IBM DS8870 Enterprise Storage News and Review

Session 17962

Warren Stanley

DS8000 Copy Services

IBM Systems Division

August 13, 2015

#SHAREorg

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

Agenda

- DS8870 review
- DS8870 Release 7.4 content
- DS8870 Release 7.5 content

Current DS8870 Family Models

DS8870 – Enterprise Class

- Dual POWER7+ controllers (2-way, 4-way, 8 -way, 16-way)
- Up to 1 TB processor memory
- 8 or 16 Gb/s host adapters
- 8 Gb/s device adapters
- 2.5” Enterprise SAS-2 and SSD drives , 3.5” Nearline drives
- Up to 1,536 2.5” drives plus 240 1.8” Flash drives
- All drives encryption capable
- All upgrades concurrent

DS8870 – Business Class

- Dual POWER7+ controllers (2-way, 4-way, 8 -way, 16-way)
- Up to 1 TB processor memory
- 8 or 16 Gb/s host adapters
- 8 Gb/s device adapters
- 2.5” Enterprise SAS-2 and SSD drives , 3.5” Nearline drives
- Up to 1,056 drives plus 240 1.8” Flash drives
- All drives encryption capable
- All upgrades concurrent

DS8870 – All Flash

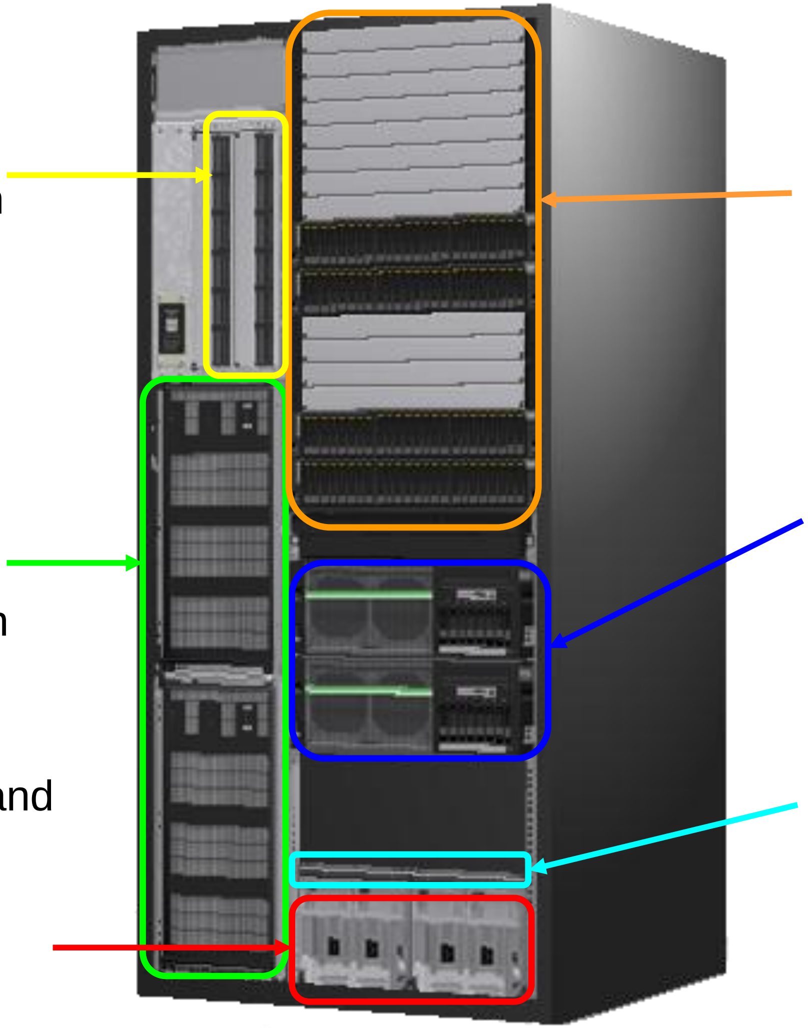
- Dual POWER7+ controllers (8-way, 16-way)
- Up to 1TB processor memory
- 8 or 16 Gb/s host adapters
- 8 Gb/s device adapters
- Up to 240 1.8” Flash drives
- All drives encryption capable
- All upgrades concurrent

DS8870 Hardware Review

High Performance Flash Enclosures provide significant performance improvements with Flash optimized RAID

Highly resilient DC-UPS power supply with 98% power efficiency and 4 second power loss ride-through (50 seconds with ePLD)

High performance Host and Device Adapters provide CPU offload



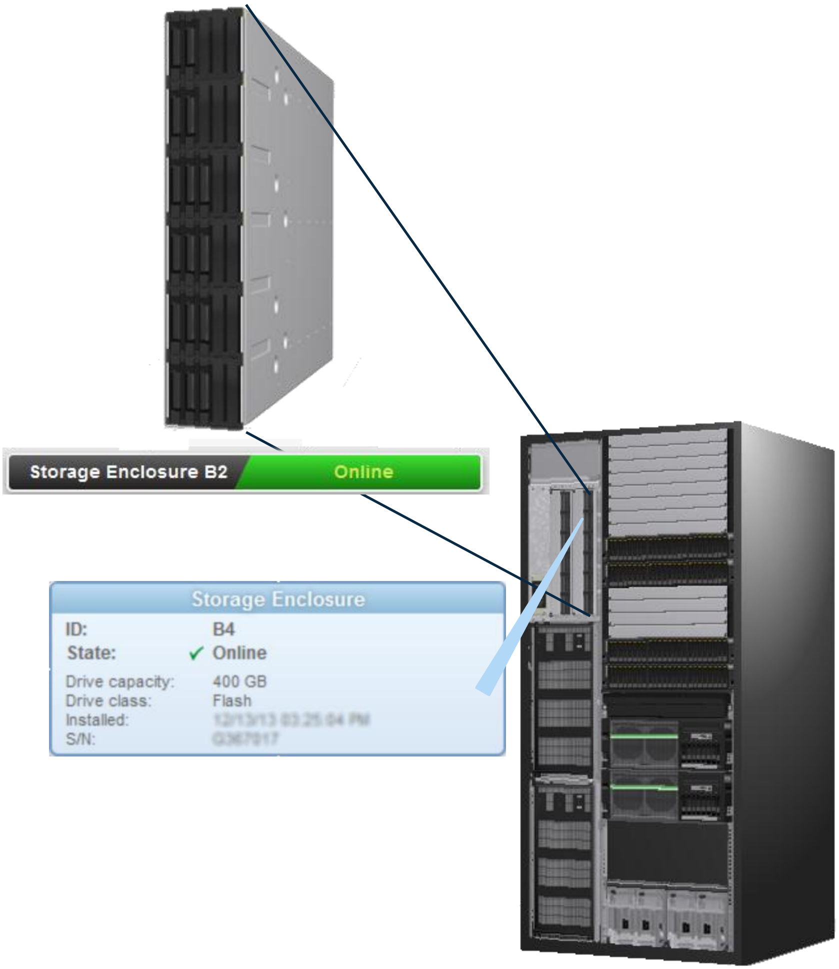
Switched Fibre Channel backend connectivity to 2.5” drives and 3.5” Nearline drives

Dual POWER7+ processors with up to 32 cores and 1TB memory in total

PCI-e fabric provides high speed resilient internal connectivity between components

High Performance Flash Enclosure

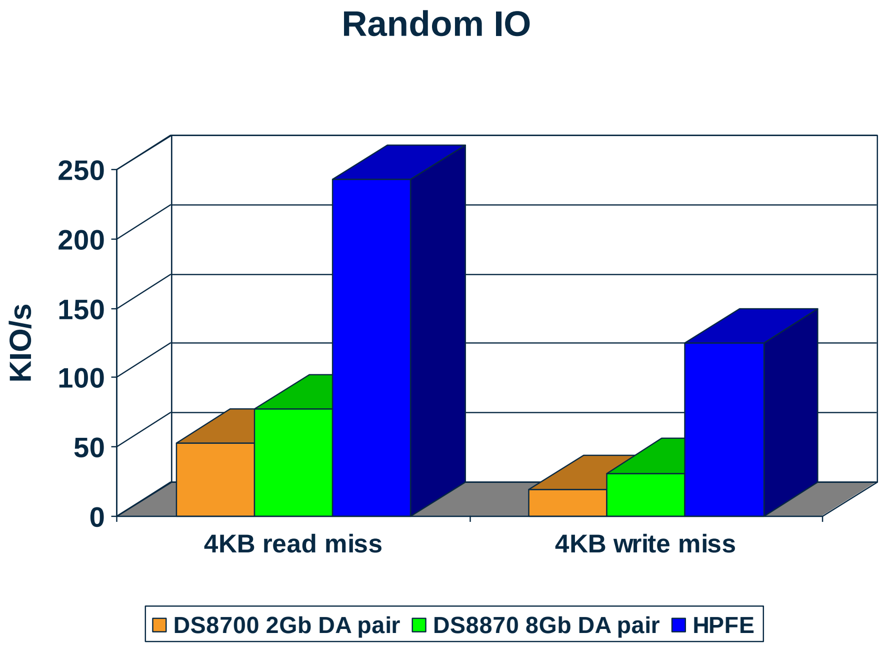
- Flash Optimized RAID for ultra-high performance
- 1U Storage Enclosure with up to 30 1.8” 400GB Flash drives
- 9.1 TiB usable RAID5 capacity per enclosure
- Fully redundant with integrated power and cooling
- Supports encryption using FDE Flash drives



Storage Enclosure

ID: B4
State: ✓ Online
Drive capacity: 400 GB
Drive class: Flash
Installed: [redacted]
S/N: [redacted]

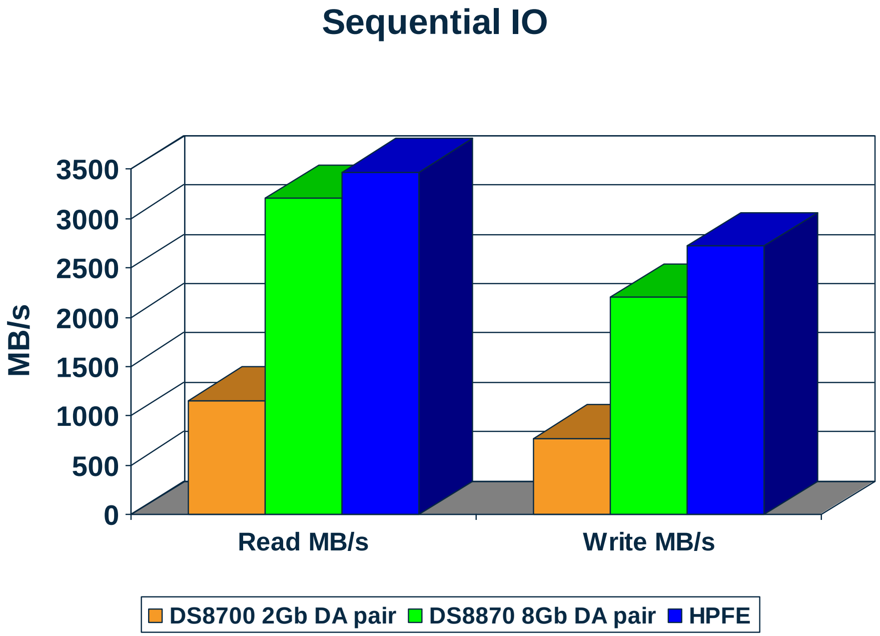
HPFE Performance



- Massive improvements in throughput for random workloads for High Performance Flash Enclosure compared to DS8870 device adapter
  - Up to 5.5x increase for random write
  - Up to 4x increase for random reads

- Single High Performance Flash Enclosure has modest increase in sequential throughput capability compared to DS8870 device adapter
  - 16% for sequential read
  - 23% for sequential write

- DS8870 device adapter already provided significant sequential throughput improvements



DS8870 Scalability

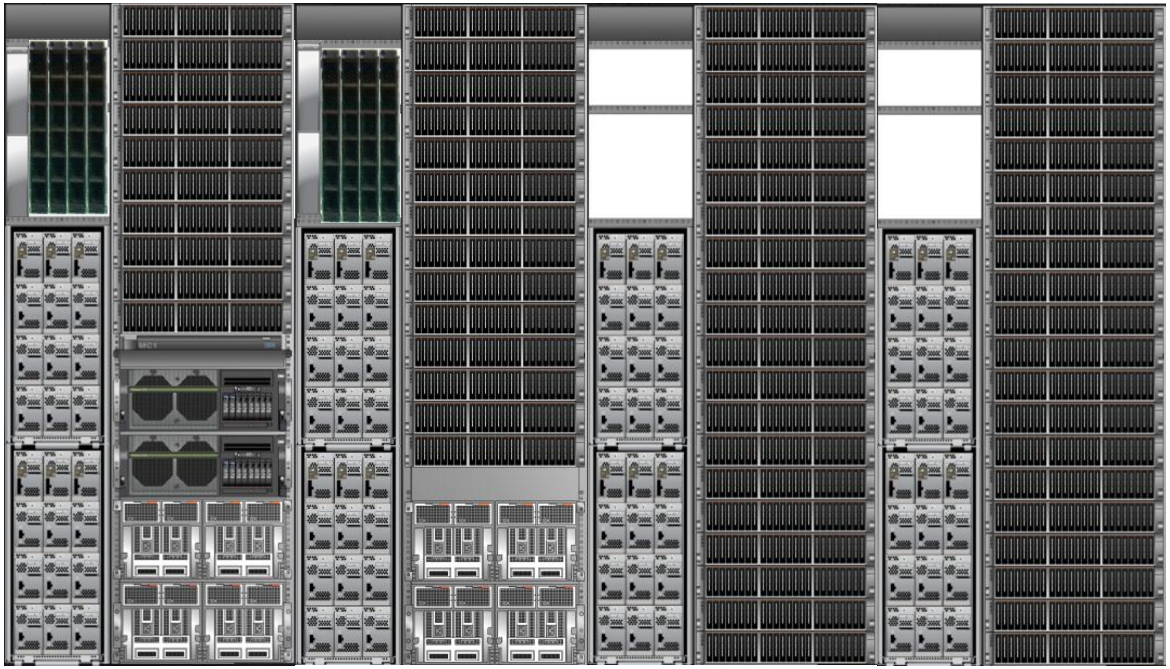
- Up to 4 frames with 1536 2.5” drives and 240 Flash drives
  - Third and fourth frames can be located separately by RPQ
- Up to 128 host adapter ports
  - Each port separately configurable as FICON or FC
- Up to 1TB Power7+ memory used for read and write cache
- Up to 65280 volumes

