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 2006 International Conference.

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

DESIGNING AND MANAGING FICON INTER-SWITCH LINK INFRASTRUCTURES

Stephen R. Guendert, McDATA Corporation
Dr. H. Pat Artis, Performance Associates, Inc.

FICON Inter-Switch Links (ISLs) were created to provide a solution to the bandwidth requirements of complex parallel sysplex environments. This paper will review basic ISL concepts, how they are defined using Hardware Configuration Definition (HCD), with emphasis on design principles for fault tolerance, and examine the measurement data available for performance management and capacity planning.

Introduction

In September 1990, the ESCON channel architecture was introduced to address the limitations of 4.5 MB/Sec parallel (bus and tag) channels. ESCON provided significant improvements in distance, data rate, switching topologies, as well as service time. From a historical perspective, ESCON was a highly successful storage network protocol for mainframe systems and was the first broadly deployed fibre channel application. However, by the end of the 1990s, 20 MB/Sec ESCON was no longer able to support the I/O bandwidth and device addressing demands of processors with ever increasing capacities. IBM responded to these issues with the announcement of FICON channels for the S/390 9672 G5 processor in May of 1998.

Over the past 8 years, FICON has rapidly evolved from 1 Gbit FICON Bridge Mode (FCV) to the 4 Gbit FICON Express4 channels that were announced in May of 2006. While the five generations of FICON channels have each offered increased performance and capacity, from the perspective of this paper the most important enhancements relate to the fibre channel upper level protocol (ULP) employed by FICON. Specifically, the first FICON ULP was called FC-SB-2. This specification supported native FICON storage devices, channel-to-channel connections, FICON directors (switches), as well as a number of other key features. However, the FC-SB-2 ULP did not support switched paths over multiple directors. When IBM introduced the FC-SB-3 ULP in January of 2003, this limitation was addressed. While FC-SB-2 employed a single byte link address that specified just the switch port on the director, FC-SB-3 employs a 2 byte address where the first byte is the address of the destination director and the second byte is the address of the port on the destination director. As will be discussed, directors discover each other during initialization and build routing tables to automatically redirect frame traffic across the ISLs that connect the directors. The focus of this paper is the definition, topology, performance, configuration design, and management of ISLs.

Cascaded FICON Defined

Cascaded FICON refers to an implementation of FICON that allows two storage fabrics to be linked via connections between pairs of directors. The director-to-director connections are known as ISLs. ISLs support processor-to-processor, processor-to-disk or tape subsystem, and subsystem-to-subsystem logical switched connections. [CRON03] Cascaded FICON directors facilitate the design and implementation of geographically distributed parallel sysplex **(GDPS)** environments and substantially reduce the infrastructure costs and complexities associated with the implementation. Cascaded directors also permit greater flexibility in the FICON architecture, more effective utilization of fibre links, and higher data availability in the enterprise. They also allow for more robust disaster recovery and business continuity as shown in Figures 1 and 2.

