System z Expo

*October 13 – 17, 2008 – Las Vegas, Nevada*

# Session Title: z/VSE V4.2 Technical Insights Part 2

## Session ID: zEG05


- Speaker Name: Ingolf Salm


Authorized IBM Training

10/01/2008

# Trademarks

**The following are trademarks of the International Business Machines Corporation in the United States, other countries, or both.**

Not all common law marks used by IBM are listed on this page. Failure of a mark to appear does not mean that IBM does not use the mark nor does it mean that the product is not actively marketed or is not significant within its relevant market.

Those trademarks followed by ® are registered trademarks of IBM in the United States; all others are trademarks or common law marks of IBM in the United States.

For a complete list of IBM Trademarks, see www.ibm.com/legal/copytrade.shtml:

*, AS/400®, e business(logo)®, DBE, ESCO, eServer, FICON, IBM®, IBM (logo)®, iSeries®, MVS, OS/390®, pSeries®, RS/6000®, S/30, VM/ESA®, VSE/ESA, WebSphere®, xSeries®, z/OS®, zSeries®, z/VM®, System i, System i5, System p, System p5, System x, System z, System z9®, BladeCenter®

**The following are trademarks or registered trademarks of other companies.**

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.
Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.
Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.
Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.
Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
UNIX is a registered trademark of The Open Group in the United States and other countries.
Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.
IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency, which is now part of the Office of Government Commerce.

* All other products may be trademarks or registered trademarks of their respective companies.

**Notes**:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput 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. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

# Agenda

- Turbo Dispatcher
- 64 bit implementation
- Capacity Measurement Tool
- SCSI support in z/VSE
- Tape encryption
- More tasks

# Turbo Dispatcher

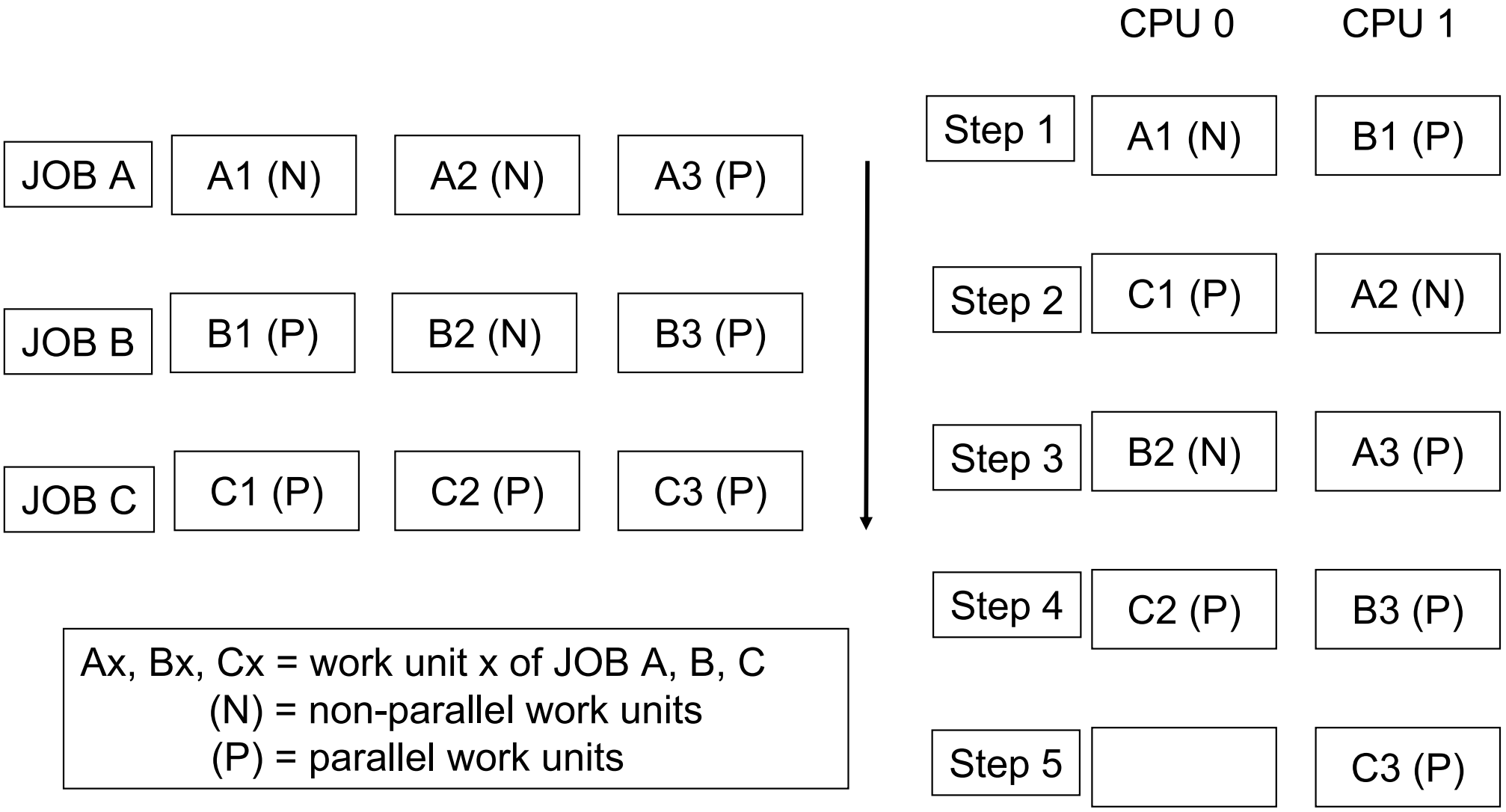
- Turbo Dispatcher history
  - Introduced in 1994, many enhancements since then.
  - Standard and Turbo Dispatcher until VSE/ESA 2.4.
  - TD the only dispatcher since VSE/ESA 2.4 (1999)

