Linux on System z Performance Update
Part 2: Networking and Crypto

Thomas Weber (tweber@de.ibm.com)

WAVV Conference 2009

Orlando, FL, May 15-19

© 2009 IBM Corporation

Trademarks

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

DB2*
DB2 Connect
DB2 Universal Database
e-business logo
IBM*
IBM eServer
IBM logo*
Informix®
System z
Tivoli*
WebSphere*
z/VM*
zSeries*
z/OS*
ECKD
Enterprise Storage Server®
FICON
FICON Express
HiperSocket
OSA
OSA Express

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

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

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

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

* 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.

© 2009 IBM Corporation

Agenda

- Networking and Crypto
  - General Networking Performance Considerations
  - Hipersockets Linux to z/OS – Recommendations
  - Networking benchmark results
  - Crypto - Linux SSL
  - CP Assist for Cryptographic Function (CPACF)
  - IPsec – example for Linux In-kernel Crypto

© 2009 IBM Corporation

General Network Performance Considerations (1)

- Which connectivity to use:
  - External connectivity:
    - LPAR: 10 GbE cards
    - z/VM: VSWITCH with 10GbE card(s) attached
    - z/VM: for maximum throughput and minimal CPU utilization attach OSA directly to Linux guest
  - Internal connectivity:
    - LPAR: HiperSockets for LPAR-LPAR communication
    - z/VM: VSWITCH for guest-guest communication
      - MTU 8992 is recommended
- For highly utilized network devices consider
  - to use z/VM VSWITCH with link aggregation (up to 8 OSA cards)
  - to use channel bonding
  - that channel bonding for high availability has low overhead

© 2009 IBM Corporation

General Network Performance Considerations (2)

- If network performance problems are observed, the buffer count can be increased up to 128
  - the default inbound buffer count is 16
  - check actual buffer count with **`lsqeth -p`** command
  - we observed that the default of 16 limits the throughput of a HiperSockets connection with 10 parallel sessions
  - a buffer count of 128 leads to 8MB memory consumption
    - one buffer consists of 16x4KB pages which yields 64KB, so 128x64KB=8MB
- Set the inbound buffer count in the appropriate config file:
  - **SUSE SLES10: in `/etc/sysconfig/hardware/hwcfg-qeth-bus-ccw-0.0.F200`**
    add `QETH_OPTIONS="buffer_count=128"`
  - **SUSE SLES11: in `/etc/udev/rules.d/51-qeth-0.0.f200.rules`**
    add `ACTION=="add", SUBSYSTEM=="ccwgroup", KERNEL=="0.0.f200", ATTR{buffer_count}="128"`
  - **Red Hat: in `/etc/sysconfig/network-scripts/ifcfg-eth0`**
    add `OPTIONS="buffer_count=128"`

© 2009 IBM Corporation

General Network Performance Considerations (3)

- Consider to switch on priority queueing if an OSA Express adapter in QDIO mode is shared amongst several LPARs. How to do:
  - **SUSE SLES10: in /etc/sysconfig/hardware/hwcfg-qeth-bus-ccw-0.0.F200 add QETH_OPTIONS="priority_queueing=no_prio_queueing:0"**
  - **SUSE SLES11: in /etc/udev/rules.d/51-qeth-0.0.f200.rules add ACTION=="add", SUBSYSTEM=="ccwgroup", KERNEL=="0.0.f200", ATTR{priority_queueing}="no_prio_queueing:0"**
  - **Red Hat: in /etc/sysconfig/network-scripts/ifcfg-eth0 add OPTIONS="priority_queueing=no_prio_queueing:0"**
- OSA card handles queue 0 first; default queue number is 2
- Note: Priority queueing on one LPAR may impact the performance on all other LPARs sharing the same OSA card.

© 2009 IBM Corporation

General Network Performance Considerations (4)

- Choose your MTU size carefully. Set it to the maximum size supported by all hops on the path to the final destination to avoid fragmentation.
  - Use **`tracepath destination`** to check MTU size
  - If the application sends in chunks of <=1400 bytes, use MTU 1492.
    - 1400 Bytes user data plus protocol overhead.
  - If the application is able to send bigger chunks, use MTU 8992.
    - Sending packets > 1400 with MTU 8992 will increase throughput and save CPU cycles.
- TCP uses the MTU size for the window size calculation, not the actual application send size.
- For VSWITCH, MTU 8992 is recommended.
  - Synchronous operation, SIGA required for every packet.
  - No packing like normal OSA cards.

© 2009 IBM Corporation

General Network Performance Considerations (5)

- System wide sysctl settings that can be changed for networking
- Temporarily by the sysctl command or permanently in the appropriate config file:
  - `/etc/sysctl.conf`
- Set the device queue length from the default of 1000 to at least 2500:
  - `sysctl -w net.core.netdev_max_backlog = 2500`
- Adapt the inbound and outbound window size to suit the workload.
  - Recommended values for OSA devices:
    - `sysctl -w net.ipv4.tcp_wmem="4096 16384 131072”`
    - `sysctl -w net.ipv4.tcp_rmem="4096 87380 174760”`
  - System wide window size applies to all network devices.
    - Applications can use setsockopt to adjust the window size.
      - has no impact on other network devices
  - >=10 parallel sessions benefit from recommended default and maximum window sizes
  - a big window size can be advantageous for up to 5 parallel sessions

© 2009 IBM Corporation

General Network Performance Considerations (6)

- Turn off Linux services that you don't need. Numerous daemons are started that might be unnecessary.
  - For example: Check whether SElinux and Linux Auditing are required for your business. If not, you can consider to turn them off.
    - `dmesg | grep -i "selinux"`
    - `chkconfig –list | grep -i "auditd"`
    - In /etc/zipl.conf
      - `parameters="audit_enable=0 audit=0 audit_debug=0 selinux=0 ....."`
    - `chkconfig --set auditd off`
  - Check for other running daemons
    - `chkconfig --list`
  - Other daemons which are not required under Linux on System z:
    - `alsasound`
    - `avahi-daemon`
    - `resmgr`
    - `postfix`

© 2009 IBM Corporation

General Network Performance Considerations (7)

- A controlled environment without an external network connection usually doesn't require a firewall.
  - The Linux firewall iptables can utilize just one CPU.
  - This a potential bottleneck and can serialize network traffic.
  - If there is no business need, switch off the firewall (iptables) which is enabled by default in Novell SUSE.
    - **`chkconfig –set SuSEfirewall2_init off`**
    - **`chkconfig -set SuSEfirewall2_setup off`**
    - **`chkconfig –set iptables off`**

© 2009 IBM Corporation

Networking - HiperSockets Linux to z/OS (1) Recommendations for z/OS **and** Linux

- Frame size and MTU size are determined by chparm parameter of the IOCDS.
  - MTU size = frame size – 8KB
- Select the MTU size to suit the workload. If the application is mostly sending packets < 8KB, an MTU size of 8KB is sufficient.
- If the application is capable of sending big packets, a larger MTU size will increase throughput and save CPU cycles.
- MTU size 56KB is recommended only for streaming workloads with packets > 32KB.

© 2009 IBM Corporation

Networking - HiperSockets Linux to z/OS (2) Recommendations for z/OS **and** Linux

- HiperSockets doesn't require checksumming because it is a memory-to-memory operation.
  - default is sw_checksumming
  - to save CPU cycles, switch checksumming off:
    - **SUSE SLES10: in /etc/sysconfig/hardware/hwcfg-qeth-bus-ccw-0.0.F200 add QETH_OPTIONS="checksumming=no_checksumming"**
    - **SUSE SLES11: in /etc/udev/rules.d/51-qeth-0.0.f200.rules add ACTION=="add", SUBSYSTEM=="ccwgroup", KERNEL=="0.0.f200", ATTR{checksumming}="no_checksumming"**
    - **Red Hat: in /etc/sysconfig/network-scripts/ifcfg-eth0 add OPTIONS="checksumming=no_checksumming"**

© 2009 IBM Corporation

Networking - HiperSockets Linux to z/OS (3) Recommendations for z/OS

- Always set:
  - **`TCPCONFIG DELAYACKS`**
- If MTU size is 8KB or 16KB set:
  - **`TCPCONFIG TCPRCVBUFRSIZE 32768`**
  - **`TCPCONFIG TCPSENDBFRSIZE 32768`**
- If MTU size is 32KB set:
  - **`TCPCONFIG TCPRCVBUFRSIZE 65536`**
  - **`TCPCONFIG TCPSENDBFRSIZE 65536`**
- If MTU size is 56KB set:
  - **`TCPCONFIG TCPRCVBUFRSIZE 131072`**
  - **`TCPCONFIG TCPSENDBFRSIZE 131072`**
  - For HiperSockets MTU 56KB, the CSM fixed storage limit is important. The default is currently 120 MB and should be adjusted to 250MB if MSGIVT5592I is observed.

© 2009 IBM Corporation

Networking - HiperSockets Linux to z/OS (4) Recommendations for Linux

- The setting of rmem/wmem under Linux on System z determines the minimum/default/maximum window size which has the same meaning as buffer size under z/OS.
- Linux window size settings are system wide and apply to all network devices.
  - Applications can use setsockopt to adjust the window size individually.
    - has no impact on other network devices
  - HiperSockets and OSA devices have contradictory demands
    - > 10 parallel OSA sessions suffer from a send window size > 32KB
    - The suggested default send/receive window size for HiperSockets MTU 8KB and 16KB are also adequate for OSA devices.

© 2009 IBM Corporation

Networking - HiperSockets Linux to z/OS(5) Recommendations for Linux

- As a rule of thumb the default send window size should be twice the MTU size, e.g.:
  - MTU 8KB
    - `sysctl -w net.ipv4.tcp_wmem="4096 16384 131072`
    - `sysctl -w net.ipv4.tcp_rmem="4096 87380 174760`
  - MTU 16KB
    - `sysctl -w net.ipv4.tcp_wmem="4096 32768 131072`
    - `sysctl -w net.ipv4.tcp_rmem="4096 87380 174760`
  - MTU 32KB
    - `sysctl -w net.ipv4.tcp_wmem="4096 65536 131072`
    - `sysctl -w net.ipv4.tcp_rmem="4096 87380 174760`
  - MTU 56KB
    - `sysctl -w net.ipv4.tcp_wmem="4096 131072 131072`
    - `sysctl -w net.ipv4.tcp_rmem="4096 131072 174760`

© 2009 IBM Corporation

Networking - Linux to z/OS, SAP recommendations

- The SAP Enqueue Server requires a default send window size of 4*MTU size.
- HiperSockets
  - SAP networking is a transactional type of workload with a packet size < 8KB. HiperSockets MTU 8192 is sufficient.
    - `sysctl -w net.ipv4.tcp_wmem="4096 32768 131072`
    - `sysctl -w net.ipv4.tcp_rmem="4096 87380 174760`
- OSA
  - If the SAP Enqueue Server connection is run over an OSA adapter, the MTU size should be 8192. If MTU 8992 is used, the default send window size must be adjusted to 4*8992.

© 2009 IBM Corporation

Network – benchmark description (1)

- AWM - IBM internal benchmark which simulates network workload
  - All tests are done with 10 simultaneous connections
  - Transactional Workloads – 2 types
    - RR – request/response
      - A connection to the server is opened once for a 5 minute timeframe
    - CRR – connect/request/response
      - A connection is opened and closed for every request/response
    - **RR 1/1** (send 1 byte from client and server and get 1 byte response)
    - **RR 200/1000** (Send 200 bytes from client to server and get 1000 bytes response)
      - Simulating **online transactions**
    - **RR 200/32k** (Send 200 bytes from client to server and get 32k response)
      - Simulating **website access**
    - **CRR 64/8k** (Send 64 bytes from client to server and get 8k response)
      - Simulating **database query**
  - Streaming workloads – 2 types
    - **STRP** - "stream put" (Send 20MB to the server and get 20 bytes response)
    - **STRG** - "stream get" (Send 20 bytes to the server and get 20MB response)
      - Simulating **large file transfers**