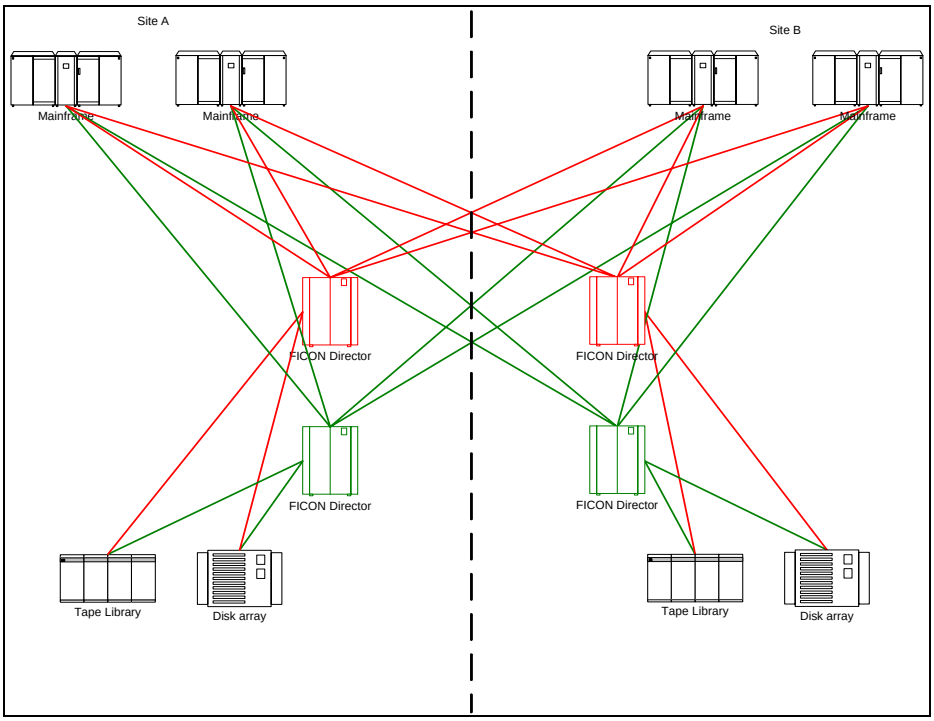

Figure 1. Two Site Non-Cascaded FICON Environment.

In Figure 1**,** processors and storage subsystems at both sites are interconnected via FICON links. The host channels are extended to the FICON directors to allow for cross-site storage access. If each line in the figure represents two FICON channels, then sixteen fibers would be required between the two sites. While this configuration provides excellent service, it presents issues related to complexity and cost. Specifically, the large number of fibers required between the two sites and the low utilization of these expensive resources.

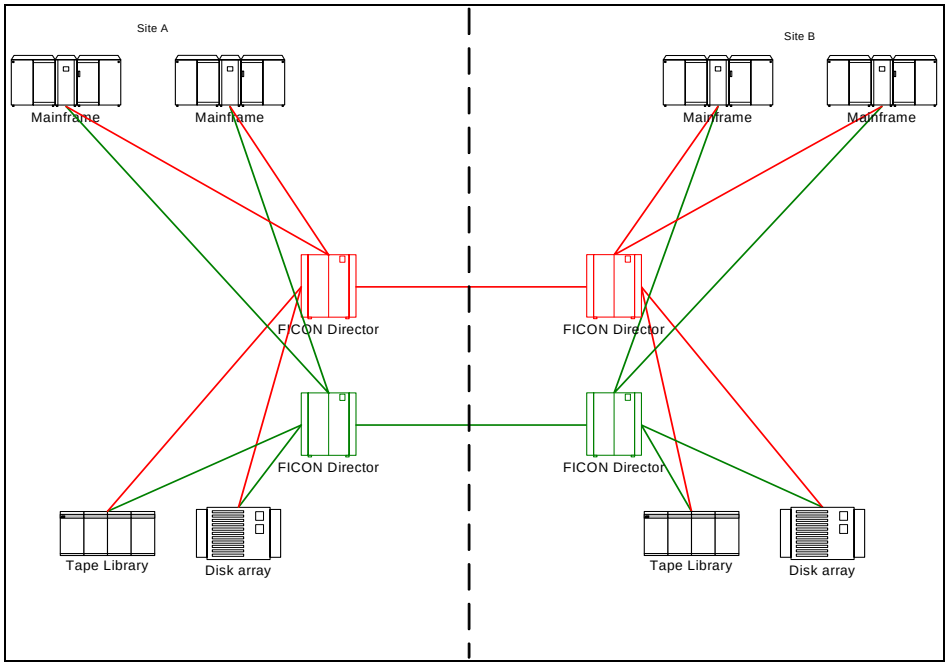

Figure 2. Two Site Cascaded FICON Environment

In Figure 2, each line represents two FICON channels. The sixteen fibre links discussed in Figure 1 between the two sites have been replaced by four ISLs. Since FICON is a packet-switched protocol (as opposed to the circuit-switched ESCON protocol), concurrent I/Os can share the bandwidth provided by the ISLs. Hence, far fewer links between the two sites are required and these expensive resources are far more effectively utilized. In addition, ISLs can be added as bandwidth demand and traffic patterns dictate. [GUEN05]

There are three hardware and software requirements specific to cascaded FICON:

- The FICON directors must support the high integrity fabric feature. Since the low level microcode implementation of this feature differs from vendor to vendor, directors connected via ISLs must be from the same vendor,
- zSeries processors must be employed since they support the two-byte addressing schema required by the FC-SB-3 ULP, and
- z/OS version 1.4 or above.

