BROCADE

WHITE PAPER

Fibre Channel Buffer Credits and Frame Management

Introduction

The basic information unit in Fibre Channel (FC) is the frame. To prevent a target device (hosts or storage) from being overwhelmed with frames, the FC architecture provides flow control mechanisms based on a system of credits (buffer credits). Each of these credits represents the availability of the device to accept additional frames (buffer credits). If a recipient issues no credit to the sender, no frames can be sent to that recipient.

This type of flow control based on credits prevents the loss of frames and reduces the frequency of entire FC sequences requiring retransmission. Having sufficient buffer credits is essential to maintaining maximum Input/Output (I/O) performance in a storage network, especially as distance increases.

All of this seems simple enough, however, in the background, buffer credits are also leveraged in various ways to help minimize I/O latency and maximize I/O performance.

The information provided by Brocade in this paper helps clarify your understanding and provide ways you can realize the benefits of properly configured buffer credits in an FC I/O environment.

Table of Contents

Fibre Channel Buffer Credits and Frame Management

The basic unit of information of the Fibre Channel (FC) protocol is the frame. Other than primitives, which are used for lower-level communication and monitoring of link conditions, all information is contained within the FC 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. When sending data via an FC network, the data is encapsulated within a frame. Managing the data transfer of a frame between sender and receiver is of vital importance to the integrity of a storage network. Buffer credits are the principal mechanism used to control point-to-point frame transmission and reception.

Flow control exists at both the physical and logical level. The physical level is called buffer-to-buffer flow control, and it manages the flow of frames between transmitters and receivers basically, from one end of an FC cable to the other end. The logical level is called end-to-end flow control, and it manages the flow of a logical operation (not frames) between two end nodes. It is important to note that a single end-to-end operation may make multiple transmitter-to-receiver pair hops (end-to-end frame transmissions) to reach its destination. However, the presence of intervening switching devices and Inter-Switch Links (ISLs) is transparent to end-to-end flow control (see Figure 1). Since buffer-to-buffer flow control is the more crucial subject in an environment using ISLs, the following section provides a more in-depth discussion.

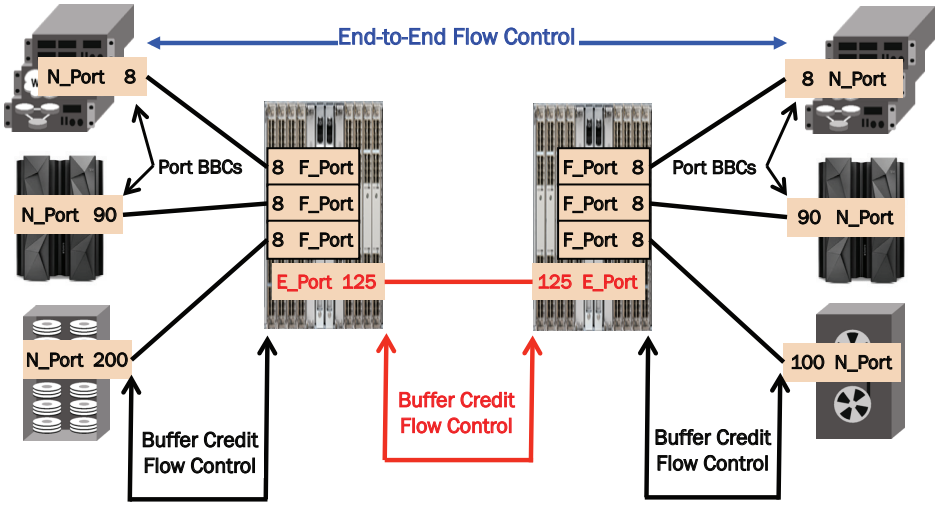


Figure 1. End-to-end flow control.

Upon arrival at a receiver, a frame goes through several steps, and its header becomes stored in the receive buffers, where it is processed by the receiving port. Frames arriving while the receiver is busy processing an earlier received frame are stored in the next free receive buffer. If the receiver is not capable of processing the frames in a timely fashion, it is possible for all of the receive buffers to fill up with received frames. At this point, if the transmitter continued to send frames, the receiver would not have receive buffer memory space available, and the frame would be lost (discarded or dropped). Losing data by

discarding it, or attempting to overwrite a buffer that is already in use, is not acceptable and cannot be allowed in a storage environment. This is why flow control is so necessary for deterministic high-performance low-latency networks.

Buffer credit flow control is implemented to limit the amount of data that a port may send, and, if properly tuned, is primarily based on four metrics: the speed of the link, the distance that the link traverses, whether the data is being compressed, and the average size of the frames being sent. Buffer credits also represent finite physical-port memory, which means that as more buffer credits are allocated on a port, more physical memory is consumed on that port to hold the buffer credits. Within a fabric, each port may have a different number of buffer credits although this is not recommended. Within a single link connection, each side of the link may have a different number of buffer credits. The configured buffer credit value on a port is not required to be symmetrical, but users should always make ISL links contain a symmetrical number of buffer credits (see Figure 1).

Buffer-to-buffer flow control is an agreed upon and known control of buffers between adjacent ports within the I/O path, in other words, transmission control across each single FC cable link. A sending port uses its available credit supply and waits to have its credits replenished by the receiving port on the opposite end of the cabled link.

Buffer credits are used by Class 2 and Class 3 FC services. When host or storage node ports (N_Ports) are connected to switch fabric ports (F_Ports), a receiver ready "R_RDY" message is the standards-based acknowledgment from the receiver back to the transmitter that a frame was successfully received and that a buffer credit can be freed by the transmitter for its next frame to use. The rate of frame transmission is regulated by the receiving port and is based on the availability of buffers to hold received frames. In short, buffer-to-buffer flow control works because a sending port is using its available buffer credit supply and waiting to have the buffer credits replenished by the receiving port on the opposite end of the link.

Brocade provides very robust deployment, monitoring, and management of all of its FC resources in any Storage Area Network (SAN). Brocade® Network Advisor with Fabric Vision™ technology makes FC fabric management, including buffer credit management, much more robust and less complex.

