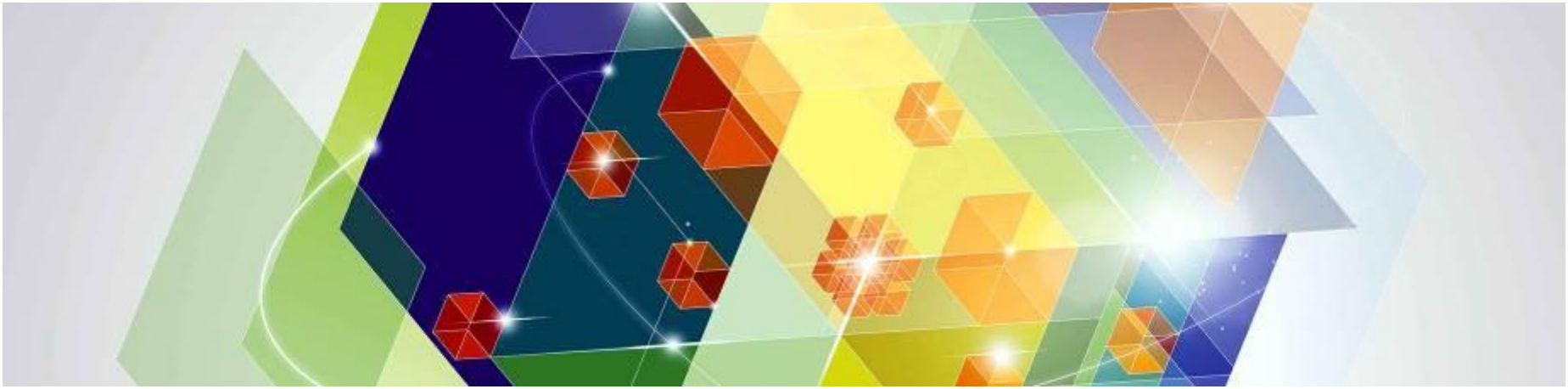
System z Flash Express

Introduction, Setup, Management, Configuration, Uses, and Benefits

SHARE in Anaheim
March 12th, 2014
Session 14726

Elpida Tzortzatos:
elpida@us.ibm.com

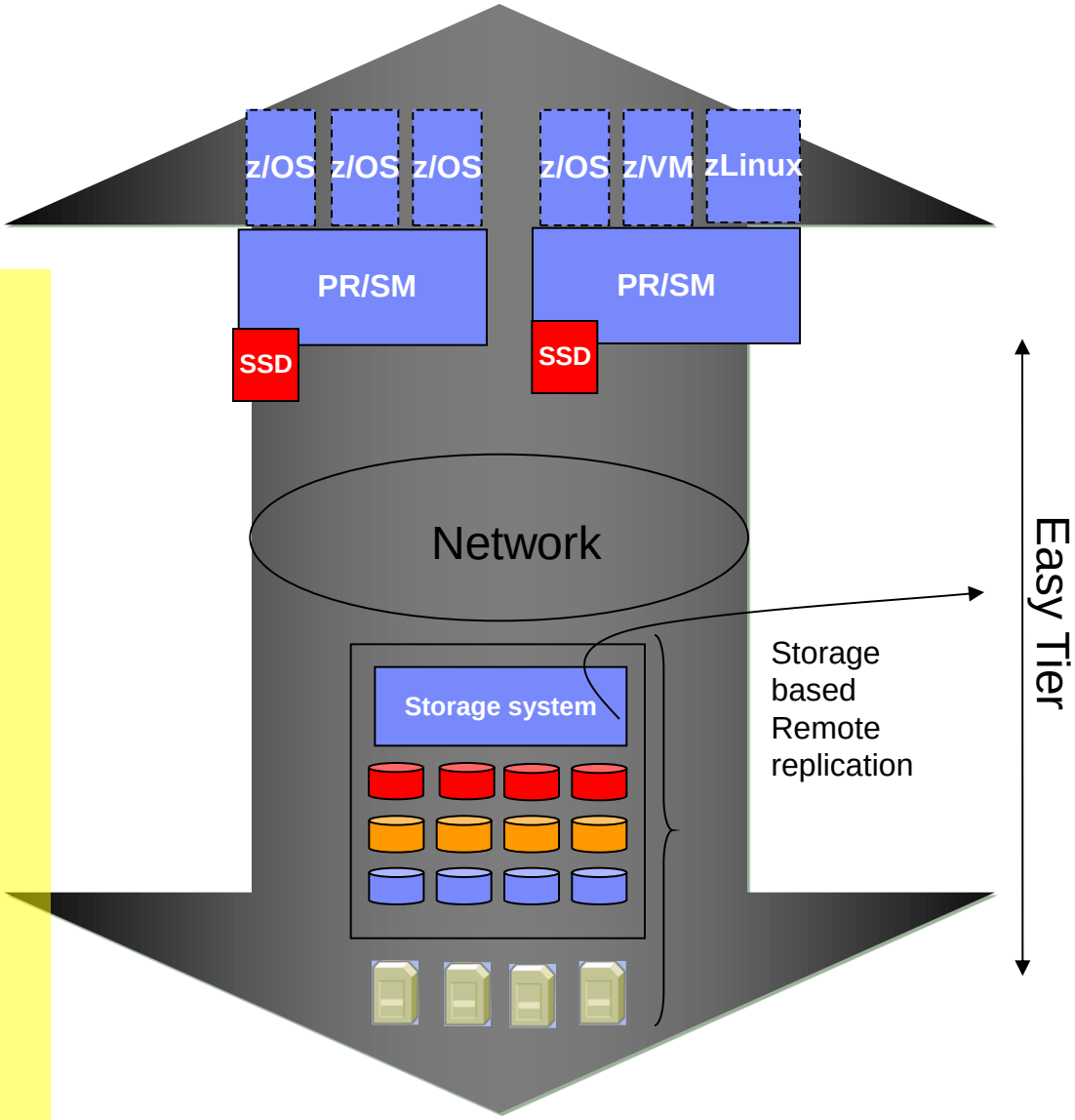


Agenda

- **z/OS Customer Value Proposition**
- **System z Flash Express and z/OS**
- **Flash Performance Results**
- **z/OS Flash Roadmap**
- **Under the Covers – Implementation Highlights**

System z Flash Express IO Adapter

- Flash Express is a PCIe IO adapter with NAND Flash SSDs (Solid State Drives)
- Flash Express is accessed using the Extended Asynchronous Data Mover
  - Optimized software path for Flash Access based on prior learning with z expanded store
- Flash Express provides continuous availability
  - RAID 10 to cover adapter failure
  - Concurrent Firmware update to cover service
- Flash Express is fully virtualized
  - A single adapter pair can provide Flash to 60 partitions on a CEC
  - Adapter RAS (call home, recovery, etc.) done at system level, not in OS.
  - Transparent migration to new adapter technology

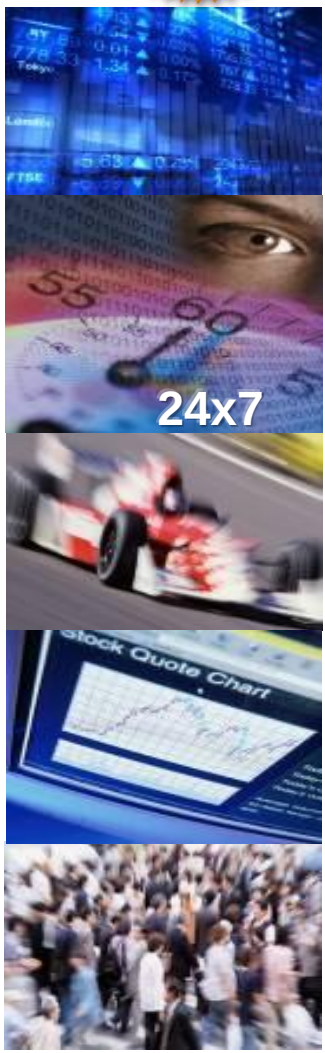


IBM Flash Express – Smarter Availability for Smarter Systems

NEW

- **Flash Express is an innovative solution designed to help you compete effectively in today's marketplace**
  - Automatically improve availability for key workloads at critical processing times
  - Drive availability and performance for workloads that cannot tolerate paging spikes or inconsistent performance
  - Slash latency for critical application processing such as diagnostics collection
- **Extends IBM's expertise in memory management introducing a new tier of memory using Flash Express**
- **Provides a secured, resilient and immediately usable solution**
- **Planned Flash Express and pageable large page exploiters:**

  - z/OS V1.13 ***Language Environment***
  -IMS 12 Common Queue Server
  -DB2 10 *
  - Java SDK601 SR4, and Java SDK7 SR3 and by extension: *- CICS Transaction Server 5.1*
  - WAS Liberty Profile v8.5**
  - IMS 12
  - DB2*

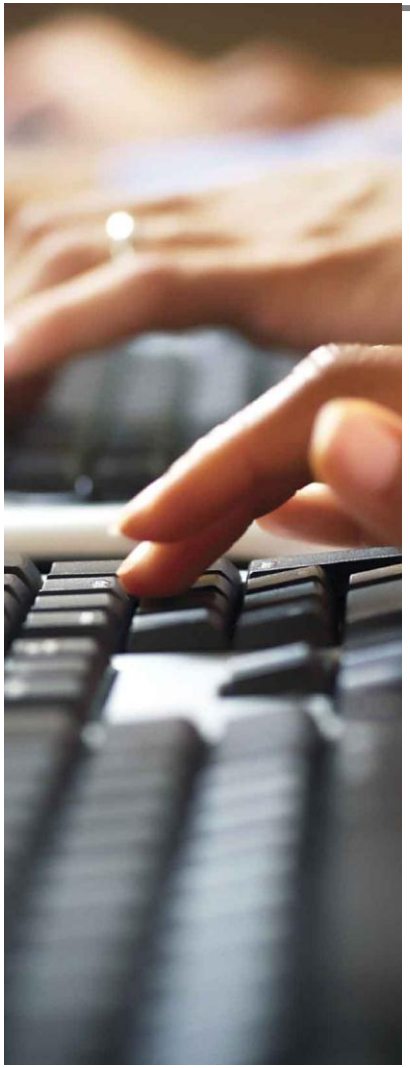


**Traditional WAS support is planned for a future date
*DB2 date to be determined. Support for V10 with APARs is planned

Flash Express Strengthens Availability

- Innovation to drive availability to exceptional levels
  - Extends IBM's expertise in memory management introducing a new tier of memory using Flash Express
  - Is an *industry unique* application of Flash to improve availability
  - Takes the next step in advanced memory management
- Flash Express can improve availability and reduce latency
  - Improves availability during transition periods and spikes
  - Helps accelerate start of day processing - batch to online
  - Enables faster snapshots of diagnostics (e.g. SVC dump, standalone dump)
  - With pageable large pages can improve performance of DB2 and Java
  - Ideal for applications with random read access and high read/write ratios
- Helps customers deliver vigorous service levels
  - Designed to help provide *exceptional* availability and fast response time
  - Delivered in tandem with pageable large pages for *superior* performance
- Minimal configuration- no special skills needed
  - Usable immediately; no special training required
  - Easy to set up and dynamically configurable

Representative Use Cases - Flash Express



Flash Express can reduce latency delays from paging to bring system availability to new heights and improve overall service levels

Application related errors will require collection of diagnostics. These diagnostics can be collected faster with Flash Express, reducing paging related delays that can impact your overall availability.

Having your working data resident in Flash can help accelerate start of day processing, and improve service for many industries at the busiest time of their work day- a time when they cannot afford disruptions.

DB2® and Java™ in memory buffer pools work to store and process application data. DB2 and Java can benefit from 1MB pageable large pages with Flash Express, improving overall performance.

Flash Express – What is it?

FLASH Express

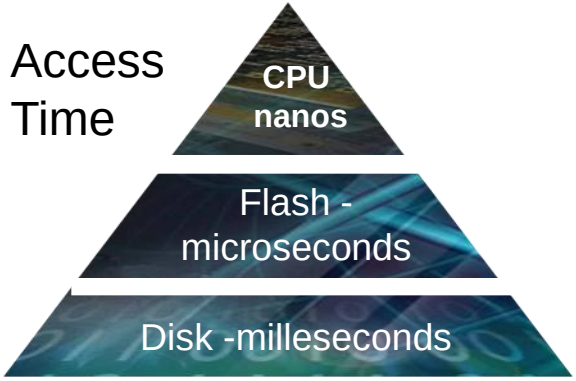
- Flash Express is a PCIe IO adapter with NAND Flash SSDs
- Physically comprised of internal storage on Flash SSDs
- Used to deliver a new tier of memory- storage class memory
- Uses **PCIe I/O drawer**

One Flash Express Card

- Sized to accommodate *all LPAR paging*
  - Each **card pair** provides **1.4 TB** usable storage (2.8 TB total)
  - Maximum 4 card pairs (4 X1.4=5.6 TB)

- Immediately usable
  - Simplifies capacity planning
  - No intelligent data placement needed
  - Full virtualization across partitions

- Robust design
  - Delivered as a **RAID10** mirrored pair
  - Designed for long life
  - Designed for concurrent firmware upgrade



Flash memory (SCM) blurs the distinction between memory and storage characteristics

- Secured
  - Flash Express adapter is protected with 128-bit AES encryption.
  - Key Management provided based on a Smart Card
  - Secure Cryptographic Erase meets audit requirements

IBM Flash Express – Smarter Availability for Smarter Systems

Outstanding Availability and Performance - Innovative Flash Express

- Companies competing for the highest quality of service in today's market must deliver **outstanding** availability and performance
- Changes in workload processing can impact service levels at critical processing times

- **Flash Express is an innovative solution designed to help you improve availability and performance to compete effectively in today's market**
  - Automatically improves availability for key workloads at critical processing times
  - Drives availability and performance for workloads that cannot tolerate paging spikes or inconsistent performance
  - Slashes latency for critical application processing such as start of day processing and also collection of diagnostics (SVC dumps, standalone dumps)
  - Delivered as a new adapter card in the PCIe I/O drawer

- **Benefits**
  - Improves availability and performance helping companies achieve highest service levels
  - Delivers a secured, resilient and immediately usable solution
  - Automatic, requires minimal setup, no special training needed

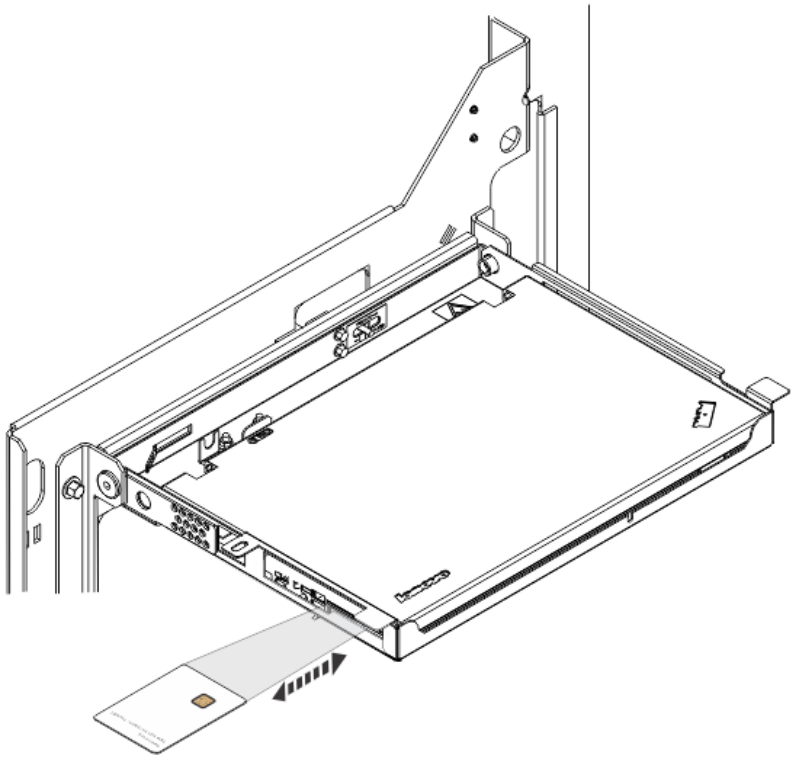
System z Flash and z/OS

Outline

- Introduction to zFlash
- Initial Setup
  - Customer Service Representative (CSR) Portion
    - Install Smart Cards in Support Elements (SE)
    - Install zFlash cards if necessary
    - Create Pair
  - Customer Portion
    - No IOCDS changes
    - Allocate zFlash memory to partition(s)
    - Configure z/OS to use zFlash
- Management
  - Management of zFlash Allocations
  - zFlash PCHID details
  - View Partition to PCHID map
  - View Flash Allocations for a specific Partition
  - View Flash (details)
  - System Activity Display (SAD) / Monitors Dashboard
  - Console Events
  - Security Logs
  - Status (Service Personnel Only)
  - Configure On/Off (Service Personnel Only)
  - Service On/Off (Service Personnel Only)
- Terminating Flash
  - Change all instances of z/OS to no longer use zFlash
  - Disband all zFlash pairs
  - Remove SE Smart Cards and destroy (optional)

Install Smart Cards in Support Elements (CSR responsibility)

- Cards will be installed by the CSR
  - During machine installation if zFlash shipped with the machine
  - Before installing the first zFlash adapters (if the machine was not shipped with zFlash)



More on Smart Cards

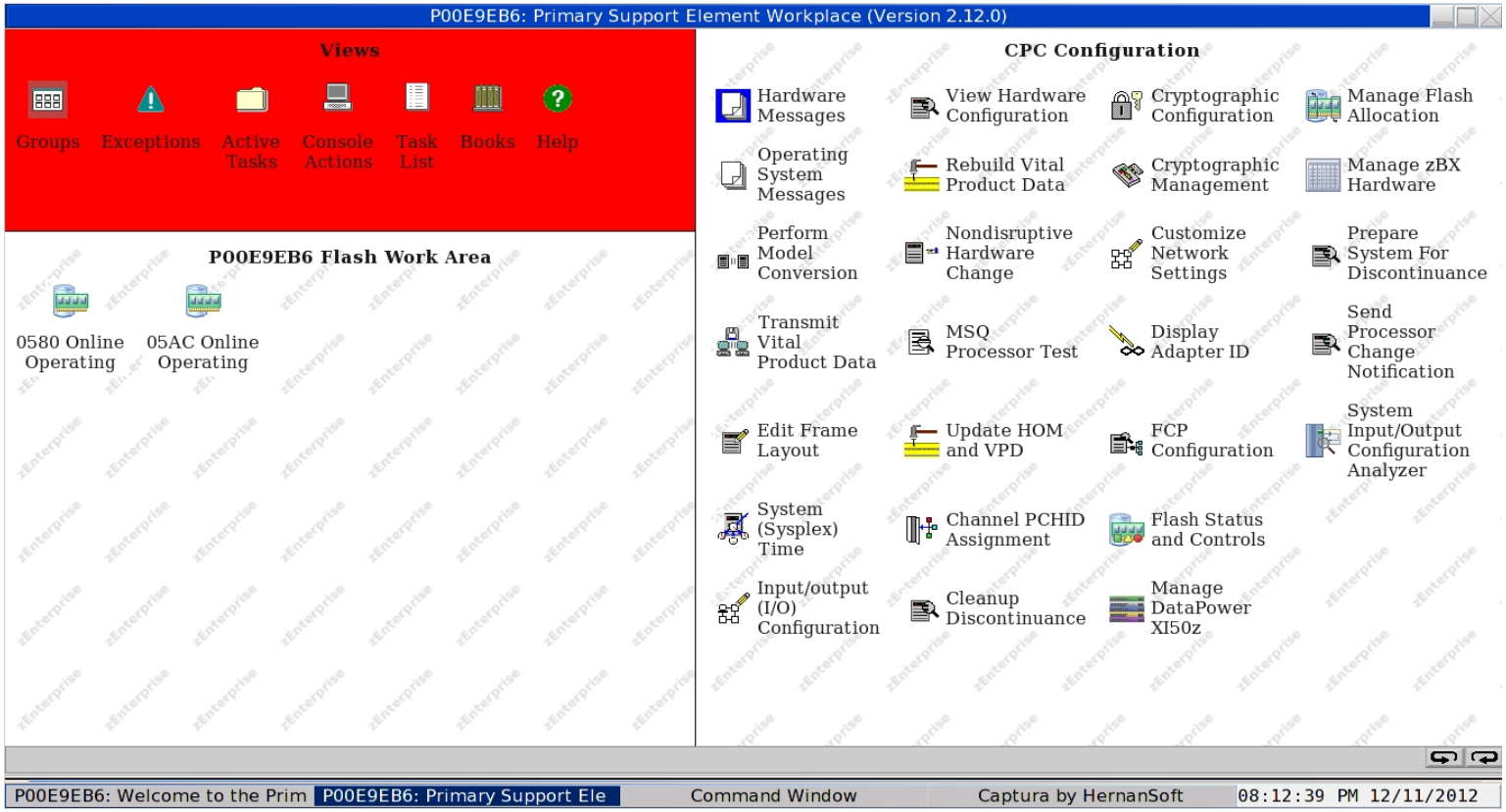
- The data on the zFlash cards is encrypted. This is done to prevent access to the data if a zFlash card is removed from the system, such as for a repair action or thru some malicious action (i.e. theft).
- The Smart Cards are an essential part of managing the encryption keys.
- The blank Smart Card is the same one used by the TKE device.
- The Primary SE will create an authentication key using the smart card and store it on the SE. The Alternate SE will uses the smart card in it to store the key sent from the primary.
- The smart card, the SE hardware, the CEC, and the generated Key are tightly coupled in order to prevent access to the data on the zFlash card in any place other than the CEC it was formatted for.
- If for some reason the smart card fails on the primary an automatic switch to the alternate will happen and a service call will occur to have the smart card or the SE serviced. There are procedures to ensure the repaired SE or Smart Card is properly updated with the encryption keys.
- The keys will not be preserved during migrations/upgrades. So, persistence of data on the zFlash adapters is thus not guaranteed. The zFlash adapters are therefore good for things like paging storage but should not be used for situations where persistence is required.
- Bottom line: The Smart Cards must be installed so that the SE is prepared to store and handle the encryption keys used to protect the data on the zFlash adapters.

