E43

VSE Performance update

Ingo Franzki (ifranzki@de.ibm.com)

zSeries Expo | Nov. 1 - 5, 2004

Miami, FL

RETURN TO INDEX

IBM eServer zSeries

VSE Performance update

Ingo Franzki (ifranzki@de.ibm.com)

10/14/2004

Trademarks

The following are trademarks of the International Business Machines Corporation in the United States and / or other counties.

CICS*
DB2*
DB2 Connect
DB2 Universal Database
e-business logo*
Enterprise Storage Server
HiperSockets
IBM*
IBM logo*
IMS
Intelligent Miner
Multiprise*
MQSeries*
OS/390*
S/390*
SNAP/SHOT*
Virtual Image Facility
VM/ESA*
VSE/ESA
VisualAge*
VTAM*
WebSphere*
xSeries
z/Architecture
z/VM
zSeries

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

LINUX is a registered trademark of Linus Torvalds

Tivoli is a trademark of Tivoli Systems Inc.

Java and all Java-related trademarks and logos are trademarks of Sun Microsystems, Inc., in the United States and other countries

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

Microsoft, Windows and Windows NT are registered trademarks of Microsoft Corporation.

SET and Secure Electronic Transaction are trademarks owned by SET Secure Electronic Transaction LLC.

Intel is a registered trademark of Intel Corporation.

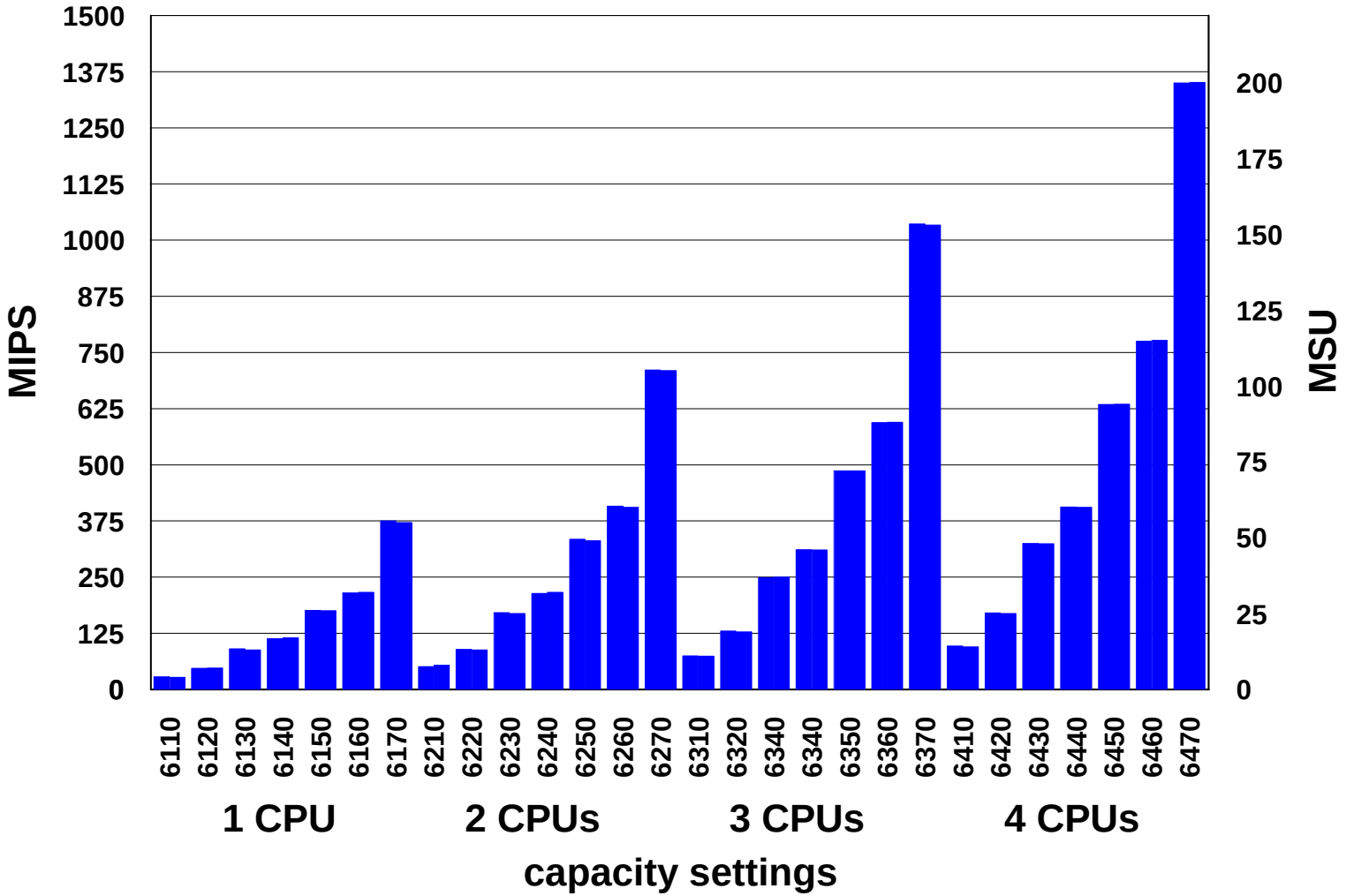
Agenda

- **News**
- **Hardware support**
- **HiperSockets**
- **Hardware Crypto support**
- **Dependencies for VSE/ESA Growth**
- **Turbo Dispatcher**
- **VSE health check**
- **Hints and Tips**

New: z/VSE 3.1 preview announcement

- **z/VSE 3.1 is planned to be able to support:**
  - z890, z800, z900, z990
  - Multiprise 3000, G5 and G6
  - Fibre Channel Protocol for SCSI — FCP channels
- **IBM plans to continue to ship CICS/VSE V2.3 together with CICS TS for VSE/ESA**
  - at no additional charge.
- **z/VSE plans to offer simplified packaging**
  - LE will become a component of VSE Central Functions
- **Fast Service Upgrade possible from VSE 2.7 and 2.6**
  - using equivalent ECKD disks
  - NOT: from ECKD to SCSI-FCP disks.

New: IBM eServer zSeries 890



z890 consists of one Model (A04) and 28 capacity settings

New: Enterprise Storage Server Model 750 (Baby-Shark)

- **The ESS Model 750 is based on the same architecture as the ESS Model 800 to support functionality, stability, and reliability**
- **up to 64 disk drives**
- **4.6 terabytes (TB) of physical capacity**
- **A two-way processor**
- **8 GB of cache**
- **2 GB of Non Volatile Storage (NVS)**
- **up to 6 Fibre Channel/FICON or ESCON host adapters**
- **Support for 72.8 GB and 145.6 GB 10,000 rpm drives**
- **configured as RAID 5, RAID 10, or a combination of both**

VSE/ESA 2.7 Hardware support

- **VSE/ESA 2.7 runs on the following machines**
- **zSeries: z800, z900, z990, z890**
- **9672 Parallel Enterprise Server (G5/G6)**
- **Multiprise 3000 (7060)**
- **equivalent emulators (Flex-ES)**
- **VSE/ESA 2.7 is based on the hardware instruction set described in the manual 'ESA/390 Principles of Operation' (SA22-7201).**
- **With VSE/ESA 2.7 it is assumed that all the ESA/390 instructions and facilities described in that manual can be used.**

Supported VSE Releases

- **VSE/ESA 2.4/2.3: already out of service**
  - runs also on zSeries (z800, z900)
  - does not run on z990, z890 (Hardwait during IPL)
- **VSE/ESA 2.5: end of service 12/31/2003**
  - runs also on zSeries (z800, z900)
  - runs also on z990 with additional PTF
- **VSE/ESA 2.6**
  - runs also on zSeries (z800, z900)
  - runs also on z990, z890 with additional PTF
- **VSE/ESA 2.7**
  - runs on zSeries (z800, z900, z990, z890, G5/G6, MP3000)
- **OSA Express: Supported with VSE/ESA 2.6 and 2.7**
- **HiperSockets and PCICA (Crypto)**
  - Supported with VSE/ESA 2.7

zSeries Remarks

- **Prior to zSeries there is one cache for data and instructions**
- **zSeries has splited data and instruction cache**
- **Performance implications:**
  - If program variables and code that updates these program variables are in the same cache line (256 byte)
    - Update of program variable invalidates instruction cache
    - Performance decrease if update is done in a loop
  - See APAR PQ66981 for FORTRAN compiler

Hardware Support

- **Queued Direct I/O**
  - Designed for very efficient exchange of data
  - Uses the QDIO Hardware Facility, without traditional S/390 I/O instructions
  - Without interrupts (in general)
  - Use of internal queues
  - With pre-defined buffers in memory for asynchronous use
- **Exploitation by TCP/IP for VSE/ESA**

HiperSockets hardware elements ('Network in a box')

- **Synchronous data movement between LPARs and virtual servers within a zSeries server**
  - Provides up to 4 "internal LANs" HiperSockets accessible by all LPARs and virtual servers
  - Up to 1024 devices across all 4 HiperSockets
  - Up to 4000 IP addresses
  - Similar to cross-address-space memory move using memory bus
- **Extends OSA-Express QDIO support**
  - LAN media and IP layer functionality (internal QDIO = iQDIO)
  - Enhanced Signal Adapter (SIGA) instruction
    - No use of System Assist Processor (SAP)

HiperSockets hardware elements ('Network in a box')

- **HiperSockets hardware I/O configuration with new CHPID type = IQD**
  - Controlled like regular CHPID
  - Each CHPID has configurable Maximum Frame Size
- **Works with both standard and IFL CPs**
- **No physical media constraint, no physical cabling, no priority queuing**
- **Secure connections**

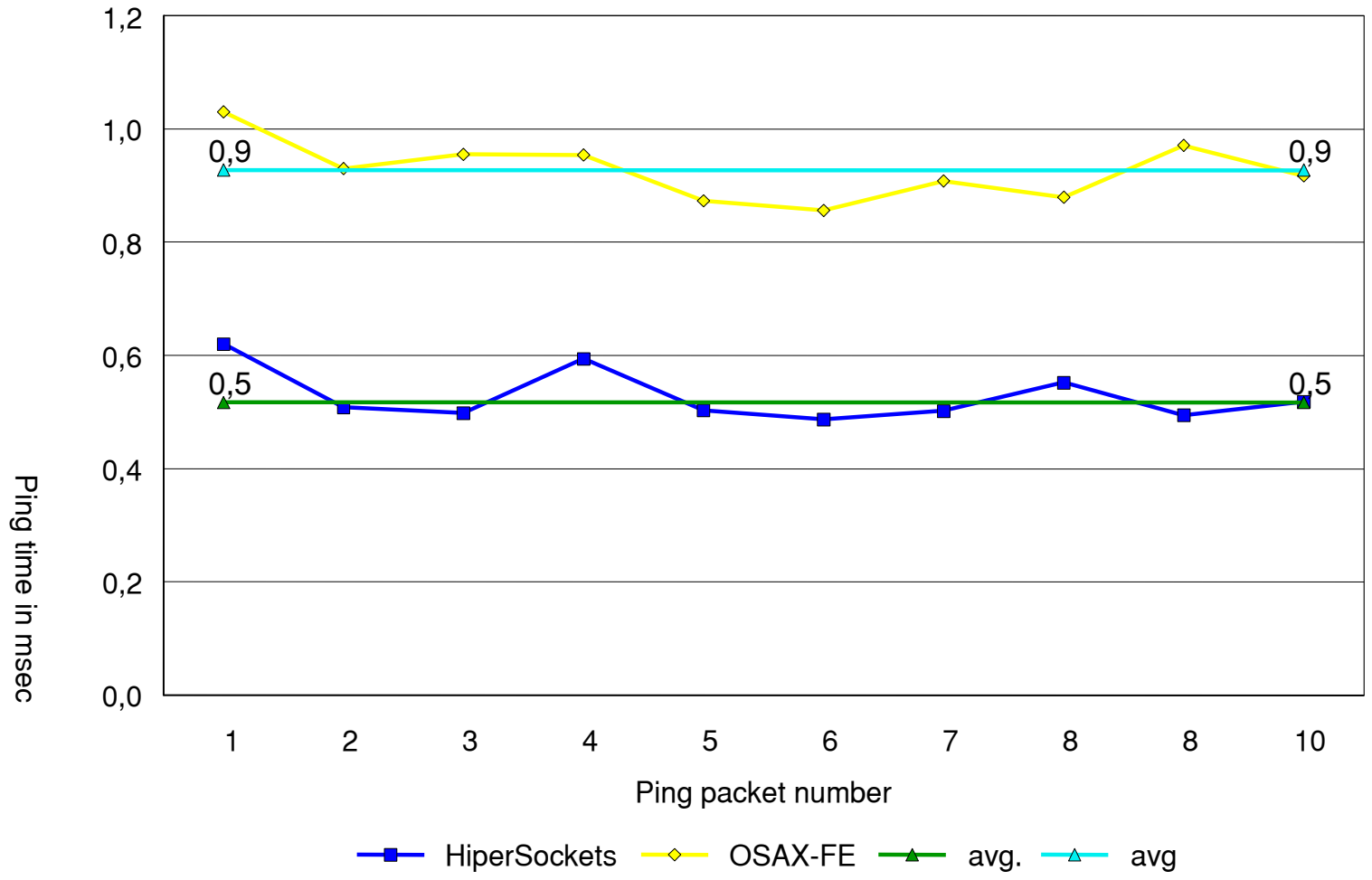
Measurement Environment

- **z800 (2066-004)**
  - 4 processors
- **VSE/ESA 2.7 GA Driver in a LPAR (native)**
  - 1 CPU active (~2066-001)
  - TCPIP00 (F7): OSA Express Fast Ethernet
  - TCPIP01 (F8): HiperSockets
- **Linux for zSeries in a LPAR (native)**
  - 3 CPUs active (shared)
  - eth0: OSA Express Fast Ethernet
  - hsi10: HiperSockets

