FICON FC-SB-4 zHPF – More Data, Fewer Frames

FICON Transport-Command IU for FC-SB-4

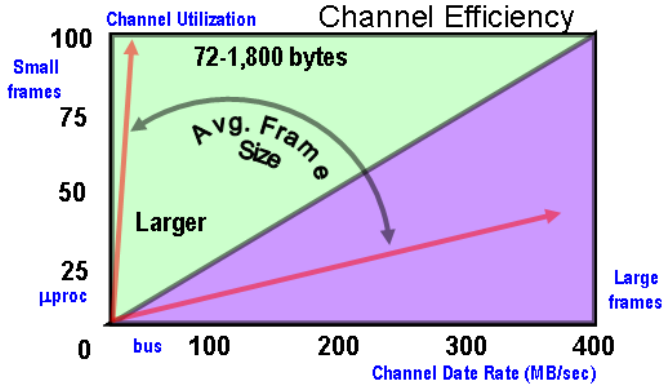


<========== 44 - 276 bytes Max ==========>

FICON Transport-Data IU for FC-SB-4 – larger average frame sizes



<==========0 - 4GB (-16 bytes) Max ==========>

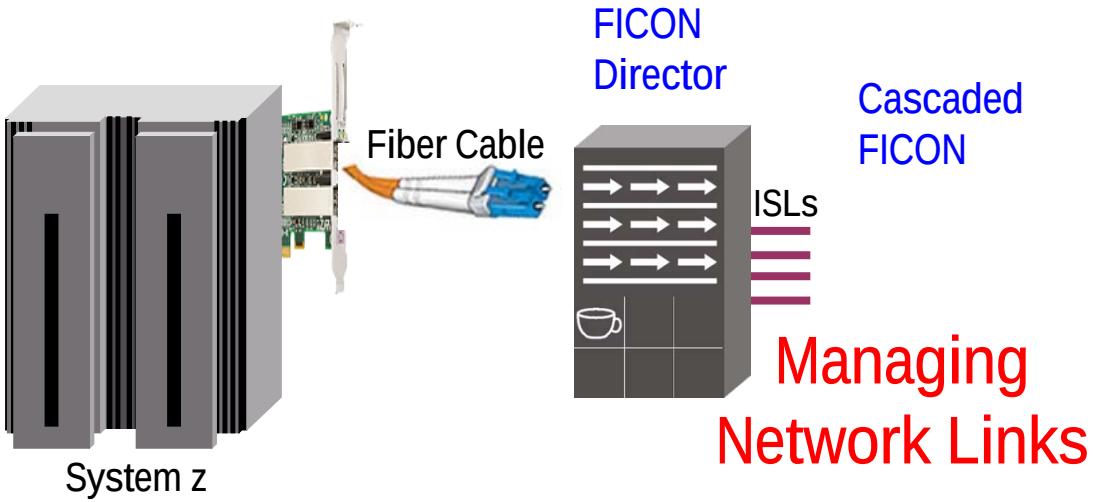


FC-SB-4

FC-SB-4 FICON tends to have an average frame size of between 72 and 1,800 bytes

FC-SB-4 used for only FICON Media Manager Datasets like VSAM, DB2, PDSE, zFS and Extended Format SAM

