System z FICON Fabric Performance Considerations

Architecting End-to-End Performance

David Lytle, BCAF
Global Solutions Architect
System z Technologies and Solutions
Brocade Communications, Inc.

Wednesday March 2, 2011
Session Number 8486

Legal Disclaimer

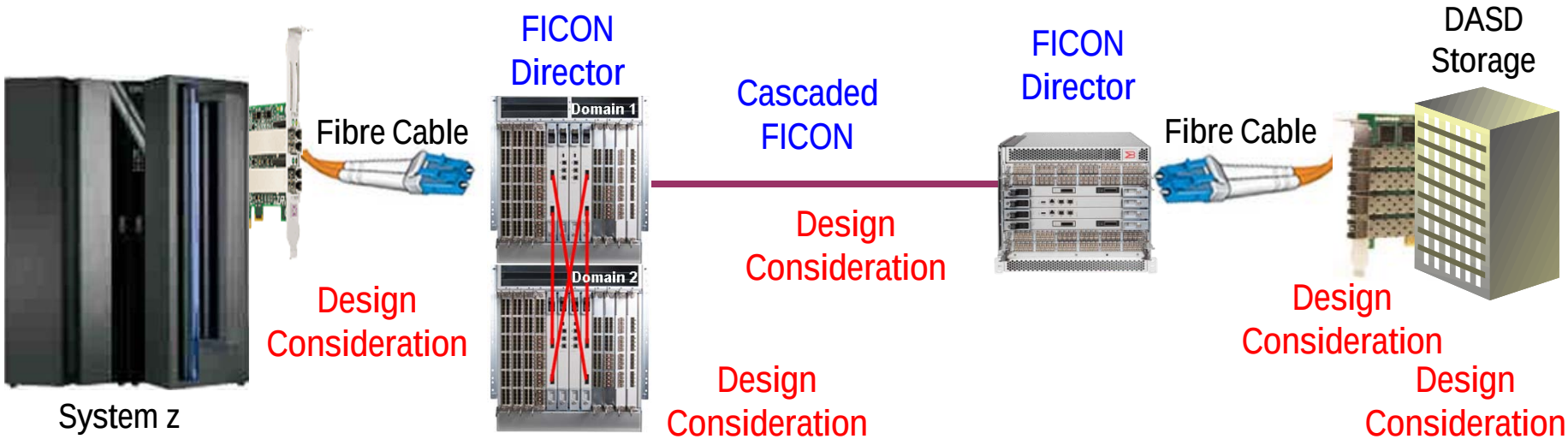
- All or some of the products detailed in this presentation may still be under development and certain specifications, including but not limited to, release dates, prices, and product features, may change. The products may not function as intended and a production version of the products may never be 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.

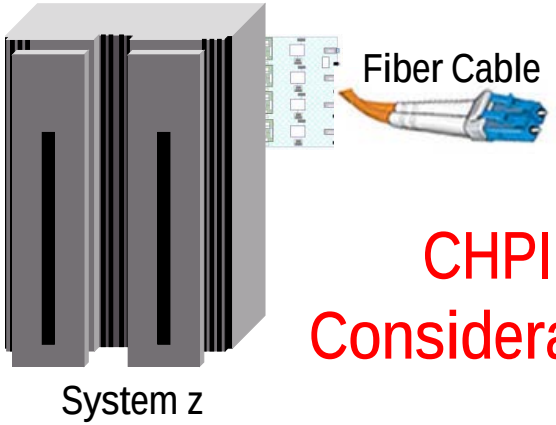
- There are slides in this presentation that use IBM graphics.

End-to-End FICON/FCP Connectivity

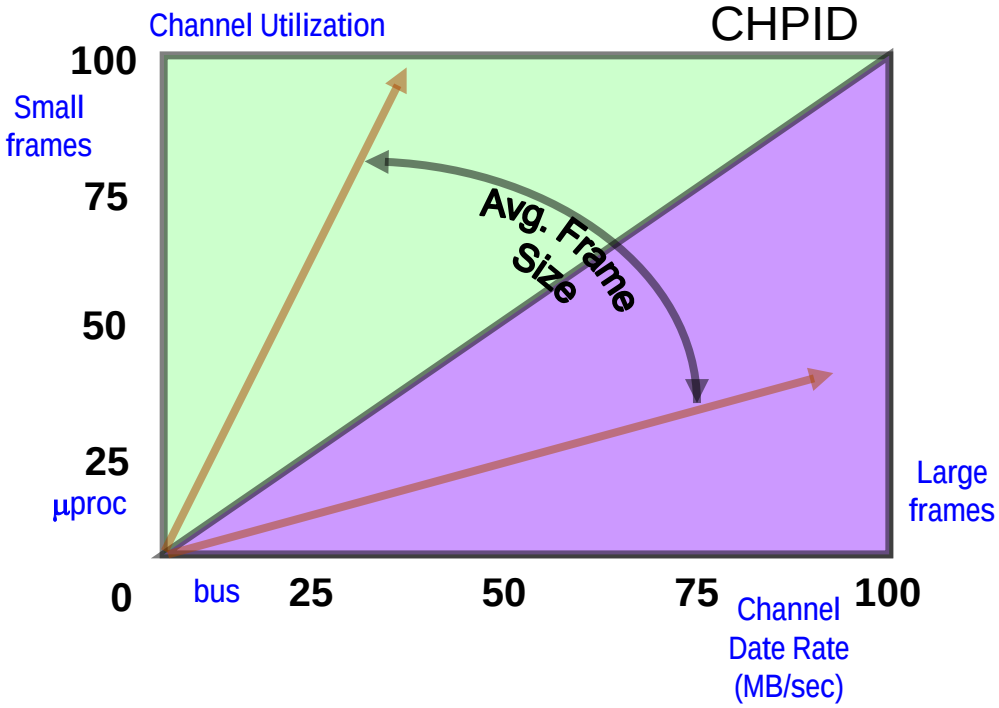


- From End-to-End in a FICON infrastructure there are a series of Design Considerations that you must understand in order to successfully meet your expectations with your FICON fabrics
- This is an OVERVIEW – 1 hour is not enough!

End-to-End FICON/FCP Connectivity

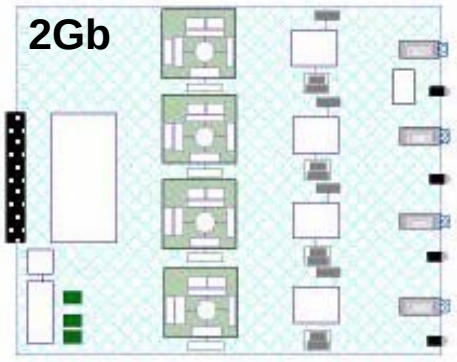


CHPID Considerations



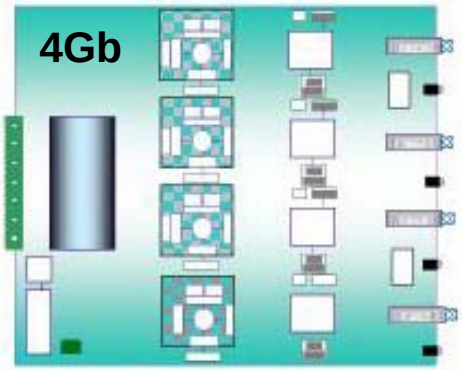
- Channel Microprocessors and PCI Bus
- Average frame size for FICON
- Buffer Credit considerations

Current Mainframe Channel Cards (Features)



FICON Express2

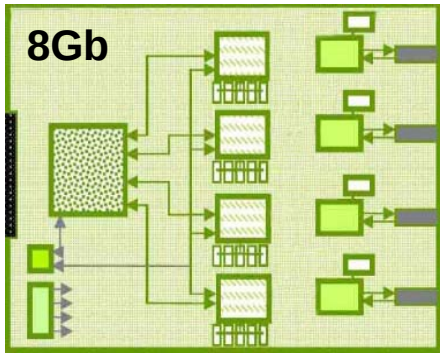
- z10, z9, z990, z890
- Longwave (LX) to 10km
- Shortwave (SX)
- 1 or 2 GBps link rate



