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

- Guendert, S. *Buffer-to-Buffer Credits and Their Effect on FICON Performance*, CMG Measure IT, March 2005.
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
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

**Fiber Cable**

- **transmit**
- **receive**

**B2B Pool**

- 107 Avail.
- 16 Avail.

**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**

<table>
<thead>
<tr>
<th>Link Speed</th>
<th>Buffer Credits</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Gbps or 100 MBps</td>
<td>2 BC</td>
</tr>
<tr>
<td>2 Gbps or 200 MBps</td>
<td>4 BC</td>
</tr>
<tr>
<td>4 Gbps or 400 MBps</td>
<td>8 BC</td>
</tr>
<tr>
<td>10 Gbps or 1200 MBps</td>
<td>24 BC</td>
</tr>
</tbody>
</table>

1st bit on the wire to last bit on the wire, the 1st bit went 2 km

Takes more data to keep pipe at 100% utilization. That is why customers went from 20% utilization to 10% utilization @ 2Gbps

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**

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

- * 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 \( @ 2 \text{Gb/s} \)

- Cron

- Hence, serious consideration must be given 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

<table>
<thead>
<tr>
<th>I/O</th>
<th>Data</th>
<th>BB Credits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Write</td>
<td>4096</td>
<td>2</td>
</tr>
<tr>
<td>EF Data</td>
<td>35</td>
<td>1</td>
</tr>
</tbody>
</table>

- 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 $\rightarrow$ more efficient usage of buffer to buffer credits

- 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

- 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

Frame Pacing Delay
R/W MBps
Error Count
For ISLs and F-ports

RMF’s FICON
Switching Device Report
Components

F-port to ISL
Interconnection

Host-to-Director
FICON

Director-to-Storage
FICON

2 / 4 / 10 Gbps
ISL’s

2 / 1 aggregation
5 / 2 aggregation
2 / 4 / 10 Gbps
ISL’s

9 Channels
FAN-IN

4 Channels
FAN-OUT

RMF 74 subtype 7 records turned on and CUP code implemented!
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

<table>
<thead>
<tr>
<th>Addr</th>
<th>Unit</th>
<th>Connection</th>
<th>Pacing</th>
<th>Addr</th>
<th>Unit</th>
<th>Connection</th>
<th>Pacing</th>
<th>Addr</th>
<th>Unit</th>
<th>Connection</th>
<th>Pacing</th>
<th>Addr</th>
<th>Unit</th>
<th>Connection</th>
<th>Pacing</th>
</tr>
</thead>
<tbody>
<tr>
<td>04</td>
<td>CHP-H</td>
<td>47</td>
<td>0</td>
<td>05</td>
<td>CU</td>
<td>0</td>
<td>0</td>
<td>06</td>
<td>CU</td>
<td>0</td>
<td>0</td>
<td>07</td>
<td>CU</td>
<td>C052</td>
<td>0</td>
</tr>
<tr>
<td>08</td>
<td>CHP-H</td>
<td>45</td>
<td>0</td>
<td>09</td>
<td>CU</td>
<td>410</td>
<td>0</td>
<td>0A</td>
<td>------</td>
<td>0</td>
<td>0</td>
<td>0B</td>
<td>------</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0E</td>
<td>CU</td>
<td>48</td>
<td>0</td>
<td>0D</td>
<td>CU</td>
<td>69</td>
<td>0</td>
<td>0E</td>
<td>CU</td>
<td>0</td>
<td>0</td>
<td>0F</td>
<td>CU</td>
<td>C053</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>CHP-H</td>
<td>46</td>
<td>0</td>
<td>11</td>
<td>CU</td>
<td>566</td>
<td>0</td>
<td>12</td>
<td>------</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>------</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>CHP-H</td>
<td>50</td>
<td>0</td>
<td>15</td>
<td>CU</td>
<td>533</td>
<td>0</td>
<td>16</td>
<td>CU</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>CU</td>
<td>C053</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>CHP-H</td>
<td>49</td>
<td>0</td>
<td>19</td>
<td>CU</td>
<td>1118</td>
<td>0</td>
<td>1A</td>
<td>------</td>
<td>0</td>
<td>0</td>
<td>1B</td>
<td>------</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1C</td>
<td>CHP-H</td>
<td>51</td>
<td>0</td>
<td>1D</td>
<td>CU</td>
<td>737</td>
<td>0</td>
<td>1E</td>
<td>CU</td>
<td>0</td>
<td>0</td>
<td>1F</td>
<td>CU</td>
<td>C052</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>CHP-H</td>
<td>4A</td>
<td>0</td>
<td>21</td>
<td>CU</td>
<td>1477</td>
<td>0</td>
<td>22</td>
<td>------</td>
<td>0</td>
<td>0</td>
<td>23</td>
<td>------</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

### Understanding the Performance Implications of Buffer to Buffer Credit Starvation: Frame Pacing Delay

Dec 2007
CMG 2007
Steve Guendert

© 2007 Brocade Communications Systems, Inc.
All rights reserved.
## Frame Pacing Delay Being Reported

<table>
<thead>
<tr>
<th>PORT</th>
<th>CONNECTION</th>
<th>AVG FRAME</th>
<th>AVG FRAME SIZE</th>
<th>PORT BANDWIDTH (MB/SEC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>04</td>
<td>SWITCH</td>
<td>3</td>
<td>71</td>
<td>1719</td>
</tr>
<tr>
<td>05</td>
<td>CHP</td>
<td>5E</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>06</td>
<td>CHP</td>
<td>C0</td>
<td>0</td>
<td>259 839</td>
</tr>
<tr>
<td>07</td>
<td>CHP</td>
<td>C0</td>
<td>0</td>
<td>678 631</td>
</tr>
<tr>
<td>08</td>
<td>SWITCH</td>
<td>----</td>
<td>71</td>
<td>1689</td>
</tr>
</tbody>
</table>
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

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
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
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.