z/VSE Performance Update

Ingo Franzki

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.

Notice Regarding Specialty Engines (e.g., zIIPs, zAAPs and IFLs):

- Any information contained in this document regarding Specialty Engines ("SEs") and SE eligible workloads provides only general descriptions of the types and portions of workloads that are eligible for execution on Specialty Engines (e.g., zIIPs, zAAPs, and IFLs). IBM authorizes customers to use IBM SE only to execute the processing of Eligible Workloads of specific Programs expressly authorized by IBM as specified in the “Authorized Use Table for IBM Machines” provided at http://www.ibm.com/systems/support/machine_warranties/machine_code/aut.html (“AUT”).
- No other workload processing is authorized for execution on an SE.
- IBM offers SEs at a lower price than General Processors/Central Processors because customers are authorized to use SEs only to process certain types and/or amounts of workloads as specified by IBM in the AUT.

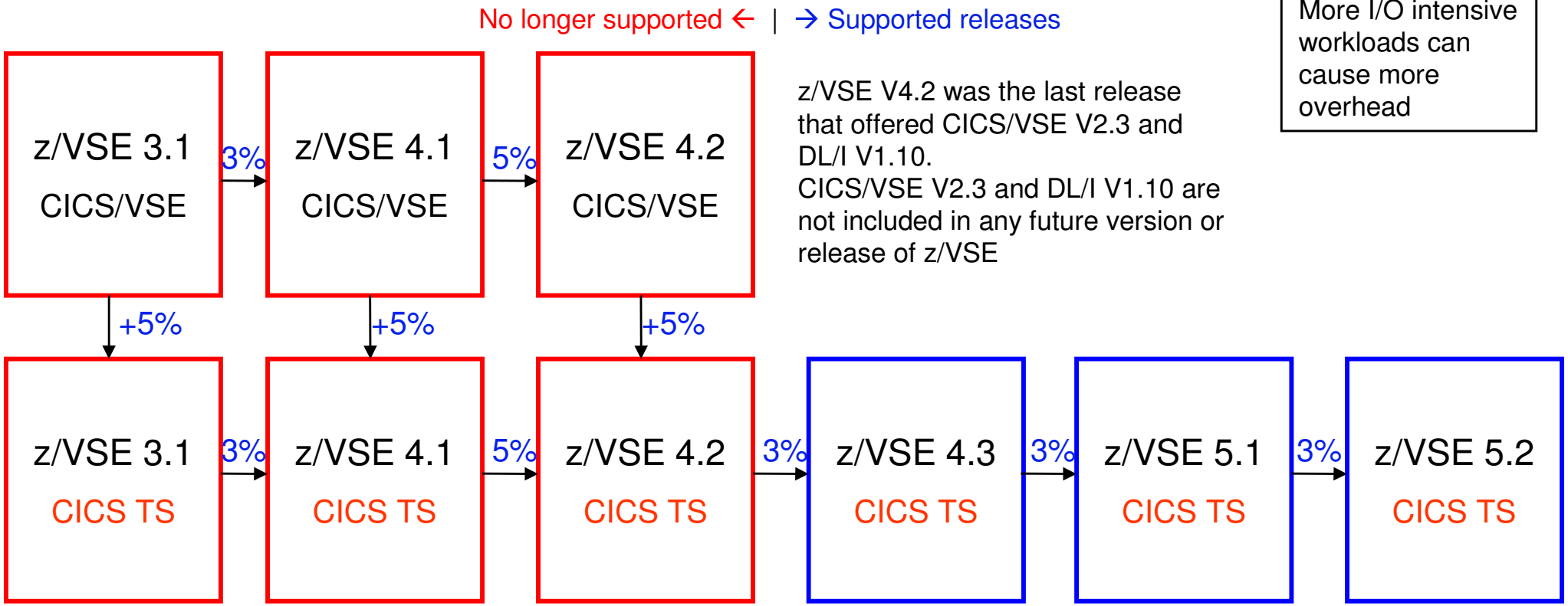
Overhead Deltas for VSE Releases

Remember that you get a lot of new functions that in most cases helps you to increase VSE system performance and throughput:

Partition Balancing, PRTY SHARE (Turbo Dispatcher), FlashCopy, Buffer Hashing, Shared data Tables (CICS TS), NOPDS with larger VSIZE

These numbers are for **a specific average I/O intensive** workload (PACEX16)

More I/O intensive workloads can cause more overhead



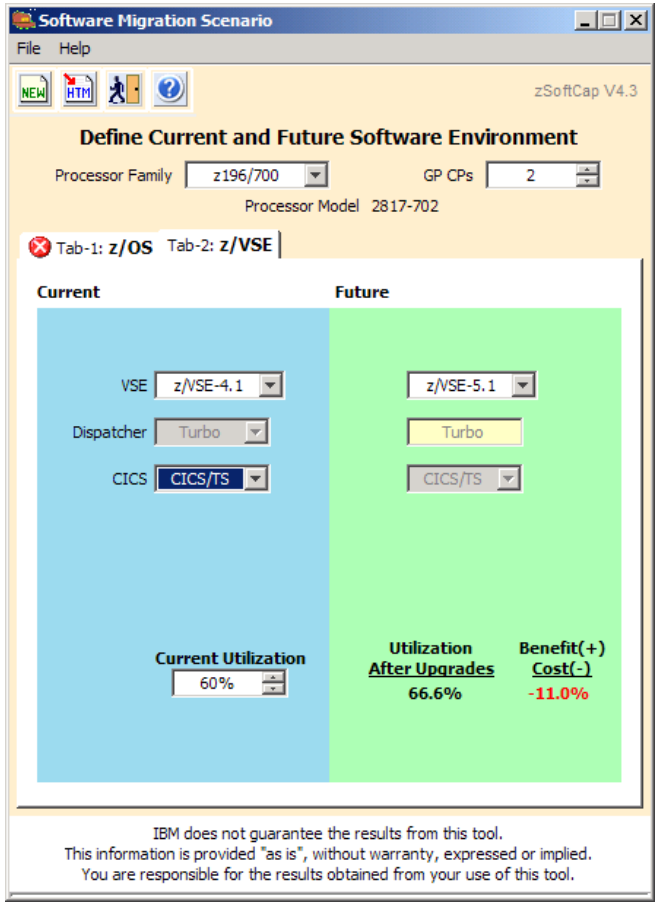
z/VSE V4.2 was the last release that offered CICS/VSE V2.3 and DL/I V1.10.
CICS/VSE V2.3 and DL/I V1.10 are not included in any future version or release of z/VSE