- Turbo Dispatcher Design
  - TD dynamically assigns partitions to CPUs
    - Assignment to one CPU lasts from dispatcher selection to next interrupt = work unit
    - If one task of a partition is active, no other task of the same partition can be selected
  - A partition (VSE/POWER job) processes many work units

# Turbo Dispatcher Design ...

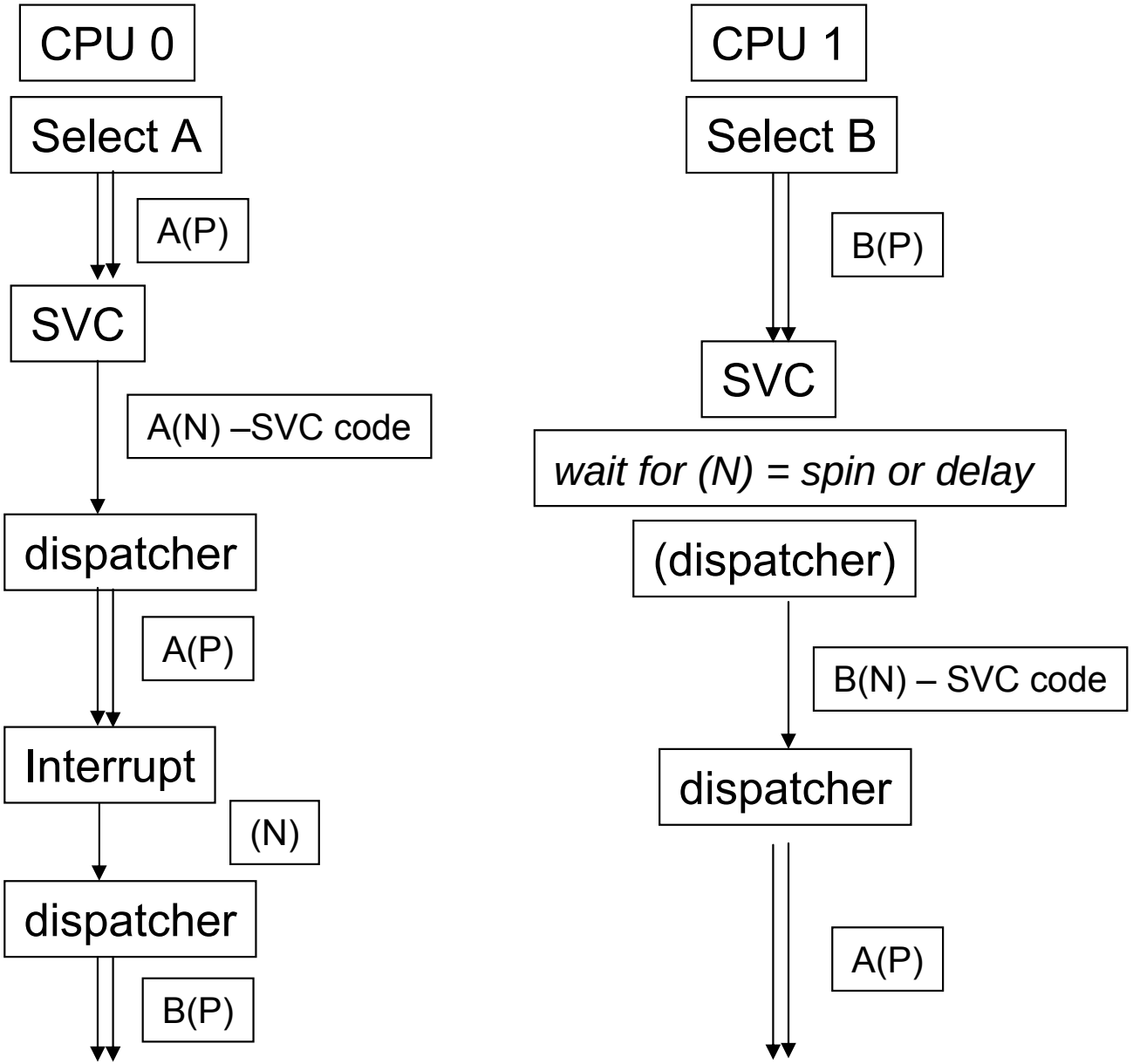
- Work units types:
  - parallel work unit
    - Application code (CICS/VSE, batch)
    - A parallel work unit may run on any CPU concurrently with other parallel or non-parallel work units.
  - non-parallel work unit
    - System code (services, ACF/VTAM)
      - As long as one non-parallel work unit is active on one CPU, no other non-parallel work unit can execute on any other CPU.
- VSE/POWER maintask has parallel or non-parallel work units

