Understanding the Performance Implications of Buffer to Buffer Credit Starvation In a FICON Environment: Frame Pacing Delay

Paper 7080

Steve Guendert

Principal Engineer

CMG 2007 International Conference

December, 2007

Abstract

- *This paper will give a brief review on buffer-to-buffer credits (BB_credits) including current schema for allocating/assigning them. It will then discuss the one method currently available to detect BB_credit starvation on FICON directors, including a discussion on the concept of frame pacing delay. Finally, the author will outline a concept for a mechanism to count BB_credit usage. The paper will conclude with a discussion of another theoretical "drawing board" concept: dynamic allocation of BB_credits on an individual I/O basis similar to the new HyperPAVs concept for DASD.*

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, the Brocade B-weave logo, McDATA, Fabric OS, File Lifecycle Manager, MyView, Secure Fabric OS, SilkWorm, and StorageX are registered trademarks and the Brocade B-wing symbol and Tapestry are trademarks of Brocade Communications Systems, Inc. or its subsidiaries, in the United States and/or in other countries. FICON is a registered trademark of IBM Corporation in the U.S. and 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.

Key References

- Cronin, C. ***Performance Considerations for Cascaded FICON Directors, www-1.ibm.com/servers/eserver/zseries/library/techpapers/gm130237.html,*** March 2003
- Artis, H.P. ***Managing Complex FICON Configurations.*** Performance Associates, Inc. 2005
- Guendert, S. ***Buffer-to-Buffer Credits and Their Effect on FICON Performance,*** CMG Measure IT, March 2005.
- Allen, A.O., ***Probability, Statistics, and Queueing Theory,*** Academic Press, 1978.
- Guendert, S. ***Taking FICON To the Next Level: Cascaded High Performance FICON,*** Proceedings of the Computer Measurement Group, 2005.
- Artis, H.P. and Guendert, S. ***Designing and Managing FICON Interswitch Link infrastructures,*** Proceedings of the Computer Measurement Group, 2006.
- Guendert, S. and Lytle, D. ***Buffer to Buffer Credit Management: An Oxymoron?*** zJournal, June/July 2007

Agenda

- Buffer to Buffer Credit Management: An Oxymoron
- Review of the basics: end to end and buffer to buffer flow control
- FICON director architectures and BB Credits
- Frame Pacing Delay
- Ideas for improvement
  - Counting BB Credits
  - Dynamic allocation of BB credits

Wikipedia defines Oxymoron

- An oxymoron is a figure of speech that combines two normally contradictory terms. *Oxymoron* is from Greek *oxy* ("sharp") and *moros* ("dull"). Thus the word *oxymoron* is itself an oxymoron.
- Oxymorons are a proper subset of the expressions called contradictions in terms. What distinguishes oxymorons from other paradoxes and contradictions is that they are used intentionally, for rhetorical effect, and the contradiction is only apparent, as the combination of terms provides a novel expression of some concept, such as "cruel to be kind".

Buffer to Buffer Credit management: an oxymoron

- There is no way to actually report/track how many BB credits are being used.
- The RMF 74-7 record comes close, but names the field something else.
- Published rules of thumb mistakenly assume full frames
- Similar to dynamic PAVs, end users tend to overkill BB credit assignment
  - Can lead to director configuration issues which may cause an outage to fix

Review of the basics

- ESCON DIBs-reviewed in paper in detail
- End to End Flow Control
- Buffer to Buffer flow control

Packet Flow

- Fundamental concepts:
  - –Prevent a transmitter from overrunning a receiver by providing real time signals back from the receiver to pace the transmitter
  - –Manage each I/O as a unique instance

End to End Flow Control

- Used by Class 1 and Class 2 service between 2 end nodes.
- Nodes monitor end to end flow control between themselves.
  – Intervening directors do not participate.
- End to end flow control is always managed between a specific pair of node ports
  – Many different values possible.

