**The Association of System
Performance Professionals**

The **Computer Measurement Group**, commonly called **CMG**, is a not for profit, worldwide organization of data processing professionals committed to the measurement and management of computer systems. CMG members are primarily concerned with performance evaluation of existing systems to maximize performance (eg. response time, throughput, etc.) and with capacity management where planned enhancements to existing systems or the design of new systems are evaluated to find the necessary resources required to provide adequate performance at a reasonable cost.

This paper was originally published in the Proceedings of the Computer Measurement Group's 2007 International Conference.

For more information on CMG please visit http://www.cmg.org

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

Stephen R. Guendert
Brocade Communications

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.

Introduction

The introduction of the FICON I/O protocol to the mainframe I/O subsystem ushered in a new era in our ability to process data rapidly and efficiently. The FICON protocol is significantly different than its predecessor ESCON protocol. And as a result of two main changes that FICON made to the mainframe channel I/O infrastructure, the requirements for a new RMF record came into being. The first significant change was that unlike FICON, ESCON did not use buffer credits to account for packet delivery. The second significant change was the introduction of "FICON cascading" which was not possible with ESCON.

While a fair amount of information is readily available, buffer to buffer credits (BB_credits) and their management in a FICON environment still appears to be one of the most commonly misunderstood concepts today. And truth be told, the phrase "buffer to buffer credit management" appears to be an oxymoron. Given their impact on performance over distances in cascaded FICON environments, this is something that needs to be better addressed. At present, there is no real way to manage/track BB_credits being used. At initial configuration a rule of thumb is used for allocating them and for management we keep our fingers crossed. Similar to how the end user manages dynamic PAVs by completely overkilling the number of aliases assigned to a base address, the typical FICON shop completely overkills the number of BB_credits assigned for long distance traffic. Just as PAV overkill can lead to configuration issues due to addressing constraints, BB_credit overkill can lead to director configuration issues which often times require outages to resolve. Mechanisms for detecting BB_credit starvation in a FICON environment are extremely limited at best.

BB_credits: Back to Basics

To get a good basic understanding of BB_credits, a brief review of the concept of flow control is in order. It also is worthwhile to take a brief historical perspective and review the concept of the device information block (DIB) from the ESCON architecture. For a more detailed discussion, the author recommends consulting Robert Kembel's The Fibre Channel Consultant.

ESCON DIBs-A brief historical perspective

The device information block (DIB) contains control, status or user data. The maximum DIB size to be employed by a configuration is defined by the storage director's (control unit's) ESCON adapters. DIBs may be 8, 16, 32, 64,128, 256, 512 or 1024 bytes long. During initialization (defined as a subsystem or path being varied online or processor IML time) a subsystem's ESCON adapters notify the channel subsystem of the DIB and data buffer size that is to be employed. The channel subsystem then employs these values for all future communications with the subsystem.

An ESCON adapter is a specialized hardware interface which manages communication between the channel subsystem and the control unit. While the ESCON adapter is a part of the control unit, we need to consider its functions distinct from those performed by the channel and the control unit during the transmission of data. **Note: insert ESCON block adapter diagram here.**

The ESCON socket connects to the 8b/10b converter, which in turn is connected to a microprocessor which manages the connection and schedules DIBs of data to and from the underlying storage director. This microprocessor also manages the data buffer which is incorporated in the ESCON adapter. The 3990-3/6 ESCON adapter's data buffer held eight 256 byte DIBs, i.e. 2048 bytes of storage. The DIB and data buffer sizes may be selected by the storage subsystem manufacturer within the constraints imposed by the ESCON architecture. Assuming that the ESCON adapter is receiving data, it must successfully pass each DIB it receives to buffer areas in the underlying storage director before it can request another DIB of data to be sent by the channel to avoid the potential of data overrun. While the channel may initially schedule a full data buffer of DIBs to be sent to the device, further DIBs may only be transmitted after a request has been received from the ESCON adapter signaling that a prior DIB has successfully been passed to the storage director.

