17118: z/OS Resilience Enhancements

Harry Yudenfriend
IBM Fellow, Storage Development
March 3, 2015

#SHAREorg

SHARE is an independent volunteer-run information technology association that provides education, professional networking and industry influence.

Agenda

- Review strategy for improving SAN resilience
- Details of z/OS 1.13 Enhancements
- Enhancements for 2014

Session 16896: IBM z13 and DS8870 I/O Innovations for z Systems for additional information on z/OS resilience enhancements.

SAN Resilience Strategy

- Clients should consider exploiting technologies to improve SAN resilience
  - Quick detection of errors
  - First failure data capture
  - Automatically identify route cause and failing component
  - Minimize impact on production work by fencing the failing resources quickly
  - Prevent errors from impacting production work by identifying problem links

FICON Enterprise QoS

- Tight timeouts for quicker recognition of lost frames and more responsive recovery
- Differentiate between lost frames and congestion
- Explicit notification of SAN fabric issues
- End-to-End Data integrity checking transparent to applications and middleware
- In-band instrumentation to enable
  - SAN health checks
  - Smart channel path selection
  - Work Load Management
  - Autonomic identification of faulty SAN components with Purge-Path-Extended
  - Capacity planning
  - Problem determination
  - Identification of Single Points of Failure
  - Real time configuration checking with reset event and self description
- No partially updated records in the presence of a failure
- Proactive identification of faulty links via Link Incident Reporting
- Integrated management software for safe switching for z/OS and storage
- High Integrity, High Security Fabric

System z Technology Summary

- Pre-2013 Items
  a) IOS Recovery option: RECOVERY,LIMITED_RECTIME
  b) z/OS 1.13 I/O error thresholds and recovery aggregation
  c) z/OS message to identify failing link based on LESB data
  d) HCD Generation of CONFIGxx member with D M=CONFIG(xx)
  e) Purge Path Extended (LESB data to SYS1.LOGREC)
  f) zHPF Read Exchange Concise
  g) CMR time to differentiate between congestion vs. lost frame to identify route cause of IFCC
  h) EC12 Channel Path Selection Algorithm
  i) Flapping Link Threshold
  j) DASD ERP Processing
  k) zHPF improves FFDC and better handles work load spikes
  l) Switch vendor error thresholds on ISLs
  m) HMC FC Analyzer Tool
  n) CUIR, ESCON Manager
- June 2013 – YE 2014
  n) Port decommissioning
  o) CUP diagnostic command
  p) Switch invoked CUP health check
  q) Switch ISL error thresholds and isolation policy
  r) Non-disruptive DASD state save
  s) I/O Timing for Tape
  t) System z I/O Exerciser
- z13 Leadership with IBM Storage
  t) FICON Dynamic Routing
  u) Fabric Priority extends WLM into the SAN
  v) FICON Express16s
  w) Forward Error Correction Codes
  x) Read Diagnostic Parameters
  y) CUP policy extensions for health check

I/O Timing for Tape

I/O Timing Facility - Overview

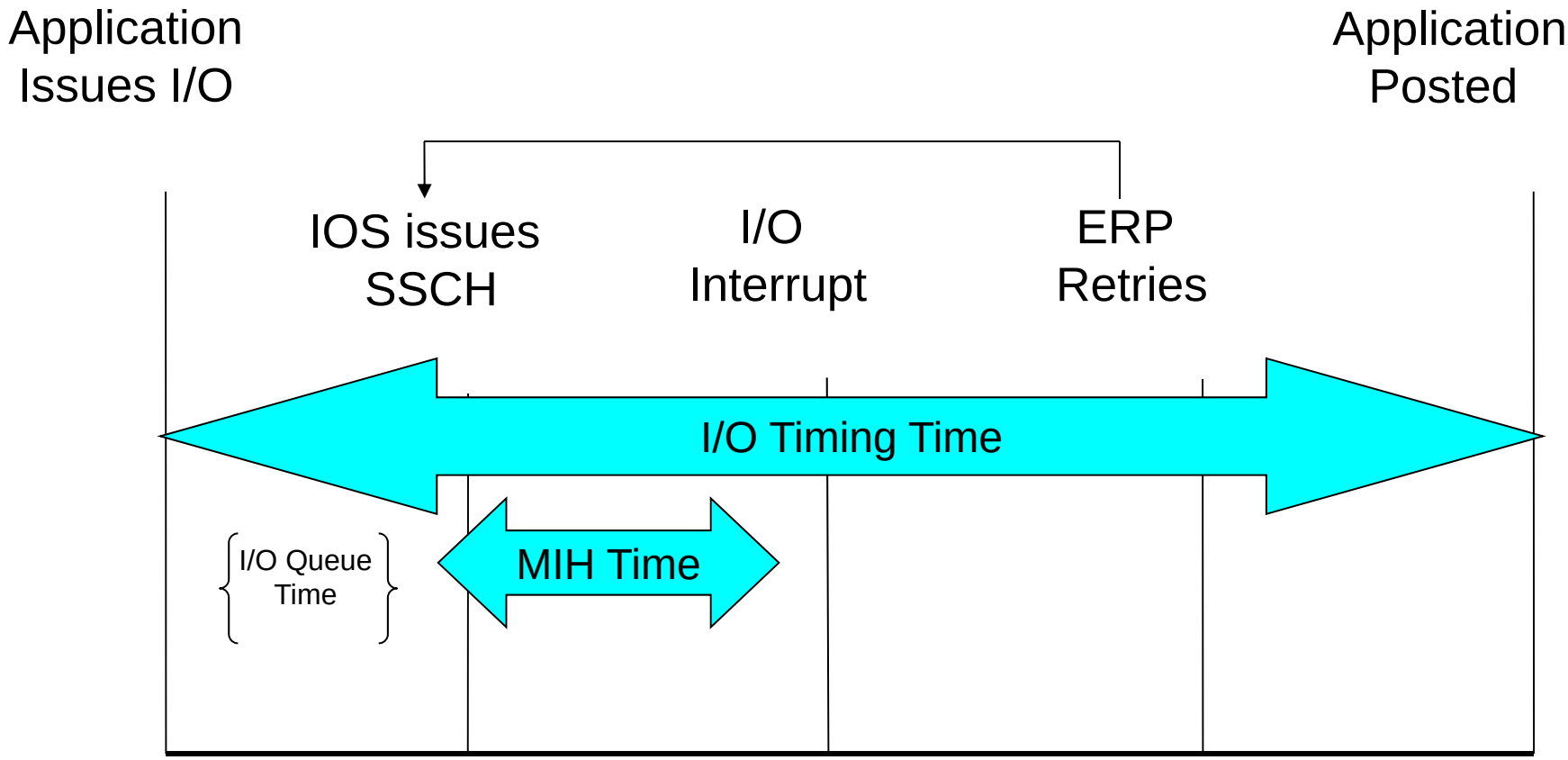


Figure 1 - MIH vs. I/O Timing

I/O Timing Facility - Overview

- Allows installations to limit how long an I/O request can remain in the system before being terminated
  - Message IOS078I for Active I/O requests (Console)
  - Message IOS079I for Queued I/O requests (Syslog)
  - MIH LOGREC Record
  - Application posted with permanent error (IOSCOD = x'53')
- Times the entire I/O request (MVS queue time, active time and ERP retries)
  - In contrast, MIH times ONLY active time
- I/O Timing time value is a device's RESPONSE time
  - Data mapped in SMF-72 records
- OA43674 provides support for tape, with MSGONLY option

Resilience with Smarter Fabrics

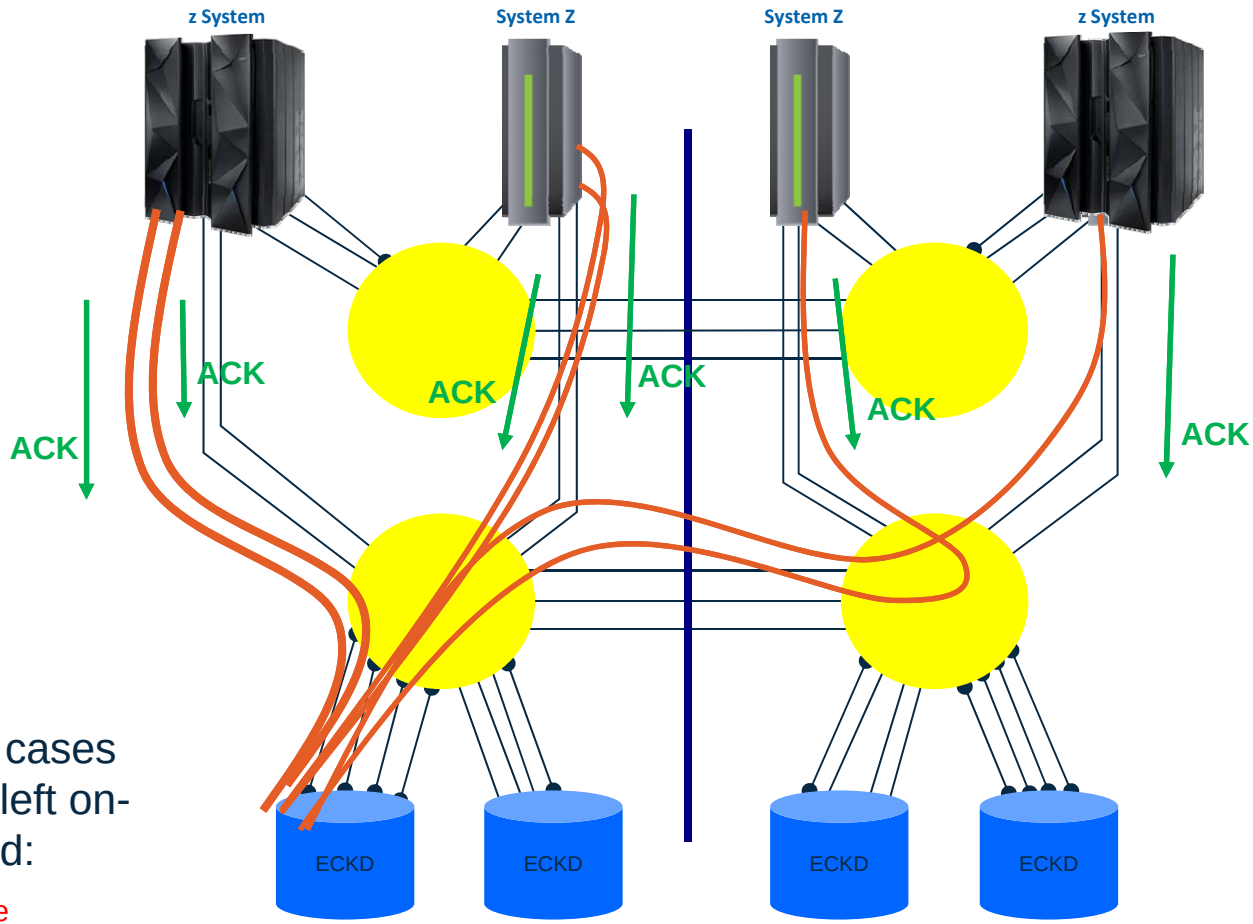
Control Unit Initiated Reconfiguration (CUIR)

CUIR helps insure that the control unit ports being serviced are the ones that have been quiesced all the sharing operating system images.

There used to be a device based option to make sure that devices affected by the service action were off-line, not in use by any applications. This option was deprecated.

We are seeing more and more cases were volumes are accidentally left on-line when actions are performed:

- Initialize VTOC – files no longer accessible
- Move VTOC – files remain, but directory has moved
- Copy/Restore Volume – all files replaced

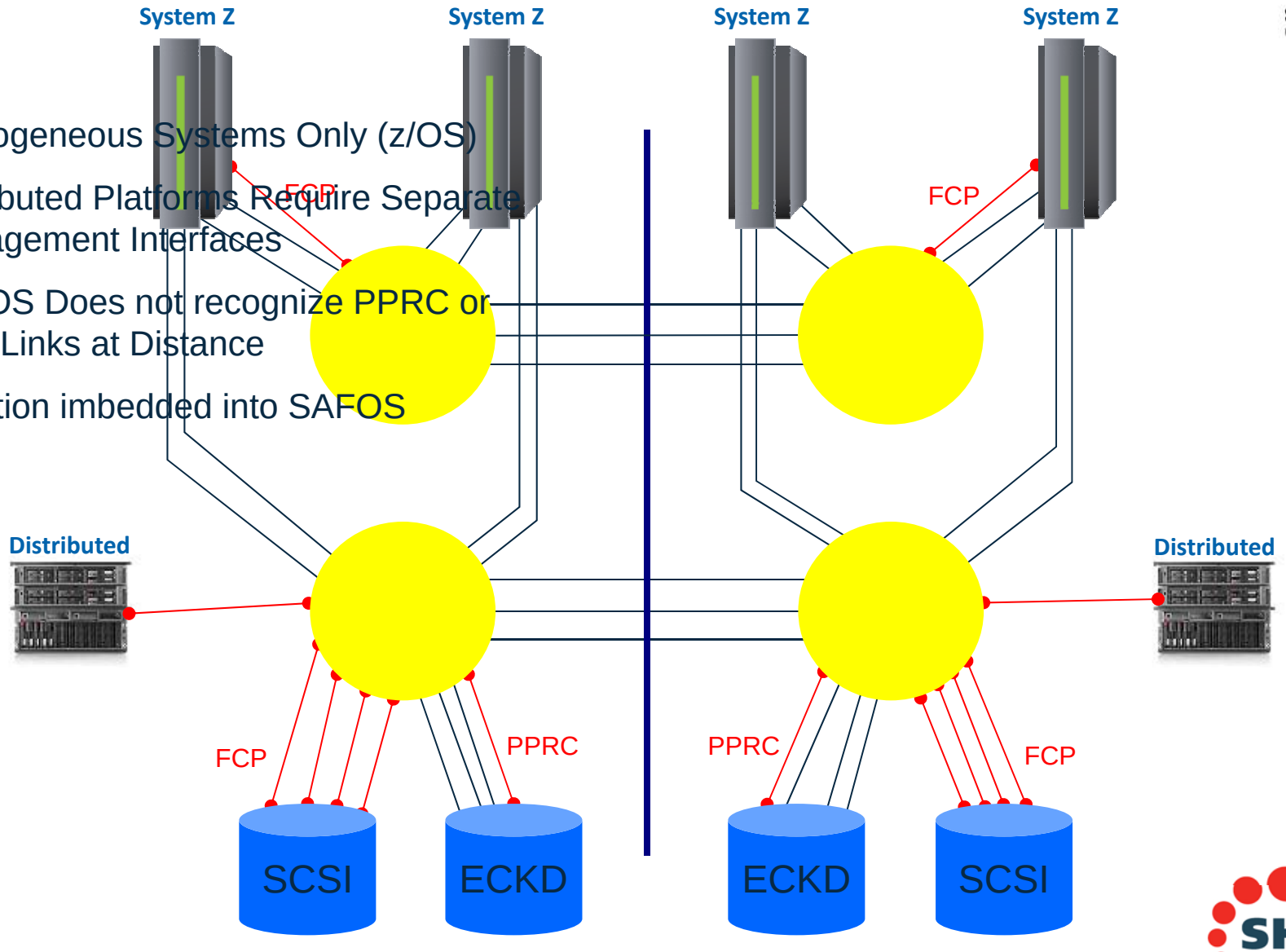


Open data sets may no longer be valid, causing data loss

Applications may have in-memory information about the location of the volume, VTOC overlay can occur