Since 1.4 is currently the minimum supported version of z/OS, host software is no longer a primary consideration.

Packet Flow and Credits

The fundamental objective of flow control is to prevent a transmitter from overrunning a receiver by allowing the receiver to pace the transmitter. Each I/O is managed 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. [ARTI05]

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. 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 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.

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 negotiate with each other during fabric login (FLOGI) to determine the number of credits available for the send and receive ports on each node. [KEMB02]

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. Hence, 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. That is, every time the distance

doubles, the number of required credits doubles to avoid transmission delays for a specified distance.

BB_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 negotiate the number of credits available during FLOGI. 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, and
- for each R_RDY received by the transmitter, the counter is incremented by one.

As long as the transmitter has available credits, it is free to continue sending data. Hence, the maximum number of frames that may be in flight over a link at any time is limited to the number of buffer credits communicated by the receiver port during FLOGI. For complex FICON connections that involve directors or pairs of directors connected by ISLs, the credits between any two adjacent sets of nodes may be viewed as participants in an electronic bucket brigade. As long as each set of adjacent ports has available credits, it passes along the frames that constitute an end-to-end logical operation. The flow of frame transmission between adjacent ports is regulated by the receiving port's presentation of R_RDYs.

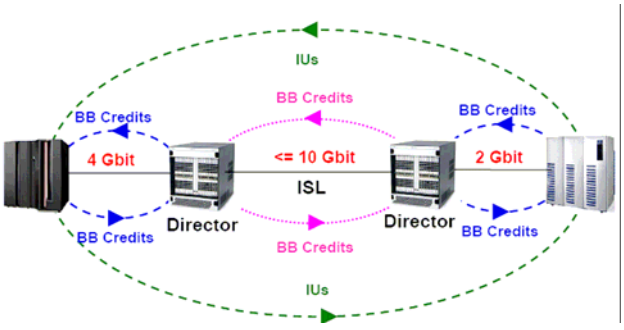

Figure 3. Buffer-to-Buffer and End-to-End Flow Control.

Another key aspect of the fibre channel architectures is that buffering allows each node-to-node segment to run at a different data rate. If fibre costs between sites are high, then the installation may chose to reduce the number of ISLs by paying the fixed initial cost of faster ISL cards rather than signing long term (recurring cost) agreements for additional fibres.

Inter-Switch Link (ISL) Definition and Addressing

As was shown in Figure 2, a logical connection over a cascaded FICON director configuration incorporates three end-to-end links. The first link connects the FICON channel N_Port (node port) to an F_Port (fabric port) on the FICON director. The second link connects an E_Port (expansion port) on the local director to an E_Port on the remote director. Finally, the third link connects the F_Port on the director to an N_Port on the subsystem.

As the readers who have employed HCD to define an I/O configuration are already aware, HCD defines the relationship between channels and directors (by switch ID) and specific switch port (as well as the destination switch ID for cascaded connections) for the switch to director segment of the link. However, HCD does not define the ISL connections. While this may initially appear to be surprising based on a reader's prior experience with the exacting specificity required by HCD, it means that the management of the traffic over ISLs is controlled exclusively by the directors. Hence, additional ISL bandwidth can be added to a configuration without any modification to the environment's HCD definitions.

HCD simply assumes that the links are present and requires that 2 byte addressing be employed to define the connectivity to storage subsystems. The first byte is the switch ID and the second byte is the port to which the storage subsystem is connected. During initialization, the switches identify their peers and create a routing table so that frames for remote subsystems may be forwarded to the correct director.