© 2009 IBM Corporation

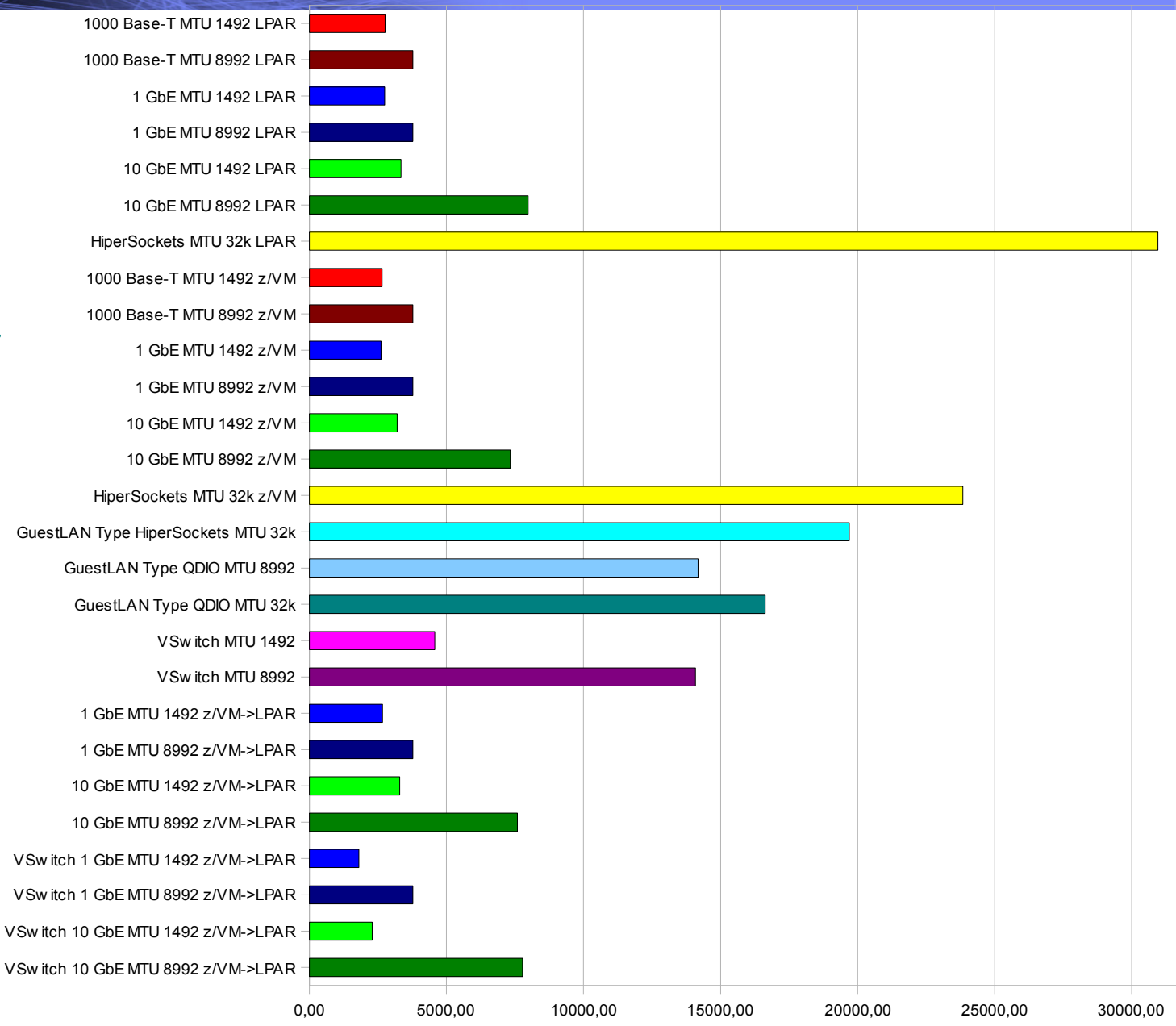
Network – benchmark description (2)

- Connection types
  - OSA 1000Base-T MTU, sizes 1492 and 8992
  - OSA 1 Gigabit Ethernet, MTU sizes 1492 and 8992
  - OSA 10 Gigabit Ethernet, MTU sizes 1492 and 8992
  - HiperSockets, MTU size 32k
  - GuestLAN type HiperSockets, MTU size 32KB (z/VM only)
  - GuestLAN type QDIO, MTU sizes 8992 and 32KB (z/VM only)
  - VSWITCH z/VM guest to guest, MTU sizes 1492 and 8992 (z/VM only)
  - OSA 1 Gigabit Ethernet dedicated z/VM guest - Linux LPAR, MTU sizes 1492 and 8992
  - OSA 10 Gigabit Ethernet dedicated z/VM guest - Linux LPAR, MTU sizes 1492 and 8992
  - OSA 1 Gigabit Ethernet VSWITCH z/VM guest - Linux LPAR, MTU sizes 1492 and 8992
  - OSA 10 Gigabit Ethernet VSWITCH z/VM guest - Linux LPAR, MTU sizes 1492 and 8992

© 2009 IBM Corporation

Database query

- SLES10 SP2 / z10
- 200 byte request
- 32k response
- 10 connections
- x-axis is number of transactions
- larger is better



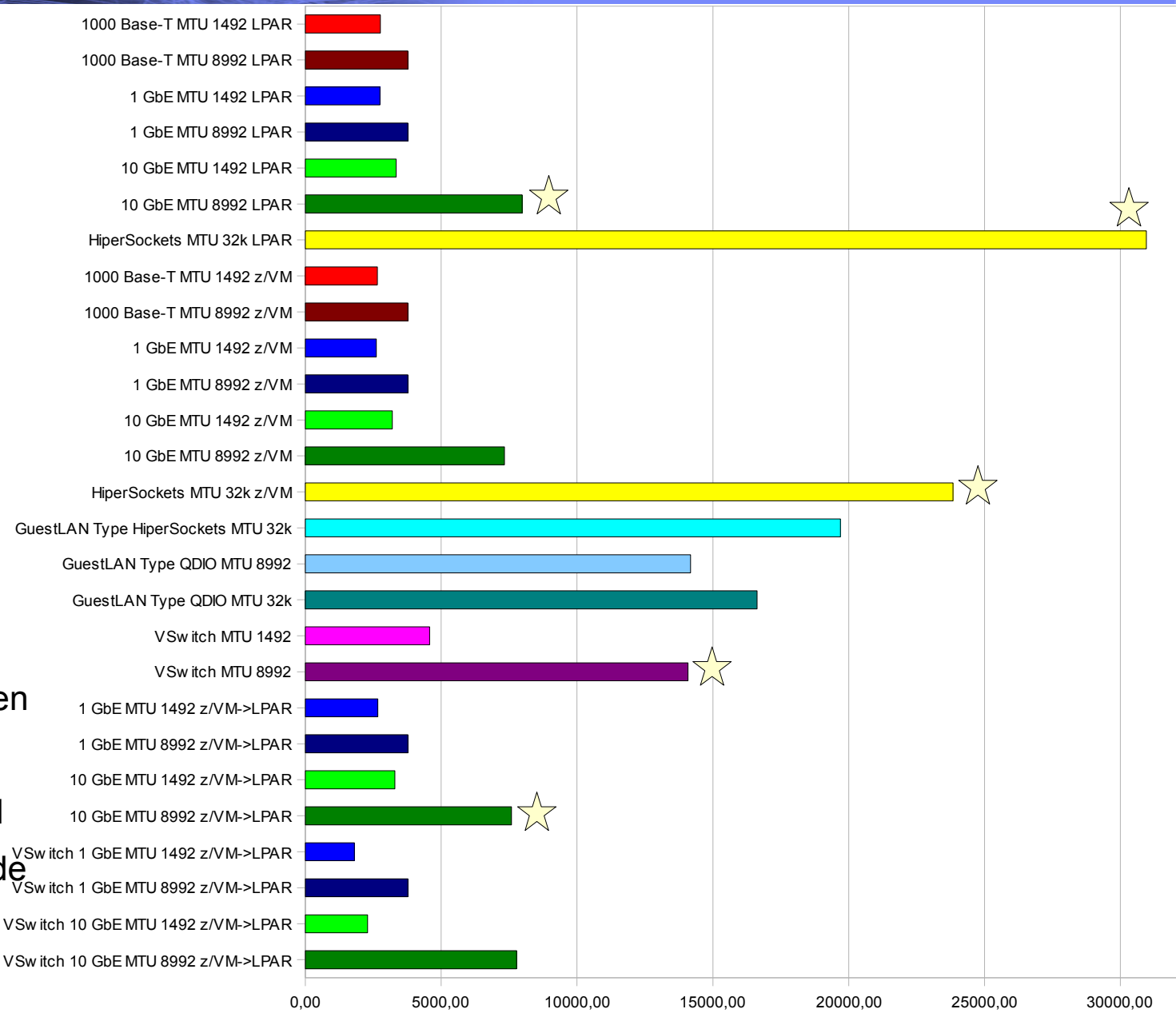
© 2009 IBM Corporation

Database query

- SLES10 SP2 / z10
- 200 byte request
- 32k response
- 10 connections
- x-axis is number of transactions
- larger is better

Recommendations:

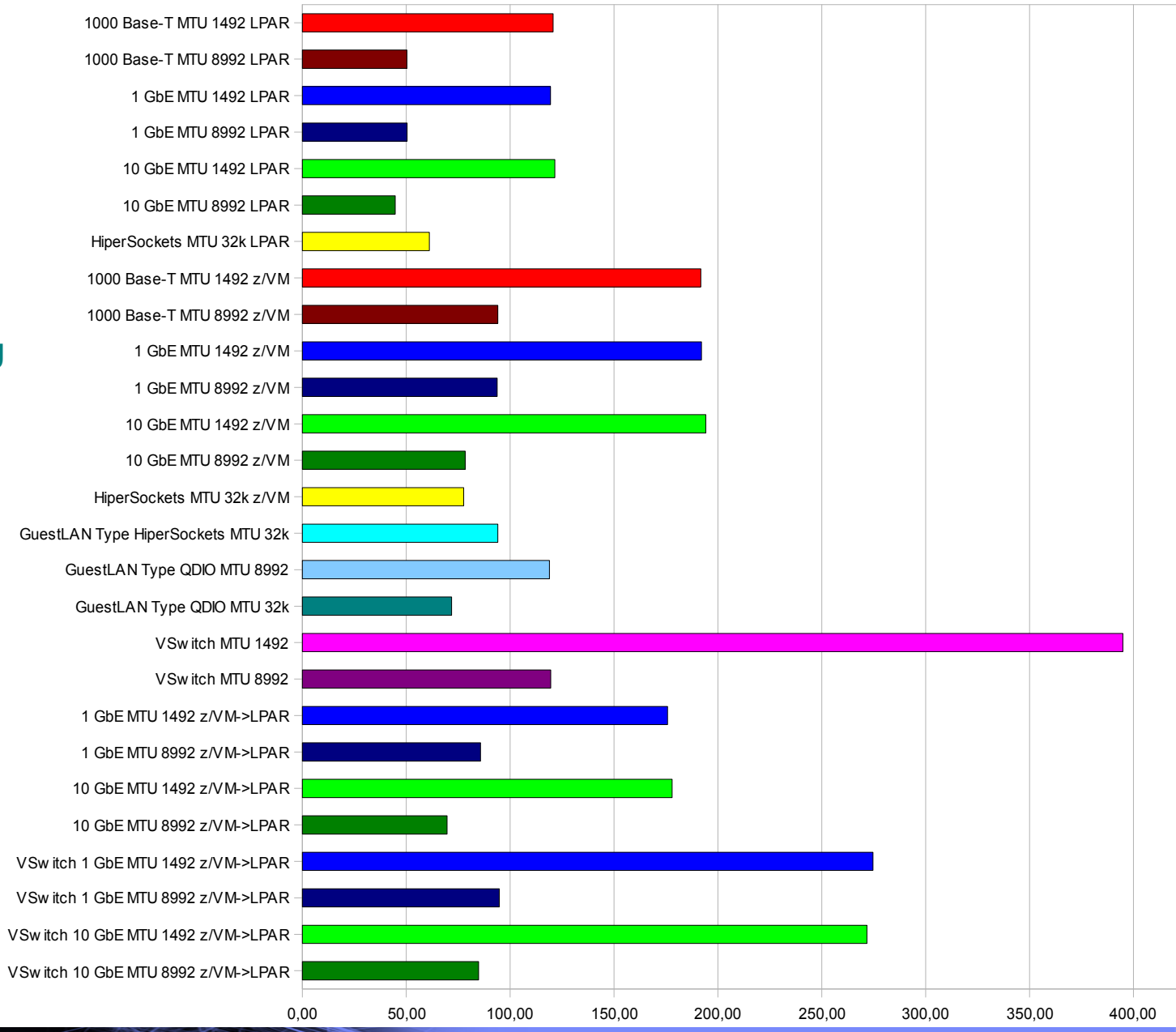
- Use large MTU size
- 10 GbE + MTU 8992
- HiperSockets between LPARs
- VSWITCH inside VM
- Direct OSA for outside connection of demanding guests



© 2009 IBM Corporation

Database query

- SLES10 SP2 / z10
- 200 byte request
- 32k response
- 10 connections
- x-axis is server CPU utilization per transaction
- smaller is better