FICON Express4

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

FICON Express4 provides the last native 1Gbps CHPID support

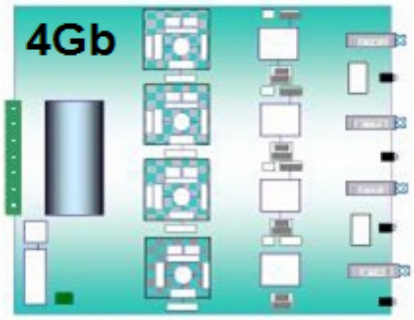


FICON Express8

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

FICON buffer credits have become very limited per CHPID

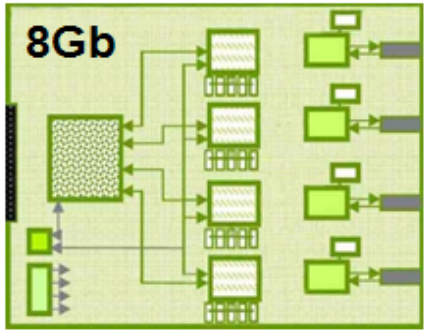
Mainframe Channel Cards



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

FICON Express4

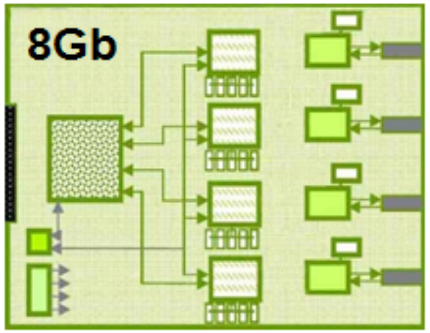
- z10, z9
- 1, 2 or 4 GBps link rate
- **Cannot Perform at 4Gbps!**
- Standard FICON Mode:
  <= 350MBps Full Duplex
  out of 800 MBps
- zHPF FICON Mode:
  <= 520MBps Full Duplex
  out of 800 MBps
- 200 Buffer Credits per port
  - Out to 50km
    assuming 1K frames



FICON Express8 – 4 ports
800MBps+800MBps = 1,600MBps

FICON Express8

- z10
- 2, 4 or 8 GBps link rate
- **Cannot Perform at 8Gbps!**
- Standard FICON Mode:
  <= 510 MBps Full Duplex
  out of 1600 MBps
- zHPF FICON Mode:
  <=740 MBps Full Duplex
  out of 1600 MBps
- **40 Buffer Credits per port**
  - Out to 5km
    assuming 1K frames

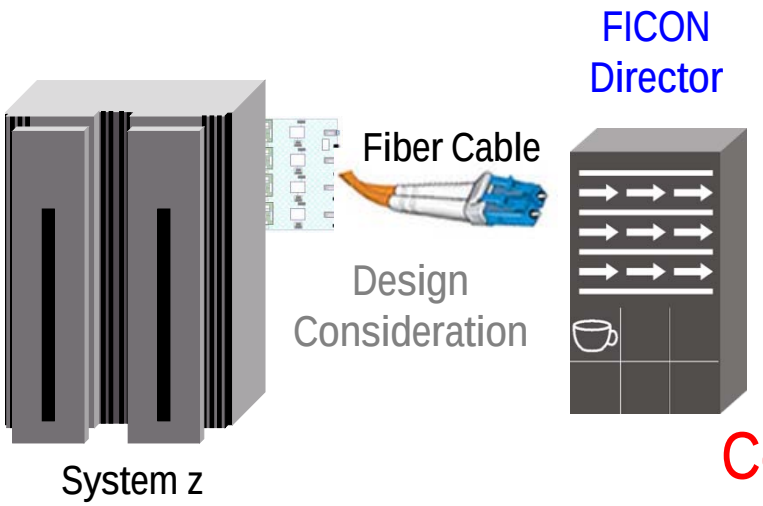


FICON Express8 – 4 ports
800MBps+800MBps = 1,600MBps

FICON Express8

- z196
- 2, 4 or 8 GBps link rate
- **Cannot Perform at 8Gbps!**
- Standard FICON Mode:
  <= 510 MBps Full Duplex
  out of 1600 MBps
- zHPF FICON Mode:
  <=740 MBps Full Duplex
  out of 1600 MBps
- **40 Buffer Credits per port**
  - Out to 5km
    assuming 1K frames

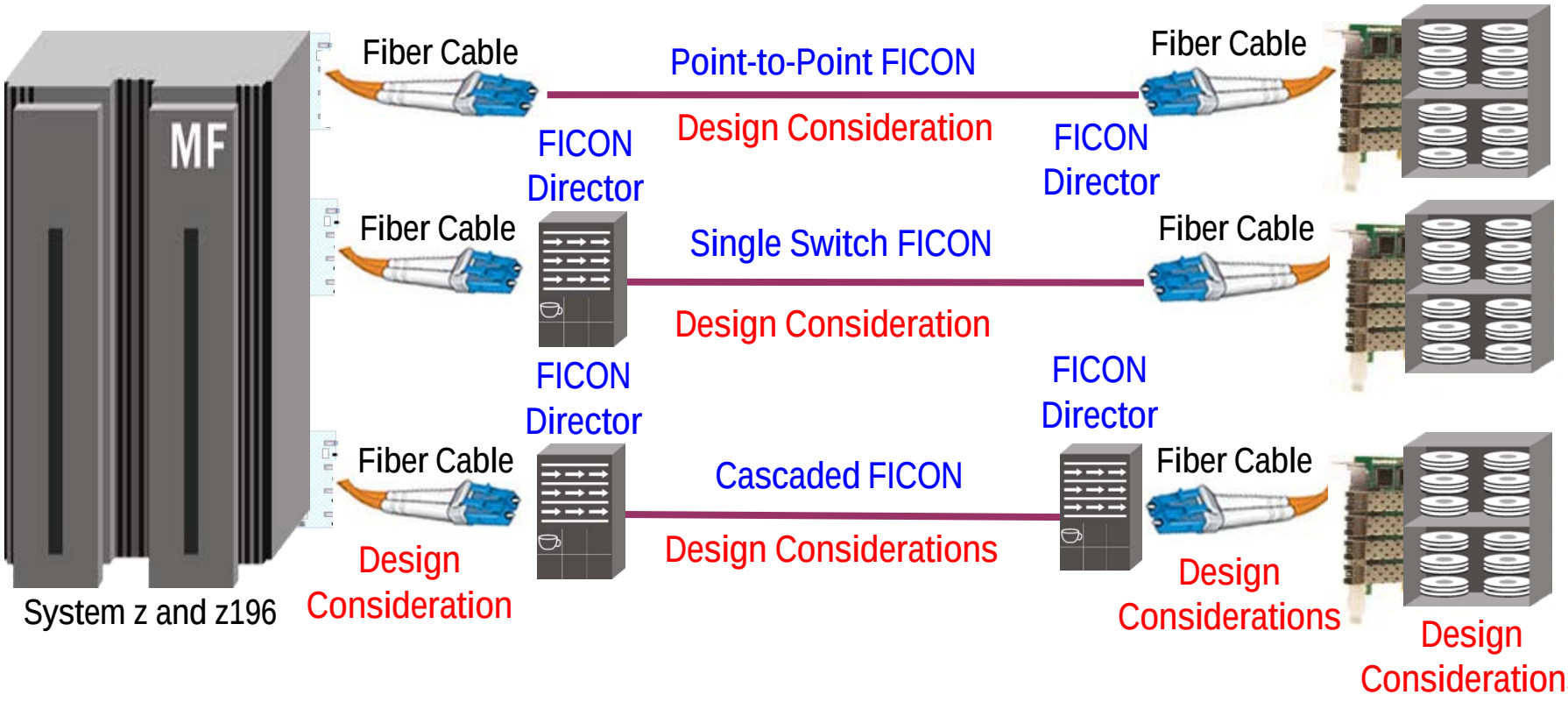
FICON/FCP Switching Devices



FICON
Director
Chassis
Considerations

- Point-to-Point versus switched FICON connectivity
- Provisioning for five-9s of availability
- Multimode cables and short wave SFP limitations
- Buffer Credits
- Control Unit Port (CUP)

End-to-End FICON Connectivity

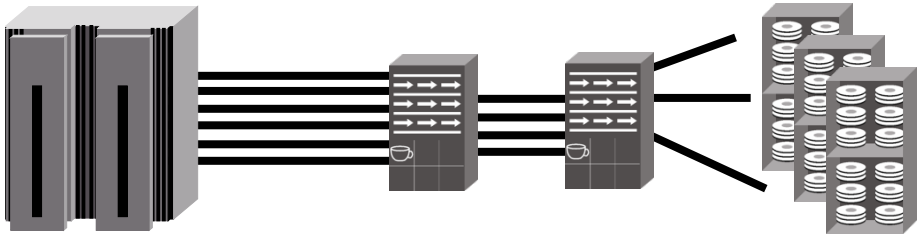


- These are the typical ways that FICON is deployed for an enterprise.
  - Long wave ports (Single Mode cables) can go from 4-100km
  - Short wave ports (Multimode cables) can go from 50-500 meters

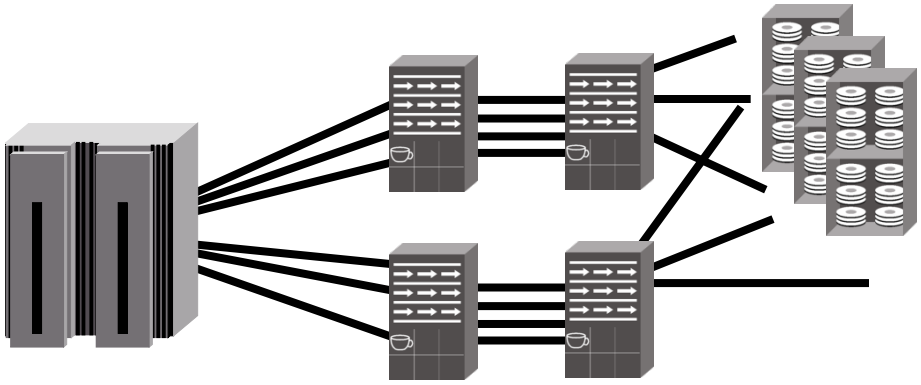
Native FICON with Simple Cascading (FC)