Latency (Round trip time) - results

- **Measurements has been done with PING command**
  - Issued at Linux side
  - 10 Pings
  - PING sends a datagram to VSE
  - VSE sends a answer back to Linux
  - Time until answer arrives is measured
    - Round trip time

Latency (Round trip time) - results

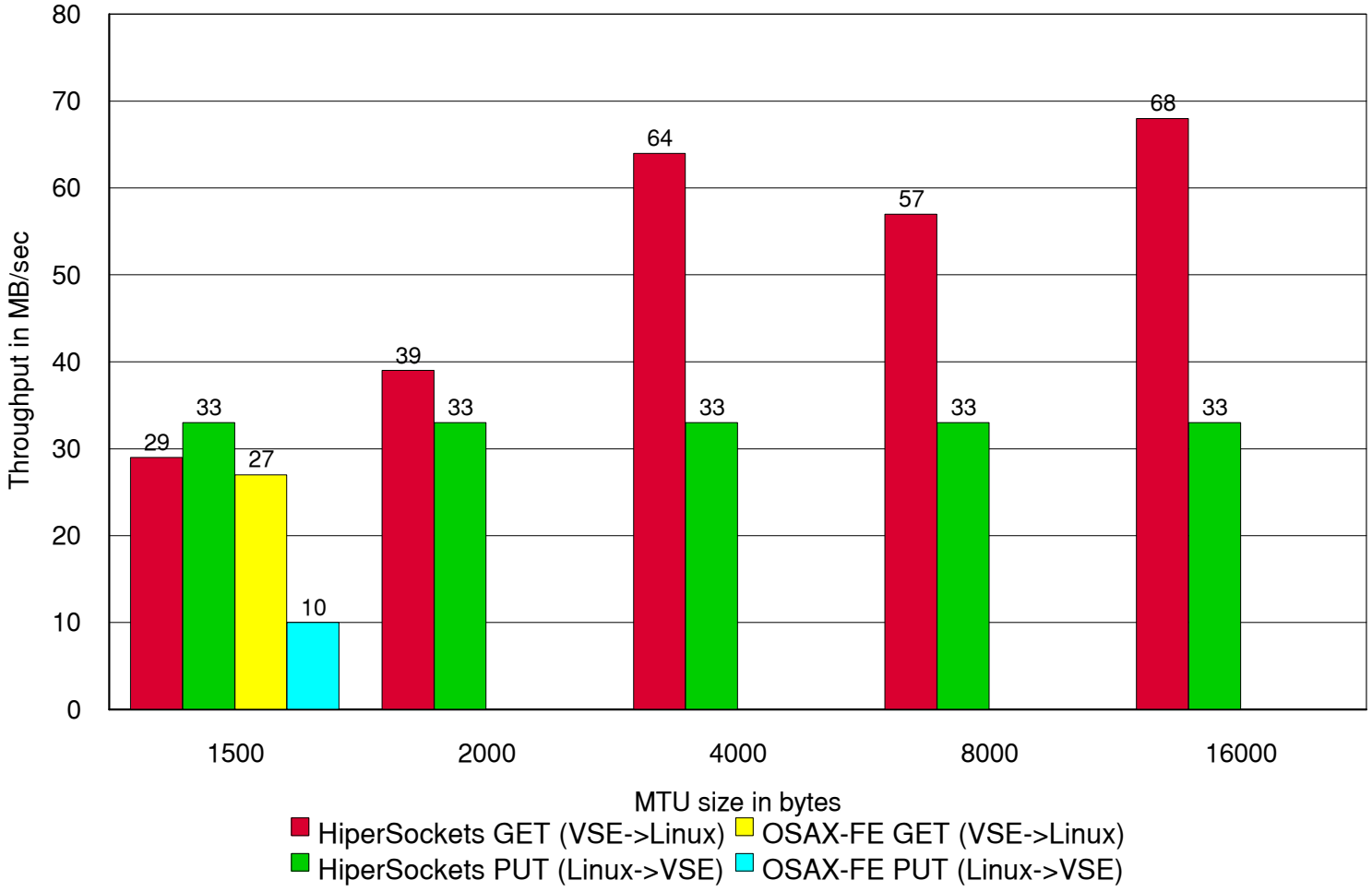


HiperSockets is about 1.8 times faster in terms of latency

Throughput (MB/sec)

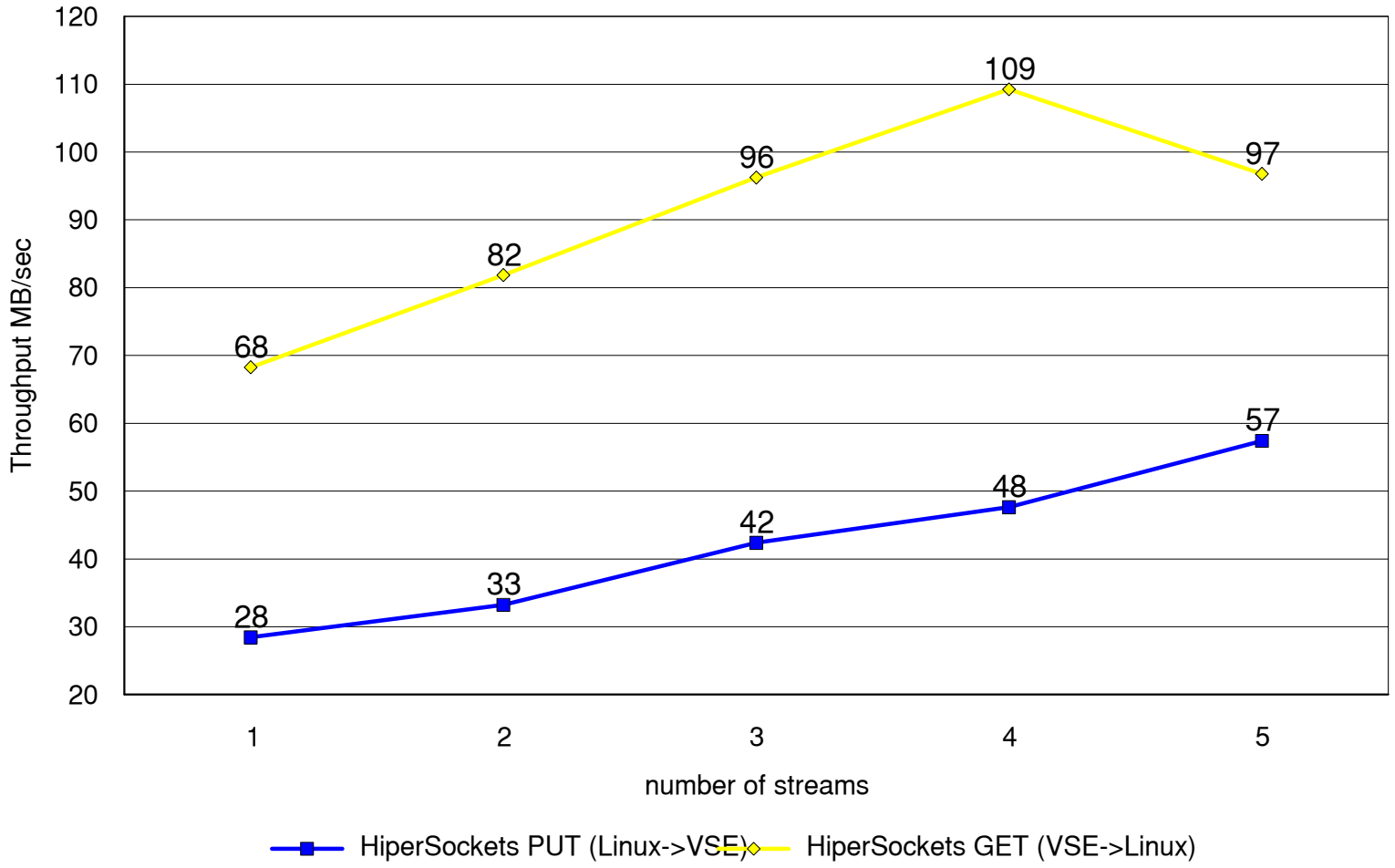
- **Measurements has been done with FTP**
  - Initiated at the Linux side
  - Transferring 1GB (1000MB)
    - without translation (binary)
    - 1 to 5 parallel streams
  - PUT: send data to VSE
    - VSE inbound
    - sending a 1GB file to $NULL file (in memory file)
    - No file I/O is done by VSE/Linux
  - GET: receive data from VSE
    - VSE outbound
    - receiving $NULL file (in memory file) into /dev/null
    - No file I/O is done by VSE/Linux

Throughput (MB/sec) - results



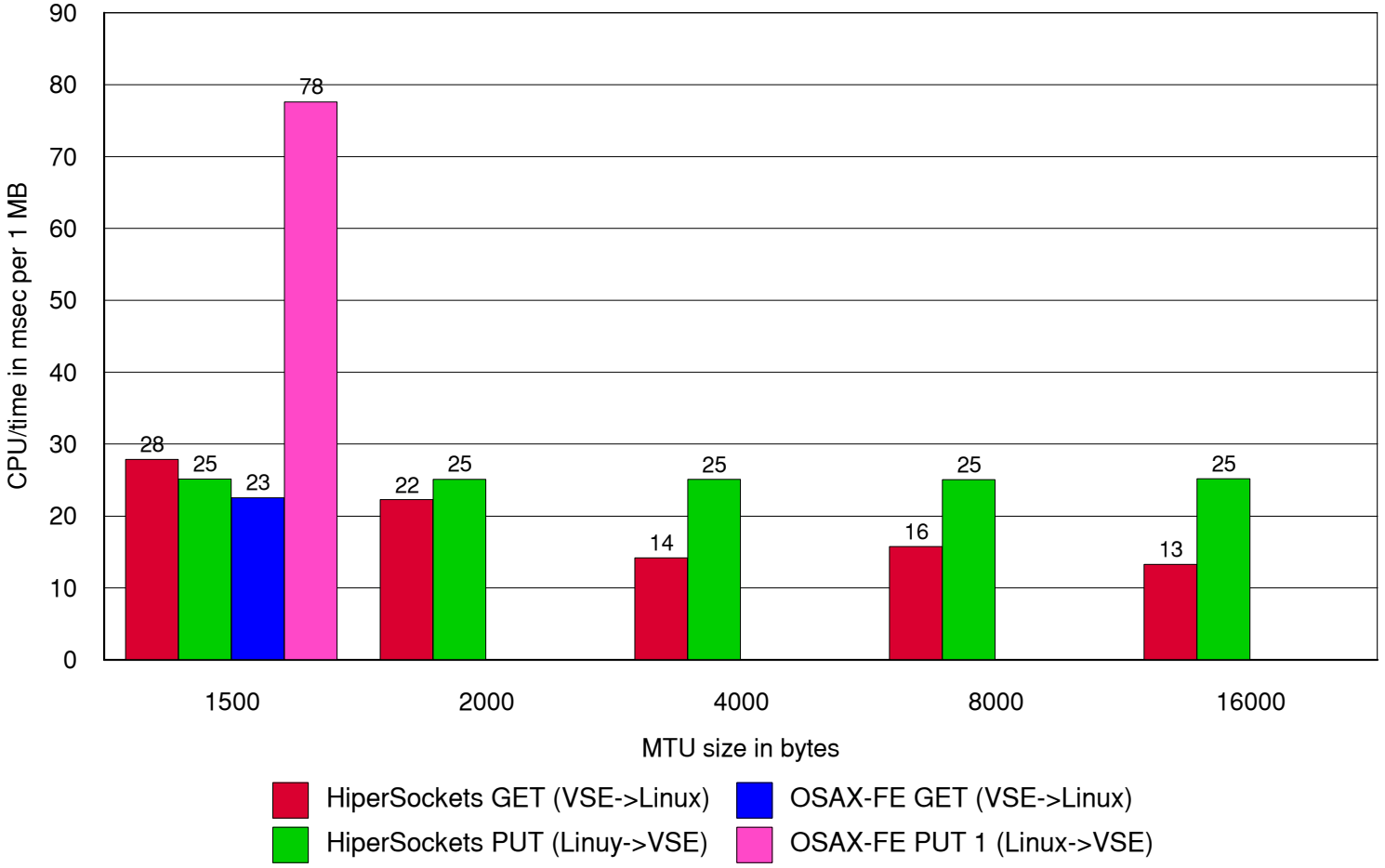
HiperSockets throughput is between 30-80 MB/sec