You can also use the zSoftCap tool to determine release migration overhead:
http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/PRS268

New: zSoftCap Tool

You can use the zSoftCap tool to determine release migration overhead:
http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/PRS268

zSoftCap is a PC-based productivity tool designed to assess the effect on capacity for IBM System z processors, when migrating to more current releases of the operating system or major subsystems. zSoftCap assumes that hardware remains constant while software releases change.



What is Performance ?

→ Performance is about:

–How fast does it run

- Job Duration
- CPU seconds
- Throughput
- Response times

–How much resources does it use

- Memory
- I/Os

–Why doesn‘t it run faster?

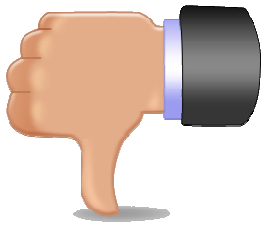
- Where is the bottleneck ?

Performance is all about comparing

- **Absolute values do not tell you much**

  Examples:
  - Job A runs 4 minutes and 10 seconds
  - Program B requires 5 MB of memory

  → Is that good? Is that bad ?



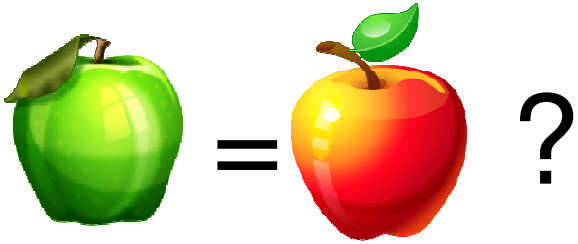
- **Comparison tells you if its good or bad**

  Examples:

  - After migrating to z/VSE X.Y to z/VSE X.Z, job A now runs 4:10 versus 3:40
    → 13% increase
  - On version X Program B now needs 5 MB, on version Y it did only need 4 MB
    → 25 % increase

Performance is all about comparing (continued)

- **When comparing, make sure you compare apples to apples**
- **Comparing A to B is only valid if**
  - Environment is the same
    - Storage layout, Sizes, Priorities, ...
  - Workload is the same
    - Same amount of data processed, same number of requests, …

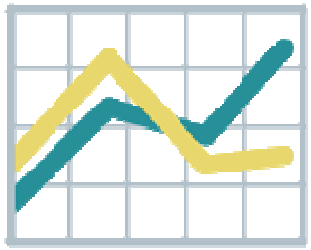


- **Little changes can cause big differences !**
- **True performance comparisons can only be done in a strictly isolated (clinic) environment**
  - Measurements in a production environment may not produce usable results
  - Results may be influenced by many different things
    - Concurrent users
    - Other work running in parallel (batch jobs)
    - Other work running on shared processors in other LPAR or z/VM guest

Performance Monitoring

- **Performance Monitor Tools**
  - Periodically gather values of certain counters (sampling)
  - Gather values of certain counters at special events (e.g. Job Accounting - At end of Job Step)

- **Data to be monitored:**
  - CPU usage (percent, CPU seconds)
  - I/O times and rate
  - Memory usage
  - Transaction rate
  - …



- **Real time monitoring**
  - Displays how the counters are NOW
  - Does not help you much to solve a problem if you are not looking at the screen at the time the problem occurs

- **History data**
  - Allows you to look at the samples over time
  - You can find out what happened when the problem occurred by looking back in history
  - You can draw charts to analyze the data

Performance Monitor tools

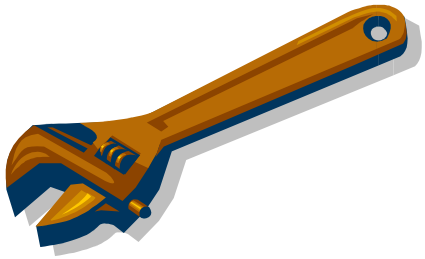
Commercial products provided by ISVs:

- TMON (ASG)
- Explore (CA)

- **Built into z/VSE**
  - Display System Activity Dialog (361/362)
  - QUERY TD
  - SIR SMF, SIR MON
  - Job Accounting
  - SNMP Monitoring Agent
  - CICS Statistics

- **Free tools**
  - CPUMON Tool

- **z/VM & LPAR**
  - z/VM Performance Toolkit
  - HMC provides very basic monitoring capabilities



QUERY TD command

- Usage:
  - SYSDEF TD,RES… eset the counters
  - // run the workloa…
  - QUERY TD,INTE… displa…

SPIN_TIME
spin CPU time waiting for a resource occupied by another CPU, not contained in TOTAL_TIME

NP_TIME
non-parallel CPU time, contained in TOTAL_TIME

TOTAL_TIME
total CPU time

```
AR 0015  CPU    STATUS     SPIN_TIME     NP_TIME  TOTAL_TIME  NP/TOT   DISP_ENTR
AR 0015   00    ACTIVE             0       76670      121778   0.629    13163221
AR 0015                   ----------------------------------------------------
AR 0015 TOTAL                      0       76670      121778   0.629    13163221
AR 0015
AR 0015                  NP/TOT:  0.629       SPIN/(SPIN+TOT):  0.000
AR 0015  OVERALL UTILIZATION:     %          NP UTILIZATION:     0%
AR 0015
AR 0015  ELAPSED TIME SINCE I        SET:     166387830
AR 0015  NUMBER OF SVC
AR 0015 1I40I  READY
```

Non-Parallel-Share (NPS) = NP / TOT

Maximum number or exploitable CPUs can be calculated as follows:

Max expl. CPUs = 0.8 / NPS

QUEUED:
Average time that an I/O request was queued in z/VSE (I/O supervisor).