- 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…there are other considerations for five-9s environments

Switched-FICON and a Cascaded FICON Fabric

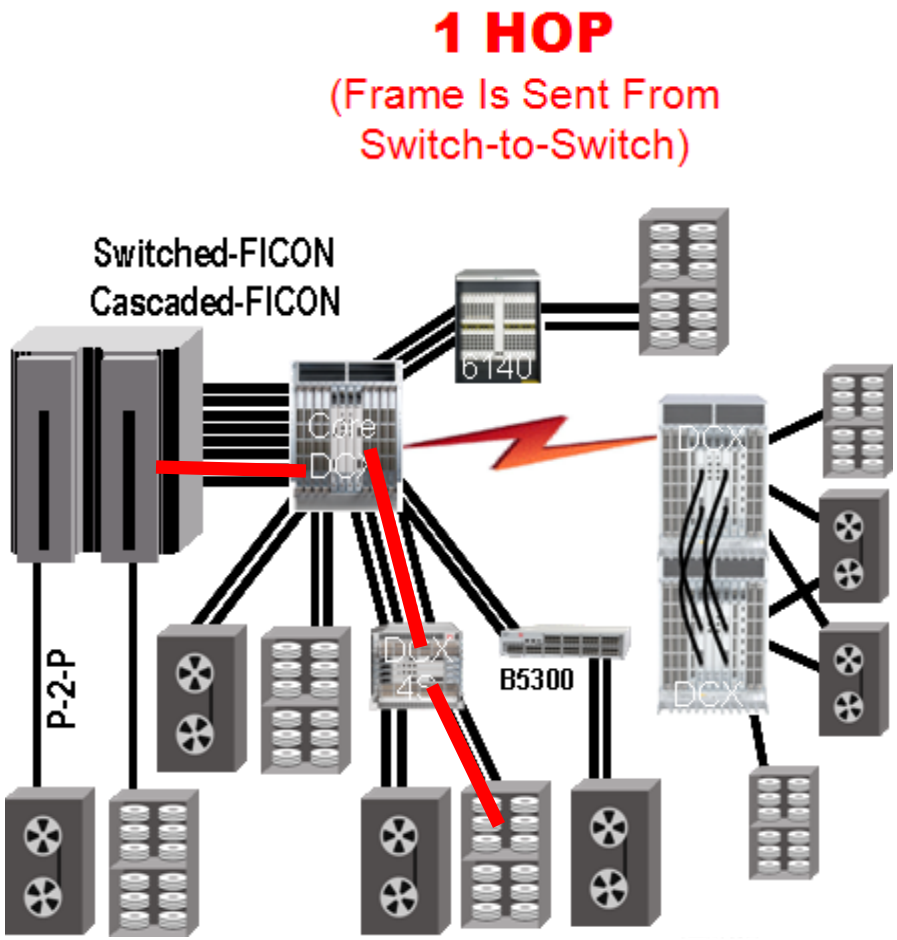


Redundant Switched-FICON and Cascaded FICON Fabrics



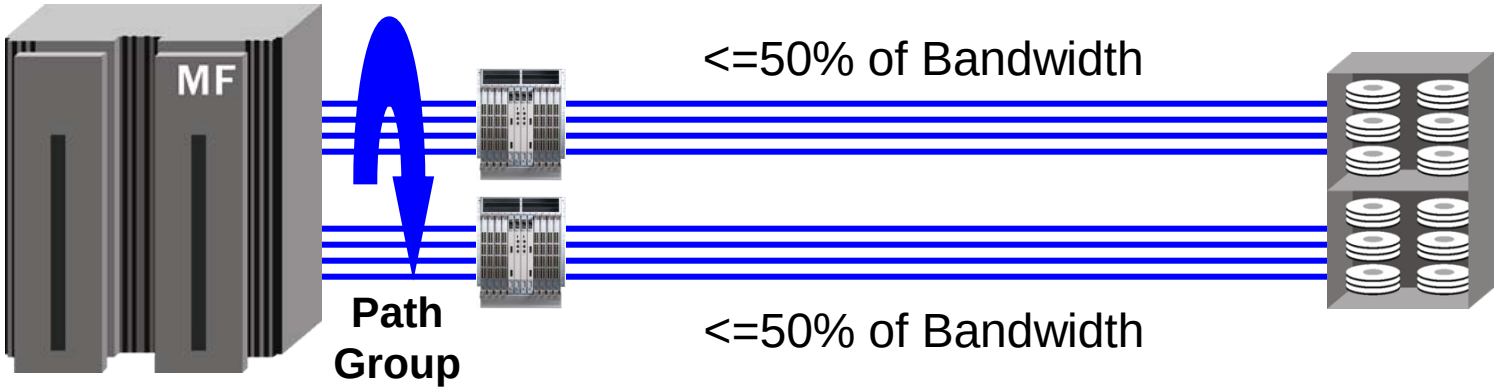
Native FICON with Cascading

- Utilizes full FICON benefits.
  It allows:
  - Scalability.
  - Multiple protocols.
  - Optimized management.
  - Supports 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



Provisioning for Connectivity Bandwidth Redundant fabrics

Provides a possible five-9's of availability environment



But each fabric really needs to run at no more than 45% busy so that if a failover occurs then the remaining fabric can pickup and handle the full workload

If either fabric runs higher than ~45% then this no longer provides you with what you consider to be five-9s

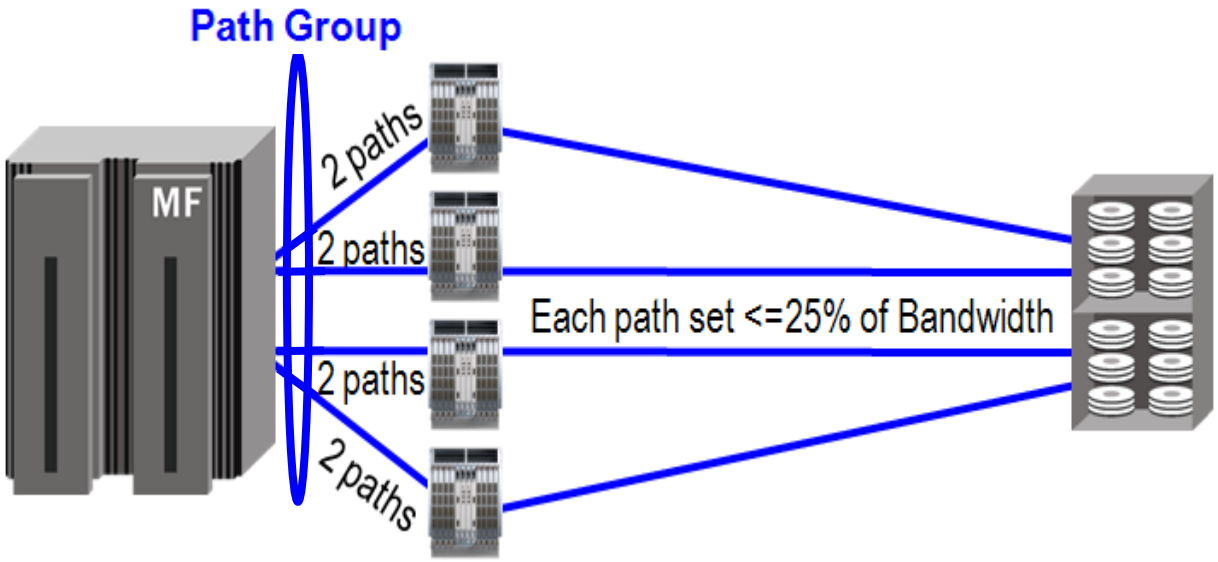
z/OS's IOS automatically load balances the FICON I/O across all of the paths in a Path Group (up to 8 channels in a PG)

Multiple FICON Fabrics – Not Just Redundant

Risk of Loss of Bandwidth is the motivator for deploying FICON fabrics like this.