Throughput (MB/sec) - results (2)



Maximum HiperSockets throughput of 109 MB/sec at 4 concurrent connections

CPU time per MB - results

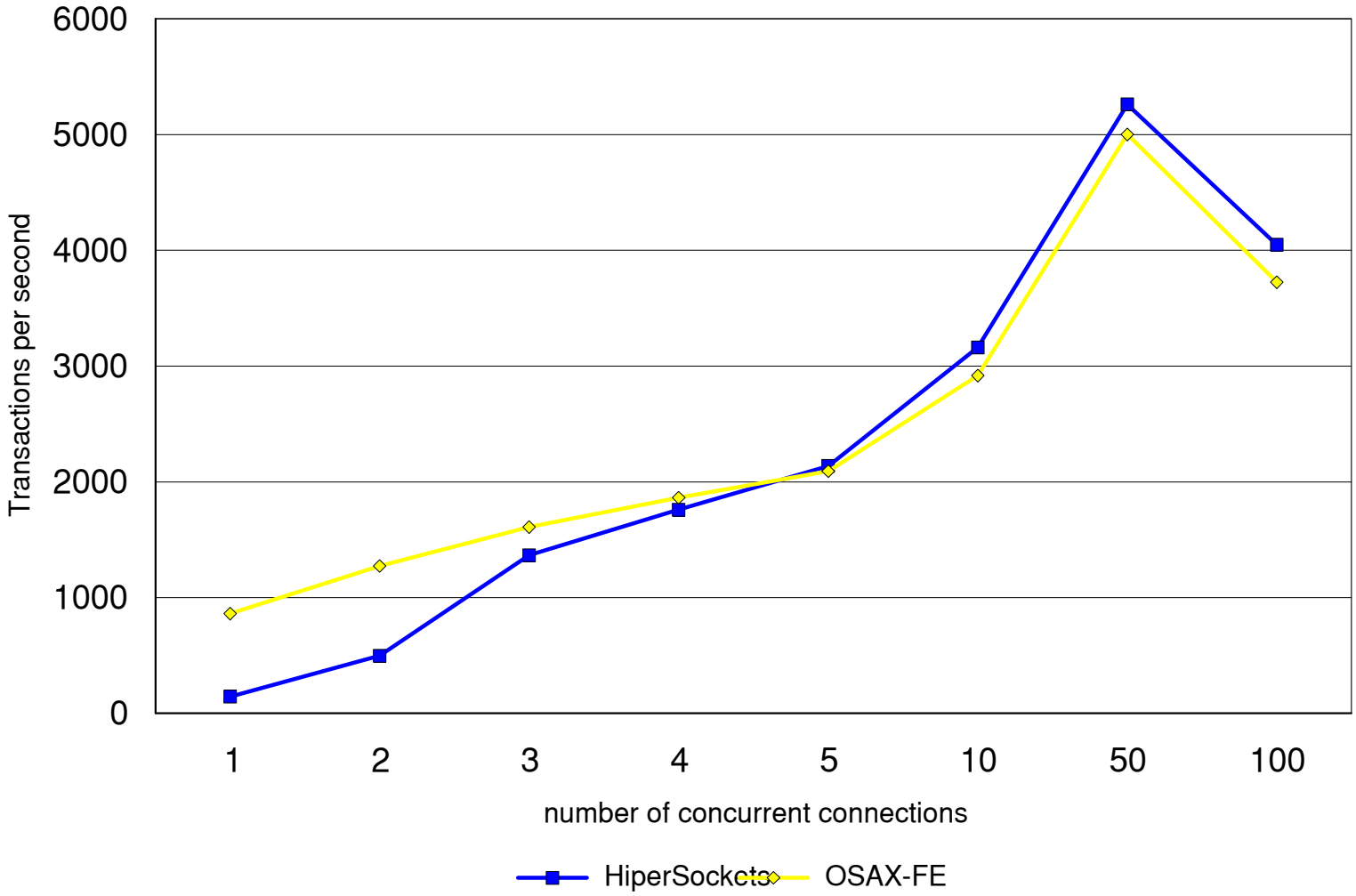


About 15-30 msec CPU time per MB for HiperSockets
(on a z800 2066-001)

Transaction per second

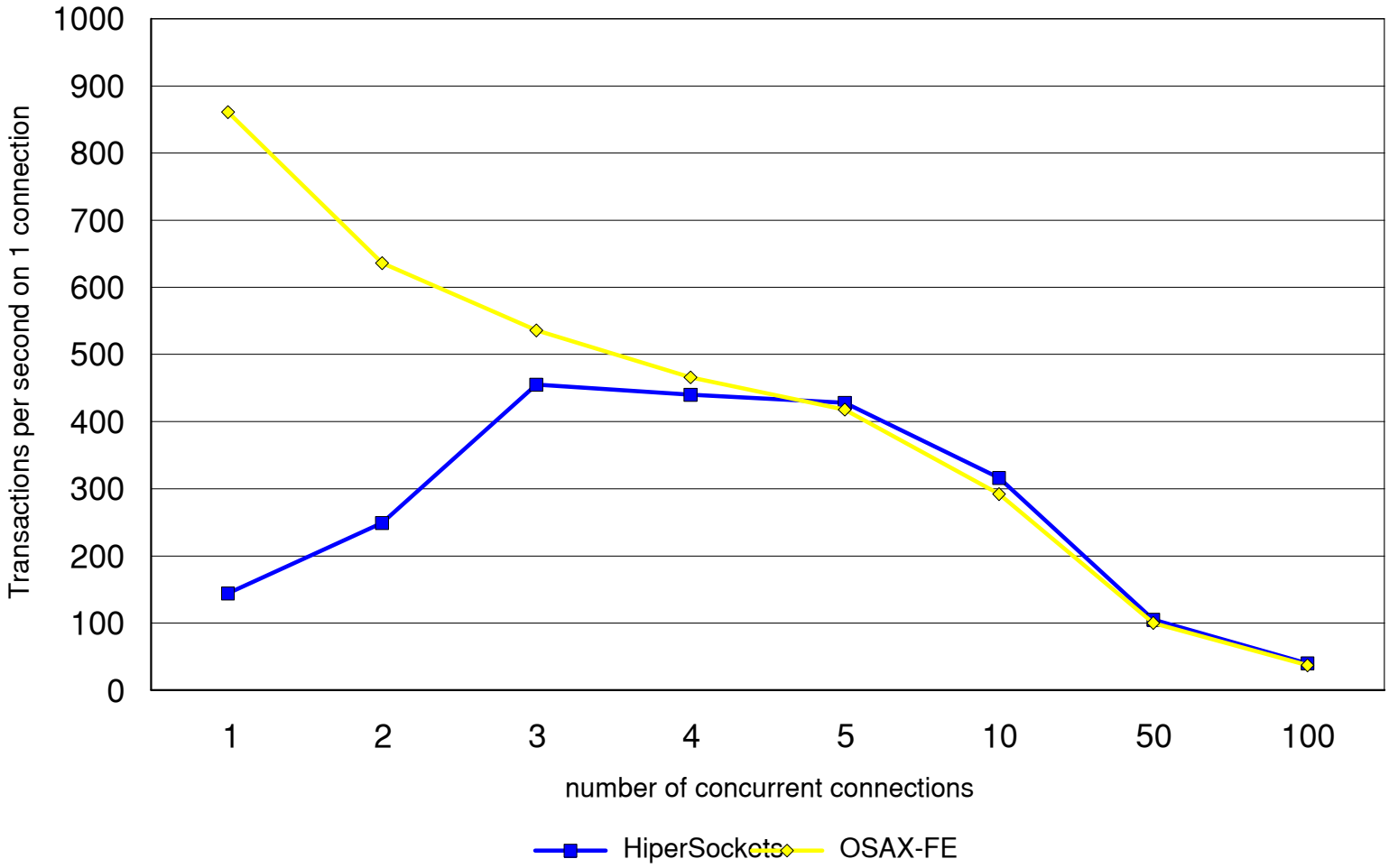
- **Measurements has been done with an ECHO server**
  - Client on Linux sends 100 bytes to server
  - Server on VSE echoes 100 bytes
  - Per TCP connection 10000 transactions are driven
  - Variations: Number of TCP connections
    - 1,2,3,4,5
    - 10,50,100
  - Measurements
    - Transactions per second
    - CPU time per transaction

Transactions per second – results



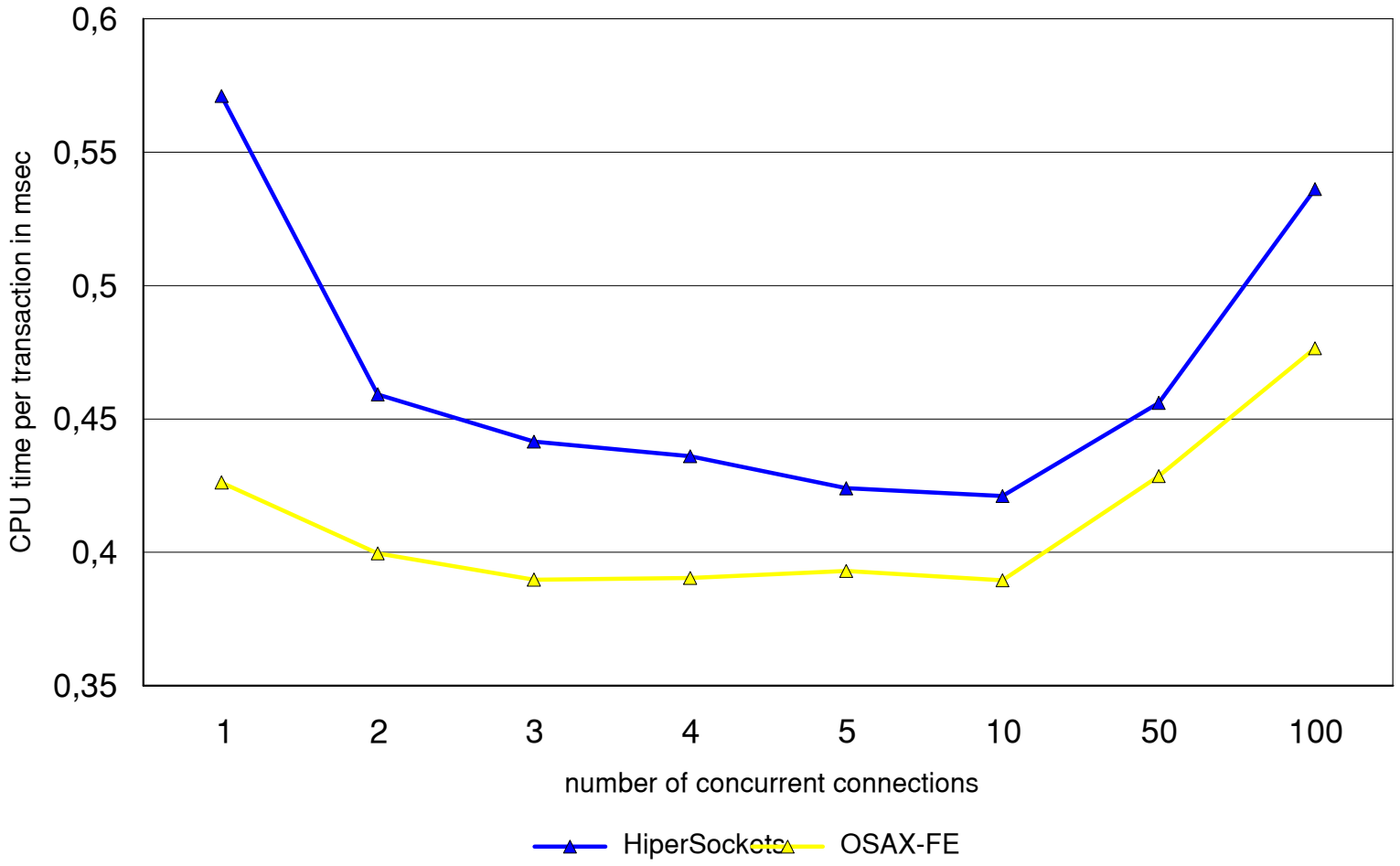
Maximum of 5200 transactions per second at 50 concurrent connections

Transactions per second on 1 connection - results



HiperSockets: Maximum of about 450 transactions per second on 1 connection (= about 2 msec response time)