If the PENDING time is not given explicitly then the QUEUED time does also include the time the request was PENDING in the Channel Subsystem.

An I/O request is queued in z/VSE from the time the I/O request is enqueued up to the point where the Start I/O operation (SSCH) has successfully been initiated. Starting with the successful initiation of the SSCH instruction, the time will be counted as PENDING in Channel Subsystem. A request would be held pending in the Channel Subsystem if e.g. a channel or a Control Unit (CU) is currently busy.

TOTAL
Average time of a complete I/O operation and is actually the sum of QUEUED (+PENDING), CONNECT and DISCONN (+DEV.BUSY).

An excessive value in this field could be the indication of a device problem.

```
AR 0015 DEVICE   I/O-CNT      QUEUED      CONNECT      DISCONN      TOTAL
AR 0015                      msec/SSCH    msec/SSCH    msec/SSCH    msec/SSCH
AR 0015
AR 0015   150          107        0.2          0.406          705        1.349
AR 0015   151          136          3          0.327           01        0.792
AR 0015   152                                  0.3
AR 0015 1I40
```

CONNECT:
Average time that a device is logically connected to a channel for purposes of transferring information between it and the Channel Subsystem.

DISCONN
Average time that a device is logically disconnected from the Channel Subsystem while the device is still busy and has not yet presented primary interrupt (Channel End) status.

In case a DEV.BUSY time is not outlined explicitly, then the DISCONN time does also include the DEV.BUSY time which is the time between the primary status (CE) and the secondary device-end status (DE).

– SIR SMF, Queue

```
AR 0015 TIMING VALUES FOR 200 BASED ON       20
AR 0015 MAXIMUM I/O QUEUE   2
AR 0015
AR 0015   QUEUED      PENDING      CONNECT
AR 0015  msec/SSCH   msec/SSCH    msec/SSCH    n                    CH
AR 0015      0.001        0.000        0.604        0.000        0.000        0.606
```

SIR MON command

- Usage:
  - SIR MON=ON ← enable MON
  - SYSDEF TD,RESETCNT ← reset the counters
  - // run the workload
  - SIR MON ← display the counters
  - SIR MON=OFF ← disable MON

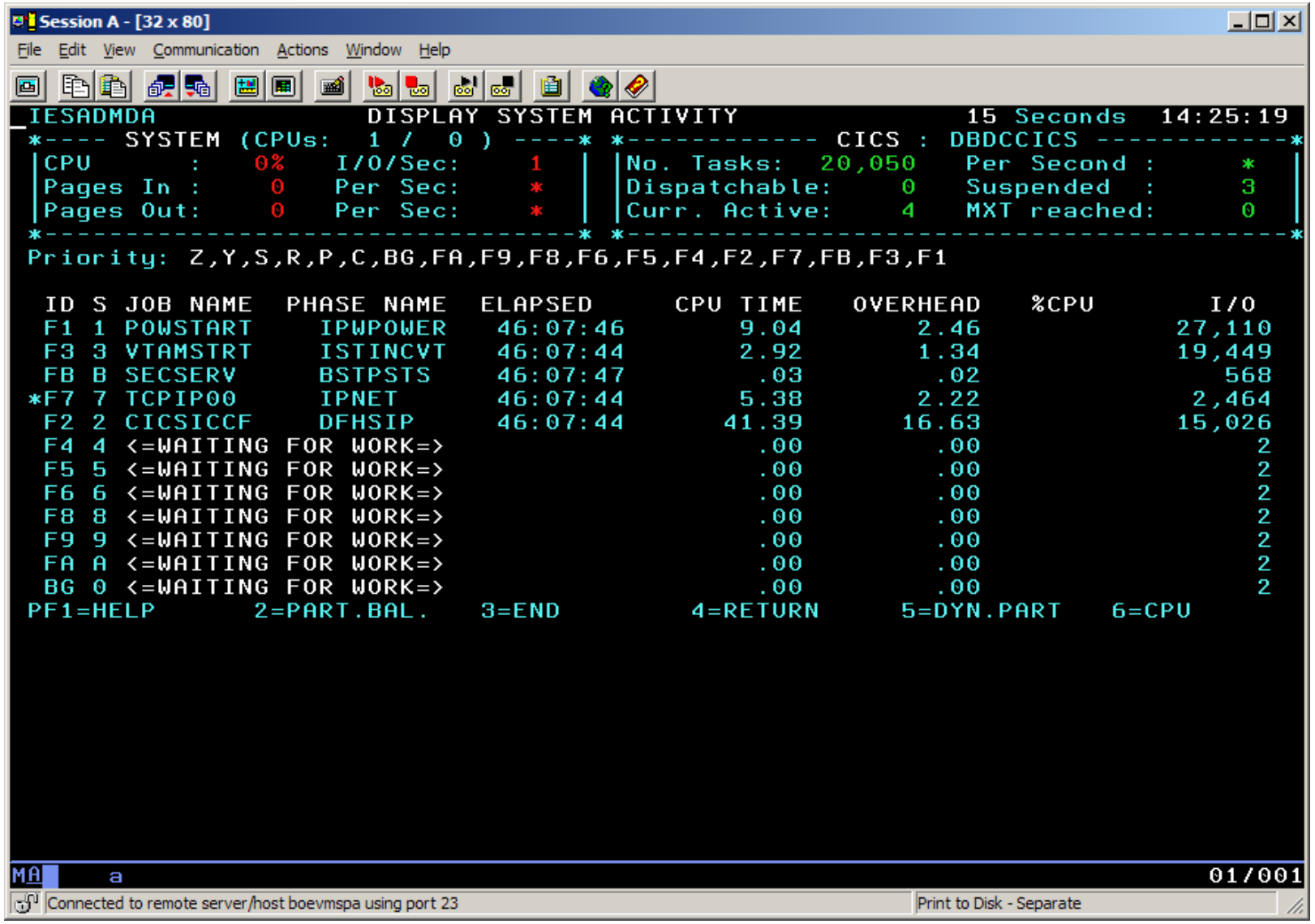
```
AR 0015                        MONITORING REPORT
AR 0015                  (BASED ON A 0000:00:13.338 INTERVAL)
AR 0015                          SVC SUMMARY REPORT
AR 0015 EXCP      =         195  FCH-$$B   =          14  SVC-03    =           1
AR 0015 LOAD      =         138  WAIT      =         405  SETIME    =          22
AR 0015 SVC-0B    =           5  SVC-0C    =         231  SVC-0D    =         236
AR 0015 EOJ       =           2  SYSIO     =         118  EXIT IT   =          28
AR 0015 SETIME    =          15  SVC-1A    =           4  WAITM     =          25
AR 0015 COMREG    =        1125  GETIME    =          25  FREE      =           1
AR 0015 POST      =         122  DYNCLASS  =           2  SVC-31    =          53
AR 0015 HIPROG    =           1  TTIMER    =           3  SVC-35    =         487
AR 0015 INVPART   =           2  GETVIS    =         708  FREEVIS   =         626
AR 0015 CDLOAD    =          11  RUNMODE   =           1  REALAD    =           1
AR 0015 SECTVAL   =         137  SETLIMIT  =           5  SVC-5B    =           1
AR 0015 XECBTAB   =           1  EXTRACT   =           7  GETVCE    =          30
AR 0015 EXTENT    =           2  SUBSID    =           1  FASTSVC   =        2992
...
```