Safe Switching via SAFOS IOOPS (aka ESCM)

- Homogeneous Systems Only (z/OS)
- Distributed Platforms Require Separate Management Interfaces
- SAFOS Does not recognize PPRC or XRC Links at Distance
- Function imbedded into SAFOS

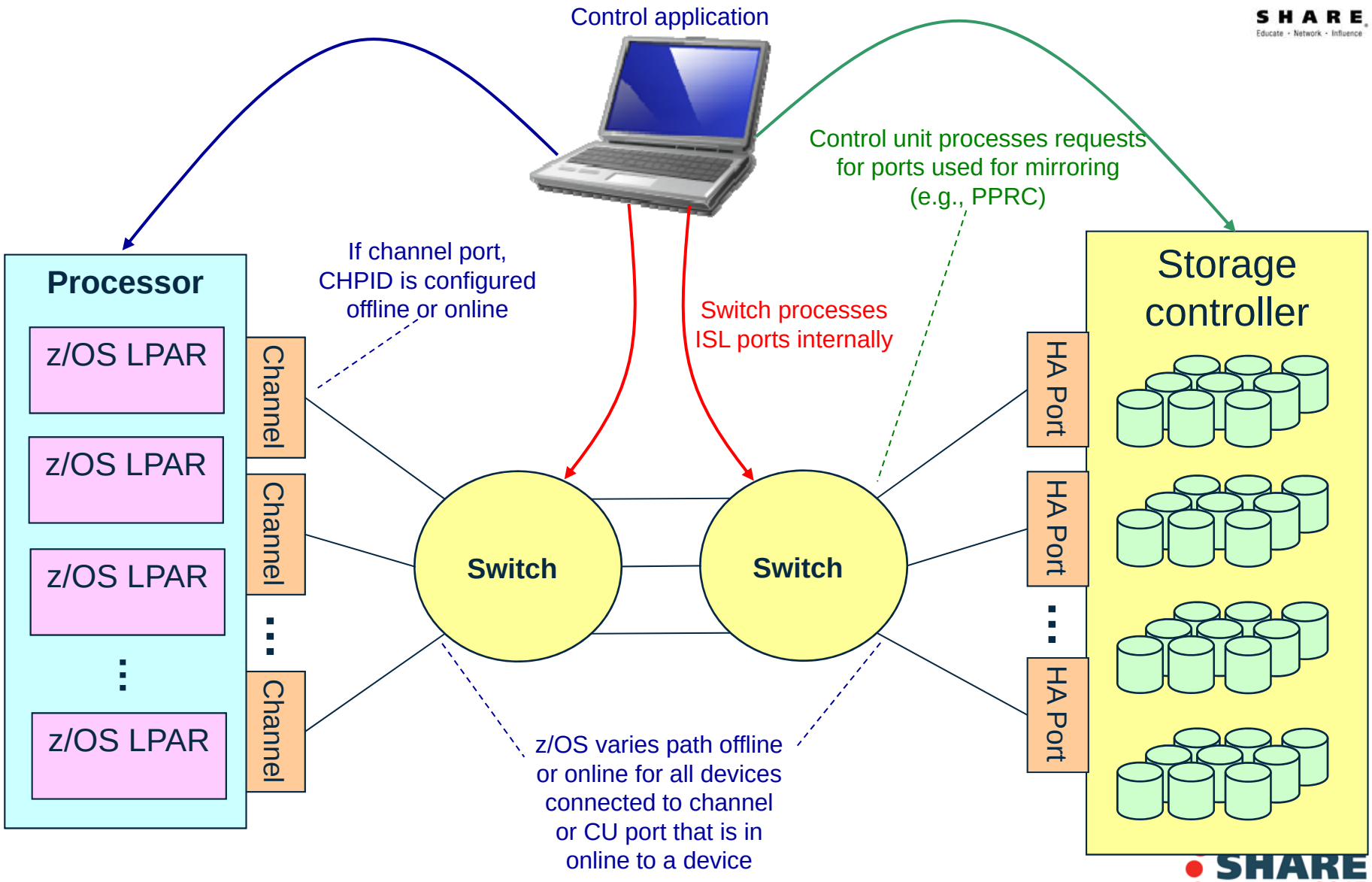


Switch Port Decommissioning

Purpose

- Automate the removal of switch ports in a fabric from use
  - Eliminate application disruption
  - Eliminate manual actions and prevent user errors
  - Similar to control unit initiated reconfiguration on disk
- Decommission - coordinate the removal of ports with users of those ports
  - Operating systems such as z/OS for FICON ports
  - Storage controllers – e.g., for PPRC ports
  - ISL ports handled internally
- Recommission – re-enable a port for use in a fabric

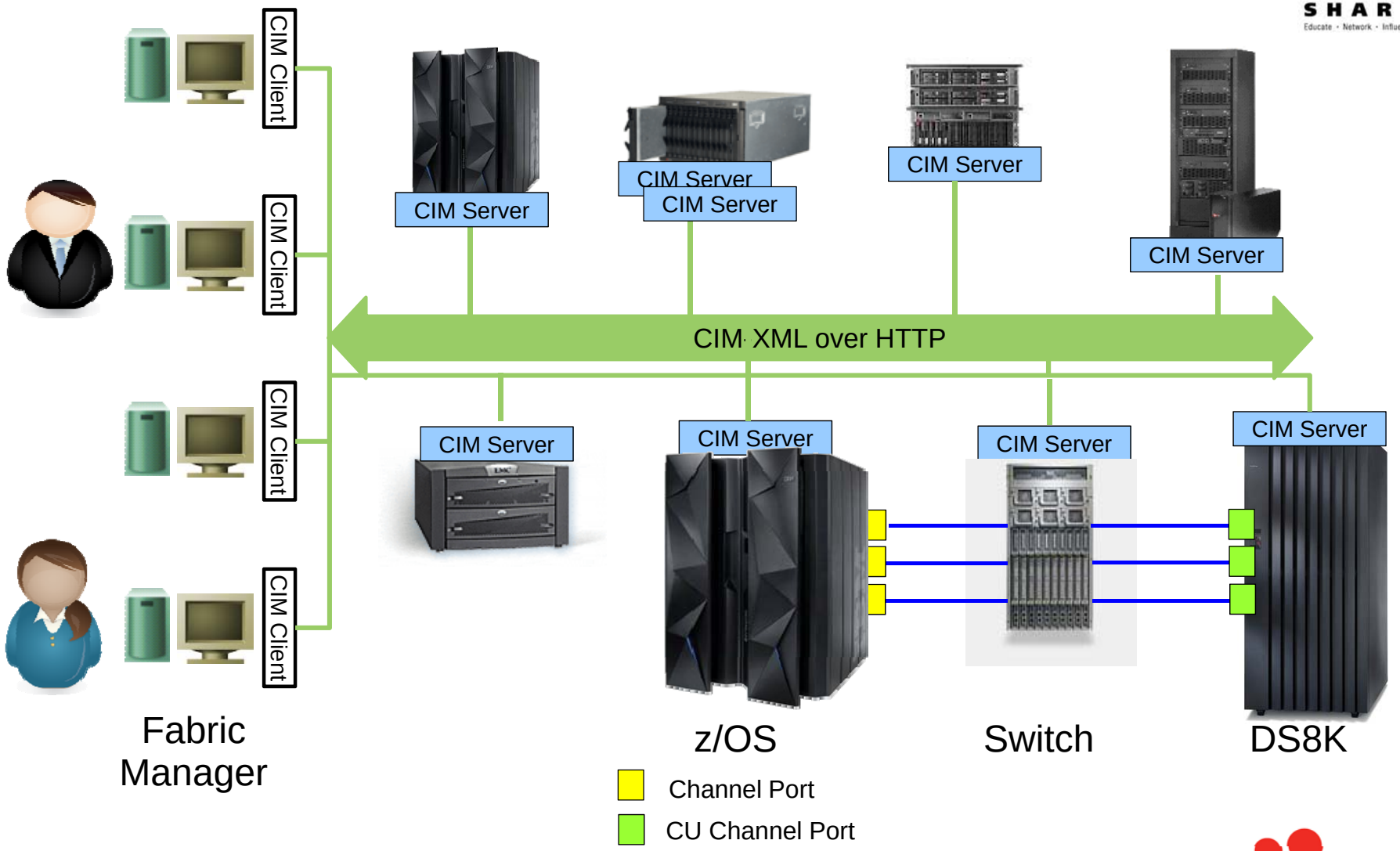
Switch Port Decommissioning



Port Decommissioning Design

- Standards Based Approach –
  The “Storage Management Initiative Specification” (SMI-S) has been updated to provide a “Control Unit Initiated Reconfiguration” type of function
  - Fabric management function notifies each CIM server of requested change
  - Server responses are aggregated by the fabric manager
- Benefits of this design:
  - Solution will work for FICON or FCP (Fibre Channel Protocol) fabrics
    - Including heterogeneous fabrics
  - Customers need to learn only one product to manage heterogeneous fabric
  - Solution will support:
    - Host to storage links
    - Inter-Switch Links (ISL)
    - Peer-to-Peer Remote Copy Links (PPRC)
  - Solution will support any operating system
  - The “orchestrating application” may be provided by any software vendor
- z/OS will delivered exploitation in 2H2013

General Topology – Port Decommissioning



z/OS Resilience Enhancements z/OS 1.9

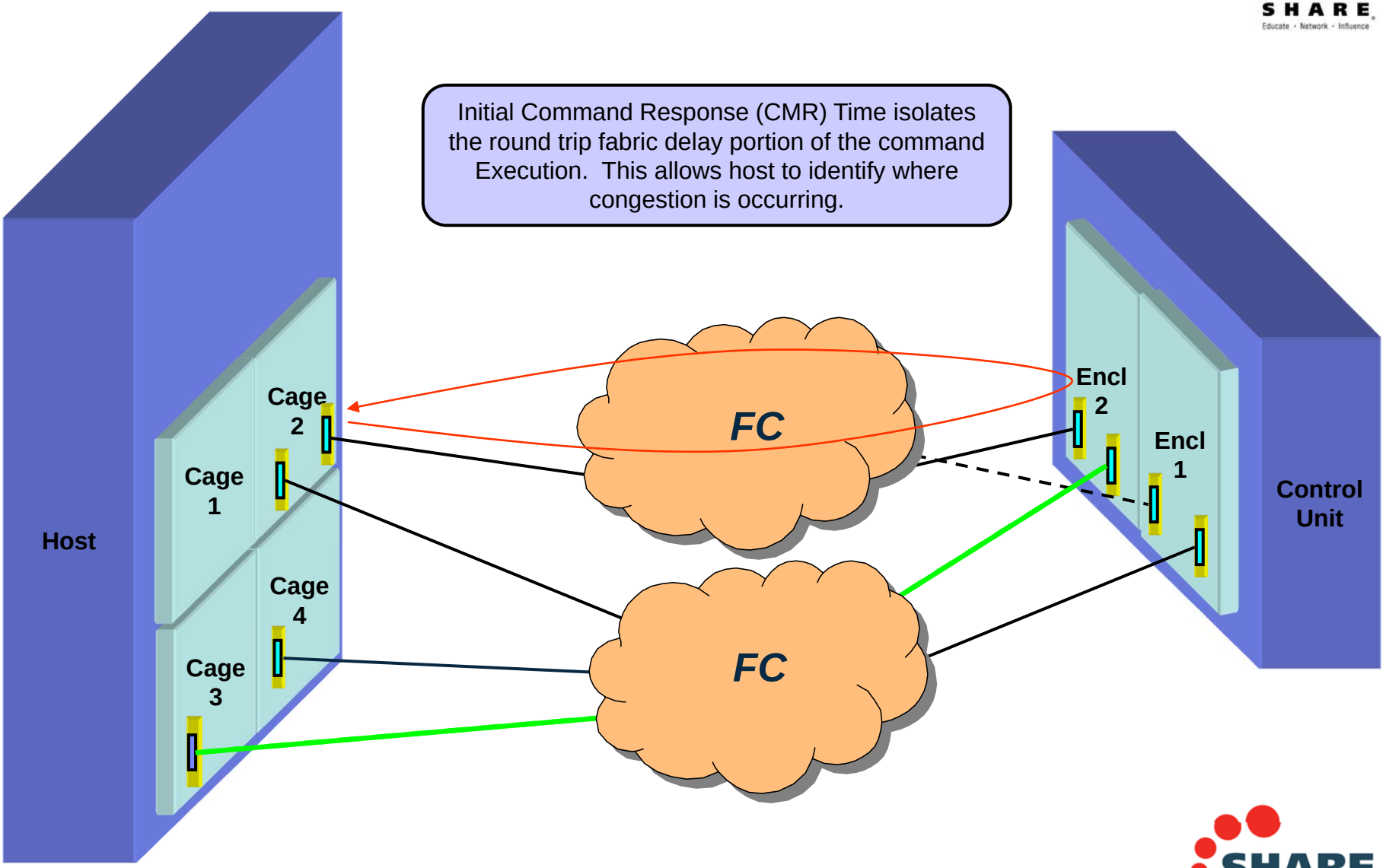
CMR ‘Health Check’

- Problem
  - Fabric issues have resulted in unacceptable I/O service times
    - RMF device activity reports show average service times to be higher than normal
    - I/O queuing reports show abnormally high “initial command response” times on a subset of the paths to a device.
      - Abnormal meaning >>5x
    - No single root cause has been identified
      - ISL failures, ucode failures, CU port congestion, CU HBA utilization,
      - CU failures, wrong laser type, ports initialize at the wrong link speed, WDM,
      - Incorrect number of buffer credits for the distance, etc.