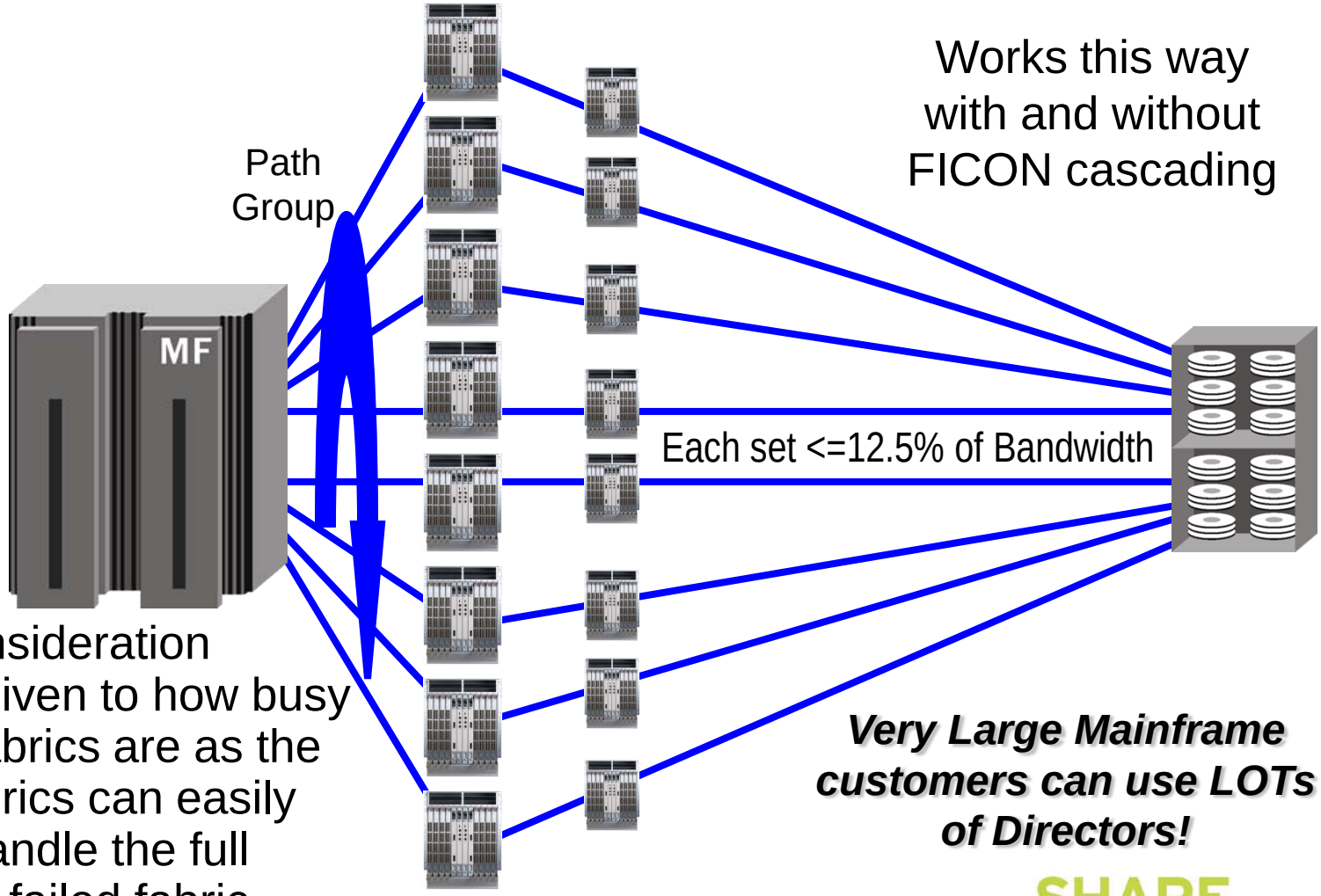
In this case, 2 paths from an 8 path Path Group are deployed across four FICON fabrics to limit bandwidth loss to no more than 25% if a FICON fabric were to fail.

Each fabric needs to run at no more than ~85% busy so that if a failover occurs then the remaining fabrics can pickup and handle the full workload without over-utilization.



z/OS's IOS automatically load balances the FICON I/O across all of the paths in a Path Group (up to 8 channels in a PG)

Provisioning for Connectivity Bandwidth
Deploying to minimize bandwidth loss



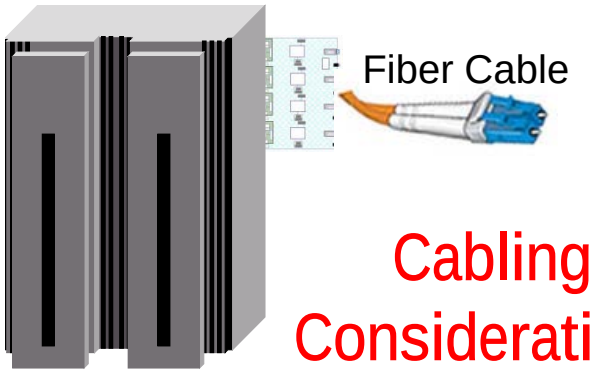
Works this way with and without FICON cascading

Provides a five-9's of availability environment

Not much consideration needs to be given to how busy each of the fabrics are as the remaining fabrics can easily pickup and handle the full workload of a failed fabric

Very Large Mainframe customers can use LOTs of Directors!

Multi-mode cable distance limitations



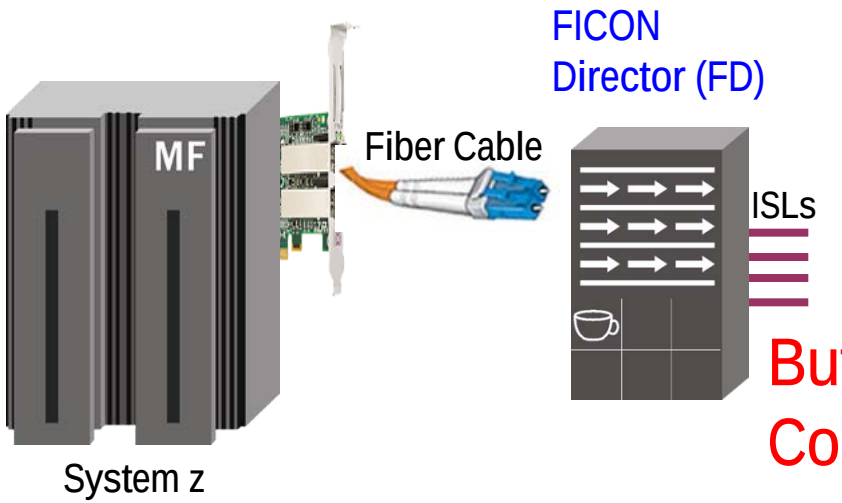
Cabling Considerations

- 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	2
2G	8b10b	2.125	492/150	984/300	1640/500	2
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



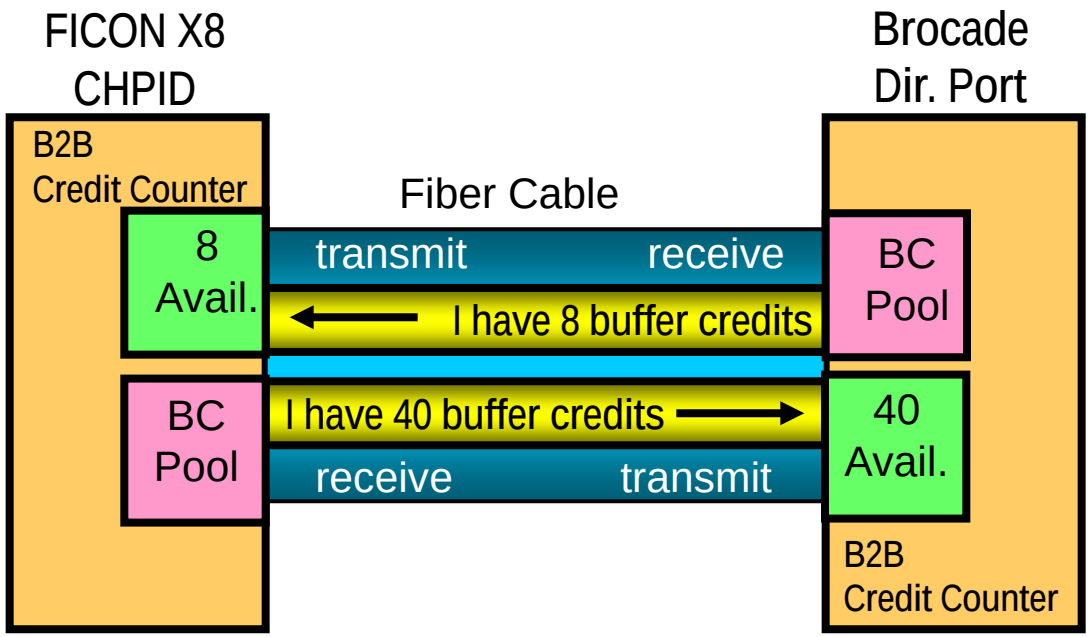
- **Express2 – 107 BCs**
- **Express4 – 200 BCs**
- **Express8 – 40 BCs**
  - **MIDAW can reduce the need for BCs**
  - **zHPF could reduce need for BCs**
  - **Drive Distance BCs via FICON Directors**

Buffer Credit Considerations

- 8G CHPIDs have enough BCs for, at best, 10KM distances
- Let us look at Buffer Credits and see how they are allocated and also why link speed, link distance and average frame size are all important to understanding your need for buffer credits
- BTW….
  - RMF only reports on those metrics with CUP