Install zFlash cards if necessary (CSR responsibility)

- Installed in pairs in Seneca cages, one per I/O domain
- Pairs are cabled together with 2 SAS cables

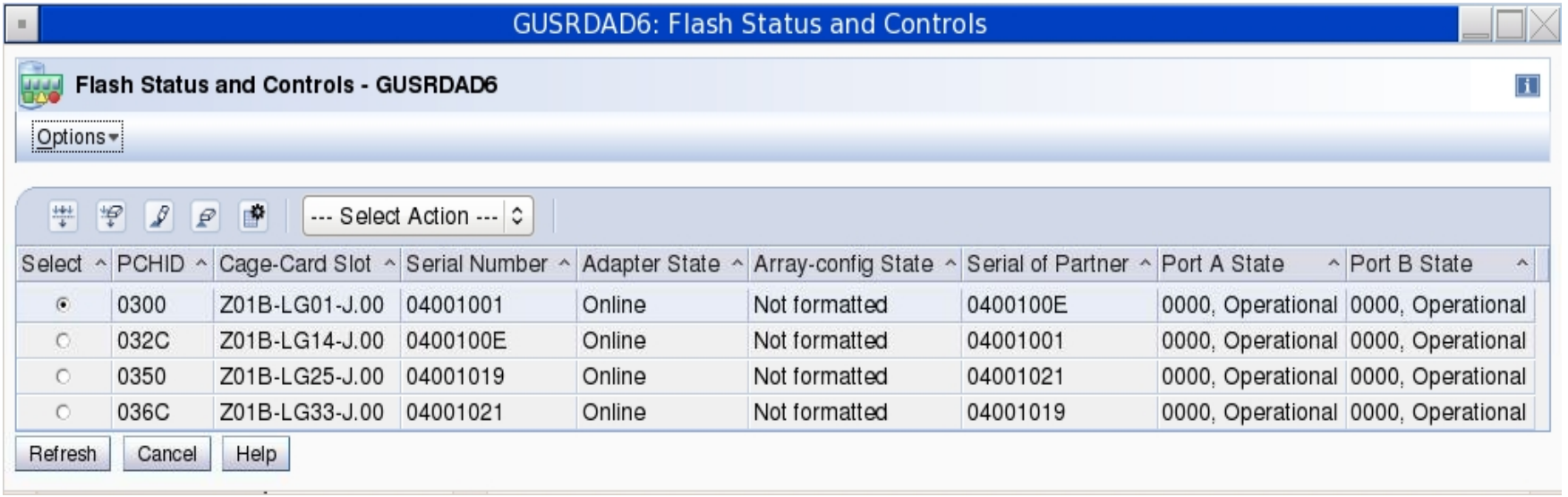
Install zFlash cards if necessary (CSR responsibility)

- Once installed, the cards are visible on the Support Element's User Interface as a PCHID



Create pair(s) of zFlash adapters (CSR Responsibility)

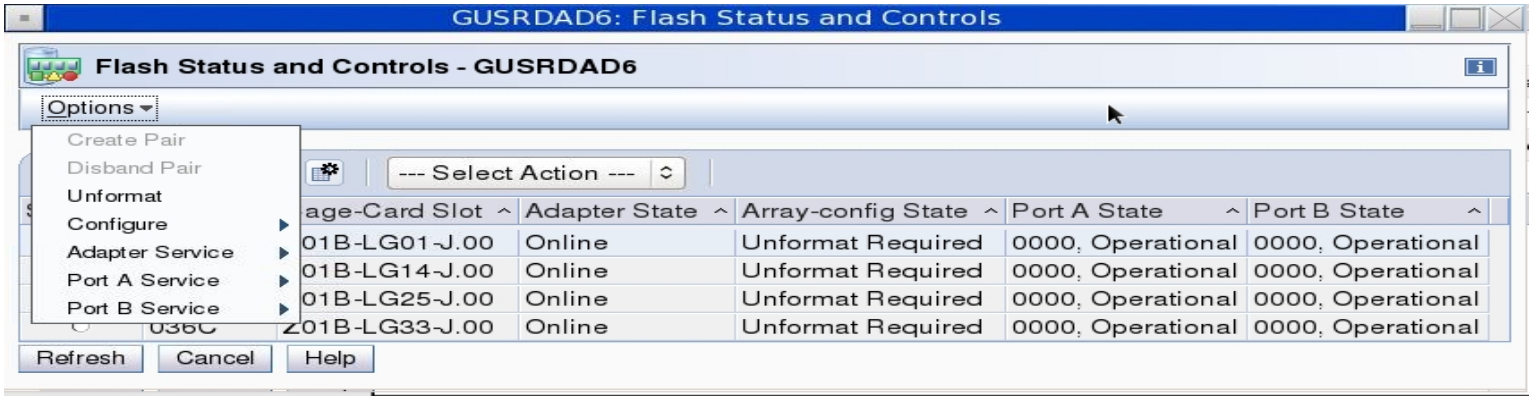
- A “create pair” operation must be performed that allows the paired adapters to initialize themselves into a pair and format the storage.
- Done via a new SE task, **Flash Status and Controls**
- Service Personnel only

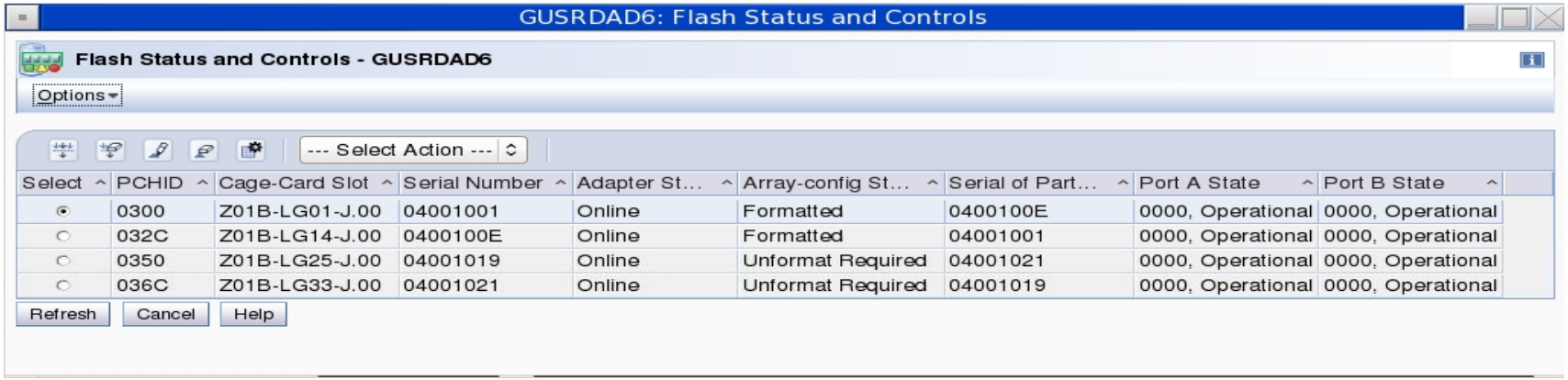


Select	PCHID	Cage-Card Slot	Serial Number	Adapter State	Array-config State	Serial of Partner	Port A State	Port B State
◉	0300	Z01B-LG01-J.00	04001001	Online	Not formatted	0400100E	0000, Operational	0000, Operational
○	032C	Z01B-LG14-J.00	0400100E	Online	Not formatted	04001001	0000, Operational	0000, Operational
○	0350	Z01B-LG25-J.00	04001019	Online	Not formatted	04001021	0000, Operational	0000, Operational
○	036C	Z01B-LG33-J.00	04001021	Online	Not formatted	04001019	0000, Operational	0000, Operational

Create Pair

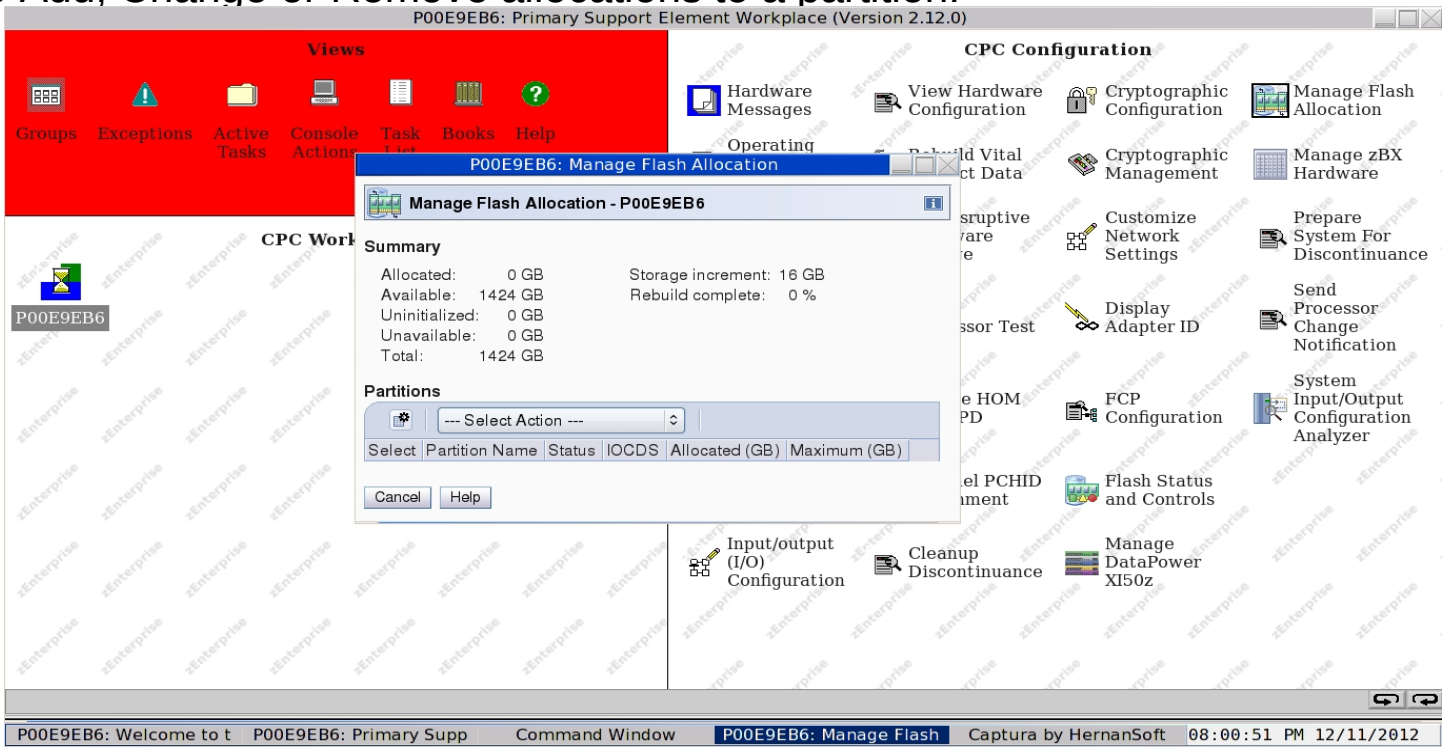
- Select **Create Pair** to create and format a pair
- Use the Refresh button to monitor the progress of the formatting.
- It takes a while (15 to 20 minutes) to complete the pairing/formatting operation.



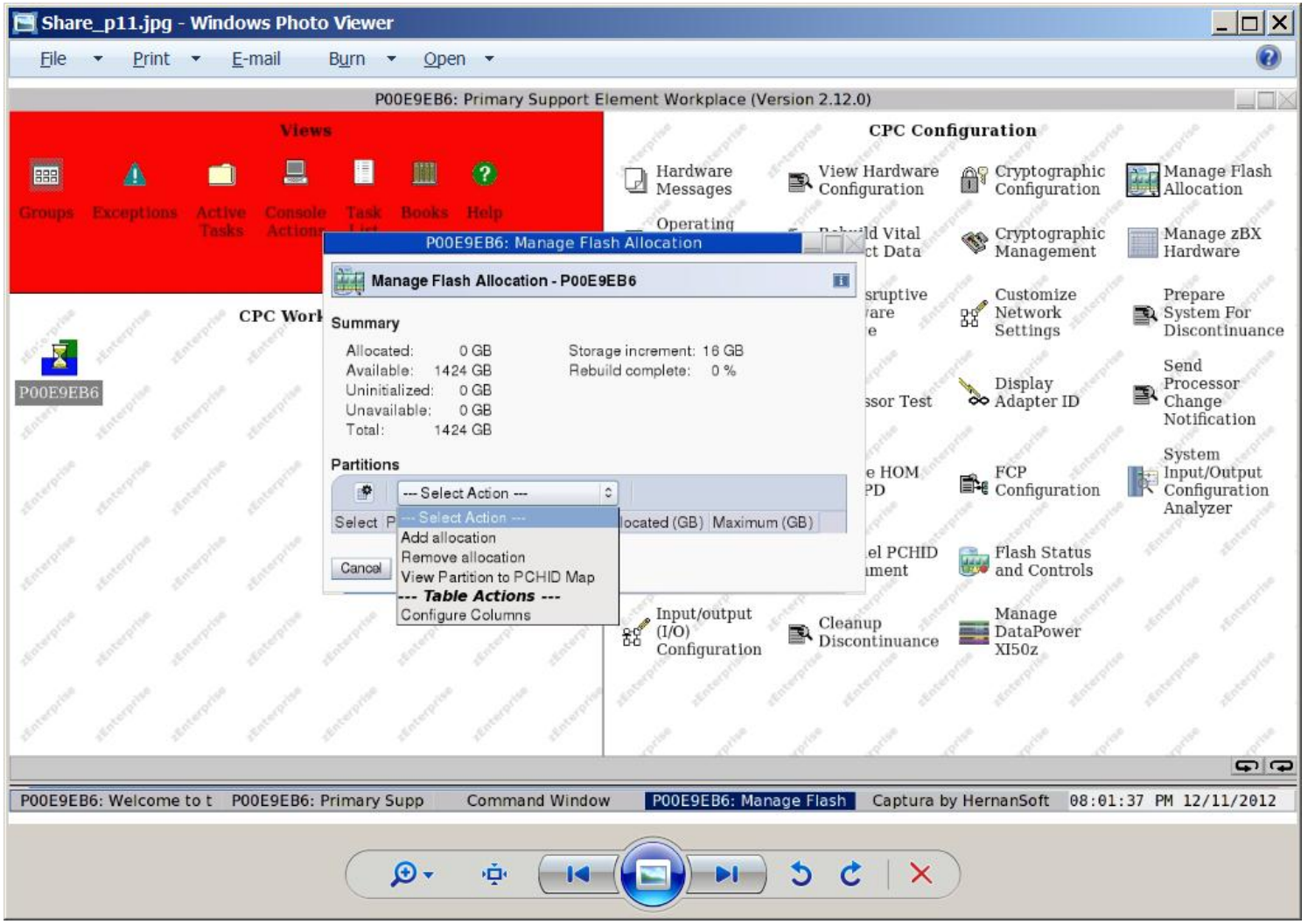


Allocate zFlash Memory to partition(s)
Manage Flash Allocation SE and HMC task

- Available on both the HMC and SE.
- Displays current summary Flash information for the system.
- Displays current Flash information by partition.
- Use to Add, Change or Remove allocations to a partition.

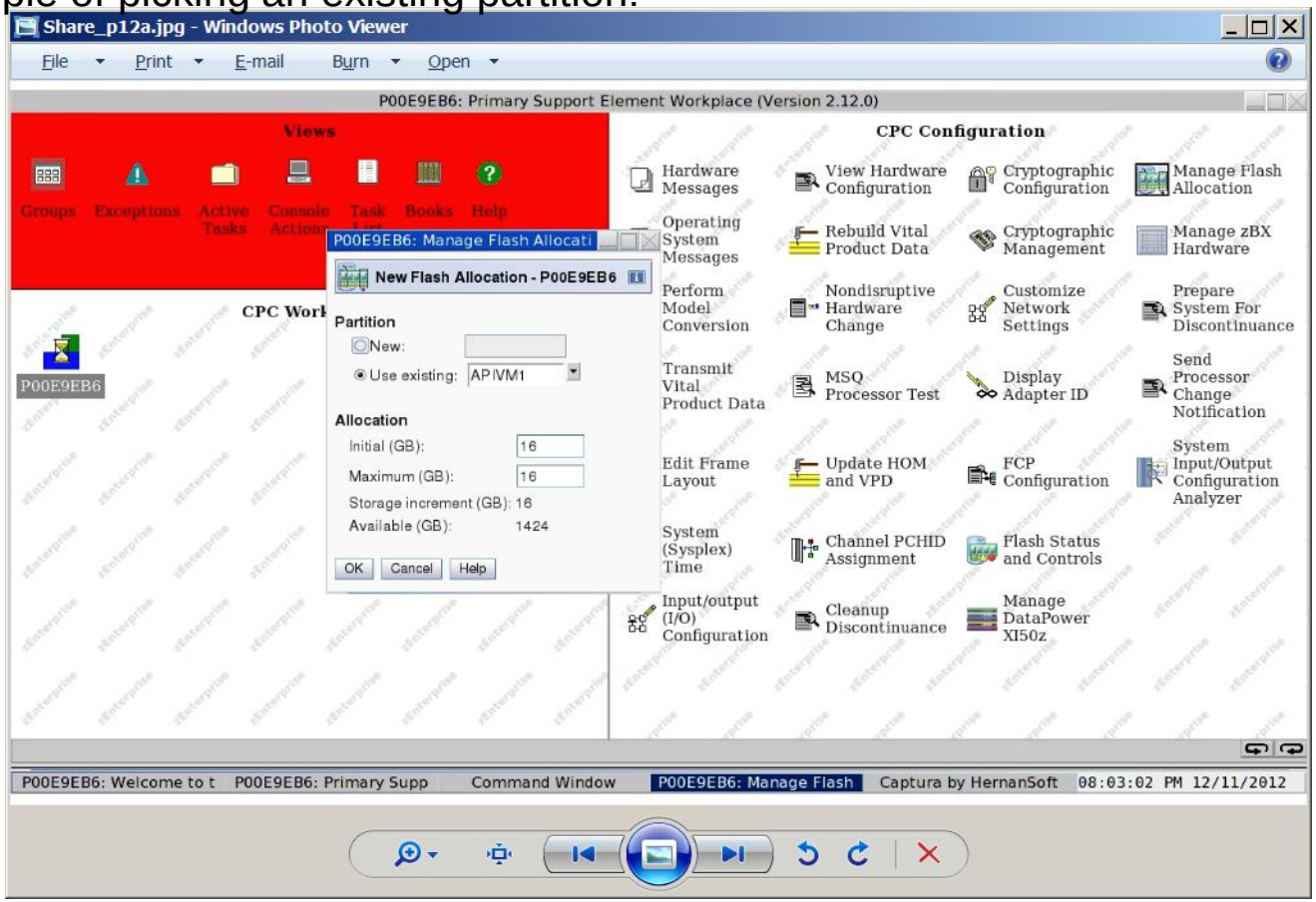


Manage Flash Allocation Task's Actions



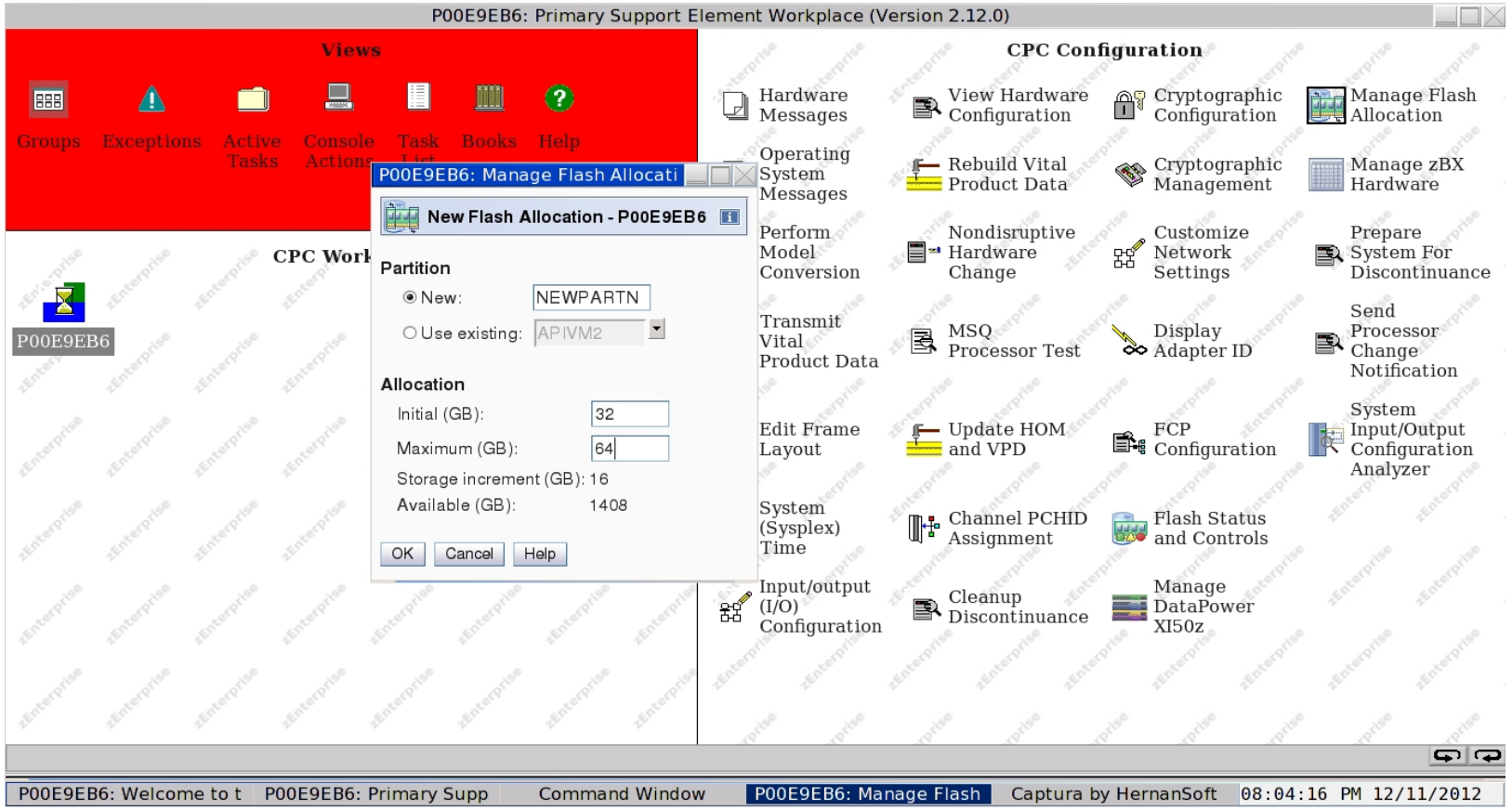
Manage Flash Allocation - Add zFlash Allocation

- Allocation can be done for a partition defined in any IOCDS or a partition not currently defined.
- May be done after initial zFlash setup as necessary (such as when a new partition is defined).
- Example of picking an existing partition:



Manage Flash Allocation - Add zFlash Allocation

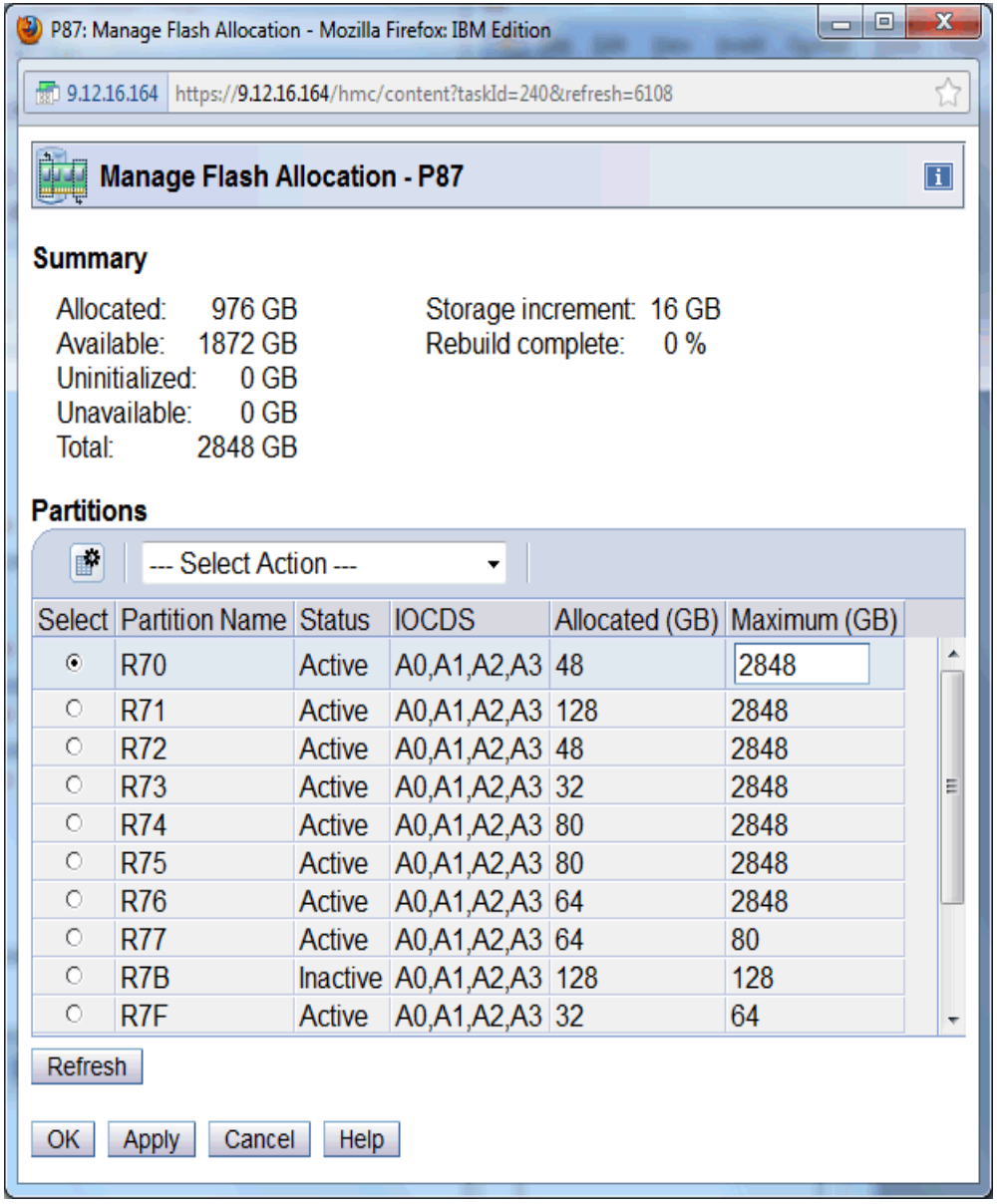
- Example of typing in a new partition name:



Allocating z FLASH

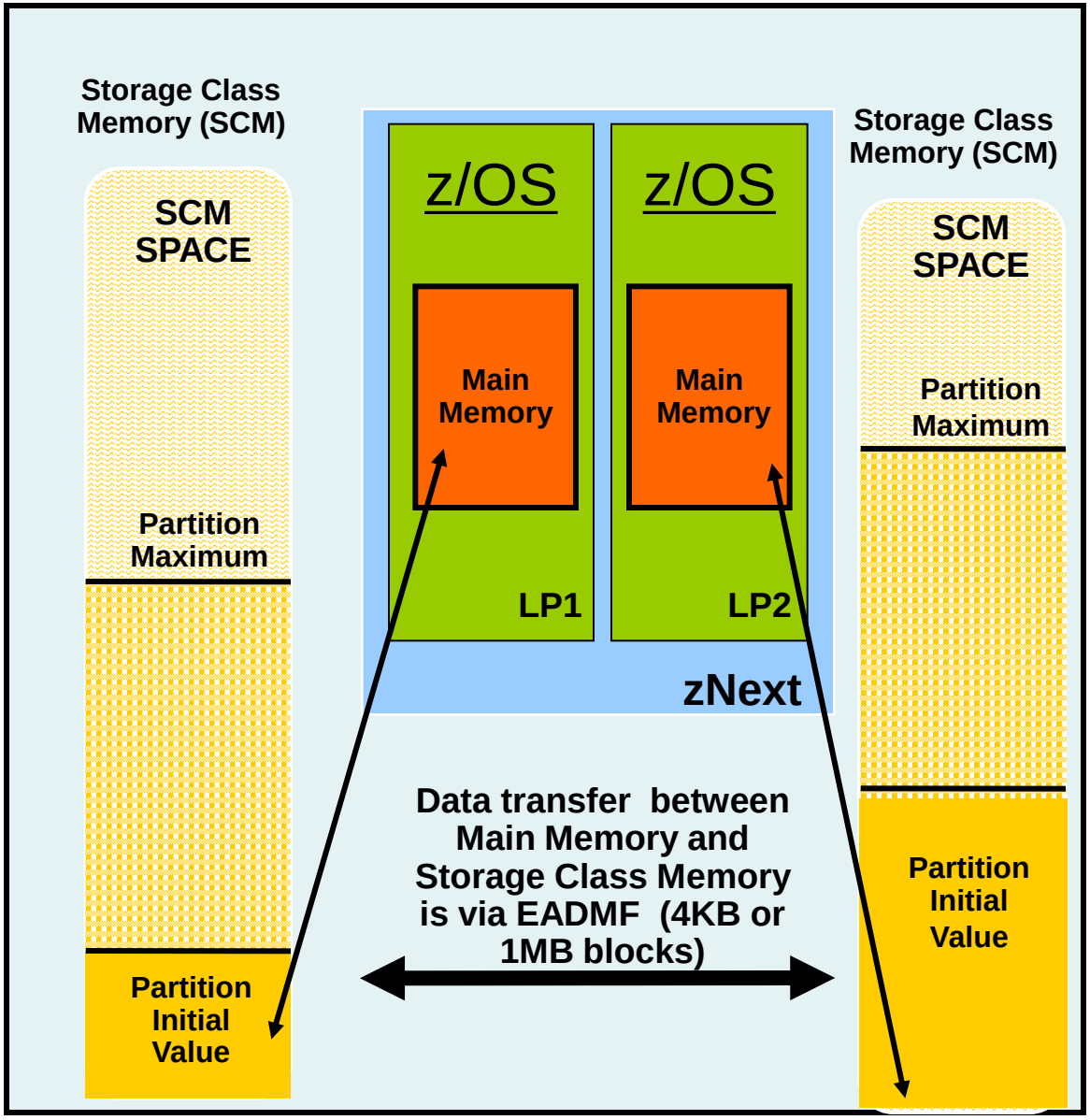
Allocating Flash to a partition

- The initial and maximum amount of Flash Memory available to a particular logical partition is specified at the SE or HMC via a new Flash Memory Allocation panel
- Can dynamically change maximum amount of Flash Memory available to a logical partition
- Additional Flash Memory (up to the maximum allowed) can be configured online to a logical partition dynamically at the SE or HMC
  - For z/OS this can also be done via an operator command
- Can dynamically configure Flash Memory offline to a logical partition at the SE or HMC
  - For z/OS this can also be done via an operator command
- Predefined subchannels, no IOCDS



z FLASH Virtualization

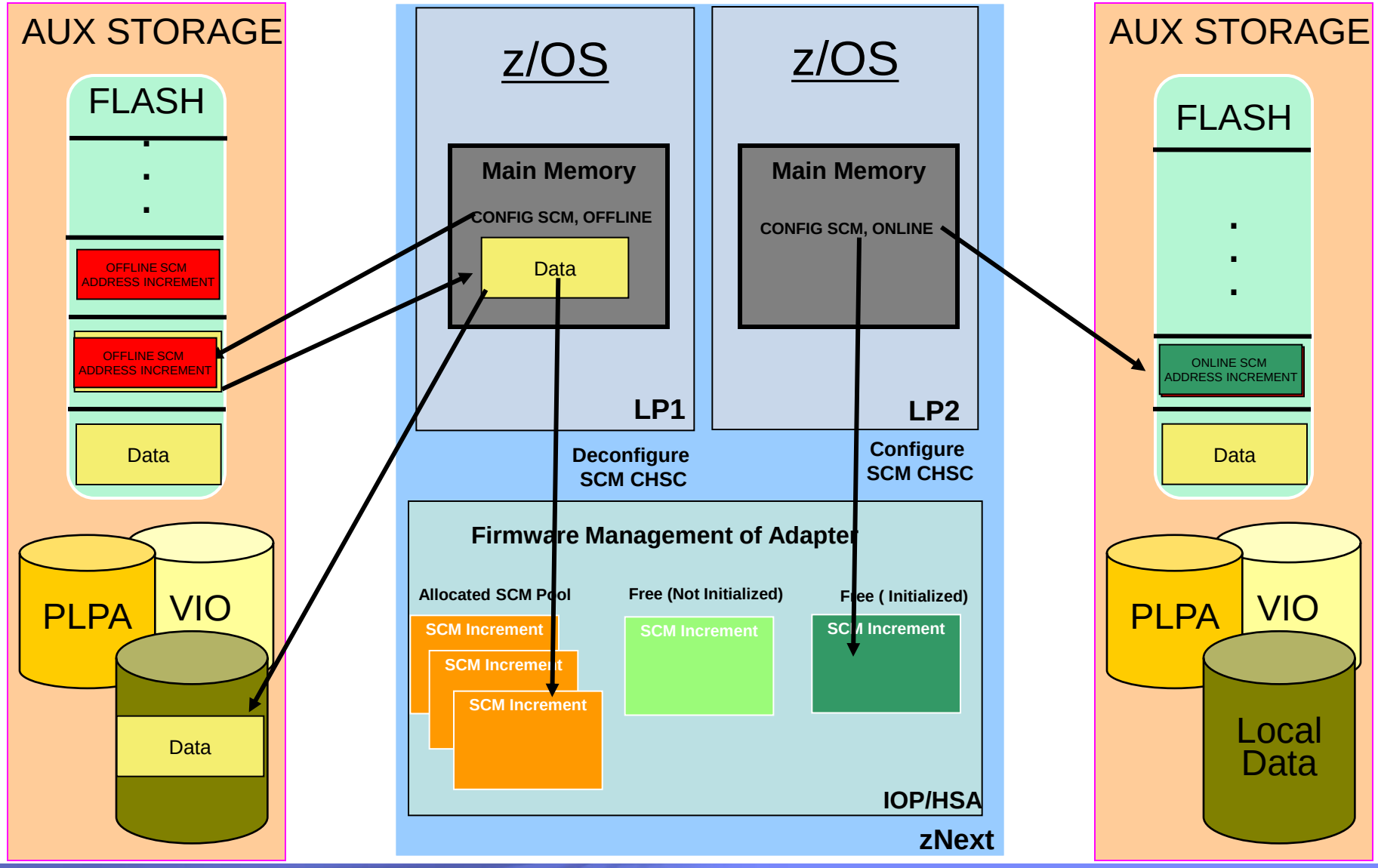
- **Full virtualization of physical Flash PCIe cards across partitions, software sees an Abstracted Flash Storage Space…**
  - Allows each logical partition to be configured with its own SCM address space
  - Allocate Flash to partitions by amount, not card size
  - Ability to change underlying technology while preserving API
- ► **No Hardware Specifics in Software.**
  - Error Isolation, Transparent mirroring, Centralized diagnostics, etc.
  - Hardware Logging, FRU Call, Recovery: Independent of software



FLASH for z/OS Paging Value

- **Flash Memory is a faster paging device as compared to HDD**
  - The value is NOT in replacing memory with Flash but replacing disk with Flash
  - Flash is suitable for workloads that can tolerate paging and will not benefit workloads that cannot afford to page
  - The z/OS design for Flash Memory does not completely remove the virtual storage constraints created by a paging spike in the system. (Some scalability relief is expected due to faster paging I/O with Flash Memory.)

A z/OS Flash Configuration



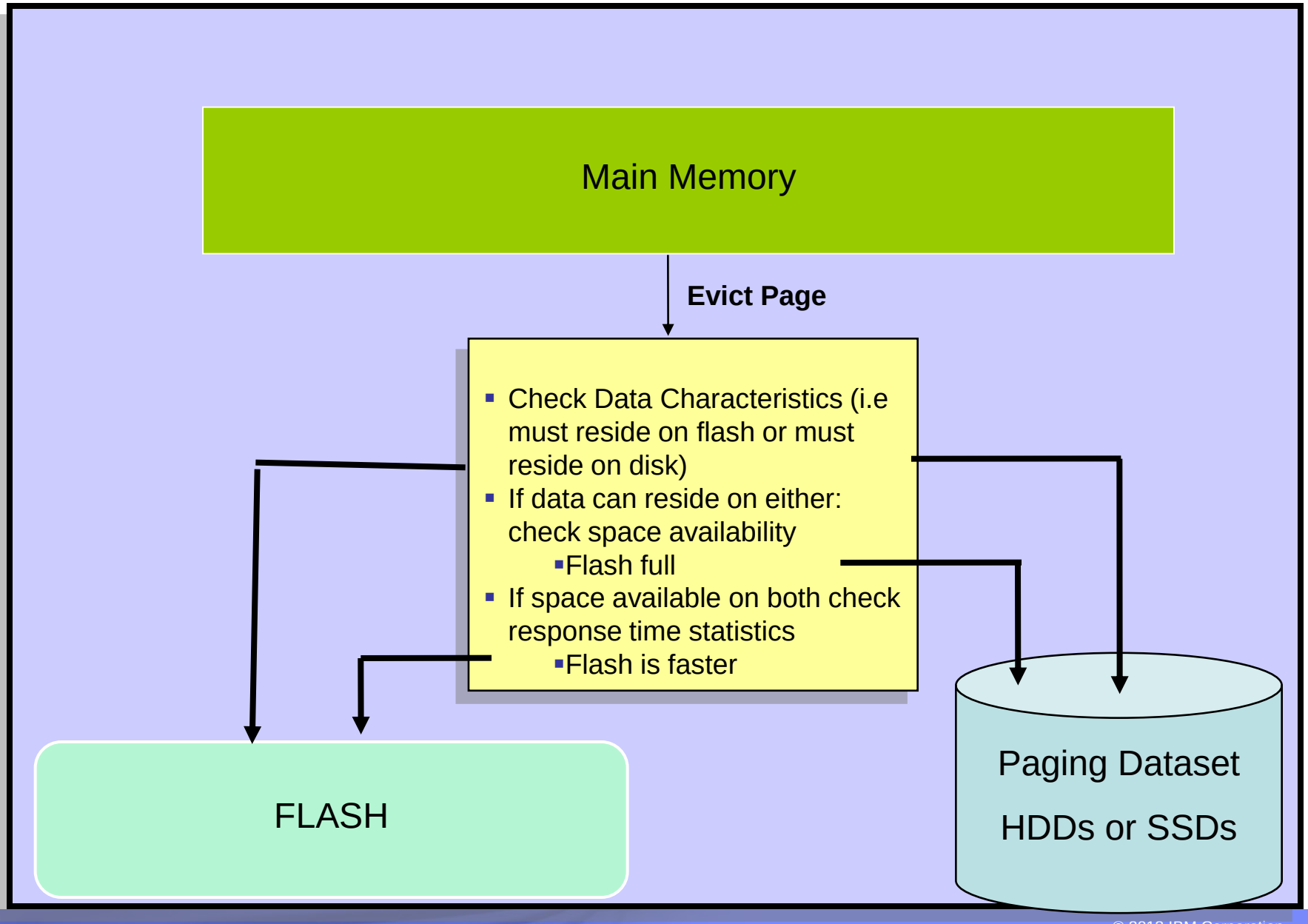
Typical Customer Configurations for FLASH

- **Flash card pair memory size is 1.4TB**
  - Min: 1 Card Pair
  - Max: 4 Card Pairs

- **Typical customer configuration is 6 to 8 LPARs per CEC and 40GB - 80GB for paging configuration dataset size**

- **Even with 10 LPARs per CEC, each LPAR has 140 GB of Flash Memory available for its paging datasets, more than double the current typical customer configuration.**
  - All paging data can easily reside on Flash
  - Data will preferably go to Flash and only go to disk (if any) when Flash is full
  - No intelligent placement of data on internal Flash needed

Flash vs Disk Placement Criteria



Flash vs Disk Placement Criteria

Data Type	Data Page Placement
PLPA	At IPL/NIP time PLPA pages will be placed both on Flash and disk.
VIO	VIO data will always be placed on disk (First to VIO accepting datasets with any spillover flowing to non-vio datasets)
Pageable Large Pages	If contiguous Flash space is available, pageable large page will be written to Flash. If Flash is not available in the system configuration pageable large pages will be backed with 4k page frames.
All other data	If available space exists on both Flash and disk then make a selection based on response time.

z/OS FLASH Use Cases

- Paging
  - **z/OS paging subsystem will work with mix of internal Flash and External Disk**
    - –Self Tuning based on measured performance
    - –Improved Paging Performance, Simplified Configuration
  - **Begin Paging 1 MB Large Pages only on Flash**
    - –Exploit Flash's random IO read rate to get CPU performance by enabling additional use of Large Pages. Currently large pages are not pagable.
  - **Begin Speculative Page-In of 4K Pages, 1MB Pages only on Flash**
    - –Exploit Flash's random IO read rate to get Improved Resilience over Disruptions.
    - –Market Open, Workload Failover

Flash Memory Usage and Invocation

- **New PAGESCM= keyword in IEASYSxx defines the amount of flash to be reserved for paging**
  - **Value may be specified in units of M, G, or T**
  - **NONE indicates do not use flash for paging**
  - **ALL (default) indicates all flash defined to the partition is available for paging**

Flash Memory Usage and Invocation (cont)…

- **New messages issued during IPL indicate the status of SCM**
  - `IAR031I USE OF STORAGE-CLASS MEMORY FOR PAGING IS ENABLED - PAGESCM=ALL, ONLINE=00065536M`

  OR

  - `IAR032I USE OF STORAGE-CLASS MEMORY FOR PAGING IS NOT ENABLED - PAGESCM=NONE`

Flash Memory Usage and Invocation (cont)…

- **The D ASM and D M commands are enhanced to display flash-related information/status**
  - **D ASM lists SCM status along with paging data set status**
  - **D ASM,SCM displays summary of SCM usage**
  - **D M=SCM display SCM online/offline and increment information**
  - **D M=SCM(DETAIL) displays detailed increment-level information**

Display ASM Command

d asm

```
IEE200I 17.17.46 DISPLAY ASM 944
TYPE      FULL STAT    DEV  DATASET NAME
PLPA      100% FULL   02E6  SYS1.PLPA.PAGCOM
COMMON     61%   OK   02E6  SYS1.COMMON.PAGCOM
LOCAL       0%   OK   098E  SYS1.LOCAL.PAGEP2
LOCAL       0%   OK   0987  SYS1.LOCAL.PAGEP3
LOCAL       0%   OK   098F  SYS1.LOCAL.PAGEP4
SCM        11%   OK    N/A   N/A
```

d asm,scm

```
IEE207I 17.35.02 DISPLAY ASM 947
STATUS       FULL      SIZE              USED          IN-ERROR
IN-USE        11%    16,777,216        2,096,144           0
```

Flash Related Commands

D M=SCM

```
IEE174I 17.57.26 DISPLAY M 230
STORAGE-CLASS MEMORY STATUS
80G DEFINED
ONLINE
 0G-64G
16G OFFLINE-AVAILABLE
14% IN USE
SCM INCREMENT SIZE IS 16G
```

D M=SCM(DETAIL)

```
IEE174I 17.57.30 DISPLAY M 232
STORAGE-CLASS MEMORY STATUS - INCREMENT DETAIL
80G DEFINED
ADDRESS  IN USE  STATUS
    0G     55%  ONLINE
   16G      0%  ONLINE
   32G      0%  ONLINE
   48G      0%  ONLINE
ONLINE: 64G  OFFLINE-AVAILABLE: 16G  PENDING OFFLINE: 0G
14% IN USE
SCM INCREMENT SIZE IS 16G
```

CF SCM(16G),ONLINE

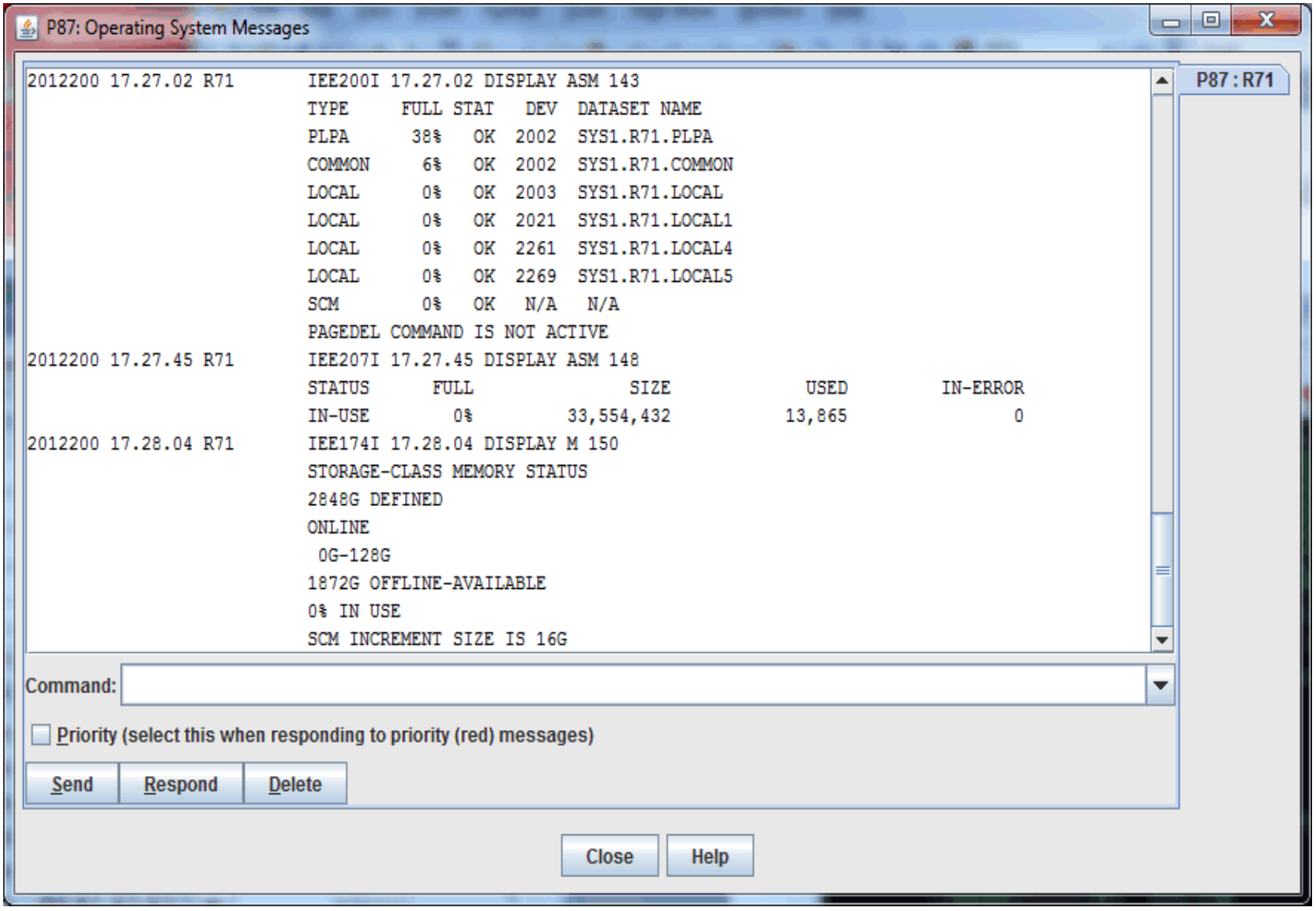
```
IEE195I SCM LOCATIONS 64G TO 80G ONLINE
IEE712I CONFIG   PROCESSING COMPLETE
```

Flash Memory Usage and Invocation (cont)…

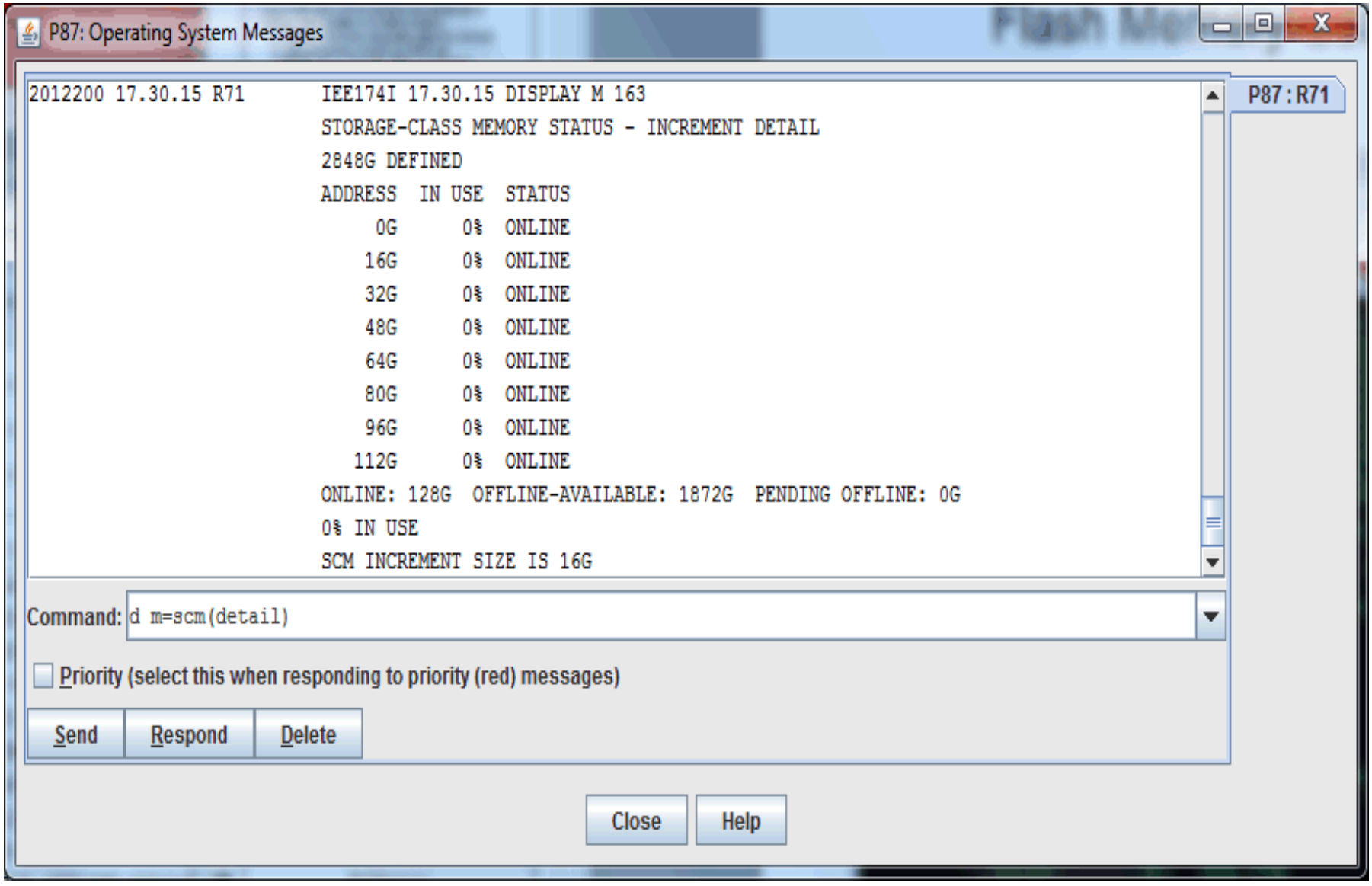
- **The CONFIG ONLINE command is enhanced to allow bringing additional SCM online**
  - **CF SCM(*amount*),ONLINE**
    - **CF SCM(16G),online**
      - **IEE195I SCM LOCATIONS 64G TO 80G ONLINE**
      - **IEE712I CONFIG PROCESSING COMPLETE**

- **The CONFIG OFFLINE command is enhanced to allow...**
  - **CF SCM(*amount*),OFFLINE**
  - **CF SCM(*start_range-end_range*),OFFLINE**
  - **Requires APAR OA40968**

Flash Memory Usage and Invocation (cont)…



Flash Memory Usage and Invocation (cont)…



RMF Updates

- **RMF Monitor II Page Data Set Activity Report includes SCM activity (RMF II → Resource → PGSP):**

```
                    RMF - PGSP Page Data Set Activity              Line 1 of 4

                       CPU=  1     UIC= 65K PR=   0         System= 4381 Tota

S VOLUME DEV  DEV      %SLOTS   PAGE   I/O REQ AVG PAGES       10:21:00
T SERIAL NUM  TYPE     IN USE TRAN TIME  RATE   PER I/O V  DATA SET NAME

P PAGCOM 02E6 33903    100.0  ------   -------    0.000    SYS1.PLPA.PAGCOM
C PAGCOM 02E6 33903    63.04  ------   -------   16.500    SYS1.COMMON.PAGCOM
L PAGEP2 098E 33903     0.00  ------   -------    0.000 Y  SYS1.LOCAL.PAGEP2
S N/A    N/A  N/A       8.58  ------   -------   10.939    N/A
```