End-to-End FICON/FCP Connectivity



Topics in this section

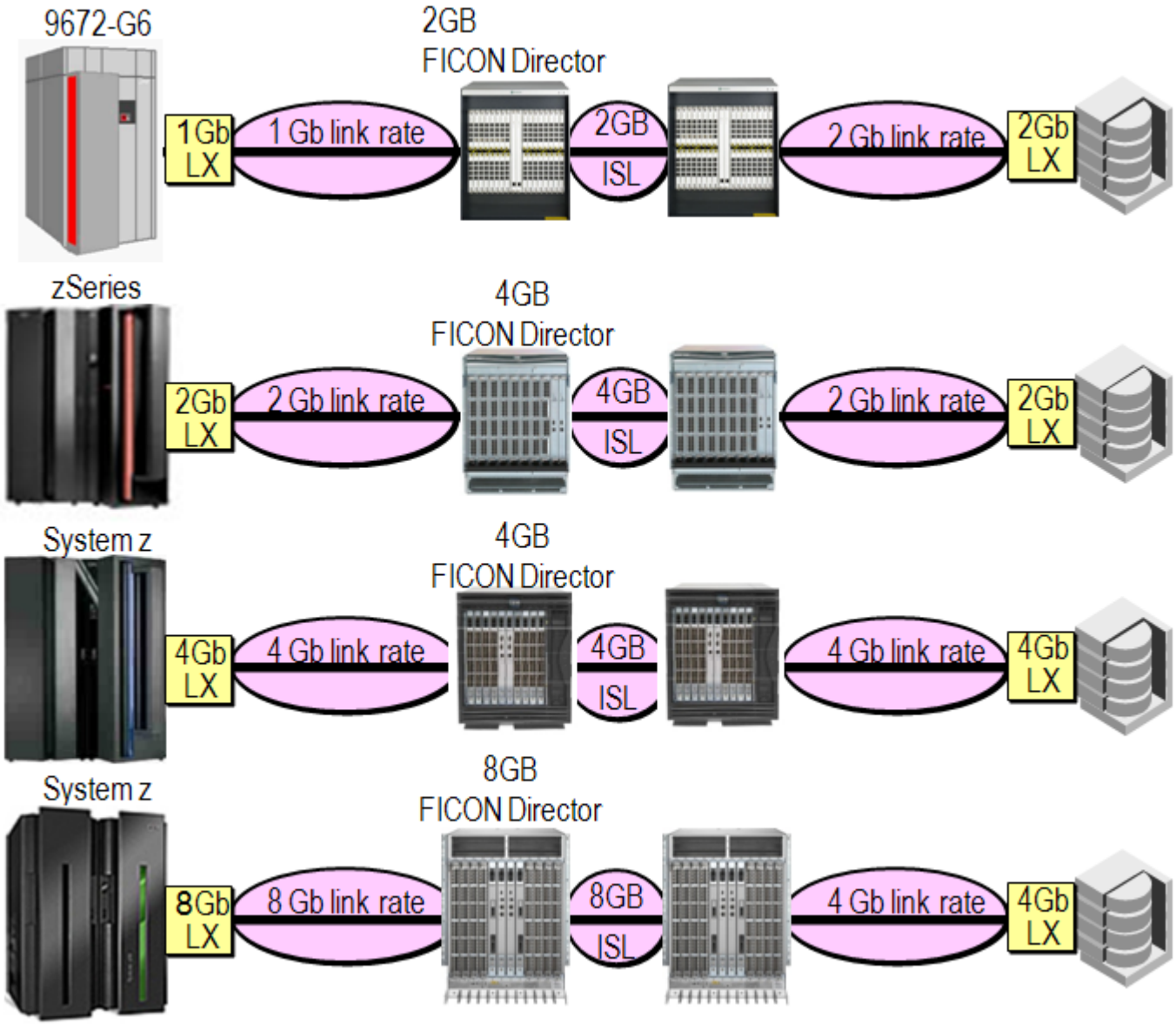
- **End-to-End Link speeds**
- **FICON Fabric Scalability**
- **Hops and hop issues**
- **Managing ISL Congestion**
- **Trunking**
- **Protocol Intermixed FICON Fabrics**
- **Buffer Credits**
- **Control Unit Port (CUP)**
- **Distance Extension**

Here we are at cascaded links (ISLs)

There are too many design considerations with switch-to-switch and data center-to-data center connectivity to do it all today