# Turbo Dispatcher Design …




Ax, Bx, Cx = work unit x of JOB A, B, C
(N) = non-parallel work units
(P) = parallel work units

# Turbo Dispatcher Design …



# Turbo Dispatcher Operation ...

- Retrieve CPU time values: QUERY TD

```
 CPU    STATUS     SPIN_TIME     NP_TIME TOTAL_TIME NP/TOT
  00    ACTIVE             0      237100     416698  0.568
  01    ACTIVE             0      157556     415229  0.379
  02    QUIESCED           0           0          0  *.***
  03    INACTIVE
                ---------------------------------------
TOTAL                      0      394656     831927  0.474

               NP/TOT: 0.474      SPIN/(SPIN+TOT): 0.000
 OVERALL UTILIZATION: 179%        NP UTILIZATION:  85%

 ELAPSED TIME SINCE LAST RESET:        463433
```

TOTAL_TIME = CPU time used by workload
NP_TIME = non-parallel CPU time, contained in TOTAL_TIME
SPIN_TIME = CPU time needed to wait for a non-parallel work unit
All above values given in milliseconds.

NP/TOT = ratio NP_TIME / TOTAL_TIME = non-parallel share
SPIN/(SPIN+TOT) = spin time ratio

# z/VSE 4.2: CPU Balancing

- When CPU balancing is activated,
  the z/VSE Turbo Dispatcher will only use CPUs required for the current workload
- Can be activated and deactivated via AR/JCL command
  - SYSDEF TD,INT=0 to deactivate, default
  - SYSDEF TD,INT=nn (=1..99) to activate and “nn” interval in seconds,
    after which the CPU utilization is inspected
- Threshold can be defined after which an additional CPU is activated
  - SYSDEF TD,THR=nn (10..99) in percent
- CPU balancing via stop or quiesce process
  - SYSDEF TD,INT=nn,STOP - the stop process to be used
    - May provide performance improvements for z/VM 5.4 guests
  - SYSDEF TD,INT=nn,STOPQ - the quiesce process to be use, default
- QUERY TD shows current settings
- CPU balancing may reduce multiprocessing overhead

# Performance Hints

- One partition can only exploit the power of a single CPU
- Use as many partitions as required for selected n-way
- Use/define only as many CPUs as really needed
- Full exploitation expected up to 3 CPUs
- Exploitation increases by reduction of non-parallel work units (e.g. by data in memory)
- Partition setup
  - Set up more batch and/or (independent) CICS partitions
  - Split CICS production partitions into multiple partitions (MRO)
  - Use a database (DB2)

# Non-Parallel Components

- A single CPU must be able to handle the non-parallel part of the total workload.
- Non-parallel code limits the maximum MP exploitation.
- QUERY TD command shows non-parallel share (NPS).
- System code (Key 0) code increases NPS.
  - Vendor code can have significant impact.
- TD searches for parallel work, when non-parallel resource is occupied.
- Overhead increases when NP code limits throughput.

# Limited Multiprocessor Benefits

- 'Largest' VSE partition requires more CPU power as available on a single CPU of the n-way
- VSE system limited by system resources other than CPU utilization, e.g. I/O, LTA, System GETVIS (24 bit), ...
- New bottleneck because of more capacity, would also appear on faster uni-processor
- Overall workload's non-parallel share too high
- Not enough partitions concurrently active

# CICS Implications

- Single CICS
  - Can consume processing power of one CPU only
  - If a CICS partition requires more CPU time than a single CPU can provide, the response time increases.
- Multiple CICS partitions (MP exploitation)
  - E.g. non-parallel share of 30 %
    - max. exploitable CPUs = 3
  - Multiple CICS workload alternatives
    - Independent CICS partitions
    - MRO transaction routing
    - MRO function shipping to file owning region
    - Mixtures of transaction routing and function shipping

# Performance Measurements

2 or 3 CPUs can be fully exploited

- Where non-parallel share ranges from 0.5 to 0.25

- Number of CPUs that can be exploited for a given workload:
  - number of CPUs = 0.9/non-parallel share
  - The value 0.9 is used here to take into account the delays caused by waiting for the non-parallel state.