RMF Updates (cont)

- **RMF Monitor III STORF Report includes SCM usage in 'Aux Slots' count (RMF III → Resource → STORF):**

```
                    RMF V2R1   Storage Frames                    Line 2

Samples: 100     System: 4381  Date: 05/09/13  Time: 10.48.20  Range: 100

           Service    -- Frame Occup.-- - Active Frames - AUX    PGIN
Jobname  C Class   Cr TOTAL  ACTV  IDLE  WSET FIXED   DIV SLOTS RATE

J273     S SYSSTC     8332  8332     0  8332   734     0     0    0
OMVS     S SYSTEM     6560  6560     0  6560   241     0     0    0
HZSPROC  S SYSSTC     3996  3996     0  3996   151     0     0    0
XCFAS    S SYSTEM     3637  3637     0  3637   872     0     0    0
PGOUT30L S SYSSTC     3551     0  3551     0    70     0    30    0
```

zFlash SVC Dump - RMF Page Data Set Report Example

- RMF Page Data Set report: average over 6 minutes

P A G E D A T A S E T A C T I V I T Y

z/OS V1R13 SYSTEM ID P41 DATE 10/09/2012 INTERVAL 05.59.585
RPT VERSION V1R13 RMF TIME 14.30.28 CYCLE 0.050 SECONDS

NUMBER OF SAMPLES = 7,190

PAGE DATA SET AND SCM USAGE

PAGE SPACE TYPE	VOLUME SERIAL	DEV NUM	DEVICE TYPE	SLOTS ALLOC	SLOTS USED MIN	SLOTS USED MAX	SLOTS USED AVG	BAD SLOTS	% IN USE	PAGE TRANS TIME	NUMBER IO REQ	PAGES XFER'D	V I O	DATA SET NAME
PLPA	41PAG0	5473	33903	98999	14655	14655	14655	0	0.00	0.000	0	0		SYS1.P41.PLPA
COMMON	41PAG0	5473	33903	89999	61	61	61	0	0.00	0.000	2	32		SYS1.P41.COMMON
LOCAL	41PAG0	5473	33903	410399	0	0	0	0	0.00	0.000	0	0	Y	SYS1.P41.LOCAL
SCM	**N/A**	**N/A**	**N/A**	**33554K**	**6030K**	**6108K**	**6061K**	**0**	**4.24**	**0.000**	**721516**	**17.19M**		**N/A**

SVC Dump Statistics

- VERBX IEAVTSFS
- Shows total dump capture time, system/task non-dispatch time, page operations required to dump requested address space (real-to-real copies, page-ins, etc)

SVC Dump Statistics (cont)

Dump start	10/09/2012 14:30:29.867495
Dump end	10/09/2012 14:30:44.224584
Total dump capture time	**00:00:14.357089**
System nondispatchability start	10/09/2012 14:30:29.870030
System set nondispatchable	10/09/2012 14:30:29.870048
Time to become nondispatchable	00:00:00.000017

SVC Dump Statistics (cont)

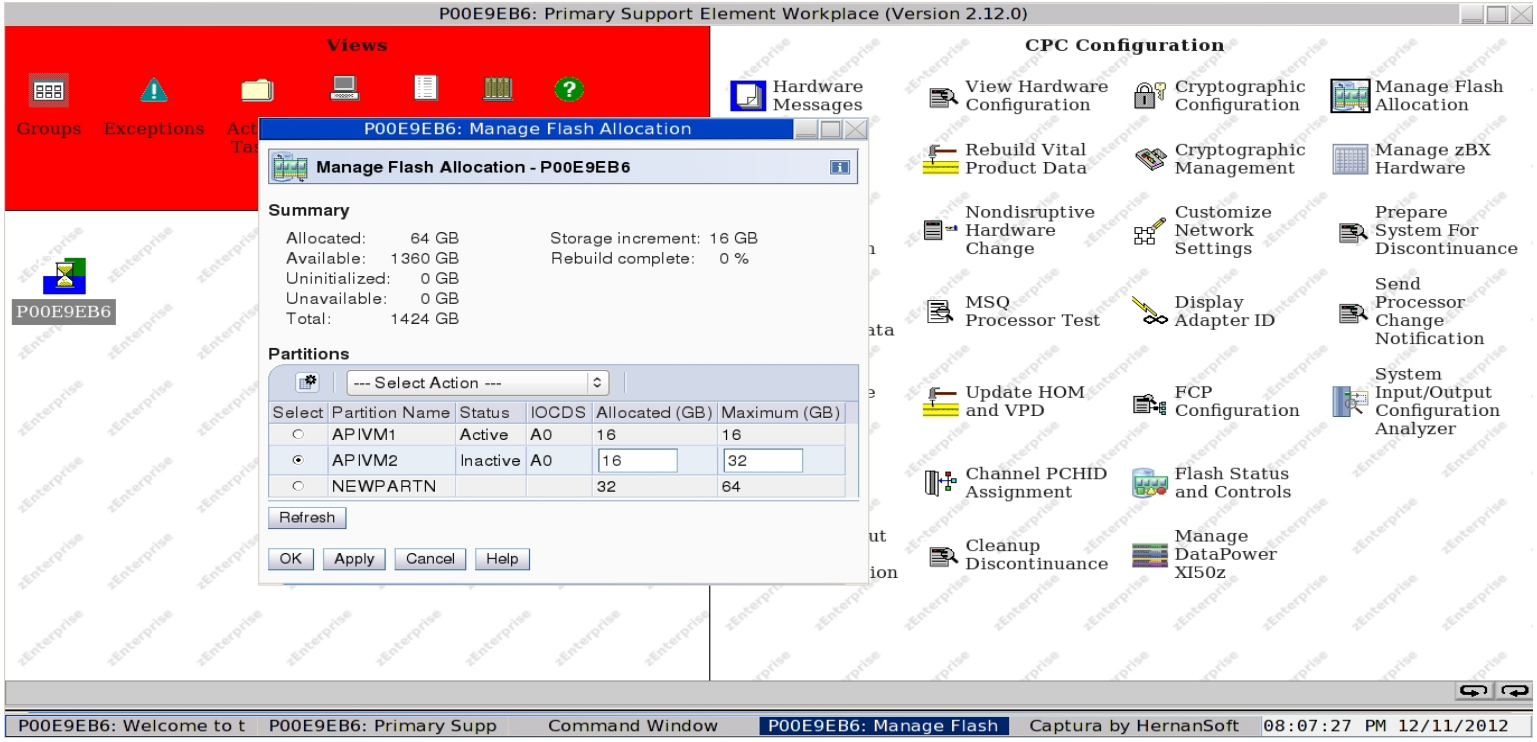
```
Asid 0071:
  Local storage start              10/09/2012 14:30:30.424083
  Local storage end                10/09/2012 14:30:43.011936
  Local storage capture time       00:00:12.587853
  Tasks reset dispatchable         10/09/2012 14:30:43.011944
  Tasks were nondispatchable       00:00:12.587861

  Defers for frame availability                0
  Pages requiring input I/O               170196
  Source page copied to target             16987
  Source frames re-assigned               566614

  Source AUX slot IDs re-assigned          15749
```

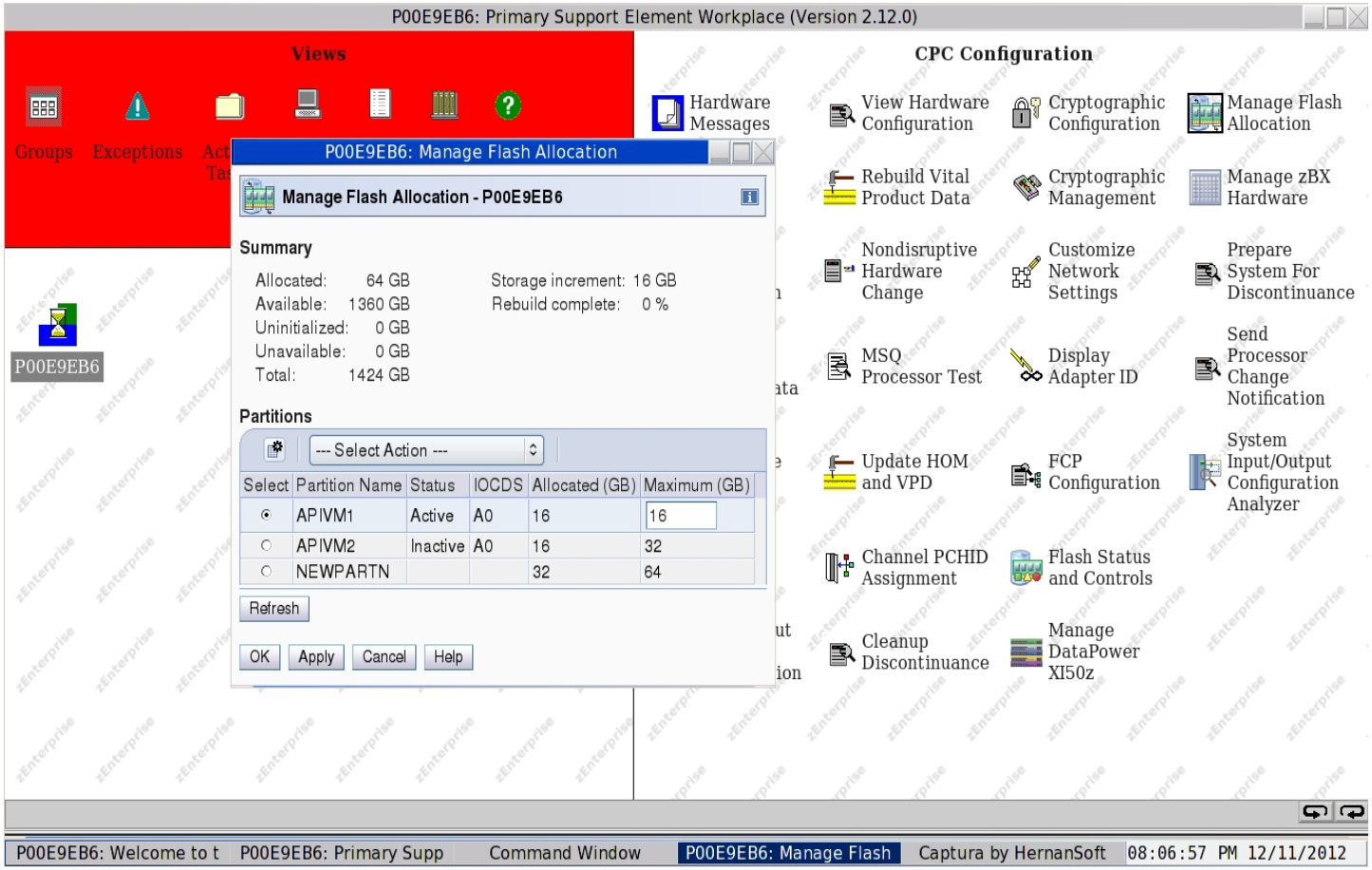
Manage Flash Allocation - Change zFlash Allocation

- Allocated can only be changed for inactive partitions (APIVM2) or undefined partitions (NEWPARTN)
- Changing Allocated results in loss of data
- Changing allocations for an inactive partition:



Manage Flash Allocation - Change zFlash Allocation

- Changing allocations for an active partition (notice only the maximum can be altered):

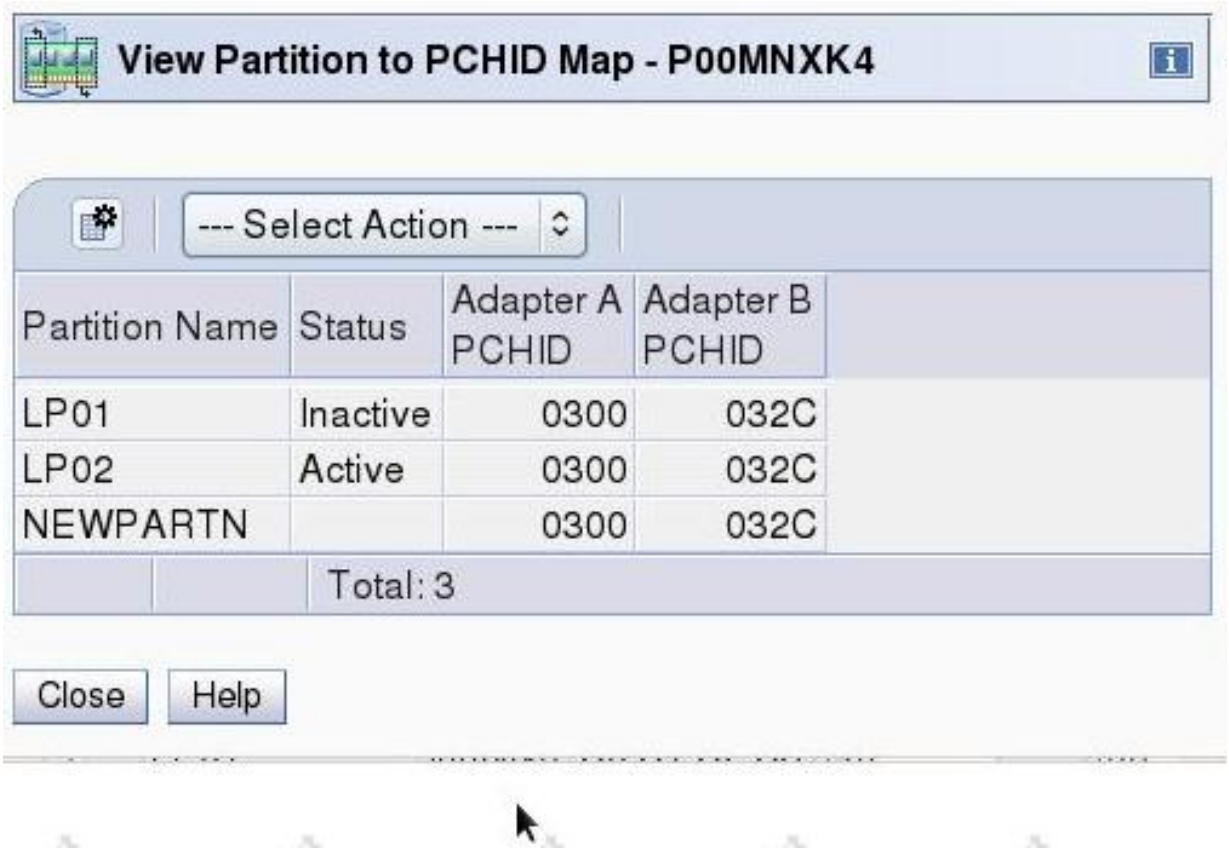


Manage Flash Allocation - Remove zFlash Allocation

- Remove Allocation can only be performed for an inactive partition
- All data will be lost
- A warning message will be issued and confirmation required before the Remove Allocation is done

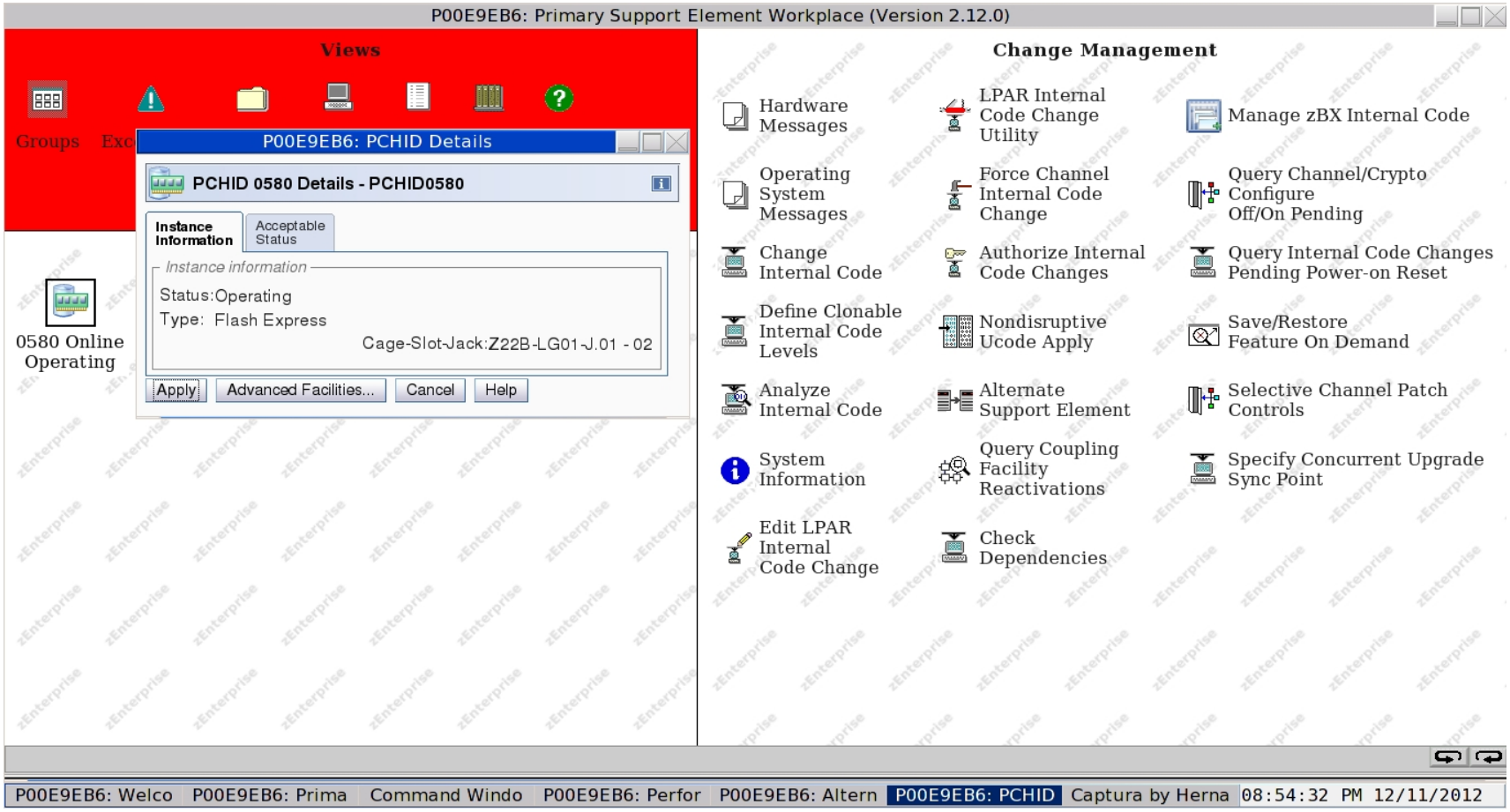
Manage Flash Allocation - View Partition to PCHID Map