CMR 'Health Check' (OA33367)

- Solution
  - IOS SAN health checking function to provide real time detection of disparate initial command response (CMR) times, a symptom of fabric I/O
  - Notify client when the condition is detected
  - Improved first failure data capture (FFDC)
  - EC12 Channel Path Selection Algorithms + New Health Check (OA40548)
  - Improved First Failure Data Capture
    - Fabric diagnostic commands to provide improved first failure data capture from the CUP
      - Vendor specific non-disruptive state save
      - z/OS SYS1.LOGREC diagnostic record
    - Compliment with CU diagnostic commands for FFDC (initial support for DASD)

SAN CMR Health Check



Symptoms of Fabric Congestion Issues – CMR Time

z/OS V1R8 SYSTEM ID SYD1 DATE 09/17/2009 INTERVAL 09.59.990
RPT VERSION V1R8 RMF TIME 20.10.00 CYCLE 1.000 SECONDS
-TOTAL SAMPLES = 600 IODF = A1 CR-DATE: 09/16/2009 CR-TIME: 15.57.12 ACT: ACTIVATE
-

LCU	CU	DCM GROUP MIN MAX DEF	CHAN PATHS	CHPID TAKEN	% DP BUSY	% CU BUSY	AVG CUB DLY	AVG CMR DLY	CONTENTION RATE	DELAY Q LNGTH	AVG CSS DLY	HPAV WAIT	HPAV MAX
0 0005	0159		76	0.040	0.00	0.00	0.0	0.0					
			6C	0.040	0.00	0.00	0.0	0.0					
			B4	0.040	0.00	0.00	0.0	0.0					
			47	0.040	0.00	0.00	0.0	0.0					
			*	0.160	0.00	0.00	0.0	0.0	0.000	0.00	0.1		
0 0008	03F0		2B	0.012	0.00	0.00	0.0	6.7					
			76	0.013	0.00	0.00	0.0	0.1					
			36	0.015	0.00	0.00	0.0	0.1					
			6C	0.013	0.00	0.00	0.0	0.3					
			B4	0.012	0.00	0.00	0.0	0.1					
			C6	0.012	0.00	0.00	0.0	0.1					
			46	0.008	0.00	0.00	0.0	3.8					
			47	0.008	0.00	0.00	0.0	0.2					
			*	0.093	0.00	0.00	0.0	1.3	0.000	0.00	0.1		
0 0009	0434		2B	0.007	0.00	0.00	0.0	4.2					
			76	0.007	0.00	0.00	0.0	0.2					
			36	0.005	0.00	0.00	0.0	0.1					
			6C	0.008	0.00	0.00	0.0	0.1					
			B4	0.008	0.00	0.00	0.0	0.1					
			C6	0.010	0.00	0.00	0.0	0.1					
			46	0.007	0.00	0.00	0.0	4.2					
			47	0.005	0.00	0.00	0.0	0.2					
			*	0.057	0.00	0.00	0.0	1.1	0.000	0.00	0.1		
0 000A	0478		**2B**	**0.012**	**0.00**	**0.00**	**0.0**	**1.6**					
			76	0.012	0.00	0.00	0.0	0.2					
			36	0.012	0.00	0.00	0.0	0.2					
			6C	0.012	0.00	0.00	0.0	0.1					
			B4	0.012	0.00	0.00	0.0	0.1					
			C6	**0.012**	**0.00**	**0.00**	**0.0**	**0.5**					
			46	**0.012**	**0.00**	**0.00**	**0.0**	**0.9**					
			47	0.012	0.00	0.00	0.0	0.1					
			*	0.093	0.00	0.00	0.0	0.5	0.000	0.00	0.1		

IOS Real Time CMR Health Check – SPE OA33367

- IOS CMR Health Check Parameters
  Default Values
  - **THRESHOLD(3), RATIO(5), XTYPE(),XCU()**
  - THRESHOLD is in units of 'ms' and indicates that the highest CMR time must be at least that value in addition to an order of magnitude (ratio) higher than the lowest CMR time. Valid values are 0 to 100.
  - RATIO is the value used to determine if the highest CMR time is greater than the lowest CMR time times that value. Valid values are 2 to 100.
  - XTYPE indicates if any device type should be excluded. Valid values are TAPE and DASD.
  - XCU is a list of any specific control unit numbers that should be excluded from the check. If XCU() is specified with no values or the parameter is excluded, the list is deleted and no CUs will be excluded. If XCU(cu1, cu2...) is specified, these new values will overwrite the old values.

IOS Real Time CMR Health Check – SPE OA33367

- IOS CMR Health Check Parameters
  - Modify Command to Change the Defaults:

```
F HZSPROC,UPDATE,CHECK(IBMIOS,IOS_CMRTIME_MONITOR),PARMS('THRESHOLD(t)',
  RATIO(r),XTYPE(devtype),XCU(cu1,cu2,cu3,...,cu40)')
```

  - HZSPRMxx Parmlib Member

```
UPDATE CHECK(IBMIOS,IOS_CMRTIME_MONITOR)
      SEVERITY(HIGH) INTERVAL(2:00) DATE(20100609)
      PARM('THRESHOLD(5),RATIO(5),XTYPE(TAPE),XCU(2000)')
      REASON('Need to change parameters and severity')
```

Or

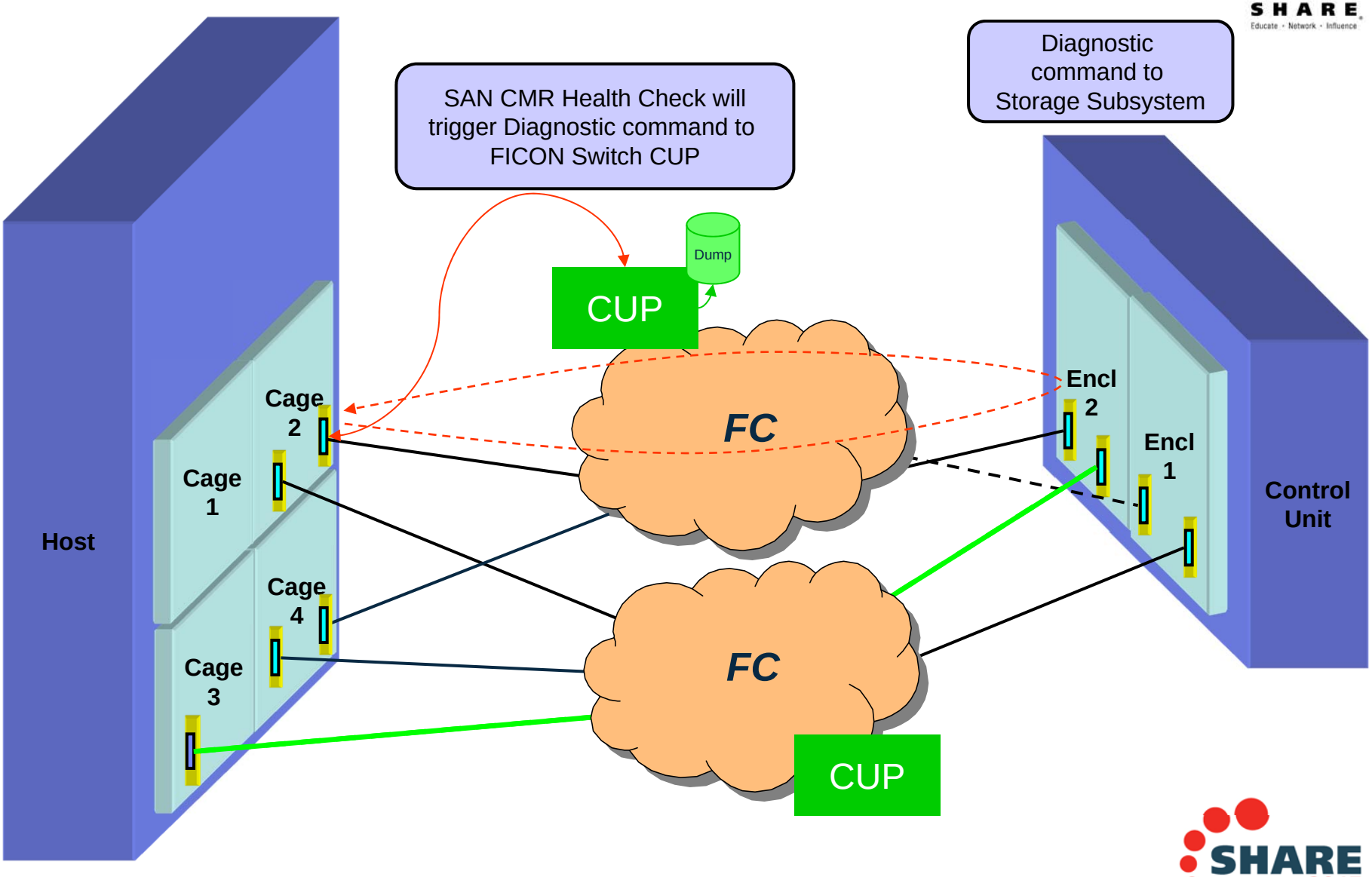
```
F hzsproc,ADD,PARMLIB(xx)
```

IOS CMR Health Check Report


```
CHECK(IBMIOS,IOS_CMRTIME_MONITOR)
START TIME: 10/29/2010 17:34:09.652404
CHECK DATE: 20100501  CHECK SEVERITY: MEDIUM
CHECK PARM: THRESHOLD(3),RATIO(5),XTYPE(),XCU()


IOSHC113I Command Response Time Report

The following control units show inconsistent average command
response (CMR) time based on these parameters:
THRESHOLD = 3
RATIO = 5



CMR TIME EXCEPTION DETECTED AT: 10/29/2010 17:31:18.301627
CONTROL UNIT = 0400
ND = 002107.000.IBM.PK.039F30100404

        ENTRY  EXIT  CU    I/O     AVG
  CHPID  LINK  LINK  INTF  RATE    CMR
   20    3C09  3DD1  0200  5.133   8.832
   10    3C08  3DD5  0230  4.731   2.688
   11    3C0B  3DD7  0240  4.420   1.152
   30    2C55  2DC3  0003  4.285   1.024
   40    2C4E  2DC7  0033  4.020   0.896
   41    2C57  2DCB  0103  3.857   0.896
```

```
* Medium Severity Exception *

IOSHC112E Analysis of command response (CMR) time detected one or
more control units with an exception.

  Explanation:  CHECK(IBMIOS,IOS_CMRTime_Monitor) determined that one or
    more control units has potentially inconsistent command response
    (CMR) time.  Each of the control units had at least one channel path
    with an average CMR time that exceeded the THRESHOLD parameter value
    and was significantly higher (as defined by the RATIO parameter
    value) than the channel path with the lowest average CMR time.  More
    specifically, when the highest average CMR time is greater than a
    multiple of the average CMR time for the path with the lowest
    average CMR time, where the multiple is the defined RATIO value. For
    example, if the RATIO value is 5 then an exception will be declared
    if the highest average CMR time is greater than 5 times the lowest
    average CMR time for a path to the control unit.
    These CMR time exceptions may happen when there is a problem
    somewhere in the fabric.
    Note: If parameters have recently been changed they will not take
    effect until the next monitor interval runs. Therefore, these check
    results may be based on a previous parameter set.

  System Action:  The system continues processing.

  Operator Response:  Contact the System Programmer.

  System Programmer Response:
    Refer to the corresponding reports for this Health Check to
    determine which control units and channel paths have delays. Use
    diagnostic tools (such as RMF or hardware diagnostics) to help
    determine the source of the problem.

  Problem Determination:  n/a
  Source:  n/a
  Reference Documentation:  n/a
  Automation:  n/a
  Check Reason:  Command Response Time Monitor for Channel Paths

END TIME: 10/29/2010 17:34:09.667626  STATUS: EXCEPTION-MED
```


Enhance First Failure Data Capture



EC12 - New Channel Path Selection Algorithm

- New channel path selection algorithm will provide improved throughput and I/O service times when abnormal conditions occur:
  - Abnormal conditions include the following:
    - Multi-system work load spikes
    - Multi-system resource contention in the SAN or at the CU ports
    - SAN congestion
    - Destination port congestion
    - Firmware failures in the SAN, channel extenders, WDMs, control units
    - Hardware failures (link speeds did not initialize correctly)
    - Miss-configuration
    - Cabling Errors
    - Dynamic changes in fabric routes (possible multi-hop cascading)
    - Insufficient number of buffer credits for the distance traversed
- When conditions occur that cause an imbalance in performance (I/O latency/throughput) across a set of channel paths to control unit 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 System z FICON and zHPF protocols and new intelligent algorithms in the channel subsystem to exploit this information
- Results show 50-100% throughput improvements with CU HBA/port congestion
  - Complements DS8870 as premier storage consolidation platform
- New Health Check to identify disparate I/O rates instead of CMR times (OA40548)

EC12 - New Channel Path Selection Algorithm

```
 CPU=  6/  1 UIC= 65K PR=   0         System= S51  Total

21:21:03  I= 40%    DCM Group       Cont Del Q  AVG   CHPID  %DP  %CU   AVG AVG
Path  DCM CTL Units MN MX DEF LCU   Rate Lngth  CSS   Taken  Busy Busy  CUB CMR

5D        3000               010B                   404.96   0.0  0.0  0.0 0.1
50        3000               010B                   404.96   0.0  0.0  0.0 0.1
3D        3000               010B                   404.95   0.0  0.0  0.0 0.0
C6        3000               010B                   404.96   0.0  0.0  0.0 0.1
52        3000               010B                   404.96   0.0  0.0  0.0 0.0
D1        3000               010B                   404.96   0.0  0.0  0.0 6.5
95        3000               010B                   404.96   0.0  0.0  0.0 6.1
C2        3000               010B                   404.19   0.0  0.0  0.0  11
                             010B   0.0  0.00  0.1  3238.9   0.0  0.0  0.0 2.9
5D        3100               010C                   404.57   0.0  0.0  0.0 0.1
50        3100               010C                   404.57   0.0  0.0  0.0 0.1
3D        3100               010C                   404.55   0.0  0.0  0.0 0.0
C6        3100               010C                   404.57   0.0  0.0  0.0 0.1
52        3100               010C                   404.57   0.0  0.0  0.0 0.0
D1        3100               010C                   404.56   0.0  0.0  0.0 6.5
95        3100               010C                   404.56   0.0  0.0  0.0 6.1
C2        3100               010C                   403.86   0.0  0.0  0.0  11
                             010C   0.0  0.00  0.1  3235.8   0.0  0.0  0.0 2.9
```

```
 CPU=  7/  2 UIC= 65K PR=   0         System= S51  Total

22:08:33  I= 57%    DCM Group       Cont Del Q  AVG   CHPID  %DP  %CU   AVG AVG
Path  DCM CTL Units MN MX DEF LCU   Rate Lngth  CSS   Taken  Busy Busy  CUB CMR

5D        3000               010B                   820.56   0.0  0.0  0.0 1.2
50        3000               010B                   757.73   0.0  0.0  0.0 1.3
3D        3000               010B                   998.91   0.0  0.0  0.0 1.1
C6        3000               010B                   913.45   0.0  0.0  0.0 2.0
52        3000               010B                   850.12   0.0  0.0  0.0 1.6
D1        3000               010B                   370.59   0.0  0.0  0.0 1.6
95        3000               010B                   576.98   0.0  0.0  0.0 1.2
C2        3000               010B                   637.72   0.0  0.0  0.0 1.2
                             010B   0.0  0.00  0.2  5926.1   0.0  0.0  0.0 1.4
5D        3100               010C                   485.75   0.0  0.0  0.0 1.3
50        3100               010C                   543.74   0.0  0.0  0.0 1.4
3D        3100               010C                  1395.7    0.0  0.0  0.0 1.1
C6        3100               010C                  1172.8    0.0  0.0  0.0 1.8
52        3100               010C                  1237.1    0.0  0.0  0.0 1.4
D1        3100               010C                   303.98   0.0  0.0  0.0 2.0
95        3100               010C                   296.60   0.0  0.0  0.0 1.5
C2        3100               010C                   246.21   0.0  0.0  0.0 1.3
                             010C   0.0  0.00  0.3  5681.8   0.0  0.0  0.0 1.4
```