Back to the Basics of Buffer Credits

One of the major strengths of Fibre Channel is that it creates lossless connections by implementing a flow control scheme based on buffer credits. Ports need memory—in other words, buffer credits—to temporarily store frames as they arrive and until they are assembled in sequence and delivered to the Upper Layer Protocol (ULP). The number of buffers—that is, the number of frames a port can store—is its buffer credit capability.

When an exchange (read or write) is to be transmitted down a Fibre Channel path, that record is split up into a number of frames, each of which can have a maximum size of 2,148 bytes. These data-carrying frames are transmitted through the fabric and then reassembled at the far end to recreate the complete data record. To make sure the FC links are used efficiently, every switch port has a number of buffers (that is, buffer credits)

associated with it. If necessary, the switch can then store several blocks of incoming data, while waiting to pass the data on to the next node. When a receiver is ready to accept information, it signals to its sender, then decrements its own BB_Credit (buffer credit) count. When the data block is passed on to the next receiver, the buffer credit is incremented again. This means that buffer credits are also used to throttle back the data transmission flow when devices or links get too busy.

As long as the transmitter's buffer credit count is a non-zero value, the transmitter is free to continue sending data. The zero credit counter increments when the last buffer credit is allocated. A buffer credit is allocated just prior to the transmission of a frame so, technically, the switch continues to transmit the last frame while at zero buffer credits. The flow of frame transmission between adjacent ports is regulated by the receiving port's presentation of acknowledgments. In other words, buffer credits have no end-to-end component, except for one end of a cable to the other end of that same cable. The sender increments its buffer credit count by 1 for each acknowledgment it receives. As stated earlier, the initial value of the buffer 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 specification allows the transmitter to be initialized at zero, or at the value BB_Credit, and to either count up or down on frame transmit. Various switch vendors may handle this with either method, and the counting is handled accordingly.

Buffer credit management affects performance over distances; therefore, allocating a sufficient number of buffer credits for long-distance traffic is essential to performance.

During initial login, each optical receiver informs the optical transmitter at the other end of the link of how many receive buffers it has available. This is called the BB_Credit value. This value is used by the transmitter to track the consumption of receive buffers and pace transmissions, if necessary.

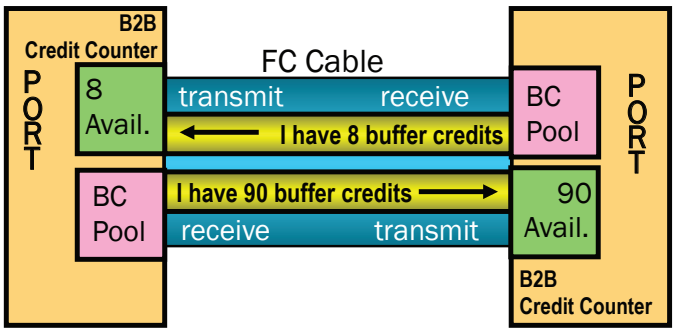


Figure 2. Buffer credit information exchange.

The terms "frame pacing" and "flow control" are often used interchangeably to describe how buffer credits provide link management. Regardless of the term that is used, the fundamental objective is to prevent the transmitter (Tx) from over-running its receiver (Rx) with an I/O workload, which causes frames to be discarded and causes potential

data integrity issues. In the process that prevents data overrun, the receiver side of the link informs the transmitter side of that link of how many frames it (the receiver) can accept and then handle correctly.

At this point, the transmitter has a process by which it sets its buffer credit counter and storage to correspond to the receiver's request. It then adheres to how many frames the receiver allows to be sent. Simply stated, a transmitter cannot send more frames to a receiver than the receiver can store in its buffer memory.

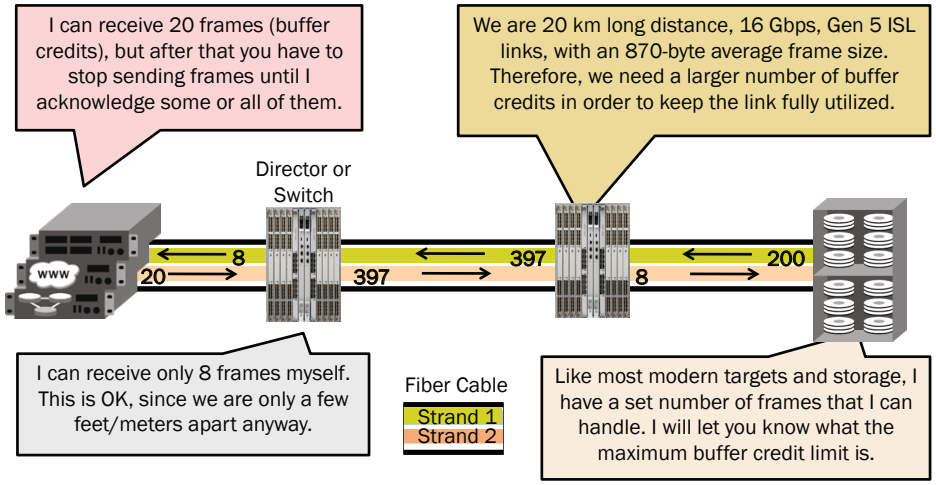


Figure 3. Buffer credit negotiation.

Each of these buffer credits represents the ability of the device to accept additional frames. If a recipient issues no credits to the sender, the sender does not send any frames. Pacing the transport of subsequent frames on the basis of this credit system helps prevent the loss of frames and reduces the frequency of entire FC sequences that need to be retransmitted across the link.

Thus, the transmitter decrements its counter based upon a logical usage of receive buffers and increments the counter based upon physical signaling from the receiver. In this way, the connection allows the receiver to "pace" the transmitter, managing each I/O as a unique instance. Pacing, as presented here, means that a frame cannot be sent from an egress port transmitter to an ingress port receiver, unless the transmitter has a buffer credit available to assign to the frame. Once a frame has been successfully accepted into the receiver, the receiver acknowledges that receipt back to the transmitter, so that the transmitter can increment its buffer credit counter by one. It then releases the buffer credit memory that it assigned to that successfully transmitted frame.