Figure 4 shows that an FC-FS 24-bit port address identifier is divided into three fields. They are domain, area, and AL_Port (arbitrated loop).

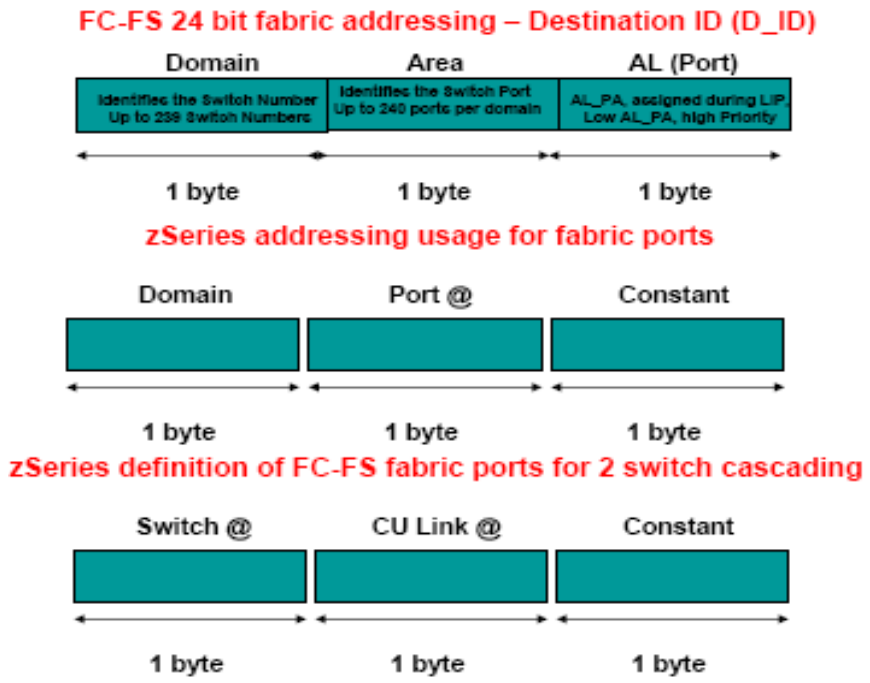

Figure 4. FC-FS Fabric Addressing.

In a cascaded FICON environment, the first 16-bits of the address are employed by the zSeries server to address a FICON subsystem. The first 8-bits of this address are the switch ID and the second 8-bits of the address define the port number on the switch.

The final 8-bits are not employed by the FC-SB-3 ULP and they are set to a constant. Figure 5 provides an example of how the addressing schema is employed to connect a channel to a subsystem over an ISL.

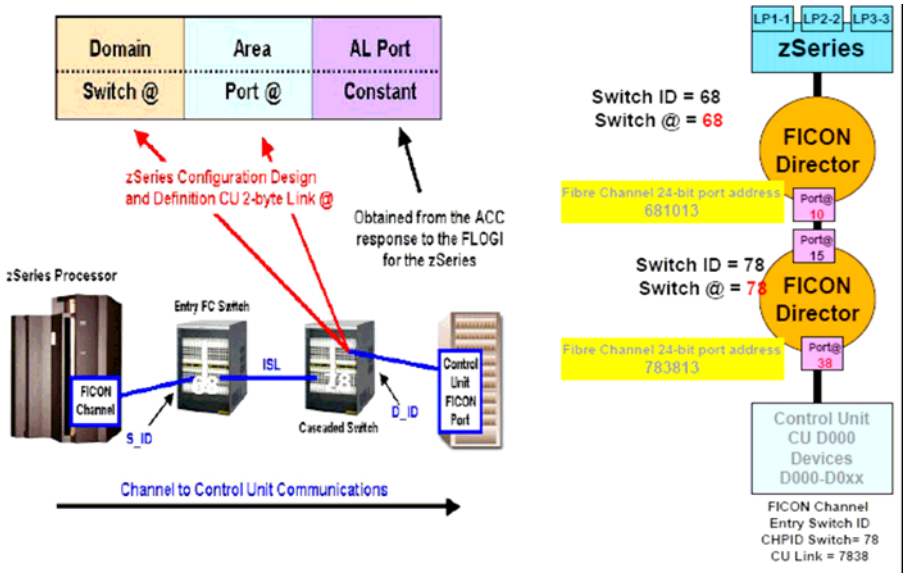

Figure 5. Addressing for Cascaded Directors.