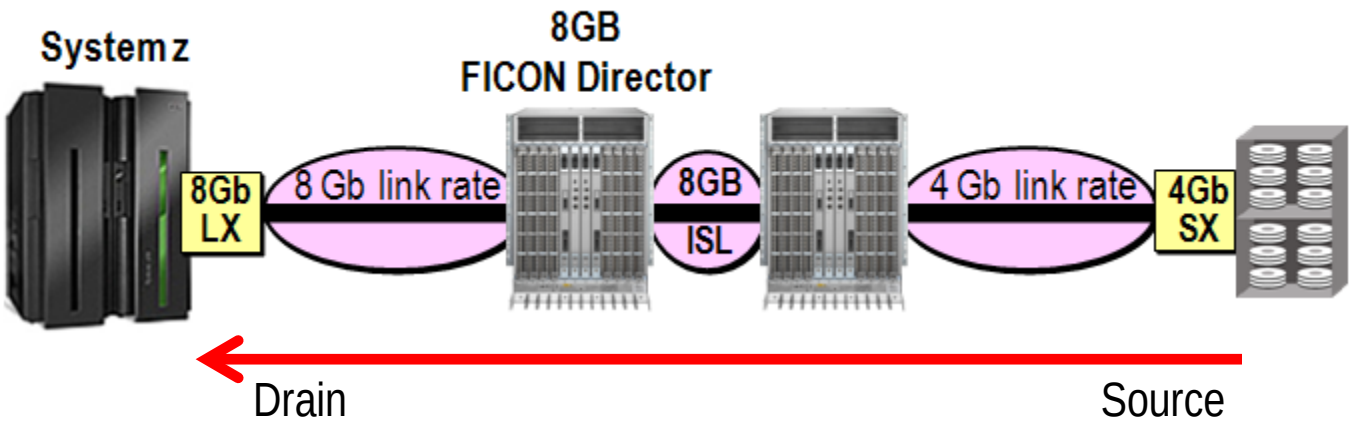
I will spend a moment to discuss end-to-end link rates.

Maximum End-to-End Link Rates



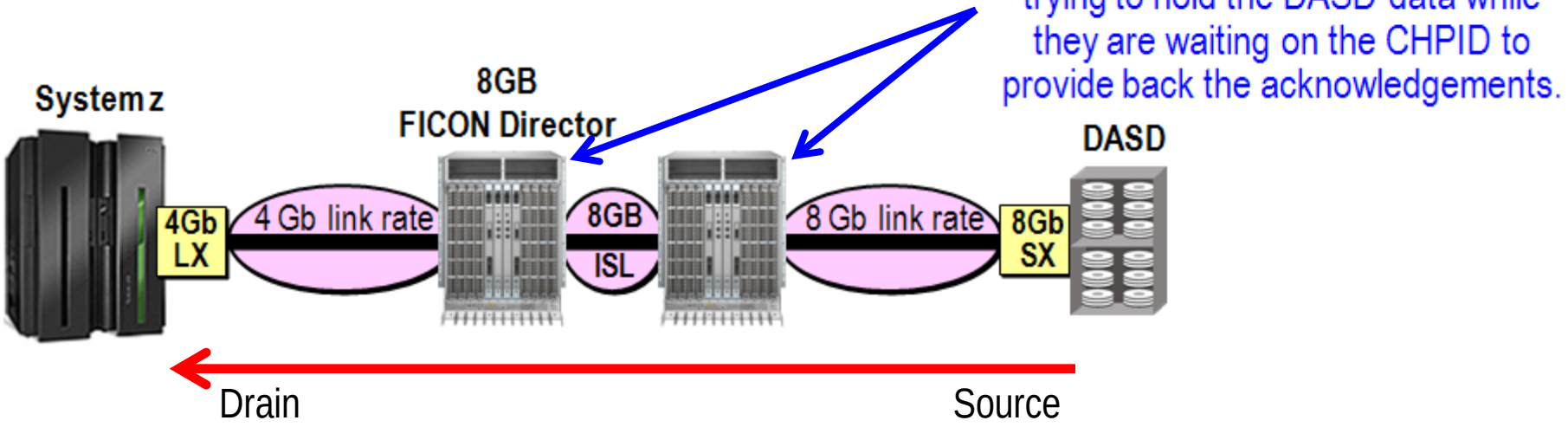
- Best link rate performance is achieved when the channel, switch and control unit all operate at the same link rate
- Link rate does not guarantee that data will flow at that speed
- Take the speed of the local …AND… cascaded links into consideration.
- Cascaded Links will flow at their rated speeds even when connected with ports of lower speed