Before

Substantial improvement in I/O rates

After

I/O Rate Health Check (OA40508)

```
CHECK(IBMIOS,IOS_IORATE_MONITOR)
START TIME: 04/22/2013 08:44:14.271360
CHECK DATE: 20120430  CHECK SEVERITY: MEDIUM
CHECK PARM: THRESHOLD(100),RATIO(2),XTYPE(),XCU()

IOSHC133I I/O Rate Report

The following control units show inconsistent I/O rates based on these
parameters:
THRESHOLD = 100
RATIO = 2

I/O RATE EXCEPTION DETECTED AT: 04/22/2013 08:44:14.254730
CONTROL UNIT = 0500
ND = 002107.000.IBM.PK.000000000002

         ENTRY EXIT  CU    I/O     AVG     IOR
  CHPID  LINK  LINK  INTF  RATE    CMR     EXC
   14    B153  B177  0001  39.603  2.560    *
   44    B353  B375  0012  101.38  2.112
   16    B055  B277  0013  98.019  2.134
   46    B154  B376  0104  50.693  2.048

•Medium Severity Exception *

IOSHC132E Analysis of I/O rates detected one or more control units with
an exception.
```

I/Os per second to the CU must be at least this high

I/O rate ratio between the highest and lowest paths

This is the exception path

Exception message appears in system log

New IOS Message Health Check message issued

```
SY1  HZS0002E CHECK(IBMIOS,IOS_CMRTIME_MONITOR):
IOSHC112E Analysis of command response (CMR) time
   detected one or more control units with an
   exception.
```

Link Incident Reports

Link Incident Reports

IOS Messages

```
IOS580E LINK DEGRADED REPORTING CHPID=nn
        INCIDENT UNIT  TM=typ/mdl SER=mmmpp-ssssss IF=xxxx
        ATTACHED UNIT  TM=typ/mdl SER=mmmpp-ssssss IF=xxxx

IOS581E LINK FAILED REPORTING CHPID=nn
        INCIDENT UNIT  TM=typ/mdl SER=mmmpp-ssssss IF=xxxx
        ATTACHED UNIT  TM=typ/mdl SER=mmmpp-ssssss IF=xxxx
```

Along with xC3 Records in Sys1.Logrec

Limited Recovery Time

Problem

- Missing interrupt condition or interface timeout (e.g., IOS051I) causes z/OS to validate the channel path
- Validation consists of issuing an I/O to test each path
- Bad paths can encounter elongated recovery times
  - I/O timeout is 15 seconds + 2 retries are performed = 45 seconds
- Application I/O is held up while validation is performed

Solution

- Provide z/OS parmlib option to limit the amount of time spent doing recovery operations
  - IECIOSxx RECOVERY,LIMITED_RECTIME=nn,DEV={DASD|IOTIMING}
  - Range 2-14 seconds
  - All DASD or only devices with I/O timing enabled
- z/OS uses the specified value to time the recovery I/Os on each path
- If recovery I/O fails with a missing interrupt start pending condition or interface timeout condition, I/O is not retried on that path

Limited Recovery Time

- IECIOSxx parmlib and SETIOS commands to enable the function

```
RECOVERY,LIMITED_RECTIME=ss,
                DEV={DASD|IOTIMING}
```

- Display IOS command to display the status:

```
D IOS,RECOVERY
IOS103I hh.mm.ss RECOVERY OPTIONS
LIMITED RECOVERY TIME IS 2 SECONDS
LIMITED RECOVERY IS REQUESTED FOR DASD
```

Limited Recovery Time – Best Practices

- IECIOSxx parmlib recommendation:

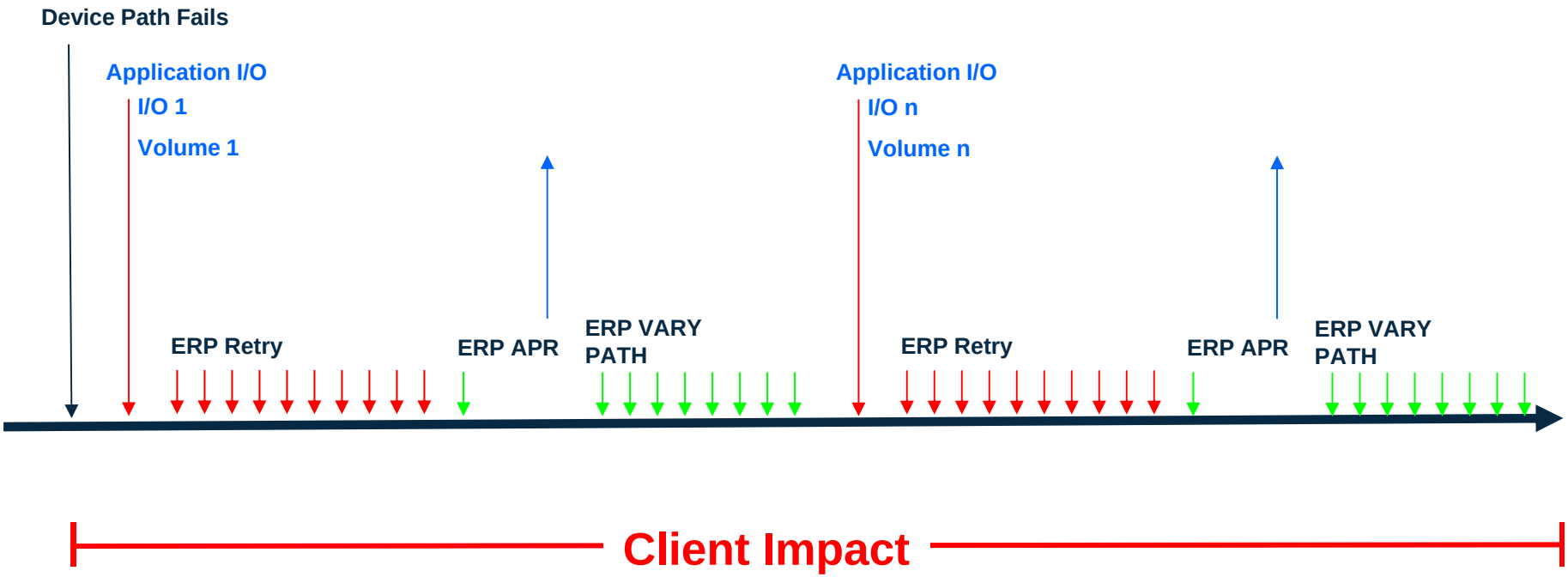
```
RECOVERY,LIMITED_RECTIME=2,DEV=DASD
```

Recent z/OS Resilience Enhancements

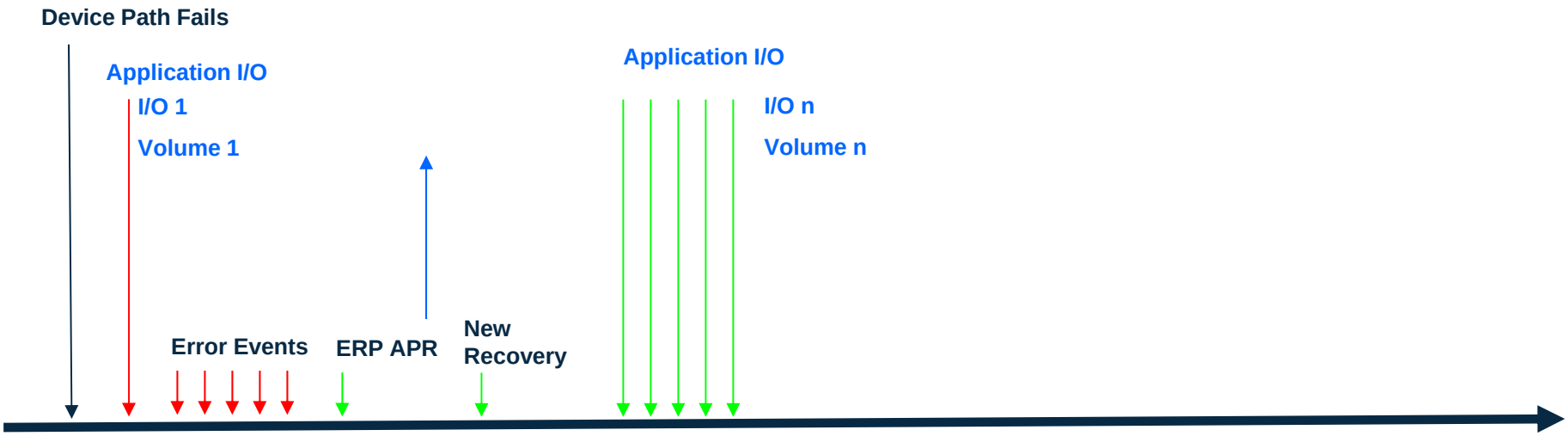
z/OS 1.13

Accelerate I/O Device Path Recovery

I/O Recovery for Failing Path - Before



I/O Recovery for Failing Path - After



Client Impact

Improved Channel Recovery

- Improved system resilience for H/W errors
- Customers would rather see path taken offline than continue to cause problems
  - IOS recovery delays application I/O even when there are other paths
- In particular:
  - IFCC and other path error thresholds
  - Proactively removing a path from all devices in an LCU
  - Identifying detecting component in IOS05xI messages
- DASD and tape only – FICON or ESCON

Improved Channel Recovery

- New IECIOSxx parmlib and SETIOS commands to enable the new function

```
RECOVERY,PATH_SCOPE={DEVICE|CU}
         PATH_INTERVAL=nn
         PATH_THRESHOLD=nnn
```

  - New display IOS command to display the status:

```
D IOS,RECOVERY
IOS103I hh.mm.ss RECOVERY OPTIONS
LIMITED RECOVERY FUNCTION IS DISABLED
PATH RECOVERY SCOPE IS BY CU
PATH RECOVERY INTERVAL IS nn MINUTES
PATH RECOVERY THRESHOLD IS nnn ERRORS
```

Improved Channel Recovery – Best Practices

- New IECIOSxx parmlib and SETIOS commands to enable the new function

```
RECOVERY,PATH_SCOPE=CU,PATH_INTERVAL=1,
                PATH_THRESHOLD=10
```

Improved Channel Recovery

- IFCC Thresholding
  - Remove path for intermittent errors
  - Default: at least 10 IFCCs per minute over a 10 minute period
  - Remove the path from all devices in the LCU
  - ERP path related error monitoring

IOS050I CHANNEL DETECTED ERROR ON dddd,yy,op,stat,
PCHID=pppp

IOS210I PATH RECOVERY INITIATED FOR PATH pp ON CU cccc,
REASON=PATH ERROR THRESHOLD REACHED

Improved Channel Recovery

- Proactively Removing Paths – DPS Validation Error
  - Dynamic Pathing Validation issues SNIDs down each path to test state of the path group
  - If error occurs, path is removed from device
  - Each device trips over the error
  - If PATH_SCOPE=CU, do all devices in LCU

IOS051I INTERFACE TIMEOUT DETECTED ON ON dddd,yy,op,stat, PCHID=pppp
IOS071I dddd,cc,jjjjjjjj, START PENDING

IOS450E dddd, cc NOT OPERATIONAL PATH TAKEN OFFLINE

IOS210I PATH RECOVERY INITIATED FOR PATH pp ON CU cccc, REASON=DYNAMIC PATHING ERROR

Improved Channel Recovery

- Proactively Removing Paths – Flapping Links
  - Each device trips over the link threshold condition
  - Stray I/O may interfere recovery after customer fixes the problem
  - If PATH_SCOPE=CU, do all devices in LCU

IOS001E dddd,INOPERATIVE PATHS pp pp pp

IOS2001I dddd,INOPERATIVE PATHS
STATUS FOR PATH(S) pp,pp,pp....
LOGICAL PATH IS REMOVED OR NOT ESTABLISHED (A0)
LINK RECOVERY THRESHOLD EXCEEDED FOR LOGICAL PATH (06)

IOS210I PATH RECOVERY INITIATED FOR PATH pp ON CU cccc,
REASON=LINK THRESHOLD EXCEEDED

Improved Channel Recovery

- D M=DEV(devno,chp) will display offline reasons

```
D M=DEV(410,(48))
IEE174I hh.mm.ss DISPLAY M idr
 DEVICE 0410   STATUS=ONLINE
 CHP                    48
 ENTRY LINK ADDRESS     22
 DEST LINK ADDRESS      E0
 PATH ONLINE            N
 CHP PHYSICALLY ONLINE Y
          ...
 PATH OFFLINE DUE TO THE FOLLOWING REASON(S)]
              [PATH RECOVERY ERROR]
              [BY OPERATOR]
              [CONTROL UNIT INITIATED RECOVERY]
              [CONFIGURATION MANAGER]
```