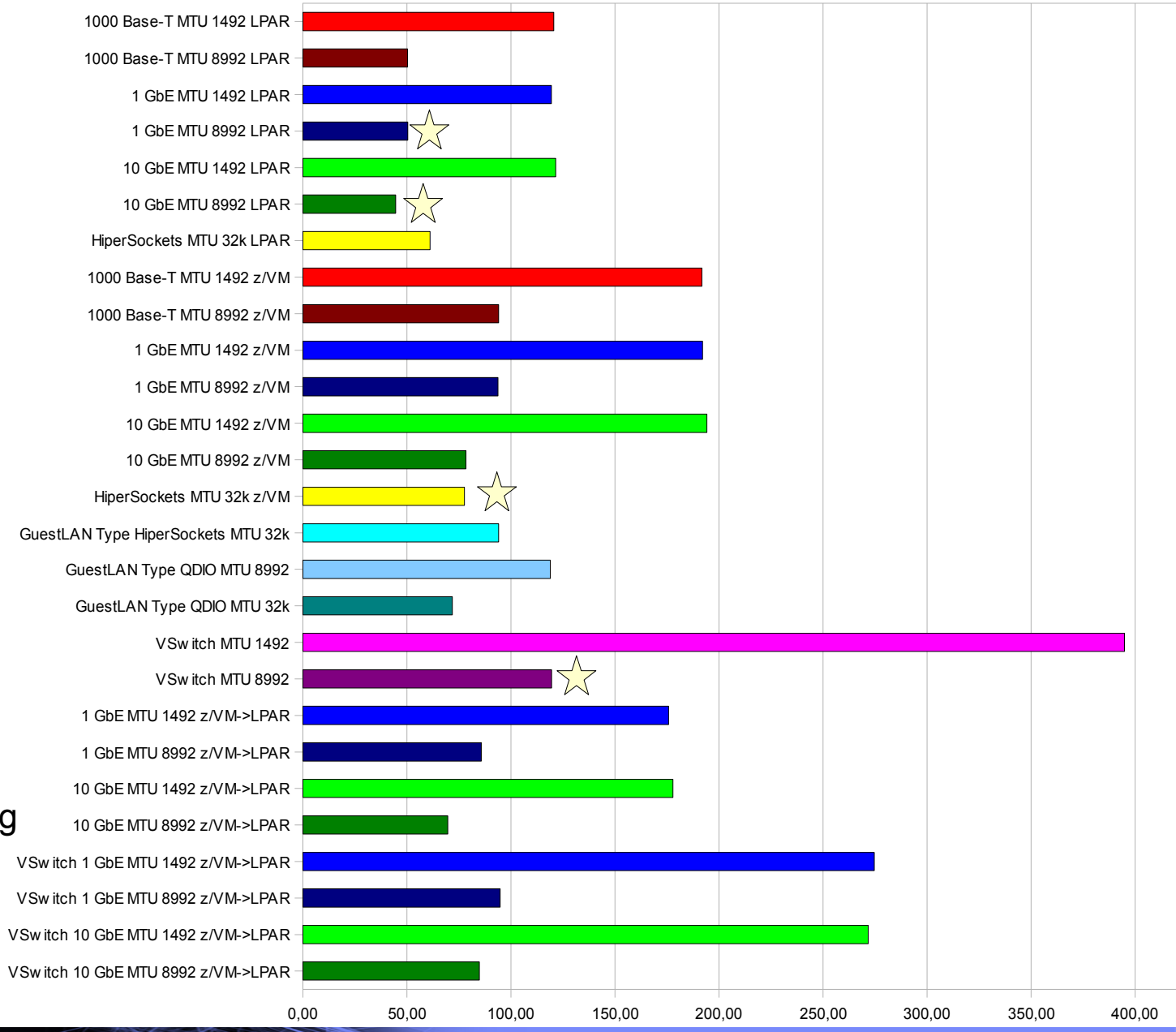
© 2009 IBM Corporation

Database query

- SLES10 SP2 / z10
- 200 byte request
- 32k response
- 10 connections
- x-axis is server CPU utilization per transaction
- smaller is better

Recommendations:

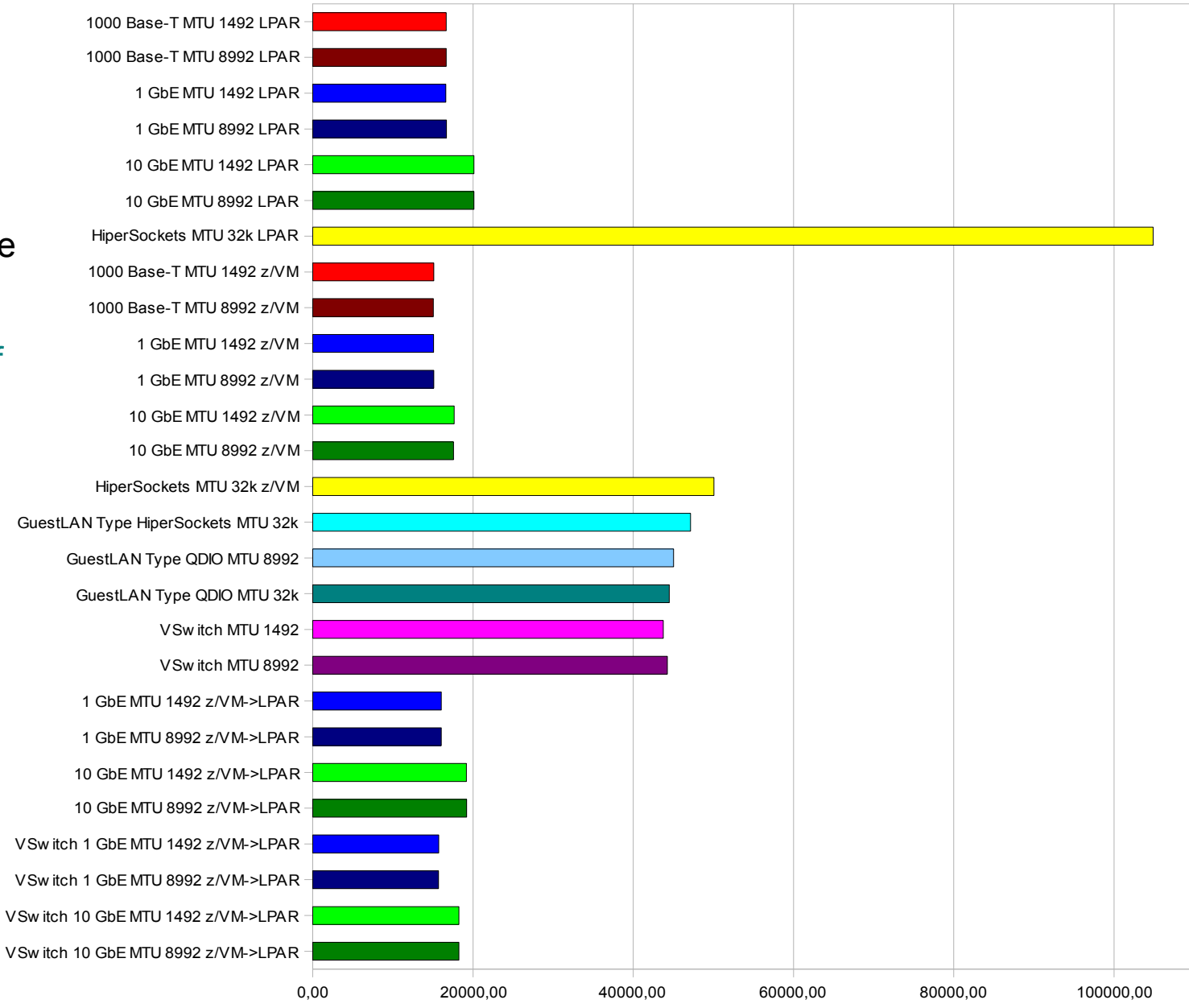
- Use large MTU size
- GbE + MTU 8992
- VSWITCH inside VM
- VSWITCH for outgoing connections has its price



© 2009 IBM Corporation

Online transaction

- SLES10 SP2 / z10
- 200 byte request
- 1000 byte response
- 10 connections
- x-axis is number of transactions
- larger is better



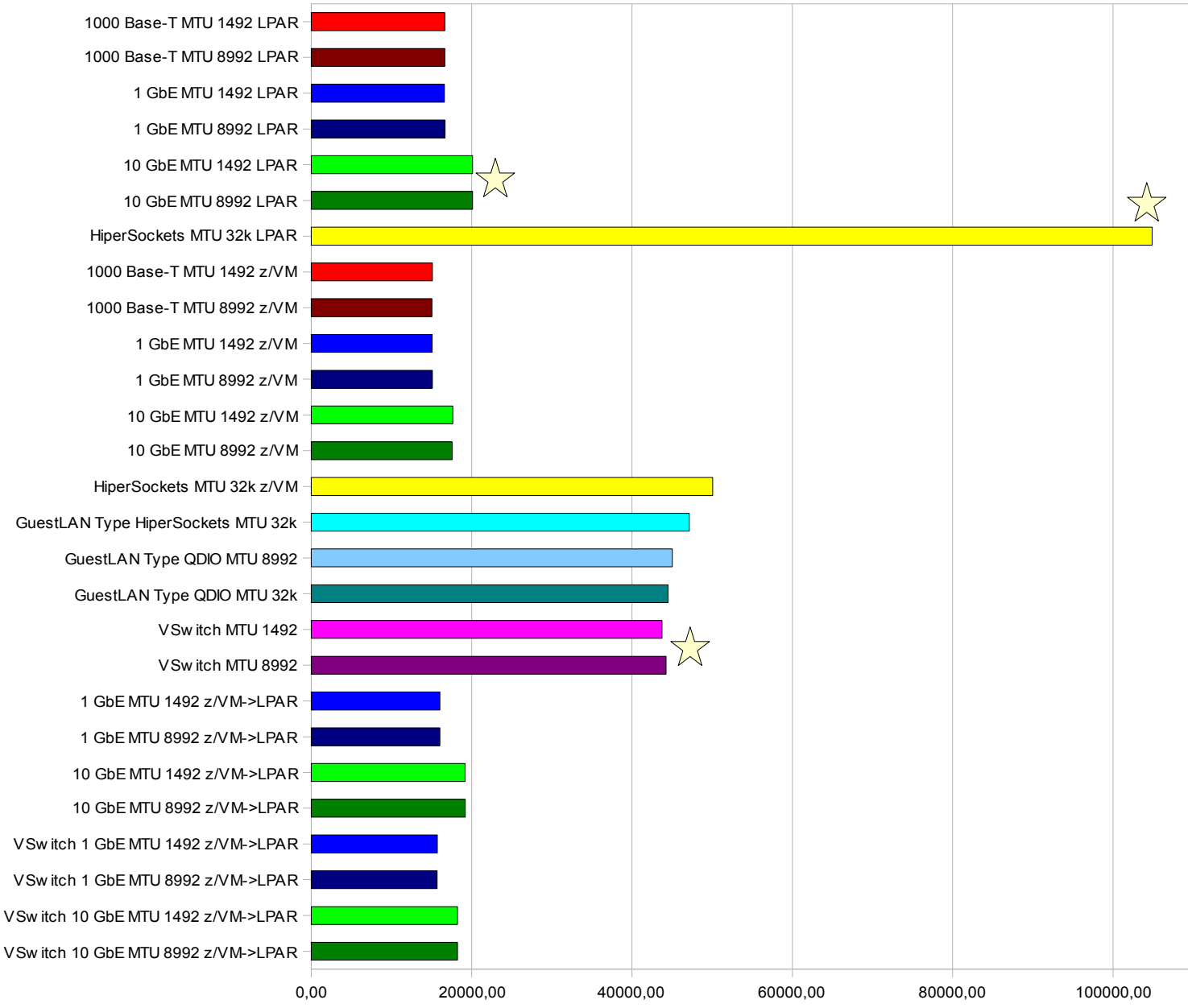
© 2009 IBM Corporation

Online transaction

- SLES10 SP2 / z10
- 200 byte request
- 1000 byte response
- 10 connections
- x-axis is number of transactions
- larger is better

Recommendations:

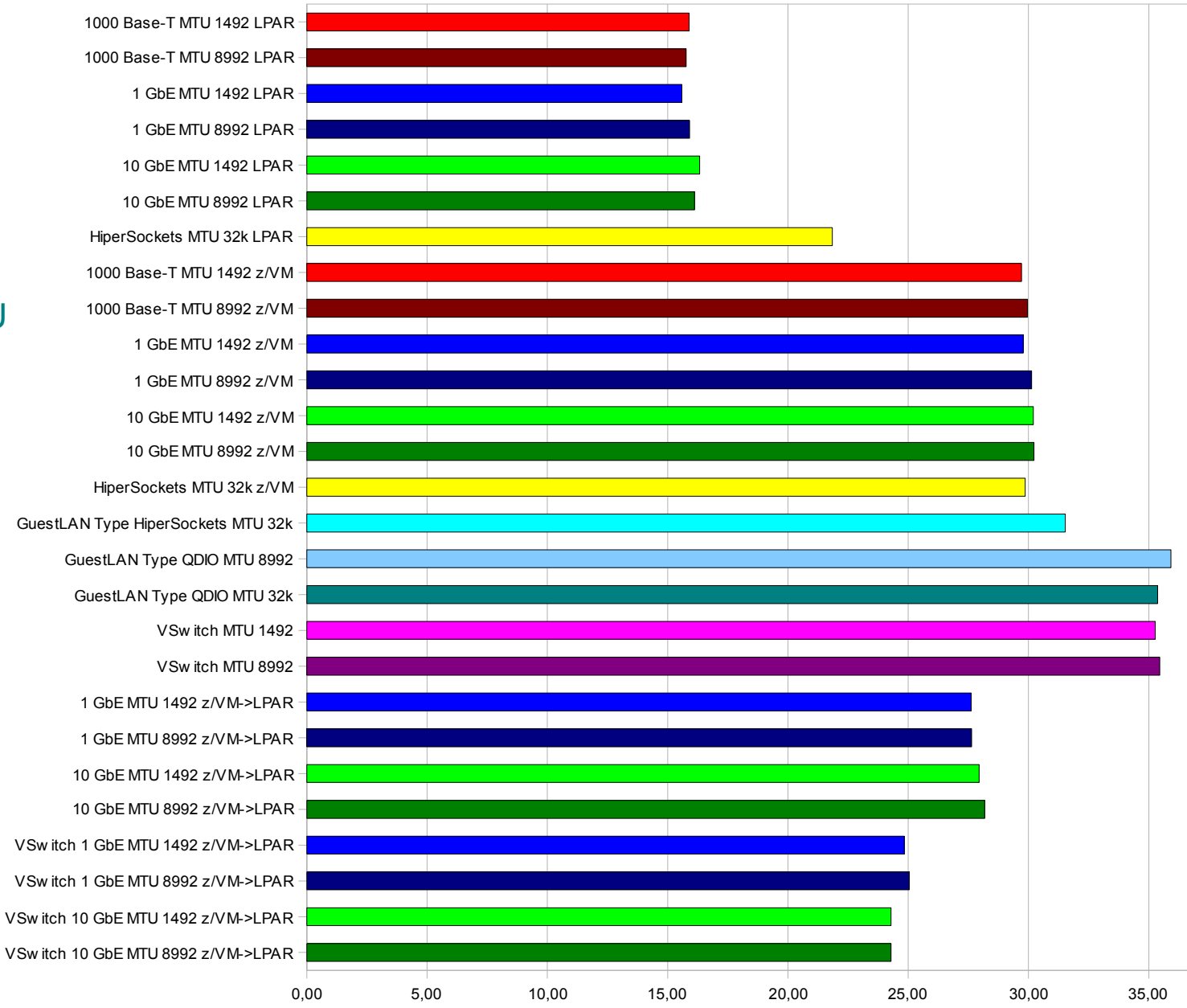
- 10 GbE better than 1 GbE
- HiperSockets between LPARs
- VSWITCH inside VM
- VSWITCH little bit slower for outside connections