CPU time per transaction



HiperSockets: About 0.45 msec CPU time per transaction for 2-50 connections

Measurement Results - conclusion

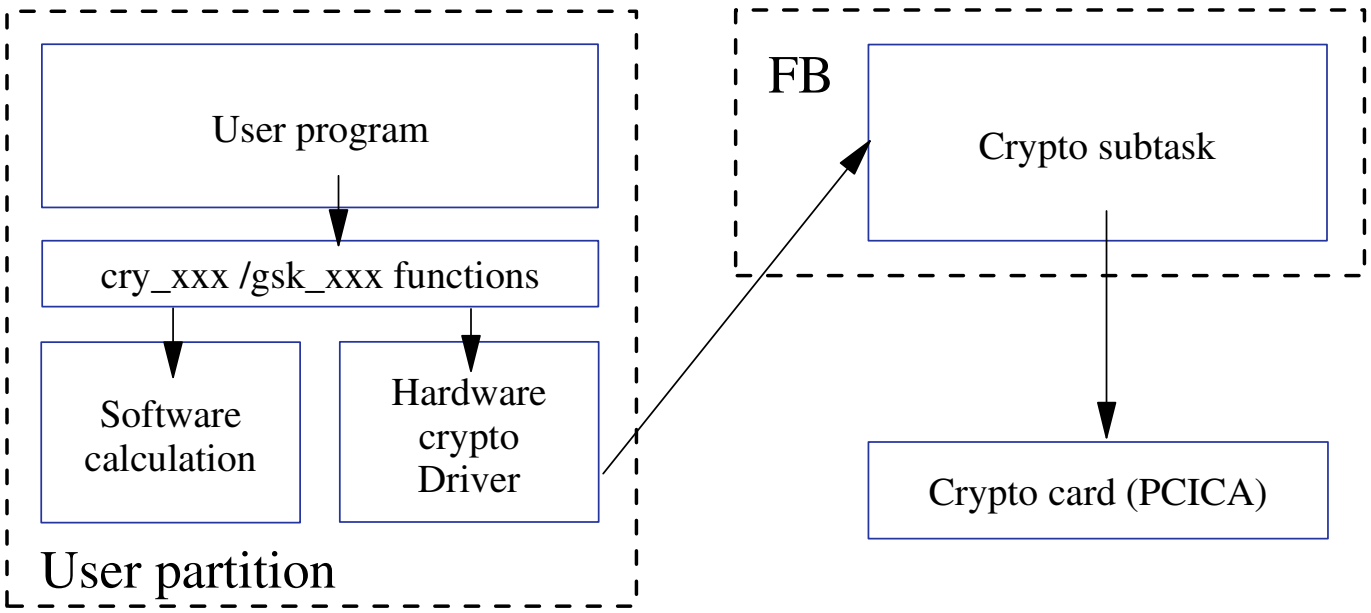
- **HiperSockets**
  - Throughput
    - Between 30-80 MB/sec
    - Maximum throughput of 109 MB at 4 connections
    - About 15-30 msec CPU time per MB
  - Transactions per second
    - Maximum of 5200 Transactions per second at 50 connections
    - About 0.4-0.45 msec CPU time per transaction

Hardware Crypto Overview

- **Requires VSE/ESA 2.7 and TCP/IP for VSE/ESA 1.5**
- **Supported crypto cards**
  - PCI Cryptographic Accelerator (PCICA)
    - Feature code 0862
    - Available for zSeries (z800, z900)
- **The crypto card is plugged into the Adjunct Processor**
- **Currently only RSA (asymmetric) is supported**
  - Of benefit for Session initiation (SSL-Handshake)
- **Also supported with**
  - z/VM 4.2 + APAR VM62905
  - z/VM 4.3

Hardware Crypto Overview - continued

- **New crypto subtask in Security Server (SECSERV) running in FB**
  - Or as separate job if no SECSERV is running
  - Crypto card is polled by crypto task



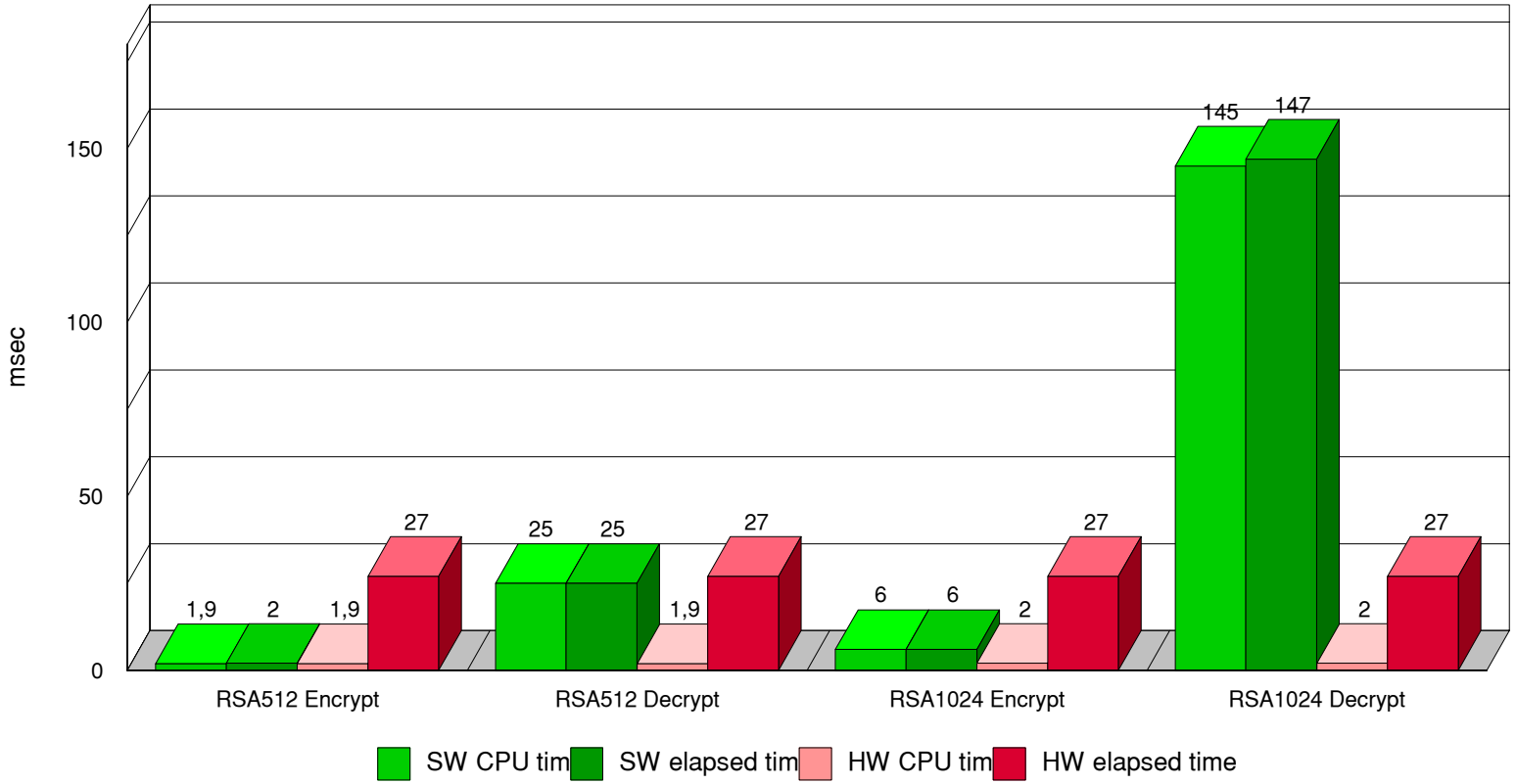
Measurement Environment

- **VSE/ESA 2.7 running on a z900 (2064-109)**
  - on 1 processor (~2064-101)
  - with a PCI Cryptographic Accelerator
- **Test case programs on VSE**
  - Crypto operations measurements
    - calling cry_xxx functions (RSA, DES, SHA, MD5)
    - each crypto operation is performed 10000 times
  - Secured data transfer (SSL)
    - performs SSL handshake
    - performs encrypted data transfer
    - counterpart program running on Windows (SSL-client)
- **All RSA operations are measured**
  - with Hardware Crypto support
  - with Software Crypto
    - (support already available with TCP/IP 1.4/1.5 as shipped in VSE/ESA 2.6)

Measurement Environment - continued

- **Variations**
  - RSA encrypt/decrypt
    - 512 / 1024 bit key
  - DES, DES CBC, 3DES CBC encrypt/decrypt
    - software crypto only
    - message length (128, 256, 512 bytes)
  - SHA Hash, MD5 Hash, SHA HMAC, MD5 HMAC
    - software crypto only
    - message length (128, 256, 512, 1K, 2K bytes)
  - SSL handshake/data transfer
    - 01 RSA512_NULL_MD5
    - 02 RSA512_NULL_SHA
    - 08 RSA512_DES40CBC_SHA
    - 09 RSA1024_DES_CBC_SHA
    - 0A RSA1024_3DES_EDE_CBC_SHA

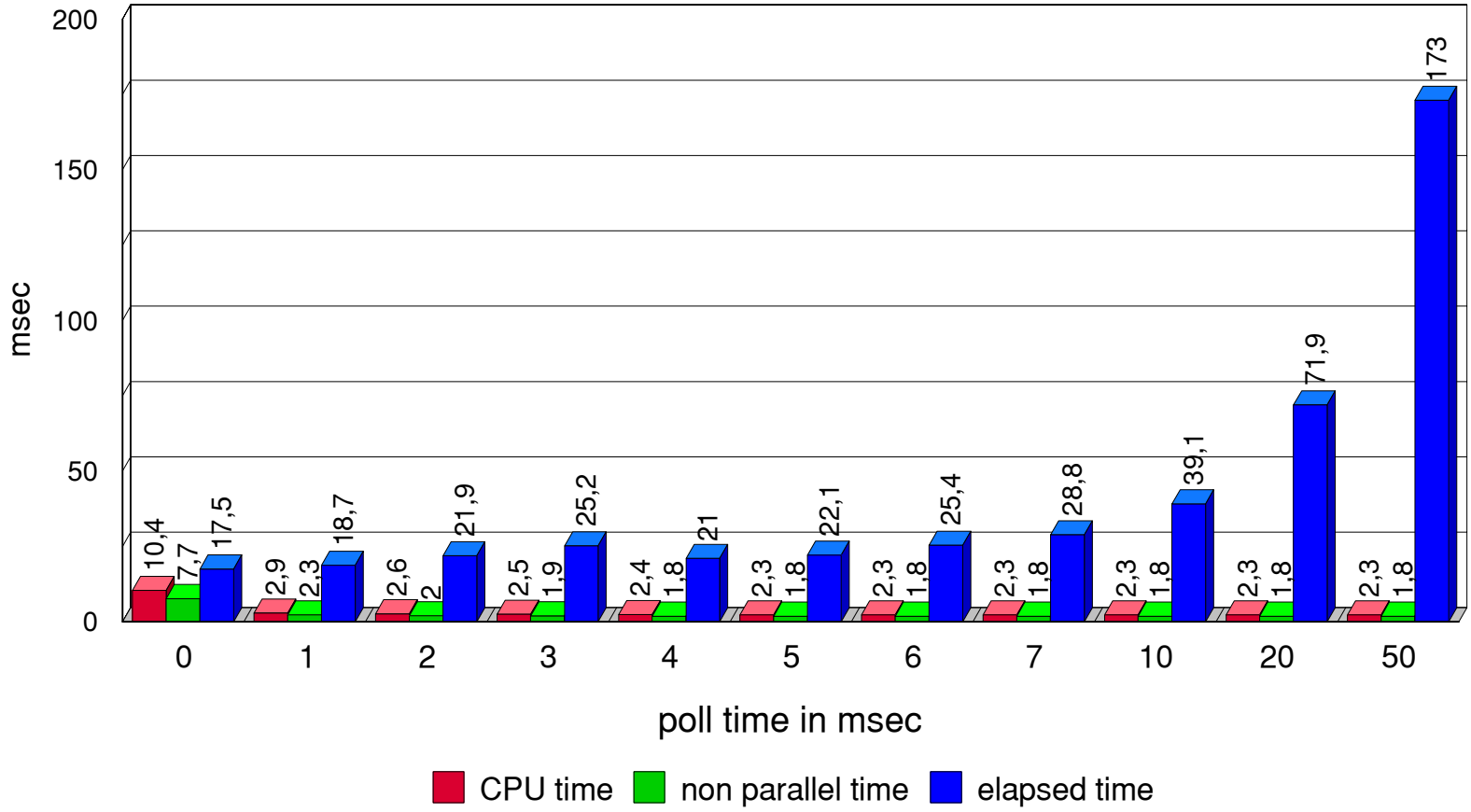
Measurements Results - RSA



HW Crypto:

- CPU time and elapsed time is independent of operation / key length
- RSA operation takes about 2 msec CPU time and 28 msec ellapsed time
- CPU time is always less than software crypto