Job Accounting

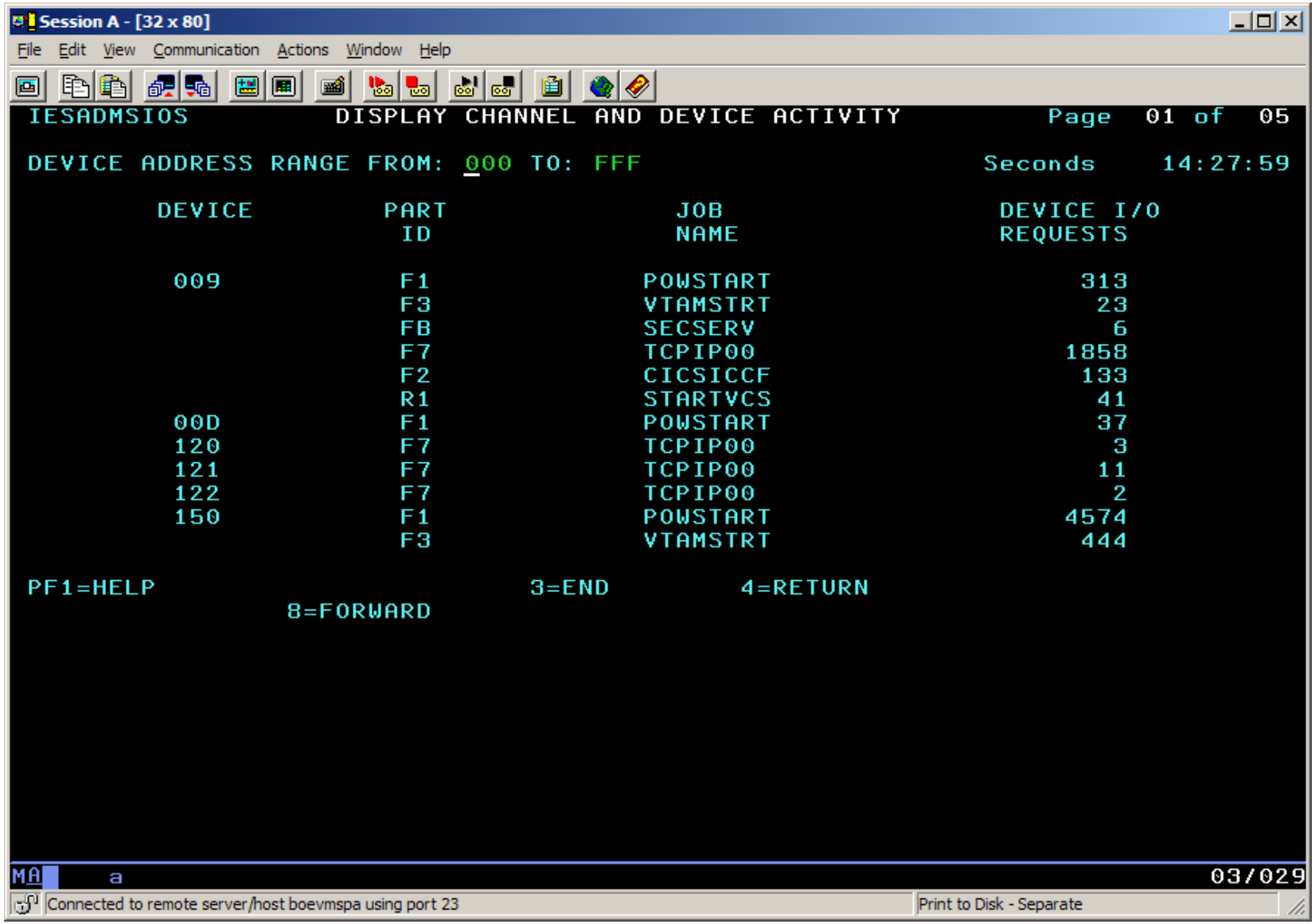
- Use skeleton SKJOBACC in ICCF Library 59
  to assemble Job Accounting routine $JOBACCT
- Prints info about CPU usage and I/Os after every job step

```
JOBNAME    = PRINTLOG  USER INFO = PR       EXEC NAME = PRINTLOG
DATE       = 11/05/99  PART ID   = BG
START      = 10:56:23  STOP      = 10:56:28 DURATION  = 5.560 SEC
CPU        =     0.060 SEC      PAGEIN SINCE IPL   = 0
OVERHEAD   =     0.017 SEC      PAGEOUT SINCE IPL  = 0
TOTAL CPU =      0.077 SEC
UNIT = E15       UNIT = FEC       UNIT = 01F      UNIT = E16
SIO  = 26        SIO  = 5         SIO  = 5        SIO  = 105
UNIT = FEE
SIO  = 4083
```