Will This Be A Slow Drain Device



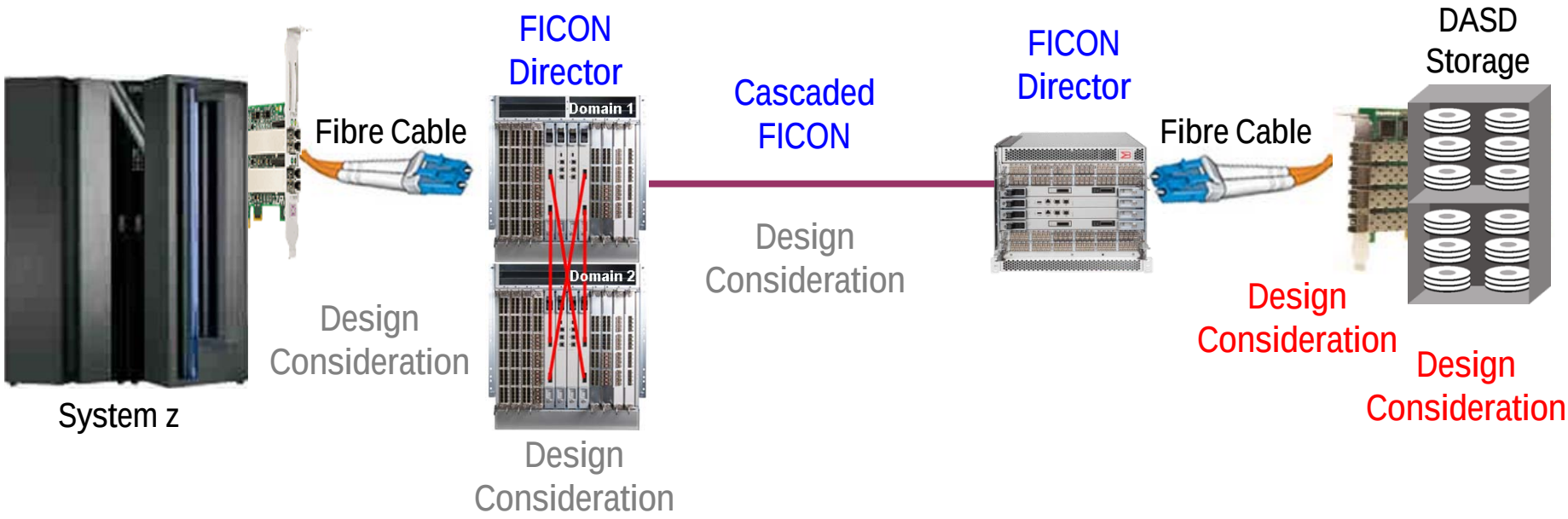
- Assuming no ISL or BC problems, and assuming the normal and typical use of DASD, is the above a good configuration?
- If you deployed this configuration, is there a probability of performance problems and/or slow draining devices or not?
- This is actually the ideal model!
- Most application profiles are 90% read, 10% write. So, in this case the "drain" of the pipe are the 8Gb CHPIDs and the "source" of the pipe are 4Gb storage ports.
- This represents an end-to-end network that will generally require the least amount of buffer credit pacing (assuming you implemented the correct number of ISLs)