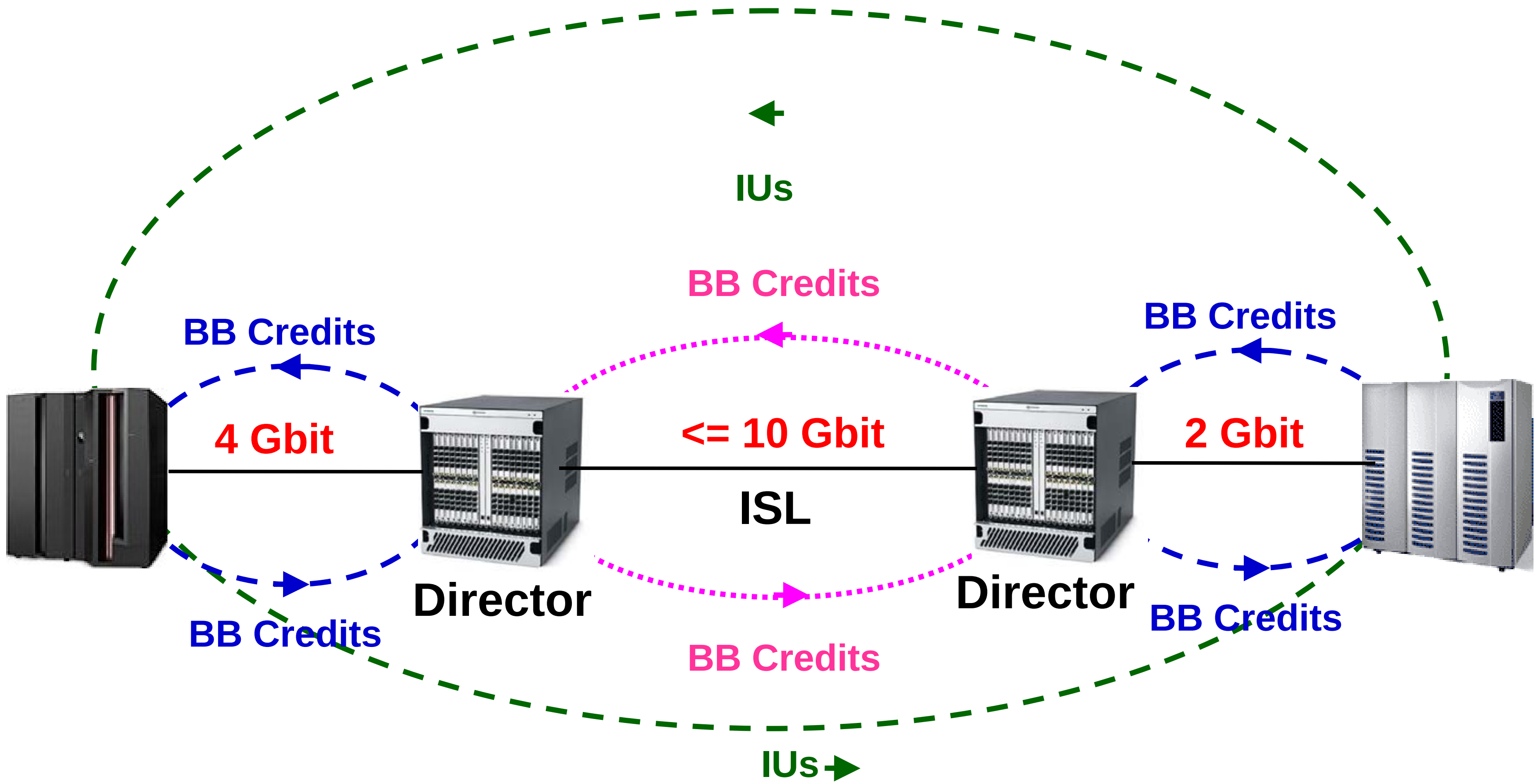
Buffer to Buffer Flow Control

- Flow control between two optically adjacent ports in the I/O path.
- Separate, independent pool of credits manages Buffer-to-Buffer flow control (BB Credits).

Buffer-to-Buffer flow control (3)

- It takes light 5 nsec to propagate through 1 meter of optical fiber
  - 50 µsec to travel 10 km.
- Faster links, longer distances leads to a performance drag similar to ESCON droop.
  - Need BB credit values>1 and frame streaming.
    - Frame streaming: allowing a sending port to send more than 1 frame without having to wait for a response to each.
    - Approach 100% link utilization

B-to-B and E-to-E Control



Buffering in the director allows each segment to run at a different data rate. When fibre costs are high between your sites, you can trade off ISL data rate versus the higher cost of the director ports

Buffer Credit Concepts

- Define the maximum amount of data that can be sent prior to an acknowledgement

- Buffer credits are physical ASIC port or card memory resources and are finite in number as a function of cost

- Within a fabric, each port may have a different number of buffer credits

- The number of available buffer credits is communicated at fabric logon (FLOGI)

Buffer Credit Concepts(2)