Display System Activity Dialog (361)



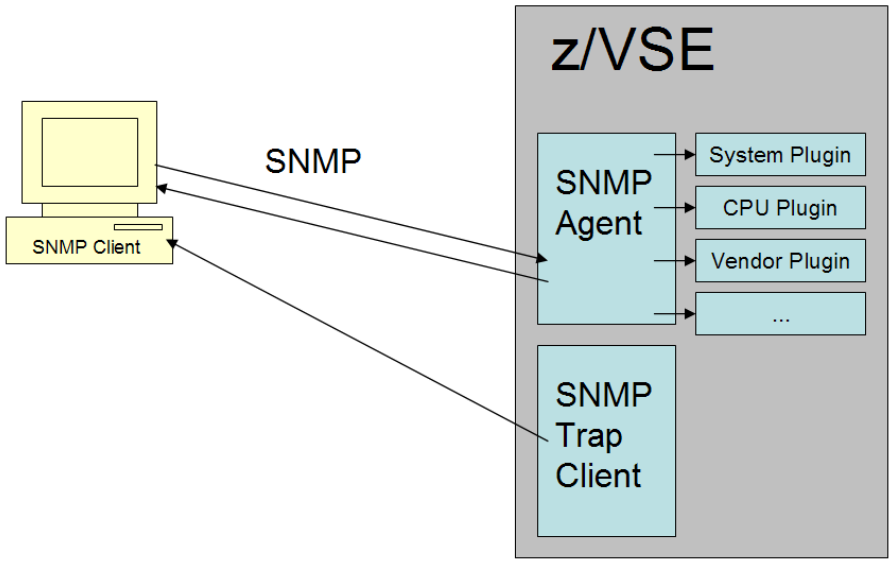
Display Channel and Device Activity (362)



z/VSE SNMP Monitoring Agent support

- **z/VSE Monitoring Agent enables customers to monitor z/VSE systems using standard monitoring interfaces (SNMP V1)**
  - Available since z/VSE V4.3
  - It also includes an open interface, which enables customers or vendors to use own programs (plugins) to collect additional data

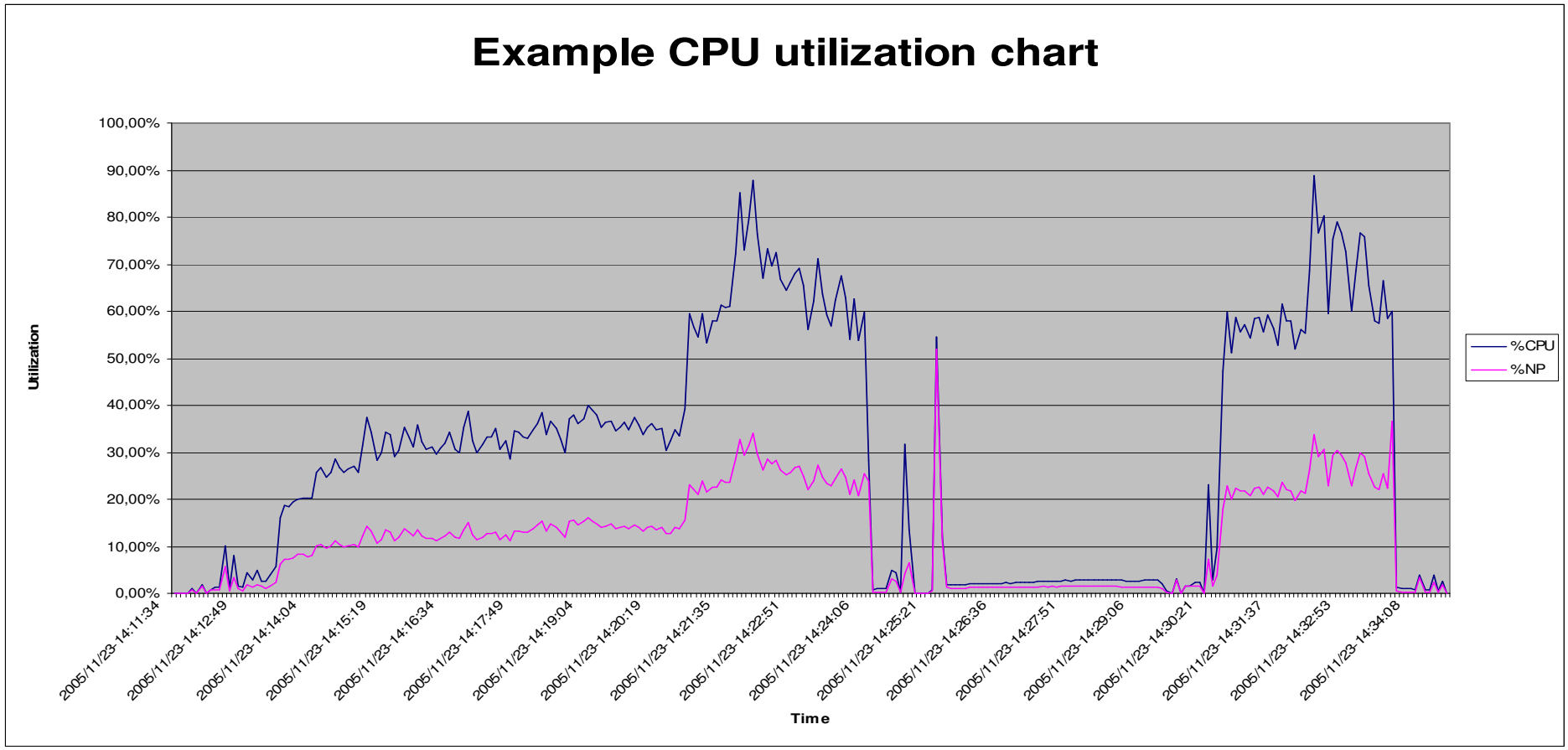
- **Data collected by the IBM provided plugins contains**
  - Information about the environment (e.g. Processor, LPAR and z/VM information)
  - Number of partitions (static, dynamic, total, maximum)
  - Partition priorities
  - Number of CPUs (active, stopped, quiced)
  - Paging (page ins, page outs)
  - Performance counters overall and per CPU
  - CPU address and status
  - CPU time, NP time, spin time, allbound time
  - Number of SVCs and dispatcher cycles



