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

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

This technical session goes into a fairly deep discussion on some of the design considerations of a FICON infrastructure.

- Among the topics this session will focus on is:
  - Congestion and Backpressure in FC fabrics
  - How Buffer Credits get initialized
  - How FICON utilizes buffer credits
  - Oversubscription and Slow Draining Devices
  - Determining Buffer Credits Required
  - FICON RMF Reporting

NOTE: I did update this session today so you might want to download the most recent copy of the presentation from the SHARE website.

This Section

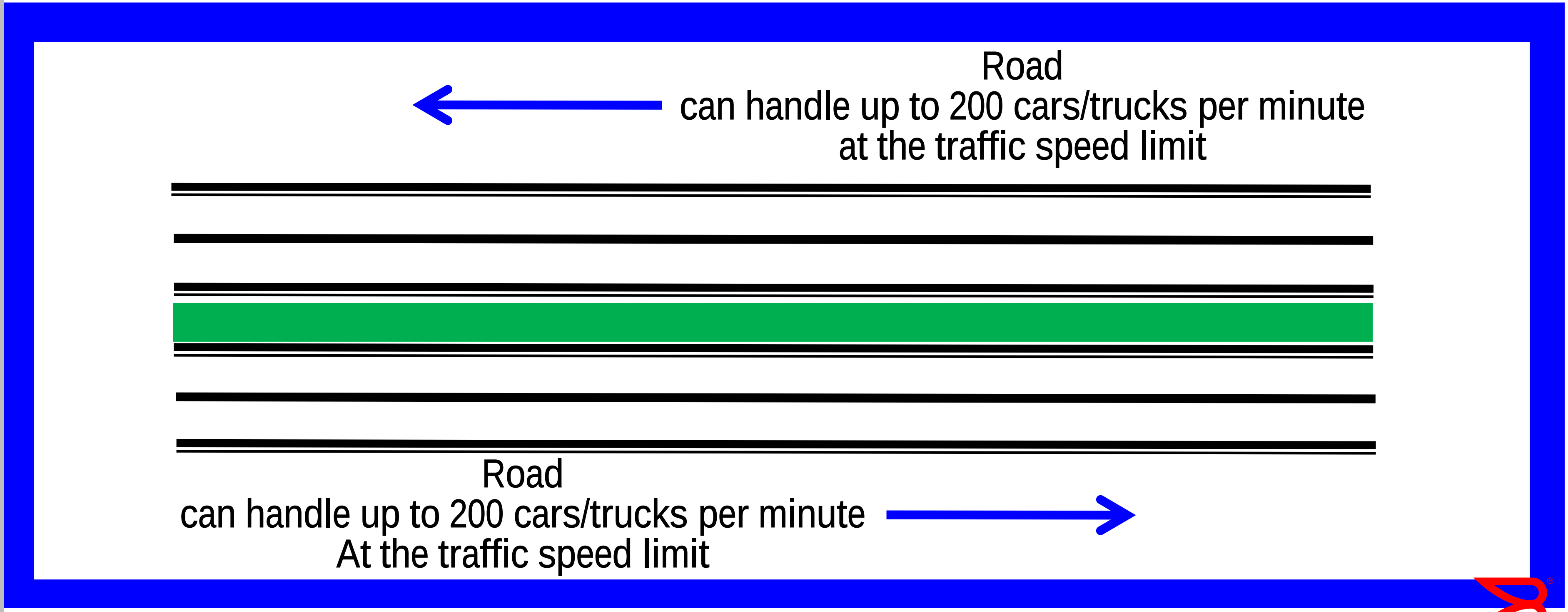
- Congestion and Backpressure Overview

Congestion and Backpressure Overview

These two conditions are not the same thing

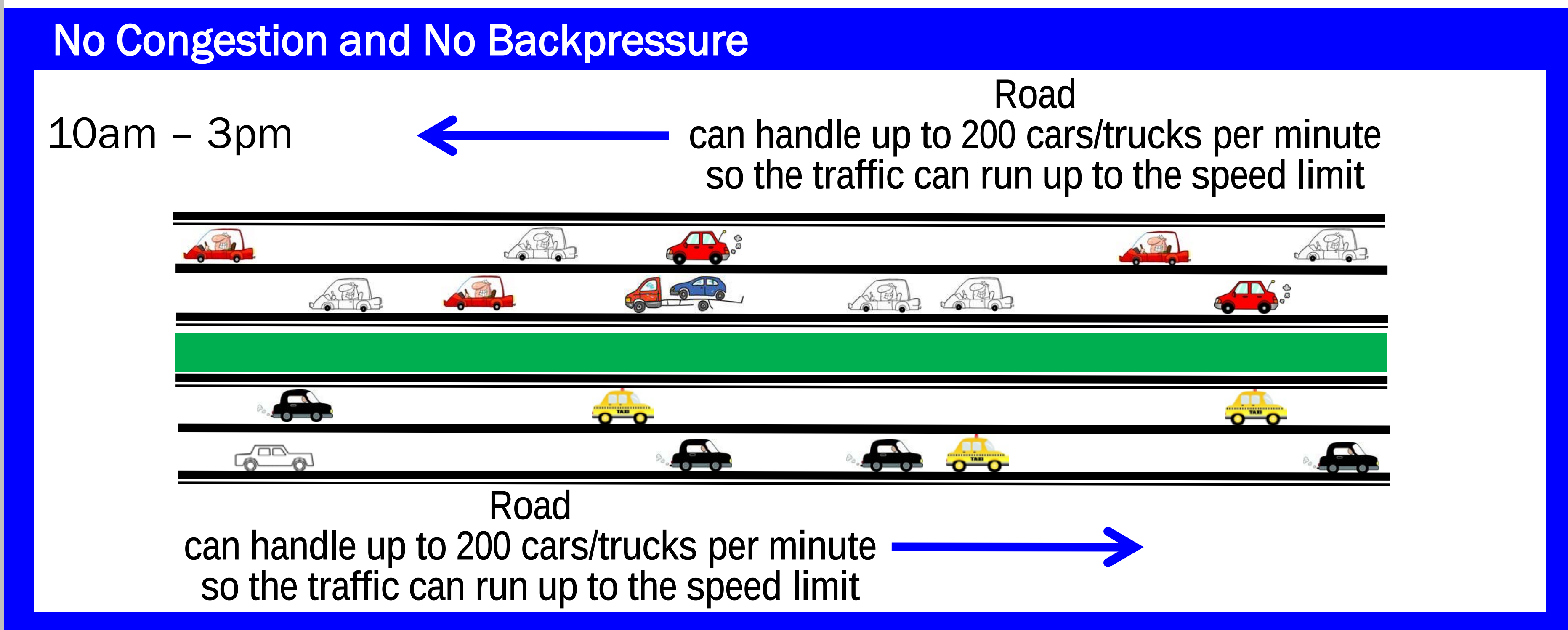
- Congestion occurs at the point of restriction
- Backpressure is the effect felt by the environment leading up to the point of restriction