240 2.5” drives
120 Flash drives

336 2.5” drives
120 Flash drives

480 2.5” drives

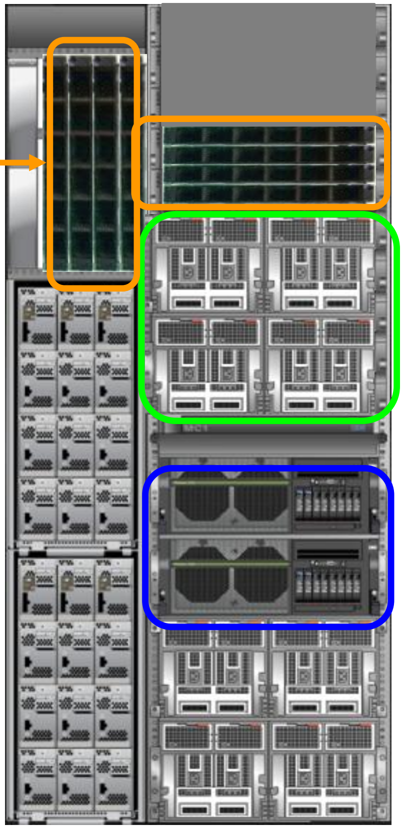
480 2.5” drives



8 Host Adapters
16 Device Adapters

8 Host Adapters
16 Device Adapters

All-Flash DS8870 with HPFE

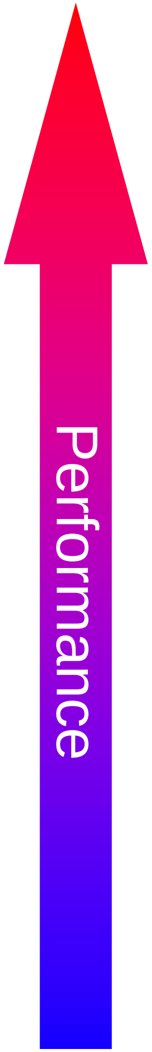


Up to 8 High Performance Flash Enclosures provide up to 96TB raw capacity with 400GB drives

All 8 IO bays installed in base frame for maximum throughput and host connectivity in single frame

8-core 256GB cache or 16-core 512GB cache or 16-core 1024GB cache

DS8870 Drive Technology

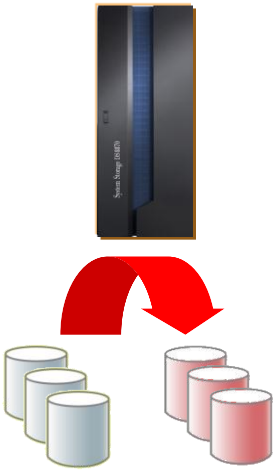


- Flash – 1.8” in High Performance Flash Enclosure
  - 400 GB drive
- SSD – 2.5” Small Form Factor
  - Latest generation with higher sequential bandwidth
  - 200/400/800GB/1600GB SSD
- 2.5” Enterprise Class 15K RPM
  - Drive selection traditionally used for OLTP
  - 146/300/600GB drives
- 2.5” Enterprise Class 10K RPM
  - Large capacity, much faster than Nearline
  - 600GB and 1.2TB drives
- 3.5” Nearline – 7200RPM Native SAS
  - Extremely high density, direct SAS interface
  - 4TB drives

DS8000 Replication Technologies

FlashCopy
Point in time copy

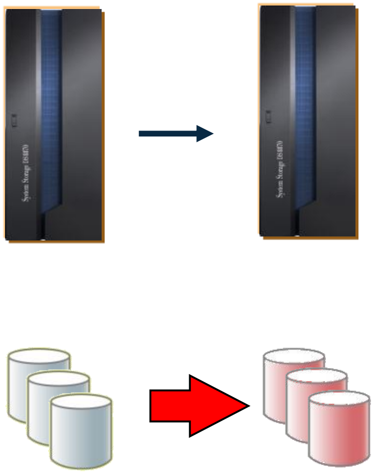
Within the same Storage System



Metro Mirror
Synchronous mirroring

Primary
Site A

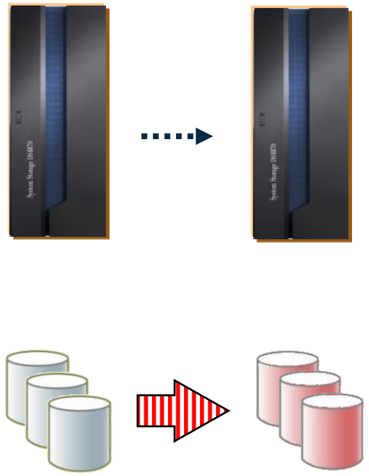
Metro distance
Site B



Global Mirror
zGlobal Mirror
Asynchronous mirroring

Primary
Site A

Out of Region
Site B

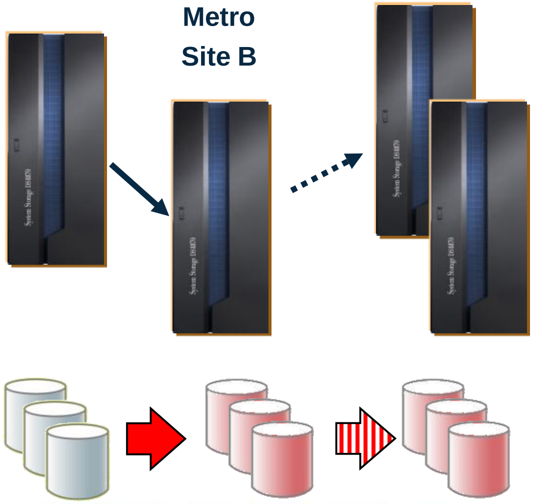


Metro Global Mirror
Metro zGlobal Mirror
Three site and Four Site synchronous & asynchronous mirroring

Primary
Site A

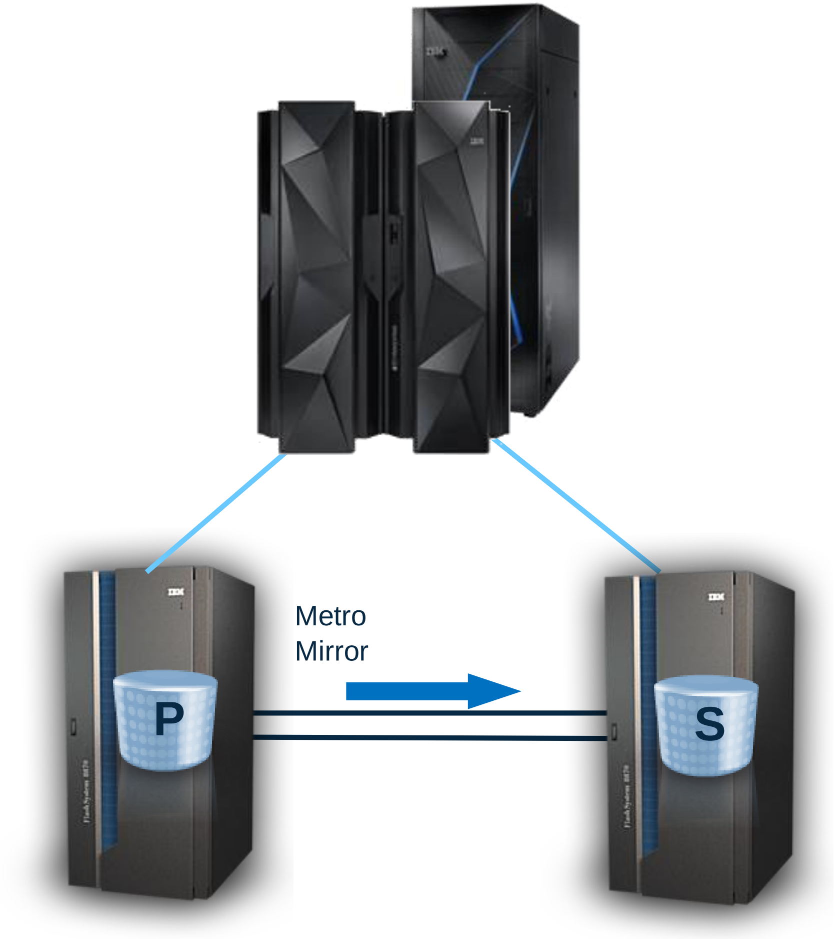
Metro
Site B

Out of Region
Site C/D



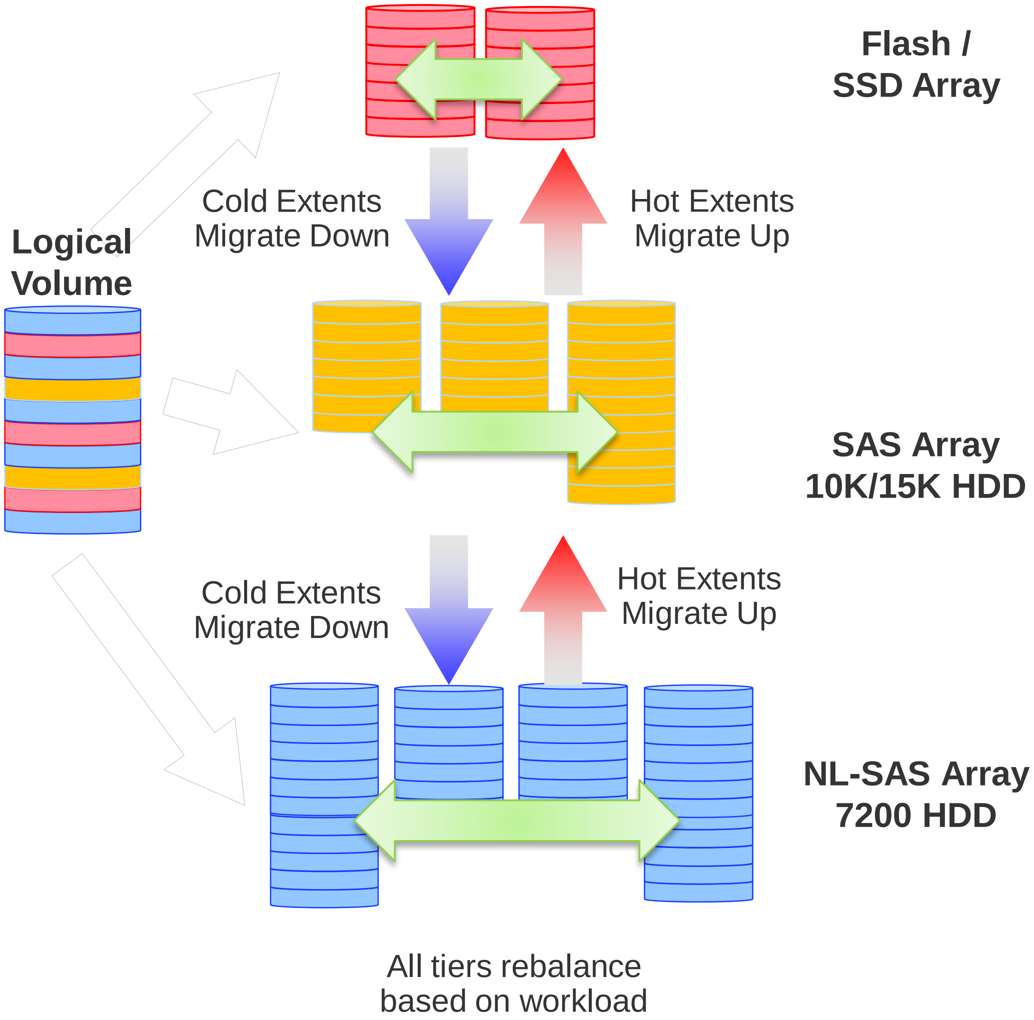
HyperSwap for Continuous Availability

- Substitutes storage secondary to take the place of failed primary device
  - Non-disruptive - applications keep running
  - Key value add to HA/DR deployments
- Unplanned HyperSwap
  - Continuous Availability in case of storage failures
- Planned HyperSwap
  - Storage Maintenance without downtime
  - Storage migration without downtime
- Typical swap times of a few seconds
- Supported on z/OS, zLinux, AIX and IBM I
  - Delivered by GDPS, TPC-R and PowerHA

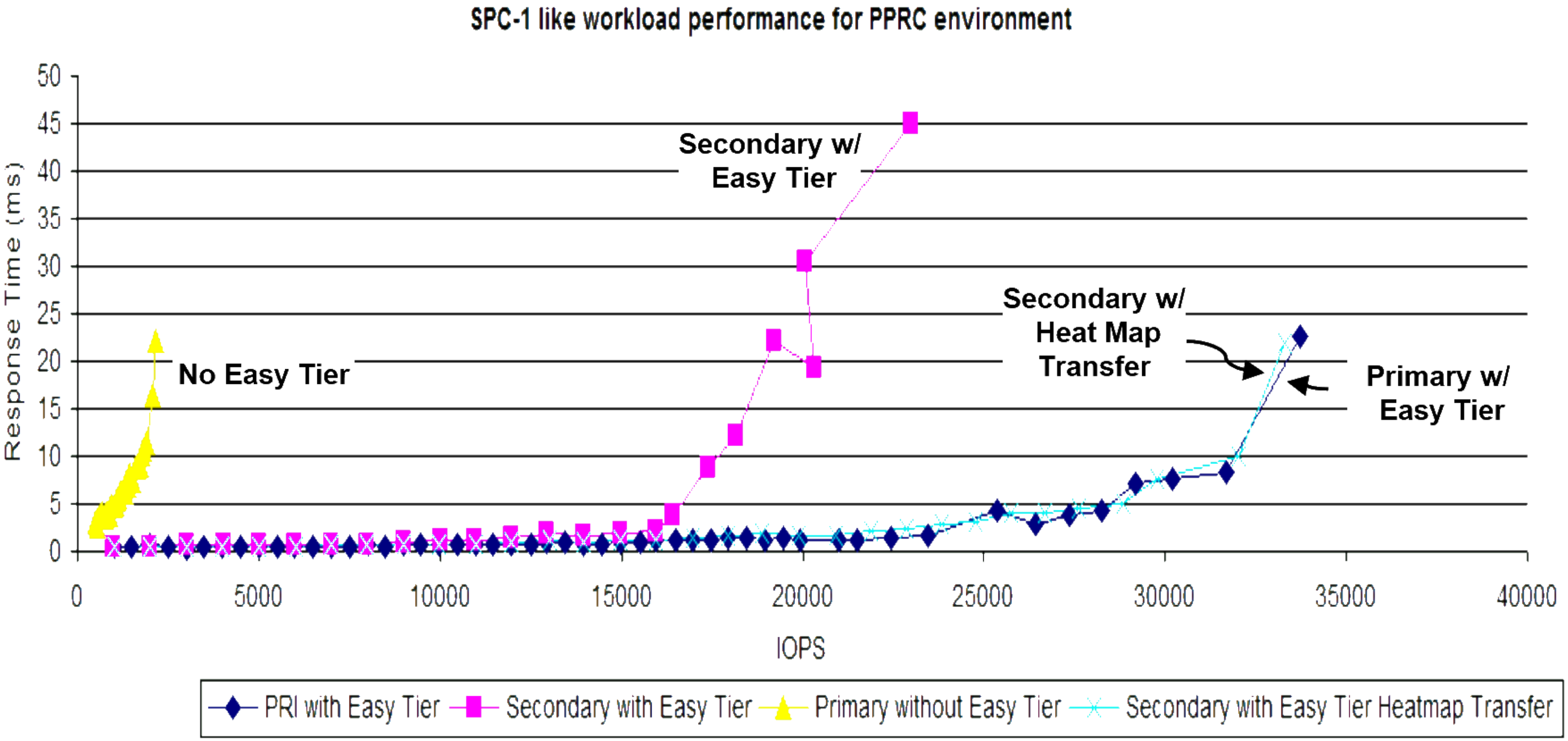


Easy Tier automated tiering

- Optimization of backend storage resources based on historical performance data
- SubLUN granularity using native DS8000 extents for any volume type
- Flexible configurations with any combination of drives of any size and speed
- Easy Tier Application provides APIs for policy and proactive actions
- Easy Tier HeatMap transfer enables workload history to be transferred for replication scenarios (DR, migration etc.)