z/VSE CPU Monitor Tool (CPUMON)

- **Intended to help customers to measure the CPU utilization of their VSE system over a period of time.**
- **The VSE CPU Monitor Tool is not intended to replace any existing monitoring product provided by partners.**
  - It provides only very basic monitoring capabilities on an overall VSE system level (same data as QUERY TD)
  - No details about CPU usage of certain applications are provided
- **Download**
  - http://www.ibm.com/systems/z/os/zvse/downloads/tools.html
  - 'As is', no official support, e-mail to zvse@de.ibm.com
- **CPUMON supports 2 different output data formats**
  - CSV Format (Comma Separated Values)
    - Good for importing into spreadsheet
    - This is the default
  - XML Format
    - Used for zCP3000 Capacity Planning tool
    - Specify XML in PARM on EXEC card
  - Conversion from one format to the other one is possible (manually)

z/VSE CPU Monitor Tool

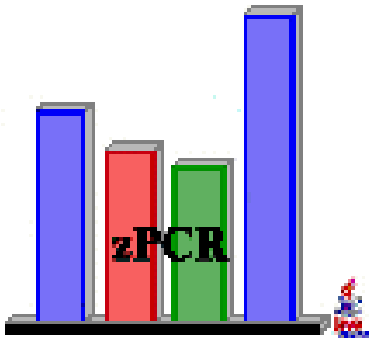


Sizing a system for z/VSE

- Sizing a system for z/VSE is different from sizing a system for z/OS
  - Although z/VSE supports multiprocessing, z/VSE does not scale as good as z/OS does
    - Do not use more than 3 active processors per z/VSE LPAR or z/VM Guest
- In general, a faster single CPU is better than multiple smaller CPUs
  - One partition can only exploit the power of one CPU
    - The largest partition (e.g. CICS) must fit into one single CPU
  - Dependent on nonparallel share (NPS) value
- Additional CPUs can be useful when multiple LPARs or z/VM Guests are used
  - Define only up to 3 CPUs per LPAR or z/VM Guest, even if more than 3 CPUs are available on the CEC
- Do **not** use MIPS tables for capacity planning purposes
  - Use zPCR Tool instead with the z/VSE workloads Batch, Online or Mixed
  - Use free of charge Capacity Planning Services from IBM

IBM Processor Capacity Reference for zSeries (zPCR)

- The zPCR tool was released for customer use on October 25, 2005
  - http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/PRS1381
  - 'As is', no official support, e-mail to zpcr@us.ibm.com
- PC-based productivity tool under Windows
- It is designed to provide capacity planning insight for IBM System z processors running various workload environments



- Capacity results are based on IBM's LSPR data supporting all IBM System z processors
  - Large System Performance Reference: https://www-304.ibm.com/servers/resourcelink/lib03060.nsf/pages/lsprindex
- For z/VSE use z/VSE workloads Batch, Online or Mixed