Packet Flow and Credits

The fundamental objective of flow control is to prevent a transmitter from over-running a receiver by allowing the receiver to pace the transmitter, managing each I/O as a unique instance. At extended distances, pacing signal delays can result in degraded performance. Buffer-to-buffer credit flow control is employed to transmit frames from the transmitter to the receiver and pacing signals back from the receiver to the transmitter. The basic information carrier in the fibre channel protocol is the frame. Other than ordered sets, which are used for communication of low-level link conditions, all information is contained within the frames. When discussing the concept of frames, a good analogy to use is that of an envelope: when sending a letter via the United States Postal Service, the letter is encapsulated within an envelope and when sending data via a FICON network, the data is encapsulated within a frame. Thankfully, service times in a FICON network are better than those of the USPS.

To prevent a target device (either host or storage) from being sent more frames than it has buffer memory to store (overrun), the fibre channel architecture provides a flow control mechanism based on a system of credits. Each credit represents the ability of the receiver to accept a frame. Simply stated, a transmitter cannot send more frames to a receiver than the receiver can store in its buffer memory. Once the transmitter exhausts the frame count of the receiver, it must wait for the receiver to credit-back frames to the transmitter. A good analogy would be a pre-paid calling card: there are a certain amount of minutes to use, and one can talk until there is no more time (minutes) on the card.

Flow control exists at both the physical and logical level. The physical level is called buffer-to-buffer flow control and manages the flow of frames between transmitters and receivers. The logical level is called end-to-end flow control and it manages the flow of a logical operation between two end nodes. It is important to note that a single end-to-end operation may have made multiple transmitter-to-receiver pair hops (end-to-end frame transmissions) to reach its destination. However, the presence of intervening directors and/or ISLs is transparent to end-to-end flow control. Since buffer-to-buffer flow control is the more crucial subject in a cascaded FICON environment, the following section provides a more in-depth discussion.

Buffer-to-Buffer Flow Control

Buffer-to-buffer flow control is flow control between two optically adjacent ports in the I/O path (i.e. transmission control over individual network links). Each fibre channel port has dedicated sets of hardware buffers for send and receive operations. These buffers are more commonly known as buffer-to-buffer credits (BB_credits).

The number of available BB_credits defines the maximum amount of data that can be transmitted prior to an acknowledgment from the receiver. BB_credits are physical memory resources incorporated in the Application Specific Integrated Circuit (ASIC) that manages the port. It is important to note that these memory resources are limited. Moreover, the cost of the ASICs increases as a function of the size of the memory resource. One important aspect of fibre channel is that adjacent nodes do not have to have the same number of credits. Rather,

adjacent ports communicate with each other during fabric login (FLOGI) and port login (PLOGI) to determine the number of credits available for the send and receive ports on each node.

A BB_credit can transport a 2,112 byte frame of data. The FICON FC-SB-2 and FC-SB-3 ULPs employ 64 bytes of this frame for addressing and control, leaving 2K available for z/OS data. In the event that a 2 Gbit transmitter is sending full 2,112 byte frames, one credit is required for every 1 KM of fibre between the sender and receiver. Unfortunately, z/OS disk workloads rarely produce full credits. For a 4K transfer, the average frame size is 819 bytes, therefore five credits would be required per KM of distance as a result of the decreased average frame size. It is important to note that increasing the fibre speed increases the number of credits required to support a given distance. In other words, every time the distance doubles, the number of required BB_credits doubles to avoid transmission delays for a specified distance.

Buffer-to-buffer credits are used by Class 2 and Class 3 service and rely on the receiver sending back receiver-readies (R_RDY) to the transmitter. As was previously discussed, node pairs communicate their number of credits available during FLOGI/PLOGI. This value is used by the transmitter to track the consumption of receive buffers and pace transmissions if necessary. FICON directors track the available BB_credits in the following manner:

- before any data frames are sent, the transmitter sets a counter equal to the BB_credit value communicated by its receiver during FLOGI
- 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 is ready to receive another data frame
- for each R_RDY received by the transmitter, the counter is incremented by one