I will use an Interstate highway example to demonstrate these concepts



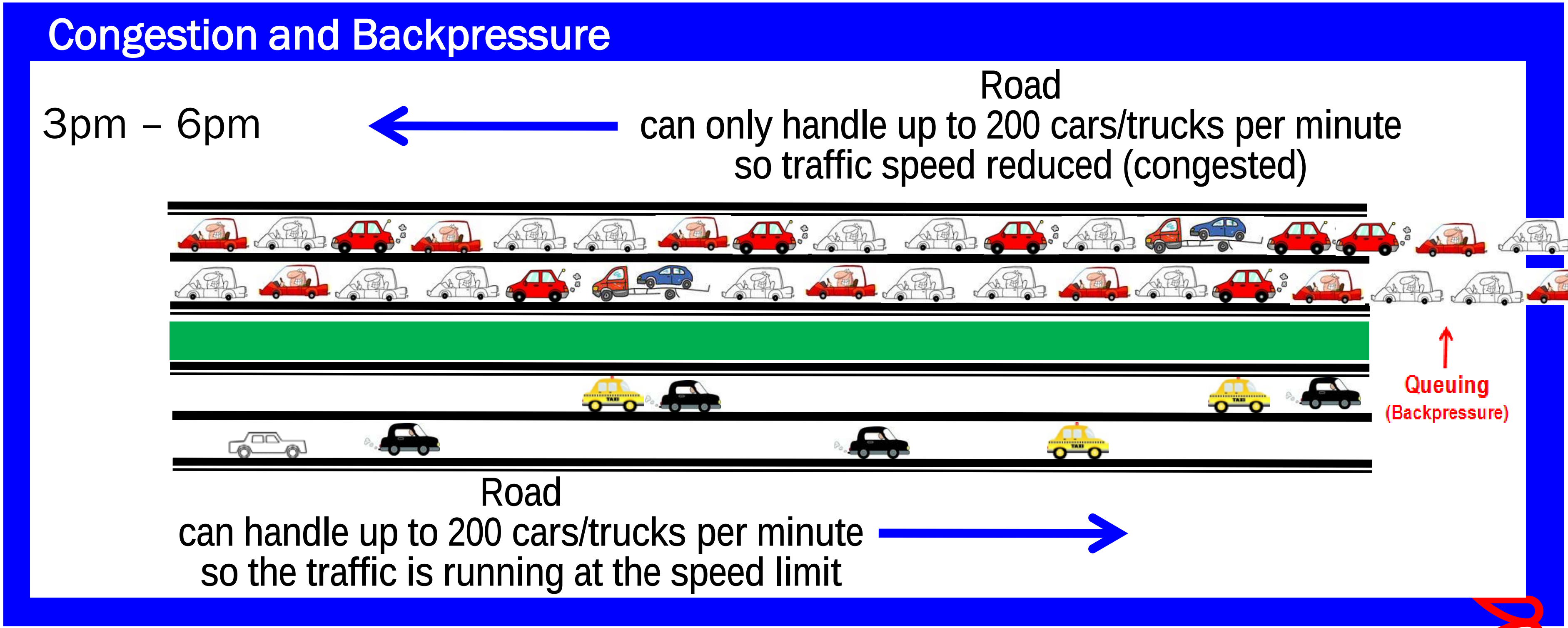
Congestion and Backpressure Overview

- No Congestion and No Backpressure
  - The highway handles up to 200 cars/trucks per minute and less than 200 cars/trucks per min are arriving
- Time spent in queue (behind slower traffic) is minimal
  - Cut-through routing (zipping along from point A to point B) works well



Congestion and Backpressure Overview

- Congestion
  - The highway handles up to 200 cars/trucks per minute and more than 200 cars/trucks per min are arriving
- Latency time and buffer credit space consumed increases
  - Cut-through routing cannot decrease the problem
- Backpressure is experienced by cars slowing down and queuing up



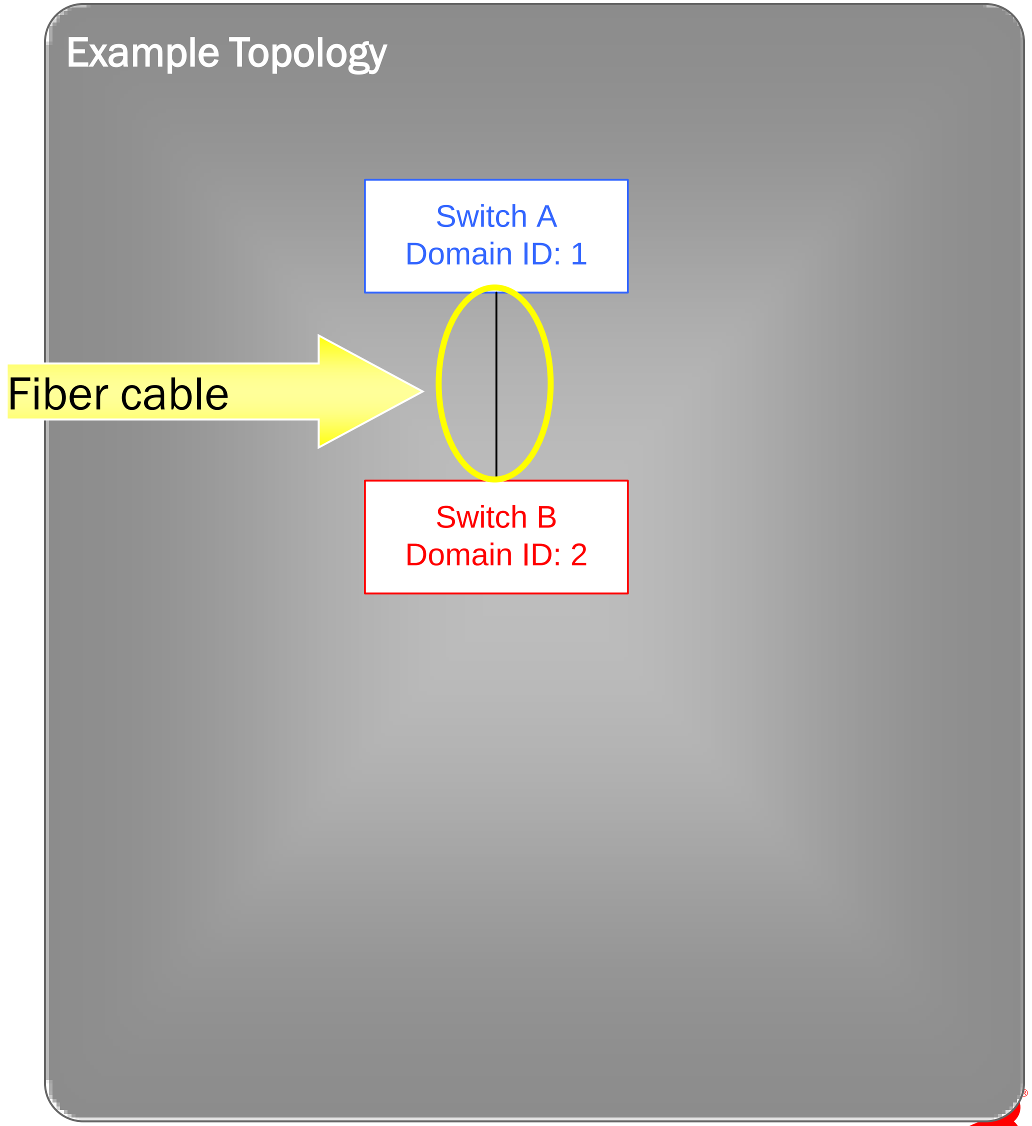
This Section

- Very basic flow for the Build Fabric process and how Buffer Credits get initialized