Figure six provides a sample of the narrative report that may be generated by HCD during the process of creating an I/O Control Program (IOCP) dataset for the processor.

```
ID (no change in ID statement for FICON)
RESOURCE PARTITION=((CSS(0),(SYSA,1),(SYSB,2),(SYSC,3)),
      (CSS(1),(SYSD,1),(SYSE,2),(SYSF,3)))

CHPID PATH=(CSS(0),80),SHARED,PART=((CSS(0),(SYSA,SYSB),(=))),TYPE=FC,SWITCH=61,PCHID=160
CHPID PATH=(CSS(0),81),PART=((CSS(0),(SYSB),(=))),TYPE=FC,SWITCH=61,PCHID=1A0
CHPID PATH=(CSS(0),82),PART=((CSS(0),(SYSB),(=))),TYPE=FC,SWITCH=61,PCHID=130
CHPID PATH=(CSS(0,1),83),SHARED,PART=((CSS(0),(SYSC),(=)),(CSS(1),(SYSD),(=))),
      TYPE=FC,SWITCH=61,PCHID=1D0

CNTLUNIT CUNUMBR=8000,PATH=((CSS(0),80,81,82,83),(CSS(1),83)),
      UNITADD=((00,256)),LINK=((CSS(0)6212,6222,6232,6242),(CSS(1),6242)),
      CUADD=0,UNIT=2105
CNTLUNIT CUNUMBR=8100,PATH=((CSS(0),80,81,82,83),(CSS(1),83)),
      UNITADD=((00,256)),LINK=((CSS(0)6212,6222,6232,6242),(CSS(1),6242)),
      CUADD=1,UNIT=2105
        *
        *
CNTLUNIT CUNUMBR=8700,PATH=((CSS(0),80,81,82,83),(CSS(1),83)),
UNITADD=((00,256)),LINK=((CSS(0)6212,6222,6232,6242),(CSS(1),6242)),
      CUADD=7,UNIT=2105

IODEVICE  (no change for FICON)
```

Figure 6. Sample IOCP.

High Integrity Enterprise Fabrics

Data integrity is paramount in a mainframe environment. End-to-end data integrity must be maintained throughout a cascaded FICON director environment to ensure that any transmission errors are identified and corrected. In addition, it is imperative to verify that the data was transmitted to or received from the proper device. While vendor implementations differ, they all meet the same standards. High integrity fabrics offer two key components:

- Insistent Domain IDs. FICON directors are prohibited from automatically changing their switch IDs. Rather, manual intervention of the switch is required. Insistent Domain IDs prohibit the use of dynamic Domain IDs. Hence, they guarantee that predictable Domain IDs are enforced within a fabric and prevents the introduction of duplicate Domain ID into a fabric. Insistent Domain IDs also provide enhanced security and integrity for the FICON fabric.
- Fabric Binding. Fabric binding prohibits the introduction of directors that do not support the high integrity fabric feature from being introduced to a fabric. Before a director may be introduced to a fabric, it must be added to the acceptable membership list for the fabric. The membership list entries include Worldwide Name (WWN) and Domain ID of each acceptable director. Using the Domain ID ensures that there will be no address conflicts (i.e., duplicate Domain IDs when the fabrics are merged) and prohibits the introduction of unknown directors. When FICON directors are connected, they exchange their membership lists. This membership list exchange is a Switch Fabric Internal Link Service (SW_ILS) function, which ensures a consistent and unified behavior across all potential fabric access points.

ISL Link Mapping

As was previously discussed, the assignment of traffic between directors over the intervening ISLs is completely controlled by the directors. There are four primary methods of assigning traffic to the ISLs. They are:

- round robin,
- prohibit dynamic connectivity mask (PDCM),
- preferred pathing, and
- ISL trunking.