As long as the transmitter count is a non-zero value, the transmitter is free to continue sending data. This mechanism allows for the transmitter to have a maximum number of data frames in transit equal to the value of BB_credit, with an inspection of the transmitter counter indicating the number of receive buffers. The flow of frame transmission between adjacent ports is regulated by the receiving port's presentation of R_RDYs. In other words, BB_credits has no end to end component. The sender decrements the BB_credit by 1 for each R_RDY received. The initial value of BB_credit must be non-zero. The rate of frame transmission is regulated by the receiving port based on the availability of buffers to hold received frames. It should be noted that the FC-FS specification allows the transmitter to be initialized at zero, or at the value BB_credit and either count up or down on frame transmit. Different switch/director vendors may handle this with either method, and the counting would be handled accordingly.

Implications to Asset Deployment

There are four implications to asset deployment to consider when planning BB_credit allocations:

1. For write intensive applications across an ISL (tape and disk replication) the BB_credit value advertised by the E_Port on the target side gates performance. In other words, the number of BB_credits on the target side cascaded FICON director is the major factor.
2. For read intensive applications across an ISL (regular transactions) the BB_Credit value advertised by the E_Port on the host side gates performance. In other words, the number of BB_credits at the local location is the major factor.
3. Two ports do not negotiate BB_credit down to the lowest common value. A receiver simply "advertises" BB_credits to a linked transmitter.
4. The depletion of BB_credits at any point between an initiator and a target will gate overall throughput.

Configuring BB_credit allocations on FICON directors

There have been two FICON director/switch architectures when it comes to BB_credit allocation. The first, which was prevalent on early FICON directors such as the Inrange/CNT FC9000 and McDATA 6064 had a range of BB_credits that could be assigned to each individual port. Each port on a port card had a range of BB_credits (for example 4-120) that could be assigned to it during the director configuration process. Simple rules of thumb on a table/matrix were used to determine the number of BB_credits to use. Unfortunately, these tables did not

consider workload characteristics, or z/OS particulars. Since changing the BB_credit allocation was an off-line operation, most installations would figure out what they needed, set it and (assuming it was correct) be done with it. Best practice was typically to max out BB_credits used on ports being used for distance traffic since each port could theoretically be set to the maximum available BB_credits without penalizing other ports on the port card. Some installations would even max out the BB_credit allocation on short distance ports "so they would not have to worry about it". However, this could cause other kinds of problems in recovery scenarios.

The second FICON director/switch architecture has a pool of available BB_credits for each port card in the director. This is the architecture that is available in the market today in products from Brocade and Cisco. Each port on the port card will have a maximum setting, however since there is a large pool of BB_credits that must be shared amongst all ports on a port card, better allocation planning must take place that what an installation could do in the past. It is no longer enough to simply use distance rules of thumb, workload characteristics of traffic need to be better understood. Also, as 4 Gbps FICON Express 4 becomes prevalent, and 8 Gbps FICON Express 8 follows, intra-datacenter distances become something that must be looked at when deciding how to allocate the pool of available BB_credits. It no longer is enough to simply say that a port is internal to the data center or campus and assign it the minimum number of credits. This pooled architecture and the careful capacity planning it necessitates make it more critical than ever to have a way to track actual BB_credit usage in a cascaded FICON environment. Simply employing a "fire and forget" approach to BB_credit allocation is no longer optimal for ensuring performance over distance.

Before proceeding further with that idea, let's discuss what happens when you exhaust available BB_credits and the concept of frame pacing delay.

Exhaustion of BB_credits and Frame Pacing Delay

Similar to the ESCON directors that preceded them, FICON directors and switches have a feature called Control Unit Port (CUP). Among the many functions of the CUP feature is an ability to provide host control functions such as blocking and unblocking ports, safe switching and in-band host communication functions such as port monitoring and error reporting. Enabling CUP on FICON directors while also enabling RMF 74 subtype 7 (RMF 74-7) records for your z/OS system yields a new RMF report called the "FICON Director Activity Report". Data is collected for each RMF interval if FCD is specified in your *ERBRMFnn* parmlib member. RMF will format one of these reports per interval per each FICON director that has CUP enabled and the parmlib specified. This RMF report is often overlooked but contains very meaningful data concerning FICON I/O performance—in particular, frame pacing delay. **It is extremely important to note that indications of frame pacing delay are the only indication available to indicate a BB_credit starvation issue on a given port.**