Build Fabric Process

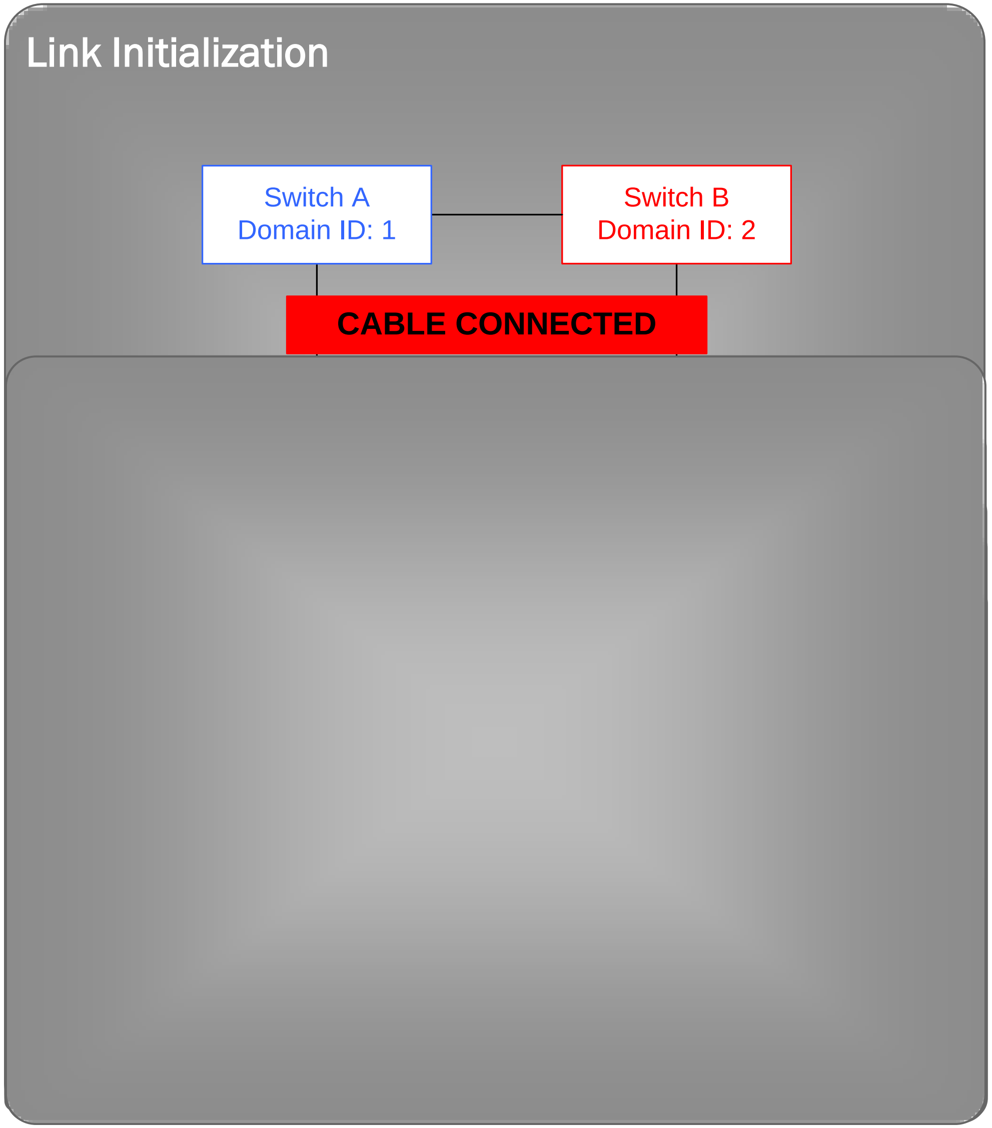
Assume

- A fiber cable will be attached between switch A and B
- This will create an ISL (E_Port) between these two devices



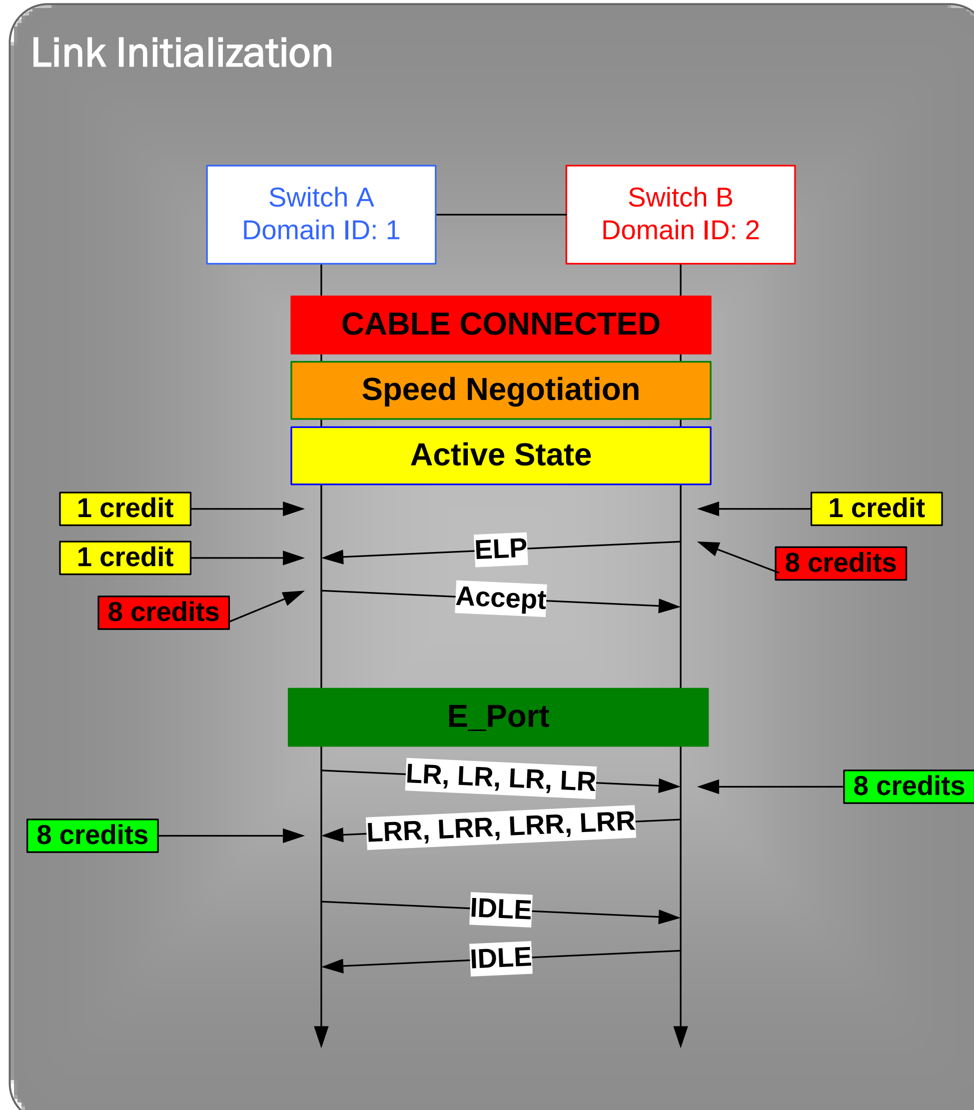
Build Fabric Process

- Cable connected
- Link Speed Auto-Negotiation
- Active state
- One credit is granted by default to allow port login to occur
- Exchange Link Parms (ELP)
  - Contains the "requested" buffer credit information for the sender
  - Assume 8 credits are being granted for this example
- Responder Accepts – does an ELP
  - Contains the "requested" buffer credit information for the responder
  - Assume 8 credits are being granted for this example
- Link becomes an E_Port
- Send Link Resets (LR)
  - Initializes Sender credit values
- Link Reset Response (LRR)
  - Initializes Responder credit values



Build Fabric Process

- Cable connected
- Link Speed Auto-Negotiation
- Active state
- One credit is granted by default to allow port login to occur
- Exchange Link Parms (ELP)
  - Contains the "requested" buffer credit information for the sender
  - Assume 8 credits are being granted for this example
- Responder Accepts – does an ELP
  - Contains the "requested" buffer credit information for the responder
  - Assume 8 credits are being granted for this example
- Link becomes an E_Port
- Send Link Resets (LR)
  - Initializes Sender credit values
- Link Reset Response (LRR)
  - Initializes Responder credit values
- Ready for I/O to start flowing

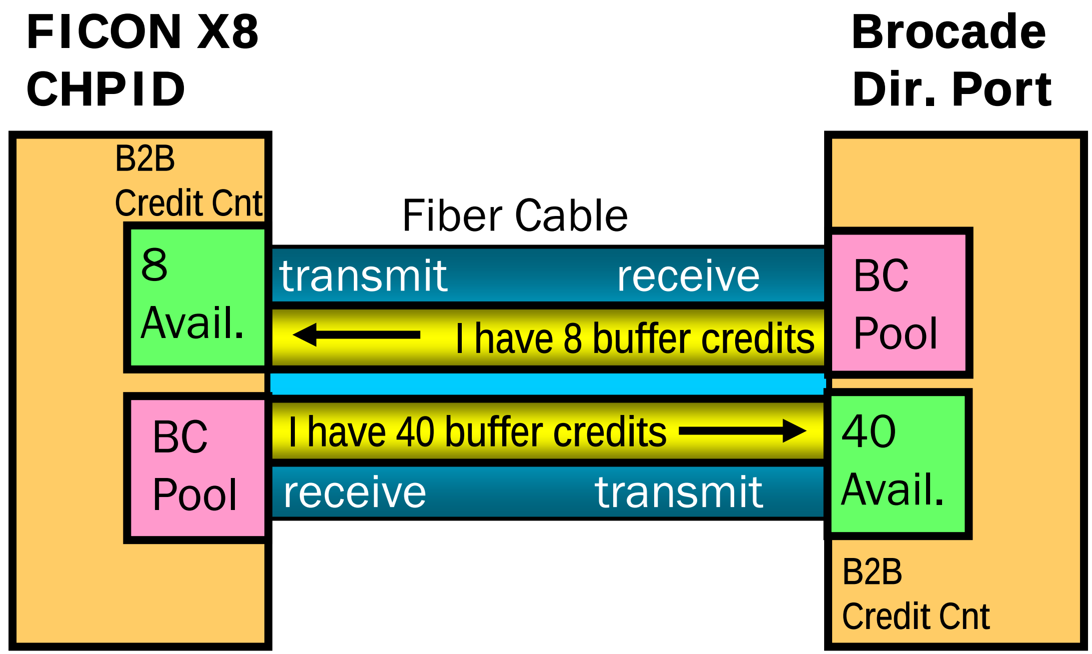


This Section

- How FICON uses Buffer-to-Buffer Credits

How 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
- DCX/DCX-4S have a BC recovery capability