- Shows information for all PCHIDs
- SE and HMC



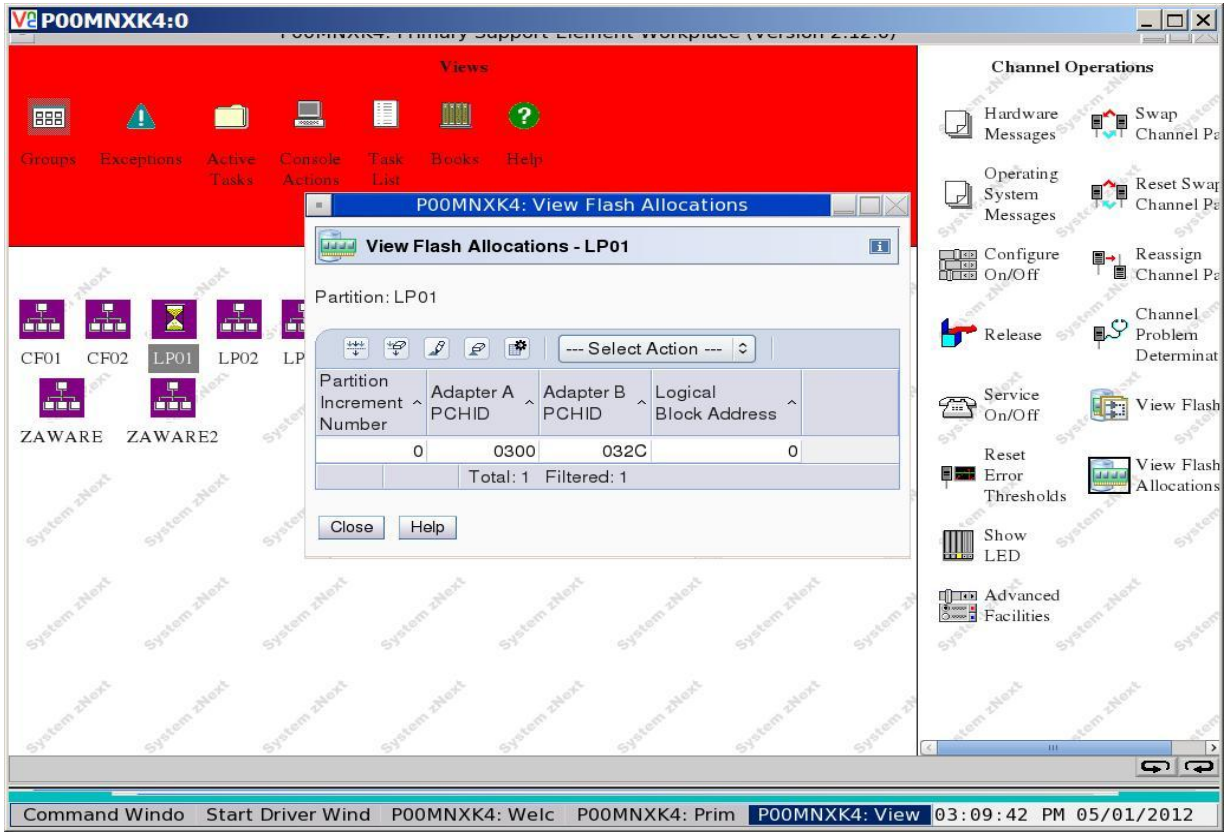
zFlash PCHID Details

- Display information for one PCHID
- SE only



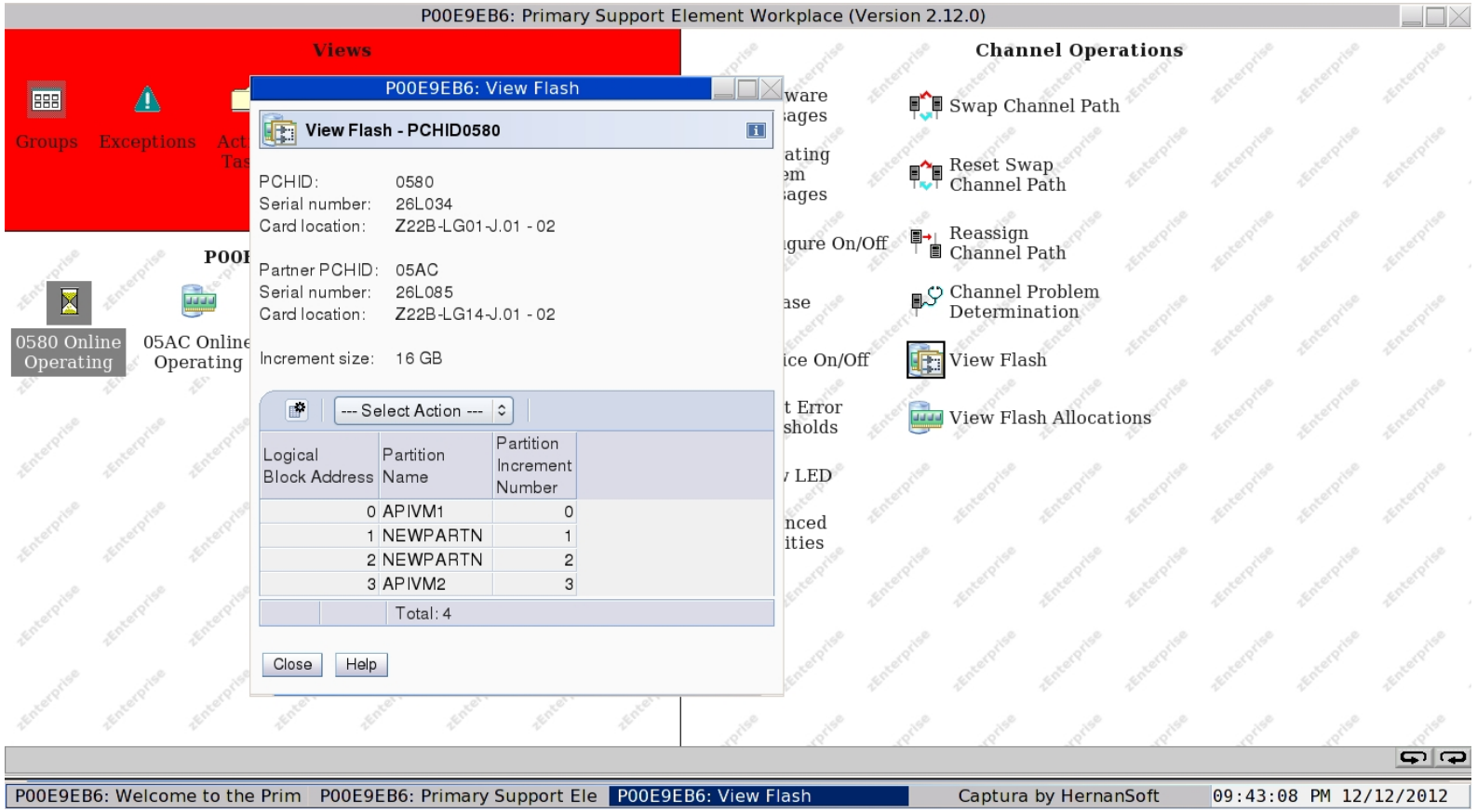
View Flash Allocations Task

- Display information for one partition
- SE only



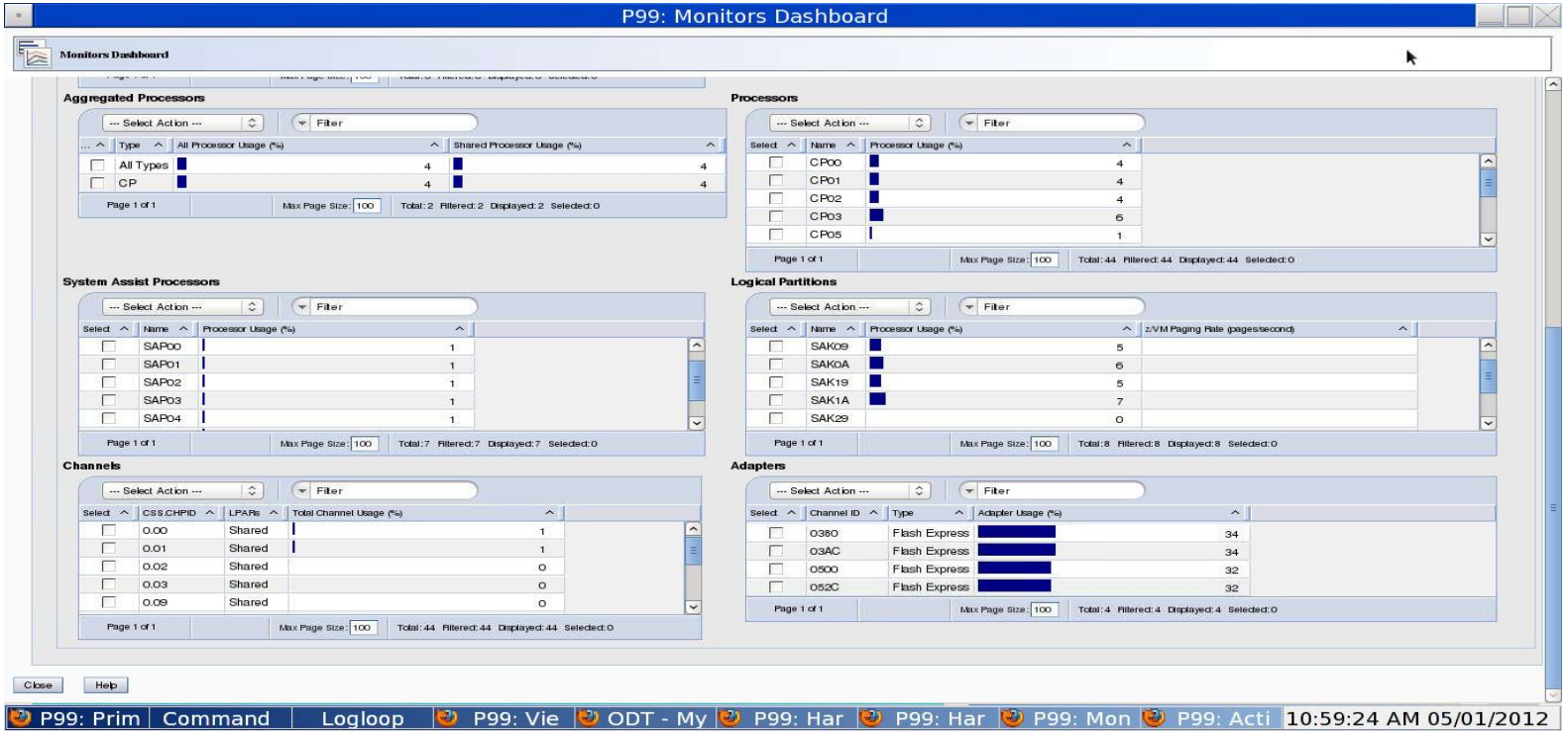
View Flash

- For the selected PCHID, shows you some physical and allocation details
- SE Only



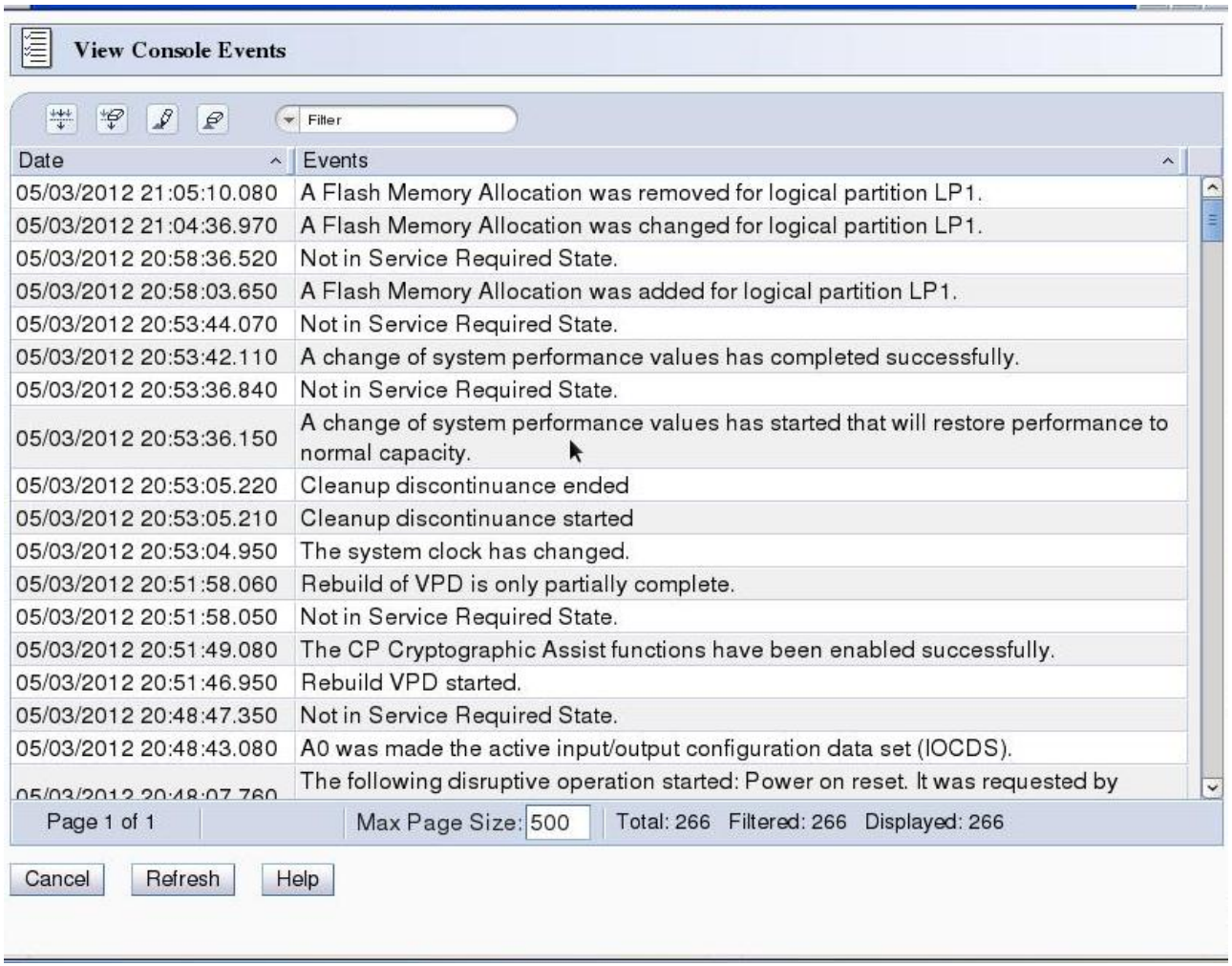
System Activity Display (SAD) / Monitors Dashboard

- Zflash is not supported by System Activity Display (SAD)
- It is supported by Monitors Dashboard. Refer to the new “Adapters” table in the lower right.



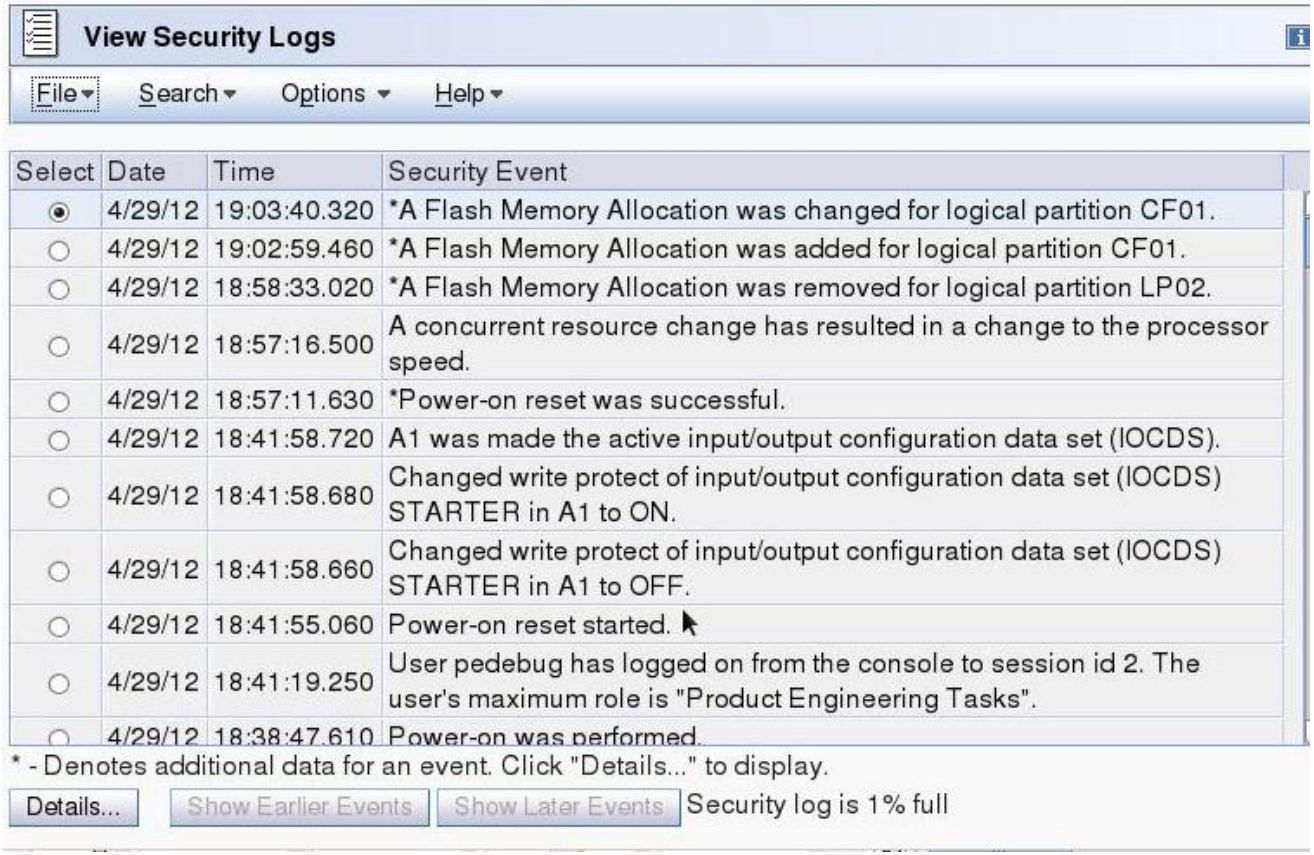
Console Events

- Event logs will be generated when a flash allocation is added, changed or removed:



Security Logs

- Appropriate security logs will be generated for zFlash-related actions.
- Examples:

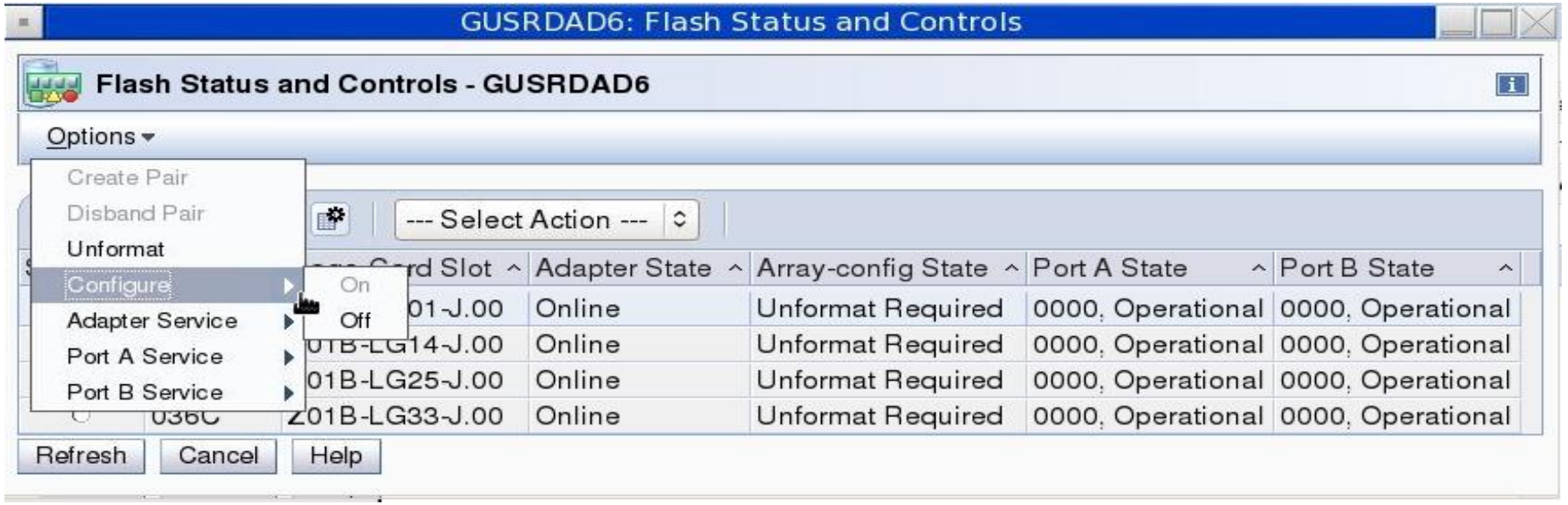


Flash Status and Controls States (Service Personnel Only)

- Adapter States:
  - Not Installed
  - Online
  - Online in progress
  - Offline
  - Offline check stopped
  - Offline in progress
  - Online check stopped
  - Service
  - Configuration error
- Array States:
  - Not formatted
  - Format in progress #% complete
  - Unformat in progress
  - Formatted
  - Configuration error
  - Rebuild in progress #% complete
  - Exposed
  - Unformat required
- Port States:
  - Unknown
  - Operational
  - Service
  - Dangling
  - Check stopped
  - Configuration error
  - Entering service mode
  - Exiting service mode
  - Repair in progress
  - Not installed

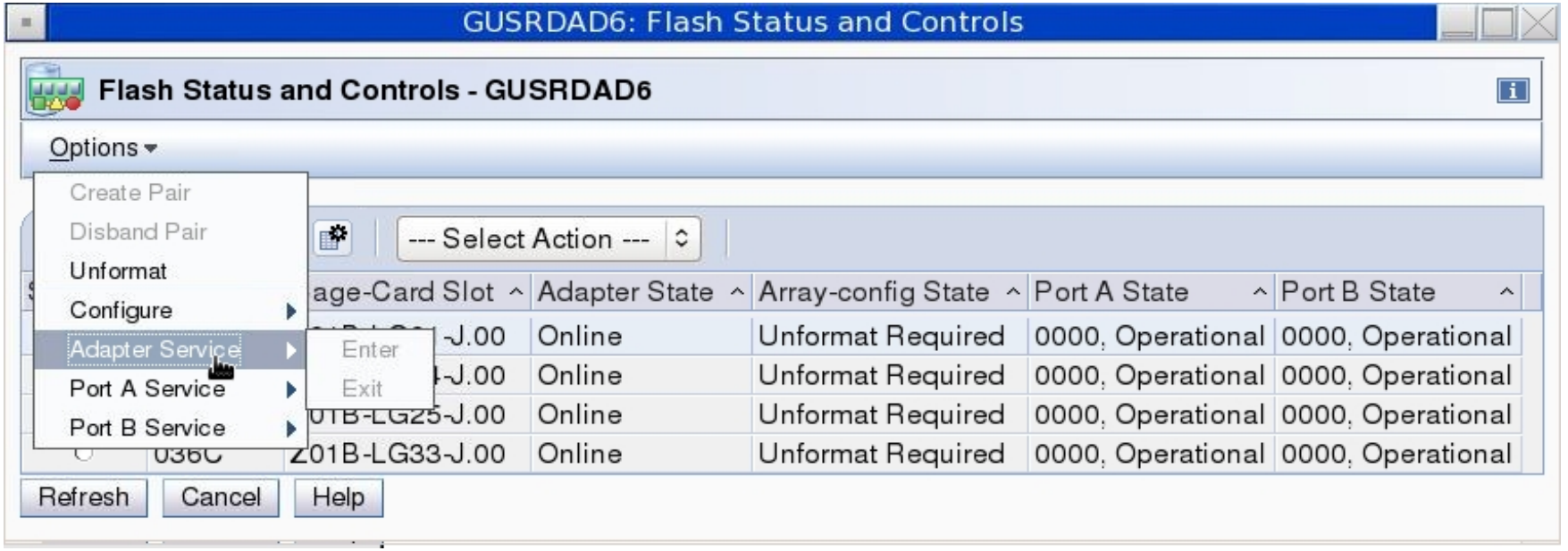
Configure the zFlash Adapter On/Off (Service Personnel Only)

Configure the adapter online or offline.



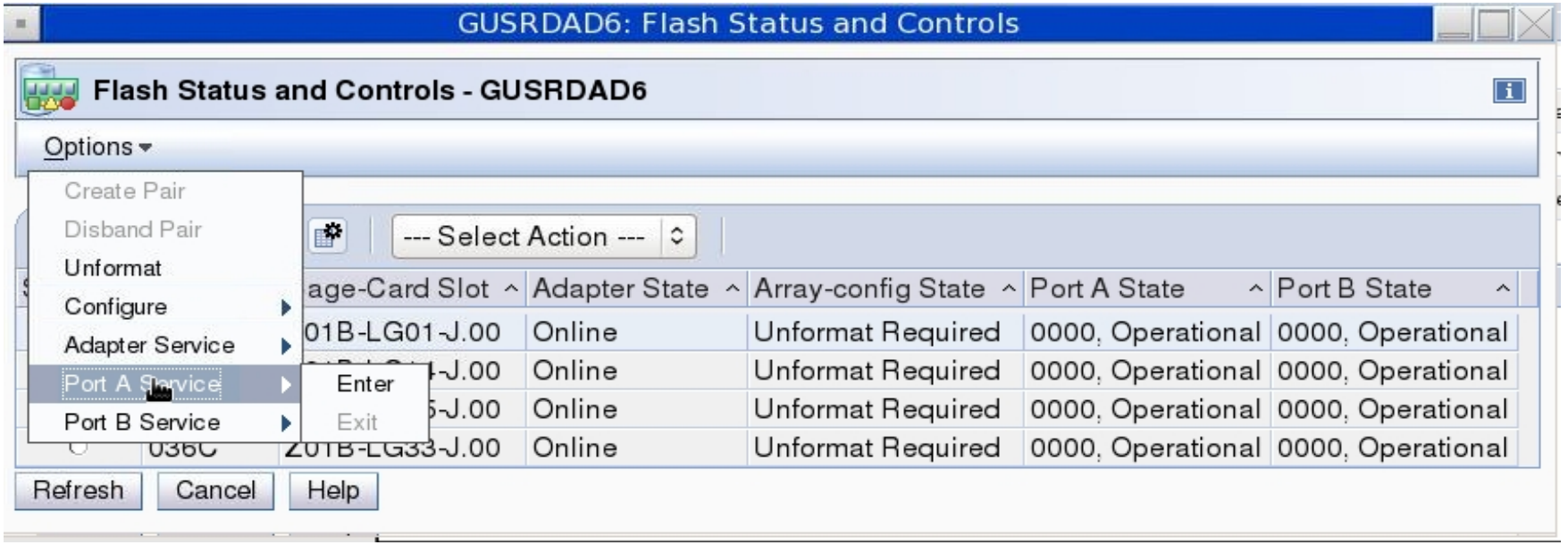
Adapter Service Mode (Service Personnel Only)

Enter adapter service mode or exit adapter service mode, they are both disabled right now because the adapter status in Online.



Port Service Mode (Service Personnel Only)

- Enter Port A service mode or exit Port A service mode, the exit is disabled currently because the port is currently in operational state.
- Port B Service behaves the same way.



Flash Express Performance Results

- **All performance information was determined in a controlled environment.**
- **Actual results may vary.**
- **Performance information is provided “AS IS” and no warranties or guarantees are expressed or implied by IBM.**

Flash Express Performance Test Setup

- z/OS Tests were designed to demonstrate flash performance under paging workloads that are typically encountered in a z/OS enterprise environment
  - SSD performance is not only about the number of IOPS but about steady performance over time and consistent latency
    - **Preconditioned SSDs** with random-write IO engage the device's wear leveling, error handling, and flash management algorithms
  - Comparison DASD Characteristics used **current device configurations**
    - DS8800 model 2107-951
    - 60 GB cache, cache hit rates of 95-100% were observed during the tests
    - DASD was not shared with any other systems and did not have any I/O traffic other than the paging traffic used for these tests
    - Configured 16 local page datasets spread across 8 LCUs

Flash Express Performance Benefits

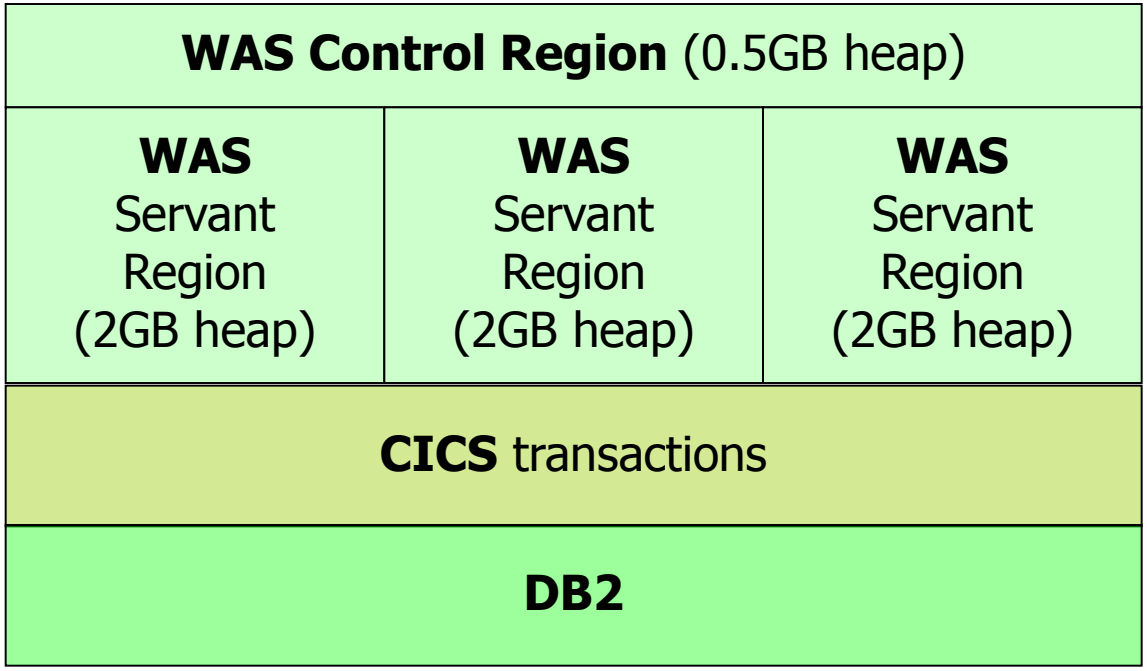
Test Results

- FLASH paging benefits
  - Improved availability through faster paging at critical times
    - Faster workload transitions (e.g.; morning startup)
      - *meaning less time to reach peak transaction rates*
    - Faster SVC dumps (reduced **non-dispatchable** time)
      - *meaning higher availability – more transactions can be run*
- Pageable Large Page benefit
  - Java realizes performance benefits from use of large 1MB pageable pages
    - Large pages benefits for JIT Code Cache, 31 bit Java applications
    - No authorization needed to access fixed large pages
    - Approximately 5-8% CPU improvement from PLP

Workload Configuration Block Diagram

Building block – A WAS instance accessing CICS and DB2

Each WAS instance has a WAS Control Region and 3 WAS Servant Regions.
Each WAS Control Region has a 0.5GB heap plus a JIT Code cache.
Each WAS Servant Region has a 2GB heap plus a JIT Code Cache.