© 2009 IBM Corporation

Online transaction

- SLES10 SP2 / z10
- 200 byte request
- 1000 byte response
- 10 connections
- x-axis is server CPU utilization per transaction
- smaller is better



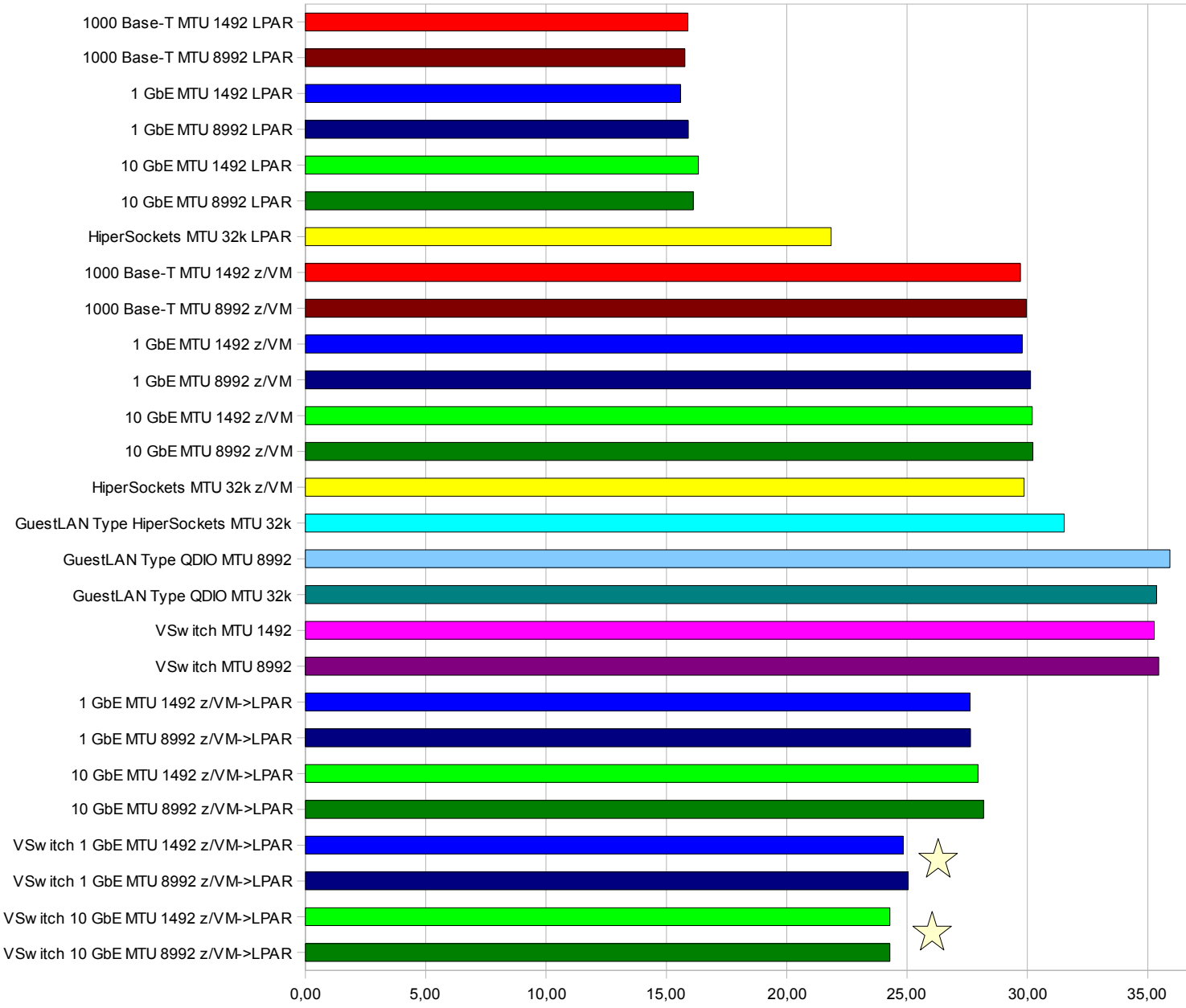
© 2009 IBM Corporation

Online transaction

- SLES10 SP2 / z10
- 200 byte request
- 1000 byte response
- 10 connections
- x-axis is server CPU utilization per transaction
- smaller is better

Recommendations:

- Internal connections are expensive
- MTU size doesn't make a difference
- VSWITCH best z/VM option for outside connection



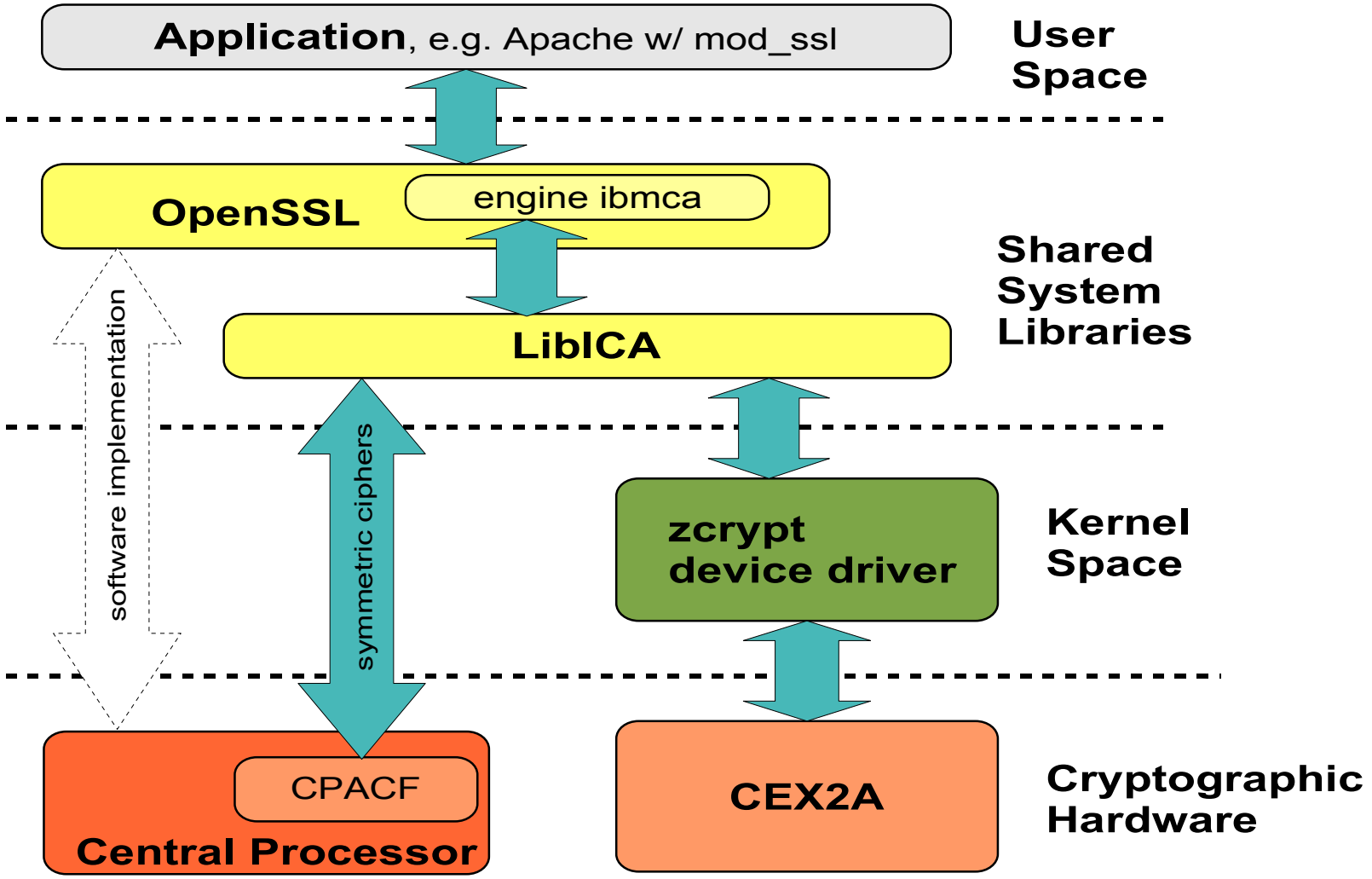
© 2009 IBM Corporation

Networking throughput overview

	Website access (crr64x8k)	Online transaction (rr200x1000)	Database query (rr200x32k)	File transfer (strp, strg 20Mx20)
Advantage of large MTU size over default MTU size	1.5x	equal	1.4x (1GbE), 2.3x (10GbE)	3.2x (only 10GbE)
Advantage of 10GbE over 1GbE	1.2x	1.2x	2.1x (large MTU)	3.4x (large MTU)
Advantage of virtual networks over OSA	2.5x (1GbE) 2.0x (10GbE)	2.8x (1GbE) 2.2x (10GbE)	3.7x (1GbE), 1.8x (10GbE)	4.8x (1GbE), 1.4x (10GbE)
Fastest connection	HiperSockets LPAR	HiperSockets LPAR	HiperSockets LPAR	HiperSockets LPAR

© 2009 IBM Corporation

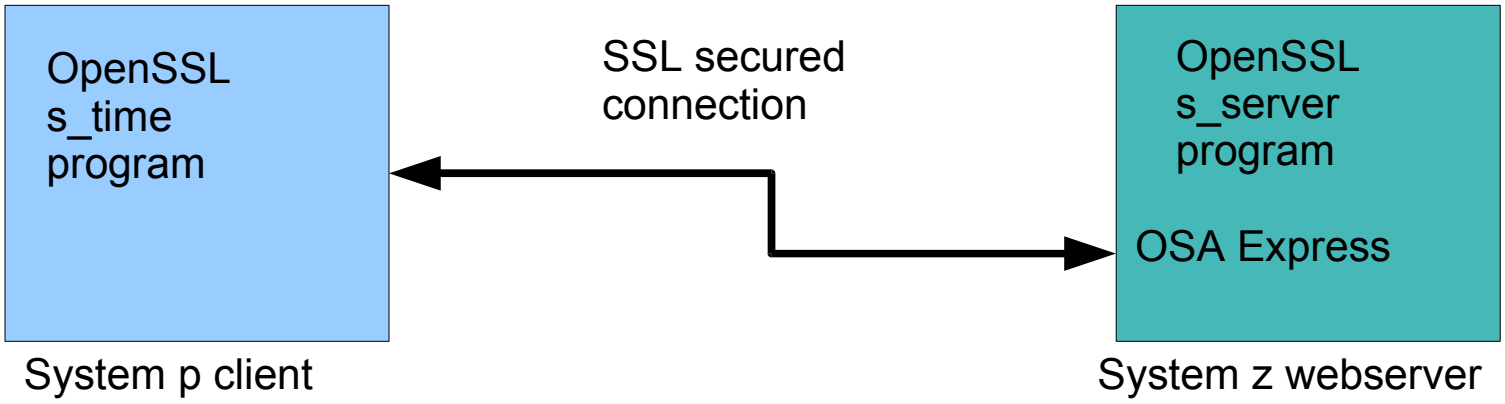
Cryptographic support – Linux SSL stack flow



© 2009 IBM Corporation

Cryptographic support – Exemplary Workload

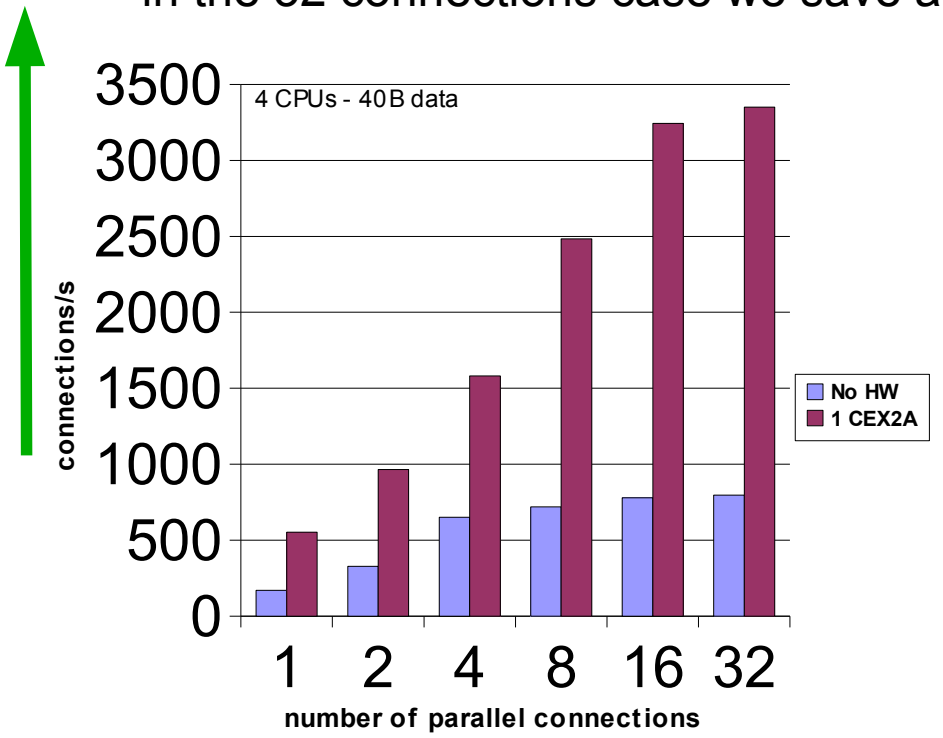
- workload emulates a secured webserver
- the connection between client and server is SSL secured
- scaling over number of parallel connections
- HTML files of different sizes are exchanged
  - 40 Bytes (SSL handshake)
  - 20 KB (small data portion)
  - 250 KB and 500KB (medium data portions)
  - 1 MB (big data portion)