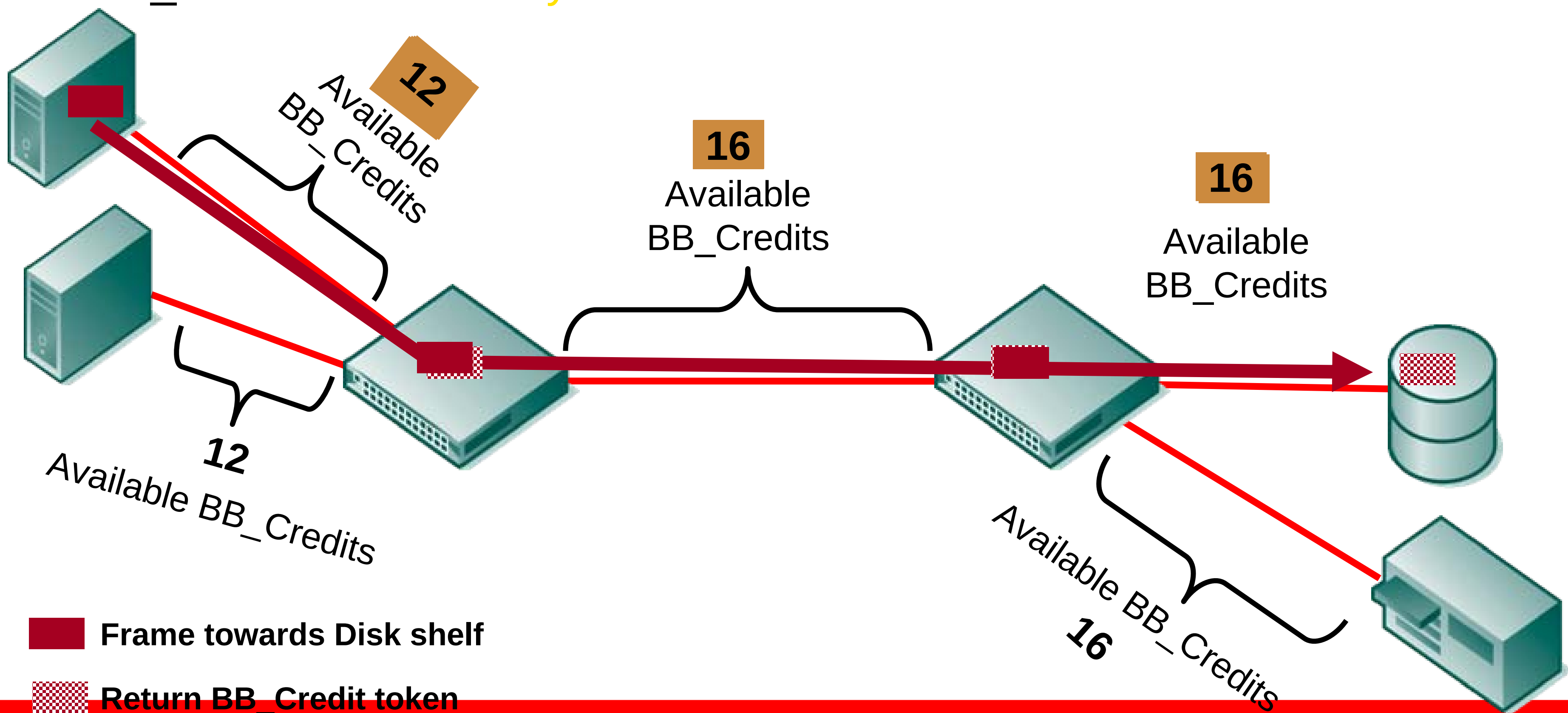
- One buffer credit allows a device to send one 2112 byte frame of data (2K usable for z/OS data)
- Assuming that each credit is completely full, you need one credit for every 1 KM of link length over a 2 Gbit fibre
- Unfortunately, z/OS disk workload rarely produce full credits. For a 4K transfer, the average frame size for a 4K transfer is 819 bytes
- Hence, five credits would be required per KM over a 2 Gbit fibre

BB credit consumption tracking process

- Before any data frames are sent, the transmitter sets a counter equal to the BB-credit value.
- For each data frame sent by the transmitter, the counter is decremented by one.
- Upon receipt of a data frame, the receiver sends a status frame (R_RDY) to the transmitter indicating that the data frame was received AND the buffer can receive another data frame.
- For each R_RDY received by the transmitter, the counter is incremented by one.

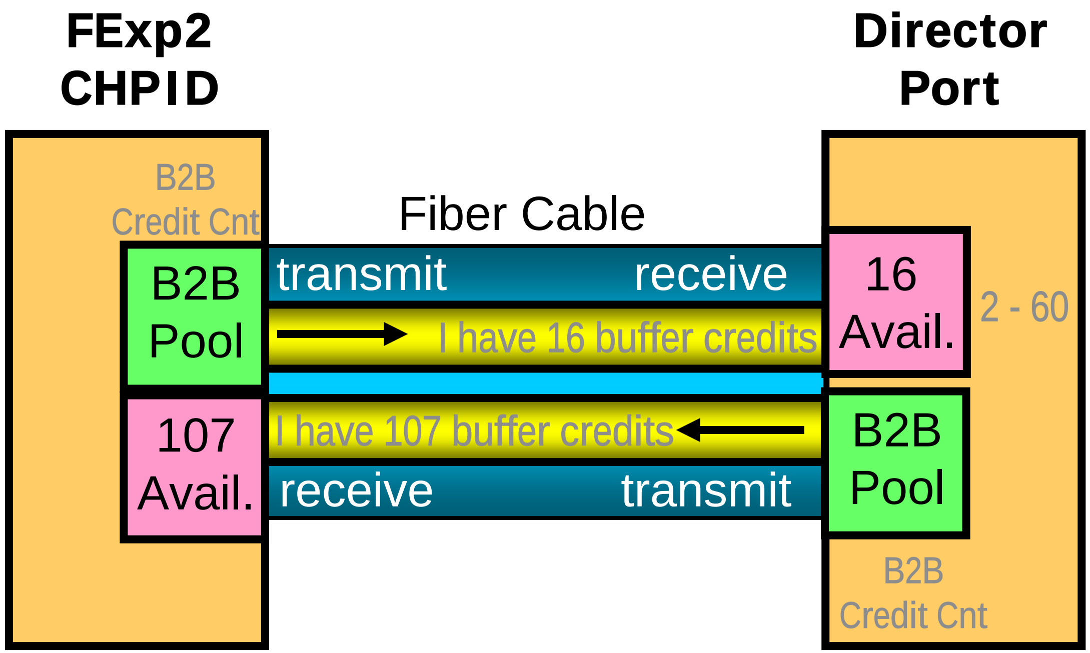
Buffer Credits

- BB_Credits are the “admission control” mechanism in FC to ensure that FC switches don’t run out of buffers (FC Switches cannot drop frames)
- For Devices operating at FC Class 3 (most devices), Buffer Credits are negotiated at login.
- BB_Credits are the only flow-control mechanism for FC Class 3.