2-book zEC12

Test Configuration

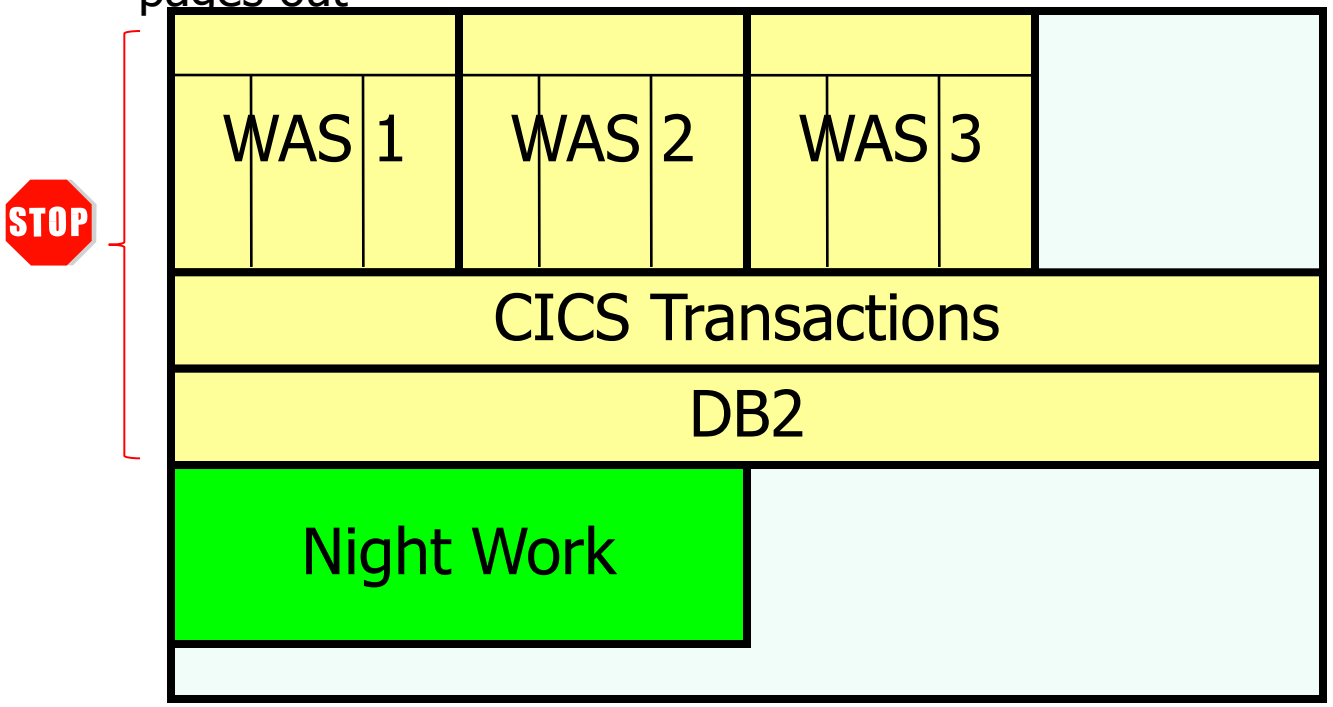
- WAS 7 (3 servants each with two GB heap) + 1 control region (.5 GB Heap)
- CICS V4.2, DB210 on a zEC12
- Storage: DS8800 2107-951 with 60GB cache, very fast device
- Tests simulated morning transition time typical of trading or call center work
- SVC dump measurements were taken for an 18 GB dump.

I. Morning Transition

Transition from night batch to OLTP

WAS workload to CICS and DB2 represents OLTP work which is then stopped

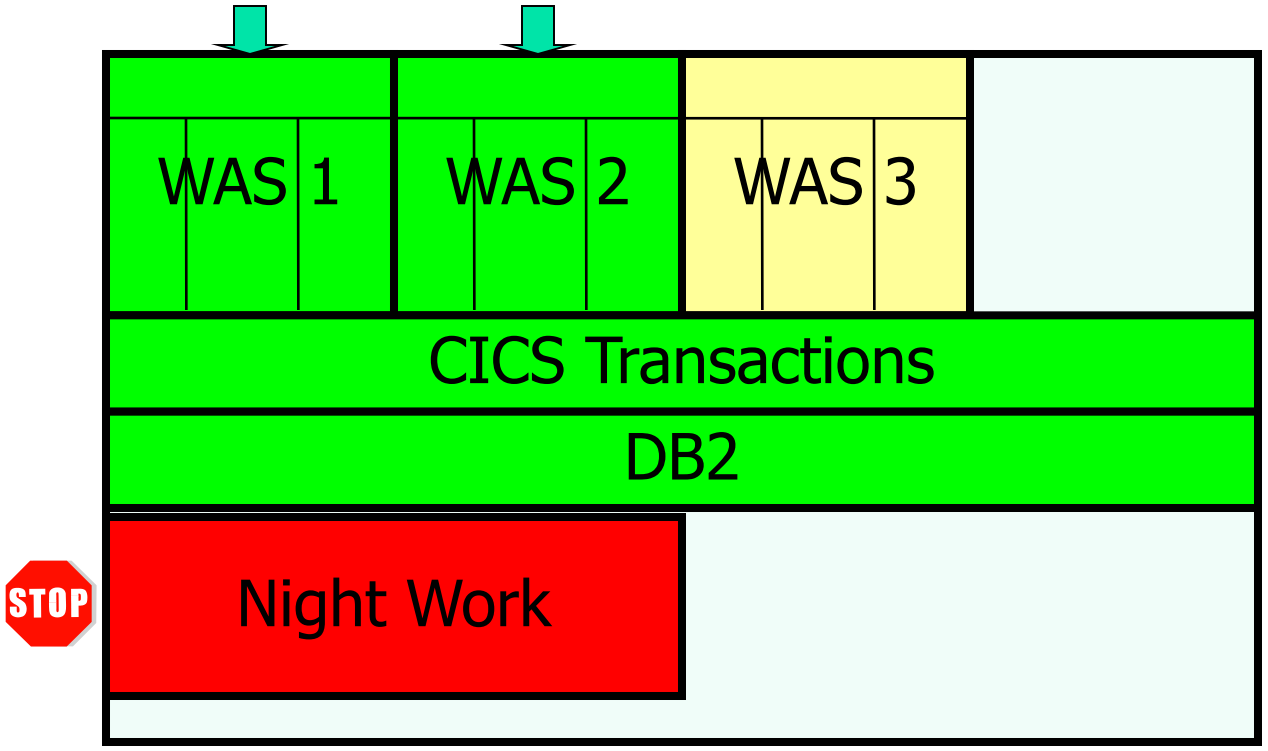
Simulated overnight work consumes real storage pushing other pages out



~21GB is pushed out to Auxiliary storage

Morning Transition - Transition from night batch to OLTP

The "Night Work" is then stopped and OLTP work is started (WAS 1 and WAS 2)
Measure the time needed to bring the OLTP work to full speed.

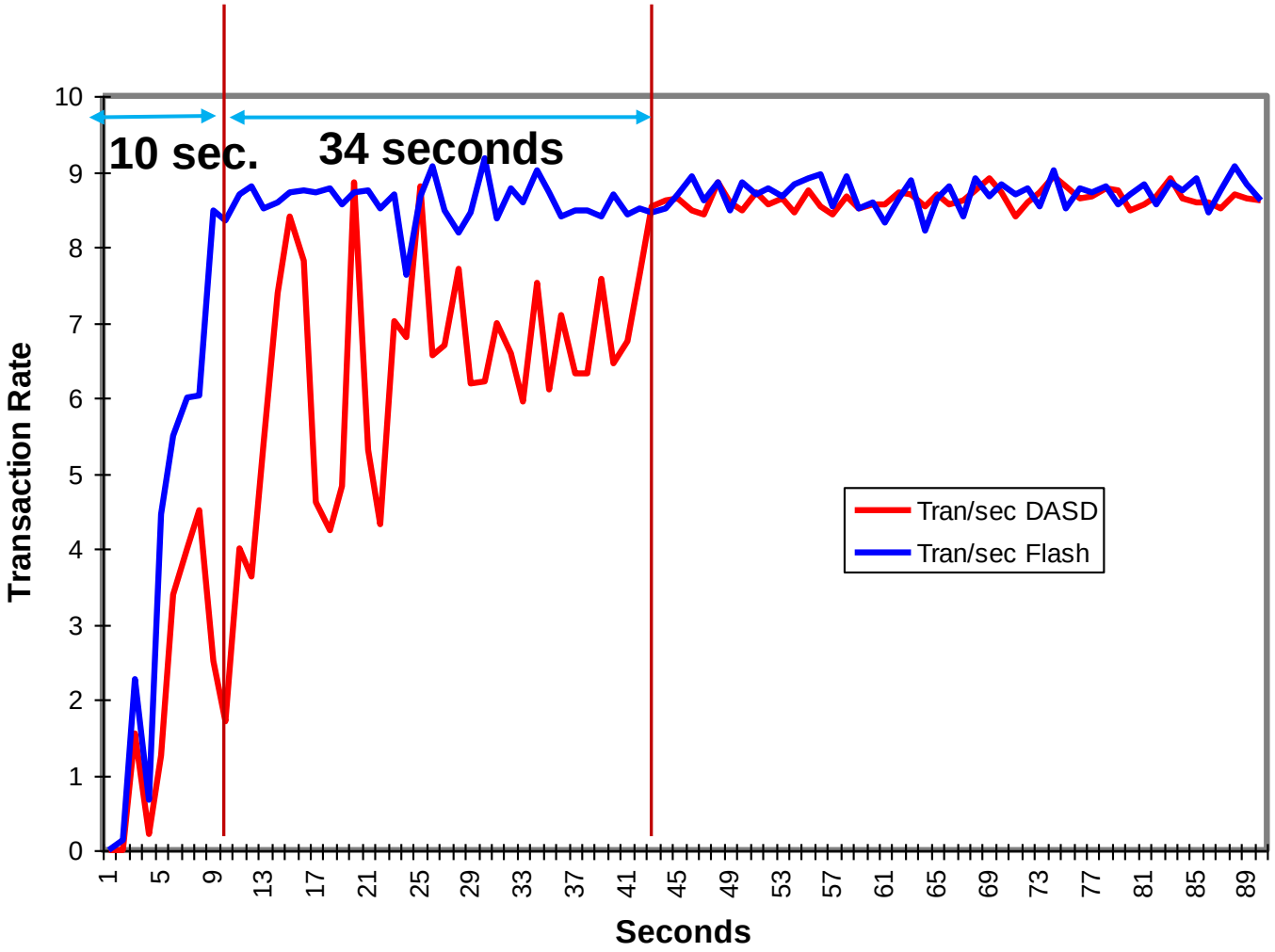


~14GB is paged back in from Auxiliary storage

Workload Transition

Morning Transition - Results

❖ **During morning transition, workloads using Flash Express reached peak throughput in under 1/4th the time**



Workload Transition

Paging to **DASD** required about **44 seconds** for the workload to reach steady state

Paging to **Flash** required only **10 seconds** for the workload to reach steady state

Morning Transition - Results Apparent in First 45 Seconds

Transaction completion & response time	DASD	Flash	Improvement
Total Transactions within first **45** seconds	251	343	**37% increase**
Average response time within first **45** seconds	0.62	0.06	**90% reduction**

Units in seconds

❖ **Paging to Flash Express during morning transition showed up to a 10 times faster response time and up to a 37% increase in throughput within the first 45 seconds**

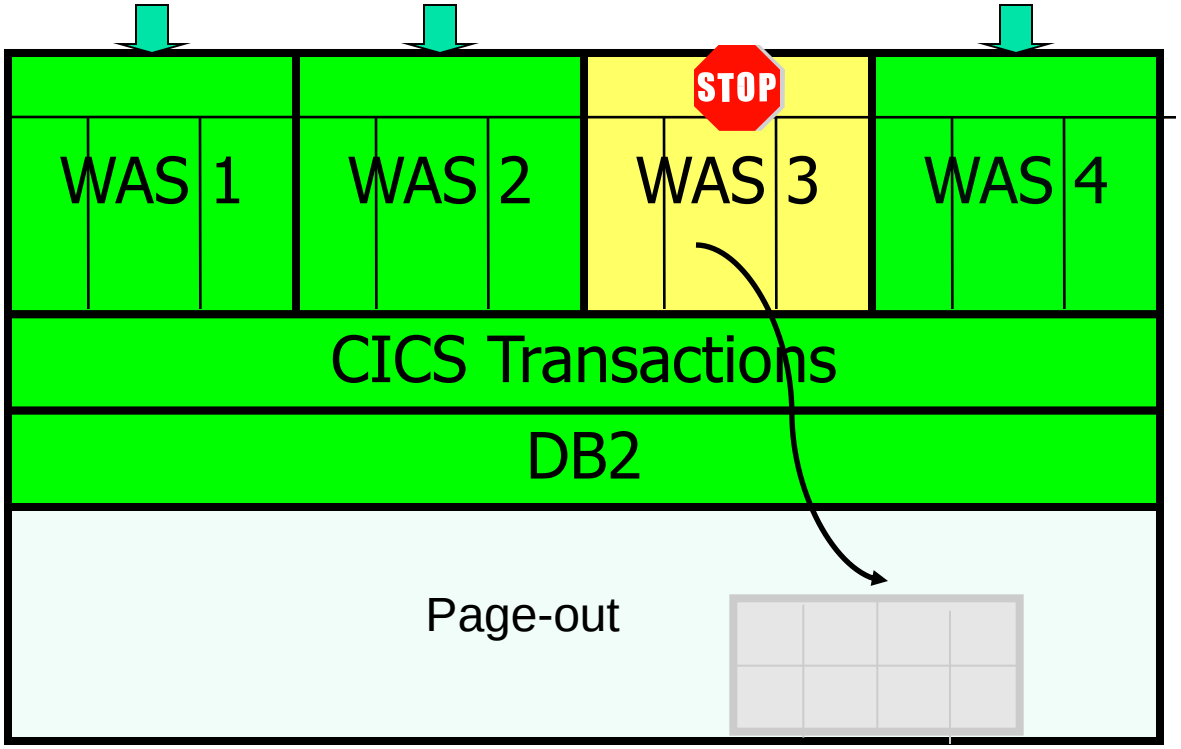
(1) Test was for the first 45 seconds of morning transition time

Workload

II. SVC Dump

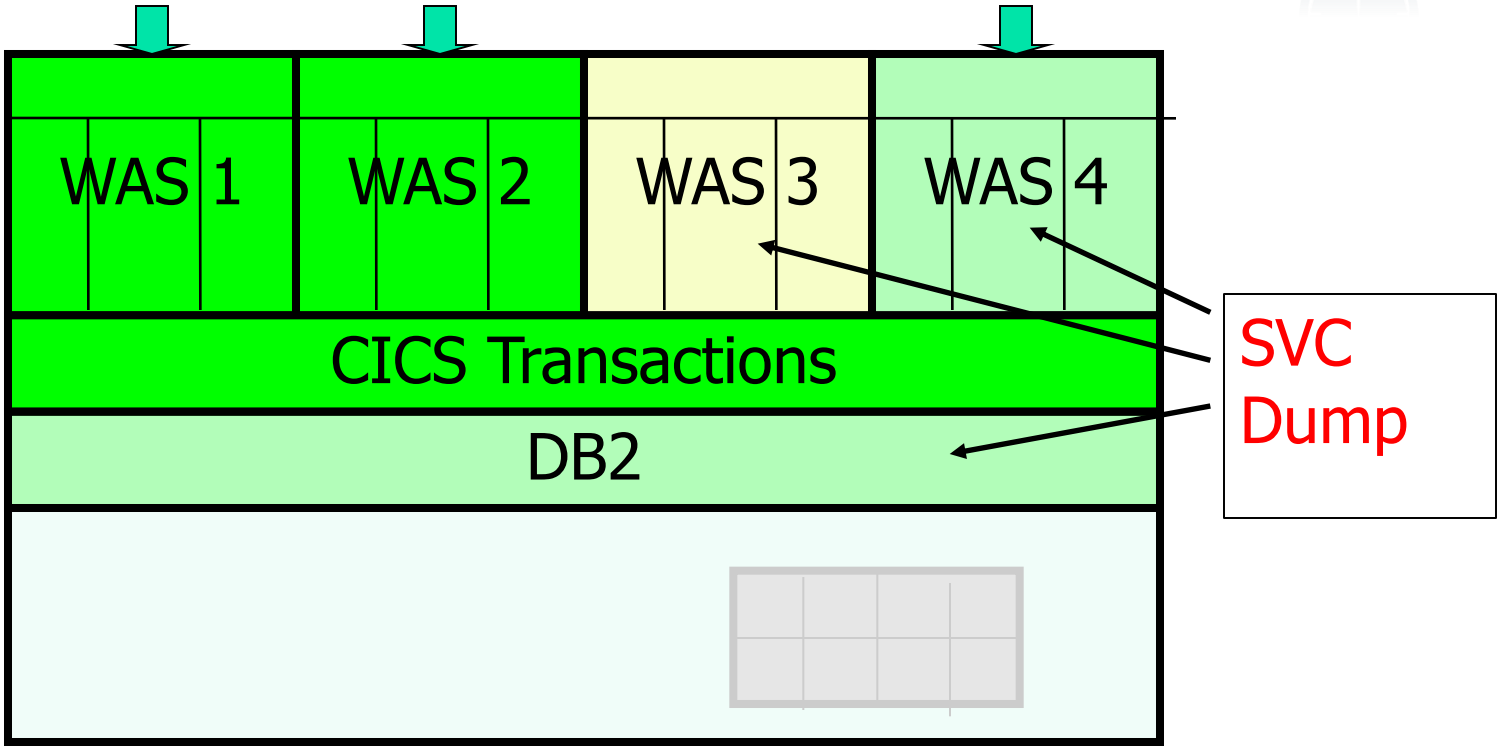
SVC dump with pages out

Three of four WAS instances were active.
One WAS instance was stopped and most pages were paged out.



SVC Dump- Diagnostics capture

Capture an SVC dump of WAS instance 3 and 4, and DB2.
Measure the capture time for the SVC dump.



SVC Dump - Results

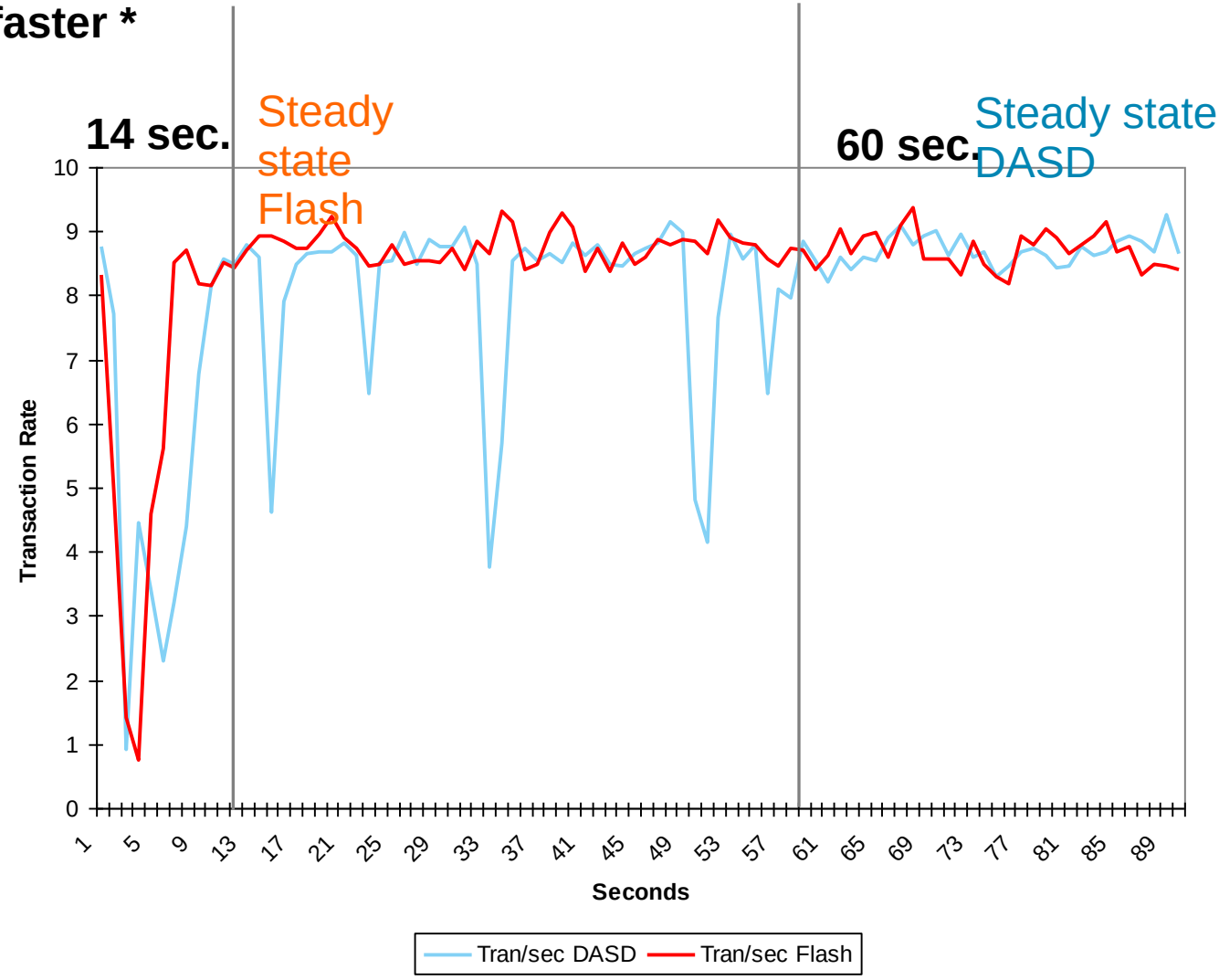
❖**Flash Express SVC dump elapsed time was up to 25% shorter**

SVC Dump Metrics	DASD	Flash
SVC Dump size (in bytes):	18GB	18GB
% of pages from Aux storage:	50%	53%
DUMP Elapsed time:	**189**	**143**
Max address space non-dispatchable seconds	**58.89**	**13.74**
System non-dispatchable seconds	**1.34**	**0.55**

Let's graph these results….

SVC Dump - Results

❖**In SVC dump test, steady state performance was achieved up to 4 times faster ***

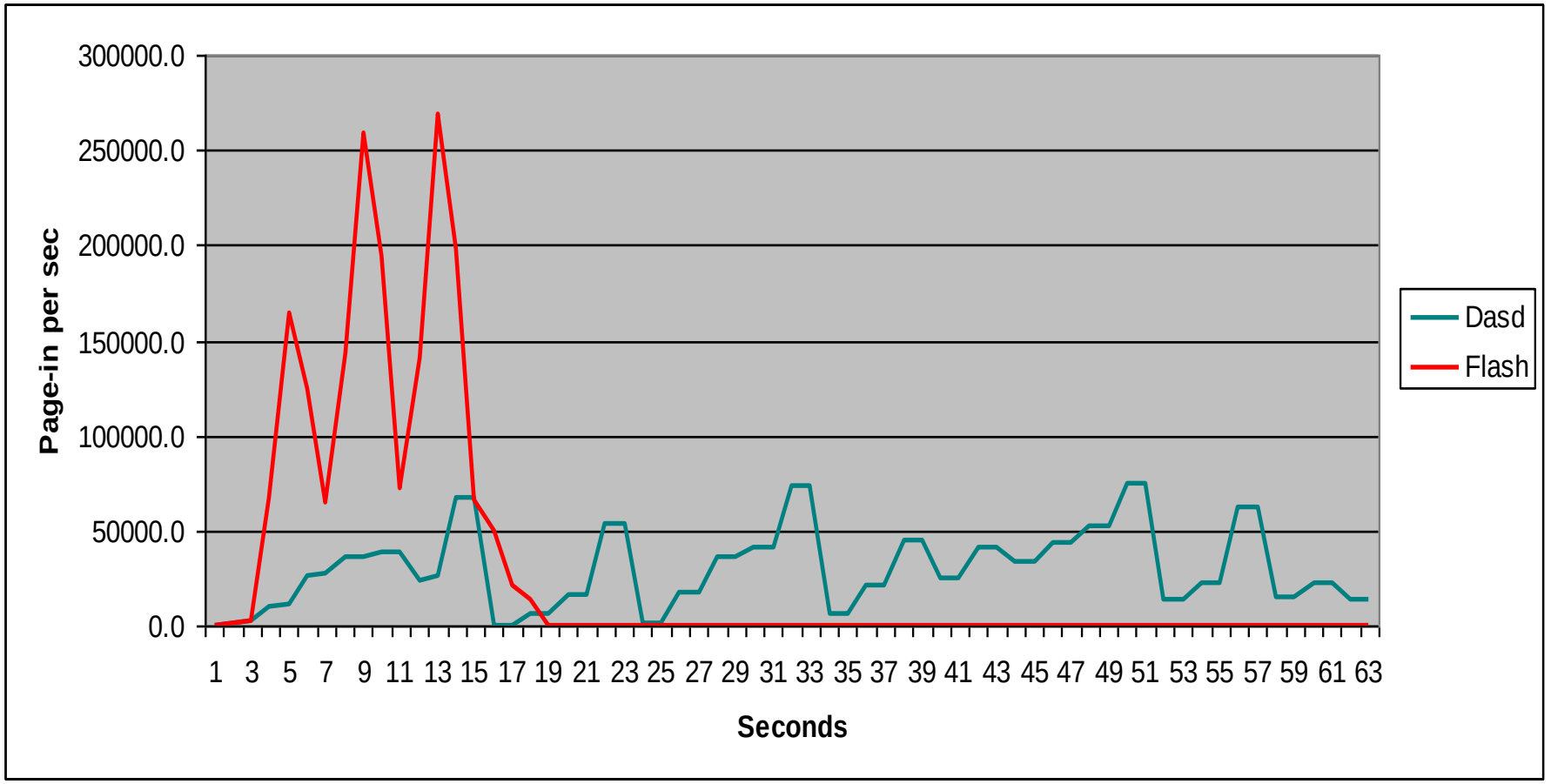


Focus on first 90 seconds.