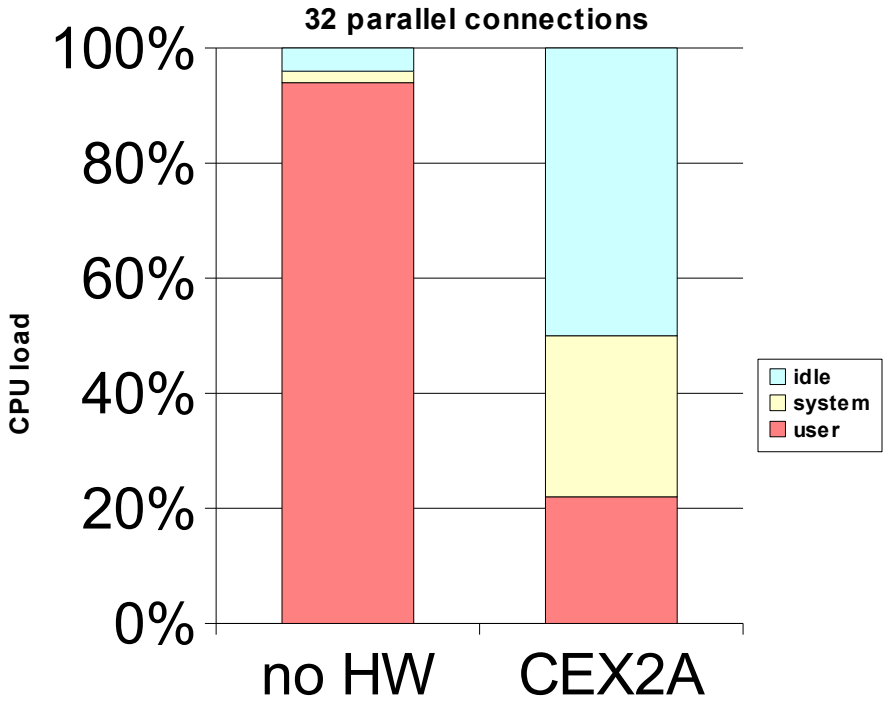


© 2009 IBM Corporation

Crypto Express2 Accelerator (CEX2A) - SSL handshakes

- CEX2A accelerates SSL handshake process (asymmetric cipher RSA)
- the number of SSL handshakes is up to 4x higher with CEX2A support
- in the 32 connections case we save about 50% of the CPU resources





© 2009 IBM Corporation

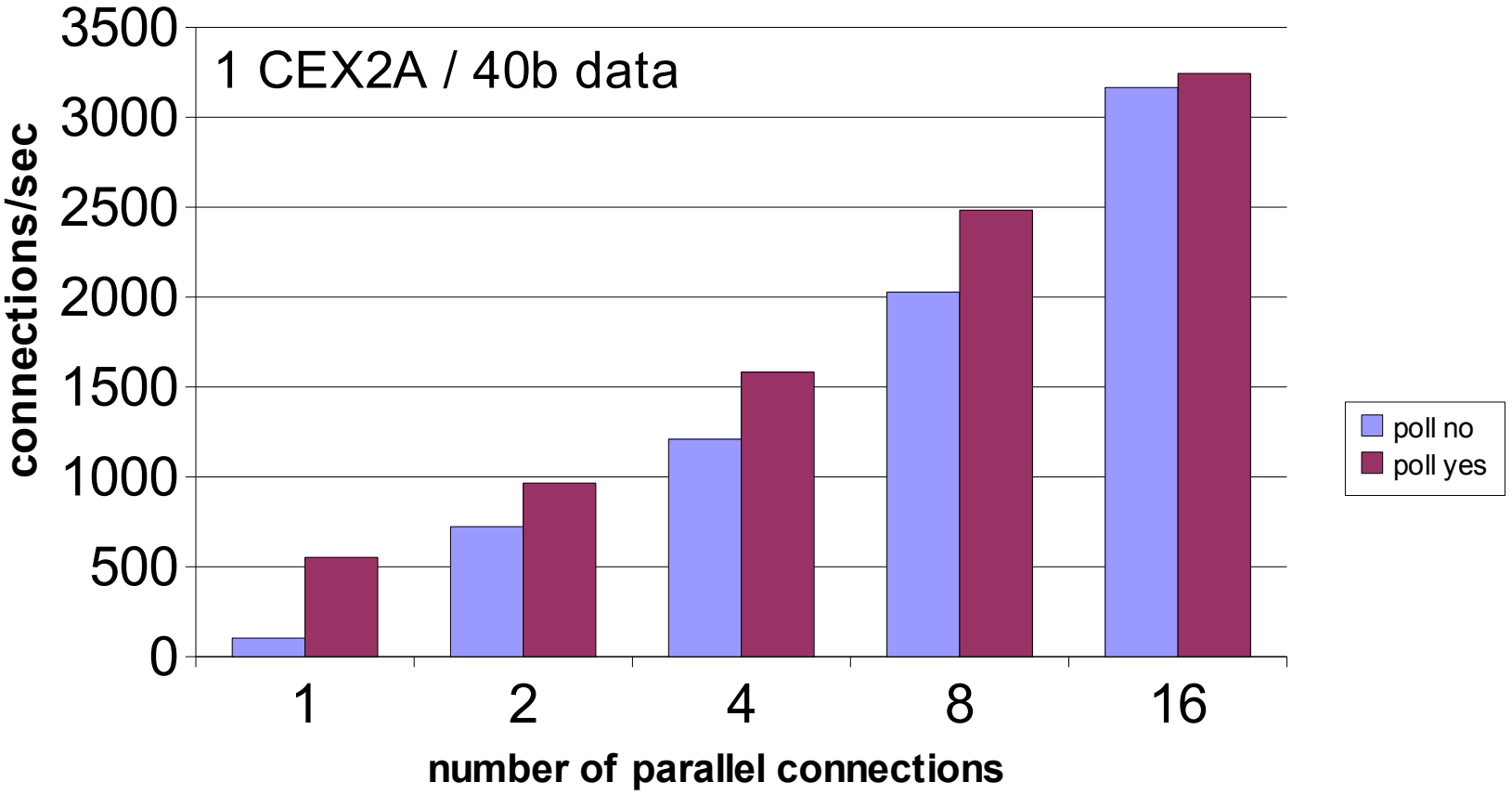
Generic cryptographic device driver polling thread (1)

- zcrypt device driver has a configurable polling thread
  - introduced with driver version 2.1.0 (SLES10 SP1 and RHEL 5.1); default was enabled
  - since SLES10 SP2 and RHEL 5.2 disabled per default
  - check state: `cat /sys/bus/ap/poll_thread` 1==enabled; 0==disabled
- **enabled**
  - polls cryptographic adapter for finished cryptographic requests
  - best utilization of CEX2 cryptographic adapter
  - uses one CPU for the thread when polling
  - only active during outstanding adapter requests
  - enable: `echo 1 > /sys/bus/ap/poll_thread`
- **disabled**
  - finished requests are fetched with Linux timer interrupt
  - poor performance when cryptographic adapter is not fully utilized
  - no further CPU costs for polling
  - disable: `echo 0 > /sys/bus/ap/poll_thread`

© 2009 IBM Corporation

Generic cryptographic device driver polling thread (2)

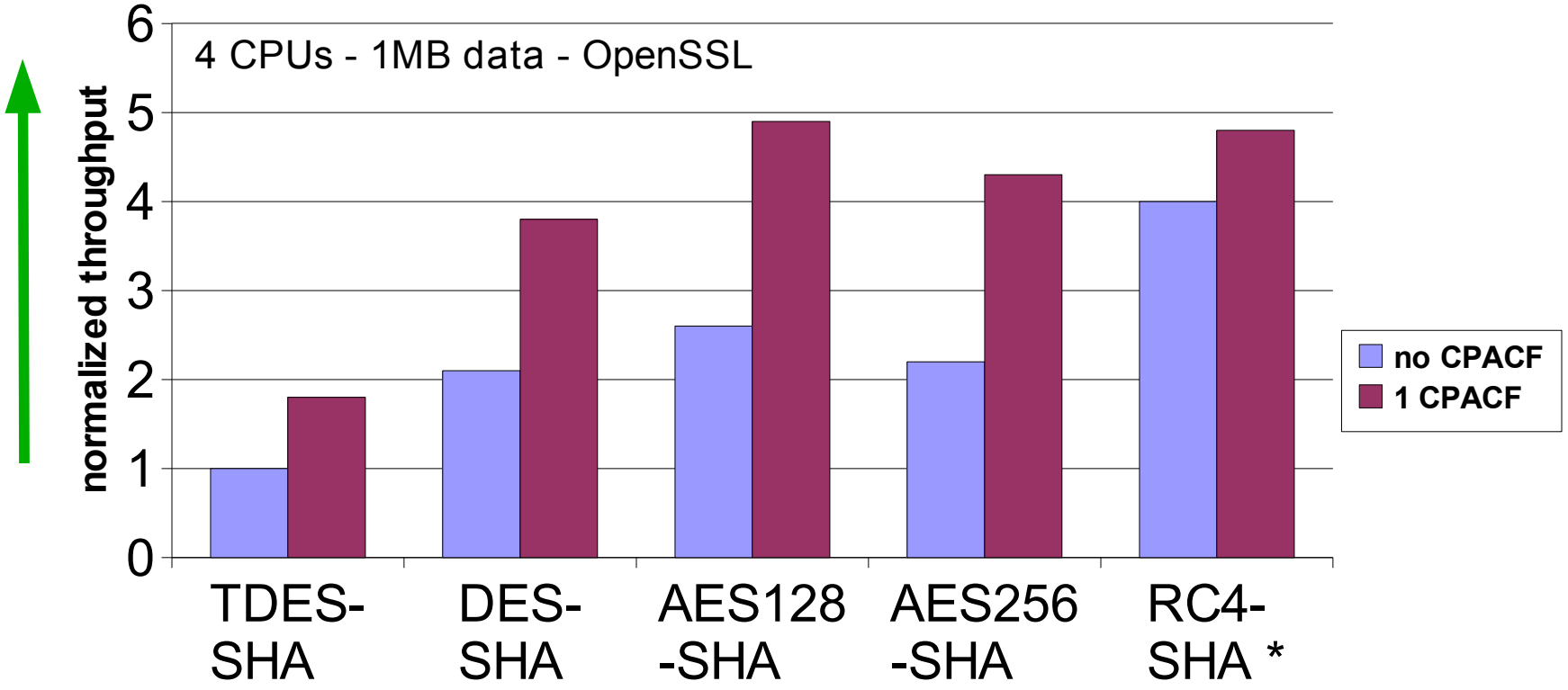
- performance degradations until adapter limit is reached for poll=no



© 2009 IBM Corporation

CP Assist for Cryptographic Function (CPACF) (1)

- z10 newly supports AES-192, AES-256
- use your cryptographic hardware! – software configuration issue