How Buffer-to-Buffer Credits Work

- A Fibre 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 through the receiver increments the B2B credit counter

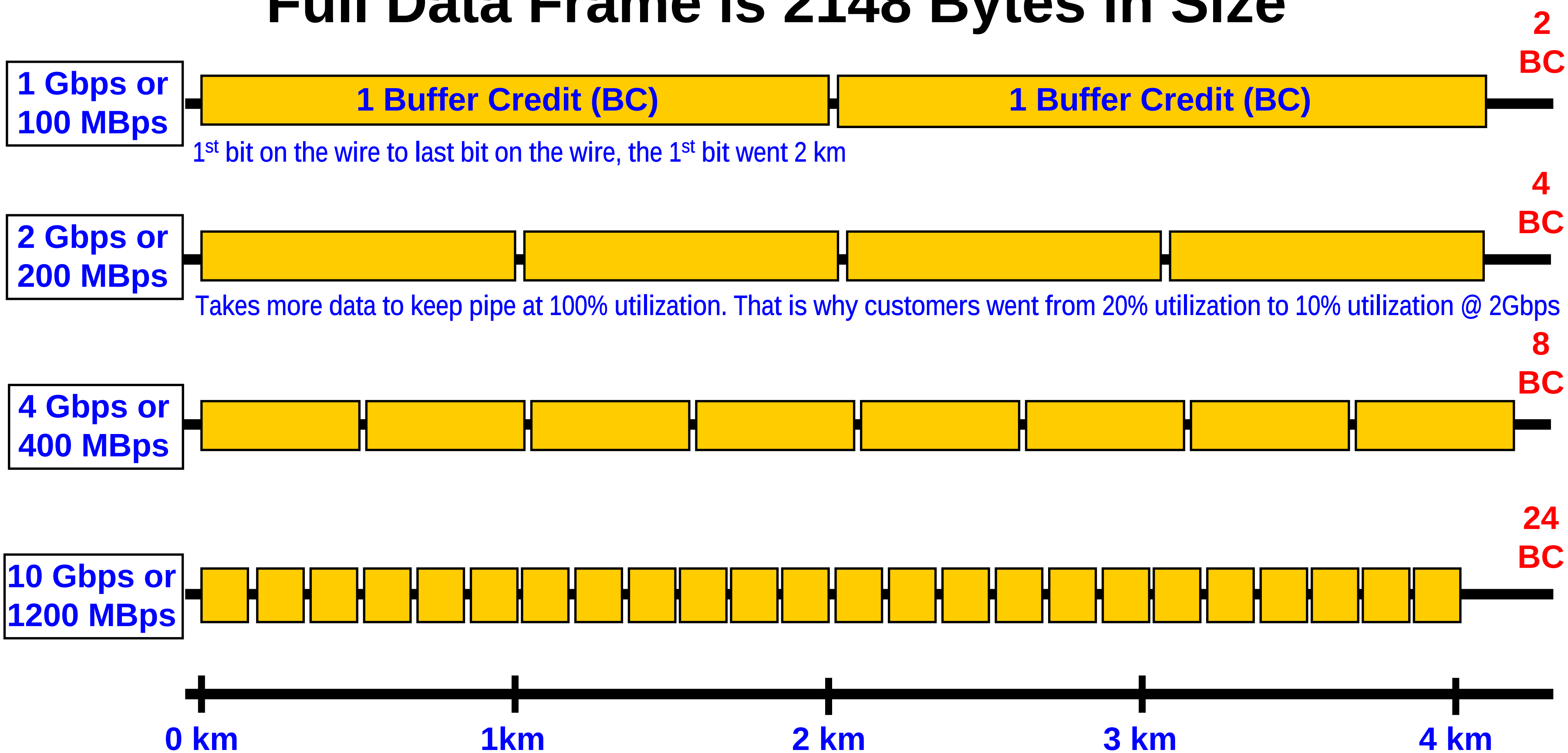


The send and receive ports each negotiate the number of available credits!

Hence, the inbound and outbound ports for a fibre pair could have different numbers of credits!

Buffer Credits relative to Link Speed

Full Data Frame is 2148 Bytes in Size



These frames never got smaller, they were always 2148 bytes – the transport is just faster each time. Therefore, to cover the same distance, at higher speeds, requires more buffer credits

Calculating the number of buffer credits

- What you must know to do this correctly is:
  - Link speed (1, 2, 4, 10Gbps) – easy to get
  - Actual fiber run distance that the frame must traverse – easy to get
  - The size of the frame – very hard to get
- Formula for assigning buffer credits (assumes 2148 frame size)
  - 1 Gbps
    - Distance (in km) / 2 + 20%
  - 2 Gbps
    - Distance (in km) + 20%
  - 4 Gbps
    - Distance (in km) x 2 + 20%
  - 10 Gbps
    - Distance (in km) x 6 + 20%
- But does this always work?

Buffer Credits Required

By Size of Frame and Link Speed

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 8b10b	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

What is the optimal number of BB Credits?