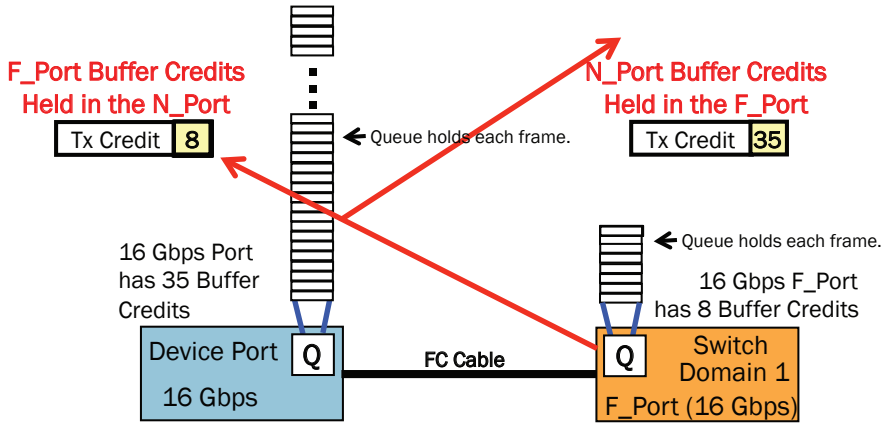


Figure 4. Buffer frame queue.

As long as the transmitter counter is non-zero, then the transmitter is free to send 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 available receive buffers.

If the transmitter exhausts the frame count offered by the receiver, it stops and then waits for the receiver to credit-back frames to the transmitter through frame acknowledgments. A good analogy is a prepaid calling card: a certain number of minutes are available to use, and you can talk until no more time remains on the card, at which point the conversation stops. In FC, you might often see this condition with slow drain devices.

A buffer credit equals one frame, regardless of frame size. Every buffer credit on a port (host, switch, storage) has the same maximum size: 2,148 bytes. However, every frame that is created can be a different size, while still consuming a full buffer credit.

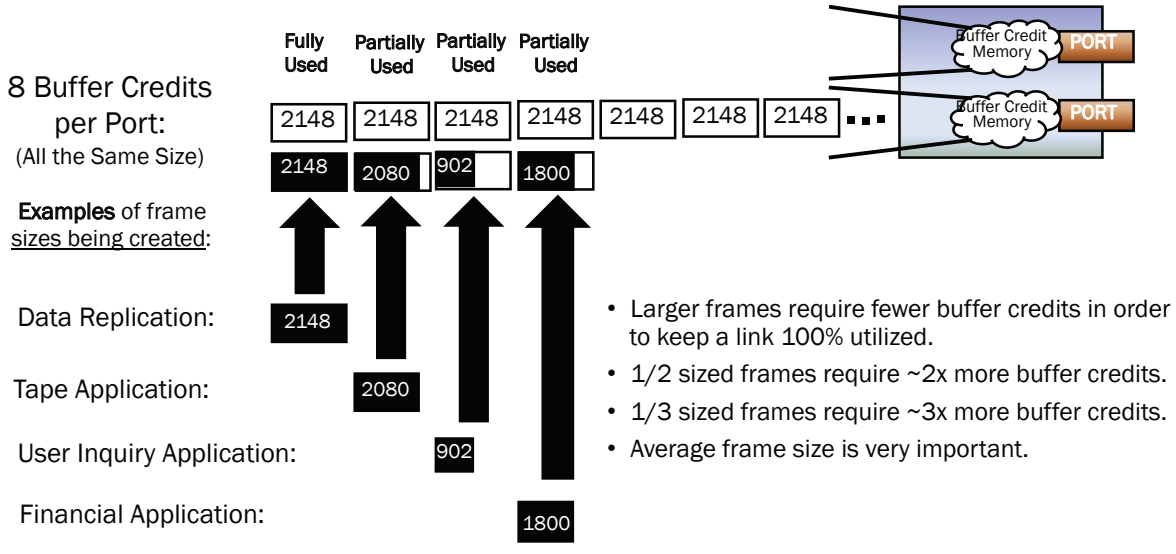


Figure 5. Effect of frame size on required buffer credits.

The round-trip link distance is also very important to understand in order to keep a link fully utilized. Although not technically accurate, you might imagine needing one-half of the buffer credits to maintain transmitter activity during data frame delivery, and the other half of the buffer credits to maintain transmitter activity while acknowledgment frames traverse the return link. Users must configure sufficient buffer credits for the total round-trip link distance, in order to ensure enough buffer credits to keep a link optimally utilized.

One of the reasons that a transmitter can exhaust the frame count is if there are not enough buffer credits initially allocated to the port. Four major metrics must be accounted for in order to allocate the correct number of buffer credits link speed, round-trip cable distance, average frame size, and compression. If one or more of these are misjudged, then too many or too few buffer credits are allocated on the link. If a transmitter runs out of buffer credits, it cannot send any more frames that it might have waiting. Also, those subsequent frames must be held in the egress port until buffer credits do become available. (Actually, the process is more complex, but this explanation suffices for now). In this case, it is not possible to optimize the full bandwidth potential of the link.

However, when ample buffer credits are allocated on the link for the full round-trip connection, the complete bandwidth potential of the link can be realized. If you could see a frame as it was being sent across the link, you would see that the slower the link speed, the more distance that a frame utilizes on that link, and that it takes fewer buffer credits to fully utilize the bandwidth. This explains the axiom that when the link speed doubles, you must also at least double the number of buffer credits allocated to that link.

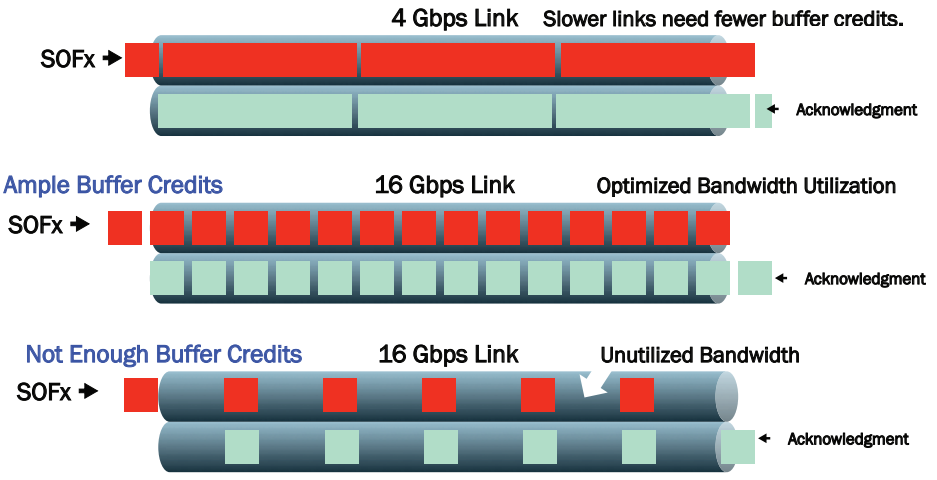


Figure 6. An example of required buffer credits.