Measurements Results - RSA polltime

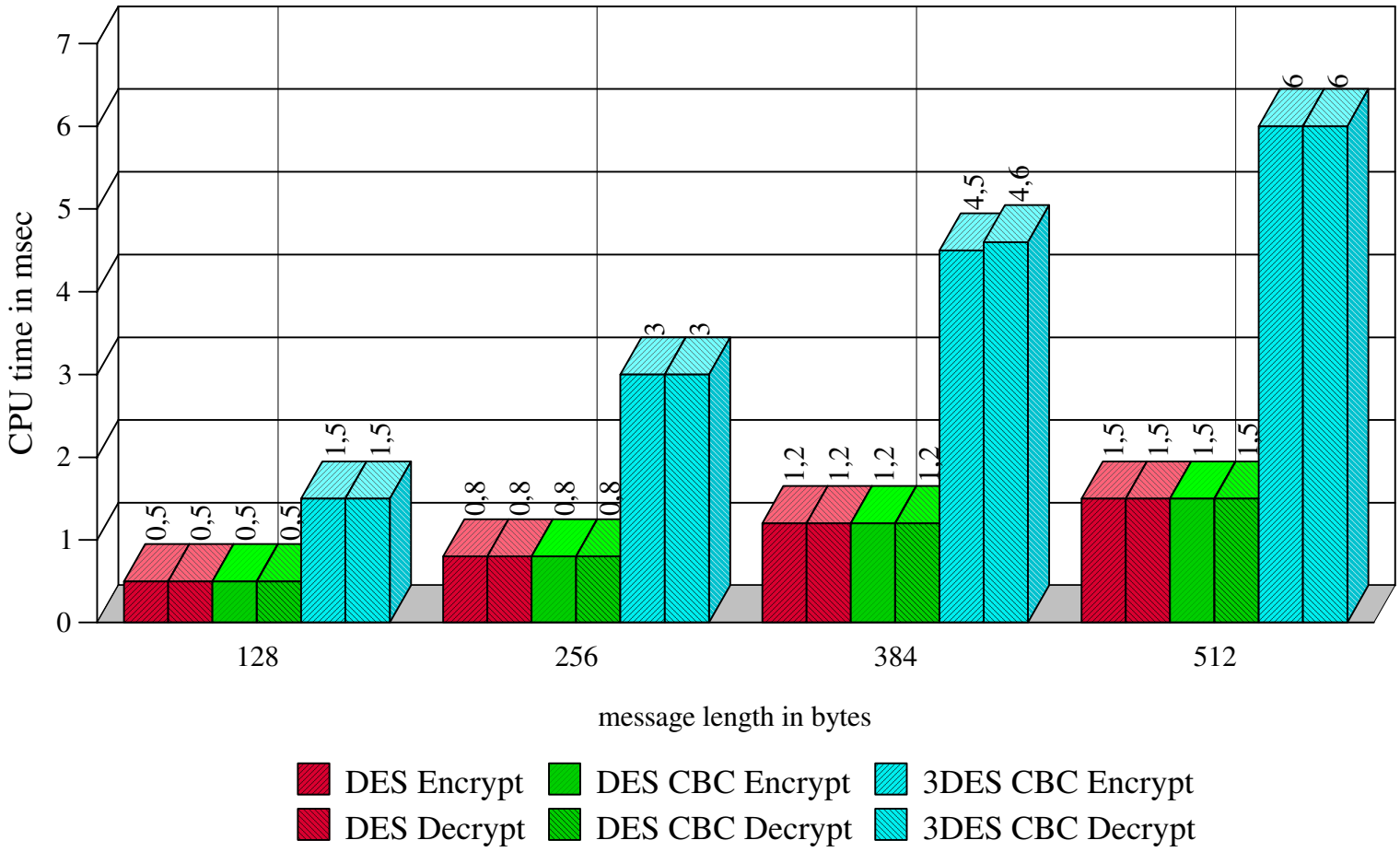


Per default a polltime of 7 msec is used.

Can be changed with: MSG FB,DATA=WAITTIME=nn

Smaller values increases CPU time, higher values increases ellapsed time

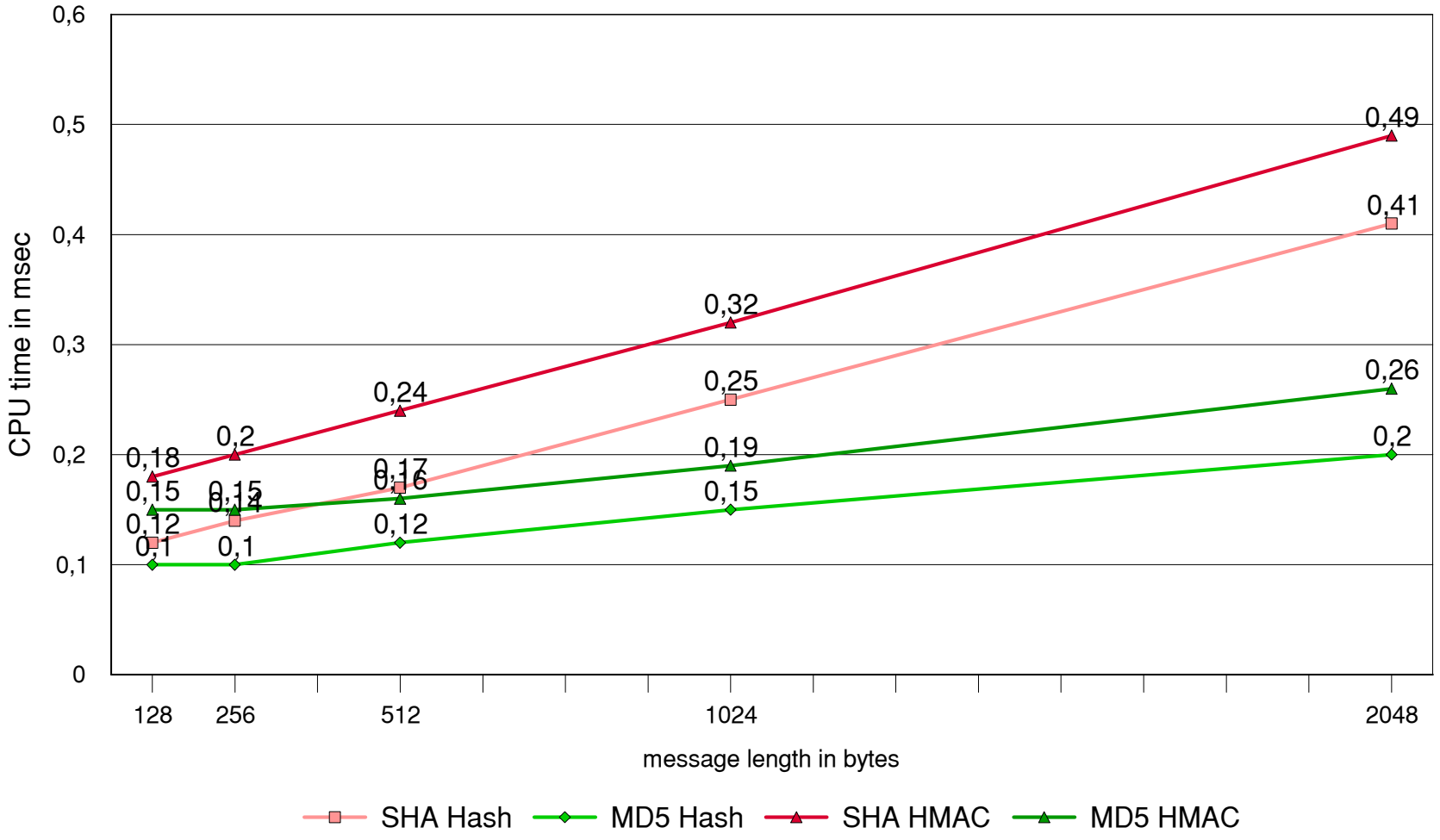
Measurements Results - DES, DES CBC, 3DES CBC



Software Crypto only!

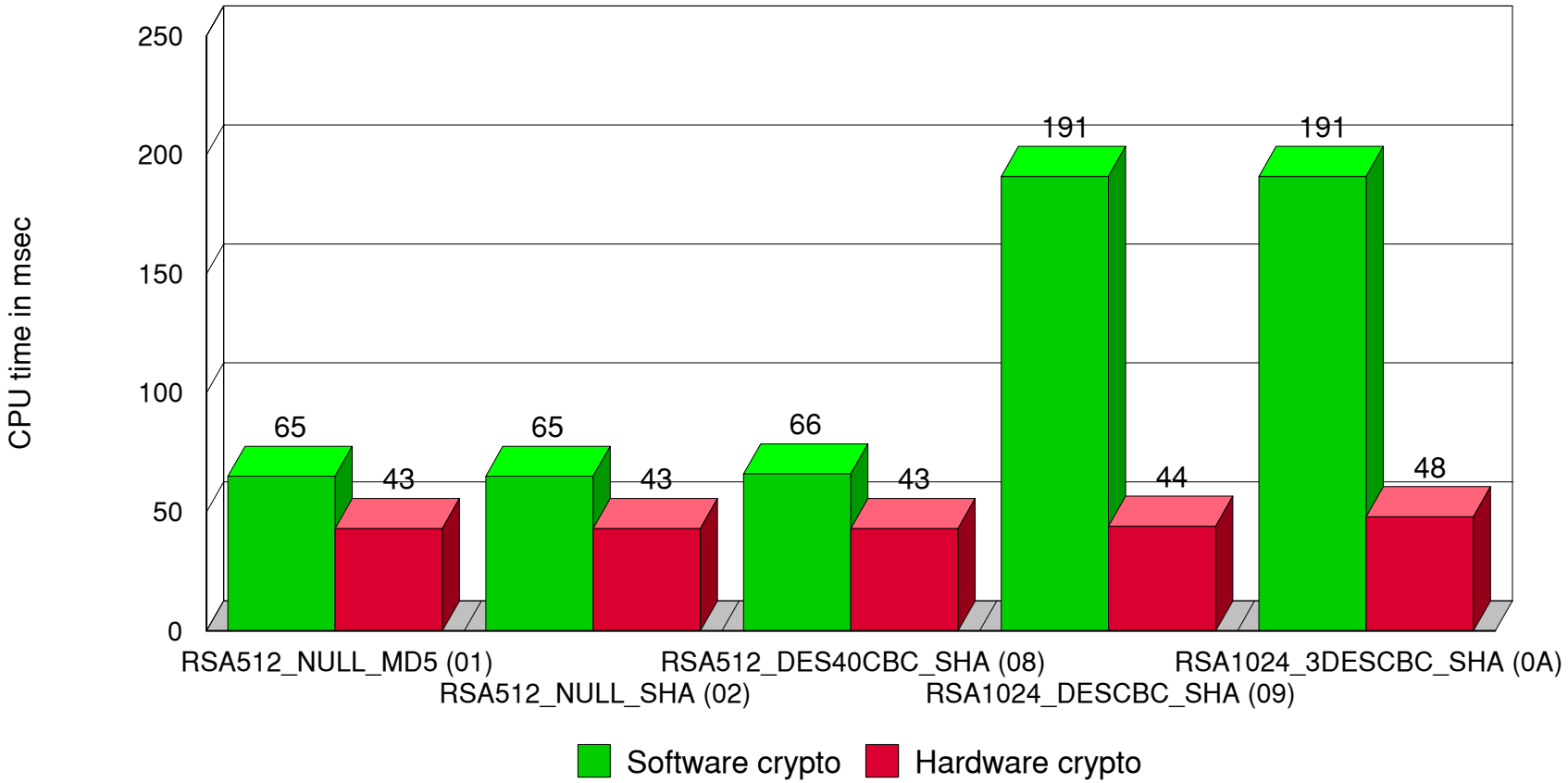
DES and DES CBC takes similar CPU times, 3DES CBC about 3.8 times