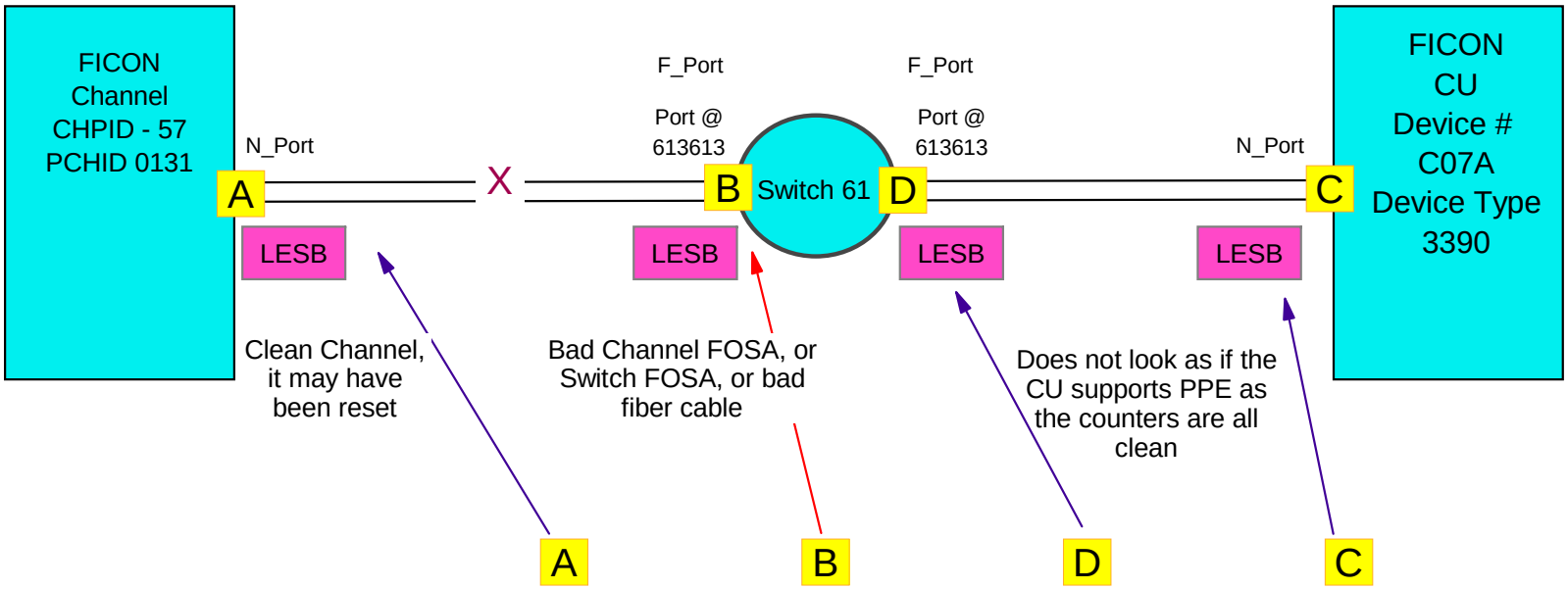
Improved Channel Recovery

- Handling of Managed Channels
  - Cannot simply vary the path offline when an error occurs. Must remove the channel.
  - DCM will remove managed channels that flagged for path recovery from all CUs in the group
    - Avoid using the CHPID.link combination for that group for 1 hour
    - Managed CHPID is eligible to be used by other groups or with other destination ports
    - Unable to precisely identify the failing component
    - FICON channels only

IOS210I PATH RECOVERY INITIATED FOR PATH pp ON CU cccc,
REASON=DYNAMIC PATHING ERROR

Fault Isolation

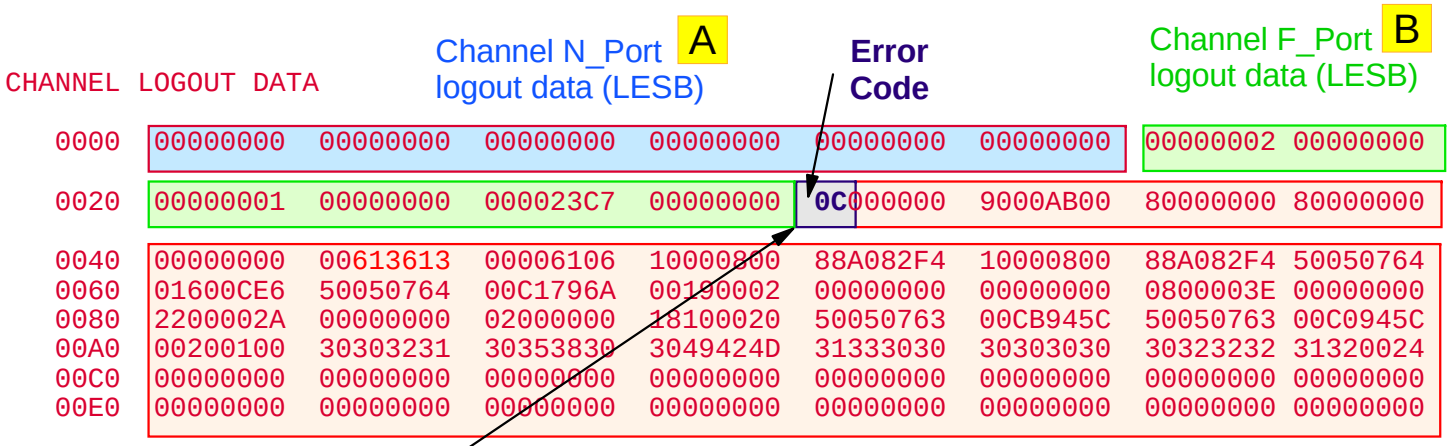
Link Error Statistics - Collection



LESB Field name	Channel N_Port LESB Counters	Channel F_Port LESB Counters	CU F_Port LESB Counters	CU N_Port LESB Counters
Link Failure Count	00000000	00000002	00000000	00000000
Loss-of-Synchronization Count	00000000	00000000	00000000	00000000
Loss-of-Signal Count	00000000	00000001	00000000	00000000
Primitive Sequence Protocol Error	00000000	00000000	00000000	00000000
Invalid Transmission Word	00000000	**00C1796A**	00000000	00000000
Invalid CRC Count	00000000	00000000	00000000	00000000

Details of the LESB contents can be found in the Fibre Channel architecture document FC-FS

Purge Path Extended – Error Codes



Error Code	Error Code meaning
00	Error-code transfer not supported
01	SB-3 protocol time-out
02	SB-3 link failure
03	Reserved
04	SB-3 offline condition
05	FC-PH link failure
06	SB-3 length error
07	LRC error
08	SB-3 CRC error
09	IU count error
0A	SB-3 link-level protocol error
0B	SB-3 device-level protocol error
0C	Receive ABTS
0D	Cancel function time-out
0E	Abnormal termination of exchange
0F - FF	Reserved

For additional details on the meaning of the PPE error code see, the Fibre Channel architecture documentation FC- SB-3 or later releases

SYS1.LOGREC SLH Record - Formatting

```
EXTENDED SUBCHANNEL LOGOUT DATA
 CHANNEL LOGOUT DATA
   N-PORT LINK ERROR STATUS BLOCK
    LINK FAILURE COUNT:         00000007   LOSS OF SYNCHRONIZATION COUNT: 0003CDA5
    LOSS OF SIGNAL COUNT:       0000000A   PRIMITIVE SEG PROTOCOL ERROR:  00000000
    INVALID TRANSMISSION WORD: 001AEC57    INVALID CRC COUNT:             00000020

   FABRIC ENTRY PORT LINK ERROR STATUS
    LINK FAILURE COUNT:         00000002   LOSS OF SYNCHRONIZATION COUNT: 00000007
    LOSS OF SIGNAL COUNT:       00000009   PRIMITIVE SEG PROTOCOL ERROR:  00000000
    INVALID TRANSMISSION WORD: 0000000A    INVALID CRC COUNT:             00000000
   ERROR CODE: 00 - Error code transfer not supported

   MODEL DEPENDENT DATA:
    0000  00000000  E000030F  A8000000  80000000  00000000  00513C00  0000413A  10000005
    0020  1E3643DE  10000005  1E3643DE  50050764  016611C1  50050764  00C22A72  00000001
    0040  00000000  00000000  0800002A  80000000  22800022  00000000  02000000  18100020
    0060  50060E80  042ABEC1  50060E80  042ABEC1  00200100  30303231  30354632  30485443
    0080  35353030  30303030  30313039  343200A4  00000000  00000000  00000000  00000000
    00A0  00000000  00000000  00000000  00000000  00000000  00000000  00000000  00000000
    00C0  00000000  00000000  00000000  00000000

CONTROL UNIT LOGOUT DATA
  N-PORT LINK ERROR STATUS BLOCK
   LINK FAILURE COUNT:         00000000   LOSS OF SYNCHRONIZATION COUNT: 00000000
   LOSS OF SIGNAL COUNT:       00000000   PRIMITIVE SEG PROTOCOL ERROR:  00000000
   INVALID TRANSMISSION WORD: 00000000    INVALID CRC COUNT:             00000000

  FABRIC ENTRY PORT LINK ERROR STATUS
   LINK FAILURE COUNT:         00000000   LOSS OF SYNCHRONIZATION COUNT: 00000000
   LOSS OF SIGNAL COUNT:       00000000   PRIMITIVE SEG PROTOCOL ERROR:  00000000
   INVALID TRANSMISSION WORD: 00000000    INVALID CRC COUNT:             00000000
  ERROR CODE: 00 - Error code transfer not supported

  MODEL DEPENDENT DATA:
   0000  00000000  00000000  00000000  00000000  00000000  00000000  00000000  00000000
   0020  00000000  00000000  00000000  00000000  00000000  00000000  00000000  00000000
   0040  00000000  00000000  00000000  00000000  00000000  00000000  00000000  00000000
   0060  00000000  00000000  00000000  00000000  00000000  00000000  00000000  00000000
   0080  00000000  00000000  00000000  00000000  00000000  00000000  00000000  00000000
   00A0  00000000  00000000  00000000  00000000  00000000  00000000  00000000  00000000
   00C0  00000000  00000000  00000000  00000000
```

Improved Channel Recovery

- Identify Detecting Component for IFCCs
  - Customers have a difficult time determining where the error is
    - Channel, switch(es), CU interface, links
  - Identify detecting component based on H/W logout data
    - FICON only
  - Not controlled by PATH_SCOPE option

IOS050I CHANNEL DETECTED ERROR ON ddddd,yy,op,stat,
PCHID=pppp

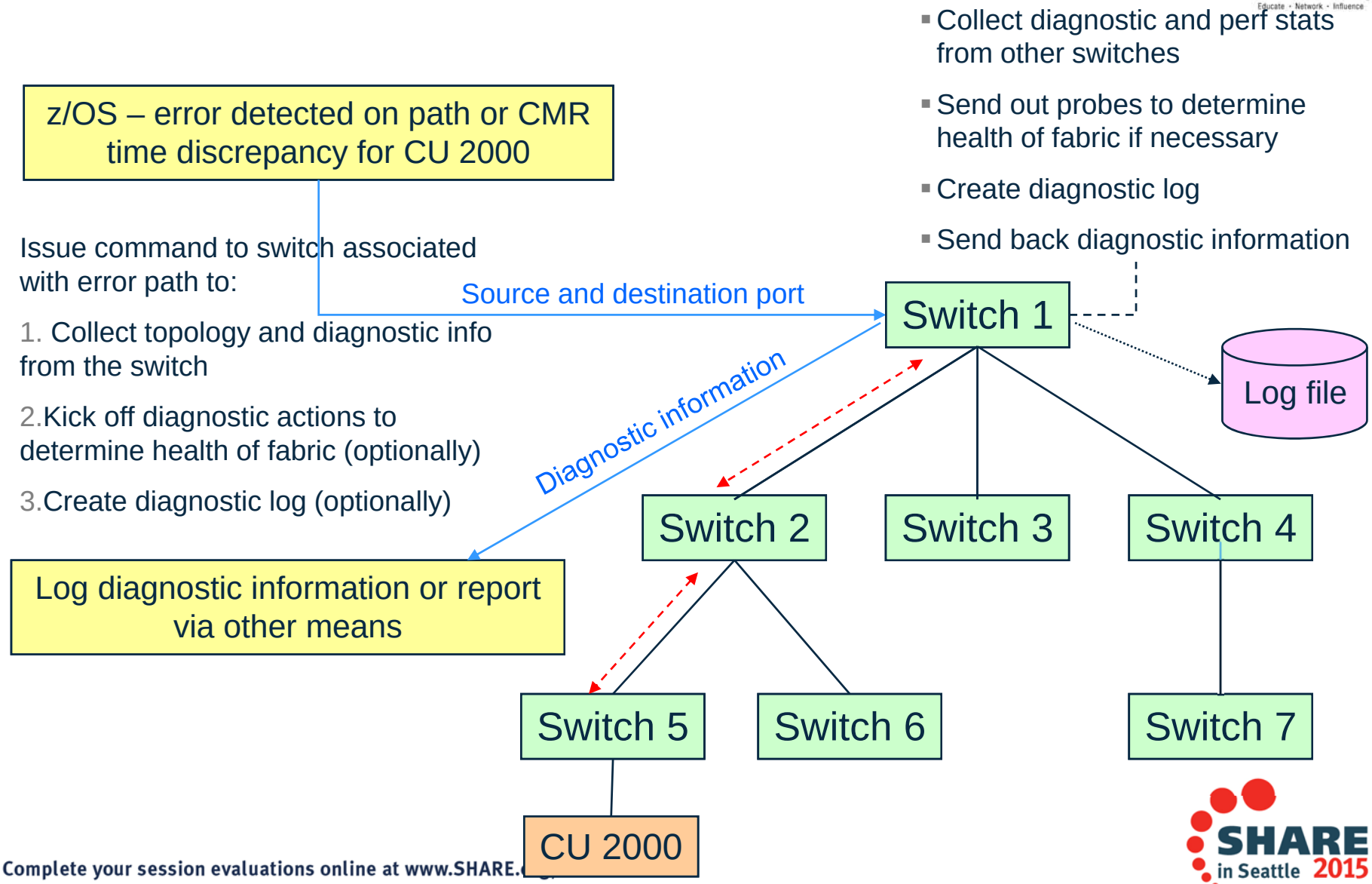
IOS054I ddddd,pp ERRORS DETECTED BY comp, comp,…

Where *comp* is one or more of the following:

CHANNEL, CHAN SWITCH PORT, CU SWITCH PORT, CONTROL UNIT

Switch Diagnostics

Enhanced First Failure Data Capture for SANs



Health Check Report

```
CHECK(IBMIOS,IOS_FABRIC_MONITOR)
START TIME: 09/03/2012 17:34:09.652404
CHECK DATE: 20120301  CHECK SEVERITY: MEDIUM
CHECK PARM: LOG(YES)

IOSHC120I Fabric Health Report

The following fabric health issues have been detected:

FABRIC HEALTH EXCEPTION DETECTED AT: 09/03/2012 17:30:18.301627
CHPID=20, Entry link=2000, Exit link=3018, Suspect link=3010

Routing information follows:
 Switch Domain=20, Type=Source
                            Group   Neg
 Port Type  From   To     Agg Dyn Speed Misc
  00  Entry Chan   Agg-1  ..  ..    8G  Static Alt=1
  01  Exit  2000   3010    1  ..    8G
  02  Exit  2000   3011    1  ..    8G
  03  Exit  2000   3012    1  ..    8G
  04  Exit  2000   3013    1  ..    8G

 Switch Domain=30, Type=Destination
                            Group   Neg
 Port Type  From   To     Agg Dyn Speed Misc
  10  Entry 2001   3018    1  ..    8G  Static
  11  Entry 2002   3018    1  ..    8G  Static
  12  Entry 2003   3018    1  ..    8G  Static
  13  Entry 2004   3018    1  ..    8G  Static
  18  Exit  Agg-1  CU     ..  ..    8G  ......
```

Health Check Report

```
Health information follows:

 Fabric Health=Port Error
 Switch Domain=20, Health=No health issues
                                %Util  %Delay   Error Count Opt Sign
 Port Health                   Trn/Rcv Trn/Rcv   Trn/Recv    Trn/Rec
  00  Port Normal                1/0     0/0       0/0      +0.1/+0.2
  01  Port Normal                2/3     0/0       0/0      +0.2/+0.2
  02  Port Normal                0/0     0/0       0/0      +0.2/+0.4
  03  Port Normal                0/1     0/0       0/0      +0.4/+0.7
  04  Port Normal                2/1     0/0       0/0      +0.2/+0.1

 Switch Domain=30, Health=Port Error
                                %Util  %Delay   Error Count Opt Sign
 Port Health                   Trn/Rcv Trn/Rcv   Trn/Recv    Trn/Rec
  10  High Error Count           2/2     0/0      93/26     +0.9/+0.8
  11  Port Normal                0/0     0/0       0/0      +0.2/+0.2
  12  Port Normal                1/0     0/0       0/0      +0.2/+0.4
  13  Port Normal                1/2     0/0       0/0      +0.4/+0.7
  18  Port Normal                0/0     0/0       0/0      +0.2/+0.1
```

D M=DEV Support

- Display new information when device and path selected
- Support displaying routing path only or routing path with health information (more overhead)

```
D M=DEV(devn,(chp))
        [,ROUTE={TODEV|FROMDEV|ALL}]
        [,HEALTH]
```

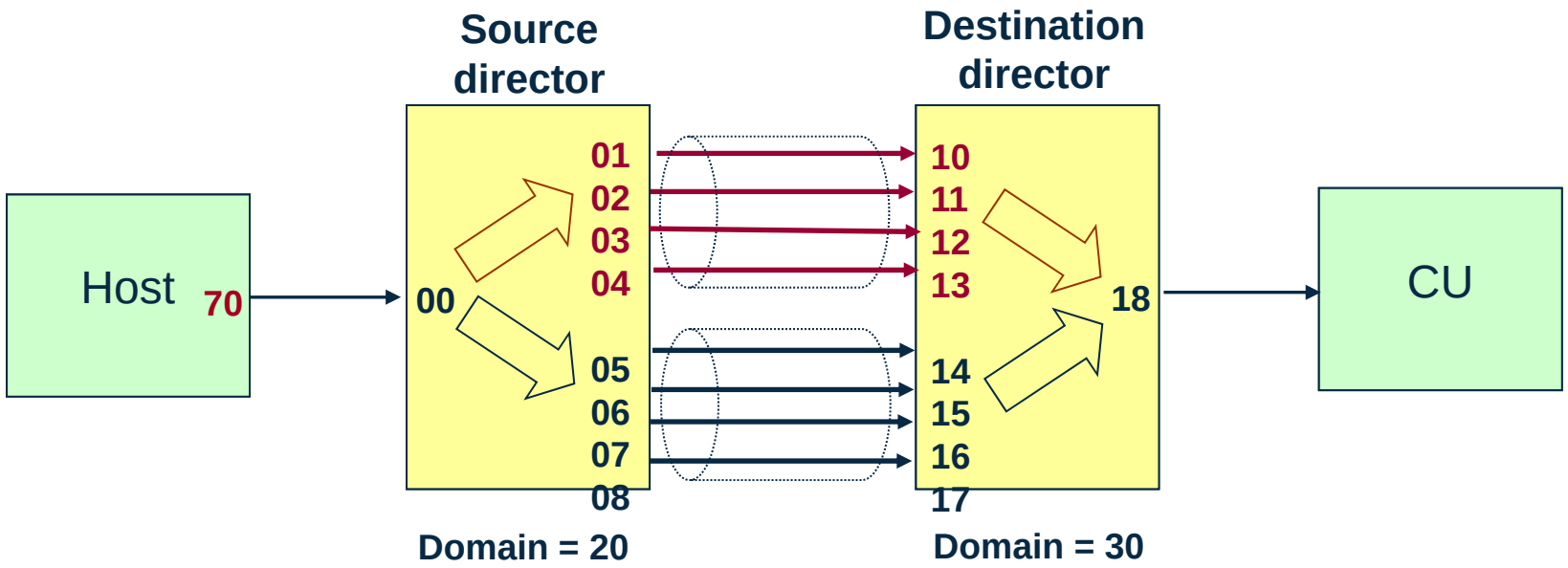
ROUTE=

- TODEV – Display the route through the fabric starting with the channel and going to the device
- FROMDEV – Display the route through the fabric starting with the device and going to the channel
- ALL – Display the route through the fabric both ways

HEALTH

- Display fabric, switch, and port heath information such as utilization, average delay, signal strength and error counts

D M=DEV Example 1 - Static Routing, Aggregate Links



Port 00 primary static path = aggregate 1 consisting of ports 01, 02, 03, 04
Port 00 alternate static path = aggregate 2 consisting of ports 05, 06, 07, 08
Only primary path will be shown in display. Number of alternates will be displayed.

Note: Aggregate = trunking

D M=DEV Example 1 - Part 1 (Routing Information)

```
D M=DEV(2000,(70)),ROUTE=TODEV,HEALTH

Routing information follows:
  Switch Domain=20, Type=Source
                                Group   Neg
    Port Type  From   To      Agg Dyn Speed Misc
    00   Entry Chan   Agg-1   ..  ..    8G  Static Alt=1
    01   Exit  2000   3010    1  ..    8G
    02   Exit  2000   3011    1  ..    8G
    03   Exit  2000   3012    1  ..    8G
    04   Exit  2000   3013    1  ..    8G

  Switch Domain=30, Type=Destination
                                Group   Neg
    Port Type  From   To      Agg Dyn Speed Misc
    10   Entry 2001   3018    1  ..    8G  Static
    11   Entry 2002   3018    1  ..    8G  Static
    12   Entry 2003   3018    1  ..    8G  Static
    13   Entry 2004   3018    1  ..    8G  Static
    18   Exit  Agg-1  CU      ..  ..    8G  ......
```

D M=DEV Example 1 - Part 2 (Health Information)

```
Health information follows:
  Fabric Health=fabric-health-description

  Switch Domain=20, Health=switch-health-description
                                 Utiliz  Delay   Error Count
    Port Health                  Trn/Rcv Trn/Rcv   Trn/Recv
    00   port-health-description uuu/uuu ddd/ddd eeeee/eeeee
    01   port-health-description uuu/uuu ddd/ddd eeeee/eeeee
    02   port-health-description uuu/uuu ddd/ddd eeeee/eeeee
    03   port-health-description uuu/uuu ddd/ddd eeeee/eeeee
    04   port-health-description uuu/uuu ddd/ddd eeeee/eeeee

  Switch Domain=30, Health=switch-health-description
                                 Utiliz  Delay   Error Count
    Port Health                  Trn/Rcv Trn/Rcv   Trn/Recv
    10   port-health-description uuu/uuu ddd/ddd eeeee/eeeee
    11   port-health-description uuu/uuu ddd/ddd eeeee/eeeee
    12   port-health-description uuu/uuu ddd/ddd eeeee/eeeee
    13   port-health-description uuu/uuu ddd/ddd eeeee/eeeee
    18   port-health-description uuu/uuu ddd/ddd eeeee/eeeee
```

FICON Dynamic Routing (FIDR)

FICON Dynamic Routing (Brocade EBR or CISCO OxID) dynamically changes the routing between channel and control unit based on the "Fibre Channel Exchange ID"

Each I/O operation has a unique Exchange ID

Therefore, work can be balanced between all available ISLs

Improves utilization of switch and ISL hardware

Easier to plan ISL bandwidth requirements

Allows sharing of switches between FICON and FCP (PPRC) and Distributed

Easier to manage

Predictable, repeatable I/O performance

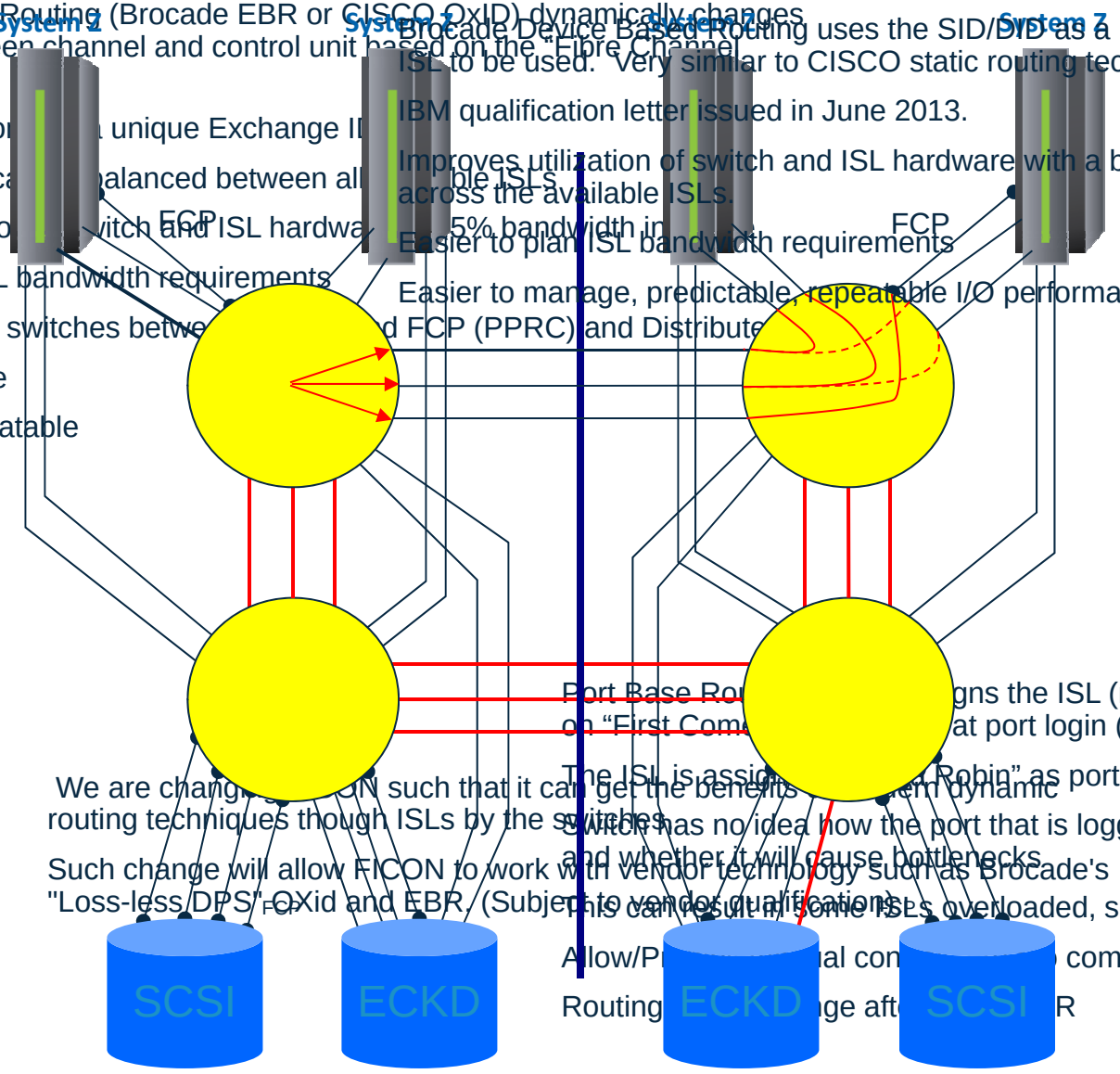
Brocade Device Based Routing uses the SID/DID as a hash to map ISL to be used. Very similar to CISCO static routing technique.

IBM qualification letter issued in June 2013.

Improves utilization of switch and ISL hardware with a better spreading of work across the available ISLs.

Easier to plan ISL bandwidth requirements

Easier to manage, predictable, repeatable I/O performance



Port Base Routing assigns the ISL (route) statically based on "First Come First Served" at port login (PLOGI) time

The ISL is assigned round-robin" as ports PLOGI (log in)

Switch has no idea how the port that is logging-in will use the ISL and whether it will cause bottlenecks.

This can result in some ISLs overloaded, some under utilized

Allow/Prohibit dynamic connectivity complicated to manage

Routing change after IOCDS

We are changing FICON such that it can get the benefits of System dynamic routing techniques though ISLs by the switches.