Buffer Credits as Integral to the ASIC Design in a Brocade Switching Device

The number of available buffer credits defines the maximum amount of data that can be transmitted prior to an acknowledgment from the receiver. Buffer credits are physical memory resources that are 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 FC is that adjacent nodes are not required to have the same number of buffer 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.

Brocade uses an intelligent ASIC architecture, called Condor3, and Gen 5 connectivity port blades have two ASICs. On some Brocade fixed switches (for example, the Brocade 6505 Switch and the Brocade 6510 Switch) there is a single ASIC, for up to 48 ports of connectivity. The Brocade 6520 Switch is architected using six Condor3 ASICs. Each Condor3 ASIC holds a pool of buffer credits that their attached ports use (see Figure 7). To be clear, each port on an ASIC makes use of the buffer credit pool rather than owning its own dedicated buffer credits. There are some dedicated buffers per port, but how they are used is proprietary information.

On each Condor3 ASIC there are 8,192 buffer credits per 16-port group when using 32-port blades, and per 24-port group when using 48-port blades. From its ASIC buffer credit pool, each port attached to that ASIC is provided with a default setting of 8 buffer credits for external data flow. Then, to provide optimal data flow across long-distance links, any one additional port attached to an ASIC on a 32-port blade can have a maximum of 5,408 buffer credits configured onto that port. Any one additional port on an ASIC on a 48-port blade can have a maximum of 4,960 buffer credits configured onto that port. Of course, other allocations of buffer credits to ports are allowable. You can find more information in the Brocade Fabric OS® Administrator's Guide in the section on "Buffer credits per switch or blade model."

Table 1. Unreserved buffers.

Switch or Blade Model	Total FC Ports Per Switch or Blade	User Port Group Size	Unsreserved Buffers with QoS (per port group)	Unreserved Buffers without QOS (per port group)
Brocade 6510 Switch	48	48	6752	7712
FC 16-32	32	16	5188	5409
FC16-48	48	24	4480	4960
FC8-64 and FC 16-64	***Extended Fabrics are not supported on these blades.***			

Using CLI Commands to Monitor Average Frame Size and Buffer Credits

Users need to periodically check to be sure that an adequate number of buffer credits are assigned to each ISL link, as workload characteristics change over time. Use the **portBufferShow** Command Line Interface (CLI) command to monitor the buffer credits on a long-distance link and to monitor the average frame size and average buffer usage for a given port.

The **portBufferShow** CLI command displays the buffer usage information for a port group or for all port groups in the switch. When portBufferShow is issued on switches utilizing trunking, they see that the trunking port group can also be specified using this command just by giving any port number in that port group. If no port is specified, then the buffer credit information for all of the port groups on the switch is displayed. You can find more information about all of the Brocade CLI commands in the Fabric OS Administrator's Guide, at www.brocade.com.

Report 1 shows the results of issuing the **portBufferShow** CLI command. In the example, you see two independent ISLs. Brocade trunking (ASIC-optimized trunking) is not in use.

The average frame size in bytes is shown in parentheses with the average buffer usage for packet transmission and reception. It is readily apparent that ports 48 and 49 were provisioned with far too many buffer credits (942) than were ever used while this statistical sample was being taken (240). This is a far greater number than the average buffer credits required (32 and 146) on these ISL links.

Report 1. Buffer credits per port

```
switch:admin> portbuffershow
User     Port     Lx     Max/Resv  Avg Buffer Usage & FrameSize    Buffer Needed    Link    Remaining
Port     Type    Mode    Buffers        Tx           Rx           Usage  Buffers   Distance  Buffers
----     ----    ----    -------    --------------------------- ------ -------   --------- ------
  48       E      LS       942      32(1162)      59( 176)         240     240        40km
  49       E      LS       942     146( 436)      72( 216)         240     240        40km
  50       F       -         8      - ( 132)      - ( 188)           8       -          -
  51       F       -         8      - (  - )      - (  - )           8       -          -
  52               -         8      - (  - )      - (  - )           0       -          -
  53       F       -         8       1(1152)       1(1860)           8       -          -
  54       F       -         8       1(1296)      - ( 424)           8       -          -
```

The Most Difficult Buffer Credit Metric: Average Frame Size

Assuming that you are setting up your ISL links for the first time, there is one fairly simple way to calculate the average frame size that you need for the BB_Credit calculation tools or BB_Credit calculation to be done manually. You can use the **portStats64show** CLI command.

These 64-bit counters are actually two 32-bit counters, and the lower one ("bottom_int") is the 32-bit counter that is used in **portStatsShow**. Each time it wraps, it increases the upper one ("top_int") by 1. It is also recommended that you provide the "–long" option on the command. The "–long" option adds the contents of the two counters together in order to provide a single result.