* ***Transaction steady state was reached in 14 seconds with Flash Express, vs. 60 seconds DASD***

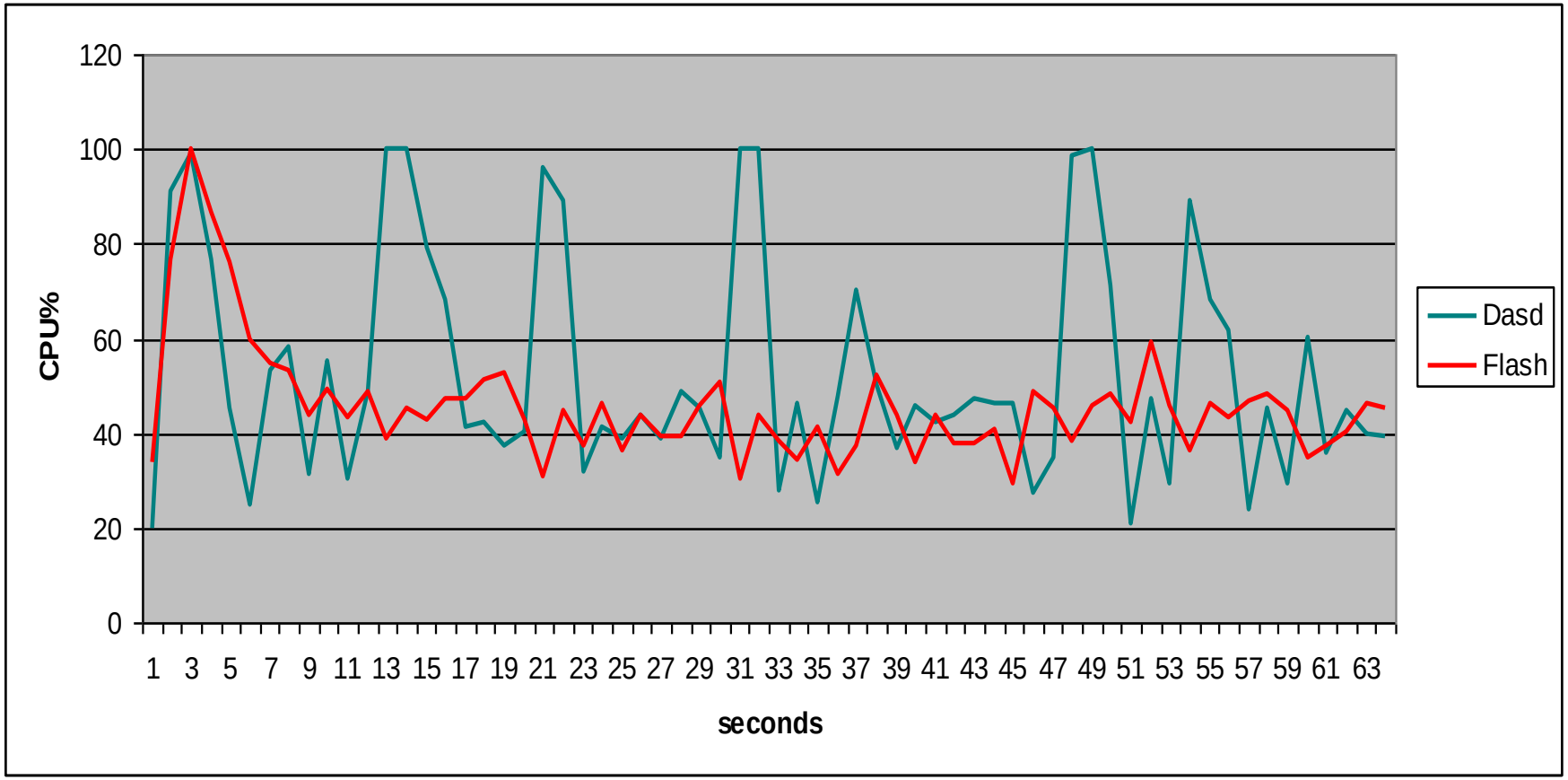
zFlash SVC Dump – Page-in Rate

- Peak page-in rate with Dasd: 75,000 pages per sec
- Peak page-in rate with SCM: 260,000 pages per sec



zFlash SVC Dump – CPU Usage

- CPU peaks correspond to peaks in page-in rates
- Several peaks when using DASD while one peak when using SCM



zFlash SVC Dump - RMF Page Data Set Report Example

- RMF Page Data Set report: average over 6 minutes

P A G E D A T A S E T A C T I V I T Y

z/OS V1R13 SYSTEM ID P41 DATE 10/09/2012 INTERVAL 05.59.585
RPT VERSION V1R13 RMF TIME 14.30.28 CYCLE 0.050 SECONDS

NUMBER OF SAMPLES = 7,190

PAGE DATA SET AND SCM USAGE

PAGE SPACE TYPE	VOLUME SERIAL	DEV NUM	DEVICE TYPE	SLOTS ALLOC	SLOTS USED MIN	SLOTS USED MAX	SLOTS USED AVG	BAD SLOTS	% IN USE	PAGE TRANS TIME	NUMBER IO REQ	PAGES XFER'D	V I O	DATA SET NAME
PLPA	41PAG0	5473	33903	98999	14655	14655	14655	0	0.00	0.000	0	0		SYS1.P41.PLPA
COMMON	41PAG0	5473	33903	89999	61	61	61	0	0.00	0.000	2	32		SYS1.P41.COMMON
LOCAL	41PAG0	5473	33903	410399	0	0	0	0	0.00	0.000	0	0	Y	SYS1.P41.LOCAL
SCM	**N/A**	**N/A**	**N/A**	**33554K**	**6030K**	**6108K**	**6061K**	**0**	**4.24**	**0.000**	**721516**	**17.19M**		**N/A**

Stand-Alone Dump

- Improvements in Stand-Alone Dump time when dumping data that are paged out
- Overall 37 second reduction in dump time due to faster page-in of data from aux when using Flash representing approximately a 19% reduction in total dump time for an 36 GB dump

Tests	Total dump time In minutes	Paging I/O wait time In seconds	Batch read rate MB/sec	Total GB dumped	GB of data from aux
DASD Page data sets (DS8800)	00:03:12.92	00:00:41.30	438.06	36.2	17.7
Flash for paging	00:02:35.03	00:00:10.38	1612.30	36.3	16.3

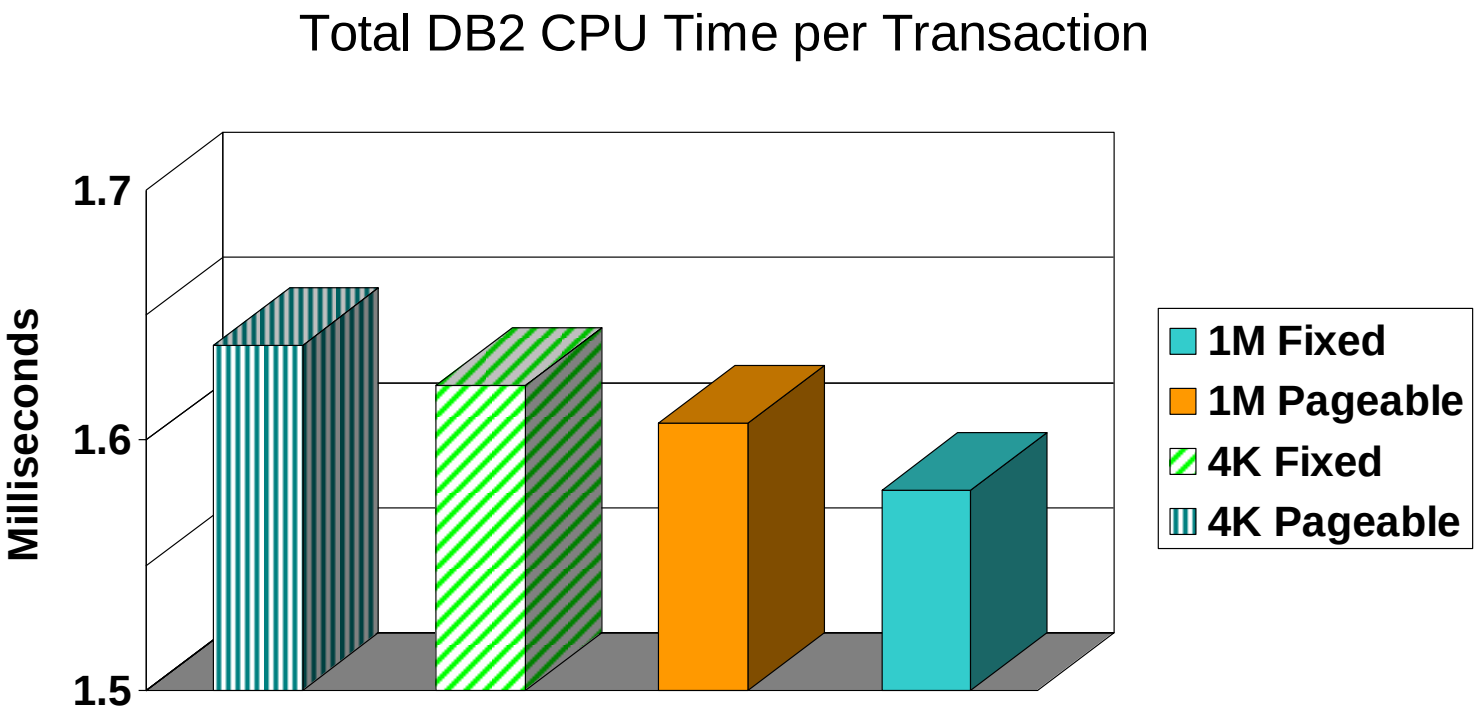
z/OS V1.13 1 MB Pageable Large Page Exploitation

- Benefits of large pages:
  - Better performance by decreasing the number of TLB misses that an application incurs
  - Less time spent converting virtual addresses into physical addresses
  - Less real storage used to maintain DAT structures
- Fixed large pages vs pageable large pages:
  - Fixed large pages are backed at allocation. Pageable large pages are backed when referenced.
  - Use of fixed large pages for unathorized users is controlled by a RACF profile (IARRSM.LRPAGES). No RACF authorization to use pageable large pages.
  - Fixed large pages stay as 1 MB pages while pageable large pages may be demoted to 4K pages in certain situations.
- Performance:
  - Java: performance with pageable 1MB large pages is equivalent to 1MB fixed large pages for java heap: up to 8% ITR impact
  - IMS using pageable large pages: up to 1% system ITR improvement.
  - DB2 using pageable large pages: up to 3% system ITR improvement.

Pageable 1MB Frames – Example from IBM Brokerage Workload

All of buffer pools are backed by real storage – DB2 10

- zEC12 16 CPs, 5000-6000 tps (simple to complex transactions)
  - 120GB real storage with 70GB LFAREA configured for 1MB measurements
- 1MB Pageable frames are 2% better than 4KB pageable frames for this workload
  - 70GB buffer pools are used, 8-10 sync I/O per transaction
- 1MB frames with PageFixed is the best performer in general



z/OS Java SDK 7:16-Way Performance Shows up to 60% Improvement

64-bit Java Multi-threaded Benchmark on 16-Way

Aggregate 60% improvement from zEC12 and Java7SR3

- **zEC12 offers a ~45% improvement over z196 running the Java Multi-Threaded Benchmark**
- **Java7SR3 offers an additional ~13% improvement** (-Xaggressive + Flash Express pageable 1Meg large pages)

WAS benchmark: z/OS Performance for Pageable Large Pages

❖ **The WAS Day Trader benchmarks showed up to an 8% performance improvement using Flash Express.**

Java 7 SR3	JIT	Java Heap	Multi Threaded	WAS Day Trader 2.0
31 bit	yes	yes	4%	
64 bit	yes		1%	3%
64 bit		yes	4%	5%

* WAS Day Trader 64-bit Java 7 SR3 with JIT code cache & Java Heap

DETAILS

- **64-bit Java heap (**1M fixed large pages (FLPs) or 1M Pageable (PLPs)) versus 4k pages
  Java heap 1M PLPs improve performance by about
  - **4%** for Multi-Threaded workload
  - **5 %** for WAS Day Trader 2.0
- **64-bit** Java 7 SR3 with **JIT code cache** 1M PLPs vs without Flash
  - **3** % improvement for traditional WAS Day Trader 2.0*
  - **1** % improvement for Java Multi-Threaded workload
- **31-bit** Java 7 SR3 with **JIT code cache and Java heap** 1M PLPs vs without Flash
  - **4 %** improvement for Java Multi-Threaded workload

** Note: This test used 64-bit Java 7 SR3 with JIT code cache & Java Heap leveraging Flash and pageable large pages.*
Also , tests used WAS Day Trader app that supports PLP; earlier version of 31-bit Java did not allocate 1M large pages

Performance Summary for Flash Express(1)

WORKLOAD TRANSITION

- **During morning transition, workloads using Flash Express reached peak throughput in under 1/4th the time**
- **Paging to Flash Express during morning transition showed up to a 10 times faster response time and up to a 37% increase in throughput within the first 45 seconds**

WAS JAVA PERFORMANCE BENCHMARKS

- **The WAS Day Trader benchmarks showed up to an 8% performance improvement using Flash Express.(2)**

** This test used 64-bit Java 7 SR3 with JIT code cache & Java Heap leveraging Flash and pageable large pages.*

Improved Availability During Diagnostics

- **In SVC dumps, availability was up to 4 times higher for workloads and up to twice as high for systems***
- **In SVC dump tests, steady state performance was achieved up to 4 times faster ***
- **Flash Express SVC dump elapsed time was up to 25% shorter**

** Transaction steady state was reached in 14 seconds with Flash Express, vs. 60 seconds DASD.*

DB2

- **Up to 28% improvement in DB2™ throughput due to faster CPU and leveraging Flash Express with Pageable Large Pages (PLP)***
- **Workloads leveraging Flash Express with PLP can see up to a 8%** price performance improvement over the z196.**

* *PLP for DB2 helps DB2 to achieve* ***"additional" up to 3% additional performance on top of zEC12 CPU expected throughput improvements of 25%.***

** *based on average 5% discount for zEC12 workloads under the AWLC pricing plus up to 3% more performance per MSU with Flash Express.*

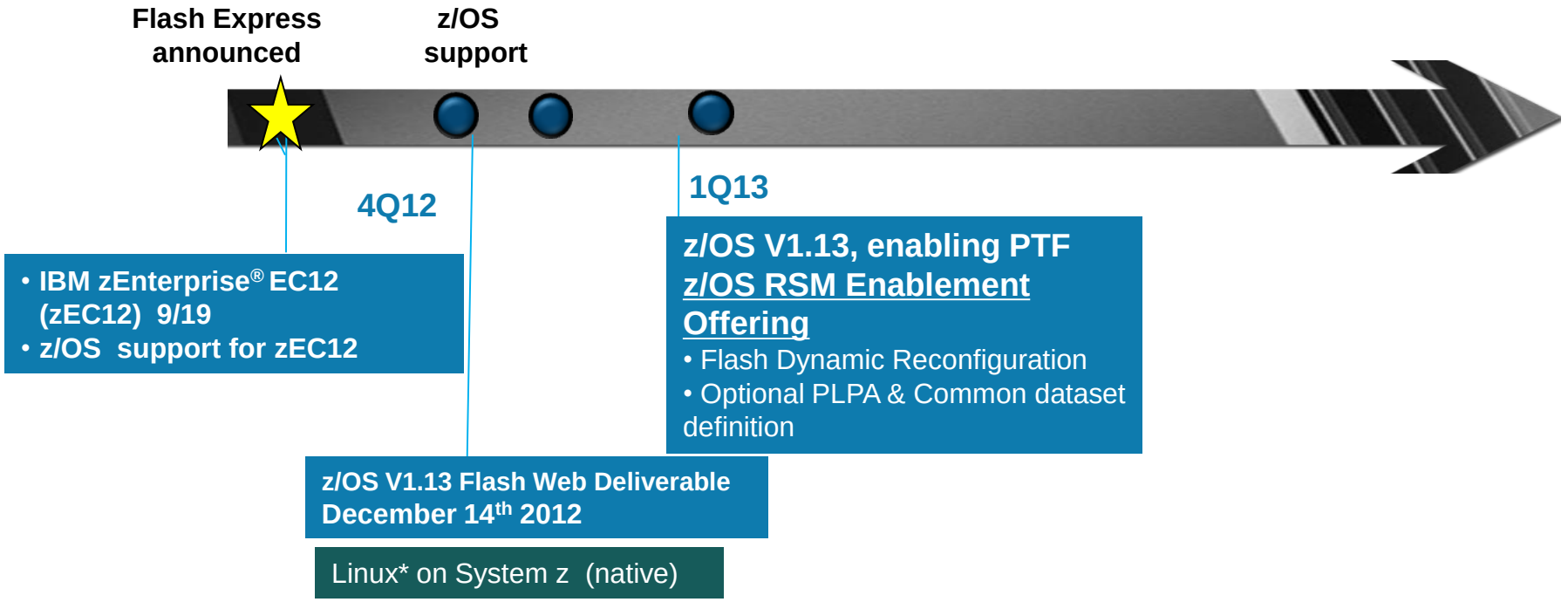
(1) All tests are comparing the use of Flash Express as compared to using DASD (DS8800)

(2) System non dispatchability and address space non dispatchability time were dramatically reduced enabling work to be processed that would otherwise have been stopped

z/OS Flash Roadmap

Flash Express Exploitation

Flash support in z/OS sets the stage for further use



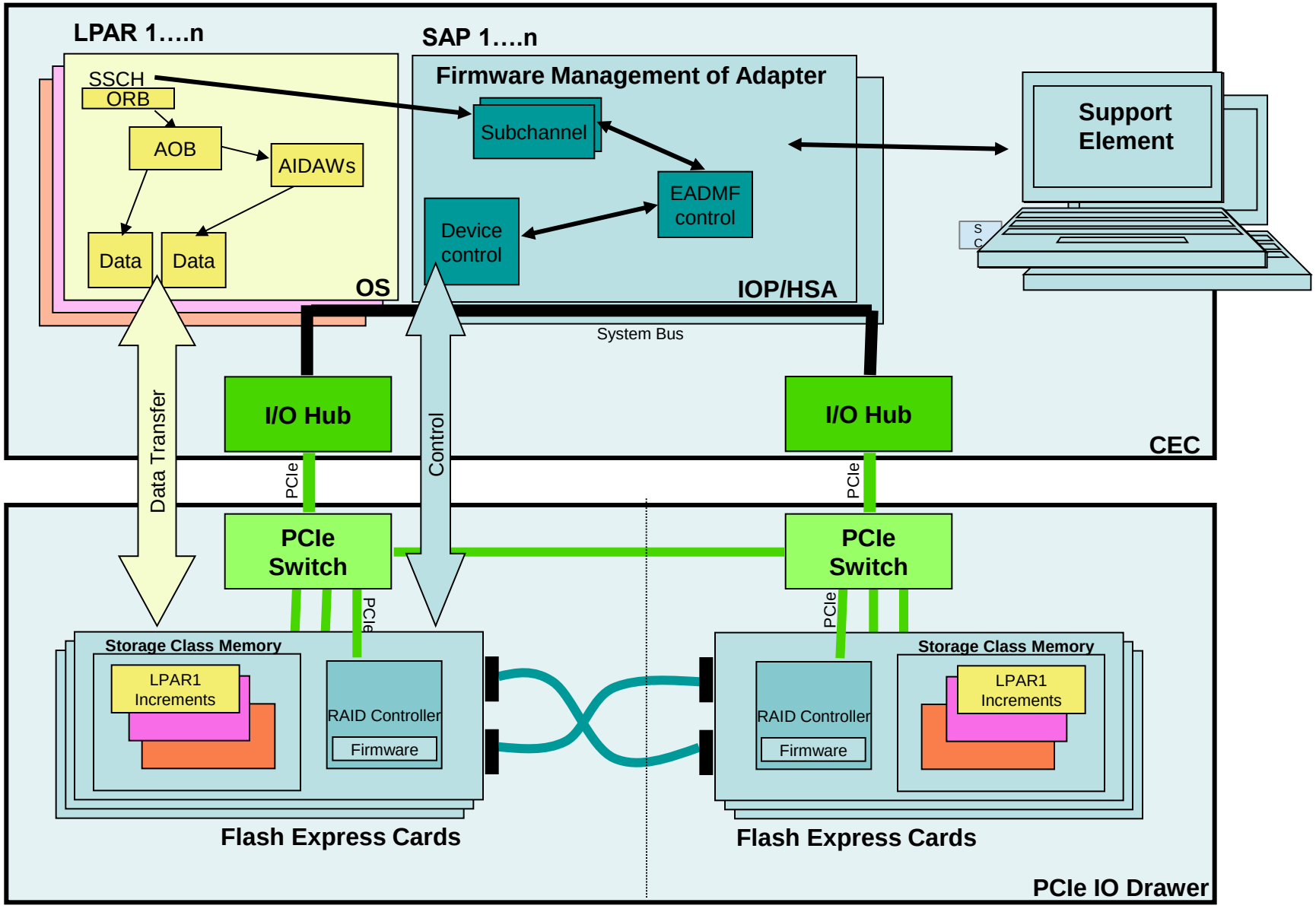
- ***Planned* Flash Express and pageable large page exploiters:**
  - DB2 for z/OS
  - Java SDK7
  - WAS Liberty Profile v8.5
  - IMS™ 12
  - z/OS V1.13 Language Environment®
  - Other (CICS®)

Expect continued middleware exploitation for 1MB pageable large pages

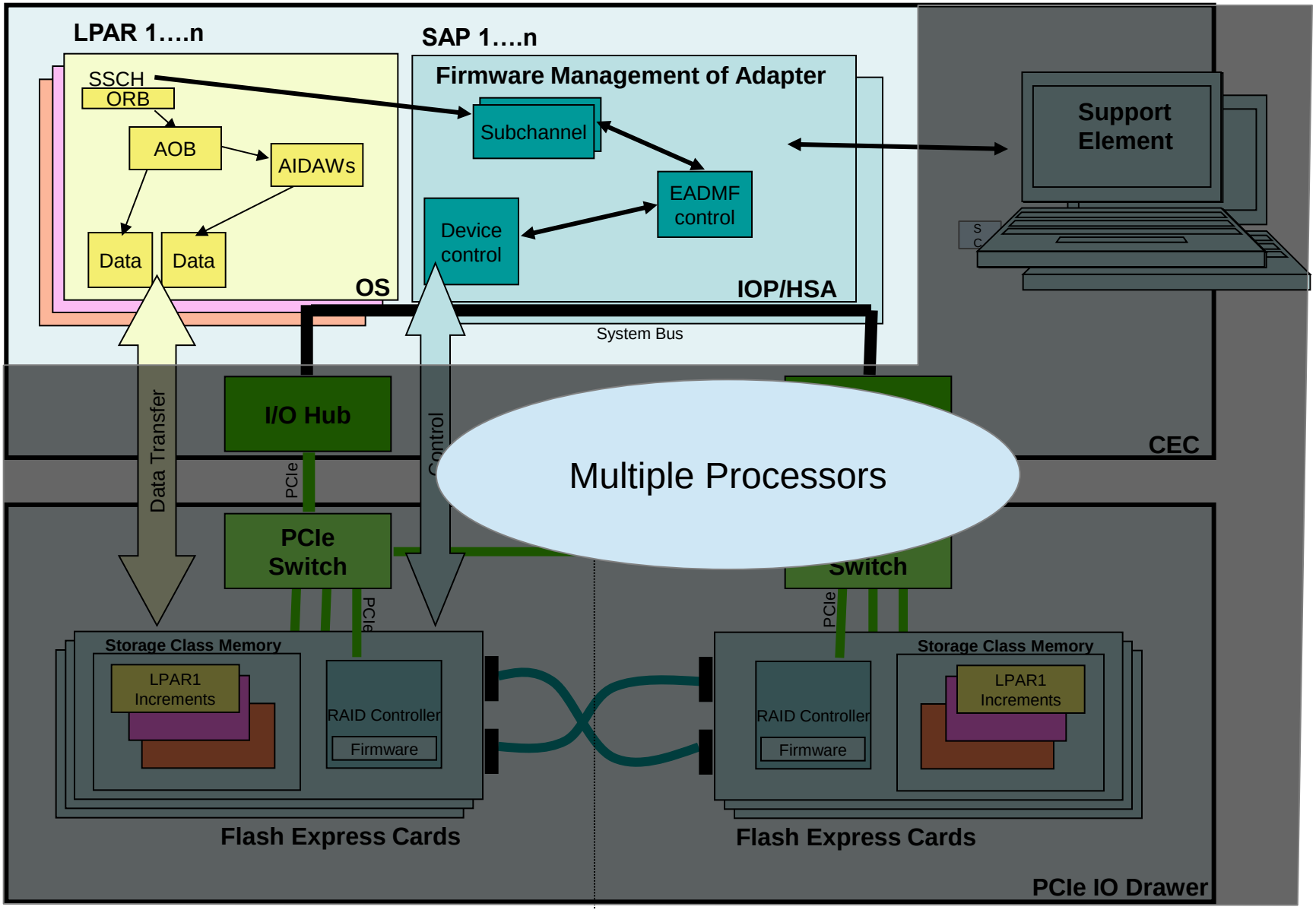
Flash Express Implementation

- ✓ System Overview
- ✓ Redundant Physical Structures
- ✓ Data Protection Mechanisms
- ✓ Data and Key Encryption
- ✓ Non-Disruptive Service Techniques