- Measurements with our workloads
  - Batch workload (16 partitions):
    - TD overhead: 15 %, NPS: 0.48, MP factor (2-way): 1.4
  - Online workload, TD overhead 4%, NPS: 0.27
    - 2-way, 3xCICS: MP factor: 1.75, utilization: 93%
    - 3-way, 4xCICS: MP factor: 2.35, utilization: 84%
  - Online Workload, more I/O intensive, TD overhead 7%, NPS: 0.31
    - 2-way, 3xCICS: MP factor: 1.65, utilization: 77%
    - 3-way, 4xCICS: MP factor: 2.17, utilization: 82 %

# Turbo Dispatcher - Summary

- VSE workload can exploit up to 3 CPUs
- One partition can only exploit the power of one CPU
- A lower non-parallel share value will allow a better multiprocessor exploitation.
- Try to minimize the number of CPUs to run your workload
  - To reduce the multiprocessor overhead

# 64 bit real

- Processor storage > 2 GB, up to 8 GB
  - **z/VSE 4.2: up to 32 GB**
- Virtual address/data space size remains at max. 2 GB
- 64 bit virtual addressing not supported
- 64 bit addressing  mode not supported for applications or ISVs
- Implementation transparent to user applications
- Performance: 64 bit real can reduce / avoid paging
- In most cases the NOPDS option can be used

# 64 bit real – z/Architecture vs ESA/390 Architecture

| z/Architecture | ESA/390 Architecture |
|---|---|
| 24-bit and 31-bit addressing (up to 2GB)<br>64-bit addressing mode (more than 2GB) | 24-bit and 31-bit addressing (up to 2GB) |
| 16-byte PSW (64-bit instruction address) | 8-byte PSW (31-bit instruction address) |
| 8-byte general purpose registers | 4-byte general purpose registers |
| 8-byte control registers | 4-byte control registers |
| 4-byte access registers | 4-byte access registers |
| Prefix area is 8K (8K low core) | Prefix area is 4K (4K low core) |
| In Prefix area: changed locations of<br>New / old PSWs<br>Interrupt information like<br>page fault address<br>Store status save area .... | |
| Location of ESA/390 PSWs not used by HW | |

# z/Architecture - z/VM Display Examples

VM CP: **D PSWG**: PSW = 0404200**0** 8**0**000000 00000000 0000CEDA (31-bit)

**D PSWG**: PSW = 0404200**0** 0**0**000000 00000000 0000CEDA (24-bit)

**D PSWG**: PSW = 0404200**1** 8**0**000000 00000000 0000CEDA (64-bit)

**D GG8** : GRG 8 = FFFFFFFF 8001760C

**D G8** : GPR 8 = 8001760C

LA 41808000 0001760C (4-byte reg. Instruction)

**D GG8** : GRG 8 = FFFFFFFF 0001760C

ST 5080xxxx --- only low-order 4-byte are stored

D VTxxxx : 0001760C

**D XG1**: CRG 1 = 0000000003F3F00

**D X1**: ECR 1 = 03F3F00

# 64 bit real - Implementation

- IPL starts in ESA/390 mode and switches to z/Architecture mode during the IPL process
- Simulation of ESA/390 low core fields
- Only the z/VSE page manager has access to the area above 2GB
- Virtual pages can be backed by 64 bit real page frames
- PFIX or TFIX requests will use real page frames below 2 GB
- Page manager control blocks below 2GB
- **z/VSE 4.2: Page manager control blocks above 2 GB**
- 64-bit page frames used directly for page-in and page-out I/O

# 64 bit real – Implementation …

- Hardware uses z/Architecture new and old PSWs and interrupt locations for interrupts
  - Interrupts: external, SVC, I/O, machine check, program check
  - Interrupt processing; hardware stores old PSW and interrupt information and passes control to interrupt new PSW

- In z/VSE z/Architechture new PSWs point to emulation code
  - Prepares ESA/390 interrupt information
  - Pass control to z/VSE interrupt handlers
  - ESA/390 interrupt information is not used by hardware

- Task save areas are not extended, therefore only selected system routines can run in 64 bit mode or use 8 byte registers

# 64 bit real – ESA/390 Emulation

- In most cases system programs use ESA/390 locations
  - Such as ESA/390 old PSWs
  - Emulation guarantees that system code runs unchaged

- When an interrupt occurs, emulation code provides
  - Transalation of z/Architecture old PSW into ESA/390 old PSW
  - Setup of ESA/390 interrupt information
  - Continuation at ESA/390 new PSW address (z/VSE interrupt handler)