Such change will allow FICON to work with vendor technology such as Brocade's "Loss-less DPS" OXid and EBR. (Subject to vendor qualification)

Objectives for Health Check

- Assumption
  - All down stream switches in the SAN run with the same dynamic routing mode set (simplifying assumption not required by current implementation)
- Host Verification
  - If any FICON switches are found to have dynamic routing enabled (as indicated by CUP) then verify the host CEC supports dynamic routing
  - If the dynamic routing state changes in the SAN (e.g. off -> ON), re-verify host and devices
- Device Verification
  - When the machine and SAN are running with dynamic routing enabled
    - Verify that each device accessed over a SAN with Dynamic Routing enabled has been qualified for FICON dynamic Routing
    - Assumes that all switches in the SAN have the same dynamic routing mode

Example HCD Changes Enable Real Time Checking

```
Supported Processors
CBDPSPR1                                                Row 318 of 612
Command ===> ____________________________________________________ Scroll

Select one to view more details.
```

Processor Type-Model	Support Level ID	Supported Protocols	WI	RI	DP	PCIe Fct	FIDR
2828-H13	H130331	D,S,S4	Yes	Yes	Yes	Yes	No
2964-NC9	H150111	D,S,S4	Yes	Yes	Yes	Yes	No
2964-NC9	H150112	D,S,S4	Yes	Yes	Yes	Yes	Yes
2964-NE1	H150111	D,S,S4	Yes	Yes	Yes	Yes	Yes
2964-NE1	H150112	D,S,S4	Yes	Yes	Yes	Yes	Yes

Example HCD Changes Enable Real Time Checking

```
  Switch List          Row 2 of 2 More:       >
Command ===> _____________________________________________ Scroll ===> PAGE

Select one or more switches, then press Enter. To add, use F11.

                                                                 CU   Dev   FI
/ ID Type +        Ad Serial-# + Description                     Num. Num.  DR+
_ 34 2032          34 DD         d                               0034 0034  __
_ 35 2032          35 DD1        d                               0035 0036  no
_ 36 2032          36 DD2        d                               0035 0036  yes
_ 36 2032          36 DD2        d                               0035 0036  aut
******************************* Bottom of data ********************************
A new prompt will show the valid values for FICON dynamic routing.

This list can be filtered. The filter panel will be updated to filter by the FICON dynamic
routing attribute.

                  Filter Switch List
CBDPSWFI

Specify or revise the following filter criteria.

Switch type  . . . . . _____________  +
Serial number  . . . . . __________
Description  . . . . . . ____________________________________
FICON dynamic routing  . _  (Y/N/Auto)
```

Example HCD Changes Enable Real Time Checking

```
Change Control Unit Definition
Specify or revise the following values.
Control unit number . . . . 1234 +
Control unit type . . . . . OSA +
Serial number . . . . . __________ +
Description . . . . . ____________________________________
Connected to switches . . . __ __ __ __ __ __ __ __ +
Ports . . . . . __ __ __ __ __ __ __ __ +
Define more than eight ports . . 2 1. Yes
                                   2. No
FICON dynamic routing . . . . . _ (Y = Yes; N = No) +
```

Example HCD Changes Enable Real Time Checking

```
Actions on selected control units
Select by number or action code and press Enter.
__ 1. Add like . . . . . . . . . . . . . . . (a)
   2. Change . . . . . . . . . . . . . . . . (c)
   3. *Prime serial number . . . . . . . . . (i)
   4. Delete . . . . . . . . . . . . . . . . (d)
   5. Work with attached devices . . . . . . (s)
   6. View control unit definition . . . . . (v)
   7. View logical CU information  . . . . . (l)
   8. View related CTC connections . . . . . (k)
   9. View graphically . . . . . . . . . . . (h)
  10. Propagate FICON dynamic routing  . . . (r)

Propagate will bring up a new panel with following content:
(same panel as for switch)

Propagate options


Select FICON dynamic routing for the selected object
and all related control units to


_ 1.Yes
  2.No
  3.clear attribute
```

z/OS Single Point of Failure Service

- z/OS 1.10 introduced IOSSPOF service which allows you to check for single points of failure (SPOFs)
  - Check for SPOFs for a specific device
  - Check for common SPOFs between two devices
    - E.g., primary and backup XCF couple data sets
- Examples:
  - Only one online path to the device
  - All online paths go through the same switch
  - All online paths are connected to the same port or host adapter card on the control unit

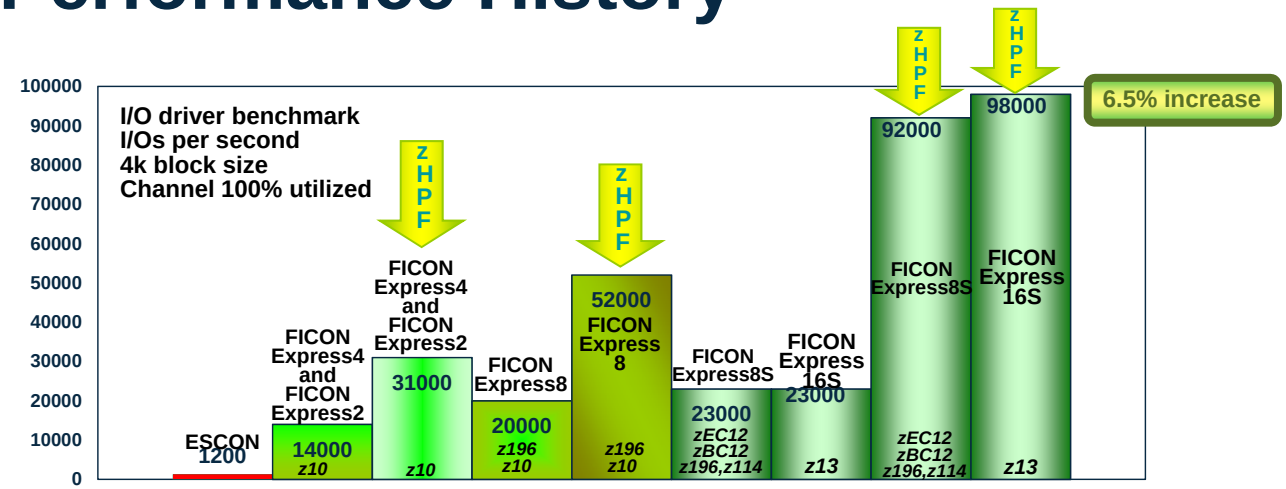
IOSSPOFD Tool

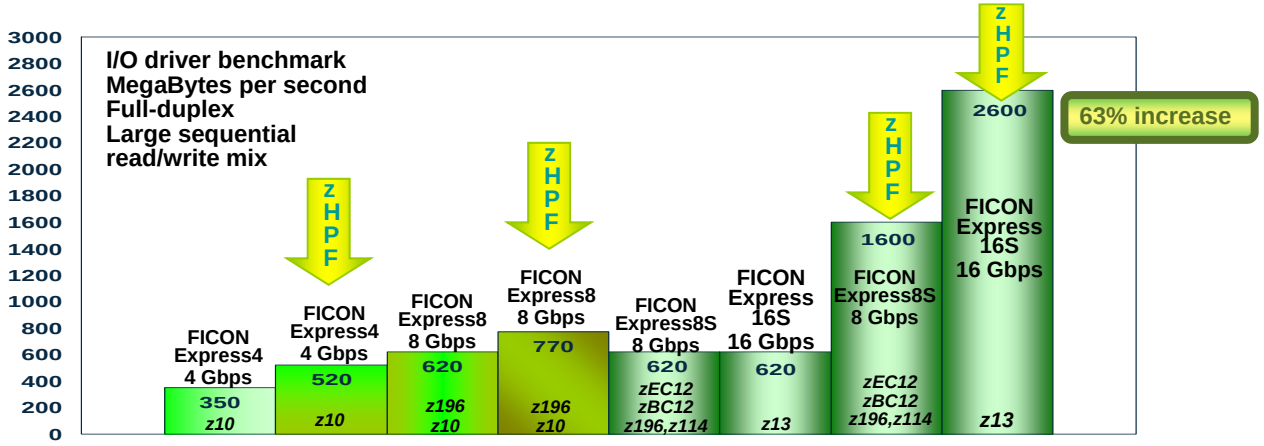
- Allows you to check for single points of failure in your own configuration
- Run as a batch job, invoked from a program, CLIST or REXX exec
- Input is a list of device numbers, volsers, or data set names
- Uses the IOSSPOF service to check for single points of failure and generate messages
- Available at z/OS tools and toys website
  - http://www-03.ibm.com/systems/z/os/zos/features/unix/bpxa1ty2.html

Review High Performance FICON (zHPF)

- Improve FICON Scale, Efficiency and RAS
  - As the data density behind a CU and device increase, scale I/O rates and bandwidth to grow with the data
    - Leverages HBA hardware optimizations done while preserving System z QOS
    - Significant improvements in I/O rates (4-5x) for small block transfer
    - Improved I/O bandwidth (ability to fill the link at 8 Gbs and beyond)
    - New ECKD commands for improved efficiency
  - Improved first failure data capture
  - Additional channel and CU diagnostics for MIH conditions
- Value
  - Enhanced resilience for work load spikes and failing components
  - Reduced job elapsed times
  - Improve First Failure Data Capture
  - Improved workload management
  - Possible to reduce the number of channels, switch ports, control unit ports and optical cables required to balance CPU MIPS with I/O capacity

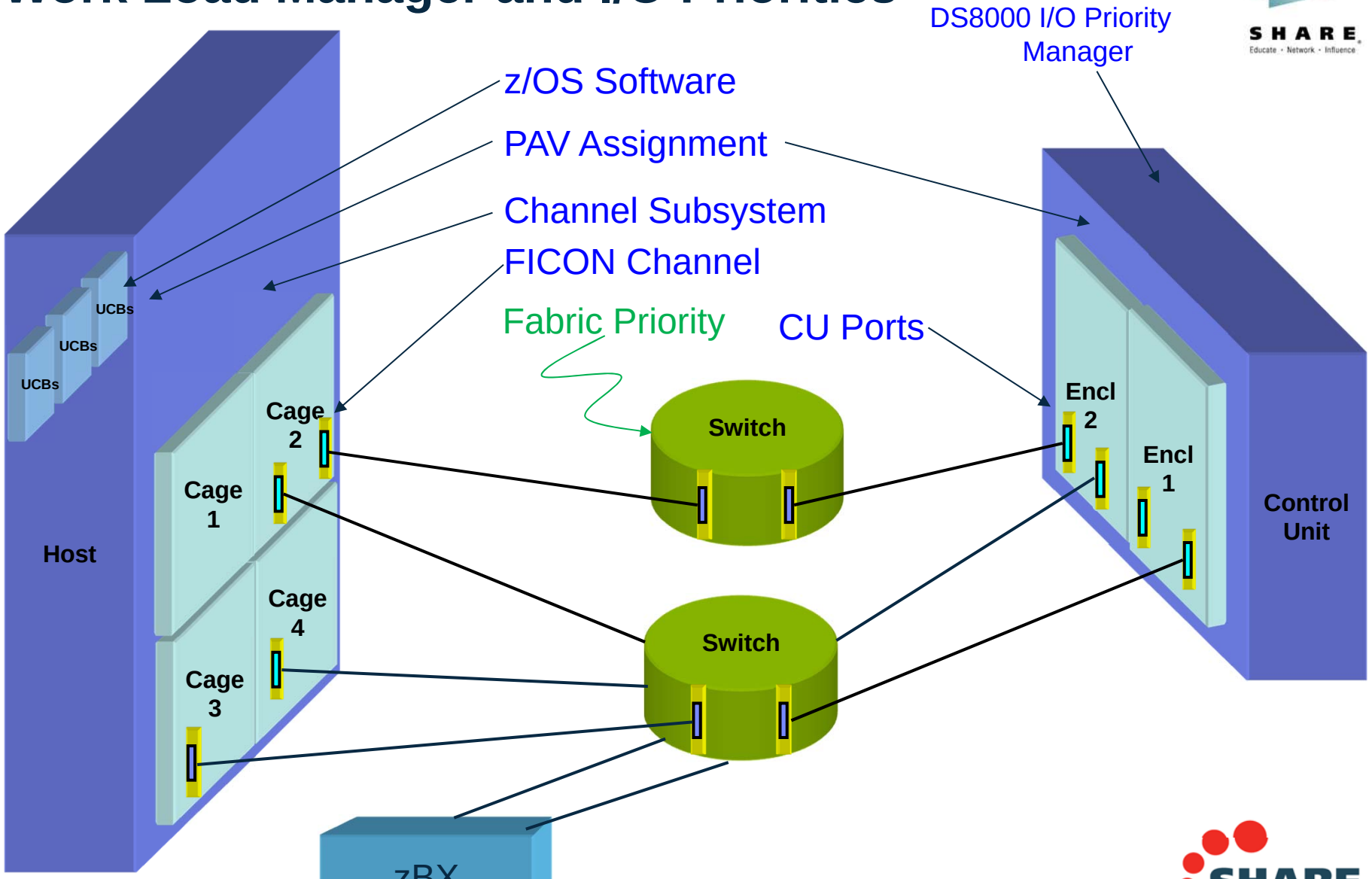
FICON Performance History





*This performance data was measured in a controlled environment running an I/O driver program under z/OS. The actual throughput or performance that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed.