Express = fixed 64 BC
Express2 = fixed 107 BC
Express4 = fixed 200 BC
Express8 = fixed 40 BC

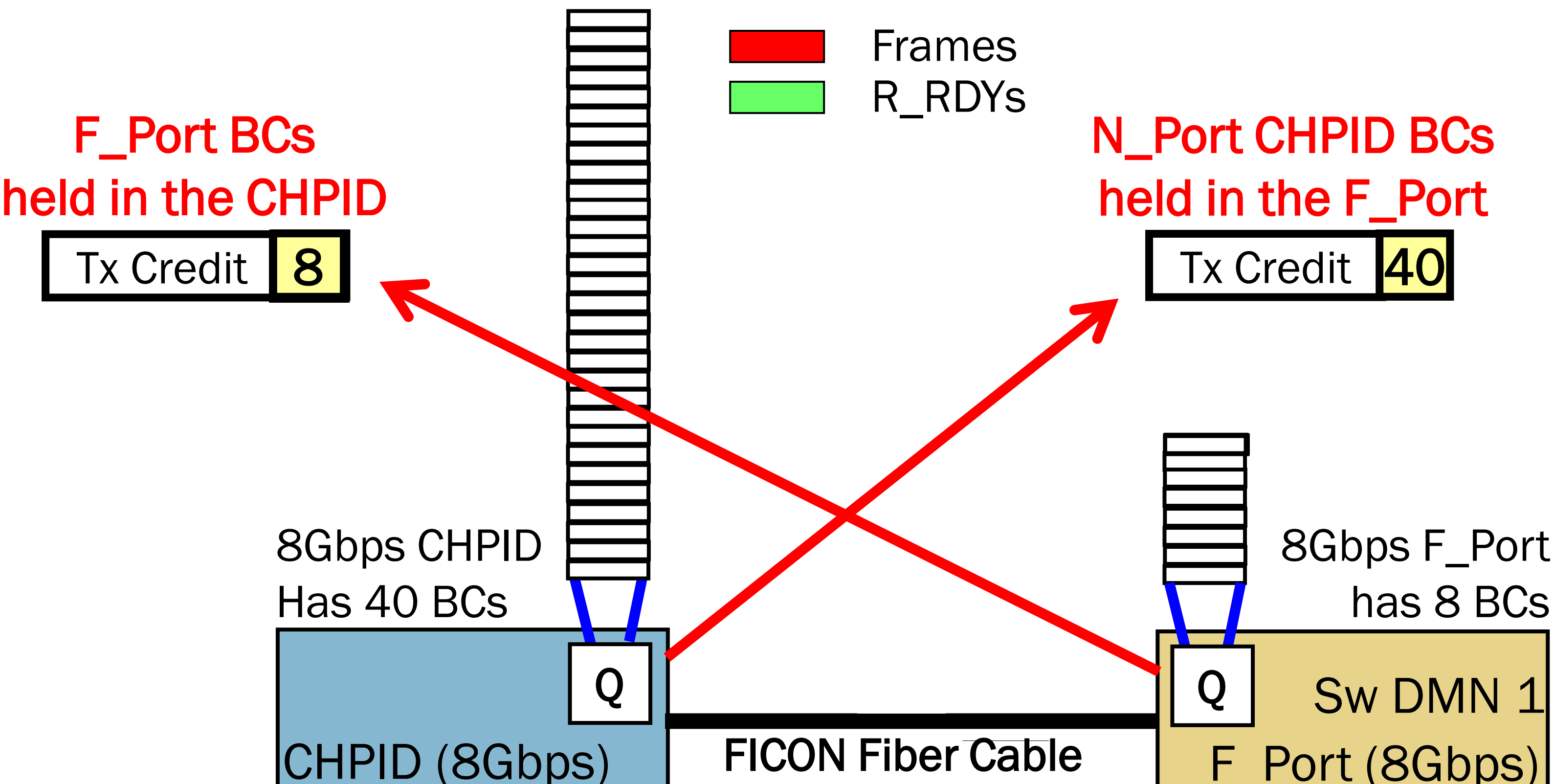
Switch has variable BCs
DASD has fixed BCs
Old Tape had variable BCs

Each receiver on the fiber cable can state a different value!

Once established, it is transmit (write) connections that will typically run out of buffer credits

Buffer-to-Buffer Credits
Buffer-to-Buffer flow control

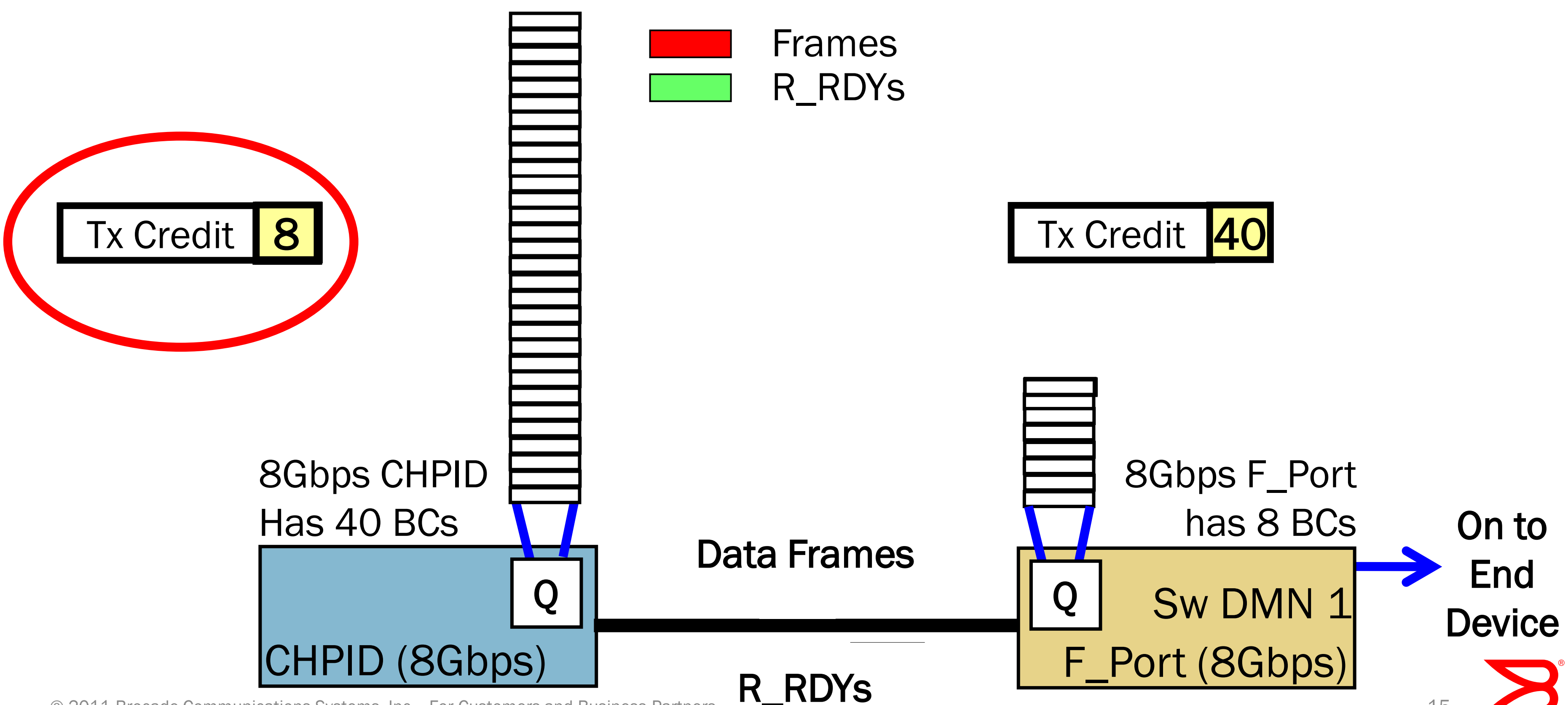
- After initialization, each port knows how many buffers are available in the queue at the other end of the link
  - This value is known as Transmit (Tx) Credit