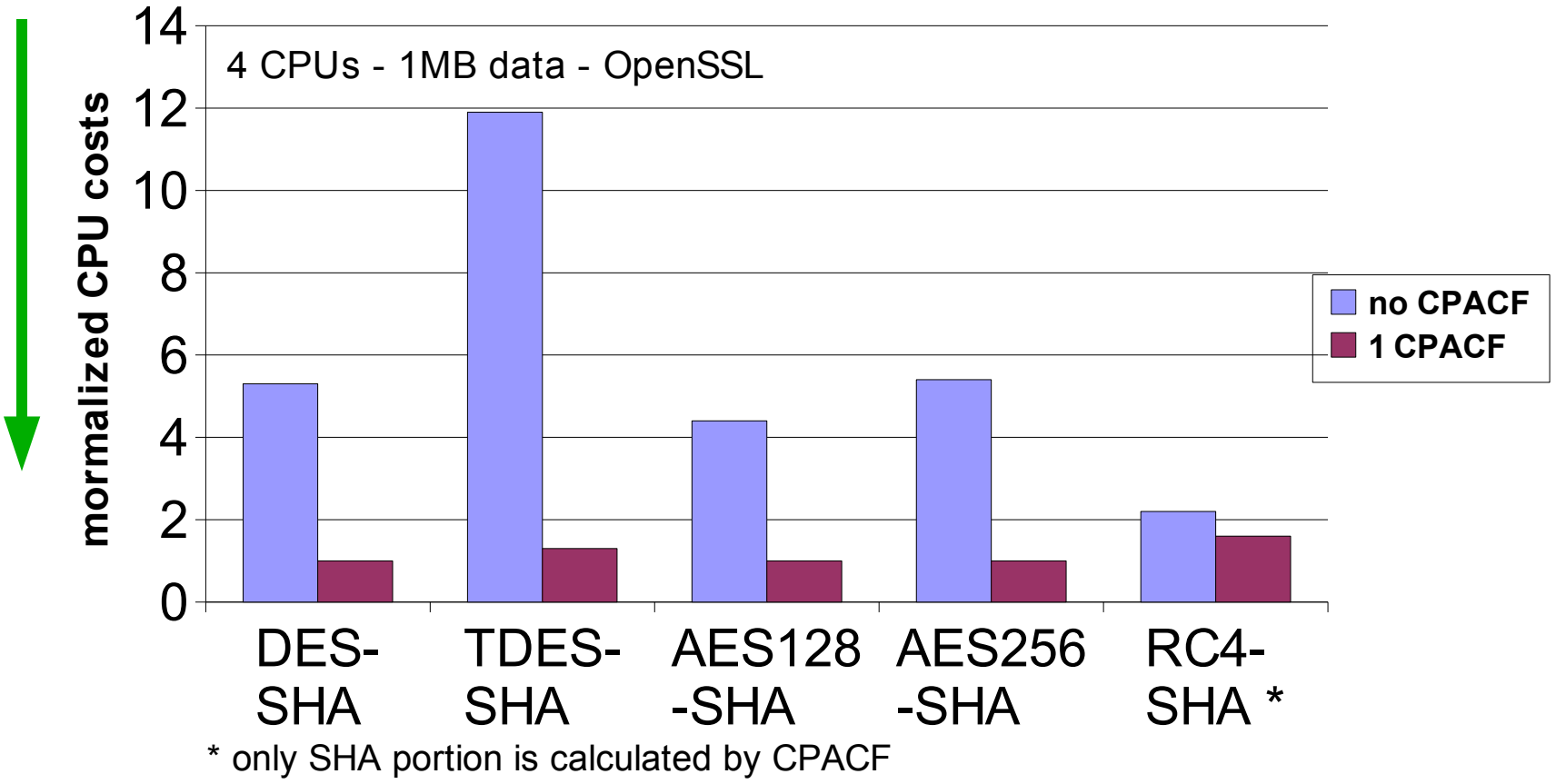


* only SHA portion is calculated by CPACF

© 2009 IBM Corporation

CP Assist for Cryptographic Function (CPACF) (2)

- reduced CPU costs for fully supported block ciphers
- TDES most expensive cipher when calculated in software



* only SHA portion is calculated by CPACF

© 2009 IBM Corporation

CP Assist for Cryptographic Function (CPACF) (3)

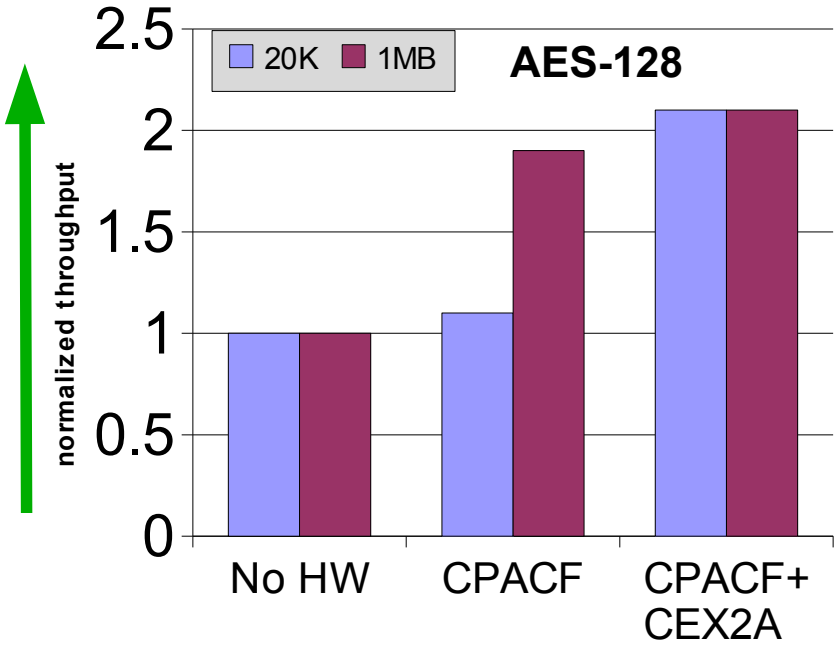
- supported ciphers and secure hash functions per System z machine

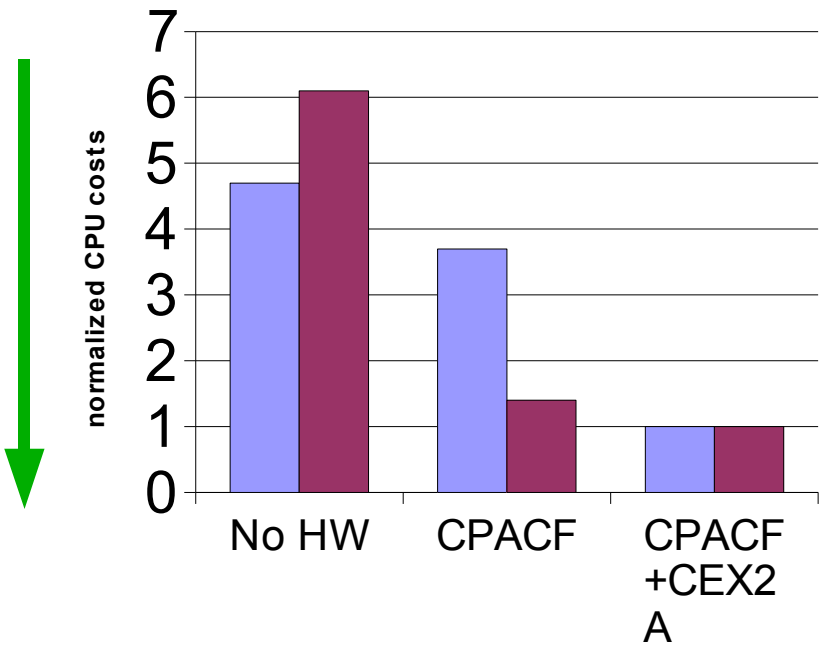
System z machine	supports
zSeries z890, z990	DES, TDES, SHA-1
z9	DES, TDES, AES-128, SHA-1, SHA-256
z10	DES, TDES, AES-128, AES-192, AES-256, SHA-1, SHA-224, SHA-256, SHA-384, SHA-512

© 2009 IBM Corporation

SSL traffic with CEX2A and CPACF

- CEX2A accelerates SSL handshakes
- CPACF accelerates data encryption (symmetric ciphers)
- use of both hardware features can double the throughput
- using pure software encryption costs up to 6x more CPU

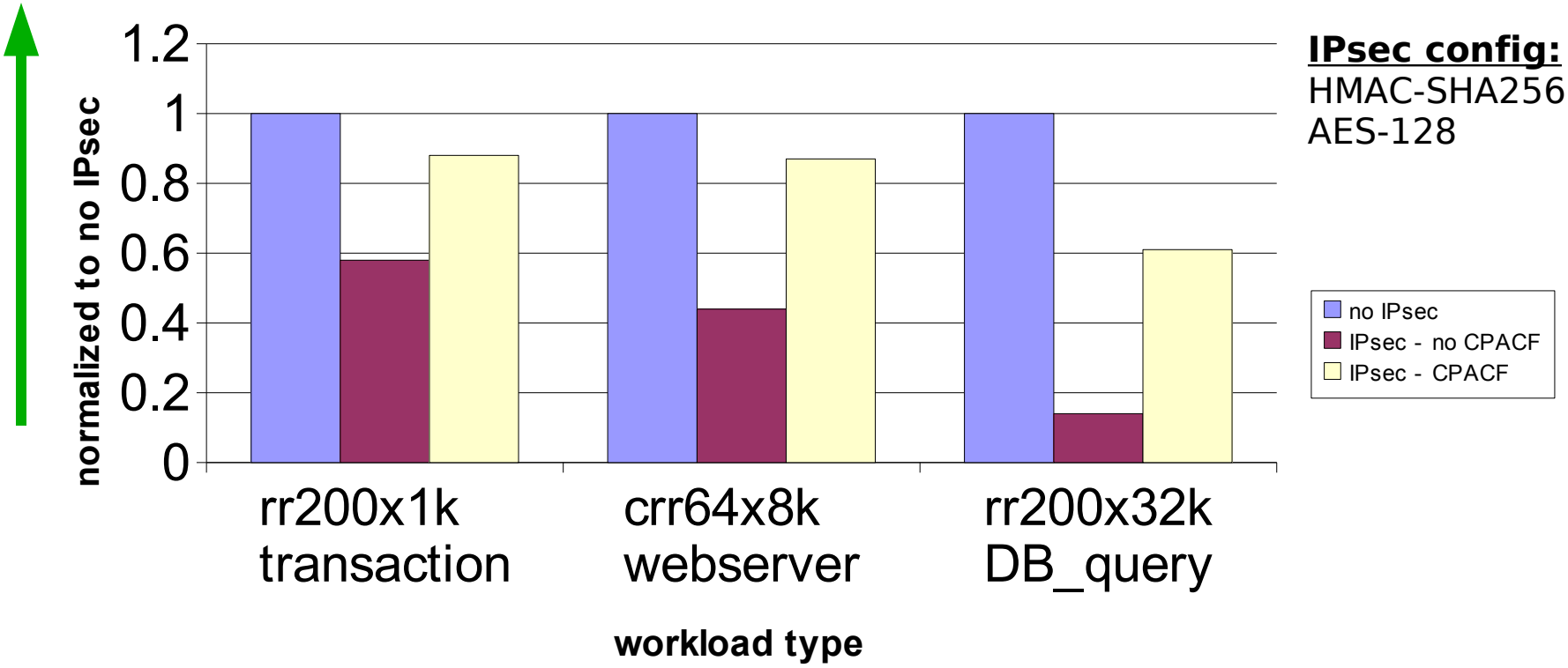




© 2009 IBM Corporation

Linux in-kernel crypto (1)

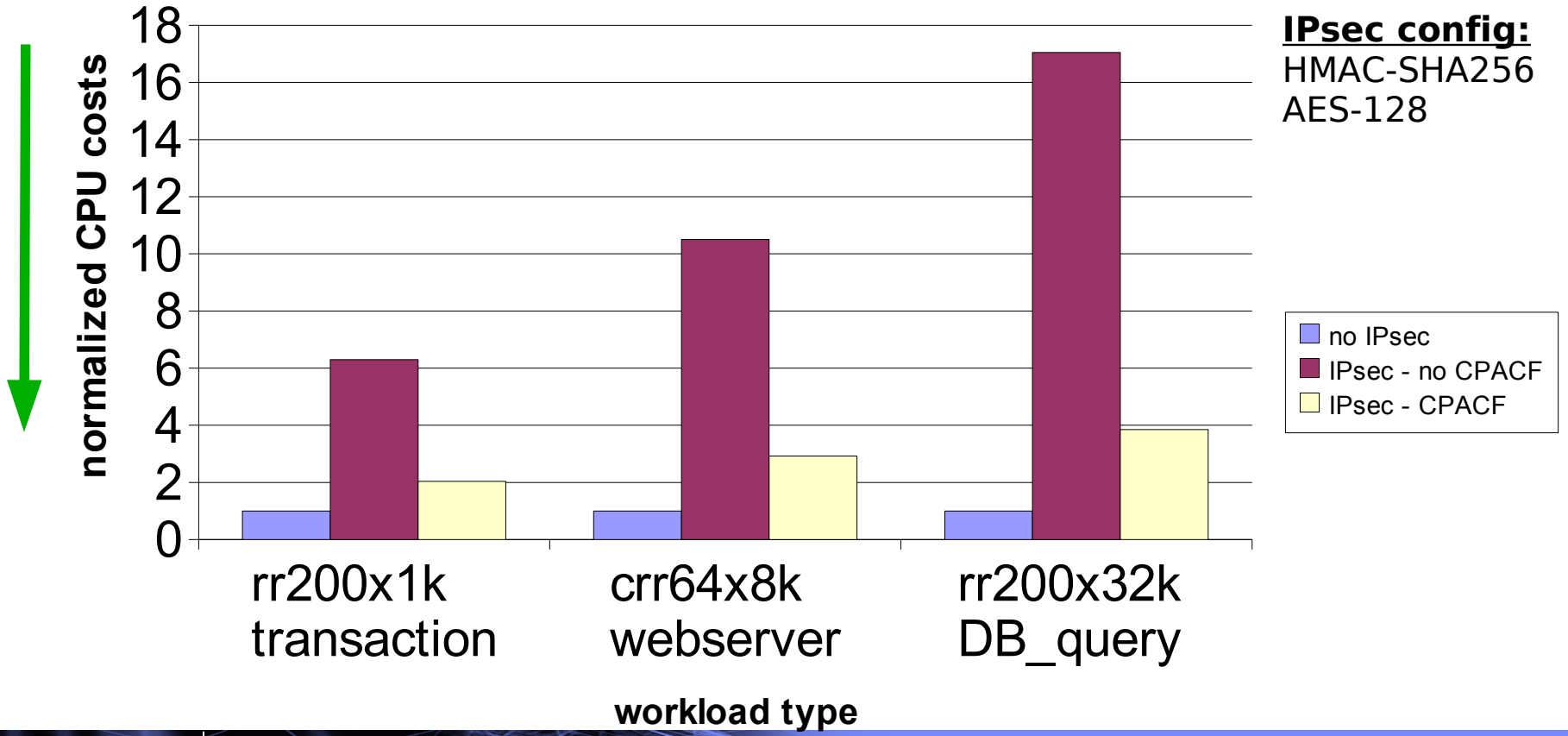
- example: IP security (IPsec) is done in the Linux kernel
- Linux kernel itself is capable of exploiting CPACF
- carefully choose a CPACF supported cipher and hash function
- strong performance impacts with pure software in-kernel crypto