- Optimal number of credits is determined by:
  - Distance (frame delivery time)
  - Processing time at receiving port
  - Link signaling rate
  - Size of frames being transmitted

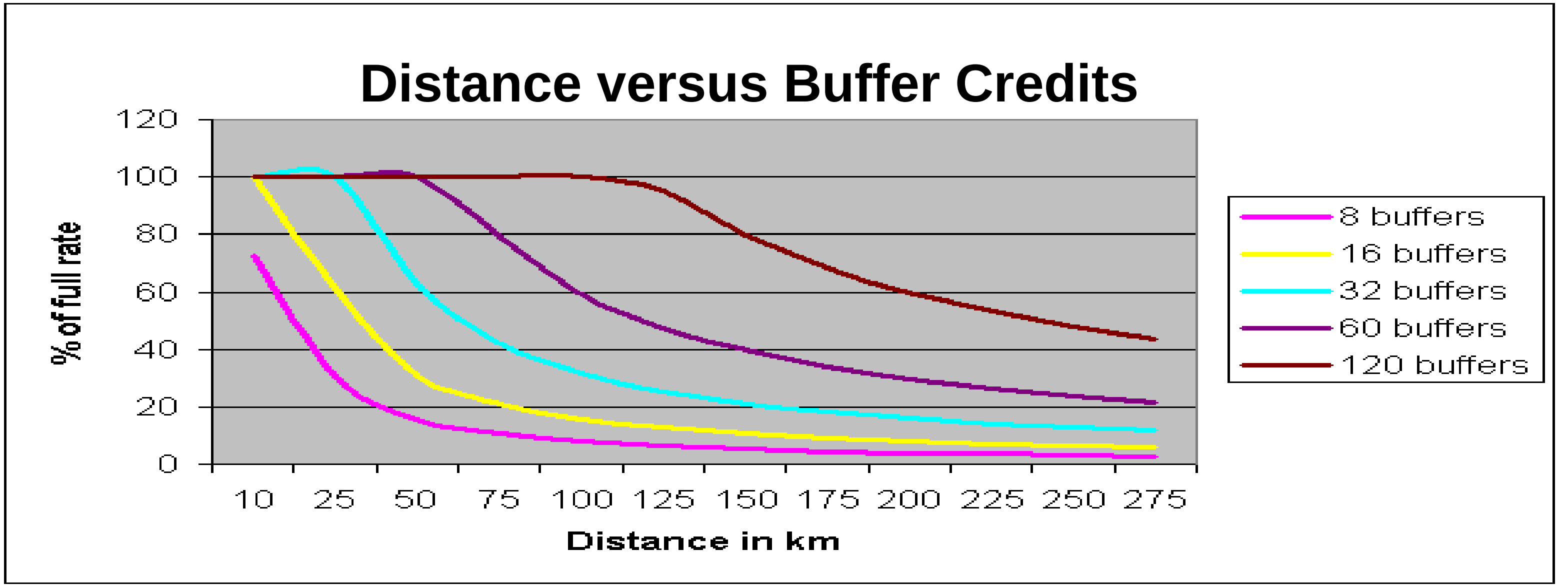
- Optimal # BB_Credit =
  - **(Round-trip receiving time+Receiving_port processing time)**
    - **Frame Transmission time**

- * As the link speed increases, the frame transmission time is reduced; therefore, as we get faster iterations of FICON such as FICON Express4 and Express8, the amount of credits need to be increased to obtain full link utilization, even in a short distance environment!

Why an optimal number?

- Analogous to DASD and cache sizing
- Law of diminishing marginal returns
- Exceeding the optimal number of BB Credits does nothing to increase performance, it merely increases your costs.
- Optimal number of BB Credits allows for performance optimized distance solutions.

Data Droop for Over Distance @ 2Gb/s



- Cron

- Hence, serious consideration must be give to the assignment of credits to ports on director architectures that share a pool of credits among the ports on a card. While relatively few credits (16) might be assigned to local devices, the bulk of the credits should be assigned to ISLs

- *For data chaining OLTP workloads, assume a worst case 512 byte average credit size to avoid any potential of droop*

- *MIDAWs and RTD/WTD substantially increase the average credit size*

How Do MIDAWs Effect ISLs?

- IDAW – Each Block Is A CCW
- MIDAW – Move Entire Chain As One I/O
- Example – 4K Block Extended Format

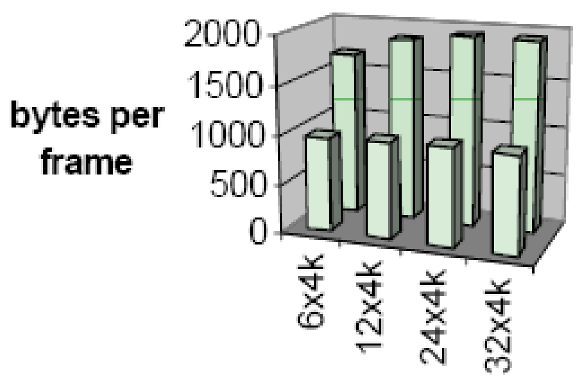
I/O	Data	BB Credits
Write	4096	2
EF Data	35	1

- Chain Of 16
  - Total Data = (4096+35) * 16 = 66096
  - IDAW, Total BB Credits = 3 * 16 = 48
  - MIDAW, Total BB Credits = 66096/2048 = 33