Measurements Results - SHA, MD5



SHA takes about 1.8 times more CPU time compared to MD5

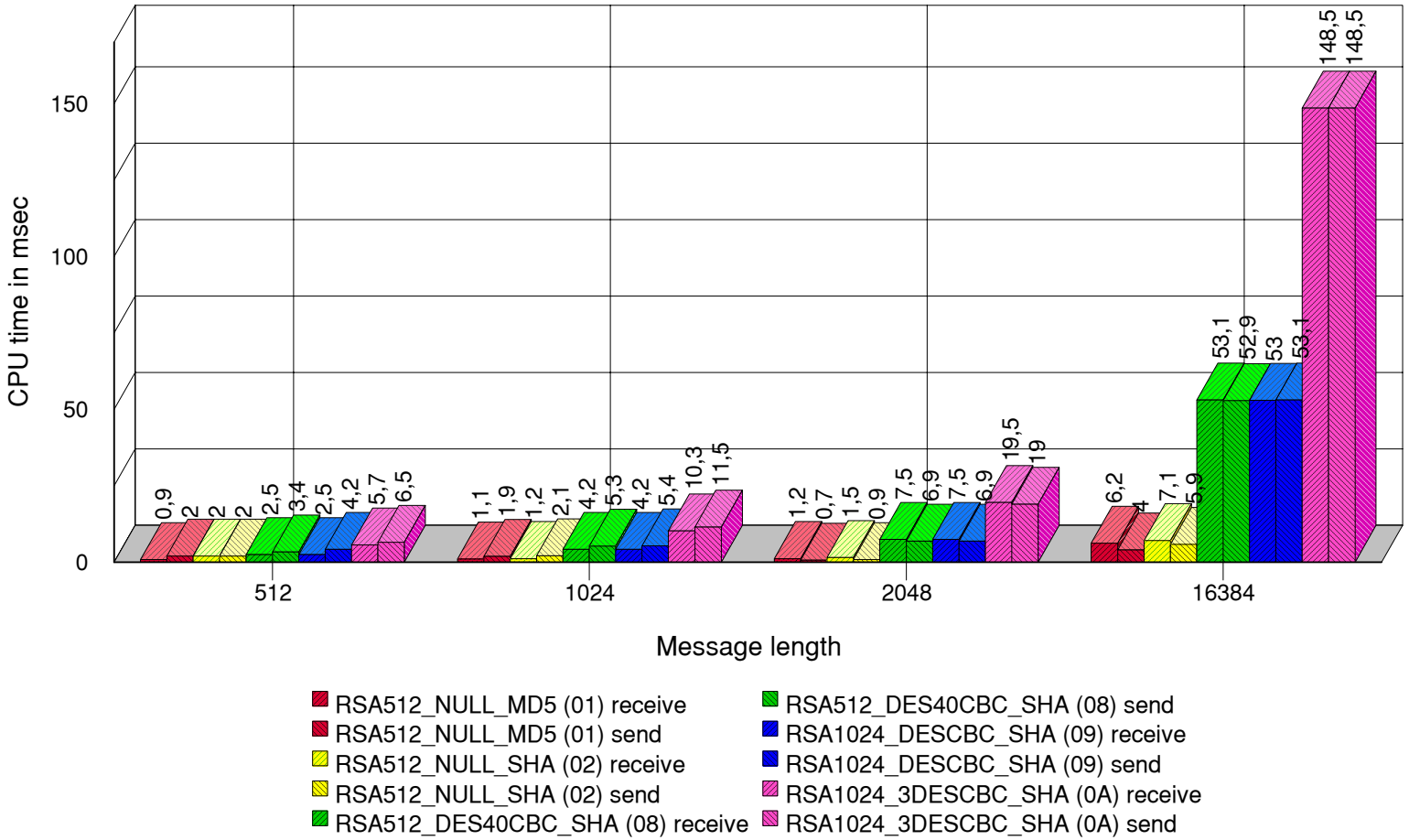
Measurements Results - SSL Handshake



HW Crypto:

- CPU time and elapsed time is independent of cipher suite used
- SSL handshake takes about 43-48 msec CPU time (connection establishment)

Measurements Results - SSL data transfer



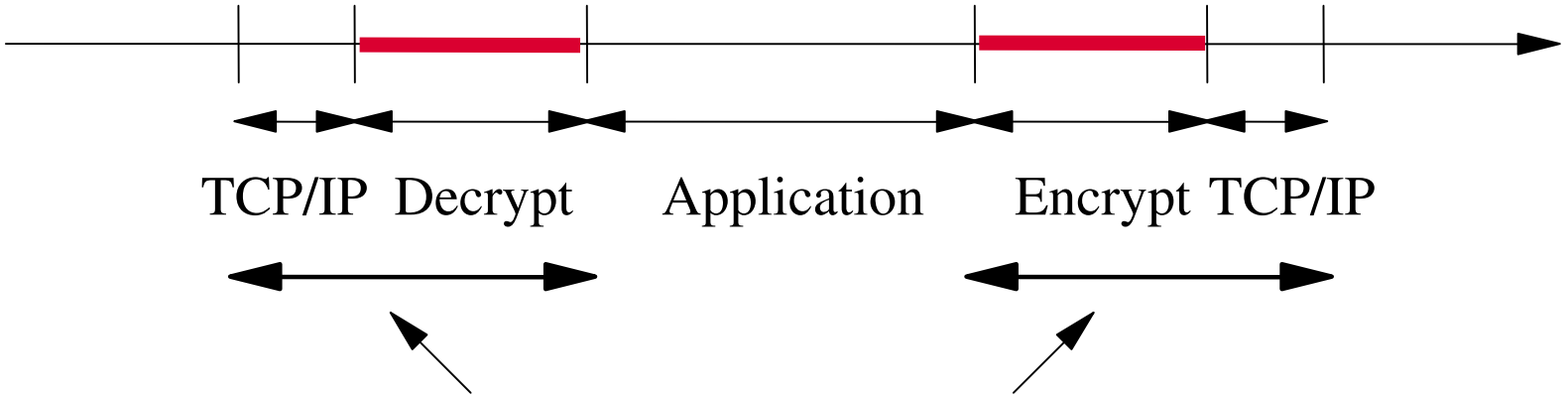
CPU time depends on used hashing (SHA/MD5) and encryption algorithm (DES/3DES)

Software Crypto only!

SSL data transfer overhead

Non SSL

SSL



this has been measured

Measurements Results - conclusion

- **HW Crypto**
  - Supports RSA operations only (e.g. used by SSL handshake)
  - CPU time/elapsed time is independent of operation and key length
  - Software RSA encryption is faster in terms of elapsed time (on large processors)
    - but hardware crypto saves CPU time
- **SW Crypto**
  - CPUtime /elapsed time is very dependent on CPU speed and utilization

SSL Performance Recommendations

- **Use SSL only if there is a need for**
  - If at least one of the following is required
    - Keeping secrets
    - Proving identity
    - Verifying information
- **Cipher Suites 01 and 02 has less CPU-time consumption, but NO data encryption**
  - RSA512_NULL_MD5, RSA512_NULL_SHA
- **If data encryption is required**
  - Use cipher suites 08, 09 or 0A
  - 08 uses 512 bit keys, others 1024
  - 1024 bit RSA key length is recommended (from a security point of view)

Turbo Dispatcher - Overview

- **Turbo Dispatcher**
  - available since 1995
  - VSE/ESA 2.1-2.3 Standard and Turbo Dispatcher
  - since VSE/ESA 2.4 only Turbo Dispatcher
  - last changes:
    - VSE/ESA 2.6.2 (APAR DY45869)
    - VSE/ESA 2.7.0 (APAR DY45926)
  - Supports basic (native), LPAR and VM mode
  - Runs on Uni- and n-Way-procerssors
    - CPUs have "equal" rights
    - more than 3 CPUs are not recommended

Turbo Dispatcher - Overview (2)

- **IPL is done on 1 CPU only**
  - after IPL other CPUs can be started
  - CPUs can be started or stopped without re-IPL
  - at least 1 CPU (IPL CPU) must always be active

```
SYSDEF TD,START=n|ALL
SYSDEF TD,STOP=n|ALL
SYSDEF TD,STOPQ=n|ALL
 QUERY TD
```

Turbo Dispatcher - Quiesced CPUs

- **SYSDEF TD,STOPQ=n to set a CPU in quiesced mode**
  - Implemented for z/VM guest systems
    - Not started guest CPUs stop IOASSIST
    - STOPQ remains IOASSIST active, and avoids TD Overhead, (CPU will no longer participate in work unit selection)
    - quiesced CPUs will not process any work units
    - quiesced CPUs will not handle any interrupt
    - quiesced CPUs can be started with SYSDEF TD,START

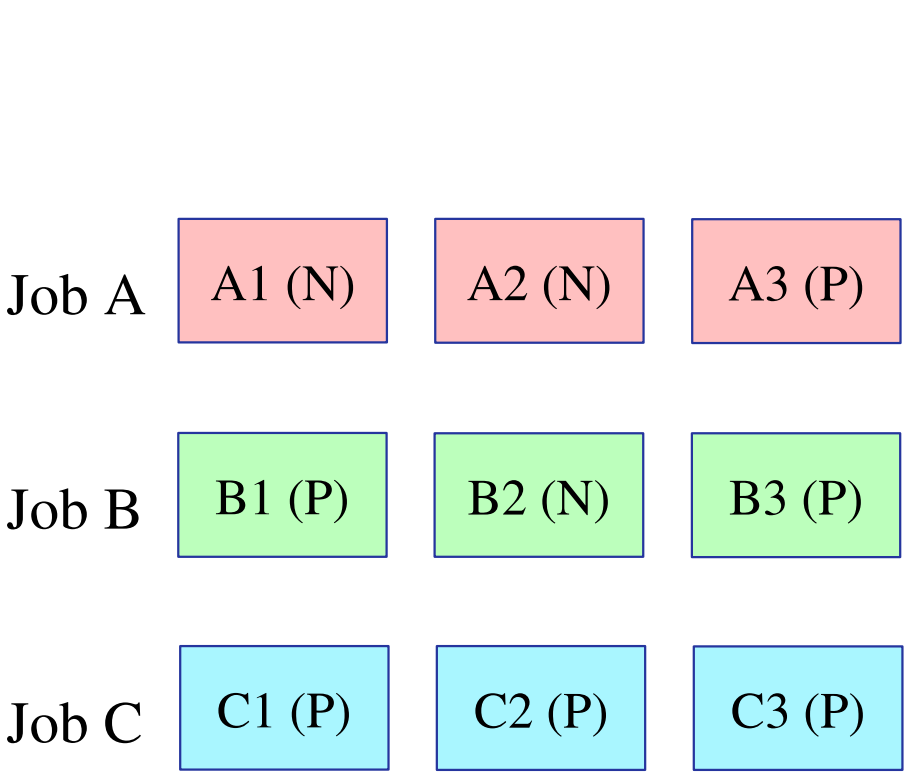
Turbo Dispatcher - Design

- **TD dynamically assigns partitions to CPUs**
  - Work unit = from assignment to one CPU until next interrupt/SVC
  - If one task (subtask) of a partition is active, no other task of the same partition will be selected
  - TD dispatches on partition-basis, not on task-basis
  - A job running in a partition is processed in several work units.

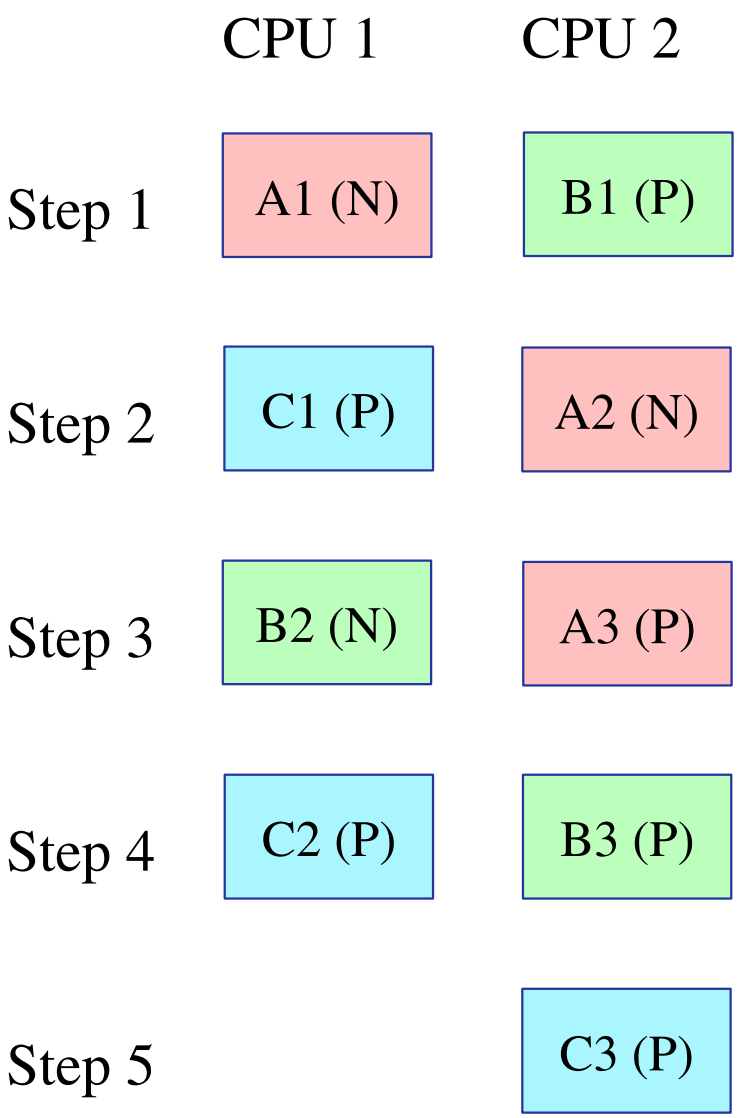
Turbo Dispatcher - Design (2)

- **parallel work units**
  - application code (CICS, Batch)
  - may run on any CPU concurrently with other parallel or non-parallel work units.
- **non-parallel work units**
  - system code (Services, VTAM, Vendor code)
  - 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.