Frame pacing delay has been around since fibre channel SAN was first implemented in the late 1990s by our open systems friends. But until the use of cascaded FICON increased, its relevance in the mainframe space has been completely overlooked. If frame pacing delay is occurring then the buffer credits have reached zero on a port for an interval of 2.5 microseconds and no more data can be transmitted until a credit has been added back to the buffer credit pool for that port. Frame pacing delay causes unpredictable performance delays, and those delays generally result in elongated FICON connect time and/or elongated PEND times that show up on the volumes attached to those links. **It is important to note that only when using switched FICON and only when CUP is enabled on the FICON switching device(s) can RMF provide the report that provides frame pacing delay information. Only the RFM 74-7 FICON Director Activity Report provides FICON frame pacing delay information. You cannot get this information from any other source today.**

```
                         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 V1R6                    SYSTEM ID SC64           DATE 10/06/2004      INTERVAL 10.00.001
                                    RPT VERSION V1R5 RMF     TIME 09.10.00        CYCLE 1.000 SECONDS
 IODF = 58   CR-DATE: 09/23/2004   CR-TIME: 15.35.18    ACT: ACTIVATE
 SWITCH DEVICE: 0061   SWITCH ID: 61     TYPE: 006064   MODEL: 001   MAN: MCD   PLANT: 01   SERIAL: 0000000119D3
 PORT    -CONNECTION-   AVG FRAME    AVG FRAME SIZE     PORT BANDWIDTH (MB/SEC)     ERROR
 ADDR    UNIT    ID     PACING       READ    WRITE      -- READ --   -- WRITE --    COUNT
  04     SWITCH  ----        0        579     889           0.04         0.03          0
  05     CHP     5A          0         71     238           0.07         0.21          0
  06     CHP     80          0         68     175           0.07         0.16          0
  07     CU      ----        0          0       0           0.00         0.00          0
  08     CU      ----        0        886      73           0.03         0.00          0
  09     CHP     5C          0        171     129           0.17         0.15          0
  0A     CHP     81          0        165      85           0.13         0.08          0
  0B     ------  ----    P O R T   O F F L I N E
  0C     CU      ----        0        829      86           0.05         0.00          0
  0D     CHP     5E          0         73     888           0.00         0.03          0
  0E     CHP     82          0        112     720           0.00         0.02          0
  0F     ------  ----    P O R T   O F F L I N E
  10     CU      ----        0        826      89           0.05         0.00          0
  11     CHP     60          0          0       0           0.00         0.00          0
```

Figure 1-Sample FICON Director Activity Report (RMF 74-7)

The fourth column from the left in figure 1 is the column where frame pacing delay is reported. Any number other than zero in this column is an indication of frame pacing delay occurring. If there is a non-zero number it reflects the number of times that I/O was delayed for 2.5 microseconds or longer due to buffer credits falling to zero. Figure 1 shows what you would always hope to see, zeros down the entire column indicating that enough buffer credits are always available to transfer FICON frames.

Figure 2-Frame Pacing Delay Indications in RMF 74-7 record

But in figure 2 above, you can see that on the FICON Director Activity Report for switch ID 6E, there were at least three instances when port 4, a cascaded link, suffered frame pacing delays during this RMF reporting interval. This would have resulted in unpredictable performance across this cascaded link during this period of time.

How does frame pacing delay occur?