Each of these techniques is discussed in the following paragraphs.

Round robin assignment is also more commonly referred to as fabric shortest path first (FSPF). FSPF

distributes traffic over the ISL link by a rotation assignment algorithm. Every time an ISL is added, the ISL traffic assignments change. While very simple, this technique is not attractive to an experienced mainframe architect since they cannot prescribe how the paths to a subsystem are mapped to the ISLs. In the worst case, all of the paths to a subsystem might be mapped to the same ISL. Moreover, once a path is assigned an ISL, that assignment is persistent. Hence, an enterprise just can't simply add another ISL to a FSPF managed configuration to address an immediate capacity requirement. Rather, the two directors must go through a power-on reset (commonly referred to as a fabric bounce) so that FSPF algorithm redistributes the load over the ISLs. Hence, the enterprise must endure the capacity shortfall until an outage can be scheduled.

The PDCM allows the user to block specific ISL ports from round robin assignment. For example, by blocking tape traffic from ISLs being employed for disk replication, one can attempt to protect *critical traffic* from bulk data transfer traffic. For readers who employed ESCON directors, the definition of an allow-prohibit matrix is a familiar process. Unfortunately, PDCM is a very labor intensive and complex process. Moreover, an installation's best intentions for protecting critical traffic can actually reduce the resilience of the enterprise's fabric. As was the case with FSPF, a fabric bounce is required to redistribute the traffic after additional ISLs are added to a configuration.

Rather than the PDCM approach of specifying prohibited links, preferred pathing allows an installation to proactively assign the ISL routes for critical traffic. In the extreme case, preferred pathing can also incorporate PDCM to protect critical traffic. As was the case with PDCM, preferred pathing is labor intensive and provides opportunities for unintentionally reducing the resiliency of the enterprise's fabric. Once again, a fabric bounce is required to add capacity.

Figure 7 provides an overview of trunking. Simply described, trunking allows the directors to manage the intervening ISLs as an aggregate resource. While there are small differences between vendor implementations, trunking automatically balances the traffic over the available ISL links. A good analogy to this process is the addition of an instruction processor (IP) to a running z/OS image. The operating system simply recognizes the presence of the new IP and balances the workload over all of the available IPs. With trunking, new ISLs are automatically incorporated into the frame traffic flow without user assignment or a fabric bounce. While some switch vendors offer facilities for prioritizing traffic, true quality of service management for ISL traffic is still an architectural goal.

Figure 7. ISL Trunking.

Regardless of which method of ISL mapping or management is utilized, the highest priority for the design of a z/OS fabric is fault tolerance. Since the complete failure of a pair of linked directors is the most significant failure that a designer can attempt to sustain, the minimum number of director pairs required to guarantee service levels may be determined by answering two questions:

(1) Does the enterprise employ 4 or 8 path Logical Control Units (LCUs)?
(2) What is the minimum number of paths required to maintain worst-case service levels?

Depending on the answers to these questions: two, four, or eight pairs of directors connected by ISLs may be required. In no case, should a single director pair be considered for an enterprise since the pair represents a single point of failure.

Cascaded FICON Measurement Data

Since ISLs are shared paths that can impact all of an environment's workloads, the management of these resources is a critical function. There are two sources of data for this task:

- the RMF 74-7 FICON Director Activity Report records, and
- vendor specific director measurement tools.

Effective capacity planning and performance management requires the use of both of these data sources.

The RMF 74-7 records report on four classes of data:

- global,
- switch,
- connector, and

- port data.