z/VSE Capacity Planning Offering

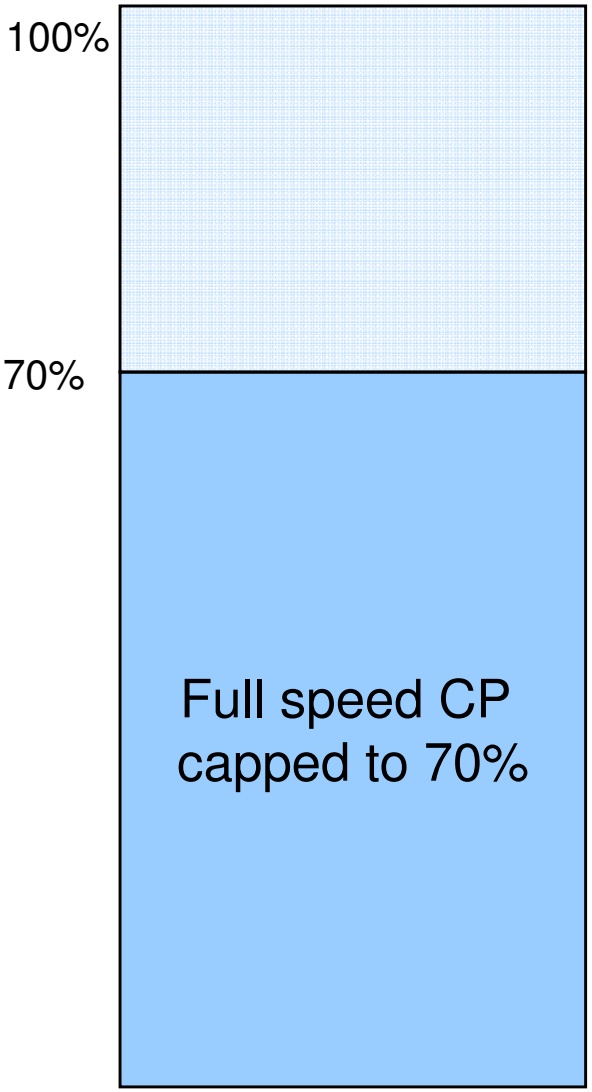
- A z/VSE Capacity Planning Offering is available
  - for Business Partners
  - and Customers
- Performance data collection is based on the XML data produced by the CPUMON Tool
- Analysis is done using zCP3000

- Contact techline@us.ibm.com and ask for z/VSE Capacity Planning Support

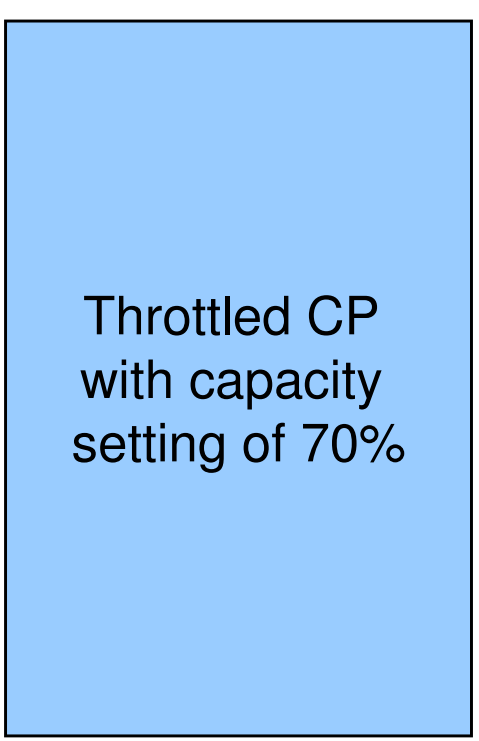
Capping versus Capacity Settings

Attention: Do not use Capping to simulate Capacity Settings !

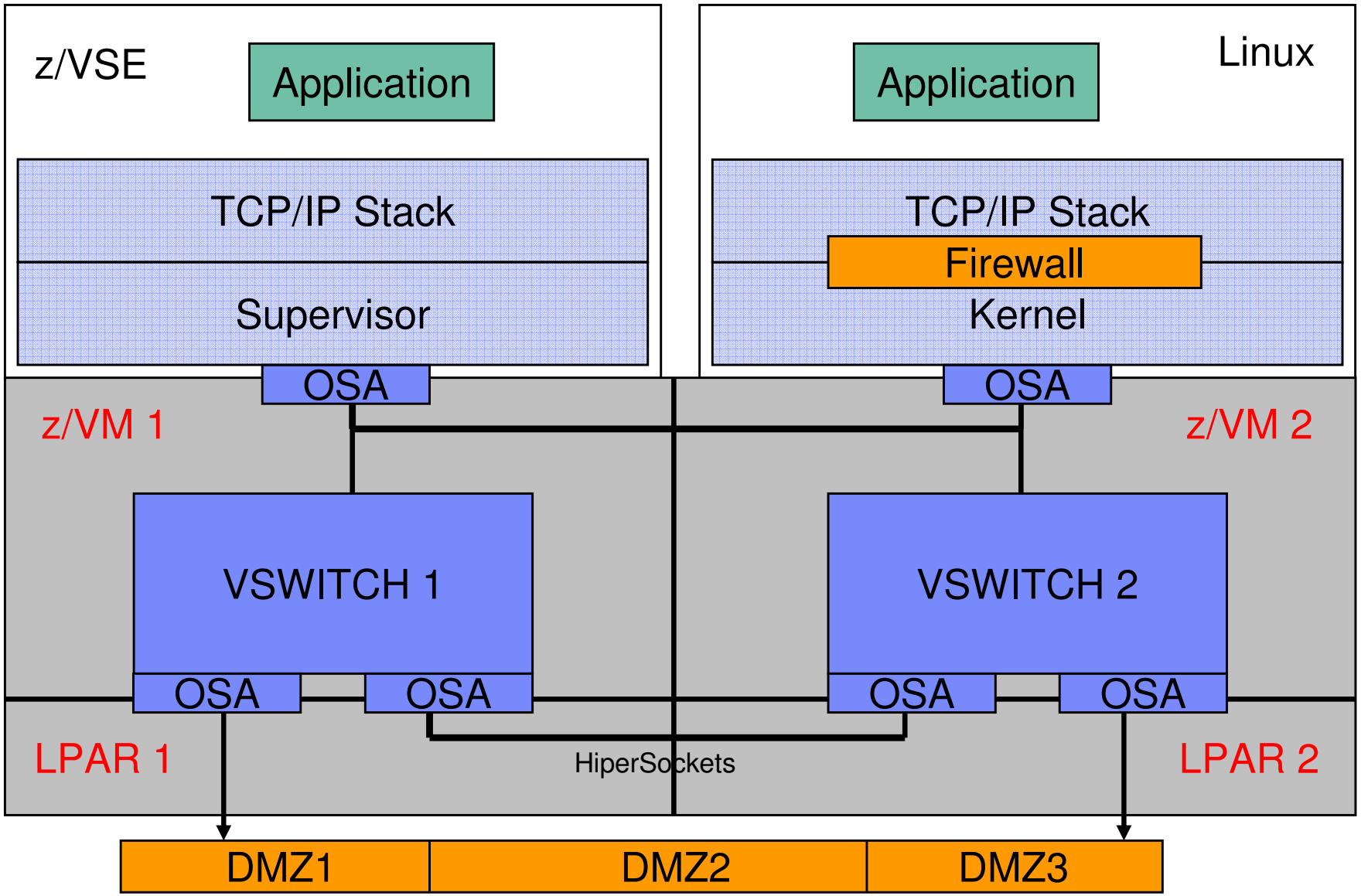
- With **Capping**, the processor runs on its full speed, until the capping stops the guest from getting dispatched by the LPAR hypervisor or z/VM (time slicing)
- With a **Capacity Setting**, the processor runs on a slower speed (and all related tasks as well, like HiperSockets memory copy, Hypervisor processing, etc)