MIDAW Uses 30% Fewer BB Credits So You Can Go 30% Farther

More bytes/frame→more efficient usage of buffer to buffer credits

Average frame size for FICON Express4 channel MIDAWs measurements



	6x4k	12x4k	24x4k	32x4k
no midaws	965	982	1000	997
with midaws	1692	1873	1943	1943

- Reference: Cathy Cronin “IBM System z9 and FICON Express 4 Performance Update. SHARE Tampa Proceedings, Feb 2006

FICON Director Architectures-old way of configuring BB credits

- Inrange/CNT FC9000 and McDATA 6000 series
- BB credits were assignable port by port
  - 1 ASIC per port
  - Each port had a range of BB credits
  - Changing your configuration was typically an offline operation

FICON Director Architectures: New way

- More ports/ASIC on a port card
- BB Credits pooled per ASIC
- This is good and bad
- Increased speeds of links has caused BB credit configuration to become a capacity planning exercise

Frame Pacing Delay

RMF 74 Subtype 7 Records

Global Data

Switch Data

Connector Data

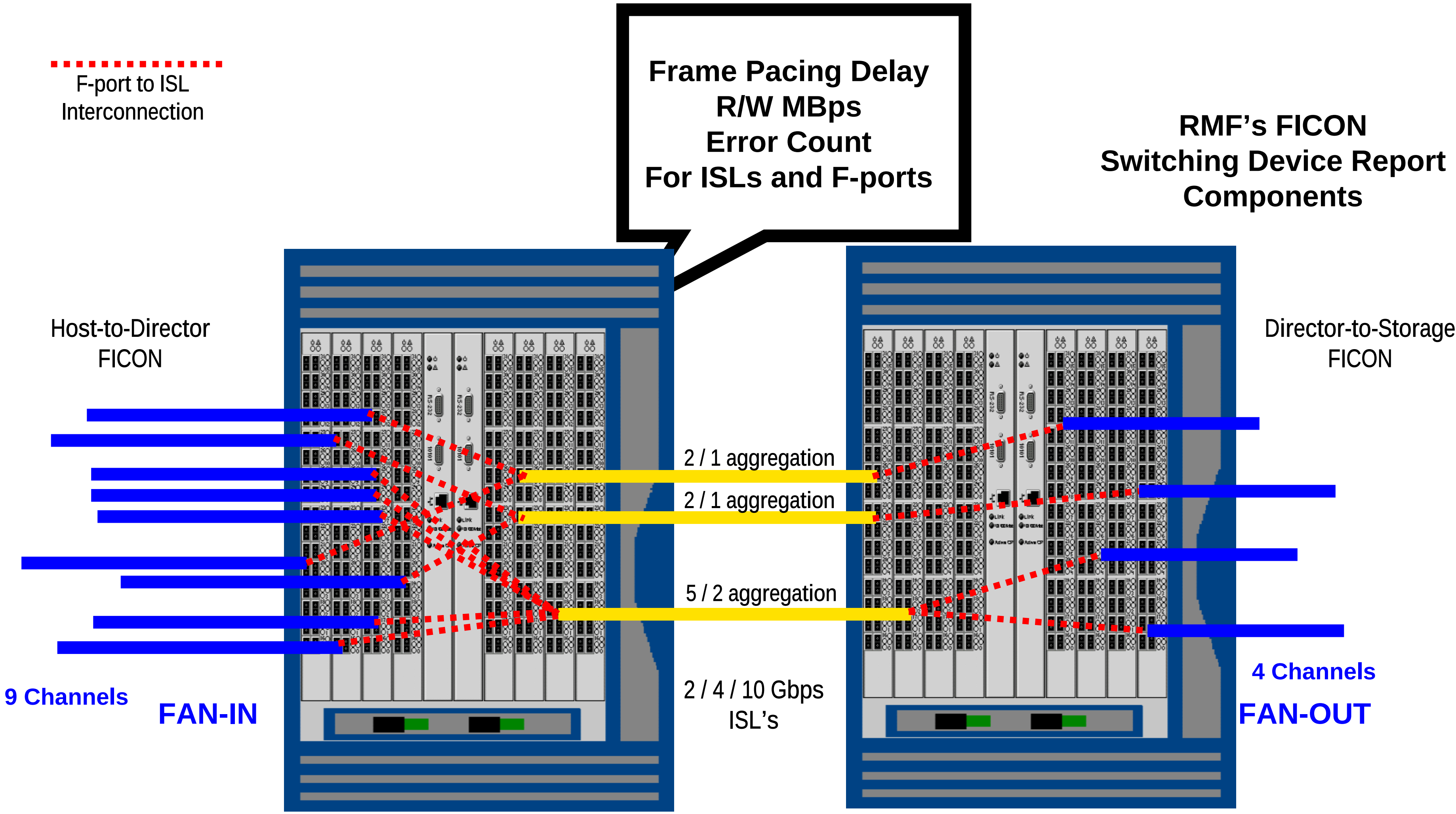
Port Data

- Four data classes of data are reported by the 74 subtype 7

- Port data includes average read/write frame sizes, average bandwidth, error count, and pacing delays for each port. *Frame pacing occurs when a director port exhausts its available credits.* Frame pacing delays are measured in 2.5 micro-second units