Will This Be A Slow Drain Device



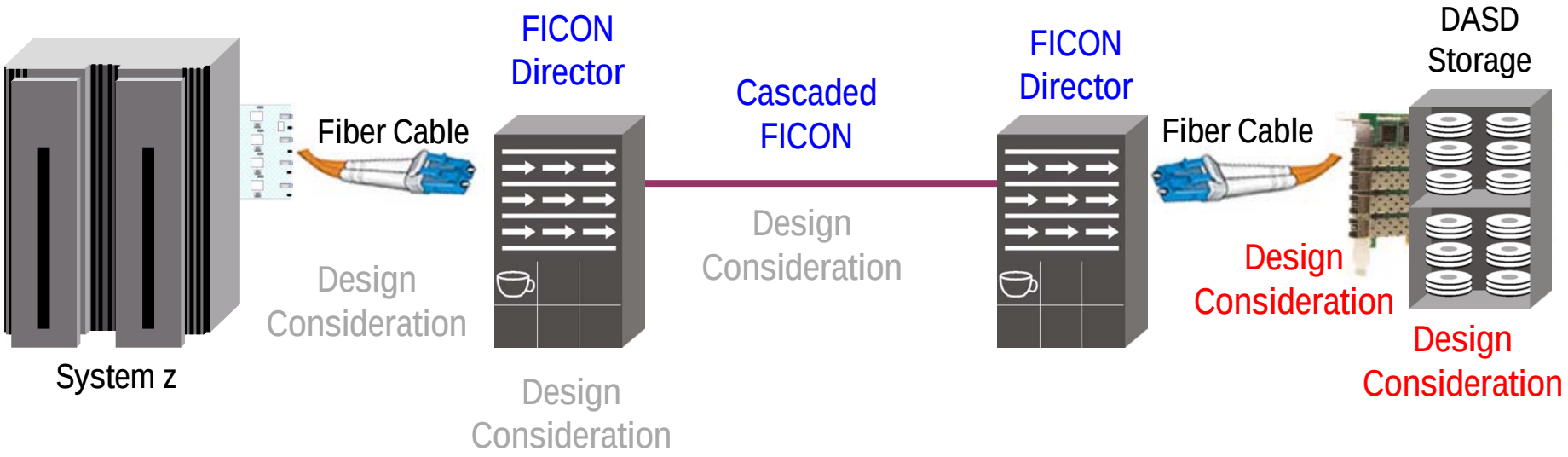
- Assuming no ISL or BC problems, and assuming the normal and typical use of DASD, is the above a good configuration?
- If you deployed this configuration, is there a probability of performance problems and/or slow draining devices or not?
- This is potentially a very poor performing, infrastructure!
- Again, DASD is about 90% read, 10% write. So, in this case the "drain" of the pipe are the 4Gb CHPIDs and the "source" of the pipe are 8Gb storage ports.
- The Source can out perform the Drain. This can cause congestion and back pressure towards the CHPID. The CHPID becomes a slow draining device.

End-to-End FICON/FCP Connectivity



- Your most challenging considerations most likely occur due to DASD storage deployment

Connectivity with storage devices



Storage adapters can be throughput constrained

- Must ask storage vendor about performance specifics
- Is zHPF supported/enabled on your DASD control units?

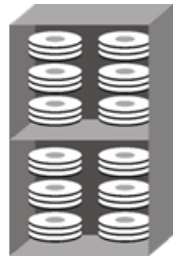
Busy storage arrays can equal reduced performance

- RAID used, RPMs, volume size, etc.
- Let's look a little closer at this

Connectivity with storage devices

How fast are the Storage Adapters?

- Mostly 2 / 4Gbps today – some 8G – where are the internal bottlenecks



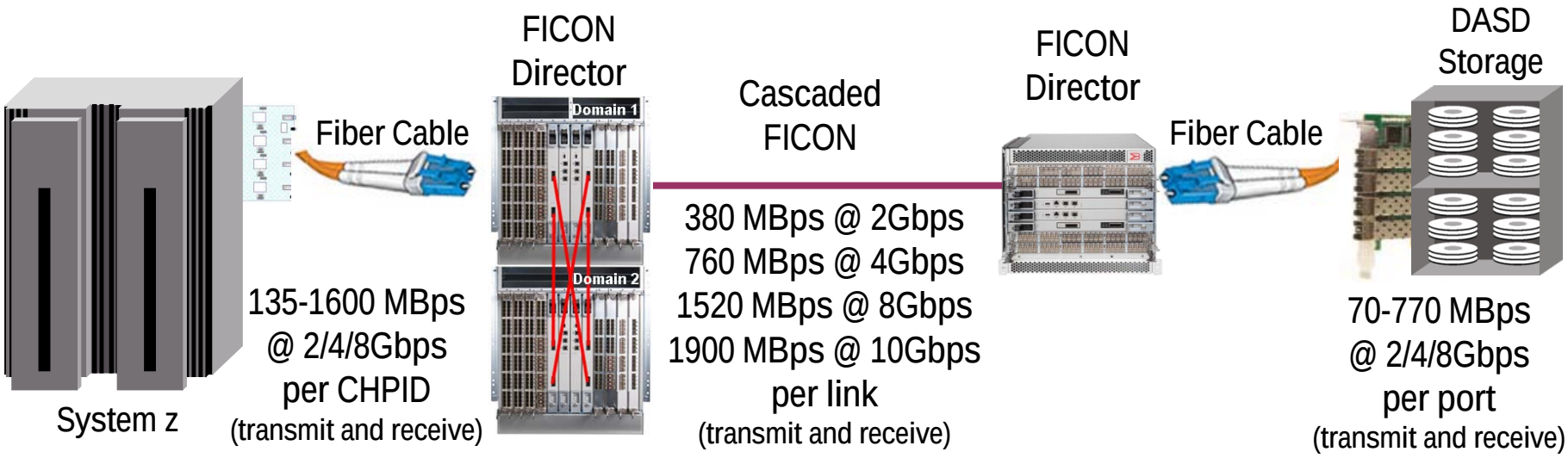
What kinds of internal bottlenecks does a DASD array h