How Buffer to Buffer Credits Work

- A Fiber channel link is a PAIR of paths
- A path from "this" transmitter to the "other" receiver and a path from the "other" transmitter to "this“ receiver
- The "buffer" resides on each receiver, and that receiver tells the linked transmitter how many BB_Credits are available
- Sending a frame through the transmitter decrements the B2B Credit Counter
- Receiving an R-Rdy (or VC-Rdy) through the receiver increments the B2B Credit Counter



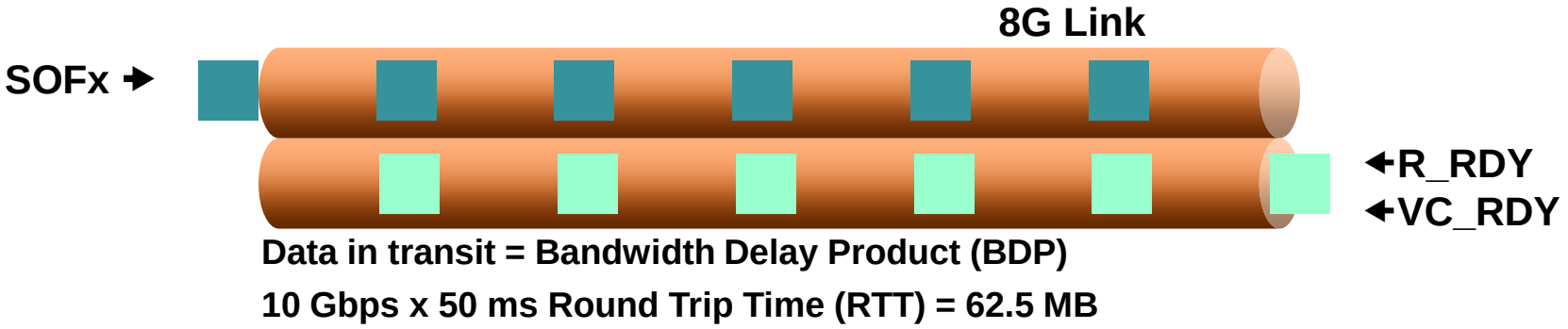
Express2 = fixed 107 BC
Express4 = fixed 200 BC
Express8 = fixed 40 BC

M6140 = 2G fixed 60 BC/port
M6140 = 4G fixed 125 BC/port
Mi10K = 2G 2 pools of 1,366 BC
Mi10K = 4G 64 or 200 BC/port
B48K = 4G 1 or 2 pools of 1,024 BC
B48K = 8G 1 or 2 pools of 2,048 BC
DCX = 8G 1 or 2 pools of 2,048 BC
DCX-4S= 8G 1 or 2 pools of 2,048 BC

- Buffer Credits are never negotiated!
- Each receiver on the fiber cable can state a different value!
- Once established, it is the transmit (write) connection that will run out of buffer credits

BB Credit Droop

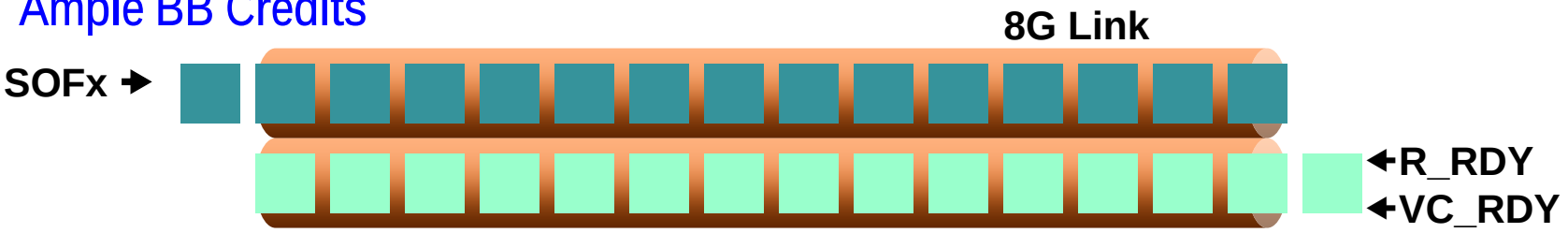
Not enough BB Credits …or…
An 8G link that has auto-negotiated to a slower link



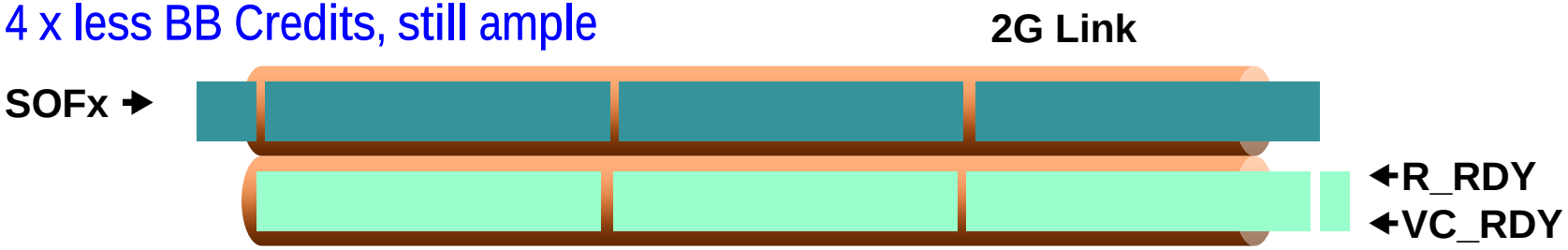
Data in transit = Bandwidth Delay Product (BDP)

10 Gbps x 50 ms Round Trip Time (RTT) = 62.5 MB

Ample BB Credits



4 x less BB Credits, still ample



Buffer Credits Required
By Size of Frame and Link Speed

As distance across a link grows, so does the need for buffer credits!

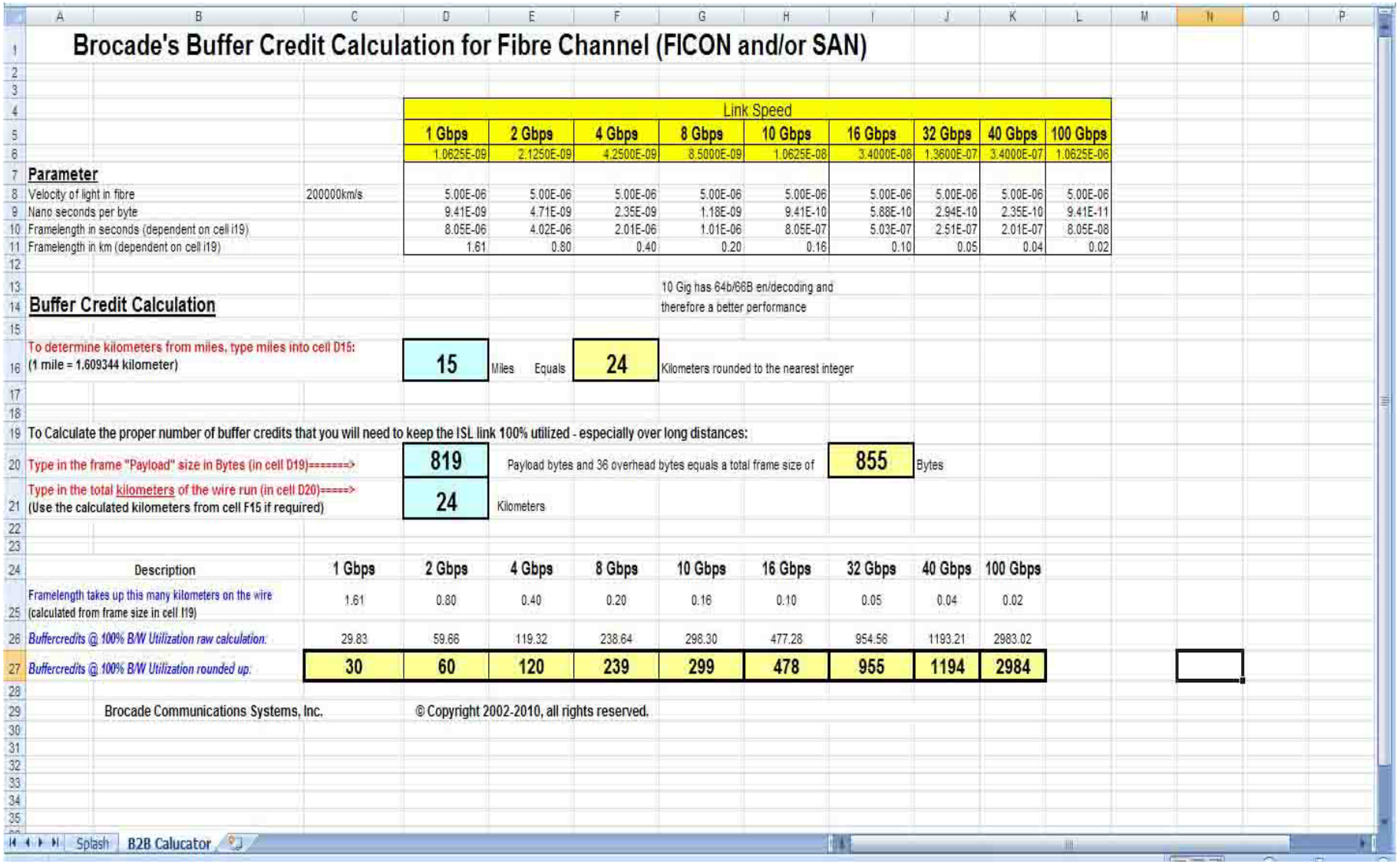
SHARE
Technology · Connections · Results