- Data is collected for each RMF interval if FCD is specified in your *ERBRMFnn* parmlib member

FICON Director Measurements



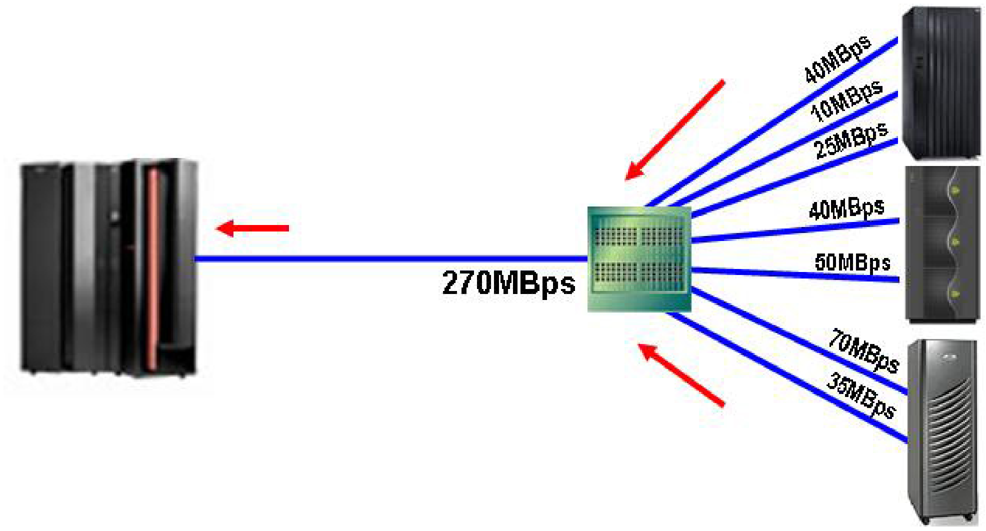
RMF 74 subtype 7 records turned on and <u>CUP code implemented!</u>

Frame Pacing Delay

- AVG FRAME PACING
  - Defined by RMF as the average number of time intervals of 2.5 microseconds that a frame has to wait before it could be transmitted due to no buffer credits being available on a given director port.
- You always want to see a zero value in this field!
  - Reporting on this value was one of the primary reason that the RMF 74-7 record was developed – it was not needed for ESCON
  - A non-zero value in the AVG FRAME PACING field indicates that you have an issue with insufficient BB Credits
  - It is critical to use CUP in any FICON environment in which distance extension is being utilized
  - 4Gbps may create more Frame Pacing Delay issues than 2Gbps
- z/OS disk workloads rarely use a "full" 2148 byte credit
  - For example, with a 4k block transfer, the average frame size for each 4k transfer is typically about 819 bytes

Where does frame pacing occur?

- Incorrect number of BB credits (not enough) assigned on a port
- Poorly architected environment:



FICON Director Activity Report

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

z/OS V1R4 SYSTEM ID KS01 START 10/03/2005-13.55.00 INTERVAL 000.05.00
RPT VERSION V1R2 RMF END 10/03/2005-14.00.00 CYCLE 1.000 SECONDS
IODF = 0C CR-DATE: 07/07/2005 CR-TIME: 18.47.41 ACT: POR
SWITCH DEVICE: 0001 SWITCH ID: ** TYPE: 006140 MODEL: 001 MAN: MCD PLANT: 01 SERIAL: 0000013A2

PORT	-CONNECTION-		AVG FRAME	AVG FRAME SIZE		PORT BANDWIDTH (MB/SEC)		ERROR
ADDR	UNIT	ID	PACING	READ	WRITE	-- READ --	-- WRITE --	COUNT
04	CHP-H	47	0	433	593	1.01	1.52	0
05	CU	----	0	830	1195	2.30	5.81	0
06	CU	----	0	140	69	0.00	0.00	0
07	CU	C052	0	591	85	0.00	0.00	0
	CU	C050						
08	CHP-H	45	0	400	577	0.98	1.55	0
09	CU	----	0	1355	374	8.42	1.00	0
0A	------	----	0	0	0	0.00	0.00	0
0B	------	----	0	0	0	0.00	0.00	0
0C	CHP-H	4B	0	431	601	1.02	1.56	0
0D	CU	----	0	776	374	3.57	1.41	0
0E	CU	----	0	0	0	0.00	0.00	0
0F	CU	C053	0	1773	78	0.08	0.00	0
	CU	C051						
10	CHP-H	46	0	366	716	1.06	2.32	0
11	CU	----	0	1099	393	4.62	0.97	0
12	------	----	0	0	0	0.00	0.00	0
13	------	----	0	0	0	0.00	0.00	0
14	CHP-H	50	0	533	832	0.32	0.58	0
15	CU	----	0	868	1223	2.25	5.55	0
16	CU	----	0	158	72	0.00	0.00	0
17	CU	C053	0	1761	77	0.09	0.00	0
	CU	C051						
18	CHP-H	49	0	378	745	1.04	2.33	0
19	CU	----	0	1118	399	4.83	0.99	0
1A	------	----	0	0	0	0.00	0.00	0
1B	------	----	0	0	0	0.00	0.00	0
1C	CHP-H	51	0	737	535	0.34	0.17	0
1D	CU	----	0	877	1230	3.22	7.91	0
1E	CU	----	0	0	0	0.00	0.00	0
1F	CU	C052	0	590	82	0.00	0.00	0
	CU	C050						
20	CHP-H	4A	0	374	756	1.04	2.40	0
21	CU	----	0	1472	413	3.51	0.36	0
22	------	----	0	0	0	0.00	0.00	0
23	------	----	0	0	0	0.00	0.00	0