An example both without and with the "–long" option follows:

```
(Without) portstats64show 3/15
Stat64_wtx      0                   top_int : 4-byte words transmitted
2308091032      Bottom_int : 4-byte words transmitted
Stat64_wrx      39                  top_int : 4-byte words received
1398223743      Bottom_int : 4-byte words received
Stat64_ftx      0                   top_int : Frames transmitted
9567522         Bottom_int : Frames transmitted
Stat64_frx      39                  top_int : Frames received
745125912       Bottom_int : Frames received

(With) portstats64show 4/10 –long
Stat64_wtx      1044696810          top_int : 4-byte words transmitted
Stat64_wrx      576487365           top_int : 4-byte words received
Stat64_ftx      4285384             top_int : Frames transmitted [Tx]
Stat64_frx      8023909             top_int : Frames received [Rx]
```

When the portStats64show -long command is executed, you do not get a counter of bytes but rather a counter of 4-byte words. Luckily, fill words on the link do not count into this number, so it is still valid for your average frame size calculation. You just need to multiply the transmitted and received words by four to turn it into total bytes. It is the Transmitted (Tx) average frame size that you want to use in your calculations to determine how many buffer credits are required:

(1044696810* 4) bytes = 4,178,787,240 / 4285384 frames = 976 (rounded up) Tx bytes average frame size

(576487365* 4) bytes = 2,305,949,460 / 8023909 frames = 288 (rounded up) Rx bytes average frame size

Port statistics are relevant only if the user knows how long the statistics have been accumulating on the port. A Power On Reset (POR) or the last time a portStatsClear command was issued (minutes, days, months?) will refresh port statistics. To make sure that the port's statistical information is relevant to peak usage or quarter-end/year-end usage, you should issue the portStatsClear CLI command before the statistics capture needs to be gathered and utilized.

Calculating Buffer Credits for an ISL Link

Although every Fibre Channel port must use buffer credits, it is really the longer-distance links that require a user to provision anything other than the default eight buffer credits per port on a switch. When a host or storage node port (Nx_Port) has no buffer credits available and has exceeded the link timeout period (Error Detect Time Out Value or E_D_TOV), a link timeout will be detected. E_D_TOV by default is 2 seconds. When a link timeout is detected, the Nx_Port or switch fabric port (Fx_Port) begins the Link Reset (LR) protocol. A switch or device port that cannot transmit frames due to the lack of credits received from the destination port for 2 seconds uses the Link Reset to reset the credit amount on the link back to its original value.

It is important to differentiate the processes of link reset and link initialization. The link initialization process is caused by cable connectivity (physical link), server reboots, or port resets. However, link credit reset can be an indication of a buffer credit starvation issue that occurs as an after-effect of frame congestion or fabric back pressure conditions. In order to ensure that a fabric runs as effectively as possible, it is vitally important that long-distance links be provisioned with an adequate and optimal number of buffer credits. There are tools that a customer will normally use, such as the Buffer Credit Calculation spreadsheet that is provided by Brocade (https://community.brocade.com/t5/Storage-Networks/bg-p/StorageNetworks). Another tool is to use the portBufferCalc CLI command.

If you know the round-trip distance (in kilometers), the speed, and the average frame size for a given port, you can use the **portBufferCalc** CLI command to calculate the number of buffer credits required on that link. If you omit the distance, speed, or frame size, the command uses the currently configured values for that port. Given the buffer requirement and port speed, you can specify the same distance and frame size values when using the **portCfgLongDistance** CLI command to actually set the appropriate number of buffer credits onto an ISL link.

In order to maintain acceptable performance, at least one buffer credit is required for every 2 kilometers (km) of distance covered (this includes round trip). Brocade also recommends adding an additional six buffer credits, which is the number of buffer credits reserved for fabric services, multicast, and broadcast traffic in the Virtual Channels of an ISL link. (These are discussed later in this paper.) The extra six buffer credits is just a static number, but it is important in order to be able to achieve maximum ISL utilization. You also must account for the actual average frame size of the frames that are being transmitted on the link, rather than just assuming that all frames sent will be at the full frame size of 2,148 bytes.

You can use this formula to approximate the number of buffer credits that an ISL link requires:

- BB_Credits = (2*(link distance in km/frame length in km)) * (full frame size/average frame size) +6

Framelength (km)	= 4.00 km @	1 gigabit per second (Gbps)
	= 2.00 km @	2 Gbps
	= 1.00 km @	4 Gbps
	= 0.50 km @	8 Gbps
	= 0.33 km @	10 Gbps
	= 0.25 km @	16 Gbps

For two sites that are 20 km (1 kilometer = 0.621 miles) apart, the approximate (rounded up) number of buffer credits required for a link is as follows:

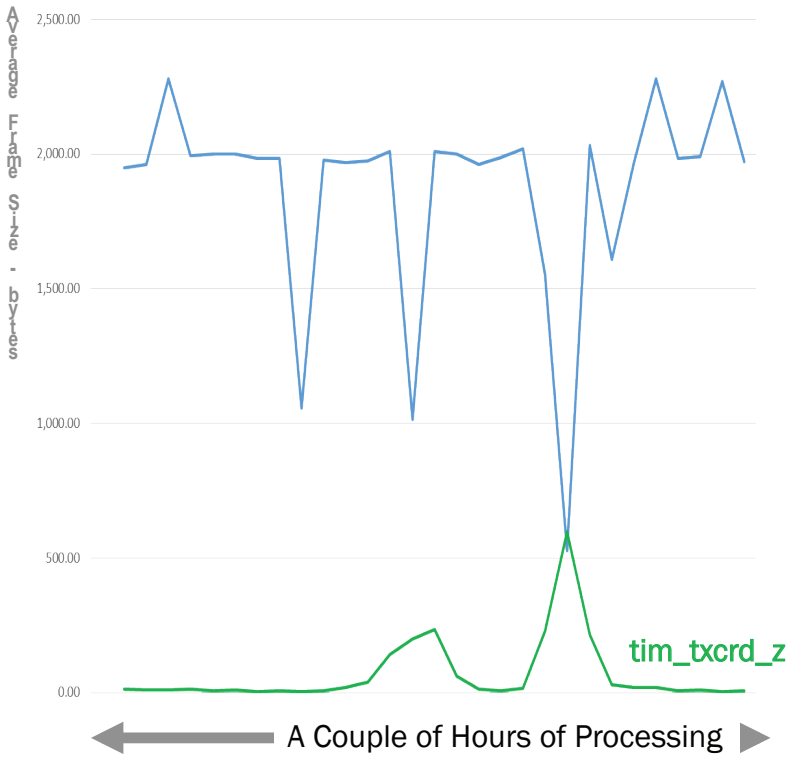
- On a 2 Gbps ISL Link: (2 * (20/2.0)) * (2148 / 870) + 6 = 56 (rounded up)
- On a 4 Gbps ISL Link: (2 * (20/1.0)) * (2148 / 870) + 6 = 105 (rounded up)
- On an 8 Gbps ISL Link: (2 * (20/.50)) * (2148 / 870) + 6 = 204 (rounded up)
- On a 10 Gbps ISL Link: (2 * (20/.33)) * (2148 / 870) + 6 = 307 (rounded up)
- On a 16 Gbps ISL Link: (2 * (20/.25)) * (2148 / 870) + 6 = 401 (rounded up)

If you want to graphically view how buffer credits and acknowledgments work, view examples starting at page 13 at this link: Buffer-to-Buffer Credits, Exchanges, and Urban Legends.