A distance of 20km with the link 100% utilized				2Gbps	4Gbps	8Gbps	10Gbps
SOF, Header, CRC, EOF	Payload	Total Frame Bytes	Smaller than full frame by x%	Buffer Credits Required 8b10b	Buffer Credits Required 8b10b	Buffer Credits Required	Buffer Credits Required 64b66b
36	2112	2148	0.000%	20	40	80	117
36	2002	2038	5.138%	21	42	84	124
36	1902	1938	9.809%	22	44	88	130
36	1802	1838	14.481%	24	47	93	137
36	1702	1738	19.152%	25	49	98	145
36	1602	1638	23.823%	26	52	104	154
36	1502	1538	28.494%	28	56	111	164
36	1402	1438	33.165%	30	60	119	175
36	1302	1338	37.836%	32	64	128	188
36	1202	1238	42.507%	35	69	138	203
36	1102	1138	47.179%	38	75	150	221
36	1002	1038	51.850%	41	82	164	243
36	902	938	56.521%	46	91	182	268
36	819	855	60.398%	50	100	199	294
36	700	736	65.957%	58	116	232	342
36	600	636	70.628%	67	134	268	396
36	500	536	75.299%	80	159	318	469
36	400	436	79.970%	98	195	390	577
36	300	336	84.641%	127	254	507	748
36	200	236	89.312%	181	361	721	1065
36	100	136	93.984%	313	626	1251	1848
36	75	111	95.151%	383	766	1532	2264
36	50	86	96.319%	495	989	1978	2922

http://www.brocade.com/san/pdf/whitepapers/Buffer_to_Buffer_Credits_and_Effect_on_FICON_Performance_WP_00.pdf

Brocade has a BC Calculator that you can use!

SHARE
Technology · Connections · Results



Brocade's Buffer Credit Calculation for Fibre Channel (FICON and/or SAN)

Parameter		Link Speed 1 Gbps	2 Gbps	4 Gbps	8 Gbps	10 Gbps	16 Gbps	32 Gbps	40 Gbps	100 Gbps
		1.0625E-09	2.1250E-09	4.2500E-09	8.5000E-09	1.0625E-08	3.4000E-08	1.3600E-07	3.4000E-07	1.0625E-06
Velocity of light in fibre	200000km/s	5.00E-06	5.00E-06	5.00E-06	5.00E-06	5.00E-06	5.00E-06	5.00E-06	5.00E-06	5.00E-06
Nano seconds per byte		9.41E-09	4.71E-09	2.35E-09	1.18E-09	9.41E-10	5.88E-10	2.94E-10	2.35E-10	9.41E-11
Framelength in seconds (dependent on cell i19)		8.05E-06	4.02E-06	2.01E-06	1.01E-06	8.05E-07	5.03E-07	2.51E-07	2.01E-07	8.05E-08
Framelength in km (dependent on cell i19)		1.61	0.80	0.40	0.20	0.16	0.10	0.05	0.04	0.02

10 Gig has 64b/66B en/decoding and therefore a better performance

Buffer Credit Calculation

To determine kilometers from miles, type miles into cell D15:
(1 mile = 1.609344 kilometer)

15 Miles Equals 24 Kilometers rounded to the nearest integer

To Calculate the proper number of buffer credits that you will need to keep the ISL link 100% utilized - especially over long distances:

Type in the frame "Payload" size in Bytes (in cell D19)=======> 819 Payload bytes and 36 overhead bytes equals a total frame size of 855 Bytes

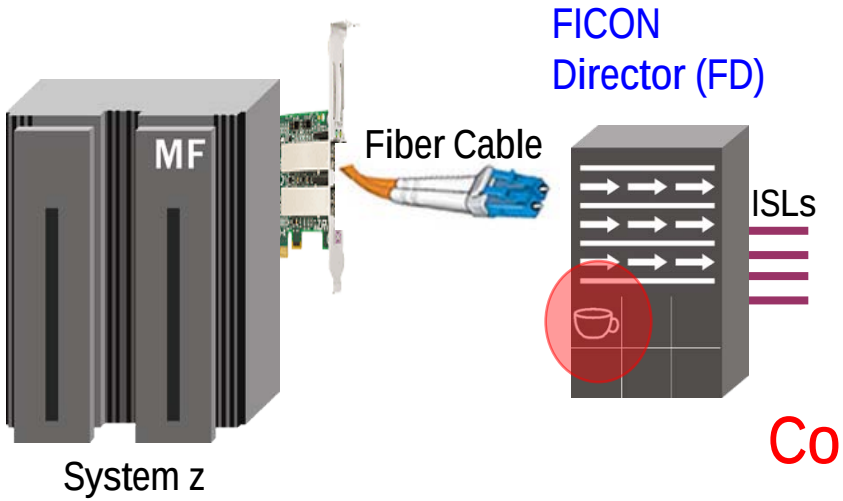
Type in the total kilometers of the wire run (in cell D20)=====> 24 Kilometers
(Use the calculated kilometers from cell F15 if required)

Description	1 Gbps	2 Gbps	4 Gbps	8 Gbps	10 Gbps	16 Gbps	32 Gbps	40 Gbps	100 Gbps
Framelength takes up this many kilometers on the wire (calculated from frame size in cell I19)	1.61	0.80	0.40	0.20	0.16	0.10	0.05	0.04	0.02
Buffercredits @ 100% B/W Utilization raw calculation:	29.83	59.66	119.32	238.64	298.30	477.28	954.56	1193.21	2983.02
Buffercredits @ 100% B/W Utilization rounded up:	**30**	**60**	**120**	**239**	**299**	**478**	**955**	**1194**	**2984**

Brocade Communications Systems, Inc. © Copyright 2002-2010, all rights reserved.

Splash | B2B Calucator

End-to-End FICON/FCP Connectivity



Control Unit Port allows z/OS to use both Systems Automation management and RMF reporting upon a FICON switching device

CUP Considerations

- Buffer credits were never used for ESCON but are for FICON
- It is a resource and buffer credits can get depleted on a link
  - 8G CHPIDs are provided with only 40 buffer credits each
- If BCs can get depleted, and cause potential performance issues, then RMF needs to be able to report on them
- IBM chose to report on FICON buffer credit usage only when switched-FICON is used and only when CUP is enabled

Certain features depend upon CUP

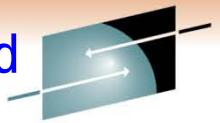
When you install the FICON Management Server (FMS) license on a FICON switching device, and then “Enable FICON Manager Server Mode”, you provide yourself with a lot of valuable tools:

- RMF can have in-band access to the FICON switching
- Systems Automation for z/OS, with the I/O OPs module implemented, can have in-band access to the FICON switching devices
- FICON Dynamic Channel Management (DCM) can dynamically add and remove channel resources at Workload Manager discretion

Regardless of whether you have single Director fabrics or cascaded fabrics the best practice is to always implement CUP for every device in a FICON fabric

FICON Director Activity Rpt

FMS Key on, CUP enabled
RMF 74-7 records captured



zHPF Enabled

F I C O N D I R E C T O R A C T I V I T Y

PAGE 1

z/OS V1R8 SYSTEM ID PRD1 START 04/12/2009-04.30.00 INTERVAL 000.15.00
RPT VERSION V1R8 RMF END 04/12/2009-04.45.00 CYCLE 1.000 SECONDS

IODF = A2 CR-DATE: 03/27/2009 CR-TIME: 16.43.51 ACT: ACTIVATE
SWITCH DEVICE: 032B SWITCH ID: 2B TYPE: 006140 MODEL: 001 MAN: MCD PLANT: 01 SERIAL: 00000131

PORT ADDR	-CONNECTION- UNIT	ID	AVG FRAME PACING	AVG FRAME SIZE READ	WRITE	PORT BANDWIDTH (MB/SEC) -- READ --	-- WRITE --	ERROR COUNT
05	CHP	05	0	849	1436	8.63	17.34	0
07	CHP-H	6B	0	1681	1395	0.87	0.32	0
09	CHP	15	7	833	1429	11.96	20.49	0
0C	CHP-H	64	0	939	1099	0.39	0.50	0
0D	CHP	6B	1	1328	1823	3.56	12.73	0
0F	CHP-H	66	0	1496	1675	1.85	2.61	0
10	CHP	64	0	644	1380	0.03	0.13	0
13	CHP-H	19	0	907	885	0.58	0.45	0
16	CHP	12	0	1241	1738	0.97	1.72	0
17	CHP	0B	0	685	1688	0.10	0.82	0
1A	CHP	15	0	1144	1664	0.65	1.18	0
1B	CHP	0D	0	510	1759	0.12	1.72	0
1E	CHP-H	05	0	918	894	0.59	0.45	0
1F	CHP	21	0	1243	1736	0.97	1.70	0
20	CU	E900	0	1429	849	17.66	8.85	0
	CU	E800						
	CU	E700						
22	CHP	10	0	923	1753	0.55	2.78	0
23	CHP	54	0	1805	69	0.80	0.00	0
24	CHP	64	0	89	1345	0.00	0.00	0
27	CHP	6B	0	1619	82	0.01	0.00	0
28	CHP	95	27	918	1589	10.32	30.56	0
2B	CHP	70	0	69	2022	0.00	0.71	0

BC Shortage:

BCs are most important to the transmitter!

