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
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

No Congestion and No Backpressure

10am – 3pm

Road can handle up to 200 cars/trucks per minute so the traffic can run up to the speed limit
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

Congestion and Backpressure

3pm – 6pm

Road can only handle up to 200 cars/trucks per minute so traffic speed reduced (congested)

Road can handle up to 200 cars/trucks per minute so the traffic is running at the speed limit
This Section

• Very basic flow for the Build Fabric process and how Buffer Credits get initialized
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 LinkParms (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 LinkParms (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

**FI CON X8 CHPI D**

- **B2B Credit Cnt**
  - **8 Avail.**
  - **Fiber Cable**
    - **transmit**
    - **receive**
    - **I have 8 buffer credits**
  - **BC Pool**
    - **I have 40 buffer credits**
    - **receive**
    - **transmit**
    - **40 Avail.**
    - **B2B Credit Cnt**

**Brocade Dir. Port**

- **BC Pool**
  - **8 Avail.**
  - **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

---

**CHPID (8Gbps)**

- 8Gbps CHPID Has 40 BCs
- Tx Credit **8**

**F_Port (8Gbps)**

- 8Gbps F_Port has 8 BCs
- Tx Credit **40**

Data Frames

R_RDYs

On to End Device
BB Credit Droop

Not Enough BB Credits

Ample BB Credits

4 x less BB Credits, still ample
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.

Available BB_Credits

- Frame towards Disk shelf
- Frame towards Tape
- Return BB_Credit token

Disk shelf capable of sustaining 800MB/sec

Tape capable of sustaining 160MB/sec uncompressed
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.
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).

Provides a potential five-9’s of availability environment.
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

Each set <=12.5% of Bandwidth

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

<table>
<thead>
<tr>
<th>Control</th>
<th>76</th>
</tr>
</thead>
<tbody>
<tr>
<td>Write</td>
<td>2048</td>
</tr>
<tr>
<td>Status Accept</td>
<td>68</td>
</tr>
</tbody>
</table>

• 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

<table>
<thead>
<tr>
<th>SOF, Header, CRC, EOF</th>
<th>Payload</th>
<th>Total Frame Bytes</th>
<th>Smaller than full frame by xx%</th>
<th>2Gbps</th>
<th>4Gbps</th>
<th>8Gbps</th>
<th>10Gbps</th>
<th>16Gbps</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Buffer Credits Required 8b10b</td>
<td>Buffer Credits Required 8b10b</td>
<td>Buffer Credits Required 8b10b</td>
<td>Buffer Credits Required 64b66b</td>
<td>Buffer Credits Required 64b66b</td>
</tr>
<tr>
<td>36</td>
<td>2112</td>
<td>2148</td>
<td>0.00%</td>
<td>20</td>
<td>40</td>
<td><strong>80</strong></td>
<td>99</td>
<td>159</td>
</tr>
<tr>
<td>36</td>
<td>1866</td>
<td>2002</td>
<td>6.80%</td>
<td>22</td>
<td>43</td>
<td>85</td>
<td>107</td>
<td>170</td>
</tr>
<tr>
<td>36</td>
<td>1824</td>
<td>1860</td>
<td>13.41%</td>
<td>23</td>
<td>46</td>
<td>92</td>
<td>115</td>
<td>183</td>
</tr>
<tr>
<td>36</td>
<td>1682</td>
<td>1718</td>
<td>20.02%</td>
<td>25</td>
<td>50</td>
<td>99</td>
<td>124</td>
<td>198</td>
</tr>
<tr>
<td>36</td>
<td>1540</td>
<td>1576</td>
<td>26.63%</td>
<td>27</td>
<td>54</td>
<td>108</td>
<td>135</td>
<td>216</td>
</tr>
<tr>
<td>36</td>
<td>1398</td>
<td>1434</td>
<td>33.24%</td>
<td>30</td>
<td>60</td>
<td>119</td>
<td>149</td>
<td>238</td>
</tr>
<tr>
<td>36</td>
<td>1256</td>
<td>1292</td>
<td>39.85%</td>
<td>33</td>
<td>66</td>
<td>132</td>
<td>165</td>
<td>264</td>
</tr>
<tr>
<td>36</td>
<td>1114</td>
<td>1150</td>
<td>46.46%</td>
<td>37</td>
<td>74</td>
<td>148</td>
<td>185</td>
<td>296</td>
</tr>
<tr>
<td>36</td>
<td>872</td>
<td>1008</td>
<td>53.07%</td>
<td>43</td>
<td>85</td>
<td>159</td>
<td>211</td>
<td>338</td>
</tr>
<tr>
<td>36</td>
<td>830</td>
<td>866</td>
<td>59.68%</td>
<td>50</td>
<td>99</td>
<td><strong>197</strong></td>
<td>246</td>
<td>393</td>
</tr>
<tr>
<td>36</td>
<td>588</td>
<td>724</td>
<td>66.29%</td>
<td>59</td>
<td>118</td>
<td>225</td>
<td>294</td>
<td>470</td>
</tr>
<tr>
<td>36</td>
<td>546</td>
<td>582</td>
<td>72.91%</td>
<td>74</td>
<td>147</td>
<td>293</td>
<td>366</td>
<td>585</td>
</tr>
<tr>
<td>36</td>
<td>404</td>
<td>440</td>
<td>79.52%</td>
<td>97</td>
<td>194</td>
<td>387</td>
<td>484</td>
<td>773</td>
</tr>
<tr>
<td>36</td>
<td>262</td>
<td>298</td>
<td>86.13%</td>
<td>143</td>
<td>286</td>
<td>571</td>
<td>714</td>
<td>1142</td>
</tr>
<tr>
<td>36</td>
<td>120</td>
<td>156</td>
<td>92.74%</td>
<td>273</td>
<td>545</td>
<td>1090</td>
<td>1363</td>
<td>2180</td>
</tr>
<tr>
<td>36</td>
<td>36</td>
<td>72</td>
<td>96.65%</td>
<td>591</td>
<td>1181</td>
<td>2362</td>
<td>2952</td>
<td>4724</td>
</tr>
</tbody>
</table>