© 2009 IBM Corporation

Linux in-kernel crypto (2)

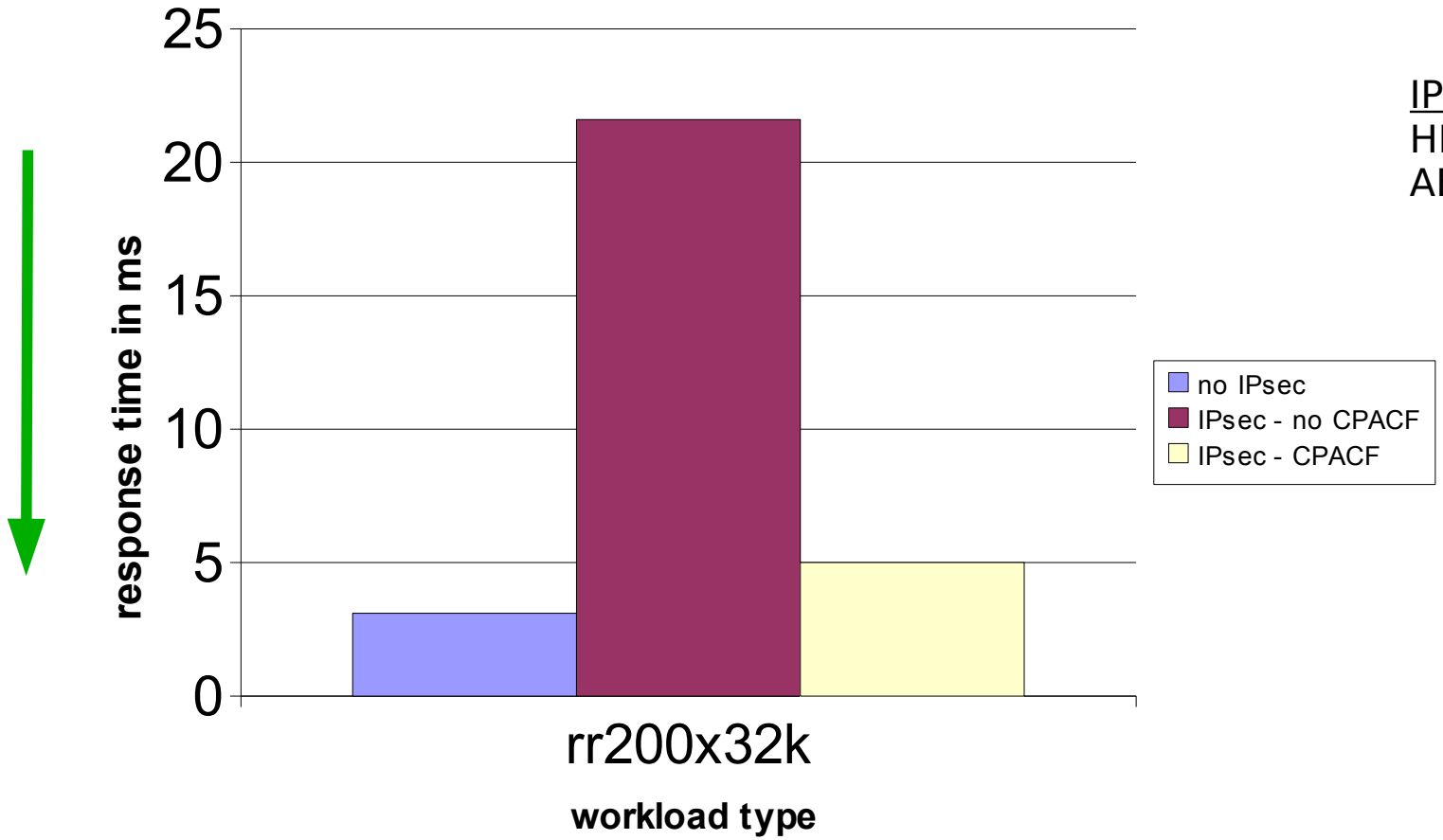
- IPsec overhead is significant
- CPU costs can be 16x higher with software in-kernel crypto
- by using CPACF the CPU costs can be reduced up to 4x



© 2009 IBM Corporation

Linux in-kernel crypto (3)

- response times <= 5 ms for emulated DB request with IPsec/CPACF



IPsec config:
HMAC-SHA256
AES-128

© 2009 IBM Corporation

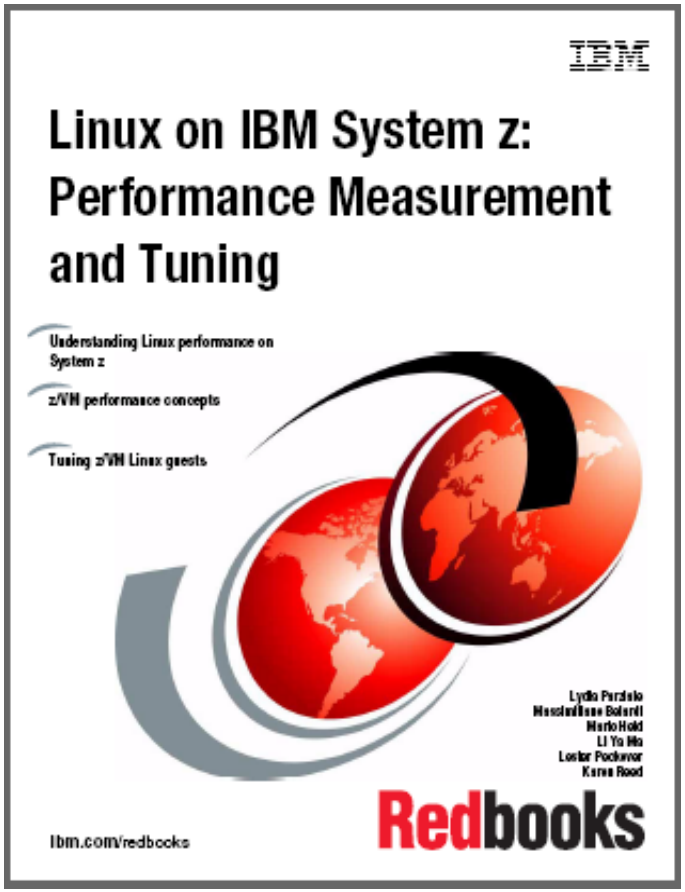
Linux cryptographic support - Summary

- Crypto Express2 Accelerator (CEX2A)
  - optional feature for System z machines
  - executes cryptographic requests asynchronously to the Central Processor (CP)
  - accelerates public key operations used for the SSL protocol (SSL handshake)
  - requires generic zcrypt device driver
  - zcrypt device driver with enabled polling thread utilizes adapter best
- CP Assist for Cryptographic Function (CPACF)
  - supports several block ciphers and secure hash functions
  - executes cryptographic requests synchronously to the Central Processor (CP)
  - Linux kernel can use CPACF as well (in-kernel crypto)
  - software must be configured appropriately to exploit the hardware
- CEX2A and CPACF can be combined
  - for example: SSL uses symmetric and asymmetric ciphers
  - best throughput results with both cryptographic features together

© 2009 IBM Corporation

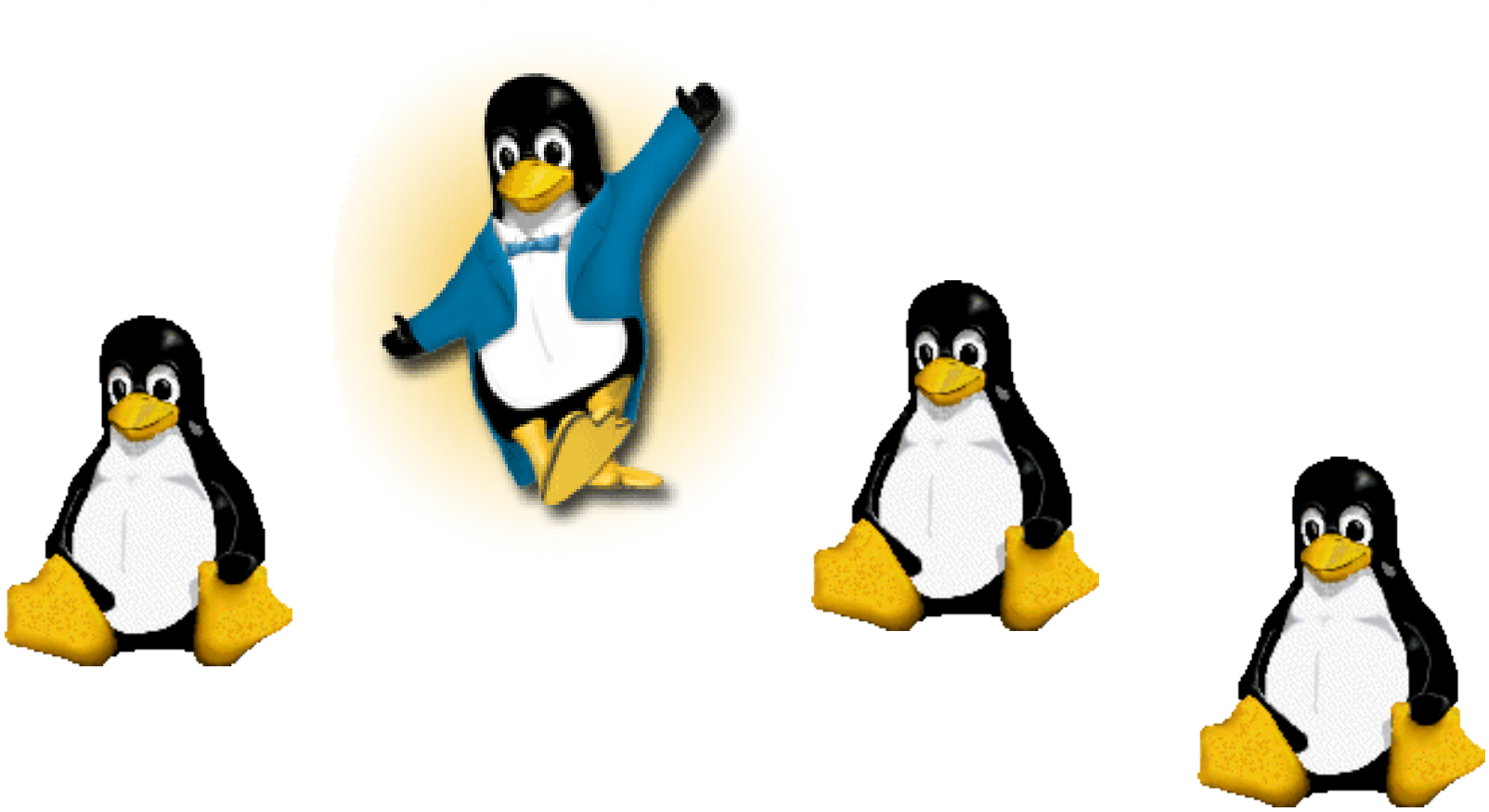
Visit us !

- Linux on System z: Tuning Hints & Tips
  - http://www.ibm.com/developerworks/linux/linux390/perf/
- Linux-VM Performance Website:
  - http://www.vm.ibm.com/perf/tips/linuxper.html
- IBM Redbooks
  - http://www.redbooks.ibm.com/



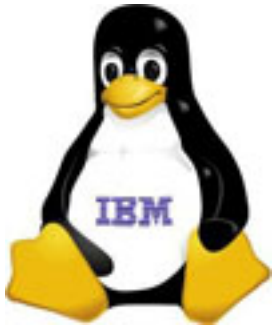
© 2009 IBM Corporation

Questions



© 2009 IBM Corporation

BACKUP



© 2009 IBM Corporation

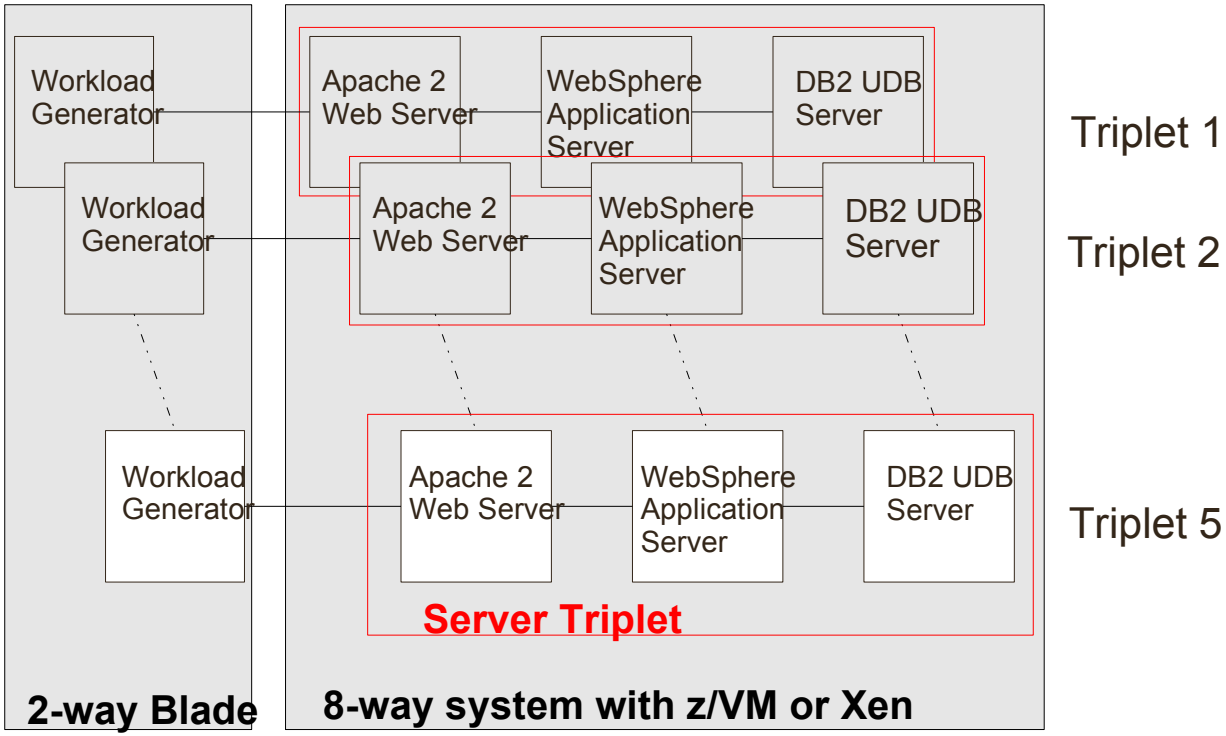
Virtualization scenario with z/VM and Xen