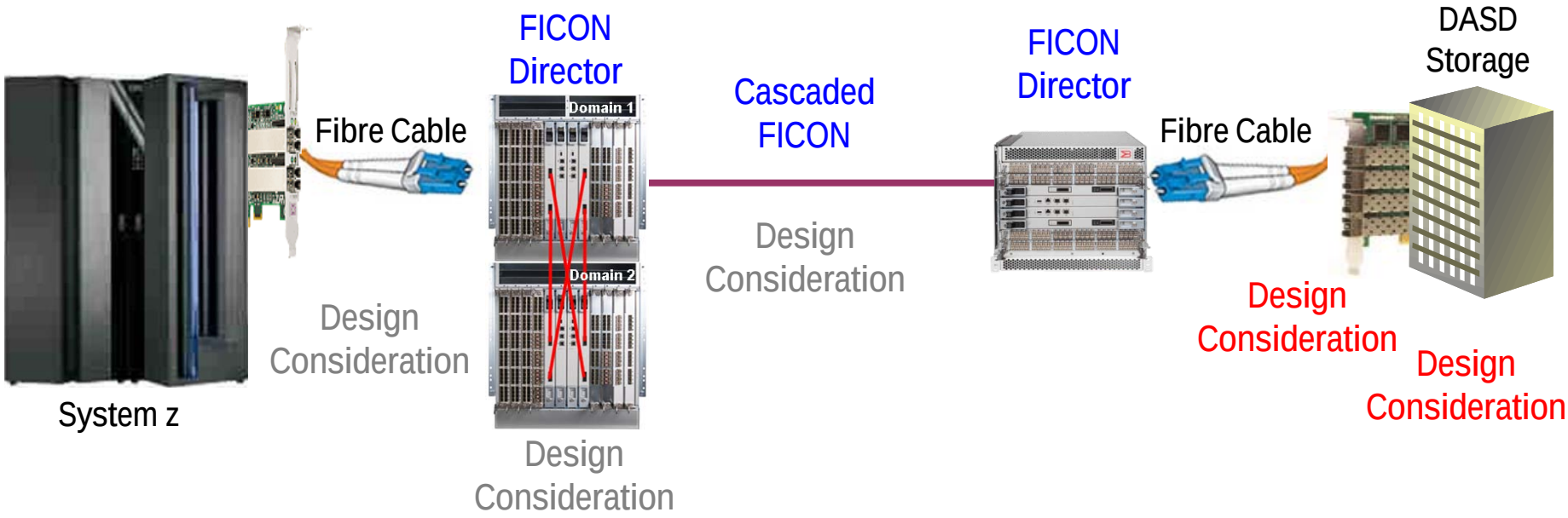
So from these reports look at the WRITE activity that is occurring and also at the WRITE MB/sec

Overall Averages: ~1116 ~1508

Note: Transport Mode results in larger frames

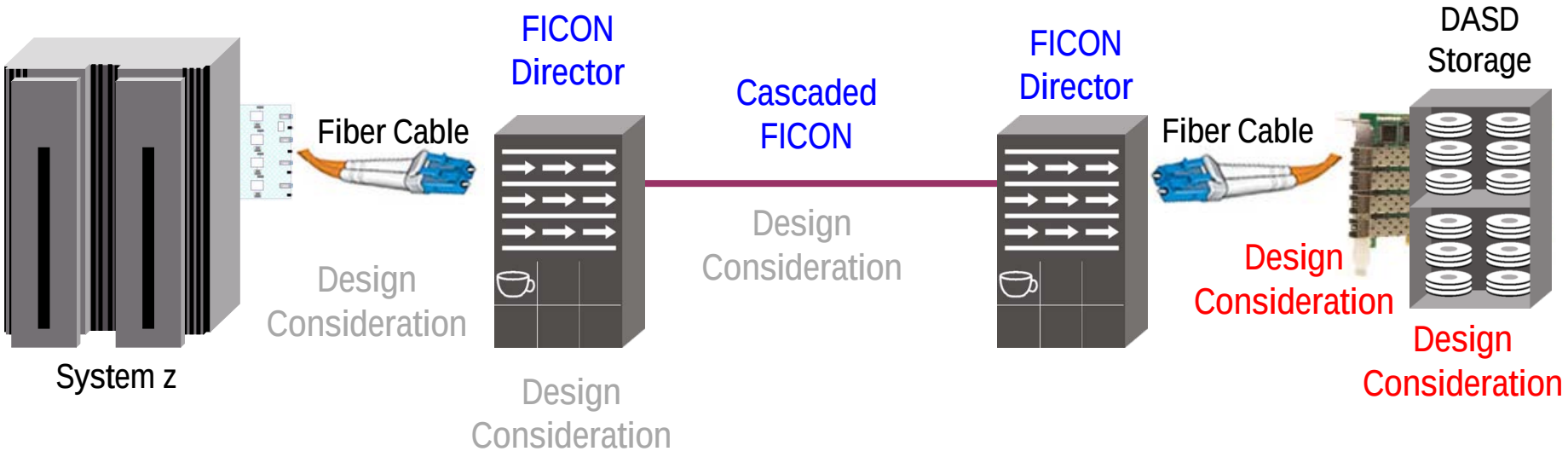
Command Mode will probably find that an average FICON frame size is 350-1000 bytes!

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?

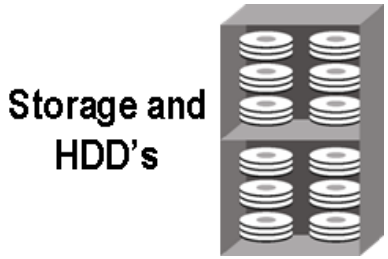
Busy storage arrays can equal reduced performance

- 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



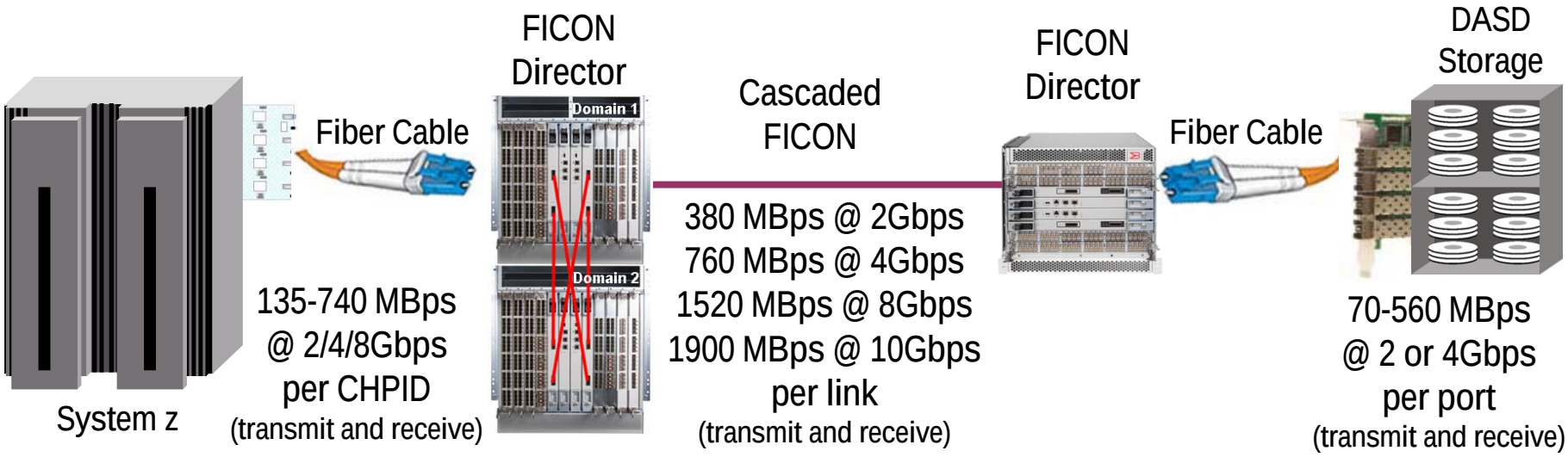
What kinds of internal bottlenecks does a DASD array h

- 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!