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
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

- 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

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)

<table>
<thead>
<tr>
<th>Link Speed</th>
<th>1 Gbps</th>
<th>2 Gbps</th>
<th>4 Gbps</th>
<th>8 Gbps</th>
<th>10 Gbps</th>
<th>16 Gbps</th>
<th>32 Gbps</th>
<th>40 Gbps</th>
<th>100 Gbps</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.6635E-05</td>
<td>3.33E-05</td>
<td>6.65E-05</td>
<td>1.33E-04</td>
<td>2.66E-04</td>
<td>5.32E-04</td>
<td>1.065E-03</td>
<td>2.13E-03</td>
<td>1.6635E-05</td>
</tr>
</tbody>
</table>

**Parameter**

- Velocity of light in fibre: 2000000000000 m/s
- Name seconds per byte: 9.41E-09
- Framelength in seconds (dependent on cell/10): 0.055-06
- Framelength in km (dependent on cell/10): 1.01

**Buffer Credit Calculation**

To determine kilometers from miles, type miles into cell D16:

<table>
<thead>
<tr>
<th>Miles</th>
<th>Equals</th>
<th>Kilometers</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td></td>
<td>24</td>
</tr>
</tbody>
</table>

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 bytes
- Type in the total kilometers of the wire run in cell D20: 24 kilometers

**Description**

- Framelength takes up this many kilometers on the wire (calculated from frame size in cell D19)
- Buffer credits @ 100% BW Utilization raw calculation:
- Buffer credits @ 100% BW Utilization rounded up:

<table>
<thead>
<tr>
<th>Link Speed</th>
<th>1 Gbps</th>
<th>2 Gbps</th>
<th>4 Gbps</th>
<th>8 Gbps</th>
<th>10 Gbps</th>
<th>16 Gbps</th>
<th>32 Gbps</th>
<th>40 Gbps</th>
<th>100 Gbps</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.61</td>
<td>0.80</td>
<td>0.40</td>
<td>0.22</td>
<td>0.16</td>
<td>0.10</td>
<td>0.05</td>
<td>0.04</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>29.83</td>
<td>59.66</td>
<td>119.32</td>
<td>238.64</td>
<td>298.30</td>
<td>477.26</td>
<td>954.58</td>
<td>1193.21</td>
<td>2983.02</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>60</td>
<td>120</td>
<td>239</td>
<td>299</td>
<td>478</td>
<td>955</td>
<td>1194</td>
<td>2984</td>
</tr>
</tbody>
</table>

Ask your local Brocade SE to provide this to you free of charge!
Brocade Certified Architect for FICON
Industry Recognized Professional Certification
We Can Schedule A Class In Your City – Just Ask!

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:

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!!

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.
Thank You!