- a Hypervisor provides
  - an environment where an operating system (=guest) runs on virtual hardware
  - the virtual hardware either by emulation in software or by dispatching requests to physical hardware
  - the possibility to run multiple guests on the same physical hardware
- Overcommitting resources
  - in a virtualized environment is it possible to assign more virtual resources to the guests than physical hardware is available
  - the most important resources to become overcommitted are CPU and memory
  - resource overcommitment is one of the major benefits of virtualization!
    - Leads to a higher utilization of the physical hardware
    - makes administration very flexible
  - **Overcommitting resources is related with effort for the hypervisor!**
- z/VM and Xen are hypervisors for different platforms (system z and x86)
  - Objective for this project is to analyze the behavior when overcommitting CPUs and memory

White paper at: http://www.ibm.com/developerworks/linux/linux390/perf/tuning_pap_VM.html#xen

© 2009 IBM Corporation

Environment



- A chain of Web Server, Application Server, and Database Server (=server triplets) is required to drive the workload
- Each server runs on one Linux guest
- Scaling the workload generators and workload triplets up to 5 triplets (total of 15 guests)

© 2009 IBM Corporation

Environment

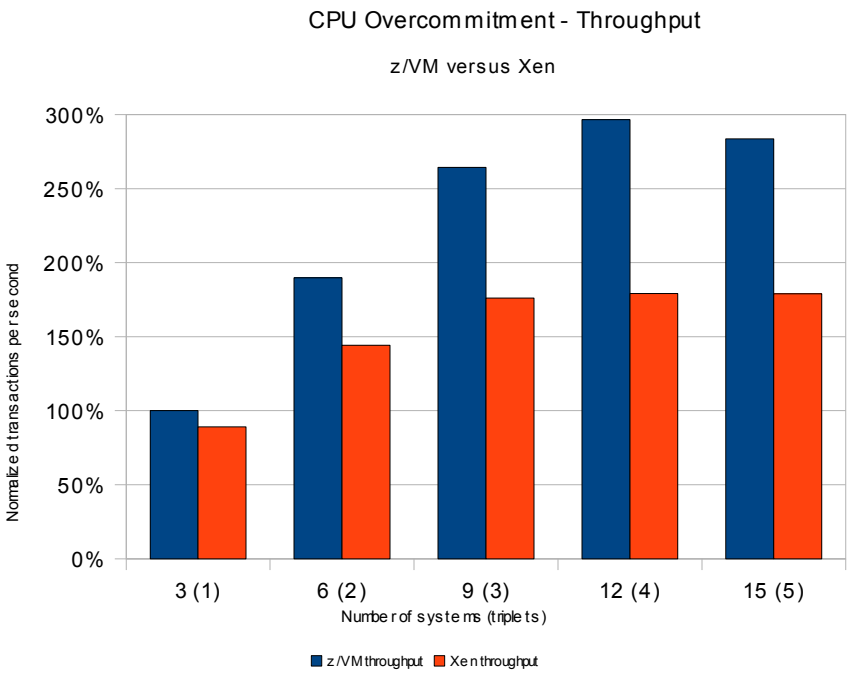
Hardware		
System	**z/VM 5.3**	**Xen**
Architecture	System z9(TM) 2094-S18	x3950, 4 Intel Xeon dual core
Processor	8-way 1.65 GHz	8 way – 3.5 GHz
Memory	20 GB + 2 GB memory	24 GB memory

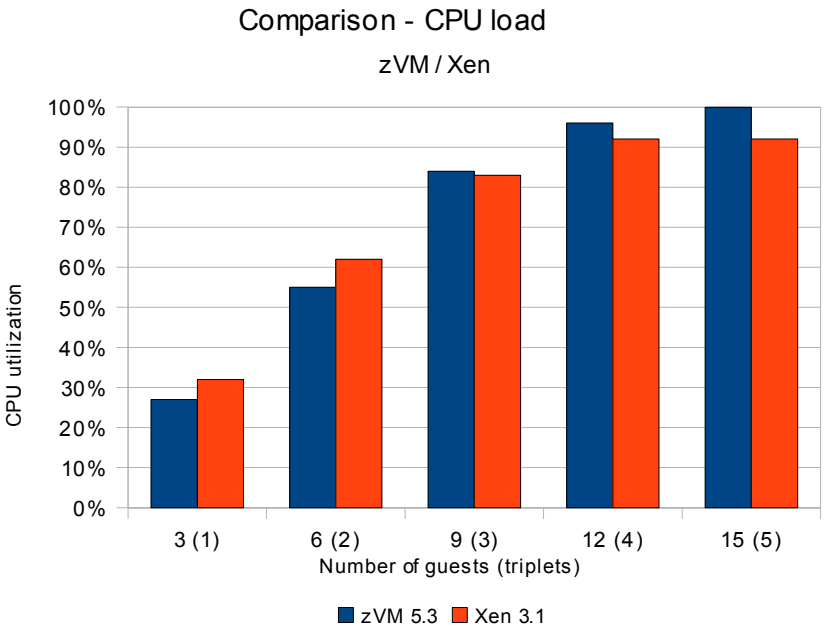
Software	
Hypervisor	**z/VM 5.3 / Xen 3.1**
Web Server	Apache 2.2.3
Linux	SLES 10, SP1
WebSphere App. Server	6.1 fixpack 11, 31-bit
IBM DB2	9.1 fixpack 3

Resource Overcommittment				
Guests (# of Triplets)	**virtual CPUs on 8 physical CPUs**	**CPU Overcommitment**	**virtual Memory [MB] on 20GB phys. Memory**	**Memory Overcommitment**
6 (2)	8	1.0:1	9196	N/A
9 (3)	12	1.5:1	13794	N/A
12 (4)	16	2.0:1	18392	N/A
15 (5)	20	2.5:1	22990	1.1:1

© 2009 IBM Corporation

CPU overcommitment

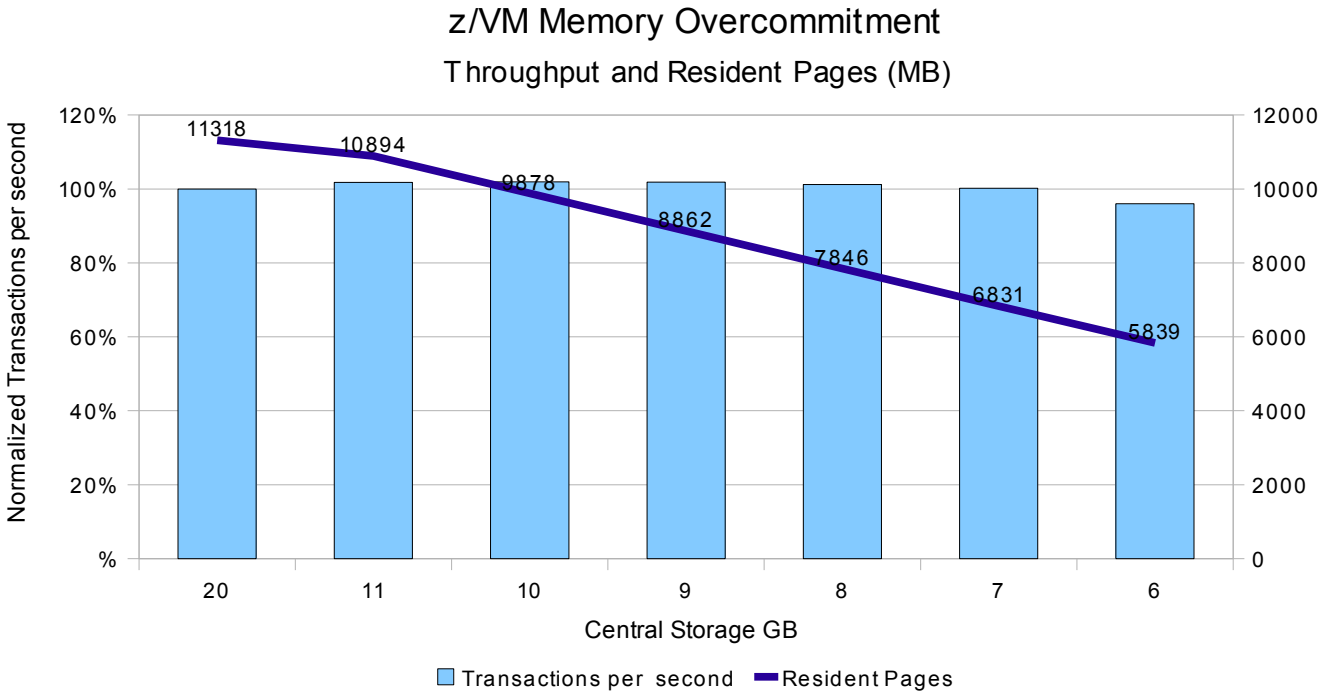




- maximum throughput on z/VM is reached with 12 guests (4 triplets) at nearly full CPU utilization
  ==> Scalability is limited by utilizing the system CPUs completely
- maximum throughput on Xen is reached with 9 guests (3 triplets) at 83% CPU utilization
  ==> scalability is limited by by CPU overcommitment (1.5:1), scaling further increases only the CPU load
  ==> Maximum CPU utilization reached was 92%

© 2009 IBM Corporation

Memory overcommitment

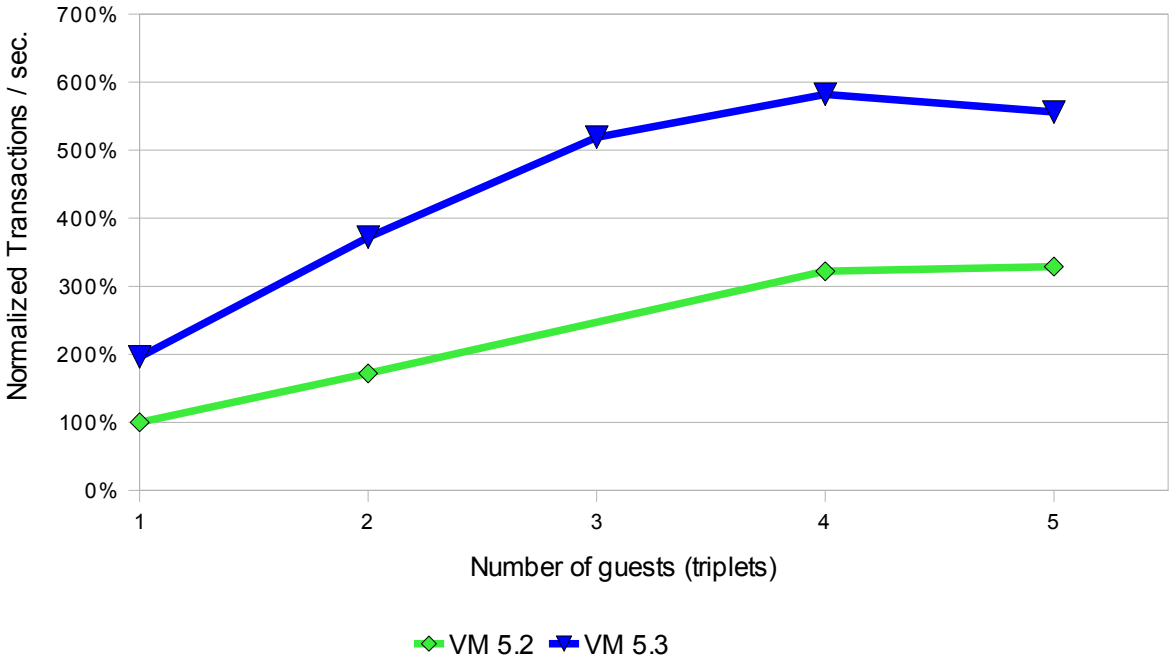


- Memory overcommitment - using 15 Guests (five triplets) and decreasing the system memory
  - Not possible with Xen 3.1, memory handling is static
  - the guests are defined with 22,990MB memory, but z/VM allocates only 11,355MB!
  - While decreasing the amount of memory to 7 GB, the throughput stays constant

© 2009 IBM Corporation

Impact of newer software releases

Virtualization Performance - Throughput Comparison



- Hardware is the same!
- Software upgraded
  - z/VM 5.2 → 5.3
  - Java 1.4 → 1.5
  - WebSphere Application server 6.0.2 → 6.1.0.11
  - DB2 8.2 → 9.1

The newer software levels provides a significant improvement in throughput!

© 2009 IBM Corporation