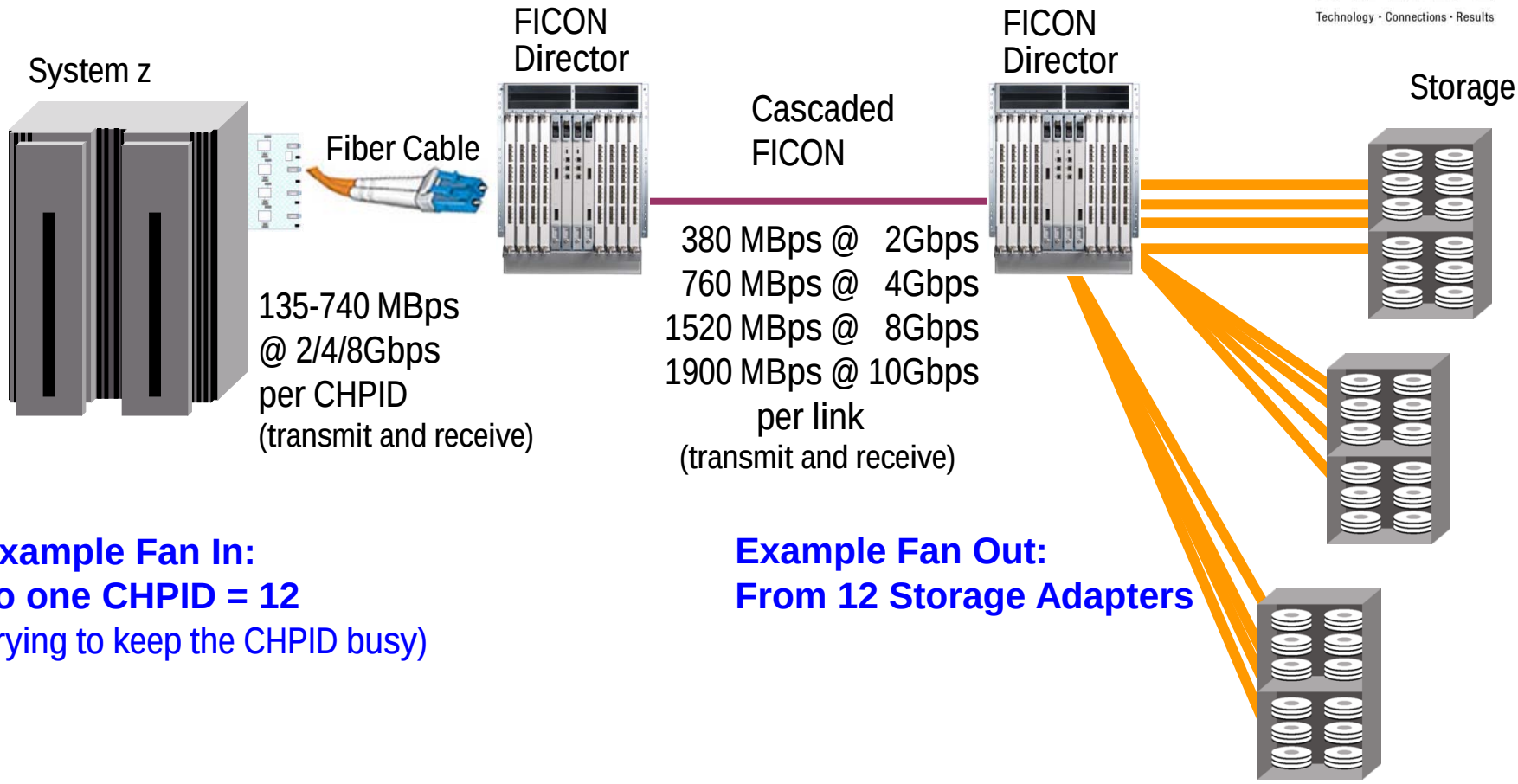
End-to-End FICON/FCP Connectivity



- 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

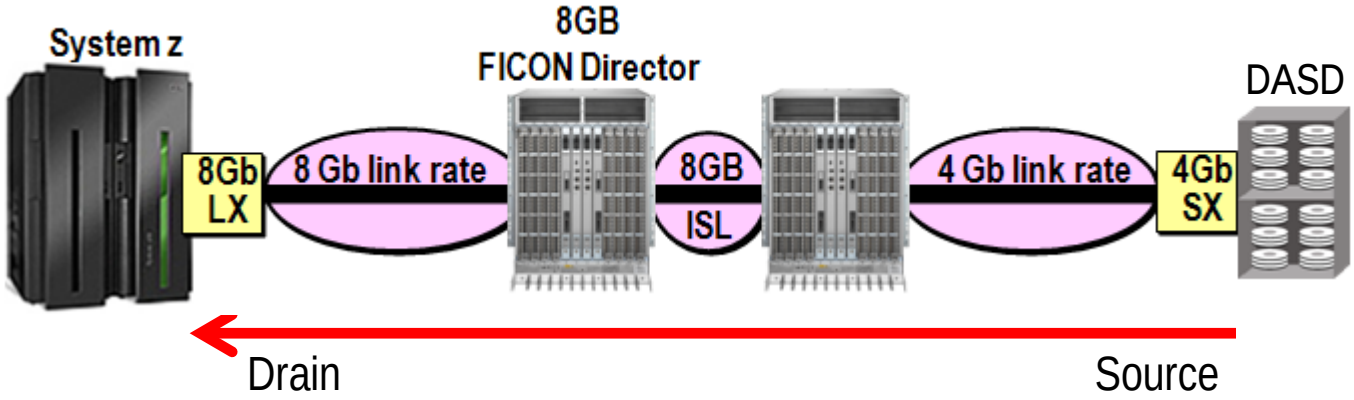


Example Fan In:
To one CHPID = 12
(trying to keep the CHPID busy)

Example Fan Out:
From 12 Storage Adapters

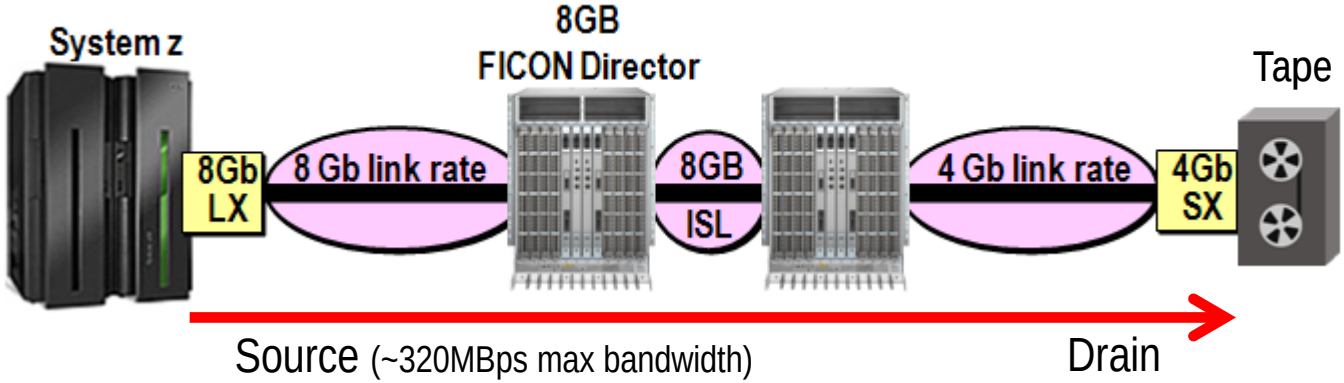
- Total FICON path usually does not support full speed
  - Must deploy Fan In – Fan Out to utilize connections wisely
    - Multiple I/O flows funneled over a single channel path

Maximum End-to-End Link Rates



- 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)
- Can have a strong Fan-Out from CHPID to storage ports

Maximum End-to-End Link Rates



BUT...
If 8G Tape runs
at 320MBps (2X)
it will take 640MBps
to feed it and the
8G CHPID <=510MBps

- Assuming no ISL or BC problems, and assuming the normal and typical use of Tape, 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?

- For 4G tape this is OK – Tape is about 90% write and 10% read on average
- The maximum bandwidth a tape can accept and compress is about 240MBps for Oracle T1000B and about 320MBps for IBM TS1130 (at 2:1 compression)
- An 8G CHPID in Command Mode can do about 510MBps
- A 4G Tape channel can carry about 380MBps (400 * .95) = 380
- So Fan-Out of a single CHPID attached to a 4G tape interface port:
  - Can run a single IBM tape drive (510 / 320 = 1.594)
  - Can run two Oracle (STK) tape drives (510 / 240 = 2.125)

BROCADE

Brocade's Mainframe Certification

Industry Recognized Professional Certification
Next class is in Atlanta – March 15-16

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

System z FICON Fabric Performance Considerations

THE END

David Lytle, BCAF
Global Solutions Architect
System z Technologies and Solutions
Brocade Communications, Inc.

Wednesday March 2, 2011
Session Number 8486