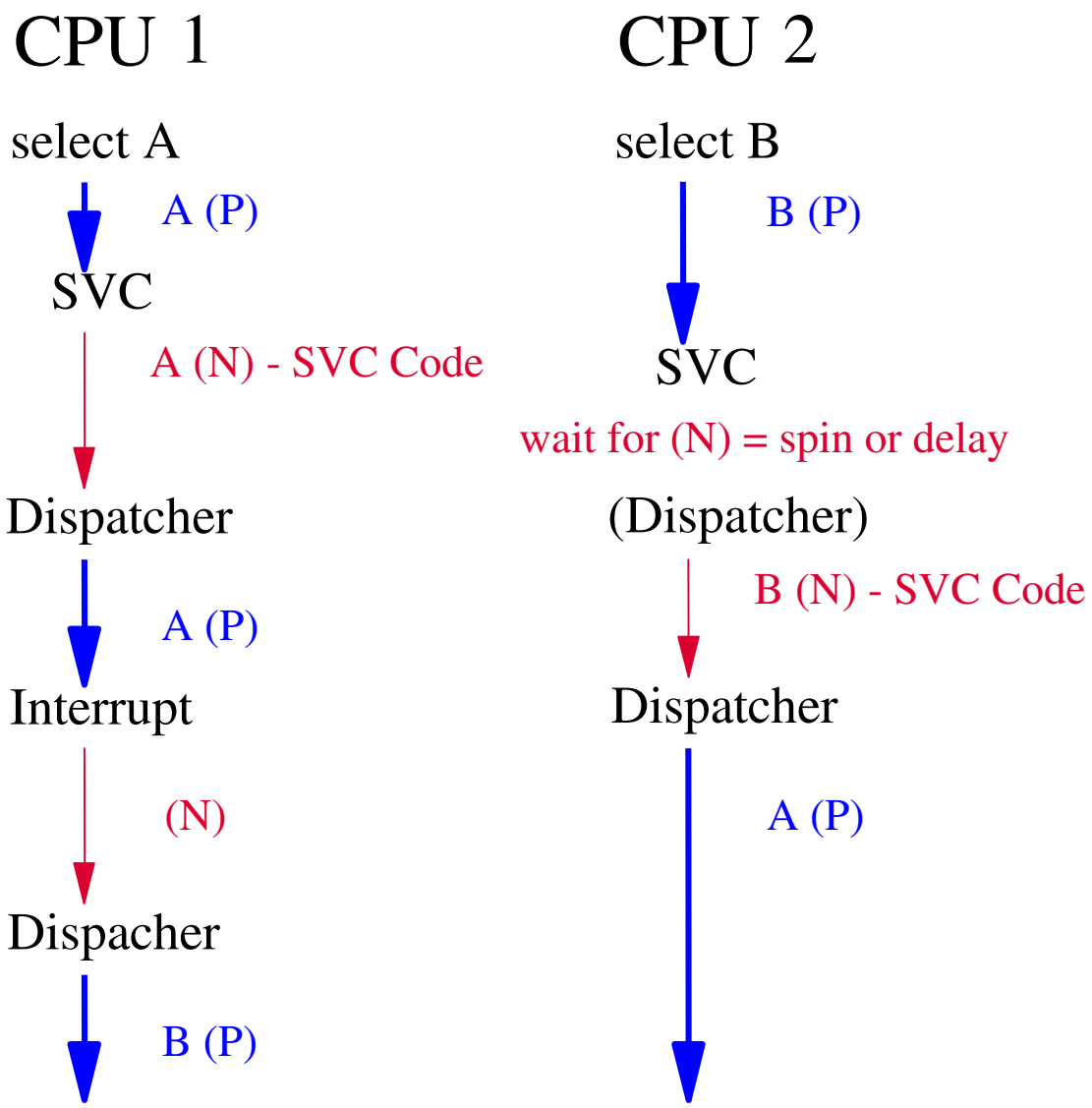
Turbo Dispatcher - Design - Example 1



Ax, Bx, Cx = workunits of job A, B, C
(N) = non-parallel work unit
(P) = parallel work unit



Turbo Dispatcher - Design - Example 2



Turbo Dispatcher - Exploitation

- **Uni-Processor**
  - new Partition Balancing Concept
    - Helps to set priorities of partitions
  - Determination of non-parallel share, to find out if a 2. or 3. CPU would be of use
- **n-Way Processors (2-3 CPUs)**
  - System tuning required for exploitation
  - Increased Capacity (dependent on workload)
    - Exploitation increases by reduction of non-parallel work units

Turbo Dispatcher - CPU time measurement

- **CPU time measurement (overall system)**
  - SYSDEF TD,RESETCNT
  - Workload (e.g. run a job)
  - QUERY TD (QUERY TD,INTERNAL)

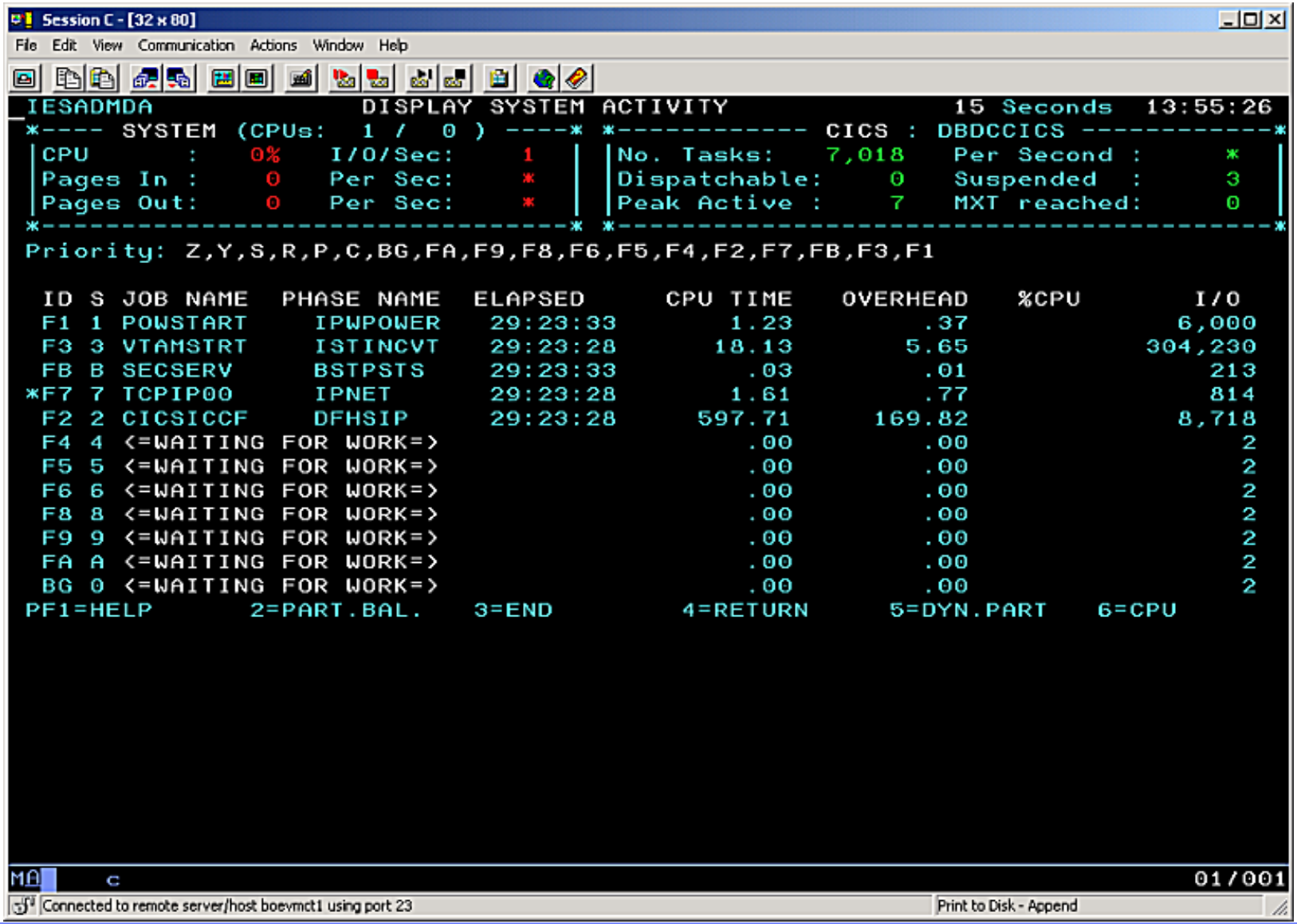
```
 CPU    STATUS     SPIN_TIME     NP_TIME TOTAL_TIME NP/TOT
  00    ACTIVE             0      237100     416698  0.568
  01    ACTIVE             0      157556     415229  0.379
  02    QUIESCED           0           0          0  *.***
  03    INACTIVE
                ---------------------------------------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

NP/TOT = non-paralell share (NPS)
SPIN_TIME = CPU time waiting for NP

Display System Activity Dialog



Migration aspects

- **Consider hard-/software requirements:**
  - Does my largest partition still fit into a single CPU of the target processor?
    - Note: a partition can only run on 1 CPU at a time!
  - Is the processor capacity and speed still sufficient to run the workload?
  - Does multiprocessing help to run the workload?
    - What about non-parallel share (on 1-Way)?
    - Are there many parallel batch jobs?
      - A large CICS partition does not benefit of a 2. CPU

Migration overhead

- **Uni-Processor**
  - increased overhead because of
    - Release migration (VSE/ESA 2.6 vs. 2.7)
    - TD overhead (Standard Dispatcher vs. TD)
    - CICS/VSE vs. CICS TS
- **N-Way Processor**
  - CPU time increases when migrating from uni to n-Way Processor (for the same workload)
    - For PACEX Workload: Factor 1.4 (2 CPUs)
    - TD overhead for multiprocessor exploitation
    - z/VM Overhead

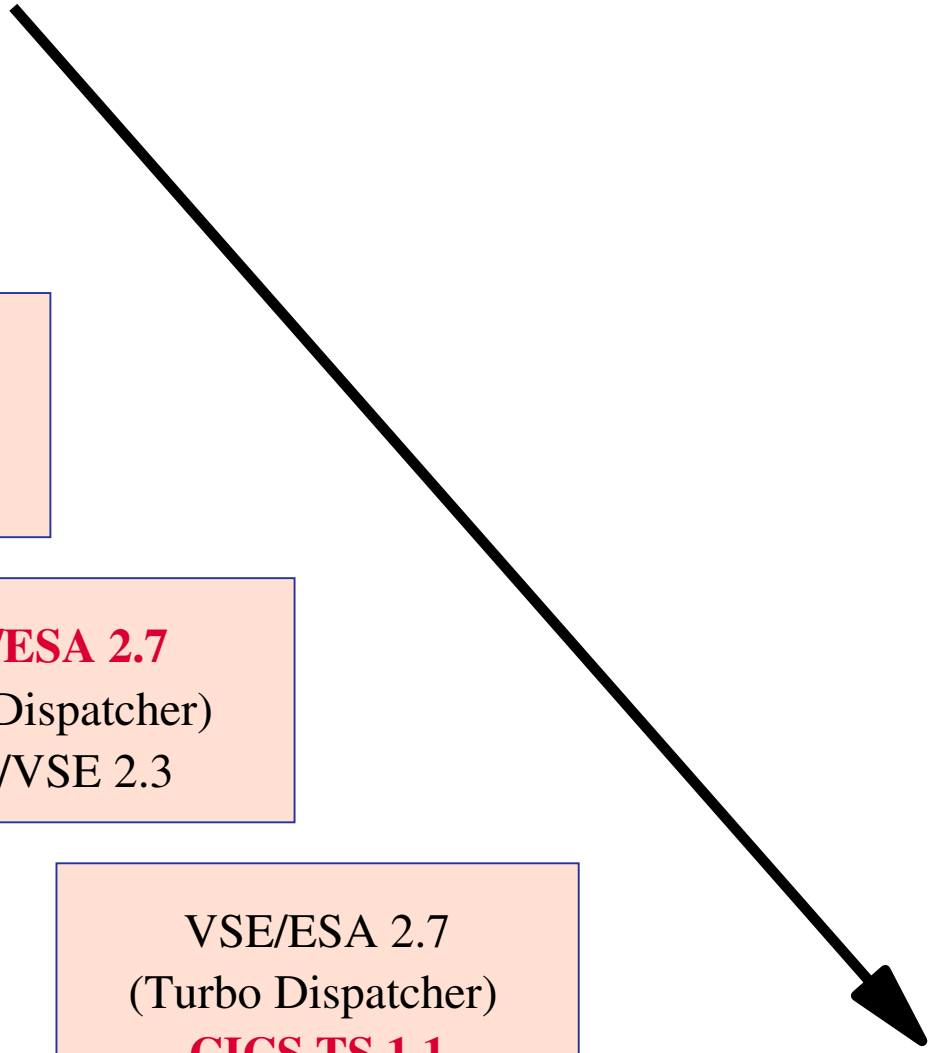
Migration path

VSE/ESA 2.3
Standard Dispatcher
CICS/VSE 2.3

VSE/ESA 2.3
Turbo Dispatcher
CICS/VSE 2.3

VSE/ESA 2.7
(Turbo Dispatcher)
CICS/VSE 2.3

VSE/ESA 2.7
(Turbo Dispatcher)
CICS TS 1.1



Change only
one thing at a time!

Allows you to see which step
has introduced a problem.

Performance Tips

- **A partition can only exploit 1 CPU at a time**
  - 2 CPUs do not have any benefit for a CICS partition
  - Use as many partitions as required for selected n-way
- **Use/define only as many CPUs as really needed**
  - additional CPUs create more overhead, but no benefit
- **Partitions setup**
  - Set up more batch and/or (independent) CICS partitions
  - Split CICS production partitions into multiple partitions

Performance Tips (2)

- **1 CPU must be able to handle all non-parallel workload**
- **Non-parallel code limits the n-Way exploitation**
  - QUERY TD: NP/TOT = NPS
  - Measure NPS before migration
  - max CPUs = 0.9 / NPS

NPS	#CPUs	NPS	#CPUs
0.20	4.5 (4)	0.40	2.2 (2)
0.25	3.6 (3)	0.45	2.0 (2)
0.30	3.0 (3)	0.50	1.8 (1)
0.35	2.6 (2)	0.55	1.6 (1)

Performance Tips (3)

- **Non-parallel code limits the maximum MP exploitation**
- **System code (Key 0) increases non-parallel share**
  - Vendor code can have significant impact
- **Overhead increases when NP code limits throughput**
- **Data In Memory (DIM) reduces non-parallel code**
  - less system calls (I/Os)
  - may increase throughput
- **In general ONE faster CPU is better than multiple slower ones**
  - Even if sum of slower CPUs is higher than one faster CPU

CICS Implications

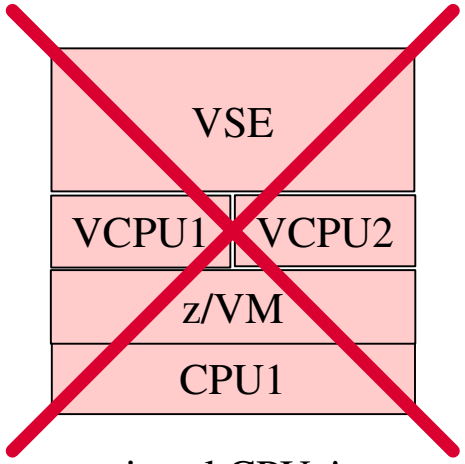
- **Single CICS**
  - Can consume processing power of one CPU only
  - parallel batch jobs may exploit 2. CPU
- **Multiple CICS partitions**
  - Number of CPUs depends on non-parallel share (NPS)
  - Function shipping and Transaction routing
    - AOR, TOR, FOR

Partition Balancing

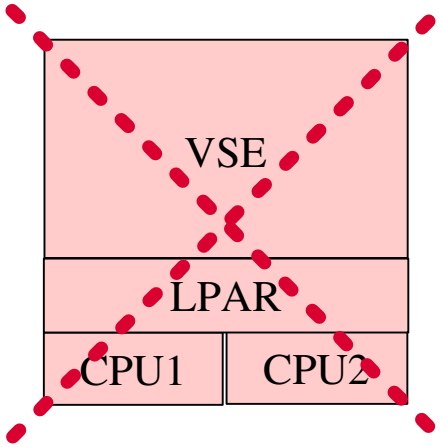
- **Balanced Group is defined with PRTY:**
  - PRTY BG,C=F5=F8,F2,F3,F1
  - Each partition/class of the group has a default-SHARE (100)
  - Dynamic partitions gets the SHARE of its class
- **To set a SHARE (1-1999)**
  - PRTY SHARE,F5=50
  - SHARE = 0 means the lowest priority within the group