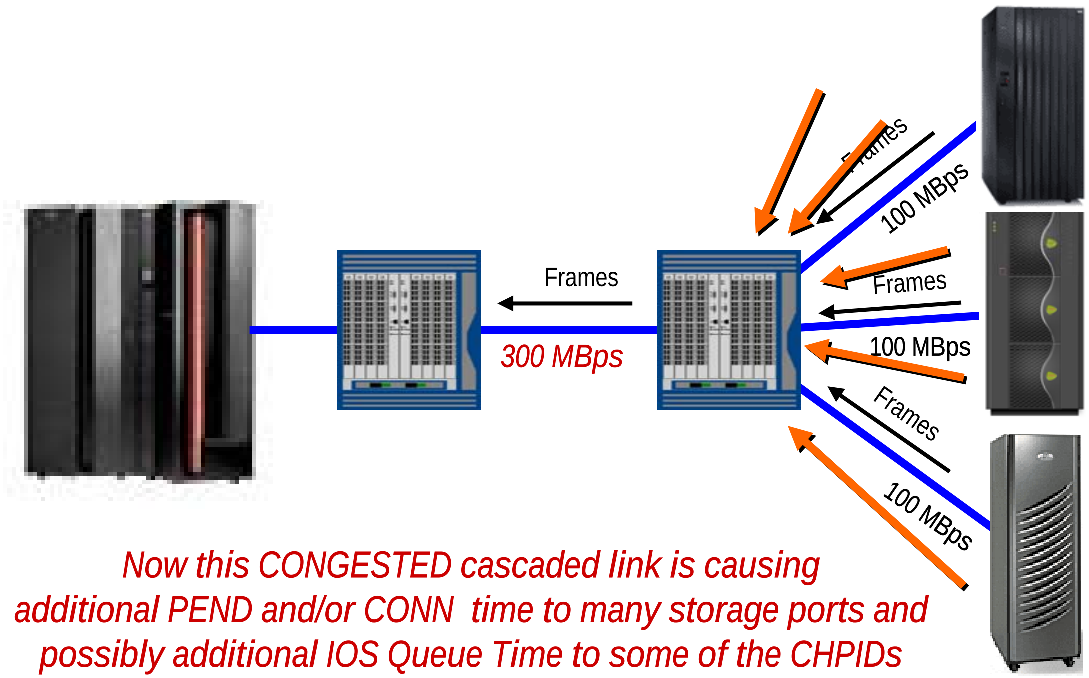
Frame Pacing Delay Being Reported

```
                                        F I C O N   D I R E C T O R   A C T I

          z/OS V1R7                SYSTEM ID PDM1            DATE 11/28/2006
                                   RPT VERSION V1R7 RMF      TIME 21.44.00

IODF = 70   CR-DATE: 09/20/2006   CR-TIME: 10.49.34     ACT: POR

SWITCH DEVICE: 006E   SWITCH ID: **     TYPE: 006140   MODEL: 001   MAN: MCD
```

PORT ADDR	-CONNECTION- UNIT	ID	AVG FRAME PACING	AVG FRAME SIZE READ	WRITE	PORT BANDWIDTH (MB/SEC -- READ --	-- WRITE -
04	SWITCH	----	3	71	1715	0.32	41.7
05	CHP	5E	0	0	0	0.00	0.0
06	CHP	C0	0	259	839	0.01	0.0
07	CHP	C0	0	678	631	0.05	0.0
08	SWITCH	----	0	71	1689	0.38	39.0

Local Frame Pacing Delay

- How can you run out of buffer credits inside a datacenter?
  - Frame pacing delays occur when multiple, heavily used paths merge into a single FICON link
  - Frame pacing delays can contribute to PEND, DISC, and CONN time measurements



Now this CONGESTED cascaded link is causing additional PEND and/or CONN time to many storage ports and possibly additional IOS Queue Time to some of the CHPIDs

Frame Pacing Delay is caused by running out of buffer credits!

UCBs serviced by these storage ports are probably experiencing additional delays usually reported as PEND Time and CONN Time and sometimes as DISC time

Frame Pacing Delay came about with FC and FICON so it is not a factor in ESCON performance!

Suggestion

- Use the RMF 74-7 record as a way to help narrow down/troubleshoot performance problems in your environment

BROCADE

NEW IDEAS

RMF 74-7 changes

- Change the field “ AVG Frame Pacing” to BB Credit starvation
- Add a field for open exchanges

Counting BB Credits in use

- Add to the director management software, and/or RMF the capability to calculate and report the number of BB credits in use during an RMF interval via CUP

Microsoft Excel
Worksheet

Dynamic Allocation of BB Credits

- Similar in concept to HyperPAVs and DASD
- Allow the z/OS I/O Supervisor (IOS) to dynamically assign the number of BB Credits required on an individual I/O basis, via CUP.

BROCADE

THANK YOU