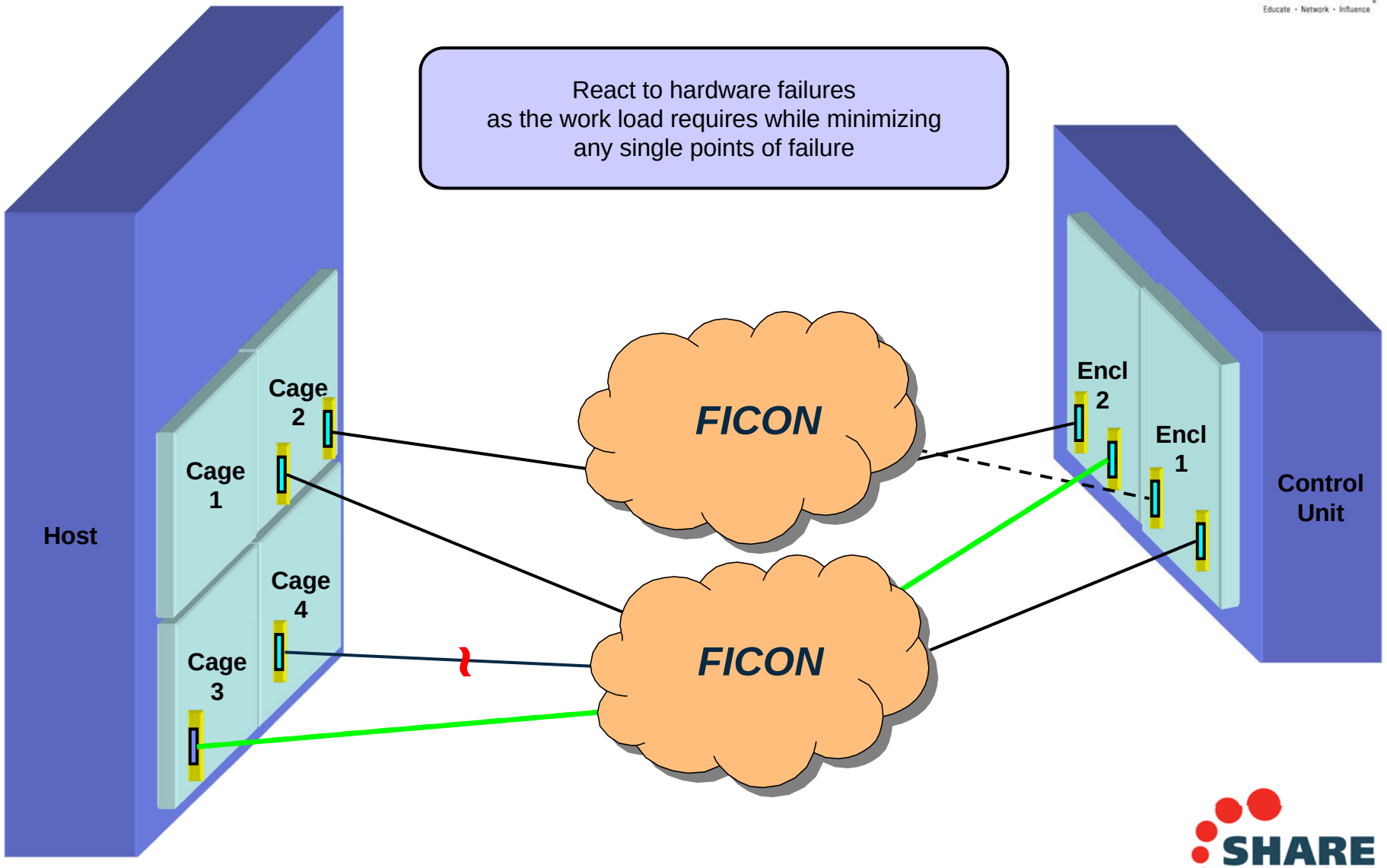
Work Load Manager and I/O Priorities



Dynamic CHPID Management Overview

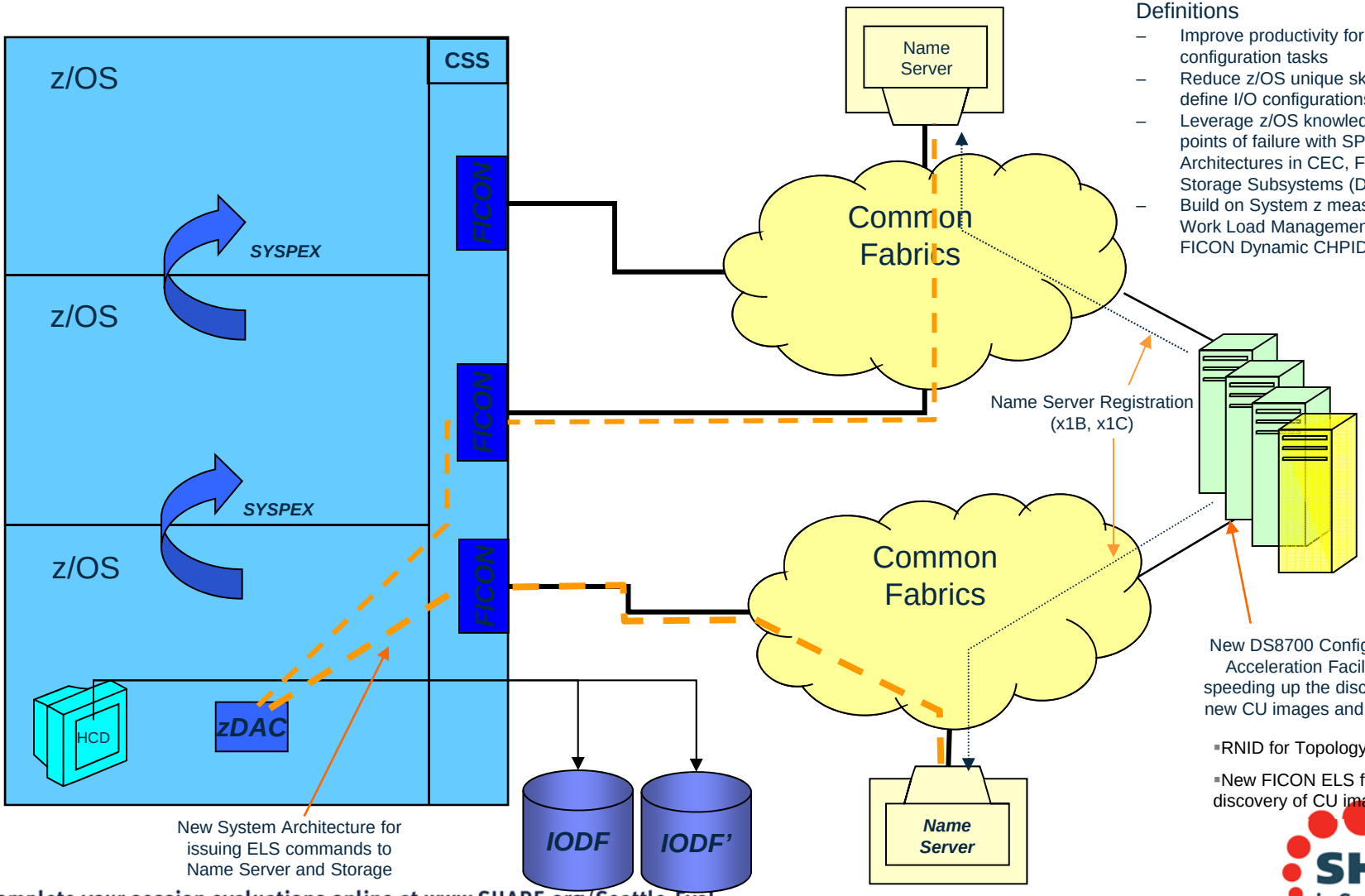
- Goals
  - Simplify I/O configuration definition task
  - Reduce customer skill set
  - Improve workload management
  - Maximize utilization of installed hardware
  - Enhance RAS
  - Complements System z Discovery and Auto-Configuration

FICON Dynamic CHPID Management



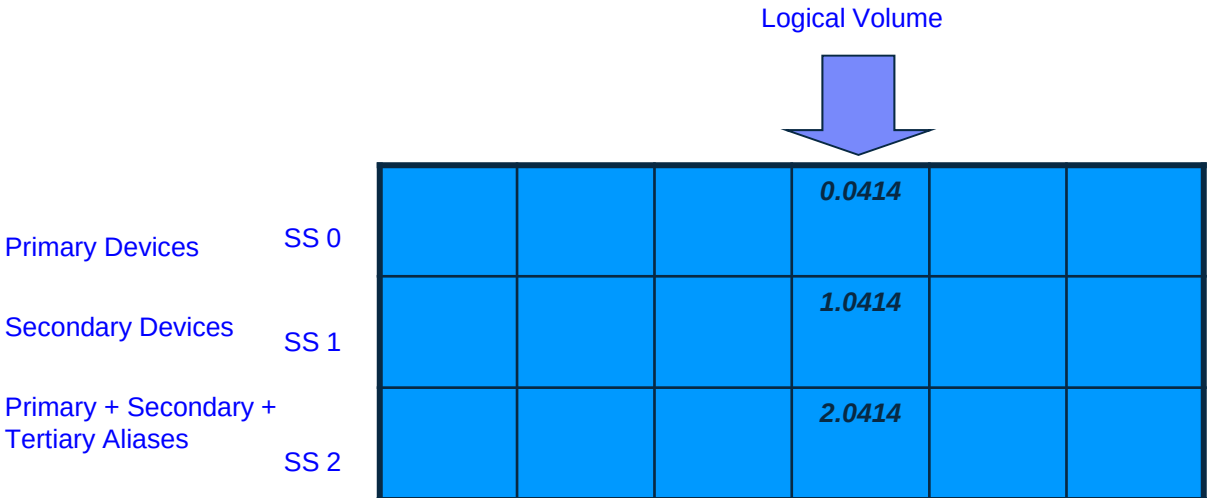
System z Discovery and Auto-Configuration (zDAC)

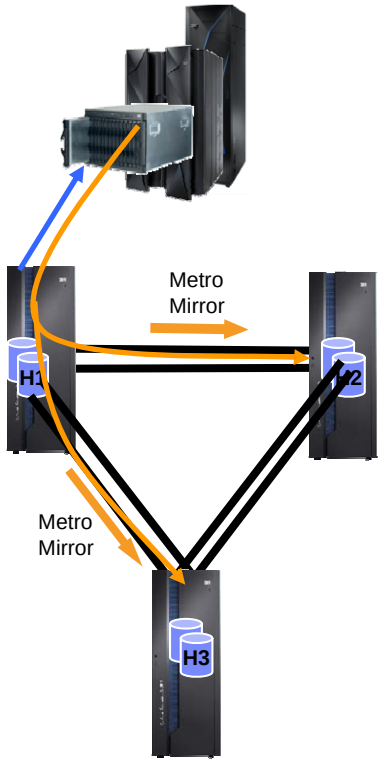
- Simplify System z I/O Configuration Definitions
  - Improve productivity for existing I/O configuration tasks
  - Reduce z/OS unique skills required to define I/O configurations for z/OS
  - Leverage z/OS knowledge of single points of failure with SPOF Architectures in CEC, Fabric and Storage Subsystems (DASD and Tape)
  - Build on System z measurements and Work Load Management Strengths with FICON Dynamic CHPID Management



Multi-Target Metro Mirror with HyperSwap

Maintain HyperSwap readiness after the primary or a secondary fails.
Device number assignment needs to be simplified:





Compliments multi-target PPRC by simplifying the configuration changes needed to define 3rd copy of data in large configurations.

HyperWrite is designed to also work with Multi-target PPRC, day 1.

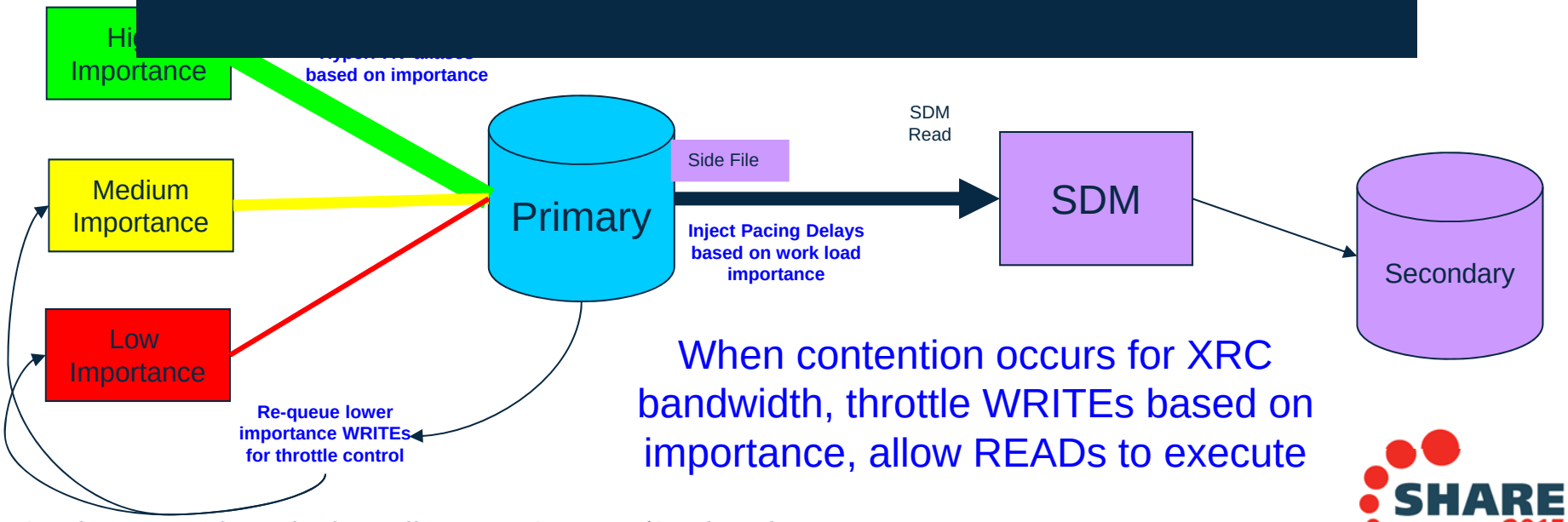
Workload Based Write Pacing with HyperPAV Support (4Q2014)

- Goals
  - Simplify management of zGlobal Mirror by

dth

to

l



When contention occurs for XRC bandwidth, throttle WRITEs based on importance, allow READs to execute

z13 provides the next generation of Mainframe I/O:

New resilient IO Infrastructure addresses Skills, Complexity, Cost and Availability

Capability	Client Value
16 Gbs FICON	Faster links will improve I/O latency. For DB2 Log Writes, 16 Gbs zHPF will **improve DB2 log write latency and throughput by up to 32%** with multiple I/O streams, resulting improved DB2 transactional latency. For multi-stream I/O bound batch jobs, clients can expect up to a **32% reduction in elapsed times**.
Forward Error Correction Codes	The faster link speed technologies are more sensitive to the quality of the cabling infrastructure. IBM is leading new industry standard to provide FEC for optical connections. This will provide the ability to correct up to 11 bit errors out of a block of 2112 bits, the same benefit that would occur as if the optical signal strength was increased 2x yielding **substantially reduced IO link errors. This technology will allow System z I/O to operate at higher speeds, over longer distances, with reduced power and higher throughput, while retaining the same reliability and robustness that FICON has traditionally been known for.**
zHPF Extended Distance II	Clients using multi-site configurations can expect **up to 50% I/O service time improvement when writing data remotely** (remote site recovery). This capability is required especially for GDPS HyperSwap configurations where the secondary DASD subsystem is in another site.
FICON Dynamic Routing	New System z host feature that allows clients to use SAN dynamic routing policies across cascaded FICON Directors. This will **simplify configuration planning, capacity planning, provide persistent and repeatable performance and be more resilient after hardware failures** by allowing the ISL links to be driven to higher utilizations before encountering queuing delays. Configuration planning is simplified and hardware costs reduced by allowing FICON and FCP (PPRC) to share the same switch infrastructure without creating separate virtual switches and adding ISLs.
Fabric Priority	With SAN Fabric Priority **important work gets done first when SAN hardware failures result in traffic congestion.** This is achieved by extending the z/OS WLM policy into the SAN fabric leveraging capabilities of the SAN vendors. z/OS and z Systems arethe first platform to provided an integrated workload management function this exploit this industry feature.
Scale	Scales to **six logical channel subsystems** (LCSS) allows for up to **85 client useable LPARs**. Up to **four subchannel sets** per LCSS for added flexibility. All FICON channels supported on z13 (FE8, FE8s, FE16s) will support up to **32K devices per channel**.
Resilience	A fourth subchannel set for each LCSS is provided to facilitate **elimination of single points of failure for storage after a disk failure** by facilitating the exploitation of IBM's DS8870 **Multi-target Metro Mirror** storage replication with GDPS and TPC-R HyperSwap.
Read Diagnostic Parameters (z13 GA2 SOD)	Integrated instrumentation to allow clients to **find potential trouble spots in the SAN without manually inserting light meters** around the machine room. This will help reduce false Repair Actions (no defect found, NDF). z/OS will also automatically be able to differentiate when errors are caused by faulty components versus dirty optical connections.


Thank You!
QUESTIONS?