Capping is NOT equivalent to Capacity Settings !



TCP/IP Tuning: A simple picture might not be that simple in reality



Shared OSA Adapter versus HiperSockets

To connect a z/VSE system with a Linux on System z you have 2 options:

1. Using a shared OSA Adapter

- All traffic is passed through the OSA Adapter
- The OSA Adapter has its own processor
  - Processing occurs asynchronous
  - Processing in OSA Adapter does not affect host processors

2. Using HiperSockets

- Direct memory copy from one LPAR/Guest to the other
- Memory copy is handled by the host processors
  - Processing occur synchronous
  - Consider mixed speed processors (full speed IFLs and throttled CPs)

    → Memory copy performed by throttled CP is slower than memory copy performed by full speed IFL

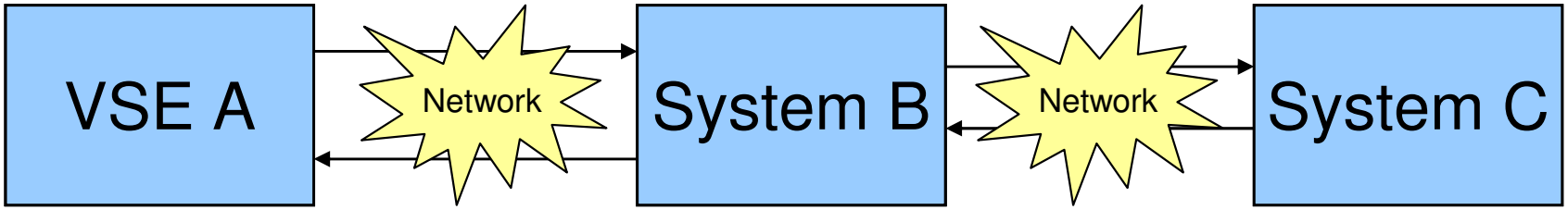
TCP/IP Tuning: Performance tuning for HiperSockets

- **When using HiperSockets to communicate between z/VSE and Linux, you may run into a "Target Buffer Full" condition**
  - This happens when z/VSE sends faster/more than Linux can receive
  - Per default Linux has 16 inbound buffers (64K per buffer = 1M per link)
  - To increase the number of buffers on Linux, use QETH option "buffer_count=128"
    - Use YAST to configure, or sysconfig scripts
    - Maximum of 128 buffers require 8MB of storage per link
- **When TCP/IP for VSE encounters this situation (BUSY), it waits 500 msec until it retries to send the packet**
  - Any additional packets to be sent are queued up
  - Problem can become dramatic, if more than 16 packets are queued up to be sent after BUSY situation
    - The resend will immediately flood the Linux buffers again, leading to the next BUSY situation, and so on....
- **You can check via QUERY STATS,LINKID=xxxx [,RESET] if you have ever run into the BUSY situation** (RESET resets the counters)

```
C1 0065 0004: IPL615I  Busy mode.............................0   ← see here
C1 0065 0004: IPL615I  Busy mode, longest....................0
```

- **You can configure a shorter BUSY wait time via DEFINE LINK command**
  - BUSY=nnn (shortest possible wait time is 100 msec)

Performance tuning in a distributed environment



- **Performance of a function in VSE may be dependent on other systems**
  - Where is the bottleneck ?
    - Is it on VSE A, or on System B or System C
    - Is it in the network ?
  - Tuning the VSE system will not help if the bottleneck is outside of VSE
  - Simple tasks on VSE may produce very time consuming tasks on other systems
- **You need to understand the whole environment**
  - With all affected systems and their dependencies

Summary

- **There is no ‘standard’ path to solve a performance problem**
  - Every customer environment is different
  - Every workload is different
- **Very seldom, a performance problem is caused by a bug in the code**
  - E.g. loops, unnecessary waits, etc
- **Mostly it is about monitoring and then suggesting tuning options**
  - Tune configuration settings
  - Find and remove bottlenecks
- **Sometimes a performance problem is because of unrealistic expectations**
  - Customer ‘thinks’ or ‘wishes’ that it should run faster
  - Customer underestimates resource/CPU requirements of a new function/workload
  - Someone ‘promised’ good performance to get a deal closed

Questions ?

THANK YOU

http://www.ibm.com/zVSE