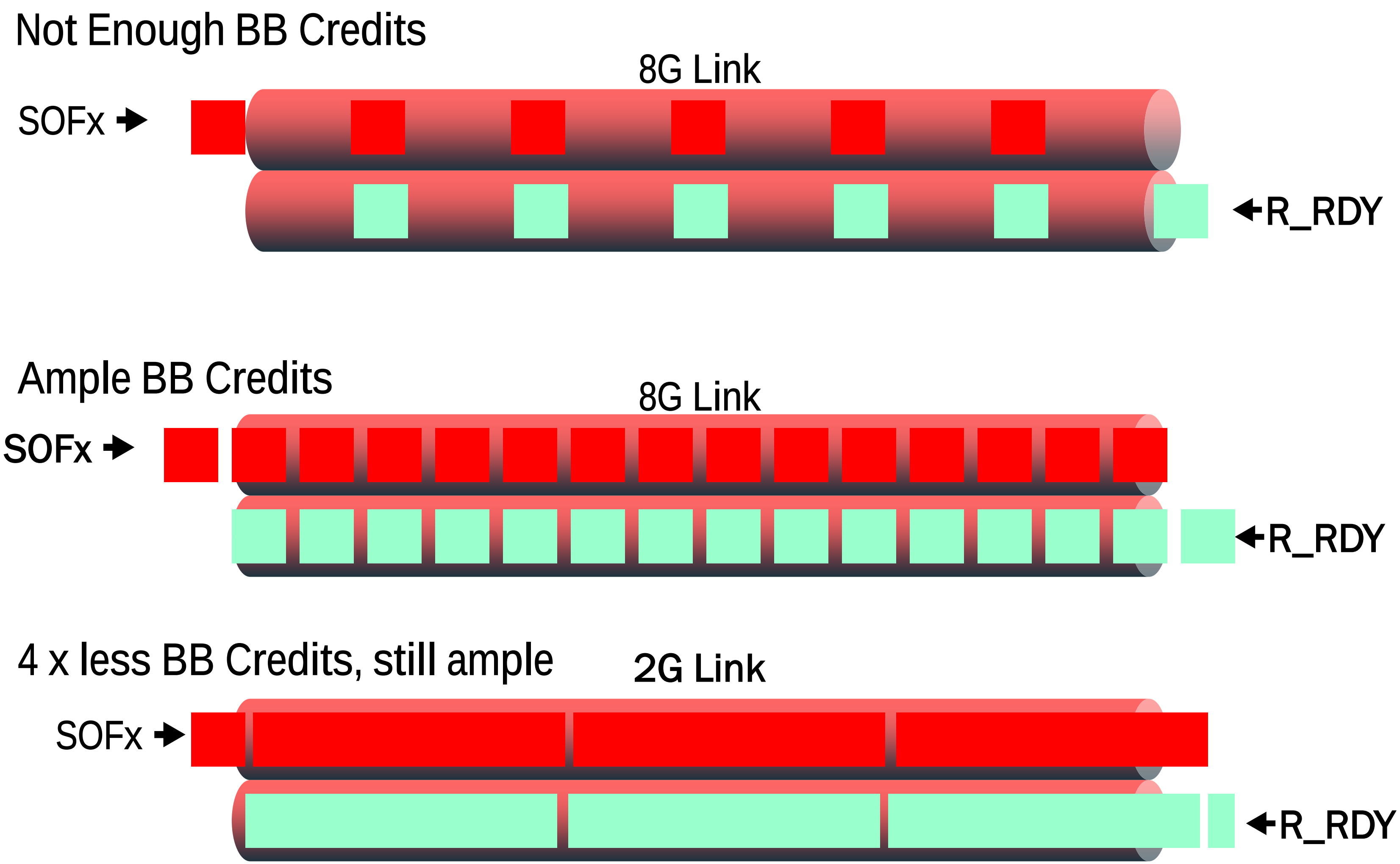
Buffer-to-Buffer Credits

Buffer-to-Buffer flow control Example

- Tx Credit is **decremented** by one for every frame sent from the CHPID
- No frames may be transmitted after Tx Credit reaches zero
- Tx Credit is **incremented** by one for each R_RDY received from F_Port



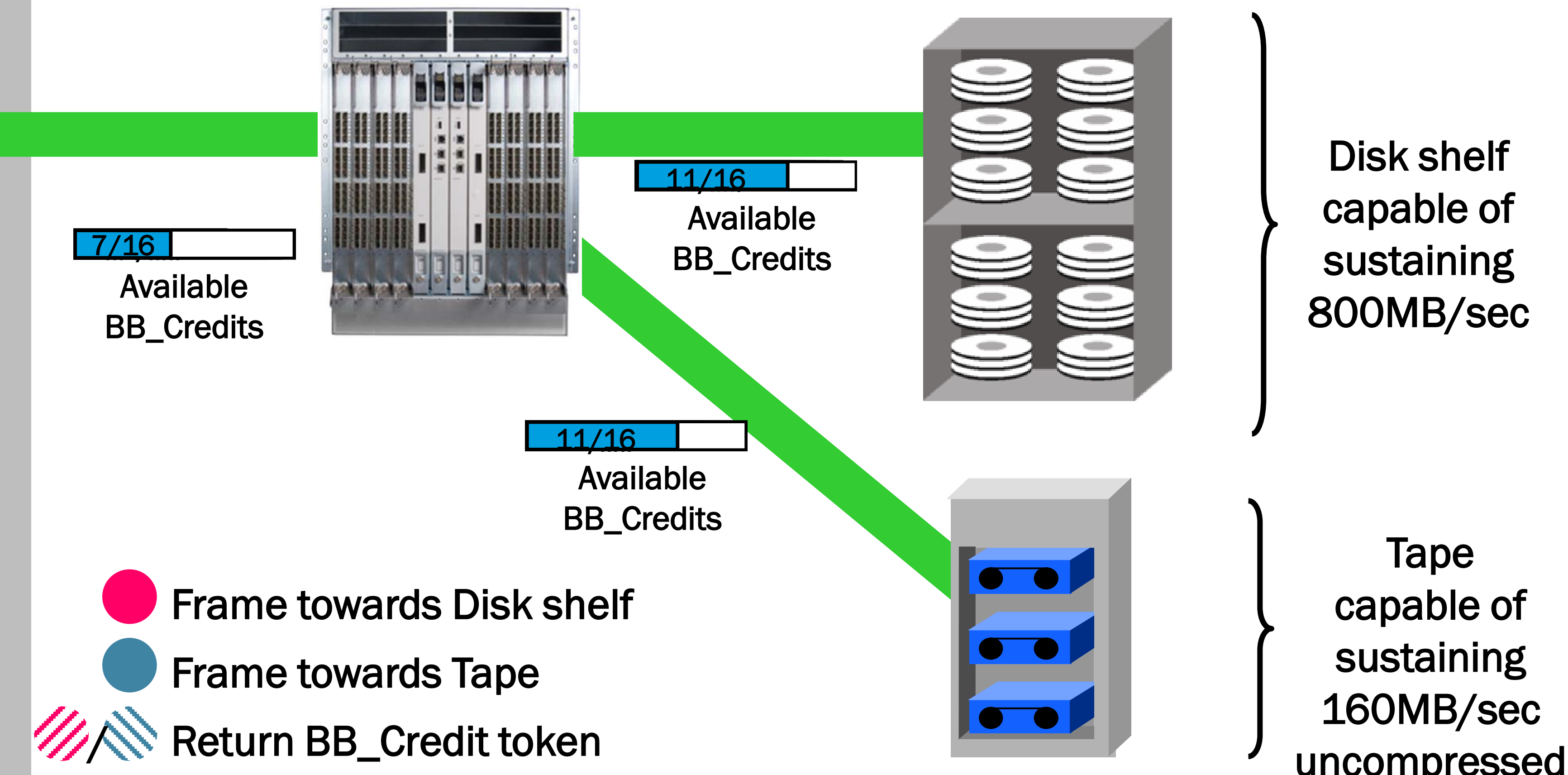
BB Credit Droop



Buffer Credits
Working Ideally

Buffer Credits are a “Flow Control” mechanism to assure that frames are sent correctly