There are two typical reasons that frame pacing delay occurs. The first is that a long distance link runs out of buffer credits, where possibly not enough buffer credits were specified to keep the link fully utilized. This can also happen even when maximum buffer credits are specified if the distance is too long for the speed of the link and the size of the frames being sent. In a local environment frame pacing delay can occur, for example, when a

FICON CHPID has issued open exchanges to a number of control units that are all responding back to the host simultaneously as shown in figure 3.
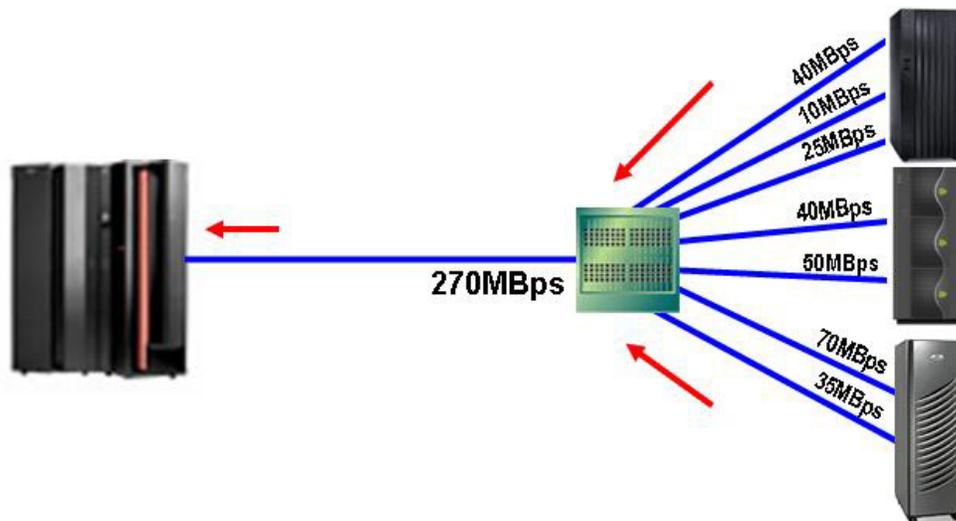

Figure 3 – Frame Pacing Delay from storage to host

In figure 3, the links between storage and the FICON Director are all running well within the limits of 2Gbps or 4Gbps links. But all of these links are attempting to converge into a single 2Gbps link that runs back to the host CHPID from the FICON Director.

From the FICON Director to storage there is no performance problem, but the 2Gbps link between the FICON Director and the mainframe has become overwhelmed. At its very best this link should operate at no more than about 150MBps and is simply unable to handle a 270MBps unidirectional frame flow on a 2Gbps link.

The port in the FICON Director for this CHPID connection quickly runs out of buffer credits. This signals all of the storage links converging to this FICON Director port to suspend I/O activity until buffer credits have been refreshed on that port.

What is the difference between frame pacing and frame latency?

Frame Pacing is an FC4 application data exchange level measurement and/or throttling mechanism. It uses buffer credits to provide a flow control mechanism for FICON to assure delivery of data across the FICON fabric. When all buffer credits for a port are exhausted, a frame pacing delay can occur. Frame Latency, on the other hand, is a frame delivery level measurement that is somewhat akin to measuring frame friction. Each element that handles the frame contributes to this latency measurement (CHPID port, switch/Director, storage port adapter, link distance, etc.). Frame latency is the average amount of time it takes to deliver a frame from the source port to the destination port.

What can you do to eliminate or circumvent Frame Pacing Delay?

If it is a long distance link that is running out of buffer credits, then it might be possible to enable additional buffer credits for that link in an attempt to provide an adequate pool of buffer credits for the frames being delivered over that link. But you might be surprised at how many buffer credits are required to handle specific workloads across distance. See figure 4.

<------Frame------>		Buffer Credits Required to go 50 KM				
Payload %	Payload Bytes	1Gbps	2Gbps	4Gbps	8Gbps	10Gbps
100%	2112	25	49	98	196	290
75%	1584	33	65	130	259	383
50%	1056	48	96	191	381	563
25%	528	91	181	362	723	1069
10%	211	197	393	785	1569	2318
5%	106	321	641	1281	2561	3784
1%	21	656	1312	2624	5248	7755

Figure 4 – Frame size, link speed and distance determine buffer credit requirements

Keep in mind that tape workloads will generally have larger payloads in a FICON frame while DASD workloads might have much smaller frame payloads. Some say the average payload size for DASD is often about 800-1500 bytes. By using the FICON Director Activity reports for your enterprise, you can gain an understanding of your own average read and write frames sizes on a port by port basis.