- Interrupt handlers/dispatcher work with ESA/390 information/locations

# ESA/390 Emulation – Program Check Example

Generated within Supervisor:
ESA/390 PC New PSW at 00000068: 000C0000 8000F142 (points to interrupt handler)
z/Arch PC New PSW at 000001D0: 00040000 80000000 00000000 0000F0B2
(points to emulation code)

Program check (page fault) occurs:
00000000000133B8 MVC D21F10009398 00506000
00000000000133B8 PROG 0011 -> 0000F0B2

Hardware sets:
z/Arch PC Old PSW at 00000150: 04040000 00000000 00000000 000133B8
z/Arch Transl. Excep. at 000000A8: 00000000 00506000 (page fault address)

Emulation code at F0B2 provides:
ESA/390 PC Old PSW at 00000028: 040C0000 000133B8
ESA/390 Transl. Excep. at 00000090: 00506000
Supervisor can continue at F142 (program check handler) as in ESA/390 mode

# Capacity Measurement Tool (CMT)

- Tool can be activated on z9 and z10 models
- z/Architecture mode required -> z/VSE 4.1 / 4.2 only
- z/VSE supported in LPAR and as z/VM guest
- Implementation
  - New system task
    - Will measure CPU usage and calculates MSUs
    - Measurement interval every 30 minutes
    - Calculation of the 4 hour rolling average
    - SMF like (SCRT89) records written to datasets
  - Datasets is input for the Sub-Capacity Reporting Tool (SCRT)
- Required for Midrange Workload License Charges (MWLC)
  - Sub-capacity option
- 13 z/VSE products participate in MWLC

# Capacity Monitoring Tool …

- CMT requires 3 (sequential BAM) disk files
  - One control file and 2 data files, size depends on configuration
- Once a month the input to the input to SCRT need to be prepared
- Sub-capacity Reporting Tool (SCRT)
  - SCRT with support for z/VSE 4.1 / 4.2
  - Analyzes SCRT89 records produced by CMT
  - SCRT Output is a report, similar to a spreadsheet report
  - Report to be send to IBM via web interface

# Capacity Monitoring Tool …

- Required steps for all z/VSE systems
  - Allocate/intitialize datasets for SCRT89 records
  - Update STDLabel procedure with DLBL and EXTENT info
  - Start capacity measurement
    - To be started after IPL complete with unique id
    - Update member USERBG.PROC with EXEC statement
    - Can be started through // EXEC IJBCMT,PARM='START ID=xxxx'
      - xxxx must be unique within the CPC
  - You may check via STATUS or SIR command if measurement is active
  - CMT can be stopped via // EXEC IJBCMT,PARM='STOP'
  - Use SCRT to analyze the measured data.
  - Send the final report to IBM once a month

# SCSI Support in z/VSE

- SCSI disks as emulated FBA disks on z/VM V5.2 or higher
  - z/VSE supports a max. size of 2 GB
- Direct attached SCSI disks
  - z/VSE supports up to 24 GB (VSAM: 16 GB)

# SCSI Support in z/VSE

- z/VSE supports SCSI disk devices only
- Impact on applications
  - Transparent to all VSE applications and subsystems,
    - Minimal impact on ISV system management tools
  - Reasons for transparency:
    - z/VSE's SCSI implementation is based on FBA support
    - Applications can not exploit SCSI commands directly
    - FBA to SCSI emulation on low level I/O interface

# SCSI Support in z/VSE

- Access SCSI devices through Fibre Channel Protocol (FCP)
  - Support available on System z processors
  - OS interfaces
    - Operating system communicates with FCP adapter
    - FCP adapter communicates with the SCSI device
  - SCSI disk devices utilize fixed block sectors
    - Therefore VSE treats them as FBA devices

# SCSI Support – Content / Limitations

- z/VSE's SCSI support includes:
  - SCSI for system and data device (SCSI only system)
  - Multipathing for fail-over
- SCSI support transparent to existing (I/O) APIs
- Block size restricted to 512 bytes,
  even if the SCSI device can be configured with larger block sizes
- Max. SCSI disk size about 24 GB, VSAM 16 GB
- FSU from SCSI to SCSI device only

# SCSI Support - Configuration

- New IPL / JCL commands and dialog to define and query a SCSI device
- Required steps to get a SCSI device known to z/VSE
  - Device configuration
  - Switch configuration
    - In case of point to point connections (System z9) no longer necessary
  - FCP Adapter to be configured in IOCDS (CHIPID type FCP)
  - FCP adapter and SCSI disk to be defined in VSE via
    - IPL ADD commands to define FCP and FBA device
    - IPL DEF or JCL SYSDEF command to define connection to LUN