In an ideal FC network all devices can process frames at the same rate and negotiate equal levels of BB_Credits)



This Section

- ISL Oversubscription

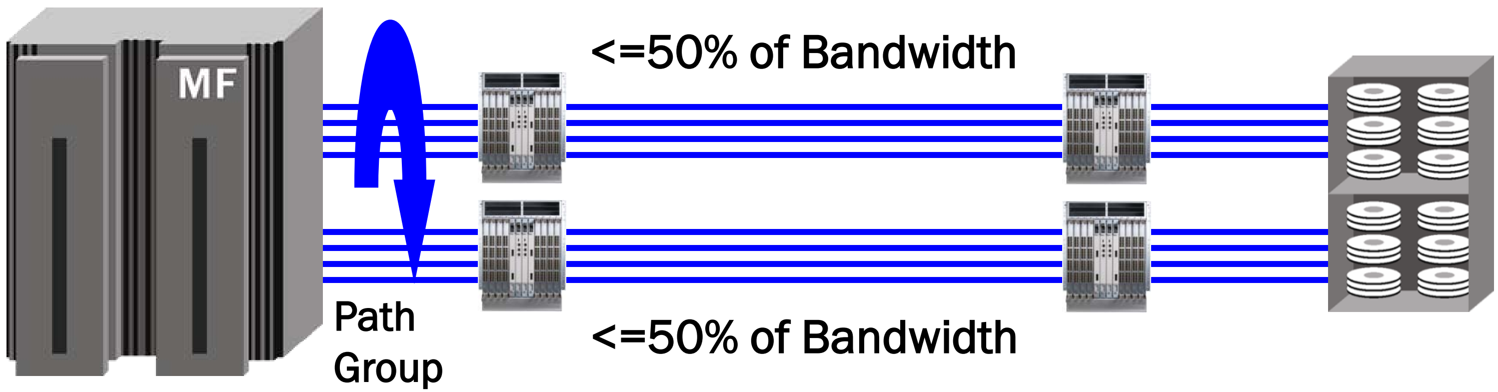
NOTE:

There is also fabric oversubscription and link oversubscription.

In this session I think that ISL Oversubscription will demonstrate how serious a concern that oversubscription really can be to the enterprise.

ISL Oversubscription – Design Architecture

Provides a potential 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

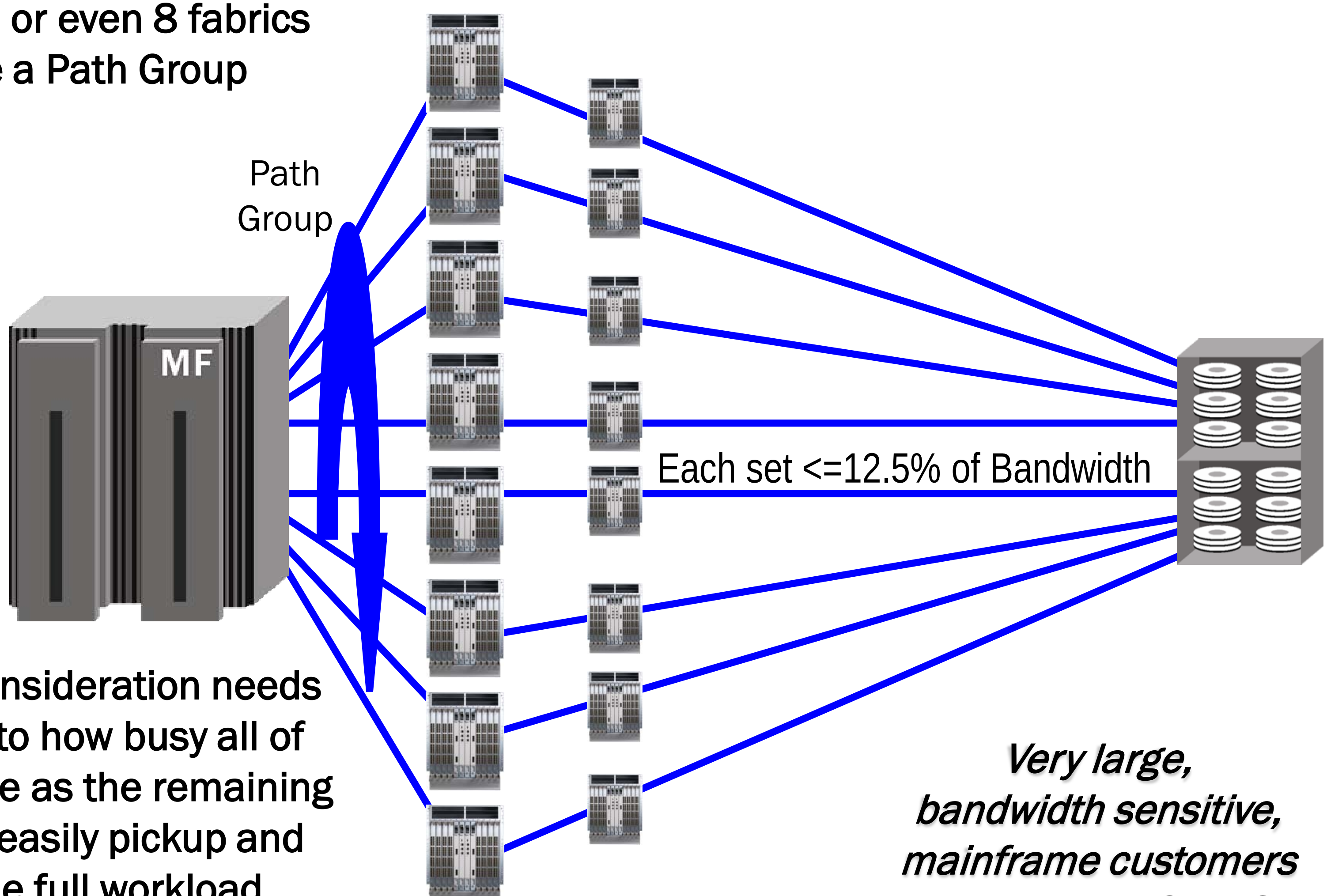
z/OS's IOS automatically load balances the FICON I/O across all of the paths in a Path Group
(up to 8 channels in a PG)