Transmit Buffer Credits Reaching zero – tim_txcrd_z

Time at zero credits (tim_txcrd_z) is generally considered to be the number of times that a port was unable to transmit frames it had in queue because the transmit BB_Credit counter was zero. But you must realize that the zero credit counter is incremented when the last buffer credit is allocated. A buffer credit is allocated just prior to frame transmission, so, technically, the switch continues to transmit the last frame while at zero buffer credits. In long-haul links this is insignificant, but in local data centers over short pieces of cable you need to look at link performance as well to understand whether a buffer credit problem exists or not. Theoretically, a short cable link can be at zero buffer credits 100 percent of the time, and the link will continue transmitting 100 percent of the time.

The purpose of this statistic is to detect congestion or a device affected by latency. This parameter is sampled on every port at intervals of 2.5 microseconds (µsec) (400,000 times per second), and the tim_txcrd_z counter is incremented if the condition is true. Each sample represents 2.5 µsec of time with zero transmitter BB_Credits and a frame waiting to be transmitted. An increment of this counter means that the frames could not be sent to the attached device for 2.5 µsec, indicating the potential of some level of degraded performance.



Report 2. Congestion detection. < Sample Report 2. >

In a high-performance environment, it is normal for buffer credits to reach zero at times, and buffer credits reaching zero does not mean there is a congestion problem. Optimization does not mean that the port can never run out of buffer credits, but rather that the port will not run out of buffer credits 99.999 percent of the time. That .0001 percent when it does run out of buffer credits will not harm performance or utilization on the link. Also, in mixed speed environments, where the switch is attached to 2 Gbps, 4 Gbps, 8 Gbps, and 16 Gbps devices, this counter typically increments at a higher rate on the lower speed devices than on the higher speed devices.

If, however, the buffer credits reaching zero is a very high value in regard to total frames sent, it could indicate a congestion problem. The absolute value is there in the counter (since the last POR or **statsClear** CLI command), but the relationship between number of buffer credits reaching zero and the total number of transmitted frames should be of interest. It is recommended that if the tim_txcrd_z counter shows a value around 10 percent or more of the total frames sent, you should take a real interest in it. A percentage much higher than 10 percent might indicate a fabric congestion problem. Keep in mind that smaller average frame sizes can contribute greatly to skewing the tim_txcrd_z counter. (See an example of this in Report 2.)

The tim_txcrd_z counter (that is, the **portStats64show** CLI command) should be used as an indicator for further investigation of buffer credit problems and not as proof of a performance problem. Also remember that port statistics are continuous until you take action to clear it. A **portStatsClear** CLI command should be executed before you attempt to use the tim_tmcrd_z, so that it can be known to be a relevant count.

Hold Time and Edge Hold Time

Fibre Channel is a network that has been architected to be deterministic, which requires devices to behave properly. By design, two link-end points tell each other how many buffers they have available to store an FC frame during link initialization in the FLOGI phase. When ISL links, hardware trunks, or operating devices in a fabric cannot return acknowledgments to replenish a transmitter's buffer credits as quickly as expected, the result can be that the transmitter runs out of its buffer credits (as discussed in the previous section). If that happens, then the transmitter must hold the frames it has in the port for longer than normal. A hold timer is the amount of time that a switch allows a frame to sit in a queue without being transmitted. When the timer expires, the frame is discarded. The timer feature that is used to mitigate ISL/trunk issues is called Hold Time (HT). A different timer feature that is used to mitigate slow draining devices is called Edge Hold Time. It makes sense to handle ISL congestion issues differently than individual devices, so these two timers have different values.

The HT value is the amount of time a Class 3 frame (for example, data frames and management frames) can remain in an ISL/trunk port queue before being dropped, while waiting for a buffer credit to be given for transmission of a frame.

Default HT is calculated from the Resource Allocation Time Out Value (RA_TOV), E_D_TOV, and maximum hop count values that are configured on a switch. At 10 seconds for RA_TOV, 2 seconds for E_D_TOV, and a maximum hop count of 7, an HT value of 500 milliseconds (ms) is calculated.

As previously discussed, time at zero credits (tim_txcrd_z) is sampled at intervals of 2.5 µsec, and the tim_txcrd_z counter is incremented if the condition is true. The HT value is 500 ms (or half of a second), at which point a frame must be discarded. How do these two facts interrelate?

• 2.5 µsec * 200,000 = 500 ms or one-half second

If an ISL/trunk port must wait 200,000 2.5 µsec intervals, due to not having a buffer credit available to send a waiting frame, then it will breach the HT mark (500 ms) for that port. Then that frame, and potentially others that are queued up behind it, must be discarded (C3 tx_timeout discard or C3TXT03).

EHT is for F_Ports and allows an overriding value for HT on devices attached to those F_Ports. EHT (220 ms) is ASIC-dependent and can discard blocked frames within an F_Port earlier than the 500 ms default hold time that is normally placed on ISL/trunk links. An I/O retry still happens for each dropped frame.

The EHT value can be used to reduce the likelihood of back pressure into the fabric by assigning a lower HT value only for edge ports (for example, host ports or device ports). The lower EHT value ensures that frames are dropped at the F_Port, where the buffer credit is lacking, before the higher default Hold Time value that is used at the ISLs expires. This action localizes the impact of a high-latency F_Port to just the single edge switch where the F_Port(s) are attached and prevents the back pressure and associated collateral damage from spreading into the fabric and affecting other unrelated flows.

When a user deploys Virtual Fabrics, the EHT value that is configured into the Base Switch is the value that is used for all Logical Switches.

Virtual Channels

Virtual Channels (VCs) are a unique feature available on 2 Gbps, 4 Gbps, 8 Gbps, and 16 Gbps Fabric Switches or director devices from Brocade.