# SCSI Configuration in z/VSE (Example)

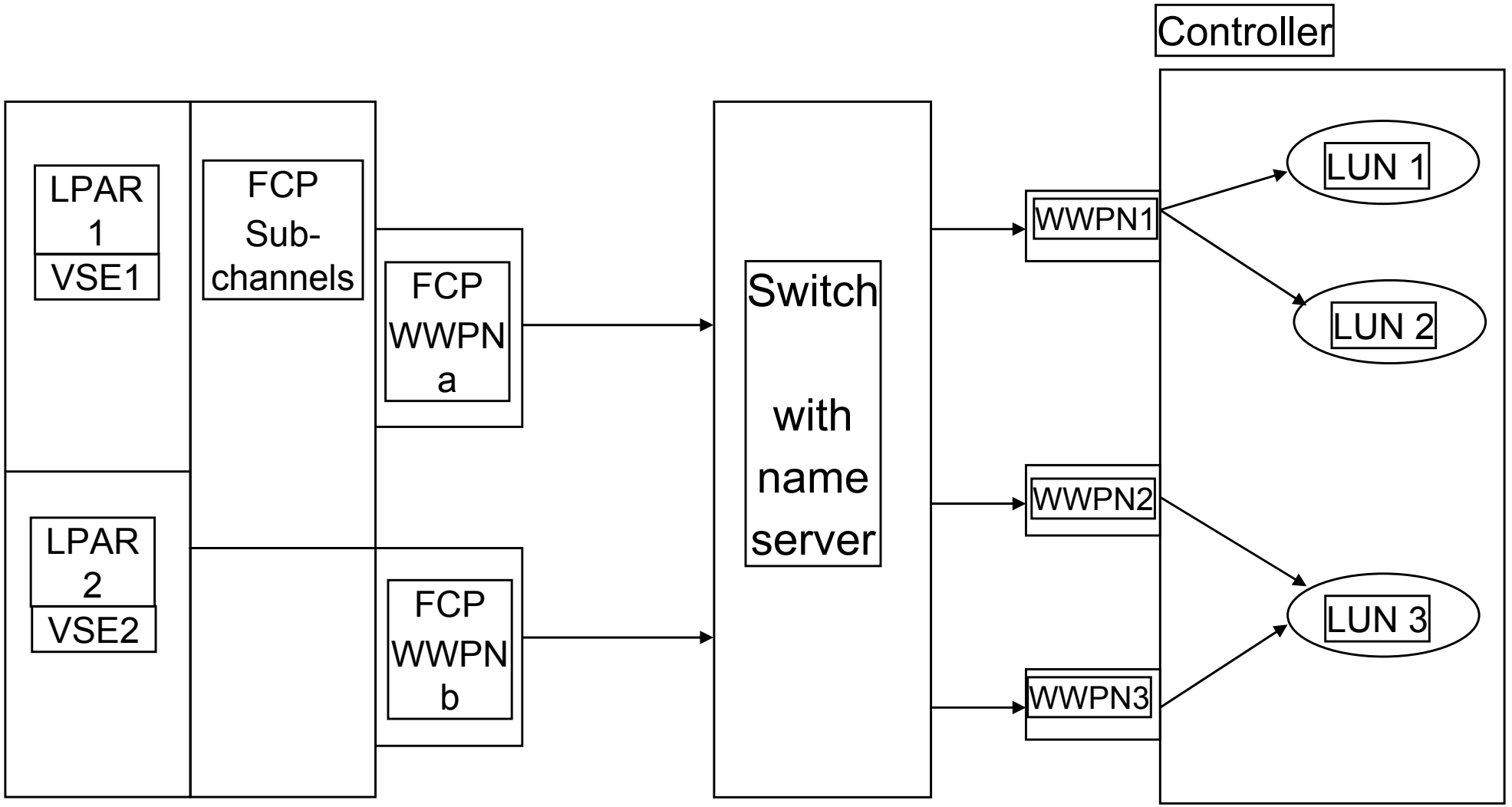
- Define FCP Devices, SCSI Disks and Connection Paths to z/VSE
  - FCP Devices
    - ADD C00,FCP (ADD C00:C0F,FCP)
    - ADD D00,FCP (ADD D00:D0F,FCP)
  - FBA Devices:
    - ADD 700:701,FBA
    - Note: these devices must not exist in the IOCP or under VM
  - Define a Connection Path (IPL)
    - DEF SCSI,FBA=700,FCP=C00,WWPN=5005076300CA9A76,LUN=5600
    - DEF SCSI,FBA=701,FCP=C00,WWPN=5005076300CA9A76,LUN=5601

      Only one FCP cuu required to access the LUNs
  - Define a Connection Path (after IPL)
    - SYSDEF SCSI,FBA=702,FCP=C00,WWPN=5005076300CA9A76,LUN=5602
    - Note: The FBA and FCP devices added during IPL.
  - IUI Dialogs are available to configure SCSI Devices