ISL Oversubscription – Design Architecture

Could have 4 or even 8 fabrics to service a Path Group

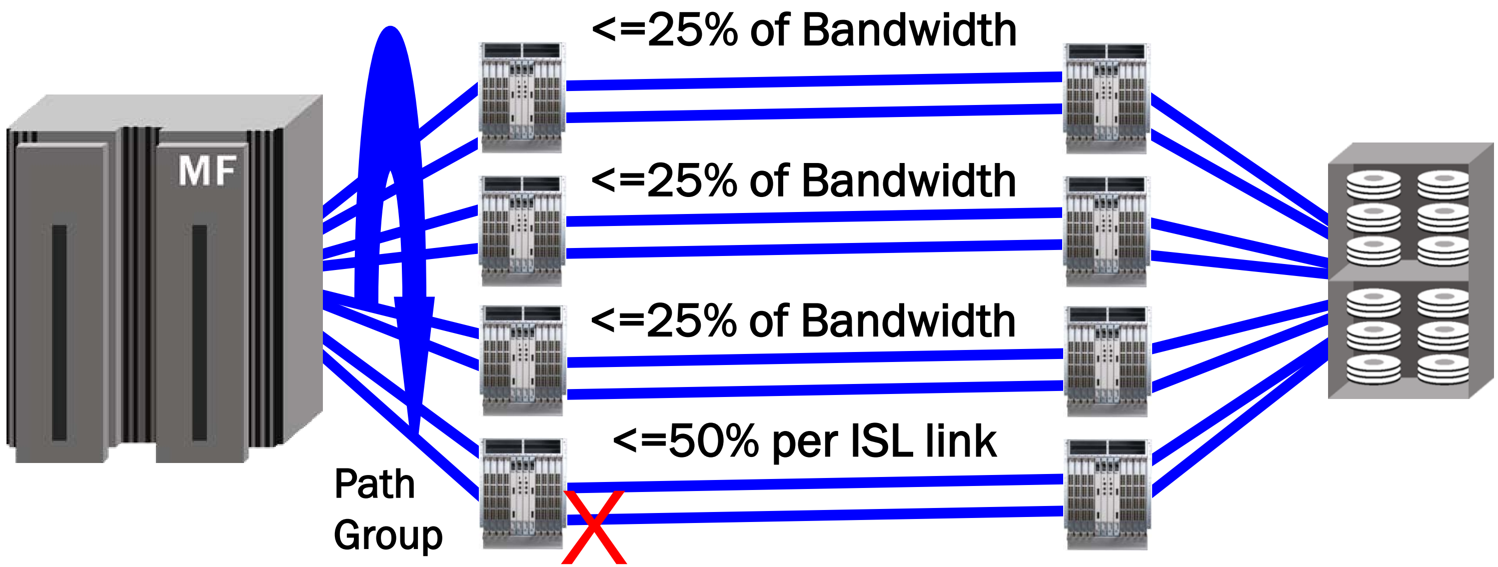
Provides a five-9's of availability environment

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



Very large, bandwidth sensitive, mainframe customers might use LOTs of Directors!

ISL Oversubscription – Design Architecture



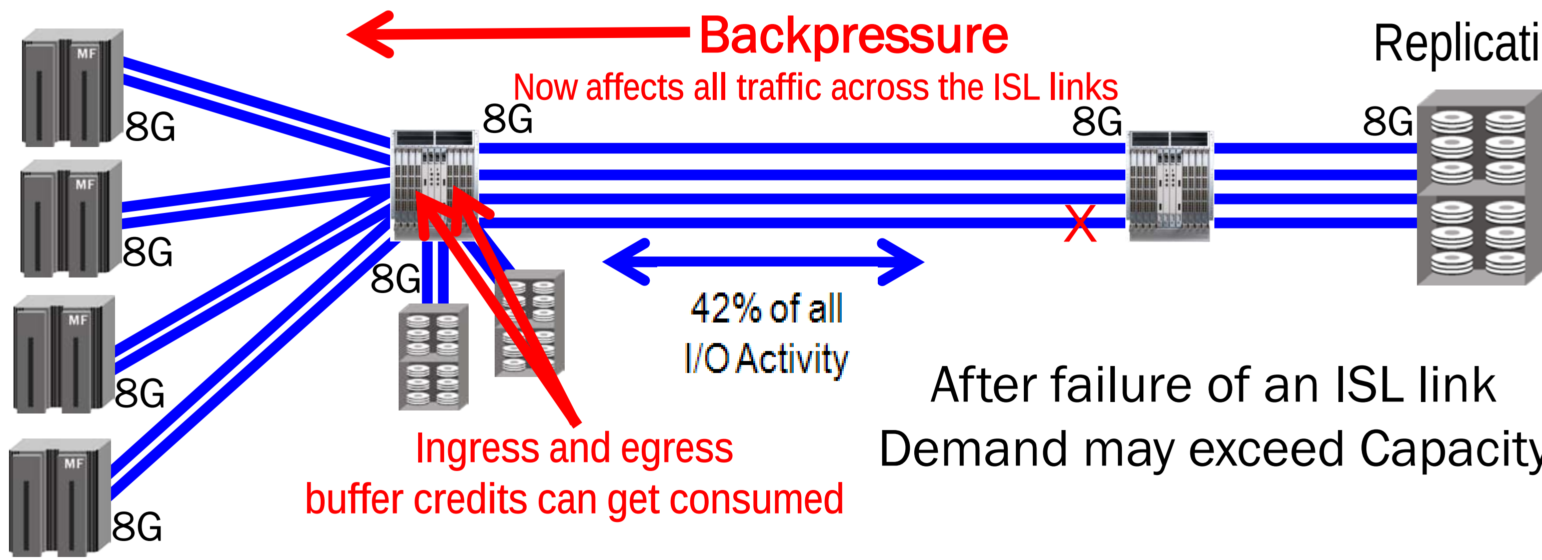
- Risk of Loss of Bandwidth is the motivator for deploying FICON fabrics like this
- In this case, 2 paths from an 8 path Path Group are deployed across four FICON fabrics to limit bandwidth loss to no more than 25% if a FICON fabric were to fail
- Each fabric needs to run at no more than ~50-60% busy so that if a failover occurs then the remaining fabrics can pickup and handle the full workload without over-utilization and with some extra utilization to spare per fabric
- If an ISL link in a single fabric fails then that fabric runs at 50% capability

ISL Oversubscription – After a Failure
Demand may exceed capacity

- In this case if a switching device fails …or… if the long distance links in a fabric fails then the frame traffic that was traveling across the now broken links will be rerouted through the other fabrics to reach the storage device
- Those remaining fabrics must have enough reserve capacity in order to pick up all of the rerouted traffic while maintaining performance
- Congestion and potential back pressure could occur if all fabrics are running at a high utilization levels – again, probably above 50% or 60% utilization
- Customers should manage their fabrics to allow for rerouted traffic

ISL Oversubscription

Creating Backpressure Problems on an ISL



- In This Example:
  - 8G CHPIDs and ISLs are capable of 760MBps send/receive (800 * .95=760)
  - Two CHPIDs per mainframe (1520MBps) and 4 mainframes (6080MBps)
  - About 42% of I/O activity is across the ISLs and requires **2550MBps**
  - Four ISLs provides 3040MBps – (760MBps * 4) – and redundancy
  - Then one ISL fails leaving only **2280MBps** – (760MBps * 3) – not enough redundancy
  - 2280MBps / 2550MBps = 89% of what is required (Congestion Will Occur)
  - Each CHPID experiences backpressure as the remaining 3 ISLs become congested and unable to handle all of the I/O traffic

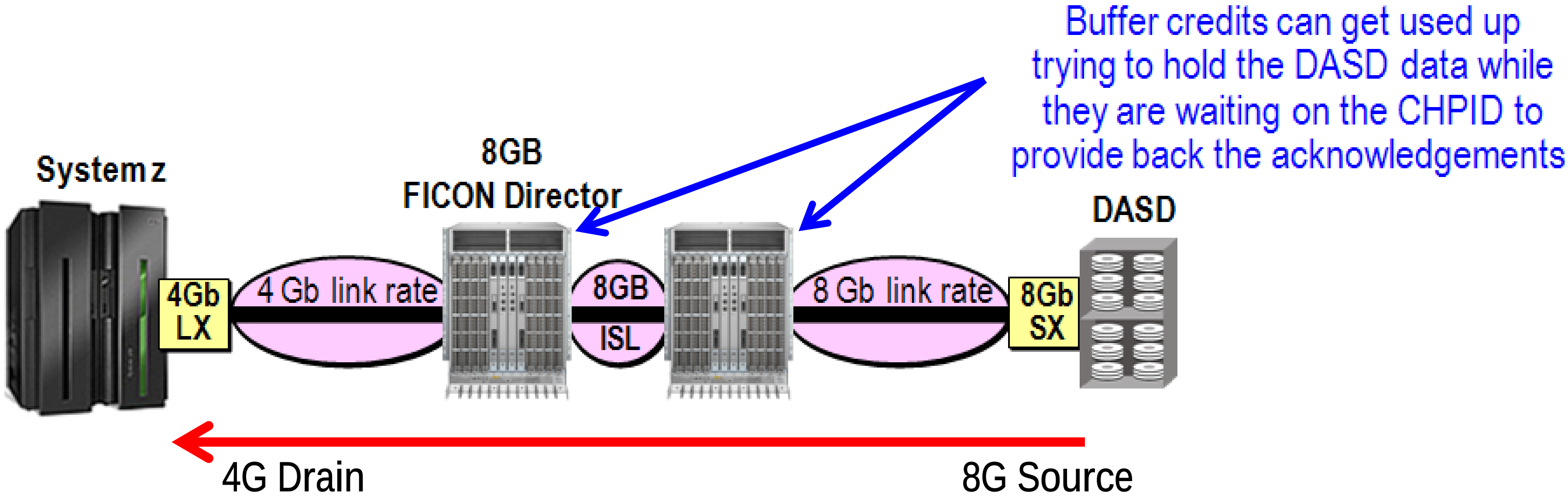
This Section

- Slow Draining Devices

Slow Draining Devices

- Slow draining devices are devices that are trying to handle more information work load than they can consume.
- A slow draining device can exist at any link utilization level where achieved throughput into the slow draining port is lower compared to intended throughput.
- It's very important to note that it can spread into the fabric and can slow down unrelated flows in the fabric.
- What causes slow draining devices?
- The most common cause is within the device or server itself. The most common cause is because a device has a slower link rate then the rest of the environment.

Slow Draining Devices



- This is potentially a very poor performing, infrastructure!
- DASD is about 90% read, 10% write. So, in this case the "drain" of the pipe is the 4Gb CHPID and the "source" of the pipe is the 8Gb storage port.
- The Source can out perform the Drain. This can cause congestion and back pressure towards the CHPID. The CHPID becomes a slow draining device.

Note: 8G Tape would typically be OK since Tape is usually 90% write and 10% read. Usually the CHPID would be the Source and the Tape port would be the Drain.

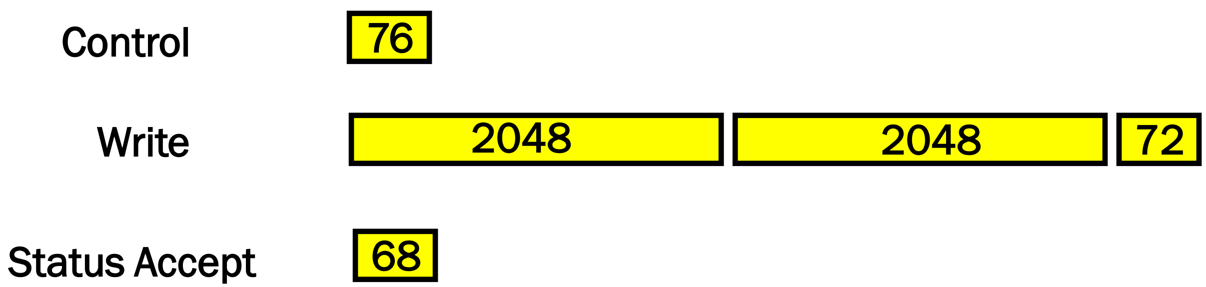
This Section

- Determining Buffer Credits Required
- RMF Reports for Switched-FICON
- Brocade's Buffer Credit Calculation Spreadsheet

Buffer Credits

Why FICON Never Averages a Full Frame Size

- There are three things that are required to determine the number of buffer credits required across a long distance link
  - The speed of the link
  - The cable distance of the link
  - The average frame size
- Average frame size is the hardest to obtain
  - Use the RMF 74-7 records report “FICON Director Activity Report”
  - You will find that FICON just never averages full frame size
  - Below is a simple FICON 4K write that demonstrates average frame size



  - Will not fit into 2 buffers because of frame headers/trailers and SB3

Average = (76+2048+2048+72+68) / 5 = 862 Bytes

Buffer Credits

The Impact of Average Frame Size on Buffer Credits

A distance of 20KM with 100% link utilization				2Gbps	4Gbps	8Gbps	10Gbps	16Gbps
SOF, Header, CRC, EOF	Payload	Total Frame Bytes	Smaller than full frame by xx%	Buffer Credis Required 8b10b	Buffer Credis Required 8b10b	Buffer Credis Required 8b10b	Buffer Credis Required 64b66b	Buffer Credis Required 64b66b
36	2112	2148	0.00%	20	40	80	99	159
36	1966	2002	6.80%	22	43	85	107	170
36	1824	1860	13.41%	23	46	92	115	183
36	1682	1718	20.02%	25	50	99	124	198
36	1540	1576	26.63%	27	54	108	135	216
36	1398	1434	33.24%	30	60	119	149	238
36	1256	1292	39.85%	33	66	132	165	264
36	1114	1150	46.46%	37	74	148	185	296
36	972	1008	53.07%	43	85	169	211	338
36	830	866	59.68%	50	99	197	246	393
36	688	724	66.29%	59	118	235	294	470
36	546	582	72.91%	74	147	293	366	585
36	404	440	79.52%	97	194	387	484	773
36	262	298	86.13%	143	286	571	714	1142
36	120	156	92.74%	273	545	1090	1363	2180
36	36	72	96.65%	591	1181	2362	2952	4724

Created by using Brocades Buffer Credit Calculator

Buffer Credit Starvation
Why not just saturate each port with BCs?

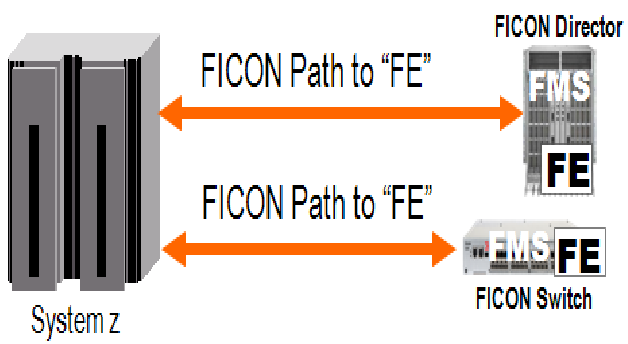
- If a malfunction occurs in the fabric ..... or....
- If a CHPID or device is having a problem...
- It is certainly possible that some or all of the I/O will time out
- If ANY I/O does time out then:
  - All frames & buffers for that I/O (buffer credits) must be discarded
  - All frames & buffers for subsequently issued I/Os (frames and buffer credits) in that exchange must be discarded
    - Remember queued I/O will often drive exchanges ahead of time
  - The failing I/O must be re-driven
  - Subsequent I/O must be re-driven
- The recovery effort for the timed out I/O gets more and more complex – and more prone to also failing – when an over abundance of buffer credits are used on ports

Buffer Credit Starvation

Detecting Problems with FICON BCs

Produce the FICON Director Activity Report by creating the RMF 74-7 records

- Option FCD in ERBRMFxx parmlib member and STATS=YES in IECIOSnn tells RMF to produce the 74-7 *records*
- A FICON Management Server (FMS) license per switching device enables the switch's Control Unit Port (CUP) – always FE – to provide information back to RMF at its request



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 04/12/2009-04.30.0   INTERVAL 000.15.00

RPT VERSION V1R8 RMF   END 04/12/2009-04.45.0   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: 00000131656G

PORT ADDR	-CONNECTION- UNIT	ID	AVG FRAME PACING	AVG FRAME SIZE READ	AVG FRAME SIZE WRITE	PORT BANDWIDTH (MB/SEC) -- READ --	PORT BANDWIDTH (MB/SEC) -- 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

- Analyze the column labeled AVG FRAME PACING for non-zero numbers. Each of these represents the number of times a frame was waiting for 2.5 usec or longer but BC count was at zero so the frame could not be sent

FICON Director Activity Rpt

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	AVG FRAME SIZE WRITE	PORT BANDWIDTH (MB/SEC) -- READ --	PORT BANDWIDTH (MB/SEC) -- 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

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!

We have a BC Calculator that you can use!

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

		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
Parameter										
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.

Splash | B2B Calcuator

Ask your local Brocade SE to provide this to you free of charge!

BROCADE

BROCADE'S MAINFRAME CERTIFICATION

Industry Recognized Professional Certification

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

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

More SAN Sessions at SHARE this week

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: 10079

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

BROCADE

Thank You !