To help you, columns five and six of figure 1 show the average read frame size and the average write frame size for the frame traffic on each and every port. These columns come in handy when you are trying to figure out how many buffer credits will be needed for a long distance link or possibly to solve a local frame pacing delay issue. But if the problem is as shown in figure 3, then your only real alternative is to reconfigure your link connections or change the way in which you access the data on the CHPID link that is being overwhelmed. Adding additional buffer credits will not solve the problem.

What is the worst performance problem than can occur as a result of frame pacing delay?

The worst form of frame pacing delay is when it occurs on one or more cascaded links as shown in figure 5.
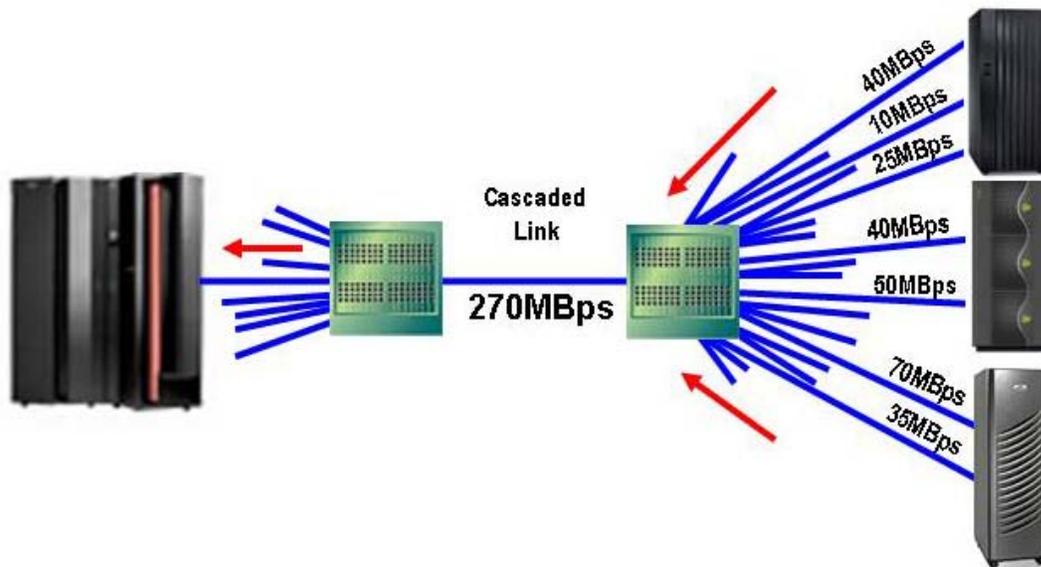

Figure 5 – Frame Pacing Delay on a Cascaded Link

Now not only are the original seven storage links affected by unpredictable performance blocking but so are the entire set of host and storage links that traverse this cascaded link regardless of whether it is a local cascaded environment or a long distance environment. RMF will undoubtedly show elongated connect time and/or

additional PEND time on the storage links. In a different scenario, if the buffer credits were to be exhausted going from the host to storage across the cascaded link(s) then possibly IOSQ time on some of the CHPID links would accrue due to frame pacing delay.

How can things be improved?

It would appear that even with the new FICON directors and the ability to assign BB_credits to each port from a pool of available credits on each port card, that the end user is still stuck. The end user can best hope they allocate correctly, and then monitor their RMF 74-7 report for indications of frame pacing delay to indicate they are out of BB_credits. They can then go ahead and make the necessary adjustments to their BB_credit allocations for crucial ports such as the ISL ports on either end of a cascaded link. However, any adjustments made will merely be a better guestimate since the exact number being used is not indicated. Imagine driving a car without a fuel gauge, and having to rely on EPA miles per gallon estimates so you could calculate how many miles you could drive on a full tank of gas. Of course, this estimate would not reflect driving characteristics. And in the end, the only accurate indication you get that you are out of gas is a coughing engine that stops running.