# SCSI Support - Configuration

- System z FCP adapter supports switched network (z/VSE 3.1), Point-to-point connection with z/VSE 4.1/4.2 and z9 BC, z9 EC, z10 EC:
  - Each FCP adapter has an associated port (WWPN)
  - FCP adapter configured in IOCDS with subchannel type FCP
  - FCP adapter connects to a switch
  - Switch connects to a controller with one or multiple ports
  - Controller accesses one or more SCSI devices (LUNs)
- **z/VSE 4.2: SAN Volume Controller (SVC) support**

# SCSI Support - Configuration



# SCSI Support – System z9 / z10 Exploitation

- N_Port ID Virtualization (NPIV) for (CHPID type) FCP channels
  - Multiple virtual FCP channels can be defined each with its own unique Fibre Channel port name and FC N_Port ID
    - Each FCP device (ADD device,FCP) has its own portname
  - NPIV allows sharing the Lock file on SCSI between multiple z/VSE systems using the same physical FCP adapter (CHPID)
    - DEF SCSI,FBA=600,FCP=C00,WWPN=5005076300CA9A76,LUN=5750 (VSE1)
    - DEF SCSI,FBA=600,FCP=C01,WWPN=5005076300CA9A76,LUN=5750 (VSE2)
      - 600 is the lock file disk. With NPIV, C00 and C01 can be on same FCP CHPID
  - To use NPIV, the Fibre Channel switch must support NPIV
  - Without NPIV,
    - Each FCP channel(device) has the portname of the FCP CHPID
    - Each z/VSE needs its own physical FCP adapter to access the lock file
- FCP point-to-point attachments
  - FCP feature can directly attach to storage devices. No switch required.

# Data Encryption (z/VSE 4.1 + PTF, z/VSE 4.2)

- IBM TS1130 Tape Drive with encryption feature
  - Supported by z/VSE 3.1, z/VSE 4.1, z/VSE 4.2
  - Supports data encryption within the drive itself
  - Using Systems Managed Encryption with the TS1130
  - z/VSE support will require the Encryption Key Manager (EKM) component running on another operating system other than z/VSE using an out-of-band connection.
    - Generation and communication of encryption keys for tape drive
    - TCP/IP connection between EKM and the tape controller
  - Data encryption is transparent to z/VSE applications
  - Data encryption
    - Data will be encrypted and compressed, when specified
    - Default: encryption disabled
  - **z/VSE 4.2: encryption re-keying support to encrypt data key of encrypted tape cartridge**
  - More details on z/VSE home page

# Data Encryption …

- Encryption Key Manager (EKM)
  - EKM is a Java application, used to generate and protect AES keys
  - On request EKM generates AES (256 bit) data keys and protects those keys
  - Key encryption key label (KEKL) identifies the encryption keys
  - The KEKL or the hash value of the public key can be stored on the cardridge.

  - You may download EKM from the internet

# Data Encryption …

- In z/VSE jobs must have an ASSGN statement and KEKL statement to access or write encrypted data
- ASSGN statement
  - ASSGN SYSnnn,cuu,mode
    - cuu = device address
    - mode =
      - 03 encryption wirte mode
      - 0B encryption and IDRC write mode
      - 23 encryption and unbuffered (compression) write mode
      - 2B encryption and IDRC and unbuffered write mode
- KEKL statement
  - // KEKL UNIT=cuu,KEKL1=key_label_1,KEM={L|H}
    - KEM = key encoding mechanism
      - L = label, H = public key hash