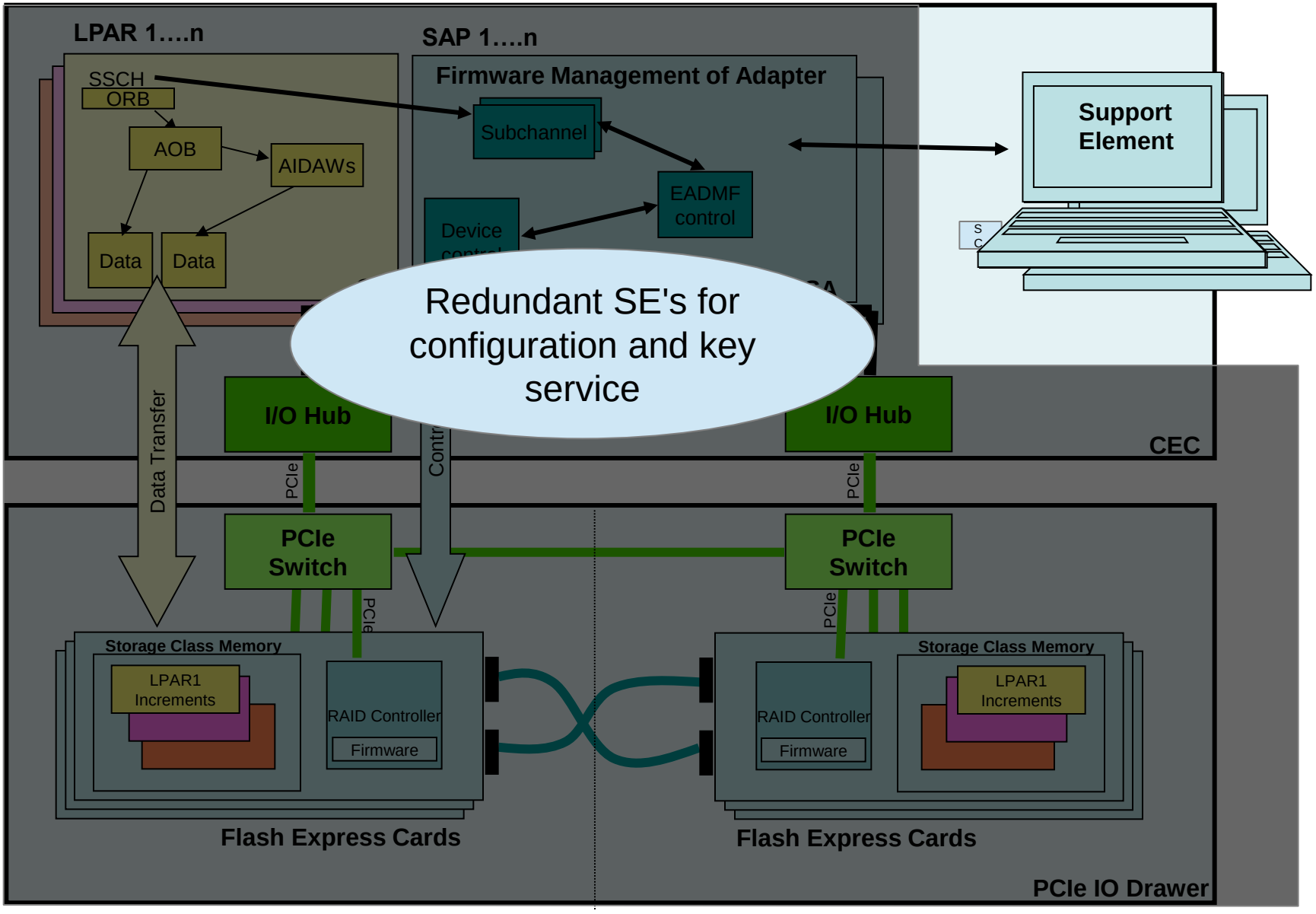
FLASH Express System Overview



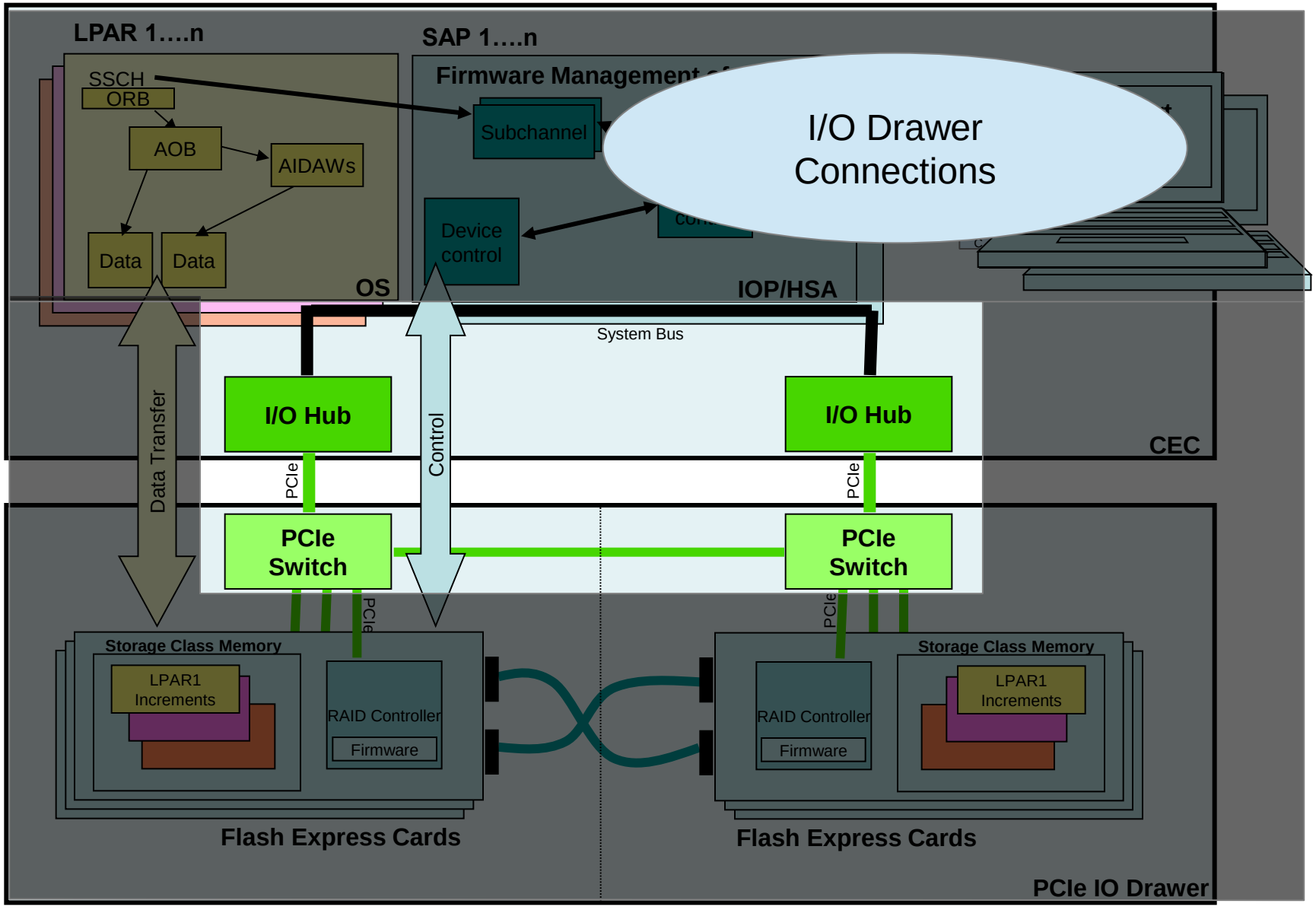
Redundant Physical Structures



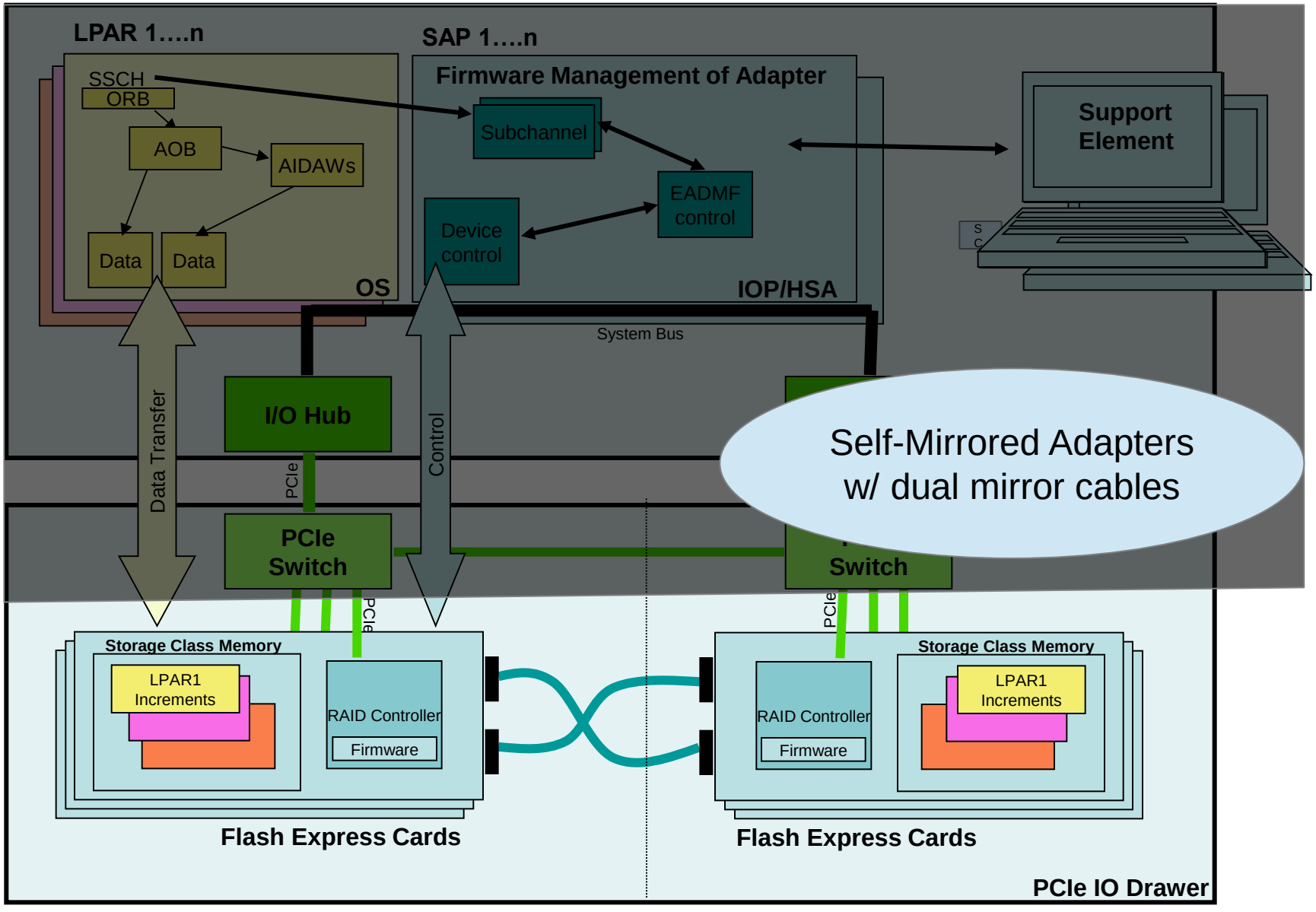
Redundant Physical Structures



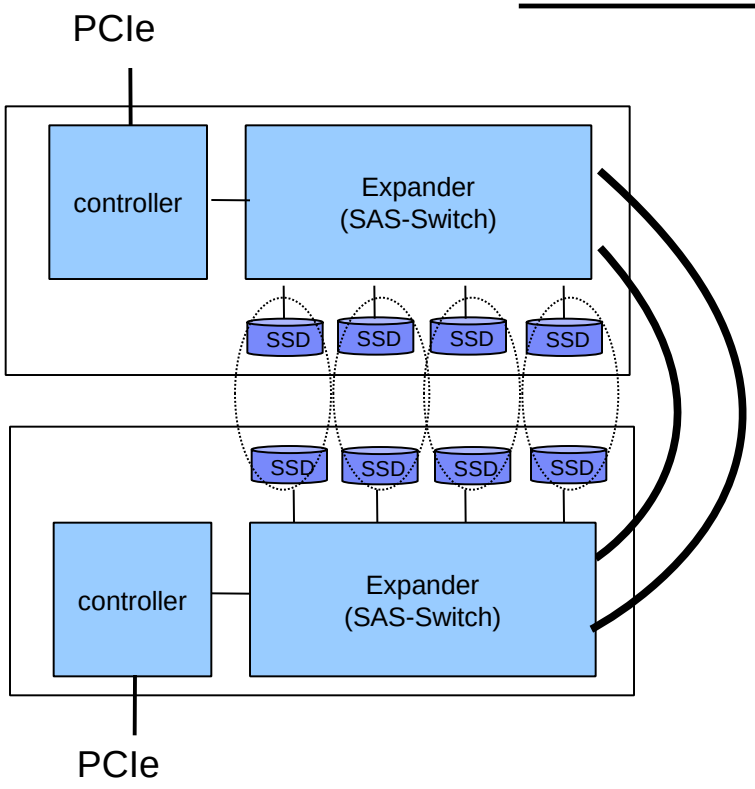
Redundant Physical Structures



Redundant Physical Structures



Data Protection Mechanisms

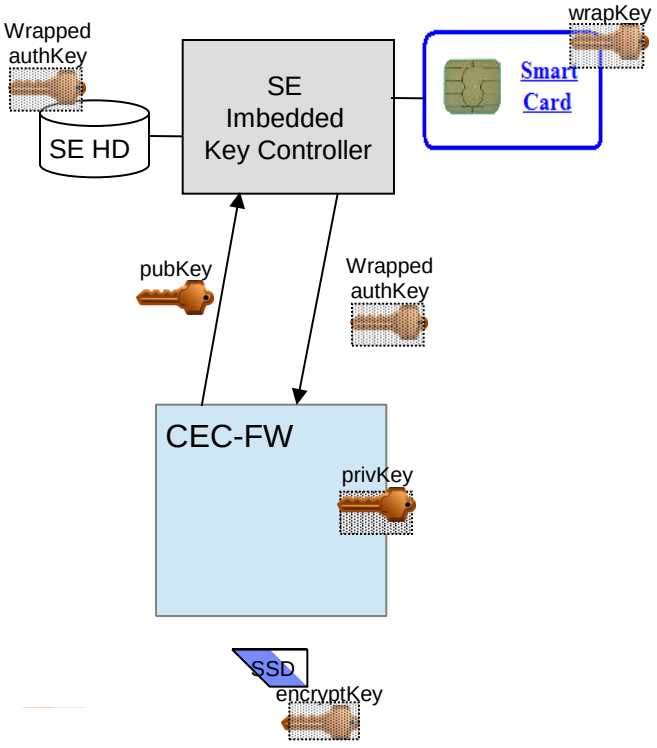


- ✓ RAID10: Protection and performance
  - RAID0 = Striping
  - RAID1 = Mirroring
  - RAID10 = Striped mirrorred data
- ✓ CRC and block seq. number stored on SSD
- ✓ Additional CRC around block transfer

- ✓ Comm links (SAS, PCIe) provide embedded protection & recovery
- ✓ CEC-based hardware address protection on communication from adapter
- ✓ ECC on internal system memory

Data and Key Encryption

- ✓ On SSD, data is protected with inline encryption (hidden encryptKey)
- ✓ Access to SSD is via authentication key (authKey) served from SE



During Flash install, in smart card on SE:

- Create authKey (aka PIN)
- Wrap authKey in an encrypted file
- wrapKey stored in smart card
- Wrapped key file stored on SE

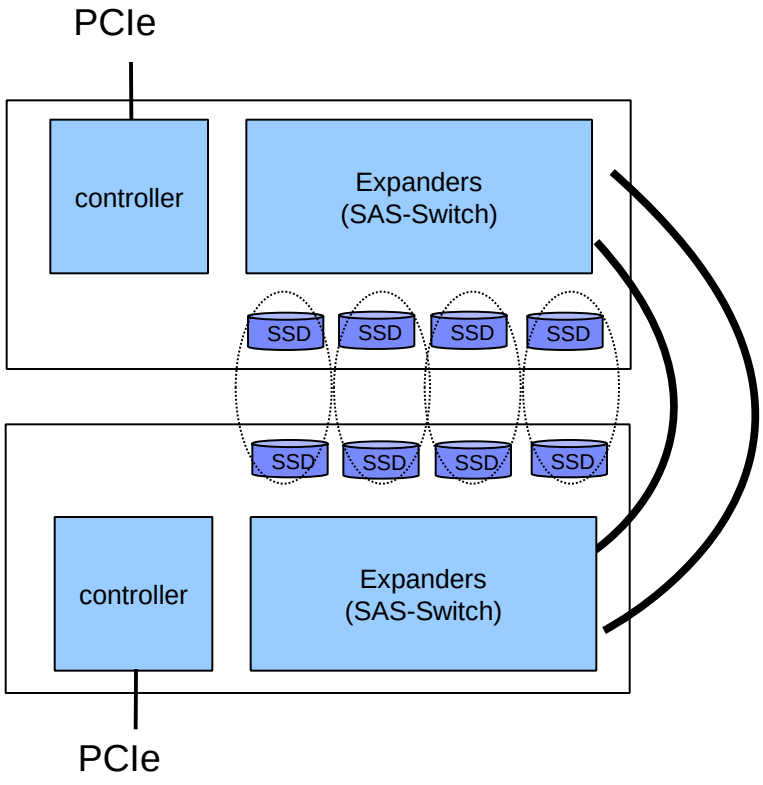
SE ⟶ CEC-FW authkey service:

- asymetric protocol – pub/private
- IOP sends public key to SE
- In smart card, Key file unwrapped then encrypted with CEC pubKey
- Encrypted authKey sent to CEC
- CEC 'unwraps' authKey using its privKey

- ✓ AuthKey used during SSD format and subsequent power cycles

Non-disruptive Service Strategy

- ✓Firmware updates
  - » Adapter
  - » SSD
  - » Expanders
  - » CEC-FW
- ✓Adapter replacement
- ✓Cable replacement
- ✓Recoveries
  - » CEC
  - » Adapter
  - » SSD



IBM Flash materials

Redbooks

SG24- 8049 - IBM zEnterprise System Connectivity Handbook (GRS ESCON / FICON CTCs, and FLASH Express, etc.)

SG24-5444 - IBM zEnterprise EC12 Technical Introduction (FLASH Express, and IBM zAware, etc.)

SG24- 8050 - IBM zEnterprise EC12 Technical Guide (FLASH Express, and IBM zAware, etc.)

Flash Express White Paper

http://public.dhe.ibm.com/common/ssi/ecm/en/zss03073usen/ZSS03073USEN.PDF

Flash Blogs

https://www-304.ibm.com/connections/blogs/systemz/entry/flashexpress?lang=en_us
https://www-304.ibm.com/connections/blogs/systemz/entry/flashexpress2?lang=en_us
https://www-304.ibm.com/connections/blogs/systemz/entry/under_the_covers_of_flash_express_implementation_highlights13?lang=en_us

Reference Documentation

- Available from “Books” group of Classic Style UI and the Welcome page of the Tree Style UI (& IBM Resource Link: Library->zEC12->Publications)
  - IBM SC28-6919: Hardware Management Console Operations Guide (Version 2.12.0)
  - IBM SC28-6920: Support Element Operations Guide (Version 2.12.0)
  - IBM SB10-7030: Application Programming Interfaces
  - IBM SC28-2605: Capacity on Demand User’s Guide
  - IBM SB10-7154: Common Information Model (CIM) Management Interfaces
  - IBM SB10-7156: PR/SM Planning Guide
  - IBM SA22-1088: System Overview
  - IBM SC27-2623 Advanced Workload Analysis Reporter (IBM zAware) Guide
- Available from IBM Resource Link: Library->zEC12->Technical Notes
  - System z Hardware Management Console Security
  - System z Hardware Management Console Broadband Remote Support Facility
  - System z Activation Profile Update and Processor Rules

System z Social Media Channels

- **Top Facebook pages related to System z:**
  - IBM System z
  - IBM Academic Initiative System z
  - IBM Master the Mainframe Contest
  - IBM Destination z
  - Millennial Mainframer
  - IBM Smarter Computing

- **Top LinkedIn groups related to System z:**
  - System z Advocates
  - SAP on System z
  - IBM Mainframe- Unofficial Group
  - IBM System z Events
  - Mainframe Experts Network
  - System z Linux
  - Enterprise Systems
  - Mainframe Security Gurus

- **Twitter profiles related to System z:**
  - IBM System z
  - IBM System z Events
  - IBM DB2 on System z
  - Millennial Mainframer
  - Destination z
  - IBM Smarter Computing

- **YouTube accounts related to System z:**
  - IBM System z
  - Destination z
  - IBM Smarter Computing

- **Top System z blogs to check out:**
  - Mainframe Insights
  - Smarter Computing
  - Millennial Mainframer
  - Mainframe & Hybrid Computing
  - The Mainframe Blog
  - Mainframe Watch Belgium
  - Mainframe Update
  - Enterprise Systems Media Blog
  - Dancing Dinosaur
  - DB2 for z/OS
  - IBM Destination z
  - DB2utor

THANK YOU

Backup Material

Registering for IBM Resource Link Access

- Registering for IBM Resource Link Access
- To view the documents on the Resource Link Web site. you need to register your IBM Registration ID (IBM ID) and password with Resource Link.
- To register:
  - Open the Resource Link sign-in page: http://www.ibm.com/servers/resourcelink/
  - You need an IBM ID to get access to Resource Link.
    - If you do not have an IBM ID and password, select the "Register for an IBM ID" link in the "Your IBM Registration" menu. Return to the Resource Link sign-in page after you get your IBM ID and password.
    - Note: If you’re an IBM employee, your IBM intranet ID is not an IBM ID.
  - Sign in with your IBM ID and password.
  - Follow the instructions on the subsequent page.

Reference Documentation

- Available from “Books” group of Classic Style UI and the Welcome page of the Tree Style UI (& IBM Resource Link: Library->zEC12->Publications)
  - IBM SC28-6919: Hardware Management Console Operations Guide (Version 2.12.0)
  - IBM SC28-6920: Support Element Operations Guide (Version 2.12.0)
  - IBM SB10-7030: Application Programming Interfaces
  - IBM SC28-2605: Capacity on Demand User’s Guide
  - IBM SB10-7154: Common Information Model (CIM) Management Interfaces
  - IBM SB10-7156: PR/SM Planning Guide
  - IBM SA22-1088: System Overview
  - IBM SC27-2623 Advanced Workload Analysis Reporter (IBM zAware) Guide
- Available from IBM Resource Link: Library->zEC12->Technical Notes
  - System z Hardware Management Console Security
  - System z Hardware Management Console Broadband Remote Support Facility
  - System z Activation Profile Update and Processor Rules

Trademarks

IBM, the IBM logo, and ibm.com® are trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at http://www.ibm.com/legal/copytrade.shtml.

Adobe is a registered trademark of Adobe Systems Incorporated in the United States, and/or other countries.

Linux is a registered trademark of Linux Torvalds in the United States, other countries, or both.

Microsoft and Windows are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Tom Mathias

IBM
mathiast@us.ibm.com

Elpida Tzortzatos

IBM
elpida@us.ibm.com

March 12, 2014 - Session 14726

Please fill out the online session evaluation at either:

SHARE.org/SanFranciscoEval, or

Aim your smartphone at this QR code: