#SHAREorg

SHARE

A First Look Into The Inner Workings and Hidden Mechanisms of FICON Performance

- David Lytle, BCAF
- Brocade Communications Inc.

- Monday August 10, 2015 – 10:00am to 11:00am

- Session Number - 17505

QR Code

SHARE in Boston

Legal Disclosure

All or some of the products detailed in this presentation may still be under development and certain specifications, including but not limited to, release dates, prices, and product features, may change. The products may not function as intended and a production version of the products may never be released. Even if a production version is released, it may be materially different from the pre-release version discussed in this presentation.

Nothing in this presentation shall be deemed to create a warranty of any kind, either express or implied, statutory or otherwise, including but not limited to, any implied warranties of merchantability, fitness for a particular purpose, or non-infringement of third-party rights with respect to any products and services referenced herein.

Notes as part of the online handouts

I have saved the PDF files for my presentations in such a way that all of the audience notes are available as you read the PDF file that you download.

If there is a little balloon icon in the upper left hand corner of the slide then take your cursor and put it over the balloon and you will see the notes that I have made concerning the slide that you are viewing.

This will usually give you more information than just what the slide contains.

I hope this helps in your educational efforts!

A first look into the Inner Workings and Hidden Mechanisms of FICON Performance

AGENDA – # 17505: 1st Look into the Inner Workings:

- Discuss FICON design and Performance considerations when using a switched-FICON infrastructure.

AGENDA – # 17506: A Deeper Look into the Inner Workings:

- Focused more on underlying protocol concepts:
  - Switch Firmware
  - Buffer Credit Flow Control
  - Fabric Congestion, Backpressure and Slow Drain Devices
  - Buffer Credit Starvation
  - Data Encoding, Forward Error Correction and Compression/Encryption

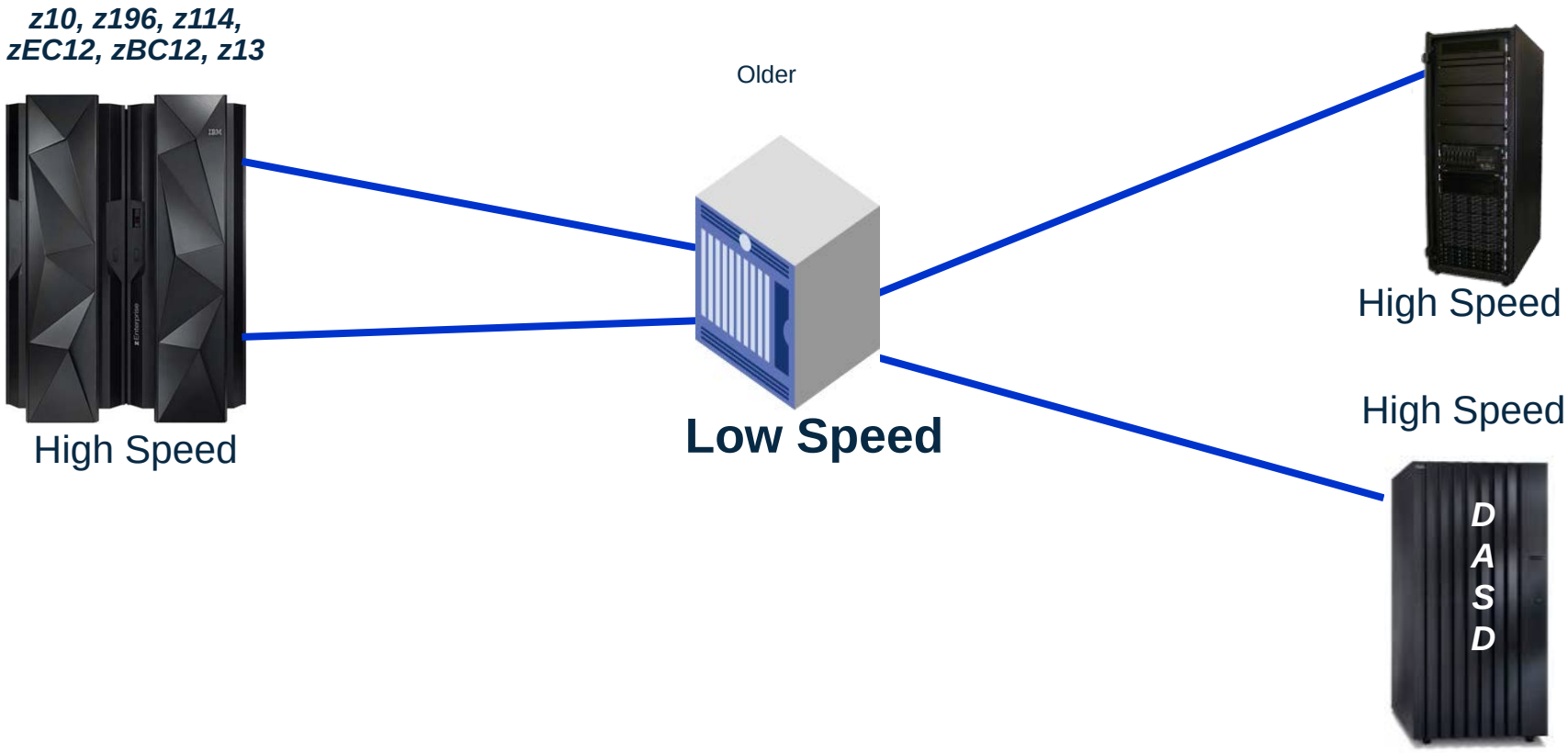
Your I/O Infrastructure Has To Keep Pace

A Rapidly Changing World is Driving us to z13

- Vastly Improved Internal Performance of System z13
- FICON Express8S and FICON Express16S Channel Cards – massive throughput
- zHPF (High Performance FICON) is being deployed by more and more shops
- DASD Vendors fully support 8 Gbps attachment and some even 16 Gbps
- Flash storage and SSD are becoming Strategic to many Customers
- Accelerated Deployment of Virtual Tape subsystems like TS7700 Grid and DLm
- And many new capabilities are beginning to show up:
  - Forward Error Correction helps to stabilize performance across links
  - FICON Dynamic Routing (Exchange-based Routing) for ISL performance
  - Fabric I/O Priority will help Un-Clog some of our Customer's I/O Infrastructures

Components Must Work In Harmony To Achieve Value

z System™ and DASD has been scaling up for performance and scaling out for capacity which means that your old I/O infrastructure needs updating!



Components Must Work In Harmony To Achieve Value

z System™ and DASD has been scaling up for performance and scaling out for capacity which means that your I/O infrastructure must provide I/O Synergy rather than an I/O blockage!

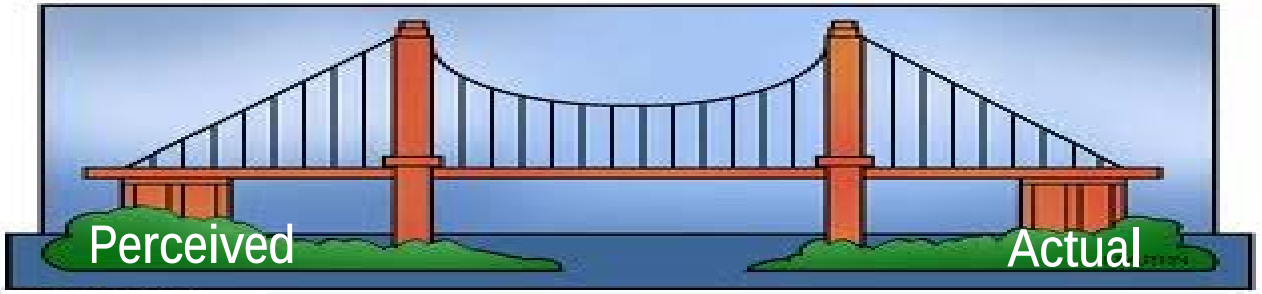
Some Things Need To Be Replaced!

Low Speed and Performance Bottlenecks!

Your enterprise needs the most robust, highest availability, fastest possible, performance oriented, ultra-scalable switching devices to be the Super-Highway between your z13 and your storage infrastructure!

Host and Storage is getting Faster and Faster

The FICON GAP



When Deploying FICON, There Is Often A Gap Between What You Expect For Its Performance And What You Actually Get!

Let Us Try To Bridge That Understanding Here!

z13 - A Superlative Processor

Requires A High Performance I/O Infrastructure

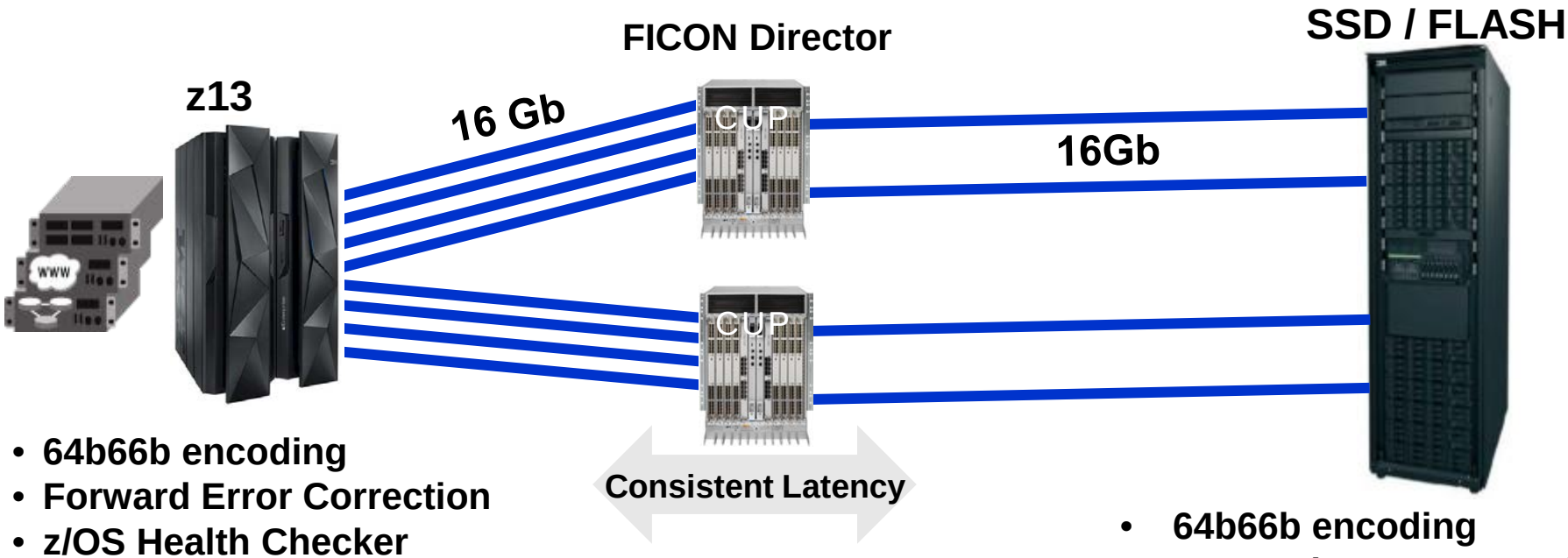
- 5 GHz Processors (among the world's fastest)
- 141 Configurable Processors (cores) available
- Up to 50% better response times
- Up to 40% capacity improvement over zEC12
- 30% better performance for Linux and Java
- 17x faster analytics for DB2-Analytics Accelerator
- Up to 8,000 virtual machines in one system
- FICON Express8S to attach 2 Gbps storage
- FICON Express16S:
  - 98,000 IOPS with zHPF
  - 2600 MBps FD with zHPF
- Up to 160 FICON Express16S cards – 320 CHPIDs
- Lots of new functions (e.g. FEC, Fabric Priority)

Must Match This With I/O Infrastructure Performance

I/O Infrastructure Performance

Maximizing System Response Time

Today's I/O Requires Low Latency and High Performance



- 64b66b encoding
- Forward Error Correction
- z/OS Health Checker
- zHPF
- Fabric I/O Priority

NOTE:
FICON Express16S is priced to be the same as FICON Express8S

- Cut-Through frame routing
- Local Switching
- Forward Error Correction
- Virtual Channels
- Bottleneck Detection
- Buffer Credit Recovery
- CUP Diagnostics

- 64b66b encoding
- Forward Error Correction
- Fabric I/O Priority

This is how a user enables their enterprise to achieve maximum benefit from storage, especially SSD storage!

I/O Latency Impacts System Response Time

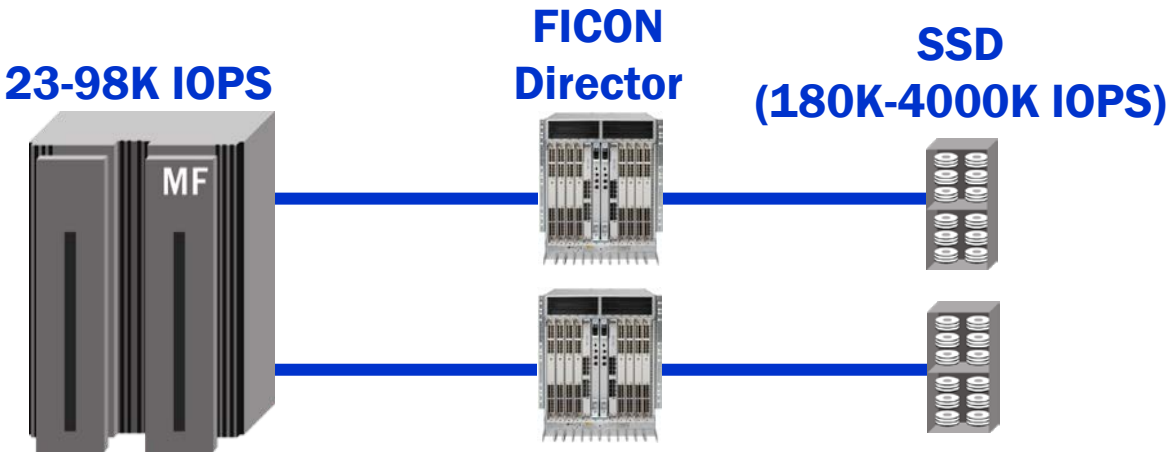
The requirement for Low Latency Networking

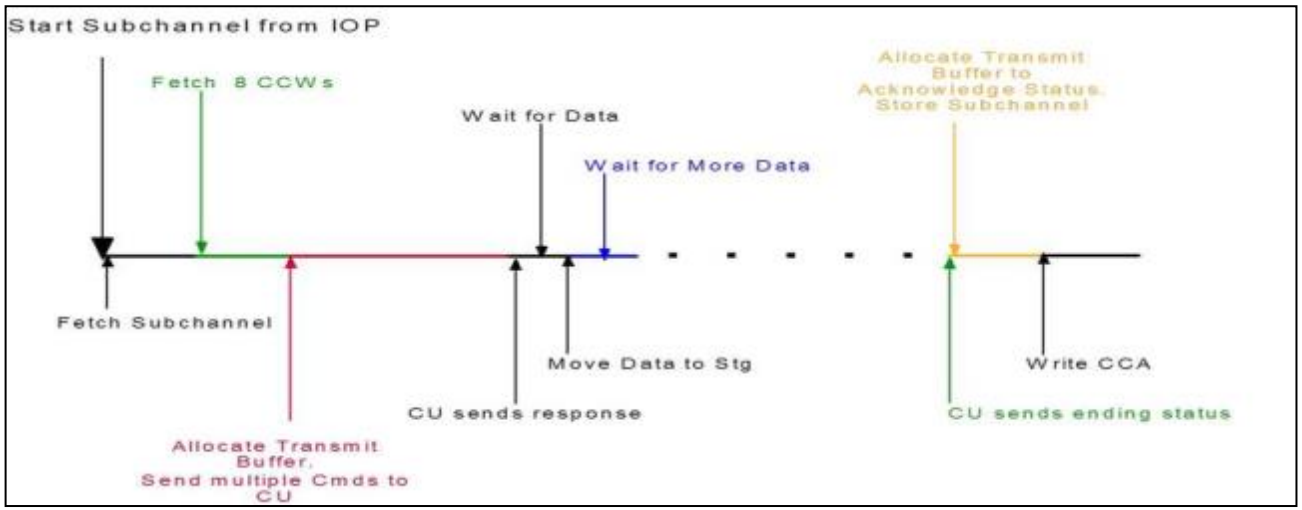
- In a world where so much computing is systematic, millisecond delays matter.
- Low-latency, high-performance, consistent Response Time networking has become a competitive imperative – but with our **data tsunami**, it is much more difficult to achieve!
  - Took from ~1940 to 2010 to store the first Zettabyte of digital data (1 million TB)
  - Took from 2010 to 2012 to store the next Zettabyte of digital data
  - Today we probably have about 7 Zettabytes of digital data stored
  - Some analysts think by 2020 there will be 44 Zettabytes of digital data stored
- It becomes evident that being just nanoseconds faster can be worth millions of dollars (euros, yen, GBP, shekels, etc.)!
- A TABB Group study, in July 2010, determined that, “A 1-millisecond advantage in a trading application can be worth US$100 million a year to a brokerage firm”.
- Simply speaking, latency is a measure of *how long* it takes for a *single* I/O request to happen from the application’s viewpoint – it is WAIT Time!

Response Time in the I/O Infrastructure

Lowering switch and infrastructure latency results in better Response Time!

- Here is a diagram as if a SINGLE FICON I/O being done on a channel link.



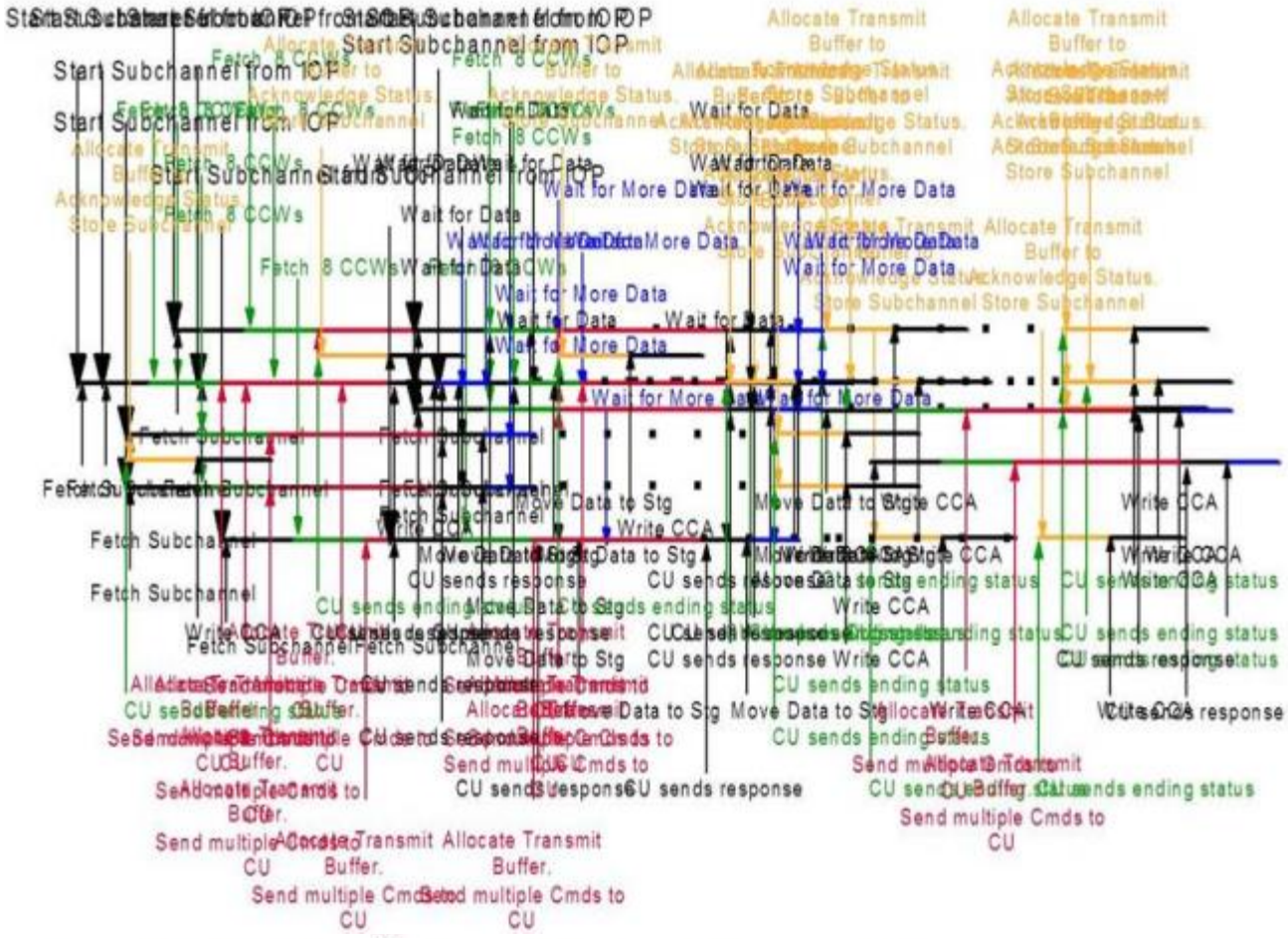


- If this example was how FICON I/O was done, that each I/O was serialized, then a user could only achieve ~400 4K I/Os per second (at 250 µs per I/O).
- But we know that is not how the z System really works so…..

Response Time in the I/O Infrastructure

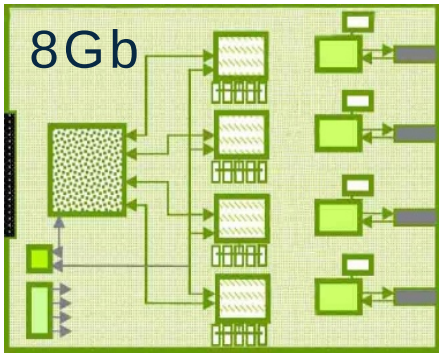
Multi-Programming is what actually occurs on the I/O Infrastructure

- Here is a diagram of a the actual multi-programming level of I/O



- **Overlapped I/O!**
- FICON is heavily overlapped, which means…
- We can only look at fabric latency in terms of ***percent of the total response time***.
- And achieving lower response time makes all users happy.
- Discussed next are some of the innovative things that can help minimize I/O response time.

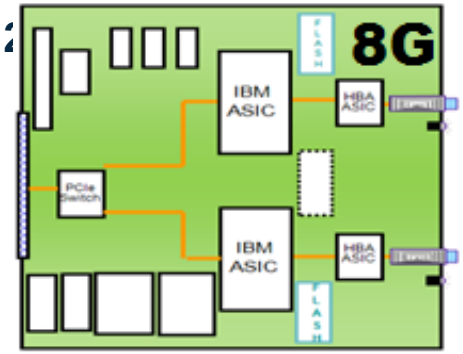
Current Mainframe Channel Cards (Features)



FICON Express8

- zXC12, z196, z114, z10
- 4 ports per feature
- <= 620 MBps for Write I/O out of 1600 MBps
- zHPF FICON Mode: <=770 MBps Full Duplex out of 1600 MBps
- 40 Buffer Credits/CHPID

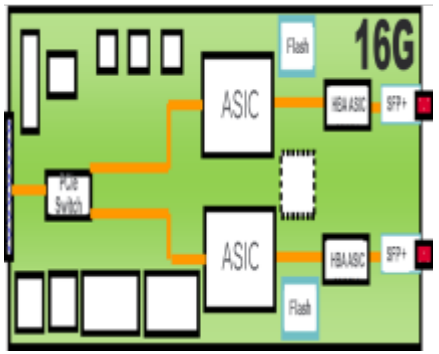
FICON buffer credits have become very limited per CHPID



FICON Express8S

- z13, zXC12, z196, z114
- 2 ports per feature
- <=620 MBps for Write I/O out of 1600 MBps
- zHPF FICON Mode: <=1600 MBps Full Duplex out of 1600 MBps
- 40 Buffer Credits/CHPID

Only 2 Ports per feature
...BUT...
Better Performance



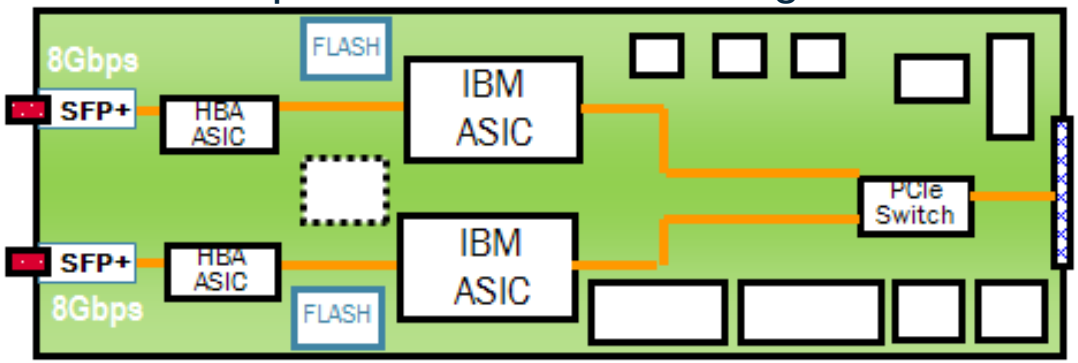
FICON Express16S

- z13
- 2 ports per feature
- <=620 MBps for Write I/O out of 1600 MBps
- zHPF FICON Mode: <=2600 MBps Full Duplex out of 3200 MBps
- 90 Buffer Credits/CHPID

Improved ASIC
...AND...
LOTS of new functionality

We Start With The 16G Channel Cards (Features)

FICON Express16S Standard, long PCIe card



- New ASIC provides 16 Gbps data rate performance and Technology benefits
- Improved FICON Express16S (FX16s) ASIC for additional functionality:
  - Enhanced zHPF functions especially over longer distances
  - Forward Error Correction (FEC) to stabilize high speed links
  - Fabric I/O Priority managed by Workload Manager/Goal Mode
  - Read Diagnostic Parameters to pinpoint physical link failures
- FX16S auto-negotiates to 4, 8, or 16 Gbps (1 and 2 Gbps NOT supported)
- FX16S also:
  - Increased bandwidth (Full Duplex 2600 MBps vs 1600 MBps onFX8S)
  - Provides substantially improved transactional latency
  - Provides up to 32% reduction in elapsed time for I/O bound batch jobs

I/O Channel Path Groups – since the 1980s

Minimizing frame latency maintains a good I/O response time

- The z System® operating system has a built in capability known as “Path Group” to balance and provide performance-oriented I/O.
- For mainframes, users can group up to 8 of their physical channels between the Channel Path IDs (CHPIDs), which are the mainframe I/O ports, out to connected switch or storage ports.
- It is the mainframe channel subsystem (CSS) algorithm that decides which path in the path group will be used by deciding which path is least busy and which paths are operational, etc.
- Path Groups allow I/O to be automatically spread evenly and fairly across a number of physical channel paths with every effort made to avoid over-subscribing any given I/O channel in the PG.
- Path Groups provide instantaneous fail over to operational links if a path group link fails.

Channel Sub-System Enhancements

zXC12 / z13 Path Group Enhancement

- When conditions occur that cause an imbalance in performance (e.g. I/O latency/throughput):
  - The channel subsystem will bias the path selection away from poorer performing paths toward the well performing paths.

- This is accomplished by exploiting the in-band I/O instrumentation and metrics of z System FICON and zHPF protocols and new intelligent algorithms in the channel subsystem to exploit this information.

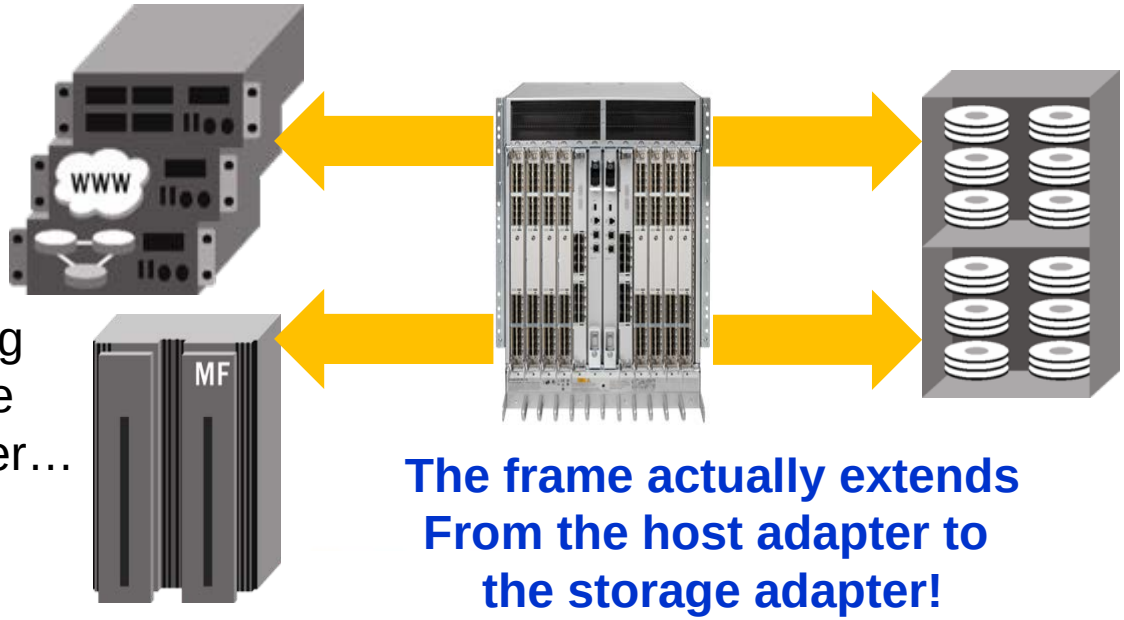
You can see that the mainframe is concerned about Latency, and makes sure that Latency is minimized on the channels in each Path Group.

FICON Switch: Cut-through Frame Routing

Minimizing frame latency maintains a good I/O response time

- Some Directors utilize “cut through” frame routing which means that a full frame does not have to reside in switch memory before it gets passed along
- “Cut through” frame routing allows them to have a low average frame latency delay for fibre channel data frames – especially significant with SSD devices!
- Reducing I/O path “Latency” provides improved I/O performance to devices like SSD!

The last of a frame’s bits are leaving the HBA or the FICON adapter…



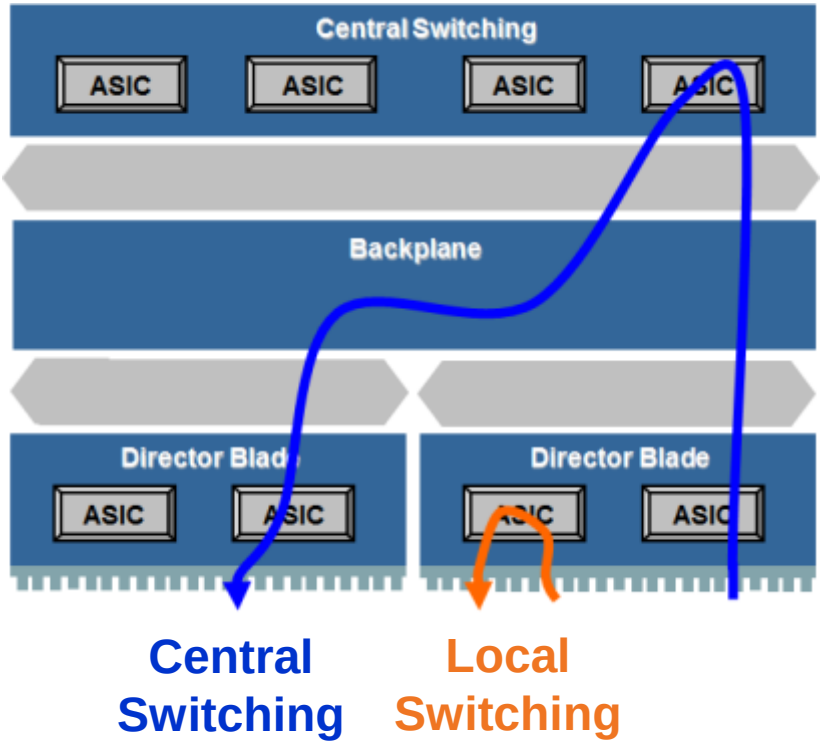
…as the first of a frame’s bits are entering the storage port and the storage systems begins to work with that same frame.

The frame actually extends From the host adapter to the storage adapter!

Local Switching = Lower Response Time

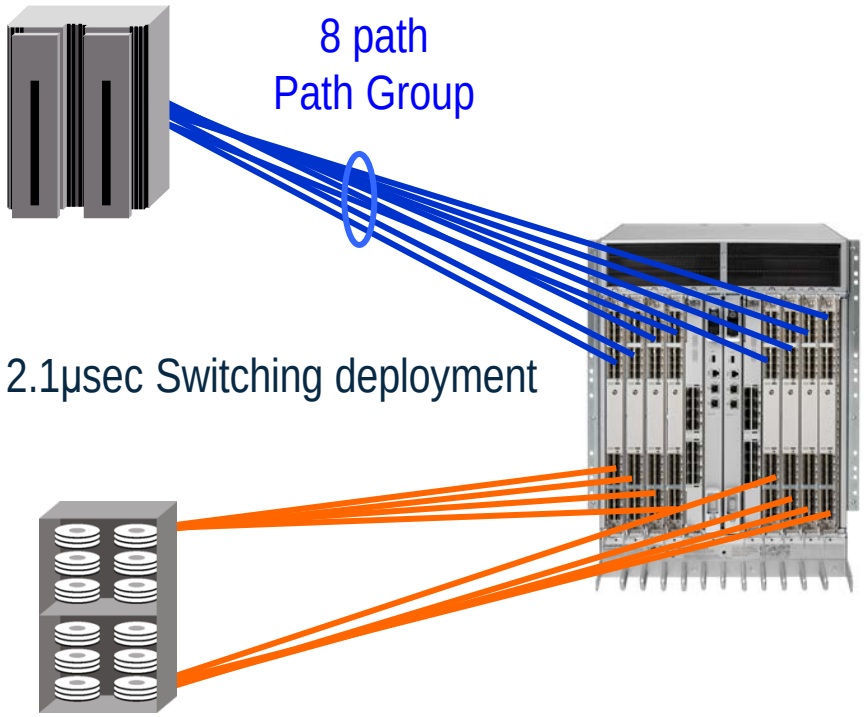
How you plug the cables into the blades determines whether local or central switching will be used!

- Standard central (backplane) switching
  - Ingress port is on one ASIC while the Egress port is on a different ASIC – even on the same blade
- ASIC sends frame from its ingress port to its egress port – very fast
- Data traffic within same port group does not cross the backplane – ASIC does switching
- Minimizing Frame latency
  - 2.1 µs between ports for central switching
  - 700 ns for "locally switched" ports



Local Switching
(Traffic in same port group doesn't consume slot bandwidth)

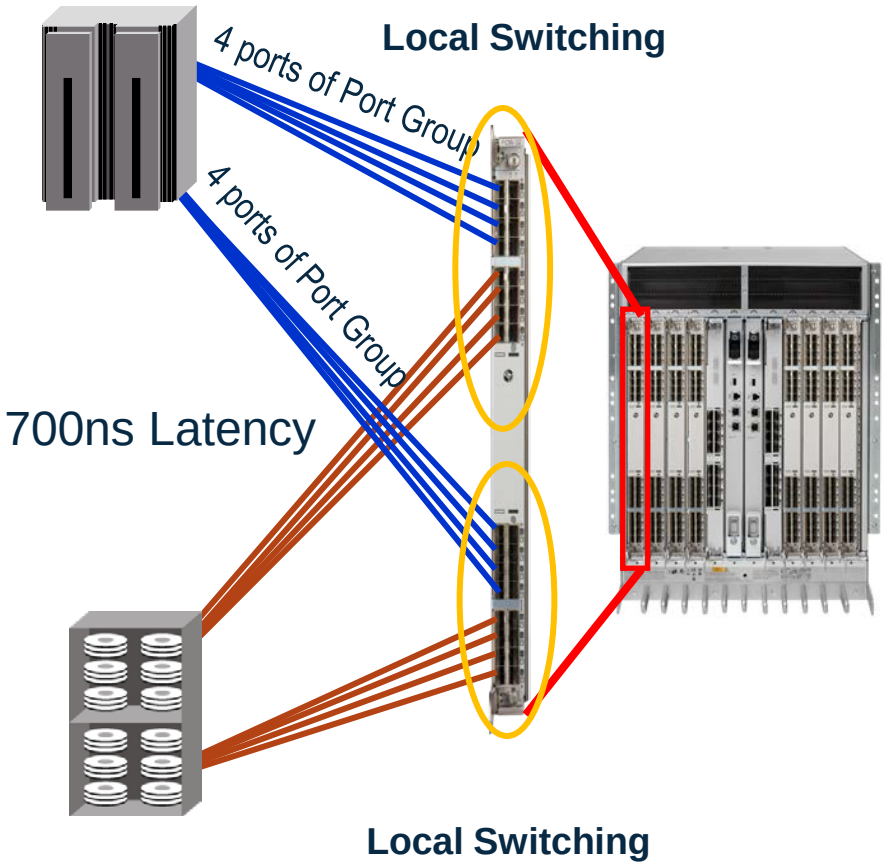
Central Switching = Excellent Response Time



Not a bad strategy…
Excellent resiliency and availability…
Just not the best performance-oriented strategy!

- A popular deployment strategy is to take a mainframe Path Group and spread it out across 8 blades – one path of the path group per blade
- In the event of a blade or port failure only 1/8th of the path group, and therefore the bandwidth, would be affected
- Spreading out the storage connections across all the blades in a chassis also minimizes the bandwidth loss in the event of a blade or port failure
- I/O would have a consistent and predictable 2.1 µs latency but this deployment precludes local switching from being utilized!

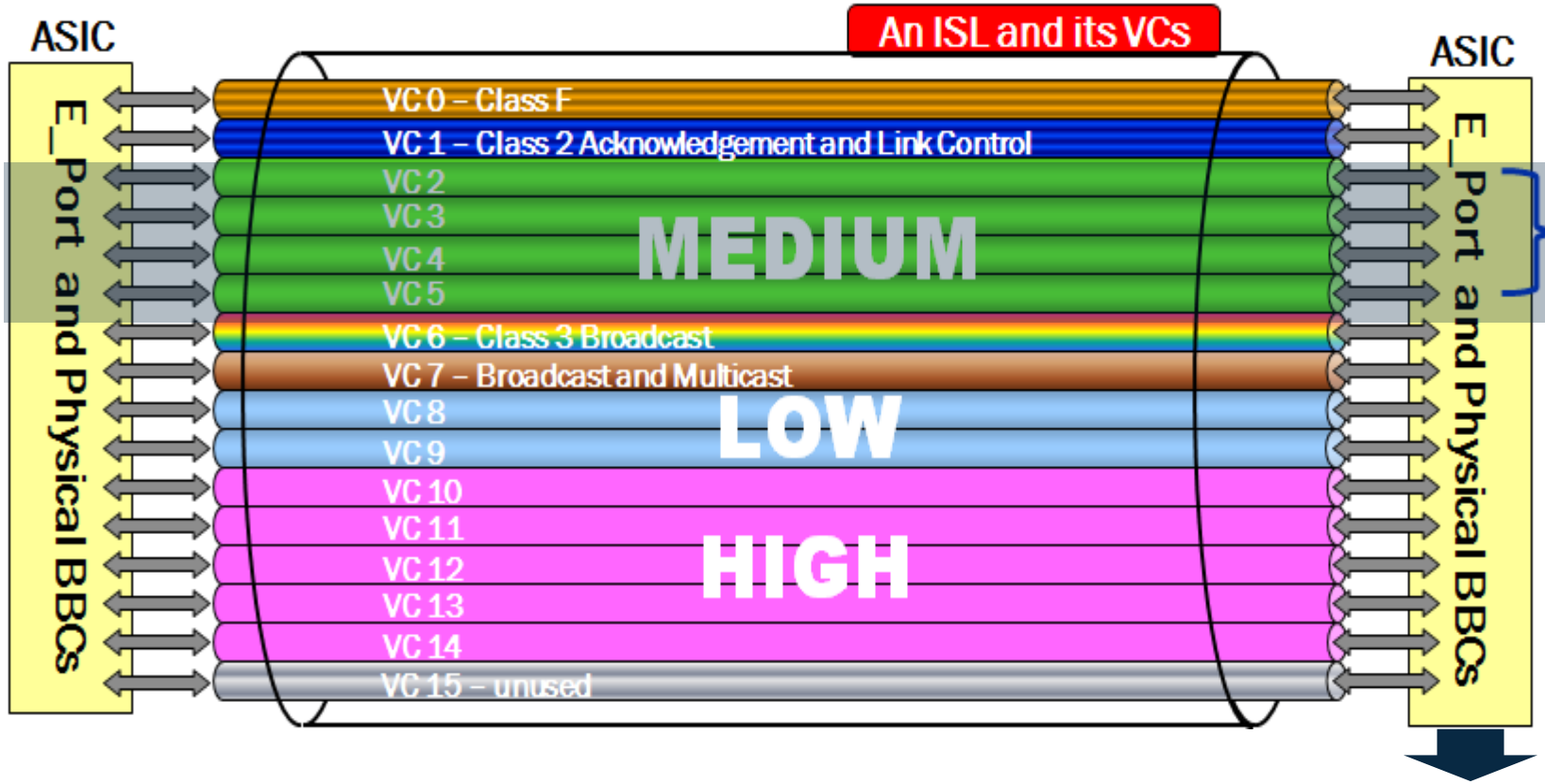
Local Switching = Best Response Time

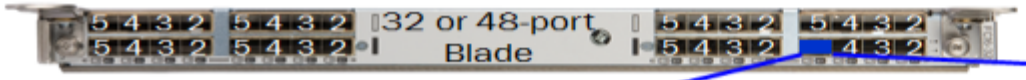


- Consider using Local Switching which means putting the ingress and egress ports are on the same ASIC of a Director's blade
- To minimize overall disruption if a blade were to fail, some customers are placing the ingress ports with their egress ports, on the same ASIC
- If the ASIC or physical port fails then that chosen set of CHPIDs and storage ports all become disabled as a group
- Not all, but a high percentage of the I/O, would probably use Local Switching in this configuration
- One of many strategies for deploying a FICON fabric

Virtual Channels (VCs) Avoid Head-of-Line Blocking

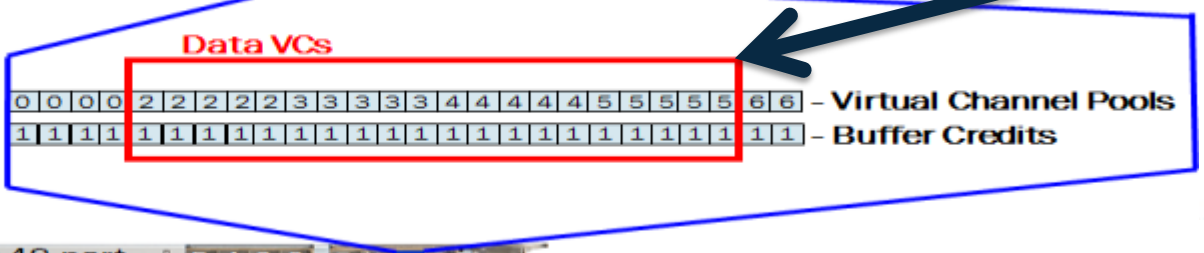
Improves I/O Response Time on ISLs and across the fabric

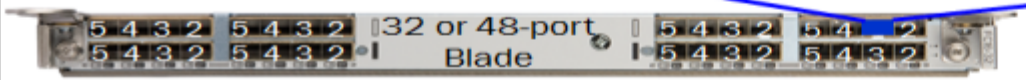




Of the total physical BBCs VC2-5 each get 5 BBCs

FC physical E_Port on both ends of the connection



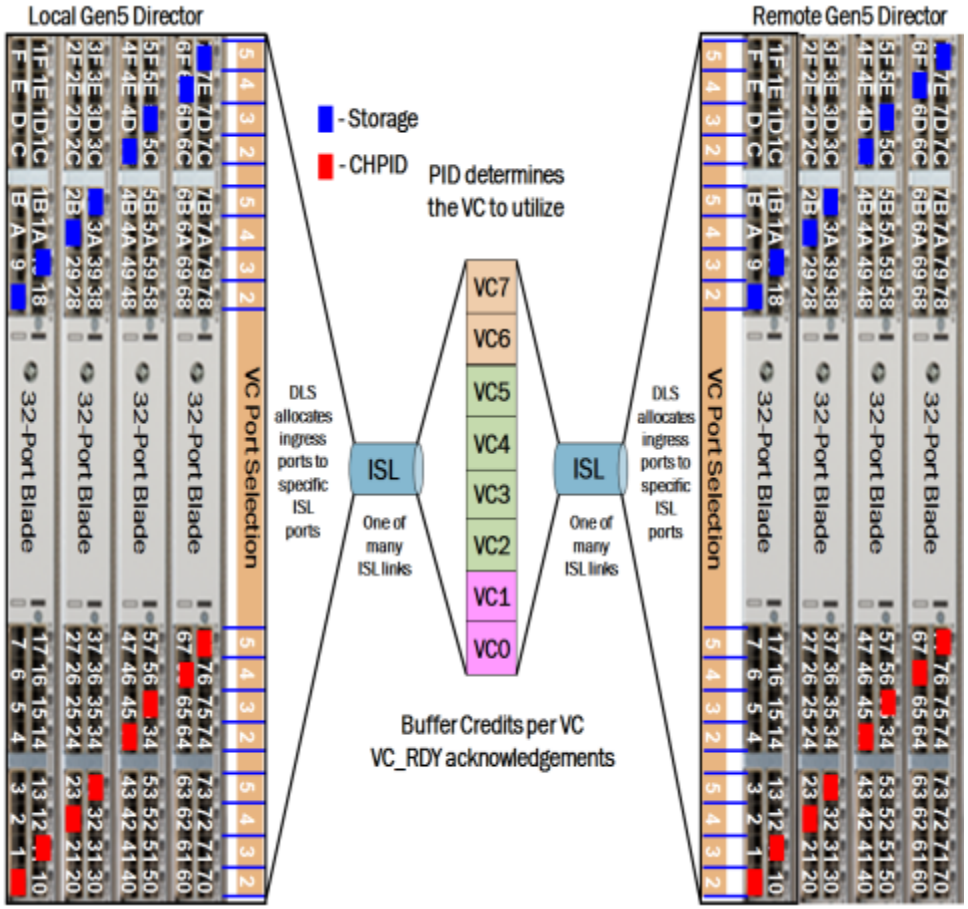


Physical BBCs are virtualized into small pools!

ISL Link Virtual Channels (VC)

Reduced Head-of-Line Blocking on ISL Links

- Reduced Head-of-Line Blocking (HoLB) on ISLs maximizes total switching fabric performance
- Can minimize HoLB on ISL links to fabric ports on a switch since 2002
- Virtual Channels can be thought of as several pools of buffer credits dedicated to an ISL port
- Each of these buffer credit pools (virtual channels) services unique F_Ports on each connectivity blade



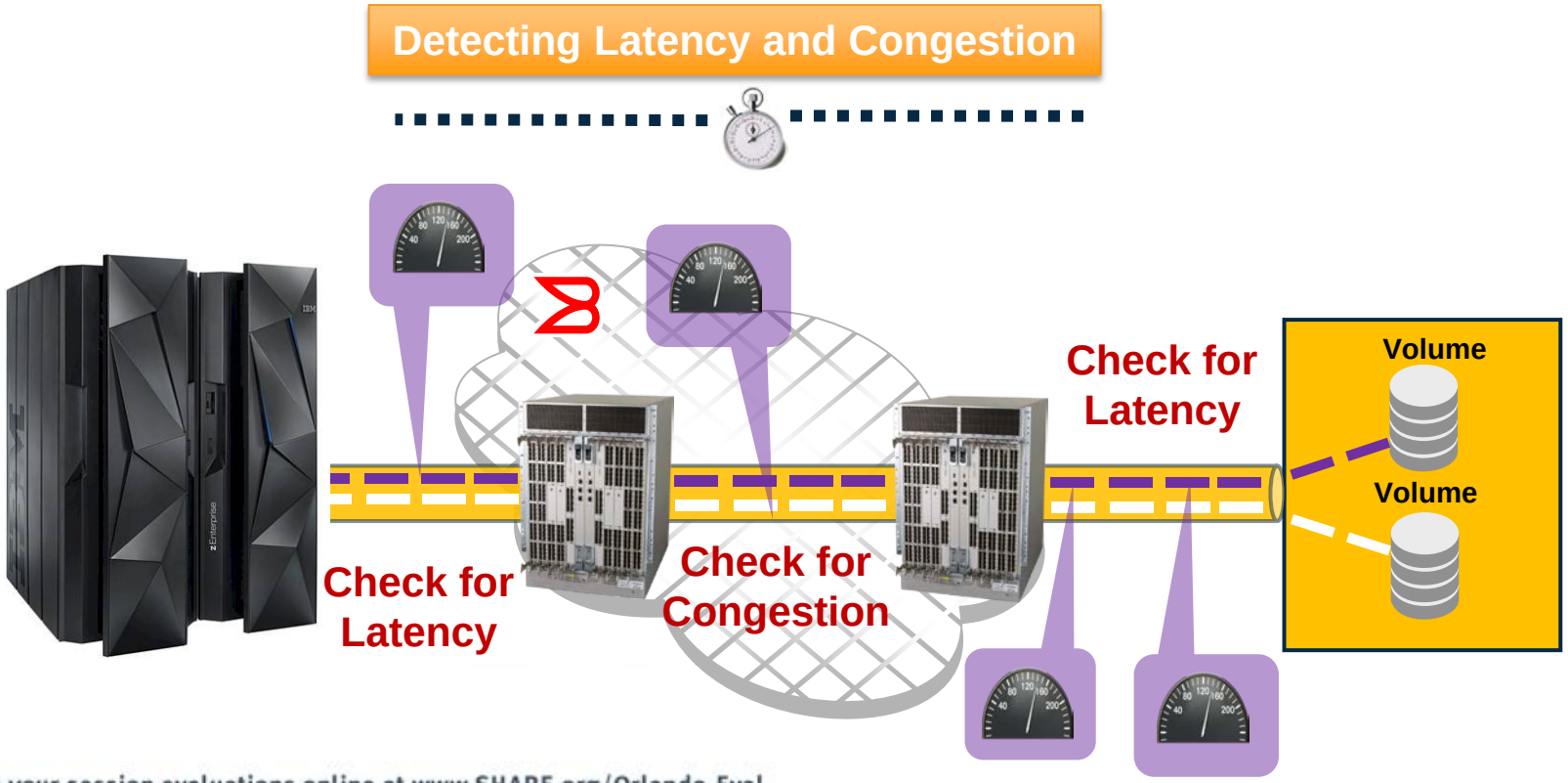
Instead of placing all CHPIDs at one end of the chassis and Storage at the other, make use of VCs by staggering their placement diagonally across the blades.

Must Maintain High Performance

Bottleneck Detection

Quickly Identify and Resolve FICON Performance Degradation

- There must be some way to monitor for latency and congestion in the I/O fabric and receive notification when problems are detected
- Resource contention, congestion, and other issues can impact user application performance and FICON performance
- Alert the user when thresholds are exceeded help with troubleshooting

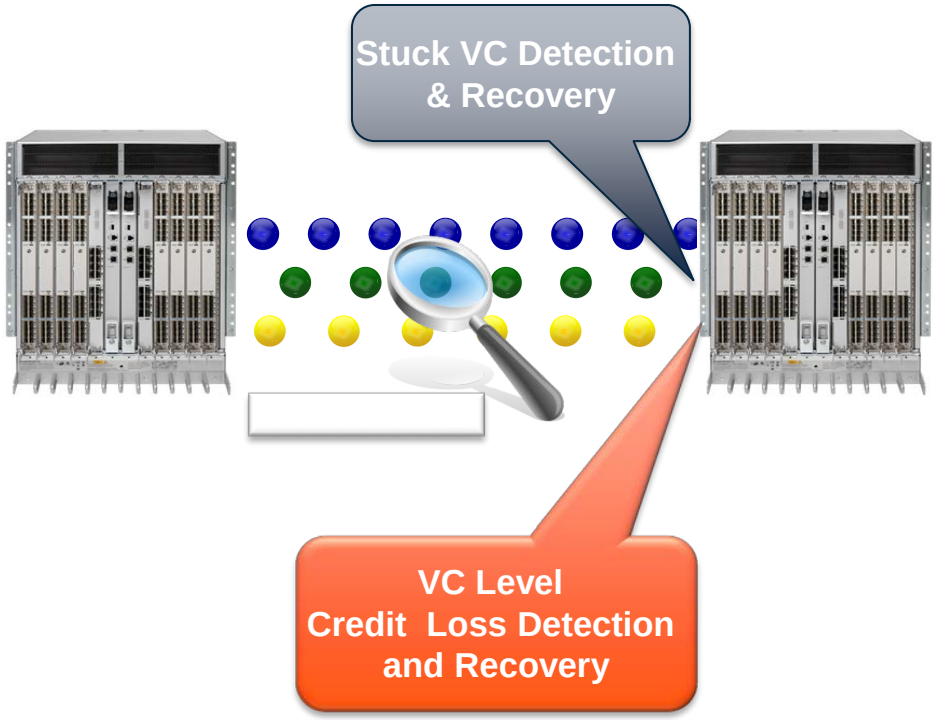


Automatic Buffer Credit Recovery

Stabilize I/O response time by maximizing link performance

- Losing acknowledgements will lead to frames never being discarded from the port's BC memory and never purged.
- Buffer credit loss can lead to performance degradation, hence early detection and recovery is essential
- Buffer credit loss should be detected and recovered automatically at the Virtual Channel (VC) level of ISLs
- Loss of single or multiple BCs and Stuck VC (no credits remain) should recover automatically

Link bit errors can corrupt acknowledgements just as they can corrupt frames.



Switched-FICON Lets You Use CUP

Control Unit Port Is A Great Tool for FICON

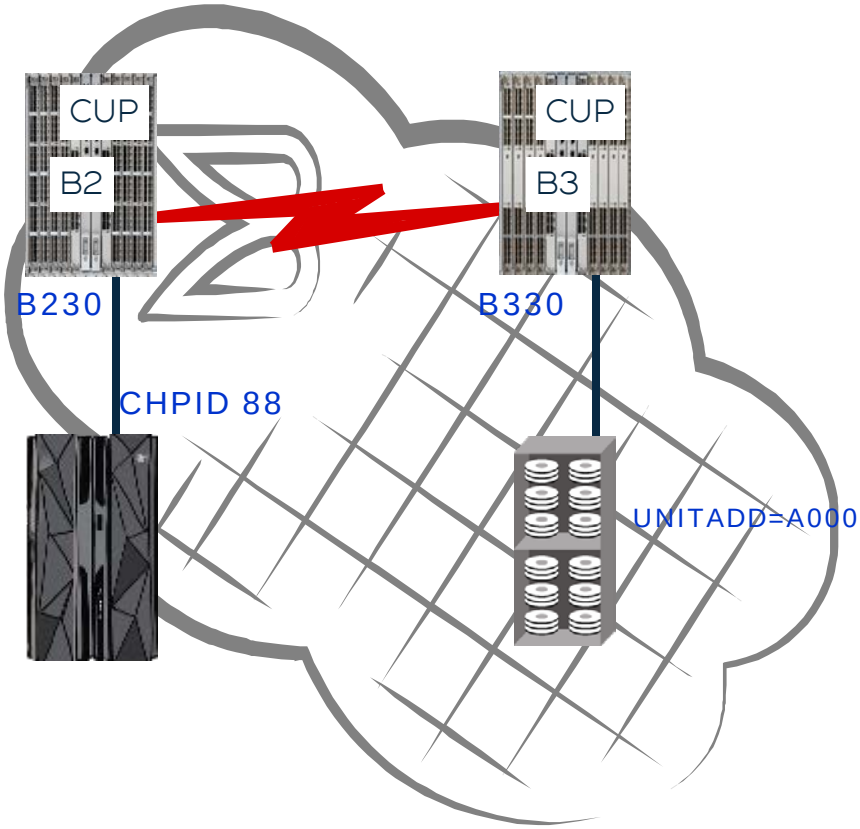
- Control Unit Port code is part of a switches firmware:
  - Used to provide FICON Director Reports in RMF (buffer credits, frame size, etc.)
  - Used for in-band communication with Systems Automation
  - Supports Prohibit Dynamic Connectivity Mask (PDCMs)
  - Supports Dynamic Channel Path Management (DCM)
  - Provides Service Information Messages (SIM) to the z/OS console for FICON device hardware failures
  - Provides CUP Diagnostics support
  - Supports IBM's z/OS Health Checker

Lets look at these two and see how they help maintain High Performance!

CUP Diagnostics

z/OS Display Matrix Command is enhanced

- ROUTE=TODEV
  - Determine the path through the fabric from the channel to the device
- ROUTE=TODEV,HEALTH
  - Adds SFP power levels, transmit/receive utilization statistics, and error counts to the report
- ROUTE=FROMDEV
  - Same as TODEV but from the device to the channel
- ROUTE=FROMDEV,HEALTH



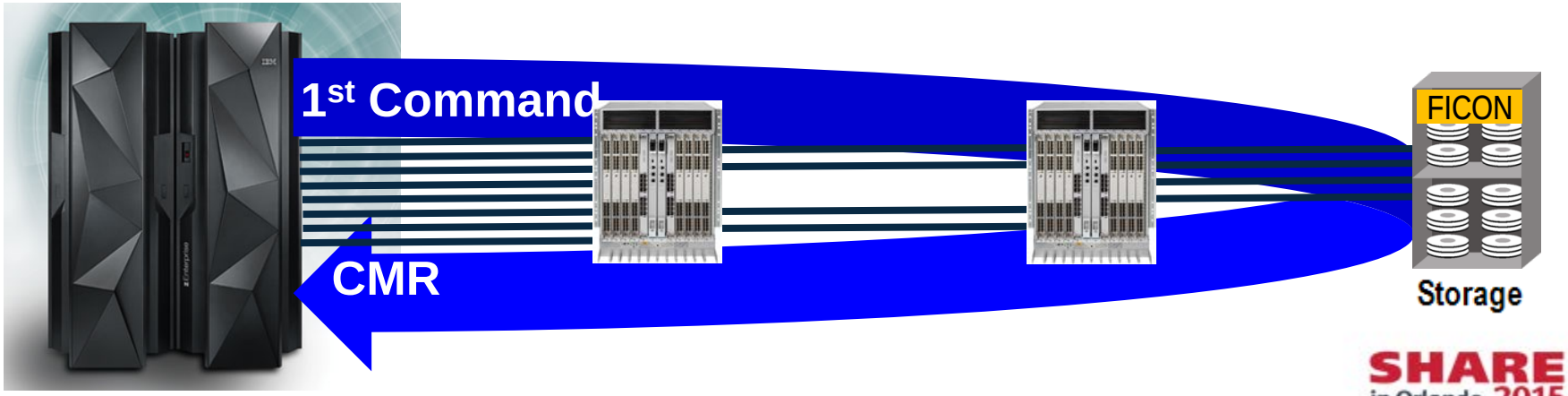
D M=DEV(A000,(88)),ROUTE=TODEV,HEALTH

Show the route that frames take when sent from CHPID 88 to Control Unit A000 Include the ISL link for that transfer and also give us SFP health information.

Measuring FICON Response Time using CMR

Used by z System Health Checker (via CUP Diagnostics)

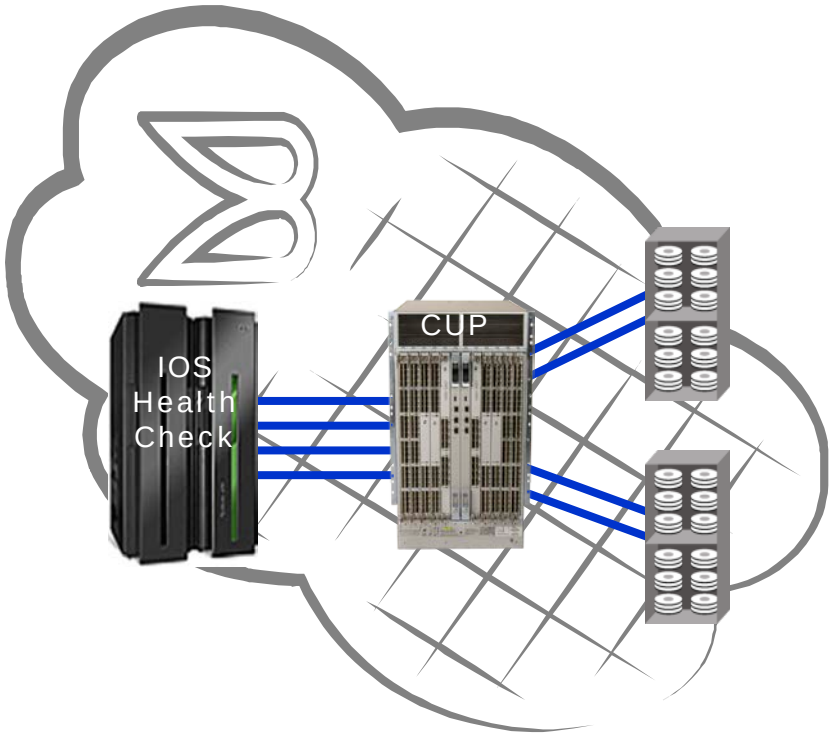
- Initial command response (CMR) time is the amount of time from when the channel sends the first command until it gets a response from the control unit
- First round trip through the fabric – a benchmark is set as the fabric first begins functioning
- AVG CMR (command response) Delay - This delay indicates the time between a successful I/O operation being initiated on the channel and the device acknowledging the connection. If delays are detected, they could be caused by contention in the fabric and/or at the ingress/egress port.
- Extreme switch latency time, ISL congestion, backpressure, slow draining devices and other situations can cause an abnormally high value for AVG CMR DELAY and is reported by RMF.



z/OS Health Checker

Points of Failure and Flow Analysis

- Leverages CUP Diagnostics
- IOS can find unacceptable fabric I/O service times that slow down response time by monitoring CMR Delay:
  - RMF Device Activity reports with high average service times
  - RMF Queuing reports with high “initial command response” time
- z/OS Health Checker tests Single Points of Failure:
  - Analyze connections and paths
  - Identify common components
- zHealth Checker can provide a Flow Description:
  - Identify fabric routes
  - Examine utilization statistics
  - Assess performance and errors



Detecting inconsistent Command Response Time (CMR)

Recent update will be available 30 September 2015

I/O Flow Monitoring

It is easier to determine What the Problem is versus Who is causing the Problem

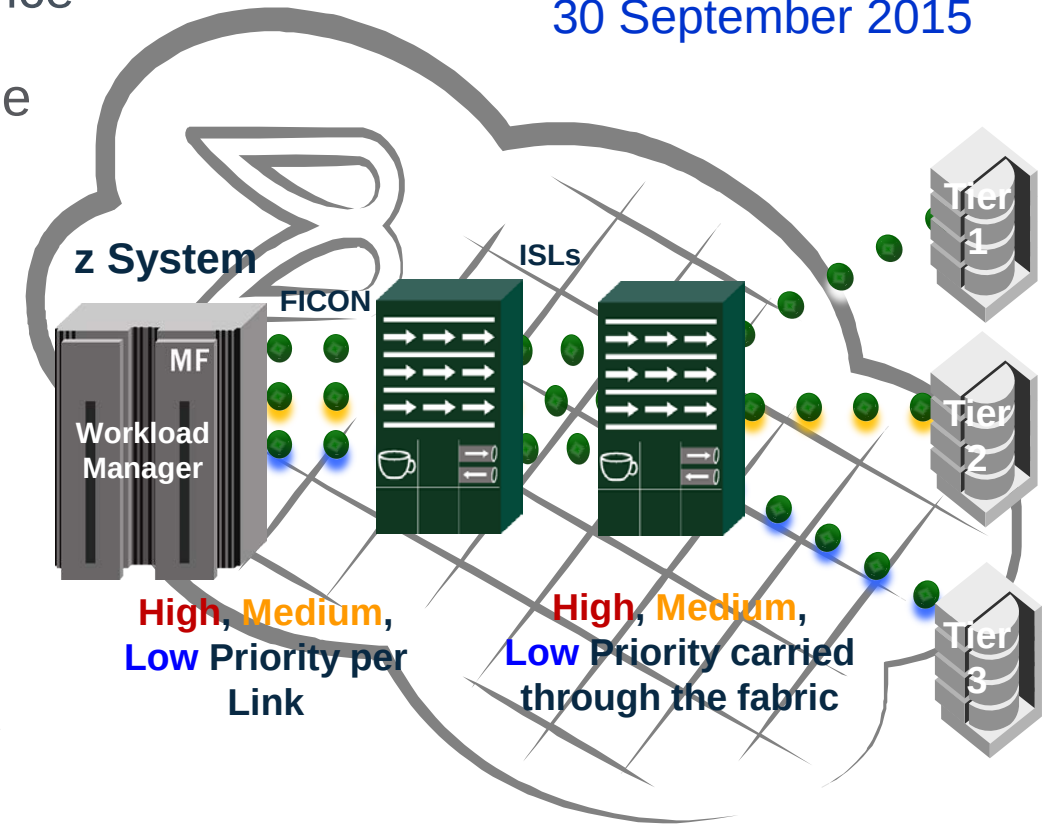
- An ability to automatically configure a Flow Monitor for all paths to a given port can be a valuable troubleshooting tool to determine where traffic to a port originates.
- Flow Monitoring automatically creates a list of **ALL** flows (SID/DID) to a given port so performance at a specific port can be analyzed by channel path to see the fabric SID/DID pairs (e.g. the “WHO”) that are using that port.
- Can then analyze a **SPECIFIC** SID-DID pair if they are suspected of not playing well in the fabric
- For example, an IOCP mistake resulting in two CHPIDs, from different CECs, contending for the same tape port can easily be identified (The WHO that is causing the problem) by setting up a Flow Monitor on the tape port to monitor all flows.
- A single mouse click should bring the user right to the port information with the RNID data that articulates the CEC serial number, CHPID, and channel subsystem (CSS).

SAN Fabric I/O Priority

Application Driven Quality of Service

Availability
30 September 2015

- Translate application importance into Fabric Priority (Quality of Service)
- Persist I/O priority levels across a fabric between host and end devices
- Fabric I/O Priority (QoS) is preserved in each frame
- Fabric I/O Priority works in conjunction with z/OS workload manager in Goal Mode
- Three levels of prioritization:
  - Prioritize the frames between a host and target with a high, medium, or low priority in Goal Mode



IBM Provides this on FICON Express16S and DS8870 16G Links Across FICON Switches

My Next Presentation:

A Deeper Look into the Inner Workings and Hidden Mechanisms of FICON Performance

Monday August 10, 2015 -- 11:15am to 12:15am – Session 17506

Please Fill Out Your Online Evaluations!!

Thank You For Attending Today!

QR Code

This was session:

17505

And Please Indicate in the evaluation if there are other presentations you would like to see us present in this track at SHARE!

Your Eval	My Reaction!
5	"Aw shucks. Thanks!"
4	"Mighty kind of you!"
3	"Glad you enjoyed this!"
2	"A Few Golden Nuggets!"
1	"You Got a nice nap!"

Monday August 10, 2015 -- 10:00am to 11:00am -- Session 17505