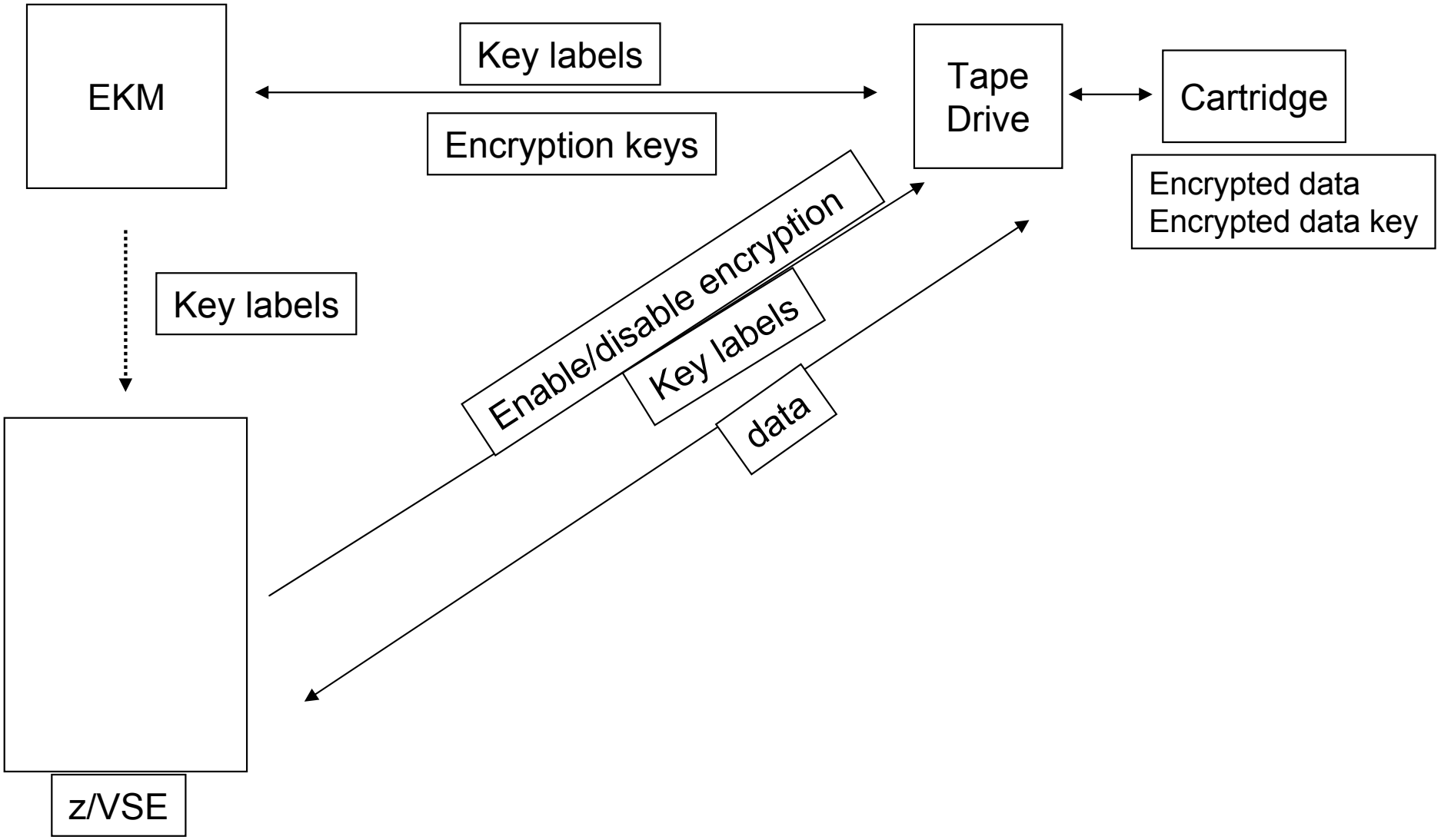
# Data Encryption …

- Write encryption data example
  - // JOB ENCRYPT
  - // ASSGN SYS005,480,3
  - // KEKL UNIT=480, KEKL1=,HUSKEK1‘,KEM1=L
  - // EXEC LIBR
    BACKUP LIB=PRD2 TAPE=SYS005
  - /*
  - /&
- Read encrypted data
  - No need to specify the ASSGN mode or KEKL
  - The control unit recognizes the encrypted tape and tries a key exchange with the EKM and the KEKL saved in cardridge memory.

# Data Encryption …

- Steps to encryption
  - 1. Load cartridge
  - 2. EKM to tape drive: specify encryption, provide key labels
  - 3. Tape drive requests data key from EKM
  - 4. EKM generates key and encrypts with public and session keys
  - 5. EKM to tape drive: Encrypted keys transmitted
  - 6. Tape drive writes encrypted data and stores encrypted data key on cartridge
- Implementation in z/VSE
  - VSE JCL enhancements
    - For encryption setting (via ASSGN)
    - Key Encryption Key Label (KEKL) may be specified
  - I/O Supervisor
    - retrieves encryption information, activates encryption and transfers KEKL

# Data Encryption …



# z/VSE 4.2: up to 512 tasks

- Up to 512 VSE tasks, still 32 VSE tasks per partition
- Task id (TID = 2 byte field) in SYSCOM and other control blocks
  - Old tasks (up to 255) = 1st byte zero, 2nd byte holds task id
    - Highest task id X'00FF'
  - New tasks: X'0100' .. X'01FF'
  - System and maintasks will always receive old task ids

# z/VSE 4.2: up to 512 tasks …

- No IPL option required
- System option (SYSDEF) to set max. number of tasks and defaults
  - SYSDEF SYSTEM,NTASKS=(nnn|MAX),TASKS=(ANY|OLD)
- EXEC parameter for compatibility mode
  - // EXEC phase,TASKS=(ANY|OLD)
- MAP/QUERY / SIR to show more task details
  - Display settings via QUERY command / MAP command

# More Information

- **… on VSE home page:**

**http://ibm.com/vse**