In switched-FC networks, frames can be sent over very complex network architectures with guaranteed flow control and an assurance of lossless frame delivery. However, without an architecture that provides fair access to the ISL wire for all of the I/O traffic streams that use it, congestion could impose a condition where one traffic stream blocks other streams on that ISL link. This is generally referred to as Head-of-Line Blocking (HoLB). Furthermore, in a Brocade switched-FC architecture using Virtual Channels, if congestion does develop in the network because of a multiplexed ISL traffic architecture, all sending traffic slows down gradually in response to the congestion.

Virtual Channels create multiple logical data paths across a single physical link or connection. They are allocated their own network resources, such as queues and buffer credits. Virtual Channel technology is the fundamental building block used to construct Adaptive Networking services.

Virtual Channels are divided into three priority groups. Priority Group 1 (P1), VC 1, is described in many documents as the highest priority VC, which is used for Class F and ACK traffic. However, engineering has never enabled it on any of the generations of Condor ASICs. Priority Group 1 (P2) is the next highest priority, which is used for data

There is no Figure 8. What Figure are you referring to...Table 2 maybe?

frames and which provides for high, medium, and low Quality of Service (QoS) on VCs 2 through 5 and 8 thorough 14. Priority Group 3 (P3) is the lowest priority and is used for broadcast and multicast traffic. This example is illustrated in Figure 8.

Each VC is given the responsibility to maintain its own buffer credits and counts. A Brocade VC_RDY is then substituted for R_RDY as the acknowledgment. But even when ISL links are being used very well, HoLB on an ISL link can create congestion and performance problems, much like toll booths on a superhighway can cause congestion and slow travel.

VCs provide a way to create a multilane connection, so that even if some frames are stopped from flowing, other frames can bypass them and go to their destinations across a single link.

To ensure reliable ISL communications, VC technology logically partitions the physical buffer credits on each ISL link into many different buffer credit pools (one for each VC). Then, if QoS is in use, VC technology prioritizes traffic to optimize performance and prevent head of line blocking. This VC technology is not just about creating smaller pools of logical buffer credits from the physical set of buffer credits. In addition, each VC is provided with its own queues, and the VC technology manages those as well.

Setting up Buffer Credits on any ISL Link

Some switch-to-switch ISL connectivity might be local, while other switch-to-switch connectivity might be over longer distances. If more than two or three buffer credits per data VC are required, due to distance or average frame size, use the **portCfgLongDistance** command to deploy an adequate number of buffer credits for the environment. Brocade Fabric OS (Brocade FOS) offers several different distance modes. You can find complete information about the use of these modes in the most current *Fabric OS Administrator's Guide* at www.brocade.com.

L0 mode is the normal (default) mode for a port. It configures the port as a regular port. Use L0 mode for local distance, up to the rated distance for the link speed that is in use.

LE mode configures the distance for an E_Port when that distance is greater than 5 km and up to 10 km, but the calculations are based on the average frame size being a full frame size of 2,148 bytes (which is generally not the case).

If a link does not use L0 or LE mode, and it requires more than 20 buffer credits, then an optional Extended Fabrics license is available and must be enabled on the switch. When Extended Fabrics is enabled, the ISLs (E_Ports) are configured with a large pool of buffer credits. The license allows the following two options to be applied to a long-distance link.

LD (dynamic) mode calculates buffer credits based on the distance measured at port initialization time. Brocade switches use a proprietary algorithm to estimate distance across an ISL. The estimated distance is used to determine the buffer credits required in LD (dynamic) extended link mode, based on the "-frameSize" option, if used, or a maximum Fibre Channel payload size of 2,112 bytes and a maximum frame size of 2,148 bytes.

LS (static) mode is the recommended mode to use to configure buffer credits over long distances, as it provides the user with the full control needed over buffer credit calculations. Configure the number of buffers by using the "-frameSize" option along with the "-distance" option.

Any time that you need a long-distance link of 10 km or more, you must enable the Extended Fabrics license (included as part of the Enterprise Software Bundle) on your switch devices, to allow for the additional buffer credits. This license, included with directors, comes at no additional charge. (See Table 2.)

Fabric Performance Impact Monitoring

A bottleneck is a port in the fabric where frames cannot get through as fast as it is expected that they should. In other words, a bottleneck is a port where the offered load is greater than the achieved egress throughput. Bottlenecks cause undesirable degradation in throughput on various links. When a bottleneck occurs at one place, other points in the fabric can experience bottlenecks as the traffic backs up.

Congestion bottlenecks are relatively easy to detect and, in effect, can be detected by several Brocade products. One of these is the Monitoring and Alerting Policy Suite (MAPS) from Brocade. However, Fabric Performance Impact (FPI) monitoring, which is discussed in the next section, is an alternative mechanism that provides much more information about the congestion itself.

Table 2. Approximate number of buffer credits required for 50 km, with and without compression and encryption.

A Distance of 50 km with 100% Link Utilization				Additional	2 Gbps	4 Gbps	8 Gbps	10 Gbps	16 Gbps
SOF, Header, CRC, EOF	Payload	Total Frame Bytes (average frame size)	Smaller than Full Frame by xx%	Compression and Head Room	Buffer Credits Required 8b/10b	Buffer Credits Required 8b/10b	Buffer Credits Required 8b/10b	Buffer Credits Required 64b/66b	Buffer Credits Required 64b/66b
36	2112	2148	0.00%	NO	56	105	204	254	402
				2.2 : 1	N/A	N/A	442	551	877
36	1540	1576	26.63%	NO	74	141	276	344	546
				2.2 : 1	N/A	N/A	600	748	1193
36	1038	1074	50.00%	NO	105	204	402	501	798
				2.2 : 1	N/A	N/A	877	1095	1748
36	834	870	59.50%	NO	129	251	495	617	984
				2.2 : 1	N/A	N/A	1081	1350	2156
36	476	512	76.16%	NO	214	422	837	1044	1667
				2.2 : 1	N/A	N/A	1833	2290	3660

Device latency-based bottlenecks, called latency bottlenecks, are much more difficult to detect. Once again, FPI monitoring provides great capabilities for user troubleshooting if this type of bottleneck is occurring.