Heat Map Transfer Measurement

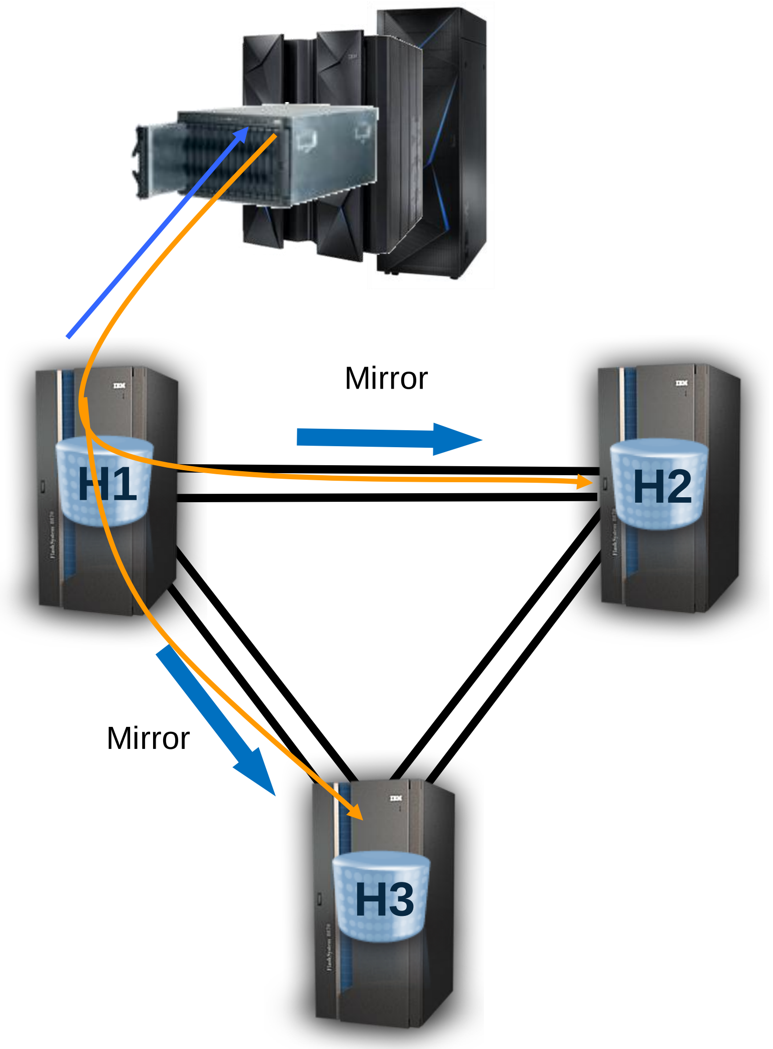


DS8870 Release 7.4 Content

- **Key Hardware Enhancements to the DS8870**
  - High Performance Flash Enclosures in First Expansion Rack
  - Encrypting 1600GB 2.5” SSDs
  - Encrypting 600GB 2.5” 15K RPM HDDs

- **Key Function Enhancements to the DS8870**
  - Multiple Target PPRC
  - PPRC Synchronization improvements
  - Global Copy Collision Avoidance
  - Multiple Incremental Flash Copy
  - zHyperWrite : DB2 Log Write Acceleration with Metro Mirror
  - zGM (XRC) Workload Based Write Pacing
  - Easy Tier Policy controls
  - zEasy Tier Application DFSMS and DB2 integration
  - Multi-thread Performance Accelerator
  - Enhanced User Interface

Multiple Target PPRC



- Allow a single volumes to be the source for more than one PPRC relationship
- Provide incremental resynchronisation functionality between target devices
- Use cases include
  - Synchronous replication within a datacentre combined with another metro distance synchronous relationship
  - Add another synchronous replication for migration without interrupting existing replication
  - Allow multi-target Metro Global Mirror as well as cascading for greater flexibility and simplified operational scenarios
  - Combine with cascading relationships for 4-site topologies and migration scenarios
- TPC-R and GDPS support for Multi-target Metro Mirror

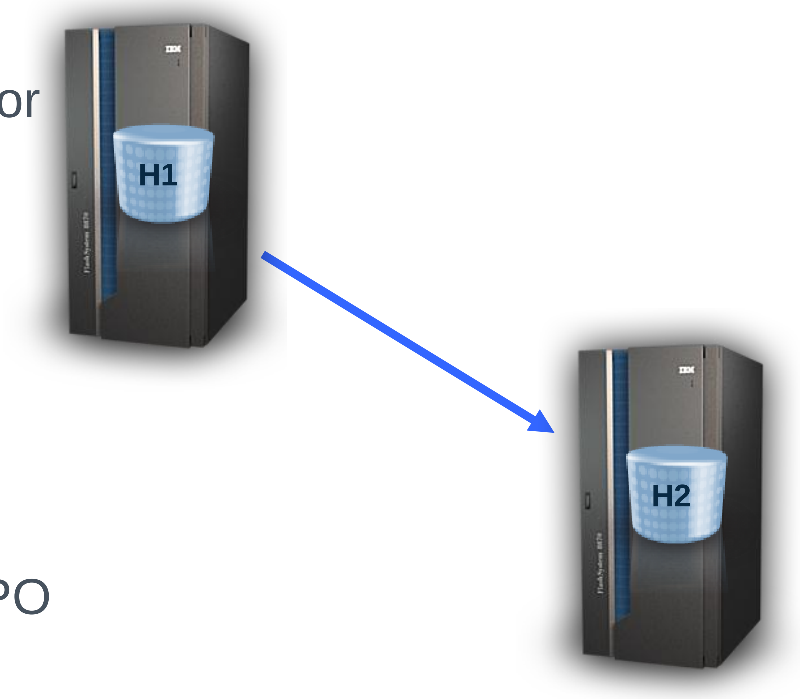
PPRC Synchronization Improvement

- The asynchronous copying of data from a PPRC primary to a secondary.

- Copies data that is out-of-sync between primary and secondary
  - Initial copy when a pair is established or resumed
  - Global Copy / Global Mirror to asynchronously transfer updated data

- To form a Consistency Group (CG), a group of changed data to the remote.
  - Synchronization is key to how fast this gets done which directly influences RPO

- New design introduced in R7.4



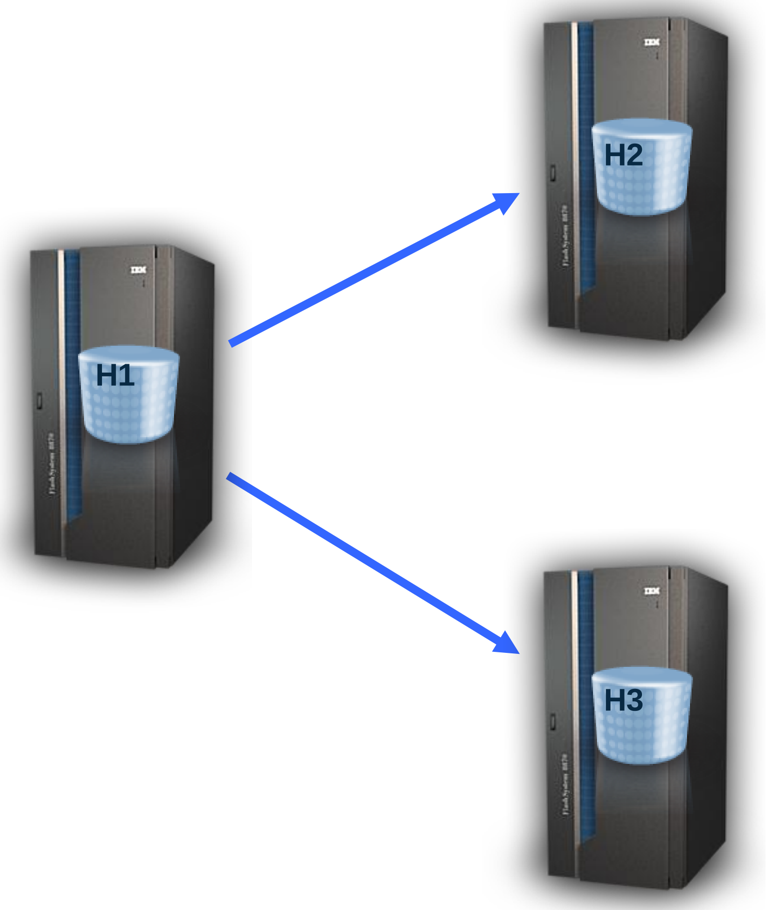
PPRC Synchronization - Previous Design

- Volume based
  - When a volume spans ranks, only the part on one rank copied at a time
- No priority mechanism
- Did not scale with volume size
  - Resources allocated per volume, regardless of size
  - Volumes broken into 5 pieces, one copy process per piece

5GB vol
5 x 1GB pieces
500GB vol
5 x 100GB pieces

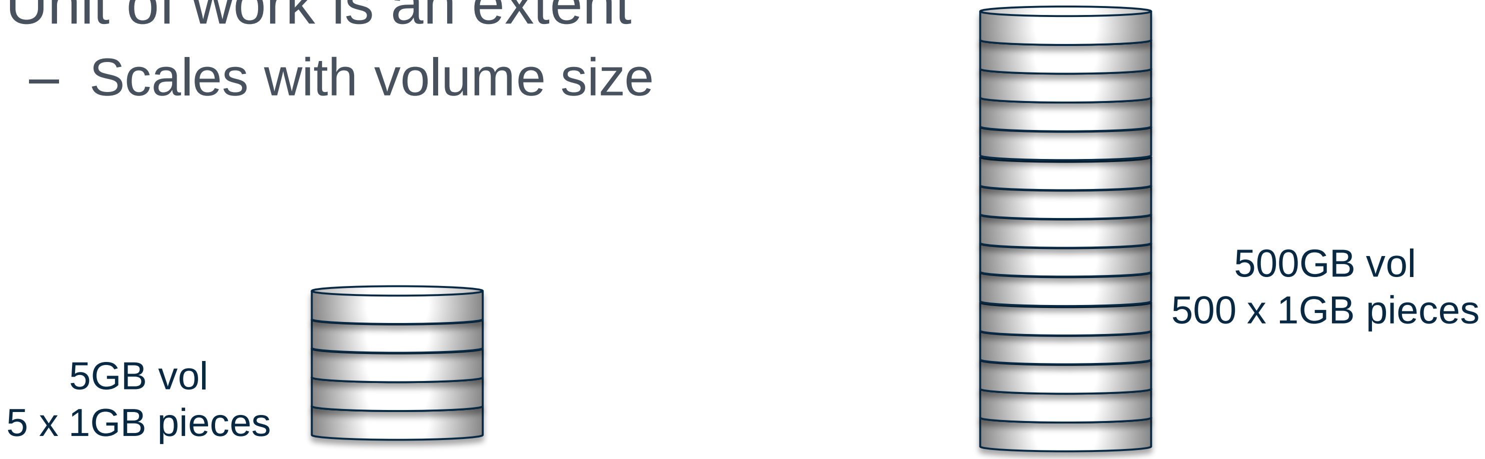
PPRC Synchronization - New Design (R7.4)

- Objectives of new design:
  - Support Multiple Target PPRC
  - Finish the copy as quickly as possible
    - Fully utilize the PPRC links
  - Minimize the impact on other work
    - Do not overdrive the ranks on the primary
    - Minimize impact on host I/O
  - Do the most important work first
    - Priority scheme



PPRC Synchronization - New Design (R7.4)

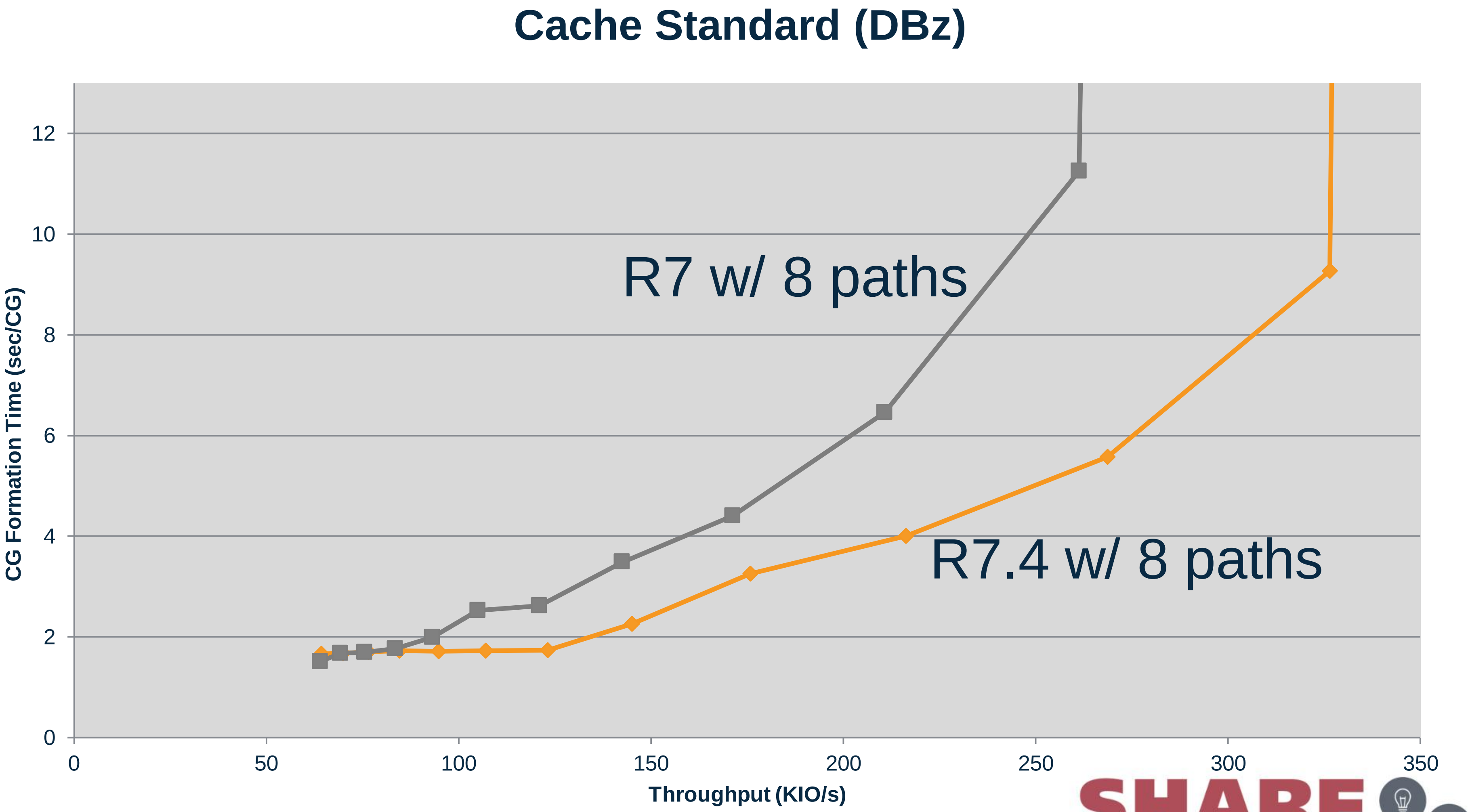
- Balances workload across:
  - PPRC Ports
  - Extent Pools
  - Device Adapters
  - Ranks
- Assigns priorities
  - For example, forming GM consistency groups > Resynchronization
- Unit of work is an extent
  - Scales with volume size