- 7200rpm, 10,000rpm, 15,000rpm
- What kind of volumes: 3390-3; 3390-54; EAV; XIV
- How many volumes are on a device? HiperPAV in use?
- How many HDDs in a Rank (arms to do the work)
- What Raid scheme is being used (RAID penalties)?
- Etc.

Intellimagic or Performance Associates, for example, can provide you with great tools to assist you to understand DASD performance much better

These tools perform mathematical calculations against raw RMF data to determine storage HDD utilization characteristics – use them or something like them to understand I/O metrics!

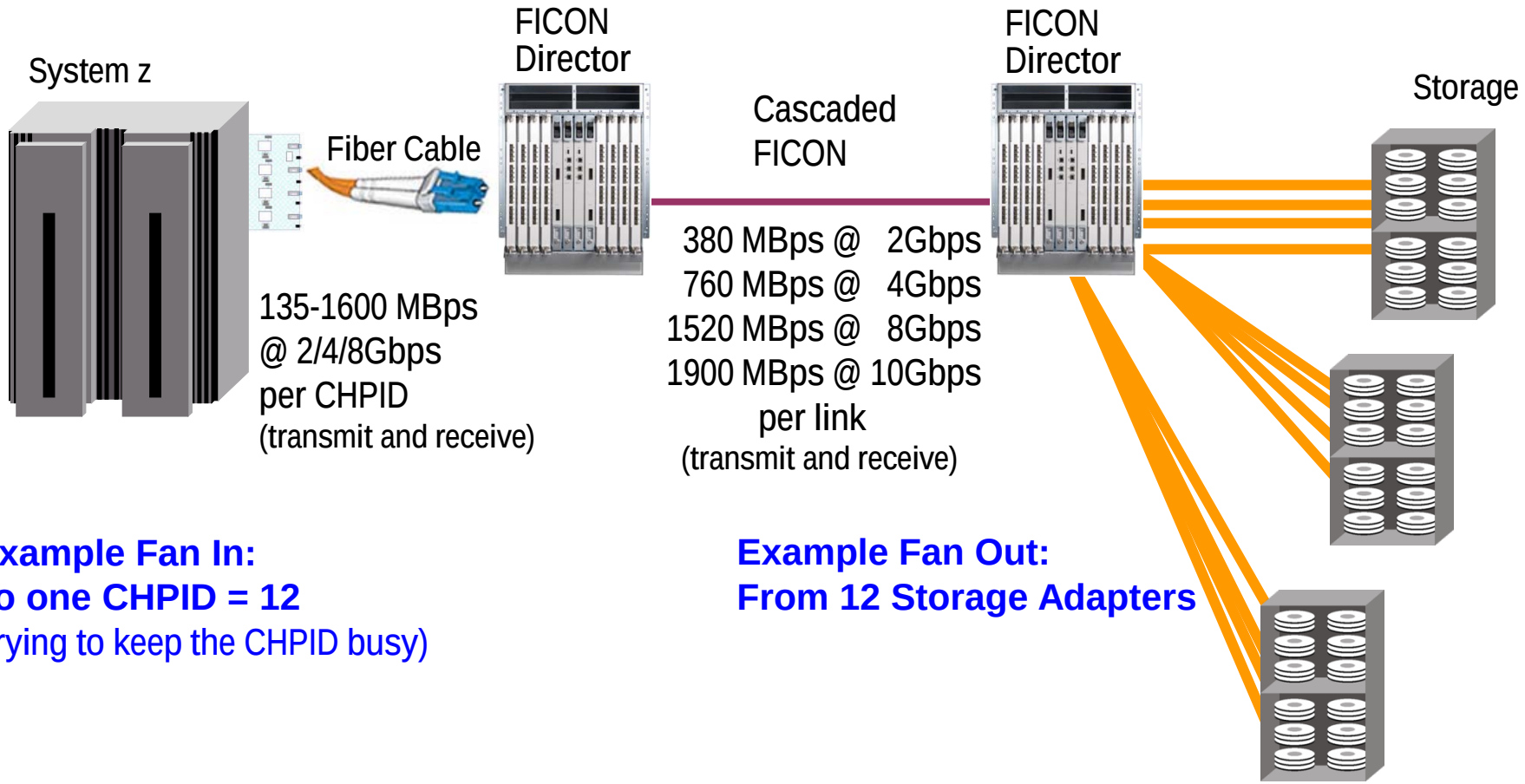
End-to-End FICON/FCP Connectivity



- In order to fully utilize the capabilities of a FICON fabric a customer needs to deploy a Fan In – Fan Out Architecture
- If you are going to deploy Linux on System z, or private cloud computing, then switched FICON flexibility is required!

FICON should just never be direct attached!

FI-FO Overcomes System Bottlenecks



Example Fan In:
To one CHPID = 12
(trying to keep the CHPID busy)

Example Fan Out:
From 12 Storage Adapters

- Total FICON path usually does not support full speed
  - Must deploy Fan In – Fan Out to utilize connections wisely
    - Multiple I/O flows funneled over a single channel path

BROCADE

Brocade's Mainframe Certification

Industry Recognized Professional Certification

We Can Schedule A Class In Your City – Just Ask!

SHARE
Technology · Connections · Results

Brocade FICON Certification

Brocade Certified Architect for FICON

Certification for Brocade Mainframe-centric Customers – Available since Sept 2008

For people who do or will work in FICON environments

Brocade provides a free on-site or in area 2-day class (Brocade Design and Implementation for FICON Environments – FCAF200), to assist customers in obtaining the knowledge to pass this certification examination – ask your local sales team about this training – also look at www.brocade.com under Education