FPI monitoring is one of the key enhancements available with Brocade FOS 7.3 and later releases. It is a simplified and advanced tool for detecting abnormal levels of latency of an F_Port in a fabric. It substantially improves monitoring for visibility of device latency in a fabric over the legacy Bottleneck Detection capability. FPI, at Brocade FOS 7.3, does not monitor E_Ports.

FPI provides automatic monitoring and alerting of latency bottleneck conditions through predefined thresholds and alerts in all predefined policies in MAPS. The advanced monitoring capabilities identify ports in two latency severity levels and provide intuitive reporting in the MAPS dashboard under a new Fabric Performance Impact category.

FPI monitors the F_Port for latency conditions and classifies them into two states. One is the I/O performance impact state, which is the condition causing I/O disruption. Another condition is I/O frame loss. This condition could potentially severely degrade throughput. The default action with those two states is to report into RASLOG.

Two significant enhancements were made to FPI with the release of Brocade FOS 7.4. Slow Drain Device Quarantine (SDDQ) is a mechanism to recognize that a slow draining device is operating in the fabric and is causing latency issues. This situation is detected, and then a Registered State Change Notification (RSCN) is sent to all of the switches in the fabric. This alerts each of the zones that can communicate with the slow drain device to only send traffic to that device across a low-priority Virtual Channel. The slow drain device is now "quarantined" away from other data traffic, which helps alleviate fabric congestion. This process is also known as Automatic VC Quarantine (AVQ). Data traffic that is transmitted to other normal acting devices continues to use the medium VC, as always.

The second enhancement is called "port toggle," which is an action that automatically recovers slow drain device conditions when they are detected by FPI monitoring. A port toggle, (which is a port disable followed by a port enable) can recover the ports from some slow drain device conditions or force traffic failover to an alternate path. Port toggle and SDDQ are mutually exclusive. If an attempt is made to use these together, it might result in unpredictable behavior.

An Overview of Buffer Credit Recovery

During normal SAN operation, FC frames, R_RDYs, or VC_RDYs may become corrupted in transport, which could be in the form of one or more bits in error. Bit errors can be caused by optics failing, bad cables, loose connections, optical budgets not within tolerances, intermittent hardware malfunctions, long-distance connections, and so on.

If the receiving side of an ISL connection cannot recognize the Start of Frame (SOF) in the header of the incoming frame, then it does not respond with the appropriate VC_RDY. Because the sender had decremented the available buffer count by 1 and, in turn, does not receive the corresponding VC_RDY, synchronization between the sender and receiver

is now skewed by the missing VC_RDY. When this condition occurs, no explicit error is generated, and it will not resolve itself without some form of automated or manual recovery mechanism. Not receiving a VC_RDY is not considered an error condition. Corrupted SOF or VC_RDY violates the Cyclic Redundancy Check (CRC) and generates an error or increments a statistical counter; however, this is not specific to BB_Credit or VC_RDY loss errors and does not initiate buffer credit recovery or adjust associated counters. Another scenario occurs when the VC_RDY that is returned by the receiver is corrupted. The result is the same as in the previous scenario. The sender will have an outstanding buffer count of 1 less than what is actually available on the receiving side.

The Brocade buffer credit recovery mechanism is activated when a port on a Brocade switch that supports 8 Gbps or 16 Gbps FC is configured for LE, LS, or LD long-distance mode. Buffer credit recovery allows links to recover after buffer credits are lost when the buffer credit recovery logic is enabled. The buffer credit recovery feature also maintains good performance. If a credit is lost, a recovery attempt is initiated. During link reset, the frame and credit loss counters are reset without disruption—but potentially with slight and temporary performance degradation.

Forward Error Correction for 16 Gbps Gen 5 Links

Forward Error Correction (FEC) allows the recovery of error bits in a 10 Gbps or a 16 Gbps data stream. This feature is enabled by default on all ISLs and Inter-Chassis Links (ICLs) of 16 Gbps FC platforms. FEC can recover from bit errors, because the Condor3 ASIC collects the hi-order check bits during the 64-byte/66-byte encoding process of the frame. FEC then uses them to enable the receiver to recover the frame integrity.

Since it is often bit errors in the data stream that cause the loss of buffer credits at the transmitter, customers find that FEC not only helps to avoid data retries but also minimizes the loss of buffer credits on the link. Users can enable or disable the FEC feature while configuring a long-distance link. This allows end users to turn off FEC where it is not recommended, such as when configuring active Dense Wavelength Division Multiplexing (DWDM) links. Transparent DWDM links can still make use of FEC, as long as the original transmitter and the target receiver are both controlled by Condor3 ASICs.

Summary

Despite the perceptions of some analysts that Fibre Channel is dead or dying, it continues to develop with innovation and continues to be the ruler of lossless, high-speed storage networking.

One of the early and lasting benefits of FC fabrics was the fact that their communications showed extremely high performance. A factor in that high performance is that all FC traffic is sent in a lossless fashion: A frame that is sent across an FC fabric is guaranteed to arrive at its intended destination.

Traditional databases such as Online Transaction Processing (OLTP) and Online Analytical Processing (OLAP) are still increasing their performance speed and I/O per Second (IOPS) capability, while continually demanding absolutely error-free performance. Both of these databases perform optimally in lossless environments. The first time

that a packet is sent to a receiver should be the only time, in order to avoid errors and maintain performance, especially in high-volume applications. Fibre Channel can promise total protection against silent data corruption. This type of corruption can happen in an environment when incomplete or incorrect data overwrites data on a Fibre Channel fabric storage device, a problem that can lead to costly downtime and even data loss.

Hopefully, you now more fully understand the benefits and value of a lossless data environment when buffer credits are properly configured in a Fibre Channel fabric.

About Brocade

Brocade networking solutions help organizations achieve their critical business initiatives as they transition to a world where applications and information reside anywhere. Today, Brocade is extending its proven data center expertise across the entire network with open, virtual, and efficient solutions built for consolidation, virtualization, and cloud computing. Learn more at www.brocade.com.

Corporate Headquarters
San Jose, CA USA
T: +1-408-333-8000
info@brocade.com

European Headquarters
Geneva, Switzerland
T: +41-22-799-56-40
emea-info@brocade.com

Asia Pacific Headquarters
Singapore
T: +65-6538-4700
apac-info@brocade.com

BROCADE