These four classes of data are collected by RMF for each interval if the **FCD** keyword is specified in the z/OS image's **ERBRMFnn** parmlib member. Port data includes average read/write frame sizes, read/write bytes transferred, error counts, and pacing delays for each port. Frame pacing delays occur when a director port exhausts its available bb_credits. These delays are measured and reported in 2.5 micro-second units. Figure 8 is a sample of a FICON Director Activity Report. [NEVI05]

```
                                  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 8. FICON Director Activity Report.

In addition to RMF, each FICON director vendor offers their own performance monitoring software. This proprietary PC-based software is intended for the real time analysis and long term trending of switched infrastructures. Some of these tools are capable of collecting data at 5-second intervals.

These tools are designed to allow the end user to optimize their existing resources by monitoring the utilization of existing links. The tools also provide real time topology views as well as graphical analyses of traffic flow. Figure 9 provides a sample view of a fabric topology utilization diagram.

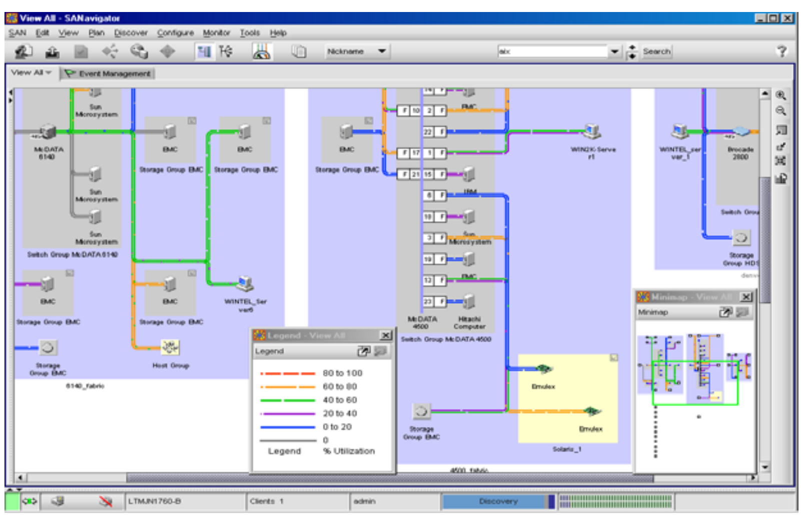

Figure 9. Sample Fabric Topology Utilization Diagram.

While the vendor-specific tools are ideal for real time problem identification, RMF is preferable for long term trending and data collection since it is already part of an enterprise's standard capacity planning processes. While it is very difficult to identify real time fabric problems with 15-minute RMF interval data, the long-term management of vendor specific measurement data is also unattractive. Hence, the best solution is to exploit the best features of each data source.

Observations

Clearly, a future announcement of FICON Express8 by IBM should not come as a surprise to any z/OS installation. That is, IBM should be expected to continue to exploit the advances in the underlying fibre channel architecture to deliver improved cost performance for z/OS environments. Moreover, terrorism and concerns about natural disasters will increase the requirements for ISL based architectures in both open system and z/OS environments. While an enterprise may initially approach ISLs from the perspective of their z/OS environment, the design and implementation of a cascaded director architecture is an enterprise-wide endeavor.

Finally, since the ISL fabric will rapidly become the most critical resources in the enterprise's infrastructure, it is important to design the fabric carefully. Once in place, there may never be an opportunity for an enterprise wide fabric outage to correct prior oversight in the design process.

References

[ARTI05] Dr. H. Pat Artis and Robert Ross, *Managing Complex FICON Configurations,* 2005. www.perfassoc.com

[CRON03] Catherine Cronin and Richard Basener, *Performance Considerations For a Cascaded FICON Director Environment Version 0.2x,* March 2003. www.ibm.com

[GUEN05] Stephen R. Guendert, *Taking FICON to the Next Level: Cascaded High Performance FICON.* Proceedings of the 2005 International Conference Of the Computer Measurement Group, December 2005.

[KEMB02] Robert Kembel, *Fibre Channel: A Comprehensive Introduction.* Northwest Learning Associates, 2002.

[NEVI05] Iain Neville, Bill White & Hans-Peter Eckham, *FICON Implementation Guide.* IBM Redbooks (2005). SG24-6497. www.ibm.com

[SCHU04] Greg Schulz, *Resilient Storage Networks.* Elsevier, 2004.

[TROW02] Ken Trowell, Bill White et al. *FICON Native Implementation and Reference Guide.* IBM Redbooks, 2002. SG24-6266-01. www.ibm.com