Certification tests a person's ability to understand IBM System z I/O concepts, and demonstrate knowledge of Brocade FICON Director and switching fabric components

After the class a participant should be able to design, install, configure, maintain, manage, and troubleshoot Brocade hardware and software products for local and metro distance (100 km) environments

Check the following website for complete information:

- http://www.brocade.com/education/certification-accreditation/certified-architect-ficon/index.page

BROCADE

Thank You !

.....My Next Presentation.....

SHARE
Technology • Connections • Results

A deeper look into the Inner Workings and Hidden Mechanisms of FICON Performance

- David Lytle, BCAF
- Brocade Communications Inc.
- Tuesday August 9, 2011 -- 3pm to 4pm
- Session Number - 10079

More SAN Sessions at SHARE this week

Tuesday:

Time-Session

1500 - 10079: A deeper look into the Inner Workings and Hidden Mechanisms of FICON Performance

Wednesday:

Time-Session

0800 - 9479: Planning and Implementing NPIV for System Z

0930 - 9864: zSeries FICON and FCP Fabrics - Intermixing Best Practices

Thursday:

Time-Session

0800 - 9853: FICON Over IP - Technology and Customer Use

0800 - 9899: Planning for ESCON Elimination

0930 - 9933: Customer Deployment Examples for FICON Technologies

1500 - 9316: SAN Security Overview

1630 - 10088: FICON Director and Channel Free-for-all

Please Fill Out Your Evaluation Forms!!

SHARE
Technology · Connections · Results

This was session: 09368

And Please Indicate On Those Forms If There Are Other Presentations That You Would Like To See In This SAN Track At SHARE.

Thank You.