```
PRTY
AR 0015 PRTY BG,C=F5=F8,F2,F3,F1
AR 0015
AR 0015 SHARE F5=  50, F8= 100,  C= 100
MSECS
AR 0015 MSECS    976      <---- influences task selection
```

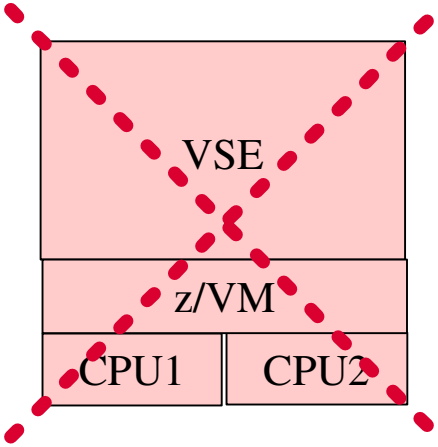
Do's and Don't Do's



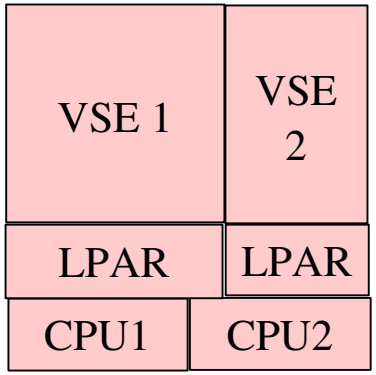
no virtual CPUs!
(creates overhead)



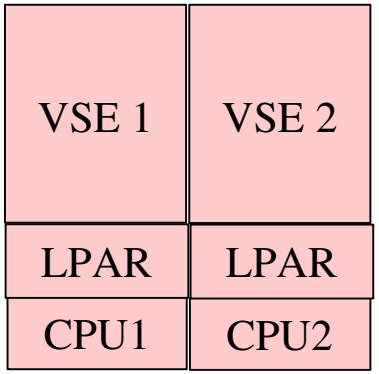
only if NPS < 4.5



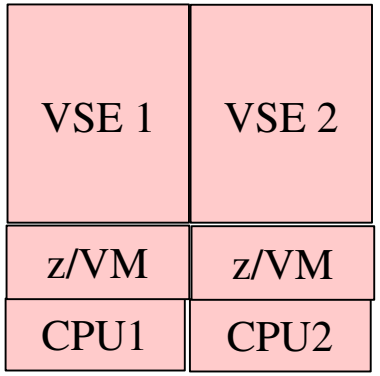
only if NPS < 4.5



VSE 1 = Production
VSE 2 = Test



dedicated CPU
per VSE

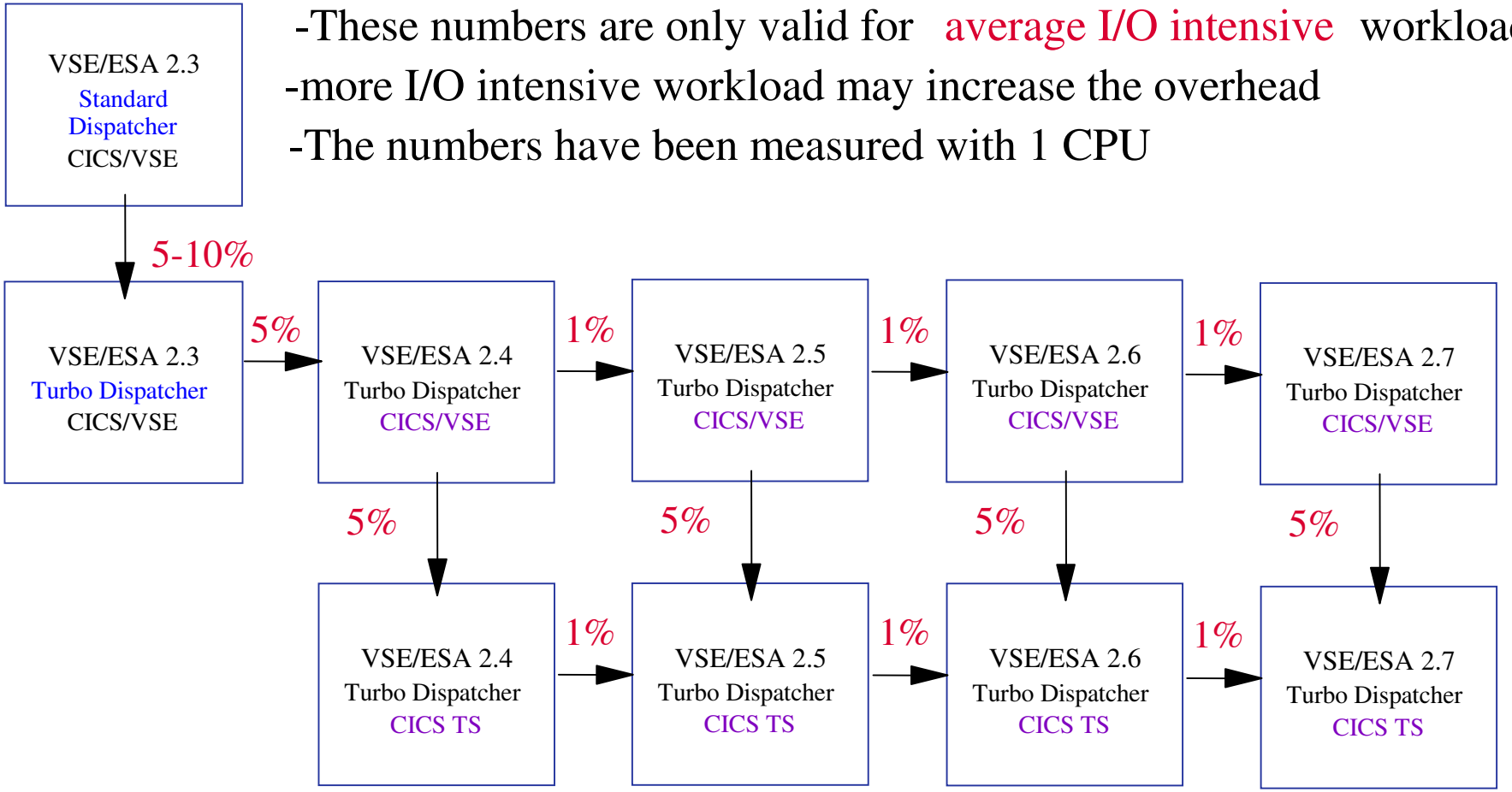


dedicated CPU
per VSE

Do's and Don't Do's (2)

The fastest uni-processor is (almost always) the best processor !

Overhead Deltas for VSE Releases



-These numbers are only valid for average I/O intensive workload

-more I/O intensive workload may increase the overhead

-The numbers have been measured with 1 CPU

New releases with new functions may increase the system overhead

BUT: Exploitataion of the new functions can increase the system throughput

VSE Health Check

- **Goals**
  - Recognize actual/upcoming problems
  - Optimize the system for new/current workload
- **A-B-C analysis**
  - A - concentrate on the essentials
    - 20 % work for 80 % results
  - B - more detailed analysis
    - 30 % work for 15 % results
  - C - analyze all details
    - 50 % work for 5 % results
- **A-B analysis takes about 2 days**
- **C analysis takes about 1 week**
- **Should be done about once a year**

VSE Health Check - continued

- **What should be checked?**
  - Processor (utilization, dispatching, z/VM, ...)
  - DASD, Tapes (I/O rate, cache, ...)
  - Network (network load, misrouted packets, ...)
  - System software
    - Turbo Dispatcher (PRTY, PRTY SHARE, ...)
    - VSAM (CA/CI sizes, share options, buffers, ...)
    - CICS (MXT, DSA/EDSA sizes, SOS, ...)
    - Storage Layout (GETVIS 24, SVA, partitions, DSPACE, ...)
    - VTAM (buffer pool)
    - POWER (DBLK, DBLKGP, ...)
    - LE runtime options (Heap size, ...)
  - Application software

Hints and Tips for Performance

- **Try to exploit Turbo Dispatcher functions**
  - Priority settings
  - Partition balancing
  - Partition balancing groups
- **Use as much data in memory (DIM) as possible**
  - CICS Shared Data Tables
  - Large/many VSAM Buffers (with buffer hashing)
  - Virtual Disks
- **Switch tracing/DEBUG off for production**

Hints and Tips for Connector and TCP/IP-Performance

- **Reduce amount of data transferred**
  - Transfer only data that is needed
  - Issue only requests that are needed
- **Use connection pooling**
  - Reduce overhead of connection establishment
- **Performance of connectors depends on**
  - Network performance
  - Performance of "server"
  - Performance of "client" or middle tier
- **Reduce misrouted packets**
- **Use a packet filter**
  - Unwanted packets increases TCP/IP and CPU load

Further Information

- **VSE Homepage:**
  - http://www.ibm.com/servers/eserver/zseries/os/vse/
- **VSE Performance Homepage: http://www.ibm.com/servers/eserver/zseries/os/vse/library/vseperf.htm**
- **Performance Documents from W. Kraemer**
  - available on the Performance Homepage

Questions ?

RETURN TO INDEX