Why do we not yet have the capability, either in RMF or in the FICON director management software, to have BB_credit usage actively monitored and counted? Earlier in this paper, the fact that the individual ports track BB_credit availability was discussed and the mechanism by which this occurs was described. So, it would appear to be a matter of creating a reporting mechanism and what we have then is a situation similar to what we have with monitoring open exchanges. In 2004 Dr. H. Pat Artis wrote a paper that discussed open exchanges and made a sound case for why open exchange management is crucial in a FICON environment. Dr. Artis proved the correlation between response/service time skyrocketing and open exchange saturation, demonstrated how channel busy and bus busy metrics are not correlated to response/service time and recommended a range of open exchanges to use for managing a FICON environment. Since RMF does not report open exchange counts, Dr. Artis derived a formula using z/OS response time metrics to calculate open exchanges. Commercial software such as MXG and RMF Magic use this to help users better manage their FICON environments.

Similar to open exchanges, the data needed to calculate BB_credit usage is currently available in RMF. All that would be needed is some mathematical calculations be performed. **The author respectfully submits that the RMF 74-7 record (FICON Director Activity Report) should be updated with these 2 additional fields and the appropriate interfaces be added between the FICON directors and CUP code. Director management software could also be enhanced to include these two valuable metrics.**

Dynamic Allocation of BB_credits

The techniques used in BB_credit allocation is very similar in concept to the early technique used in managing parallel access volume (PAVs) aliases. The simple approach used was called static assignment. With static assignment, the storage subsystem utility was used to statically assign alias addresses to base addresses. While a generous static assignment policy could help to ensure sufficient performance for a base address, it resulted in ineffective utilization of the alias addresses (since nobody knew what the optimal number of aliases was for a given base), and put pressure on the 64K device address limit. Users tended to assign an equal number of addresses to each base, often taking a very conservative approach resulting in PAV alias overkill. That situation is very similar to what we currently have with BB_credit allocation, in particular with older FICON directors.

An effort to address this was undertaken by IBM, resulting in IBM providing workload manager (WLM) support for dynamic alias assignment. Here, WLM was allowed to dynamically reassign aliases from a pool to base addresses to meet workload goals. However, this could be somewhat lethargic, so users of dynamic PAVs still tend to over configure aliases and are pushing the 64K device address limitation. Users face what Dr. Artis refers to as the PAV performance paradox: they need the performance of PAVs, tend to over configure alias addresses, and are close to exhausting the z/OS device addressing limit.

The author respectfully submits that a similar dynamic allocation of BB_credits, in particular for new FICON director architectures having pools of assignable credits on each port card would be a very beneficial enhancement for end users. Perhaps an interface between the FICON directors and WLM could be developed to allow WLM to dynamically assign BB_ credits. At the same time, since quality of service (QOS) is an emerging topic of importance for FICON, perhaps an interface could be developed between the FICON directors and WLM for functionality with dynamic channel path management and priority I/O queuing to enable true end to end QOS.

In October 2006, IBM announced HyperPAVs for the DS8000 storage subsystem family to address the PAV performance paradox; HyperPAVs increase the agility of the alias assignment algorithm. In a nutshell, the primary difference between the traditional PAV alias management is that aliases are dynamically assigned to individual I/Os by the z/OS I/O supervisor (IOS) rather than being statically or dynamically assigned to a base address by WLM. The RMF 78-3 (I/O queuing) record has also been expanded. If a similar feature/functionality and interface could be developed between FICON directors and the z/OS IOS, we then would have the ultimate in BB_credit allocation: true dynamic allocation of BB_credits on an individual I/O basis.

Conclusion

This paper has reviewed flow control, basics of buffer to buffer credit theory, basics of frame pacing delay, current buffer to buffer credit allocations methods and presented some proposals for a) counting BB_credit usage and b) enhancing how BB_credits are allocated and managed. Current methods of blindly allocating credits, and discovering after the fact via an obscure report that you don't have enough are not sufficient. BB_credit management is an oxymoron in 2007 - it does not have to be, nor should it be that way.