5GB vol
5 x 1GB pieces

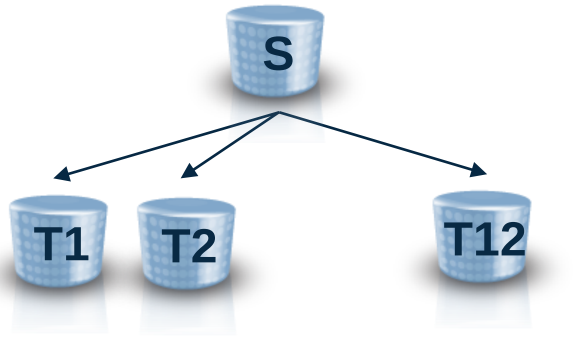
500GB vol
500 x 1GB pieces

PPRC Synchronization – GM Measurements



Multiple Incremental FlashCopies

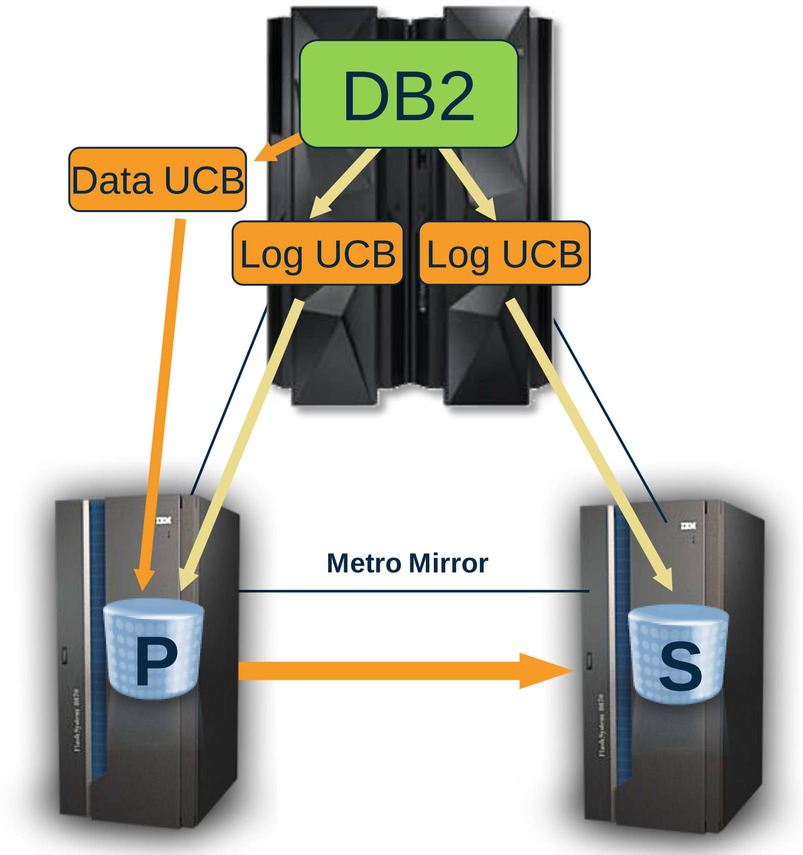
- Previously only a single incremental FlashCopy was allowed for any individual volume
- This provides the capability for up to 12 incremental FlashCopies for any volume
- A significant number of clients take two (or more) FlashCopies per day for database backup both of which can now be incremental
- The Global Mirror journal FlashCopy also counts as an incremental FlashCopy so the testing copy can now also be incremental
- The functionality is also available as an RPQ from R7.1.5



zHyperWrite – DB2 log writes

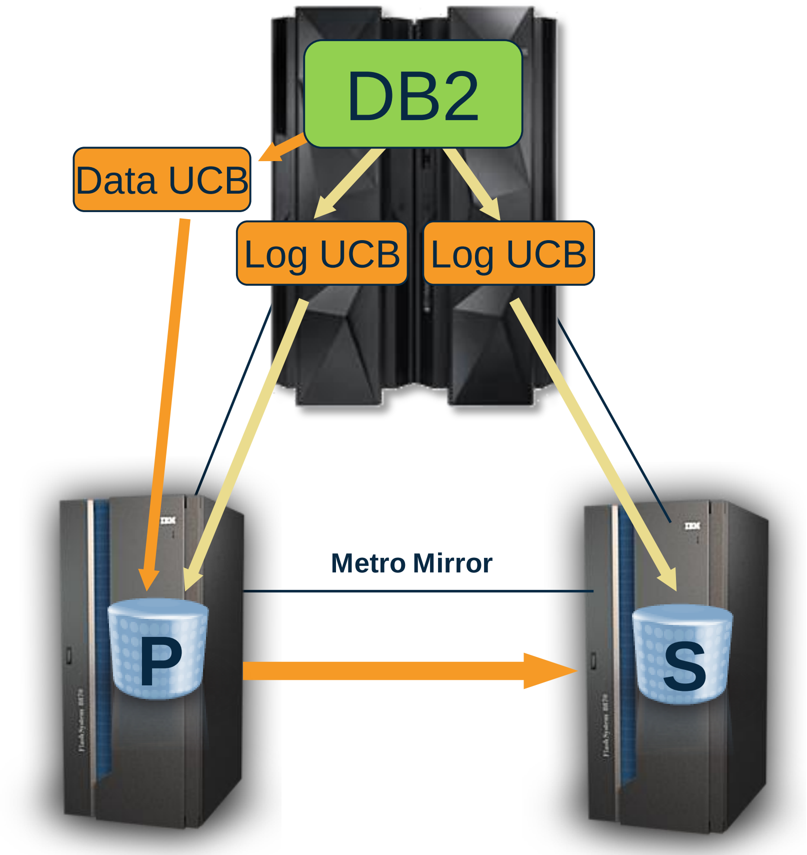
- DB2 log activity is heavy and Metro Mirror creates latency

- zHyperWrite improves DB2 Log Write Performance with DS8870 Metro Mirror
  - Reduced write latency and improved log throughput



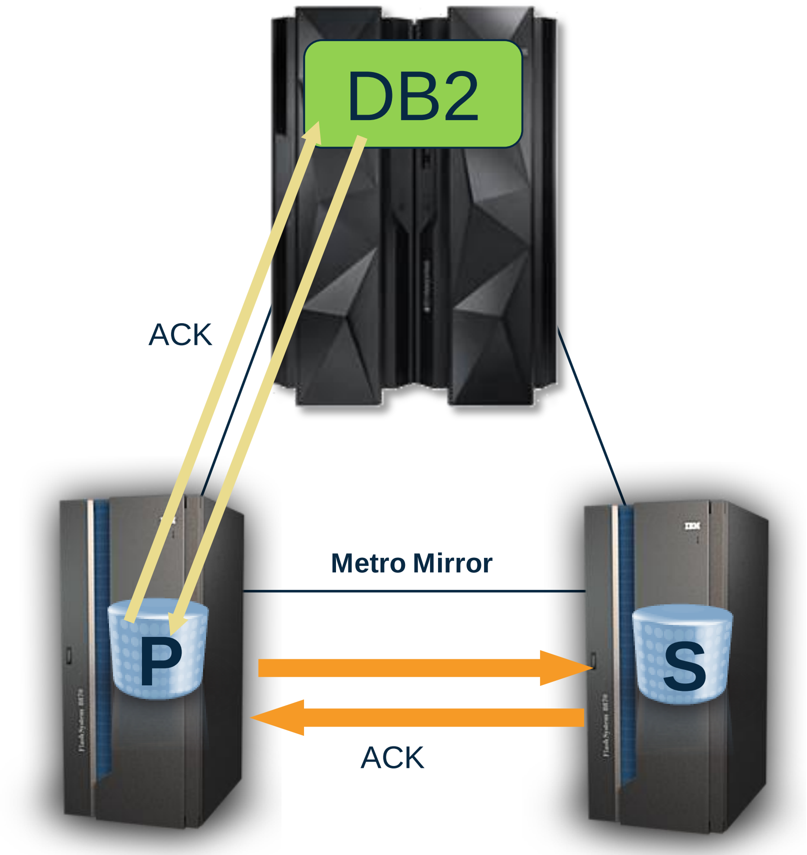
zHyperWrite – DB2 log writes

- Split DB2 Log writes into separate parallel writes to primary and secondary
- Reduce channel and mirror activity for log volumes
- Response time and throughput benefits
  - Amount of improvement varies with Metro Mirror distance and commit frequency



DB2 Log Write with Metro Mirror

1. DB2 Log Write to Metro Mirror Primary
2. Write Mirrored to Secondary
3. Write Acknowledged to Primary
4. Write Acknowledged to DB2

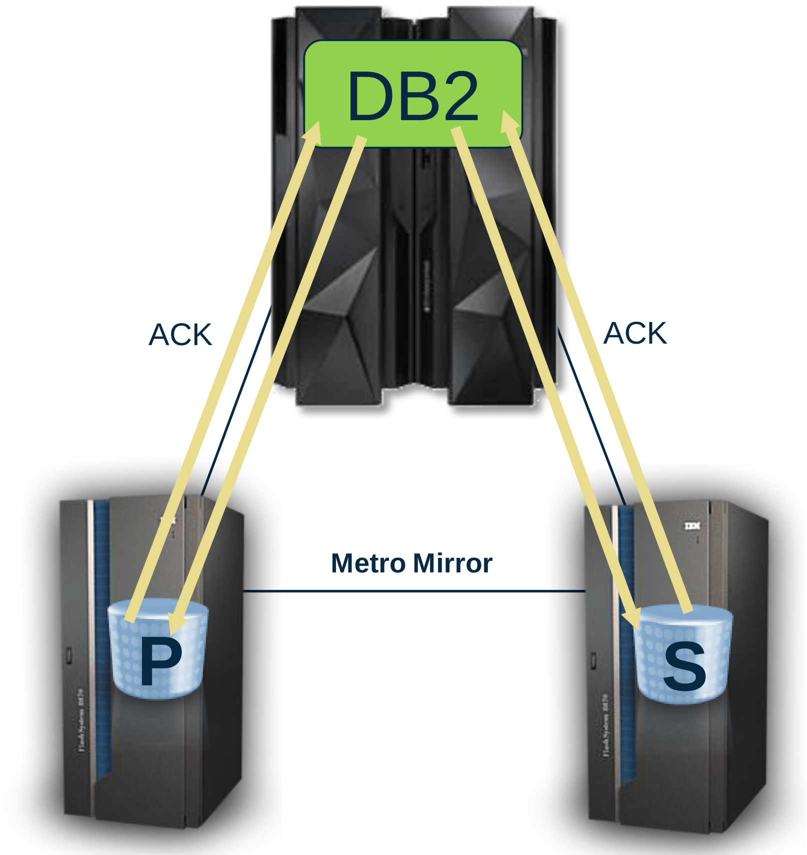


Write with zHyperWrite

1. DB2 Log Write to Metro Mirror Primary and Secondary in parallel

2. Writes Acknowledged to DB2

3. Metro Mirror does not mirror the data.



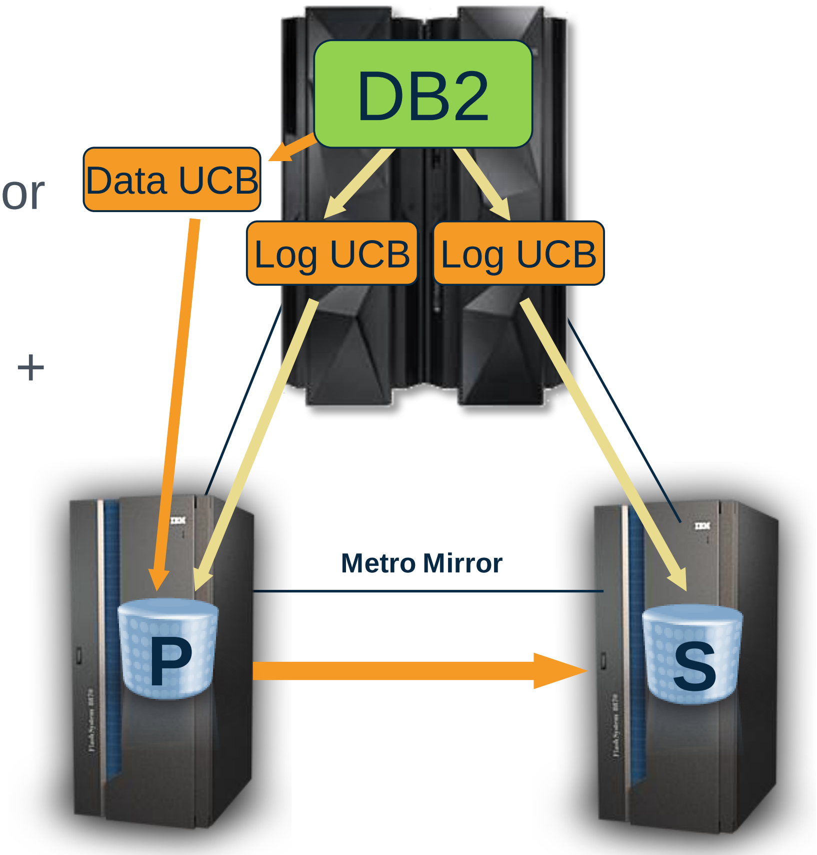
zHyperWrite – DS8870

- If the DS8870 detects that the volume pair is not in the correct state (i.e. full duplex) the write is rejected. DB2 re-drives without zHyperwrite.

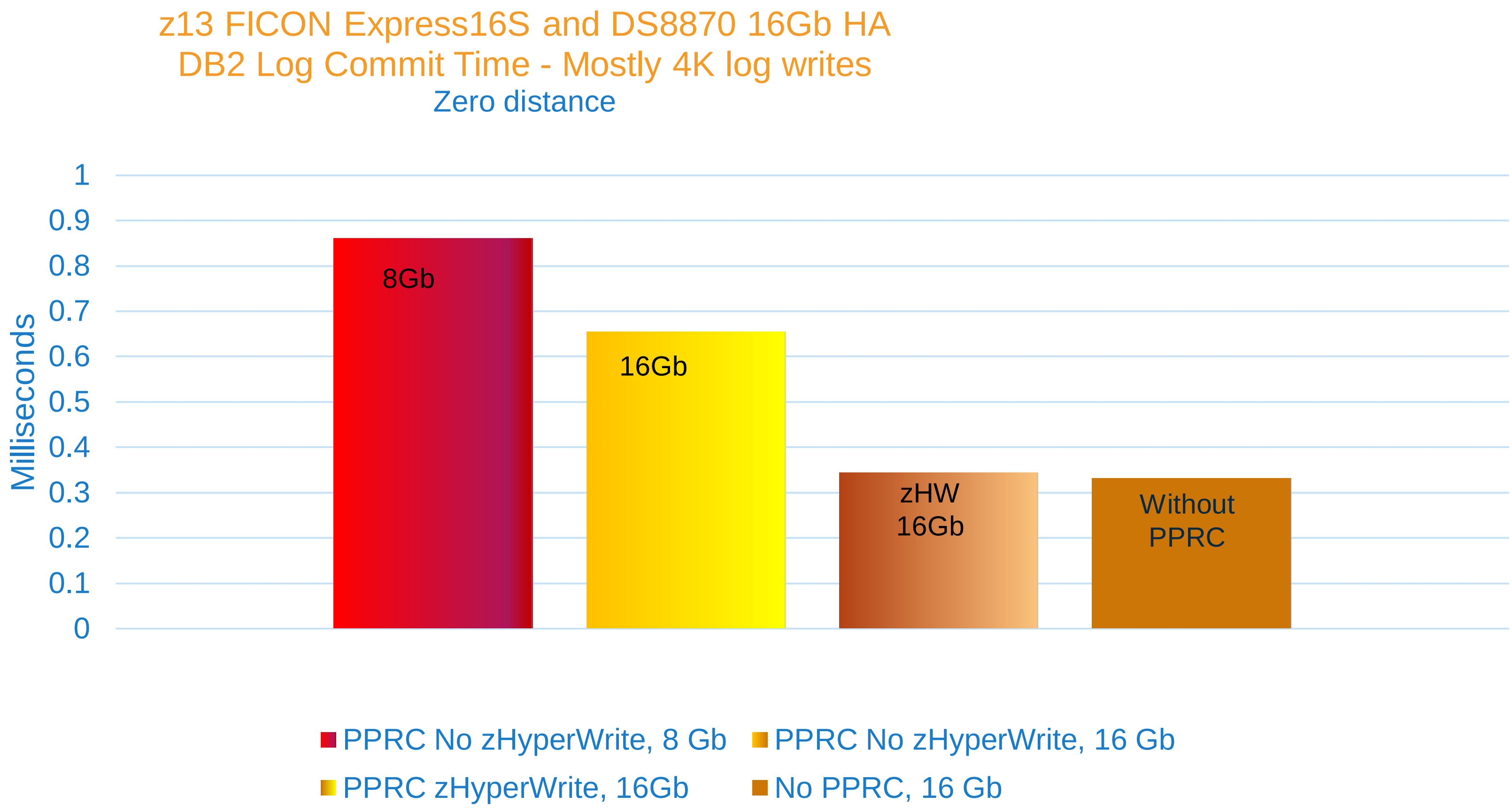
- Multiple Target Metro Mirror
  - zHyperWrite writes to both Metro Mirror targets

- Multiple Target PPRC with Metro Mirror + Global Copy/Global Mirror
  - zHyperWrite writes to Metro Mirror target
  - DS8870 mirrors for Global Copy / Global Mirror

- Supports HyperSwap with GDPS and TPC-R



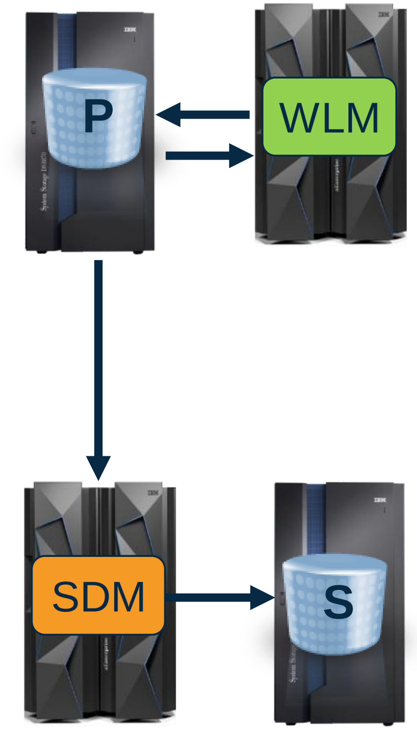
zHyperWrite Performance



- Up to 61% reduction in DB2 Commit response time
- zHyperWrite performance at zero distance is equivalent to non-PPRC

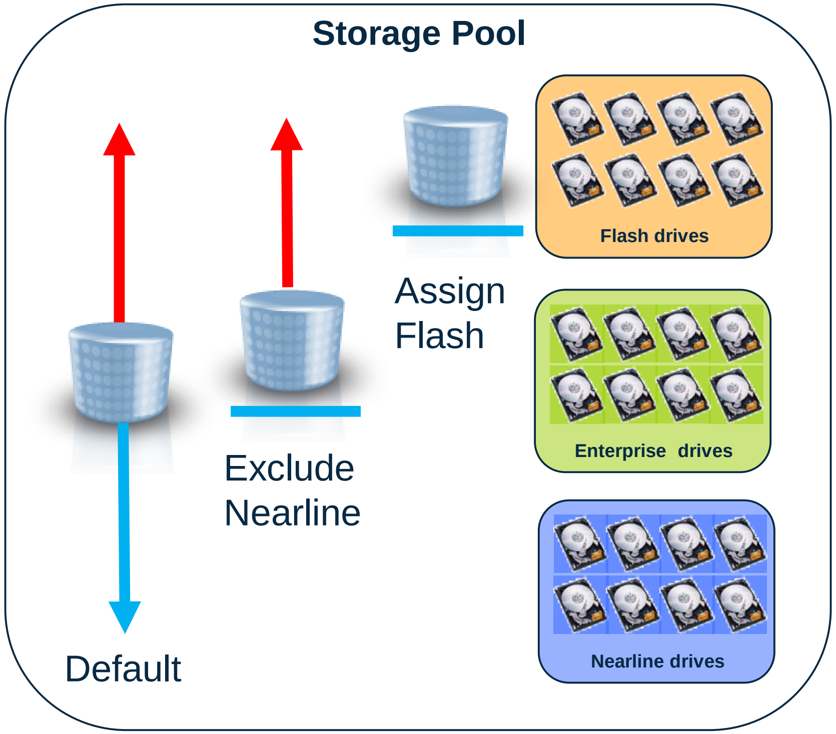
Workload Based Write Pacing for zGM (XRC)

- Software Defined Storage enhancement to allow System z Workload Manager to control XRC Write Pacing
- Reduces administrative overhead on hand managing XRC write pacing
- Avoids the need to define XRC write pacing on a per volume level allowing greater flexibility in configurations
- Prevents low priority work from interfering with the Recovery Point Objective of critical applications
- Enables consolidation of workloads onto larger capacity volumes



Easy Tier Application Policies

- New Exclude Nearline tier assignment policy
- Prevents the extents of a volume from being demoted to Nearline arrays
- If data is already on Nearline it will be promoted to Enterprise drives
- Three common use cases for Easy Tier Application policies
  - Default – optimise use of all tiers
  - Exclude Nearline – avoid potential low performance
  - Assign Flash – high performance guaranteed
- Also possible to assign to Enterprise or assign to Nearline but less common use cases



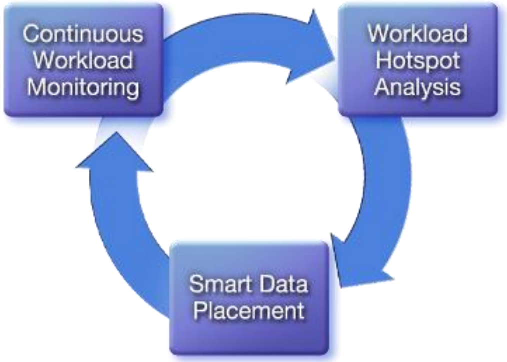
Complete your session evaluations online at www.SHARE.org/Orlando-Eval

Easy Tier Controls

- In the majority of environments Easy Tier is able to successfully use the history of workload performance to predict the future requirements
  - There are however cases where this is not true

- Easy Tier Controls provide mechanisms for proactively and reactively modifying Easy Tier behaviour to handle these situations

- Controls include
  - Pause and Resume Easy Tier learning for volume or pool
  - Reset Easy Tier learning for volume or pool
  - Pause and Resume Easy Tier migration for a pool

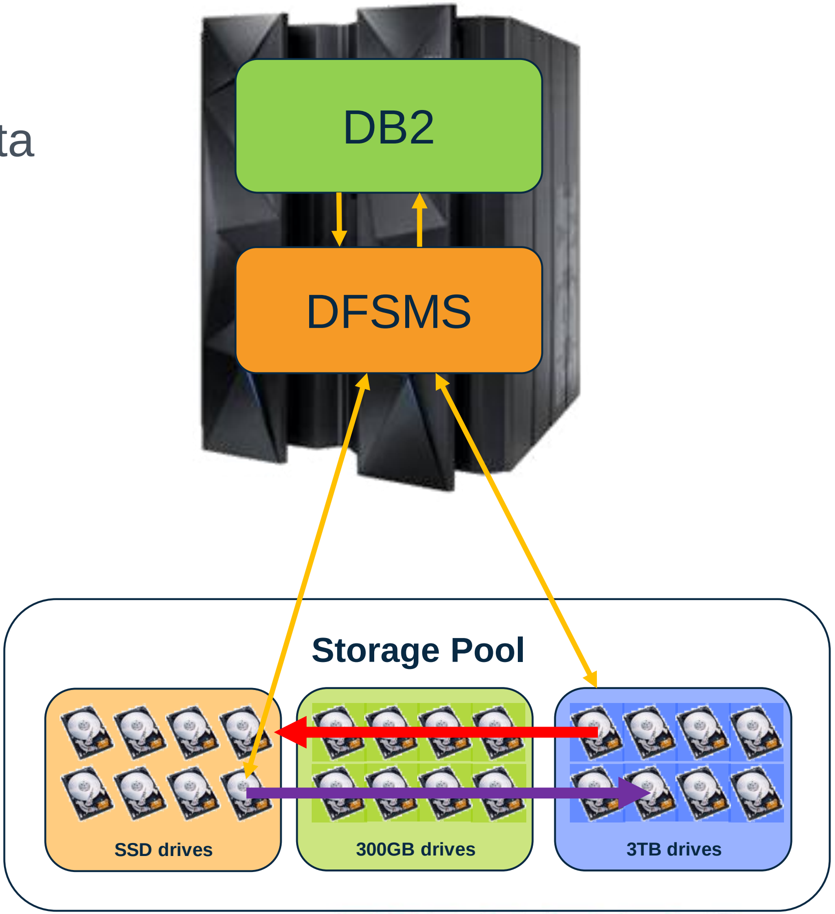


Easy Tier Application Integration with DFSMS and DB2

- Easy Tier currently optimises data placement and tiering based on workload history and this does not always reflect the future performance requirements of the data

- Easy Tier will provide interfaces to enable software such as DFSMS and DB2 to provide hints when data has been created, moved or deleted

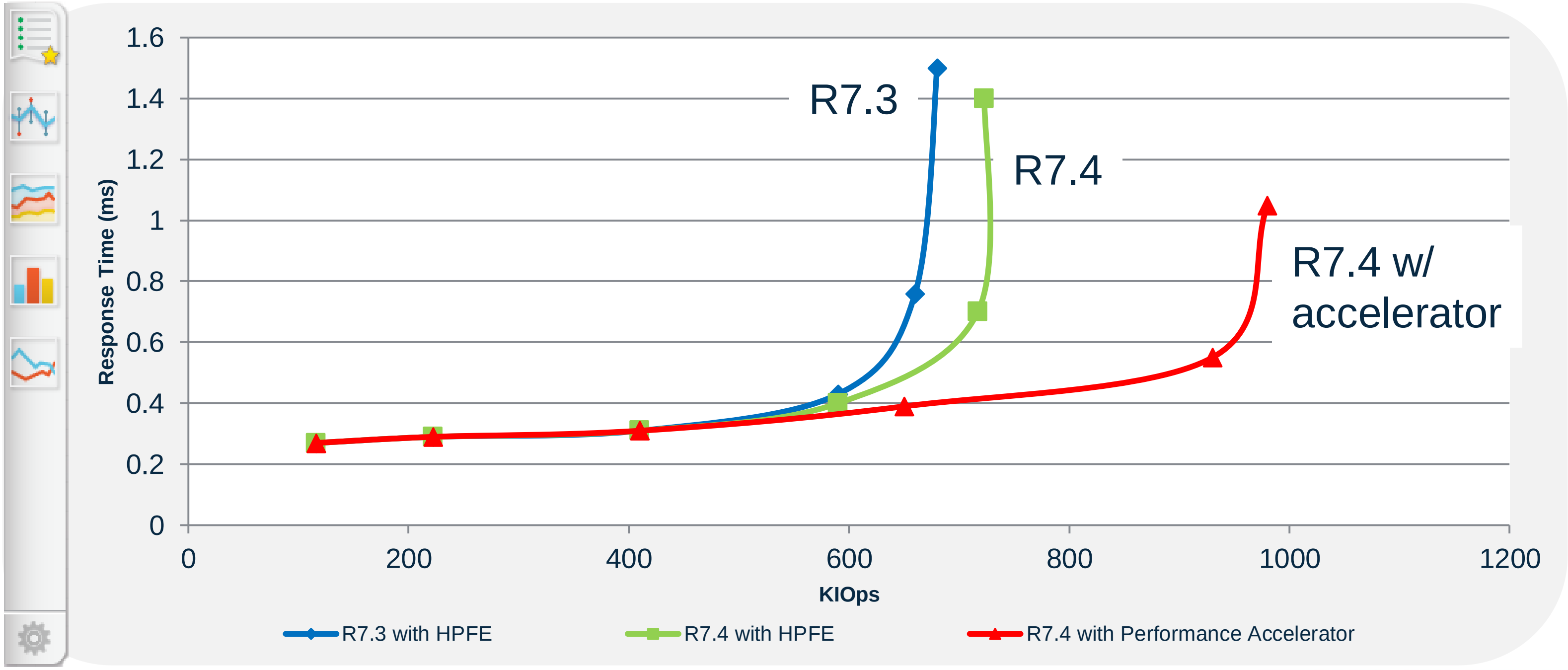
- This will avoid performance degradation following maintenance activities such as database reorganisation



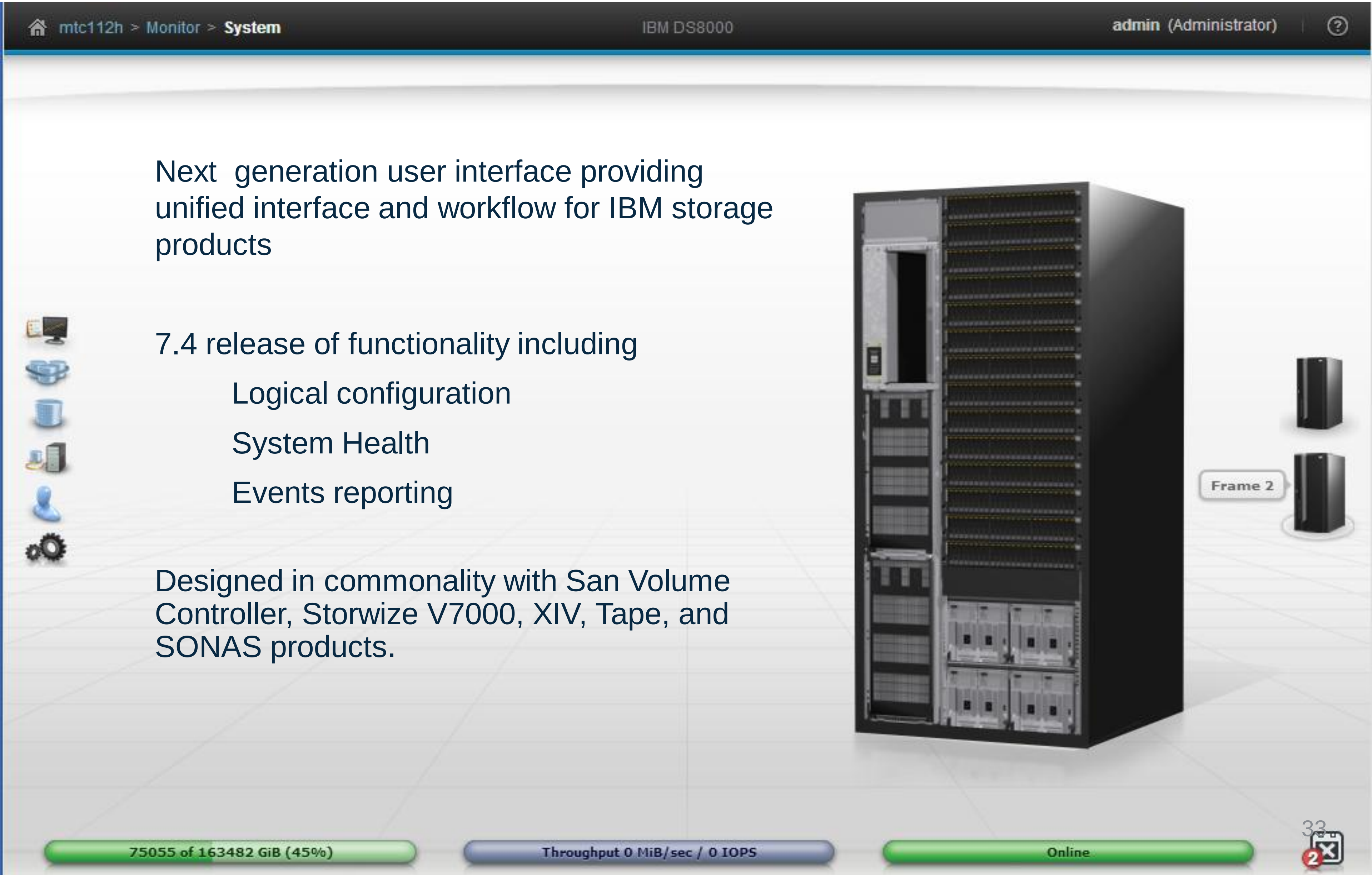
Multi-thread Performance Accelerator

- The IBM Multi-thread Performance Accelerator provides up to 45% performance improvement when used with the High Performance Flash Enclosures

- Up to 10% improvement will be seen without the Performance Accelerator



DS8000 Unified User Interface



Next generation user interface providing unified interface and workflow for IBM storage products

7.4 release of functionality including

- Logical configuration
- System Health
- Events reporting

Designed in commonality with San Volume Controller, Storwize V7000, XIV, Tape, and SONAS products.

75055 of 163482 GiB (45%)

Throughput 0 MiB/sec / 0 IOPS

Online

DS8870 Release 7.5 Content

- **Key Hardware Enhancements to the DS8870**
  - 16Gb FCP/FICON Host Adapter

- **Key Function Enhancements to the DS8870**
  - Forward Error Correction
  - FC Read Diagnostic Parameters
  - FICON Dynamic Routing
  - Fabric IO Priority
  - zHPF Performance for Distance
  - Multithread Optimization
  - Heat Map Transfer
  - Performance reporting
  - RESTful API

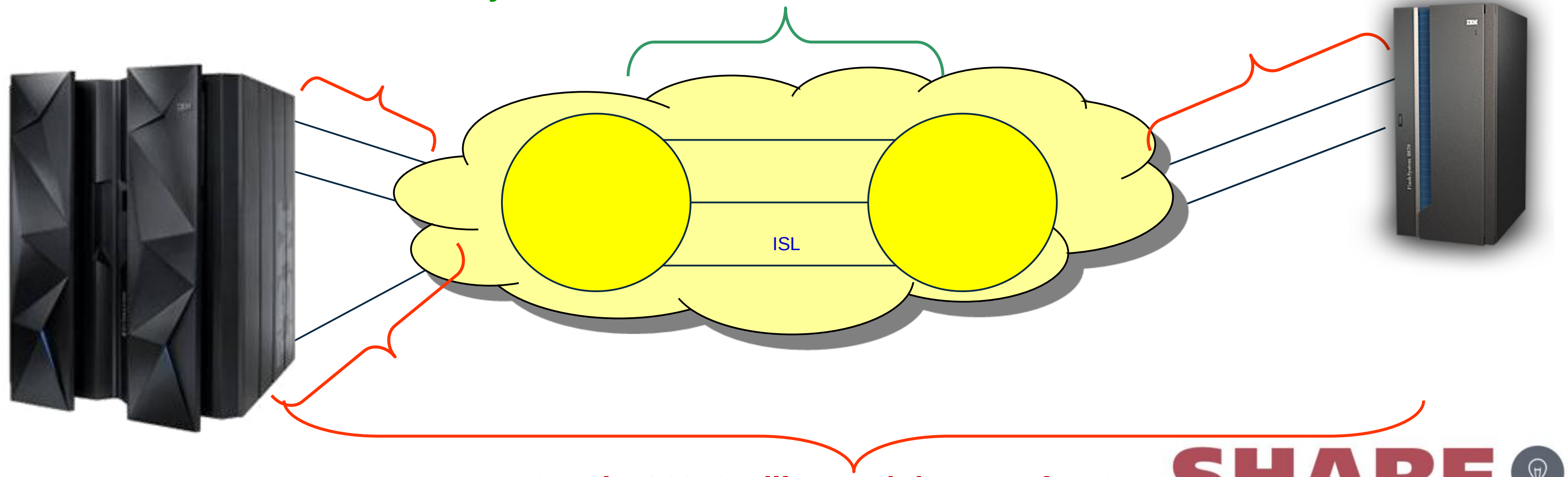
16Gb Host Adapter – FCP and FICON

- 16Gb connectivity reduces latency and provides faster single stream and per port throughput
  - 8GFC, 4GFC compatibility (no FC-AL Connections)
- Quad core Power PC processor upgrade
  - Dramatic (2-3x) IOPS improvements compared to existing 8Gb adapters (for both Z systems and distributed FCP)
- Forward Error Correction (FEC) for the utmost reliability and to smooth the transition to Gen 5 FC.
- Lights on Fastload avoids path disturbance during code loads
- Additional functional improvements for System z environments combined with 16Gb host channels
  - FICON dynamic routing
  - Fabric IO priority
  - zHPF extended distance performance

Forward Error Correction (FEC) Codes

- New standard for transmission of data on 16 Gbs Links
- T11.org FC-FS-3 standard defines use of 64b/66b encoding
  - Efficiency improved to 97% vs. 80% with 8b/10b encoding
- FEC codes provide error correction on top of 64b/66b encoding
  - Improves reliability by reducing bit errors (adds equivalent of 2.5 dB signal strength)
  - Up to 11 bit errors per 2112 bits can be corrected
  - IBM is leading new standards required to enable FEC for optical links

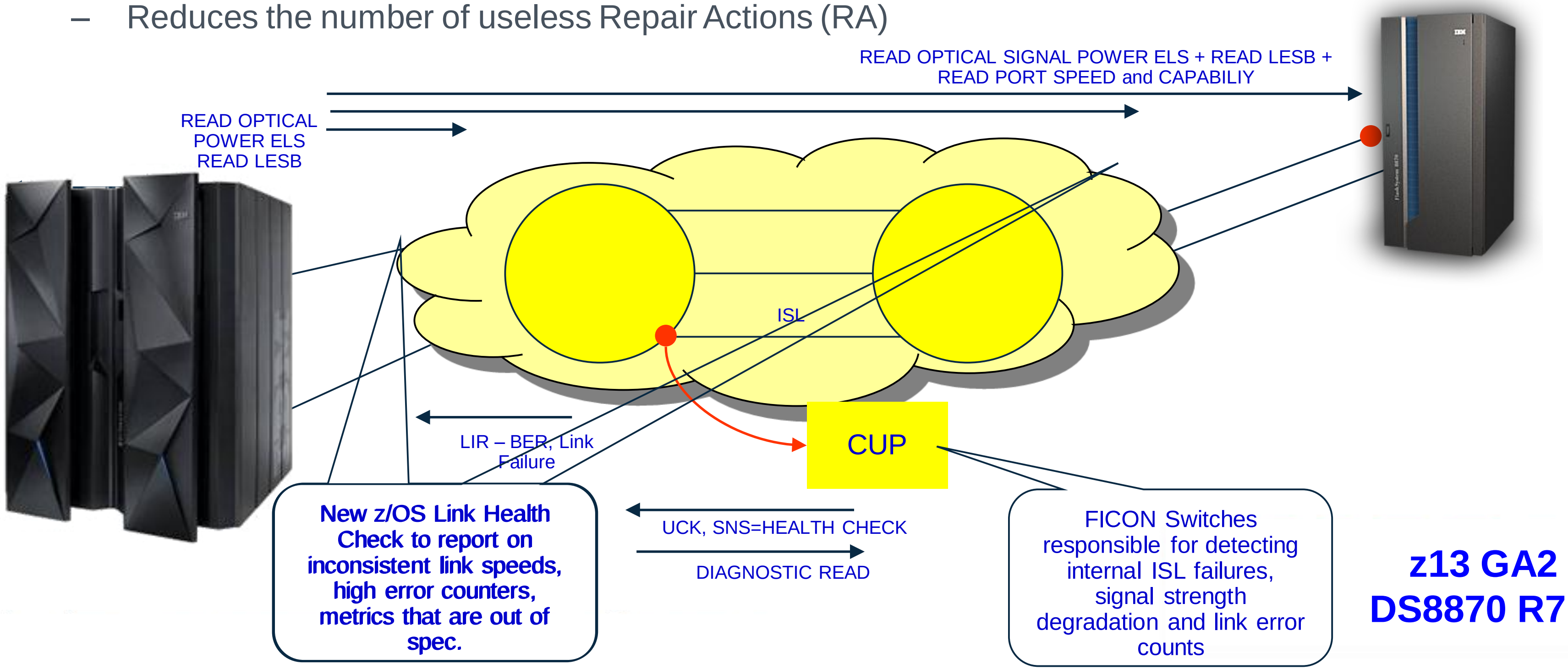
Proprietary Brocade Fabrics use FEC today to improve reliability of ISL connections in the FC fabric for 16 Gbs link



z13 and DS8870 will extend the use of FEC to the fabric N Ports for complete end-to-end coverage for new 16 Gbs FC links

Read Diagnostic Parameters

- Improved Fault Isolation
- After a link error is detected, link data returned from Read Diagnostic Parameters cab be used to differentiate between errors due to failures in the optics versus failures due to dirty or faulty links.
- Viewable on DS8870 DSCLI or Brocade switch
- Future System z Channel Subsystem Function
  - Periodic polling from the channel to the end points for the logical paths established
  - Improved fault isolation, key metrics displayable on operator console
  - Reduces the number of useless Repair Actions (RA)

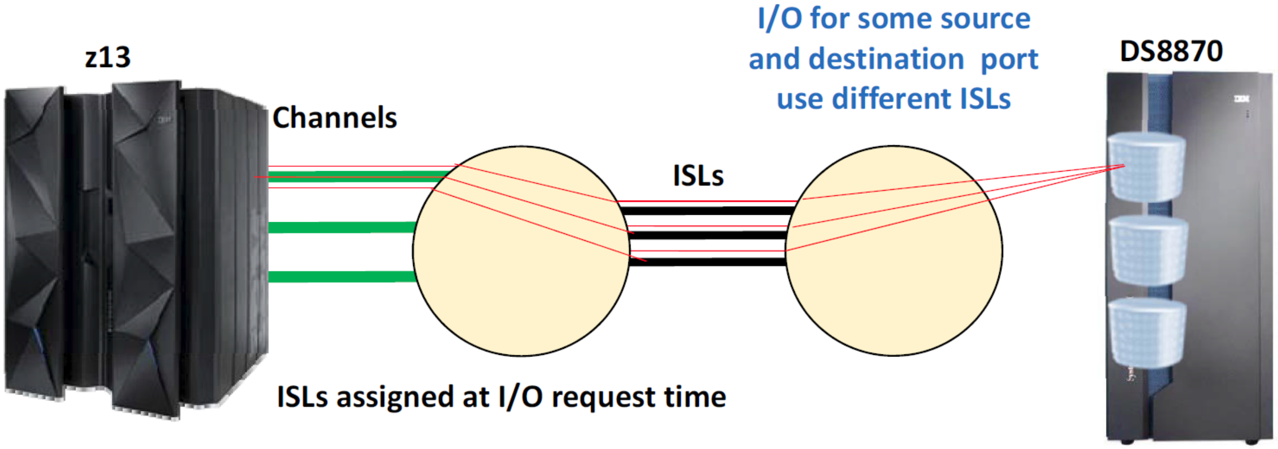

READ OPTICAL SIGNAL POWER ELS + READ LESB + READ PORT SPEED and CAPABILIY
READ OPTICAL POWER ELS READ LESB
ISL
CUP
LIR – BER, Link Failure
UCK, SNS=HEALTH CHECK
DIAGNOSTIC READ

New z/OS Link Health Check to report on inconsistent link speeds, high error counters, metrics that are out of spec.

FICON Switches responsible for detecting internal ISL failures, signal strength degradation and link error counts

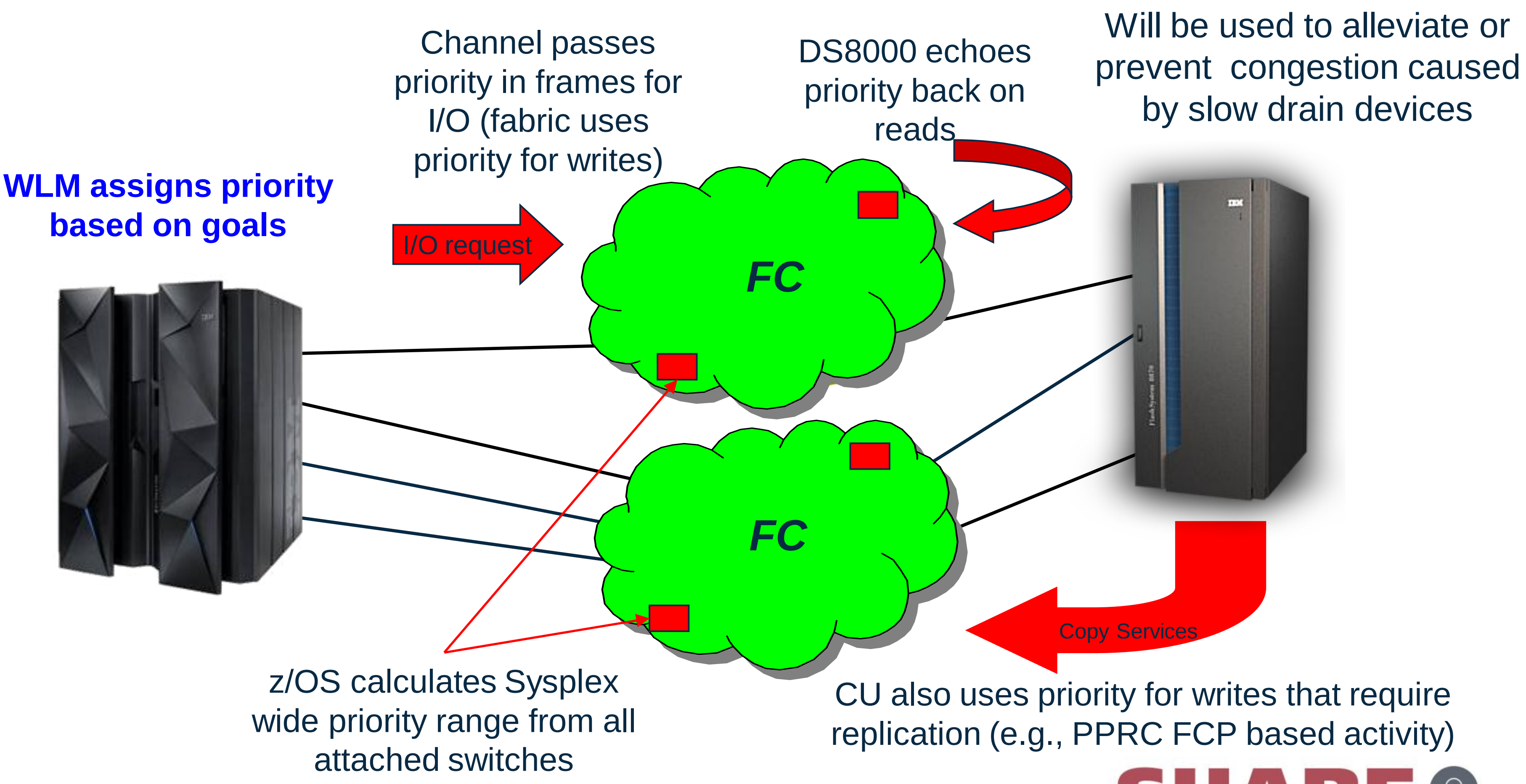
z13 GA2
DS8870 R7.5

Dynamic Routing (Exchange Based Routing)


z13
Channels
ISLs
I/O for some source and destination port use different ISLs
DS8870
ISLs assigned at I/O request time

- Dynamic Routing (Brocade EBR or CISCO OxID) dynamically changes the routing between the channel and control unit based on the "Fibre Channel Exchange ID". Each I/O operation has a unique exchange ID
  - I/O traffic is better balanced between all available ISLs
  - Improves utilization of switch and ISL hardware – ~37.5% bandwidth increase
  - Reduces cost by allowing sharing of ISLs between FICON, FCP (PPRC or distributed)
  - Easier to manage
  - Easier to do capacity planning for ISL bandwidth requirements
  - Predictable, repeatable I/O performance
  - Positions FICON for future technology improvements, such as work load based routing

Fabric I/O Priority for DS8870 and System z



WLM assigns priority based on goals

Channel passes priority in frames for I/O (fabric uses priority for writes)

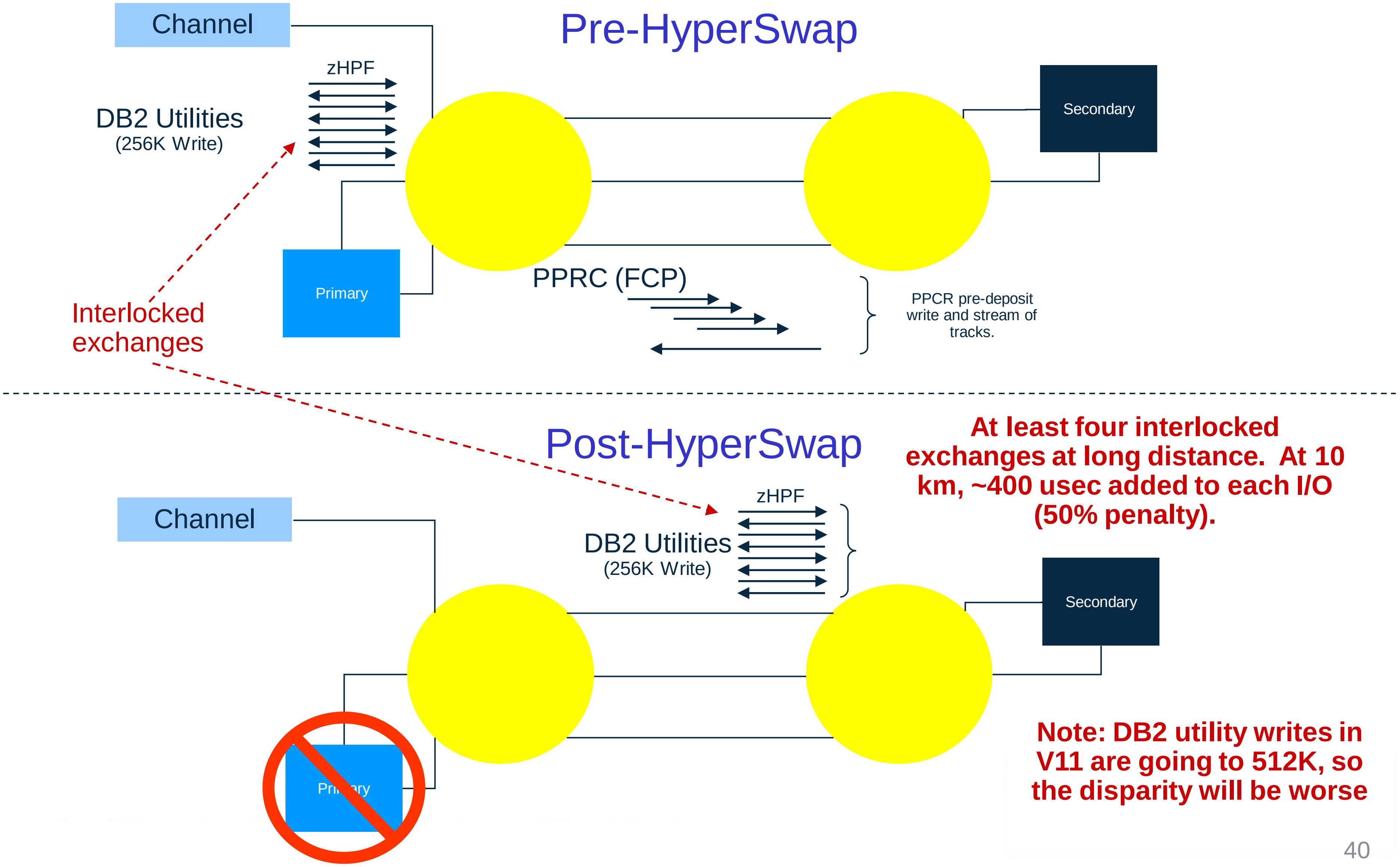
DS8000 echoes priority back on reads

Will be used to alleviate or prevent congestion caused by slow drain devices

z/OS calculates Sysplex wide priority range from all attached switches

CU also uses priority for writes that require replication (e.g., PPRC FCP based activity)

Pre-zHPF Extended Distance II



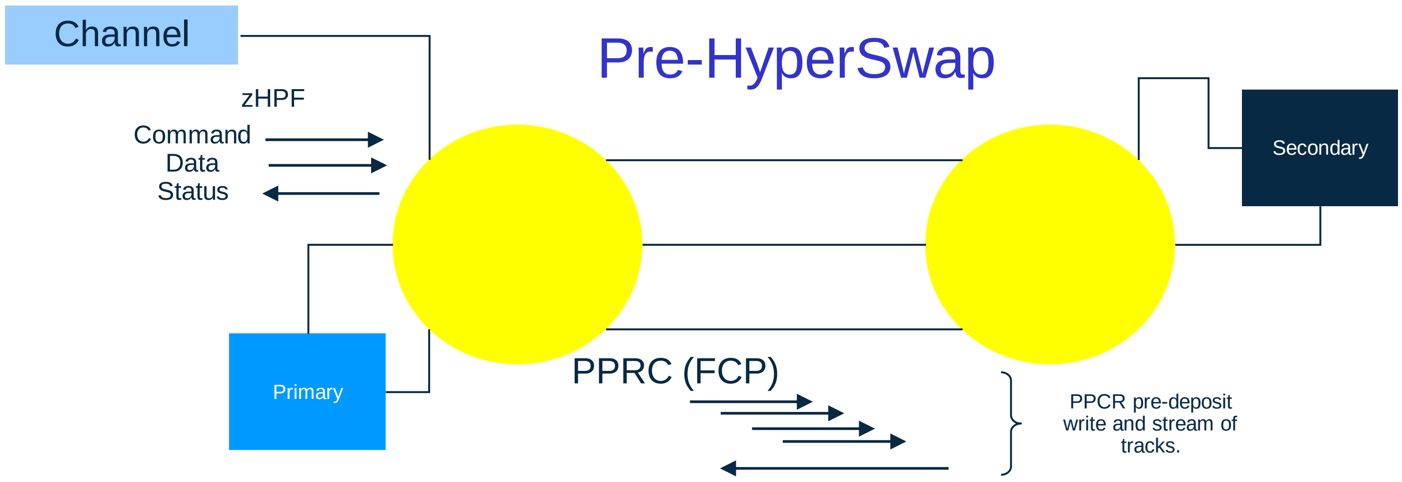
At least four interlocked exchanges at long distance. At 10 km, ~400 usec added to each I/O (50% penalty).

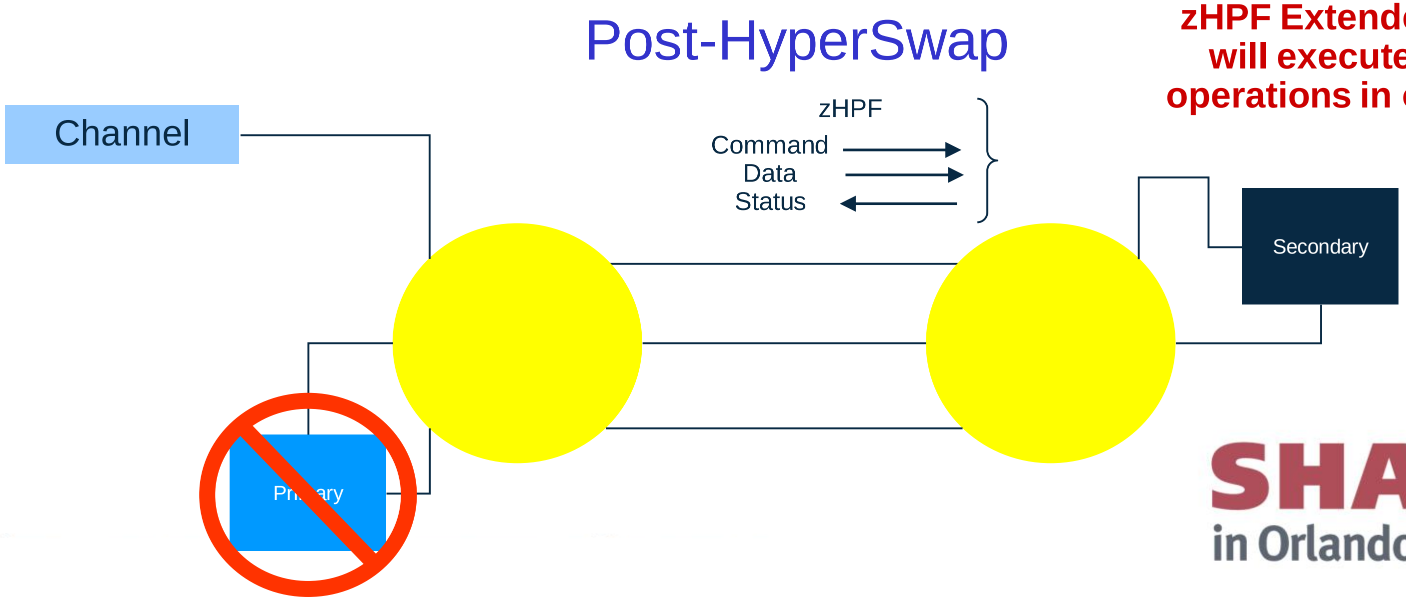
Note: DB2 utility writes in V11 are going to 512K, so the disparity will be worse

zHPF Extended Distance II

- Extended Distance I provided support for disabling first Transfer_Ready
  - Generally unused feature of FCP transport protocol
  - First Burst of data sent with command (Limited by first burst buffer size)
  - Up to 64KB for DS8000
  - Data Beyond 64KB requires one or more Transfer_Ready from control unit.
- Extended Distance II goes beyond the capabilities of FCP with dynamic use of first transfer ready and variable first burst size
  - Eliminates handshakes when control unit has the capacity
  - Significant latency reduction at distance AND without!
- Enabled through two new features in zHPF (Transport Mode) architecture – Becoming standard in FC-SB-6
  - First Transfer Buffer Credits (FTBC)
  - Transport Mode Command Retry

zHPF Extended Distance II

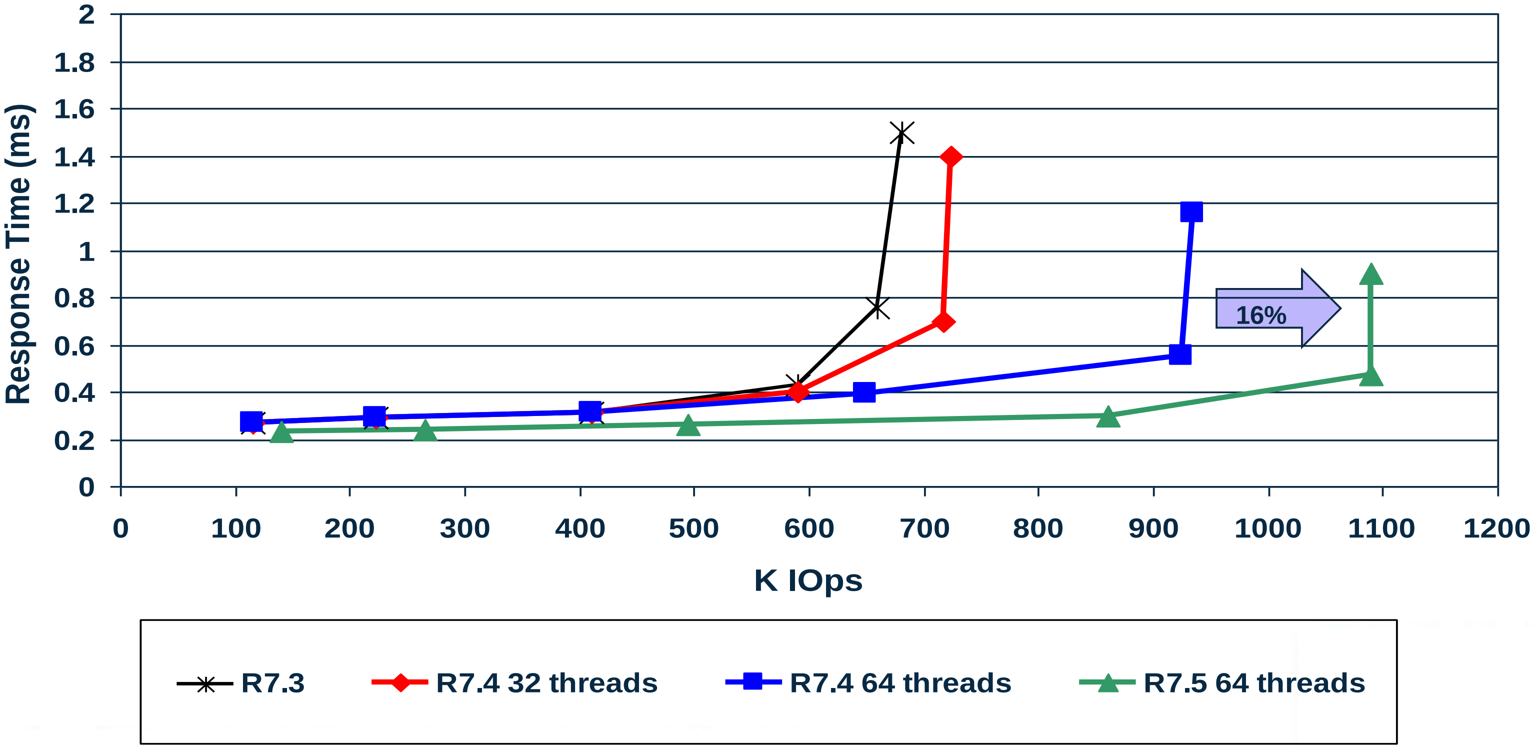




zHPF Extended Distance II will execute most write operations in one round trip

DS8000 Multi-thread Optimization

- The Multi-thread Performance Accelerator from release 7.4 enables additional Simultaneous Multithreading (SMT) exploitation to enable increased IOPS with the High Performance Flash Enclosure
- Improvements in R7.5 further increase the benefit of the additional threads and add support for Copy Services with the Performance Accelerator function enabled
- Reaching 1.1 Million IOPS!   (DBO(70/30/50) w/ 8 HPFE, 1TB Cache, 16 cores)

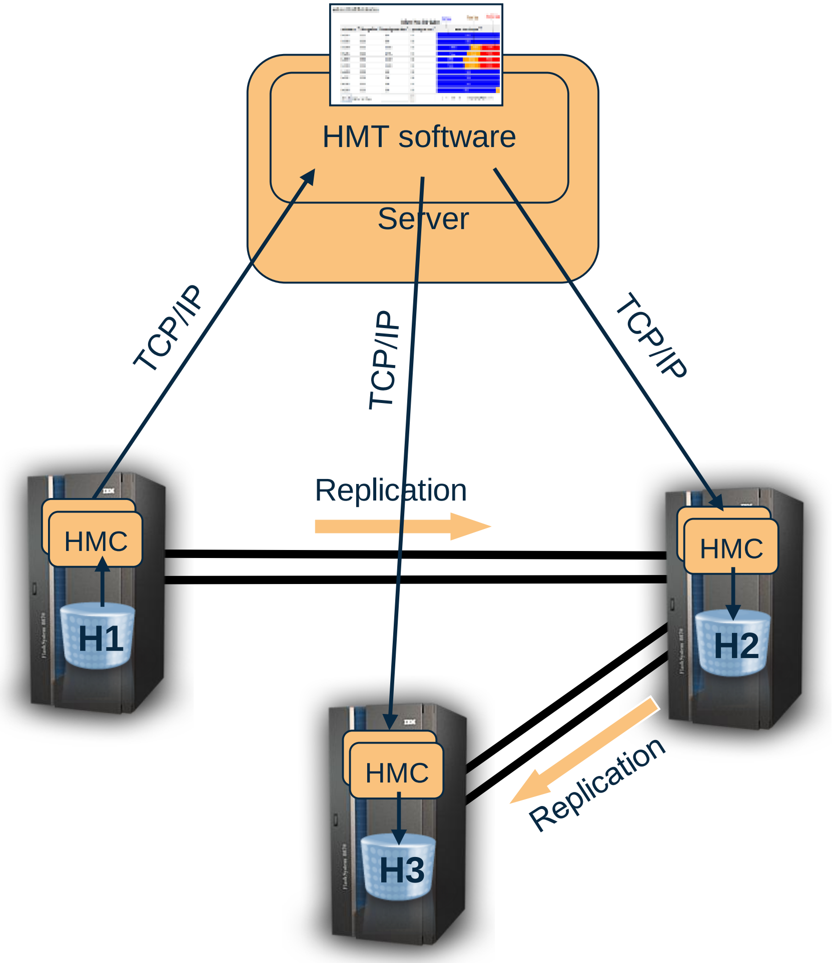


Easy Tier Heat Map Transfer Additional configurations

- Support for cascaded configurations with the Heat Map Transfer Utility (HMTU)

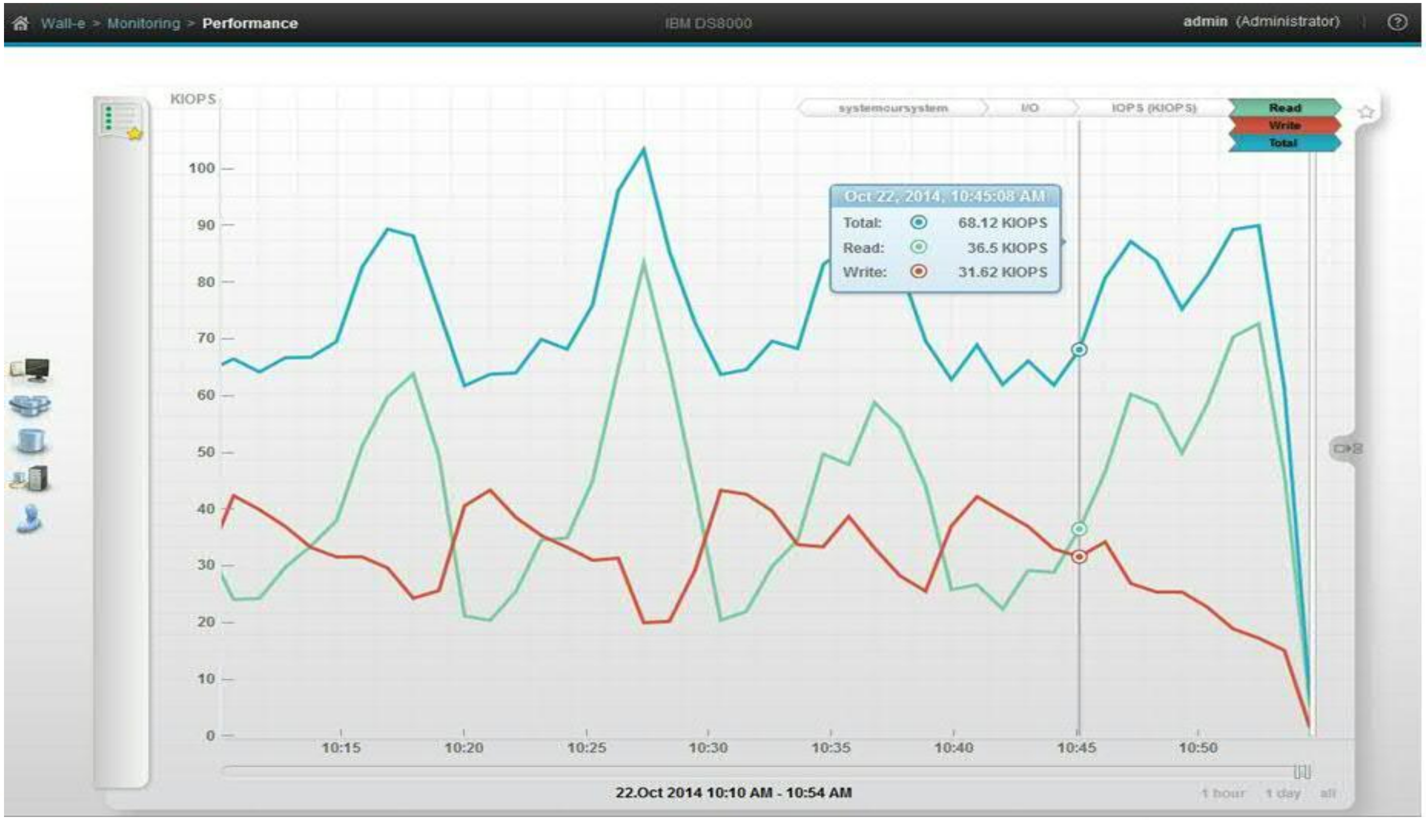
- GDPS support for additional configurations
  - GDPS/GM and GDPS/MGM 3/4-site support for FlashCopy devices

- This completes the GDPS Heat Map Transfer support for all GDPS configurations following the GDPS/XRC and GDPS/MzGM 3/4-site support in GDPS 3.12



DS8870 GUI Performance Reporting

- The DS8870 GUI will now include the ability to view performance graphs by subsystem
  - Reporting available on pools, ports and overall disk subsystem
  - Range of metrics with granularity down to 1 minute
- Also includes power, temperature and capacity reports

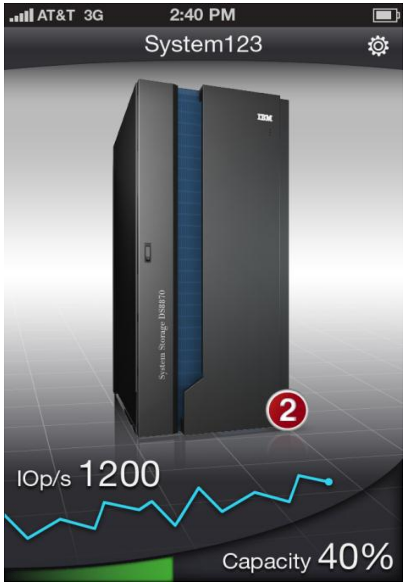


Oct 22, 2014, 10:45:08 AM
Total: 68.12 KIOPS
Read: 36.5 KIOPS
Write: 31.62 KIOPS

22.Oct 2014 10:10 AM - 10:54 AM

RESTful API

- Allows for web services to communicate, configure and control the DS8000 using industry standard HTTP.
  - Control plane applications, Mobile, custom software and more
  - Removes the need for loading DS8k specific ESSNI JAR files
- Secure Authentication: ESSNI Authentication w/ Tokens
- Ability to upgrade/extend REST API concurrently without DS8000 microcode firmware update
- Initial Release : Supported REST API constructs:
  - Logical Configuration – Query, Create, Delete, Modify
  - Event Notification
- Initial users of REST API in R7.5 timeframe
  - DS8870 Mobile Dashboard (iOS and Android support)
  - VMware VASA 2.0 support for DS8870
  - DS8870 Openstack Cinder driver – Kilo Release



DS8870 iOS Mobile App

- Allows for users to monitor multiple DS8870 R7.5 systems
  - No methods to change or modify configuration
  - Requires ESSNI Authentication / Uses Secure HTTPS
  - Uses R7.5 REST API support
- Publically available in the Apple App Store
  - Search for **"IBM Storage Mobile Dashboard"** (Free Download)
  - Shares same app as XIV, Storwize, others.

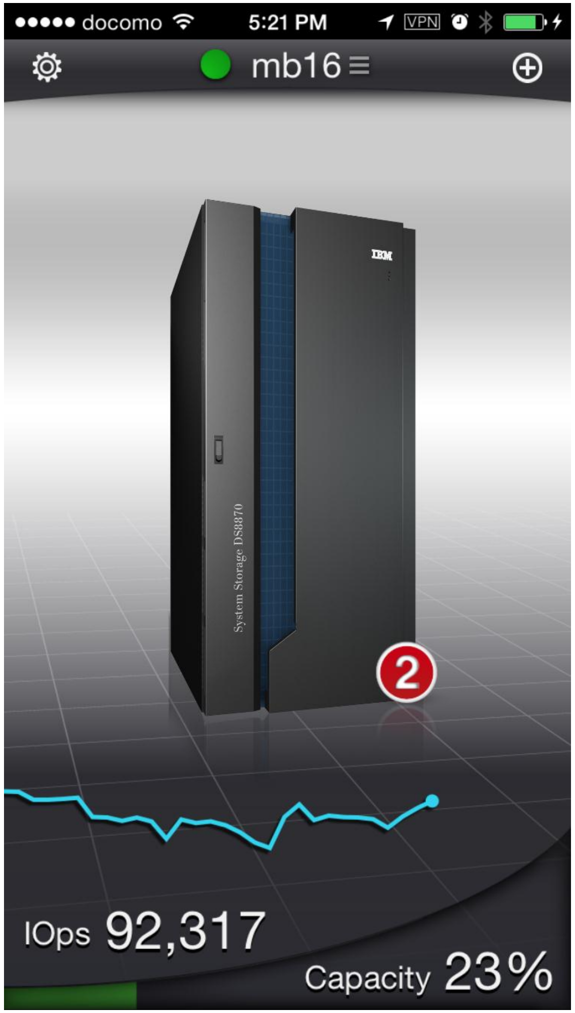
View Number of Events

Click on Icon to List Events

View full system Performance (IOPS and Latency)

Slide to switch between IOPS and Latency

Full System Capacity Percentage



Summary

- DS8870 review
- DS8870 Release 7.4 content
- DS8870 Release 7.5 content