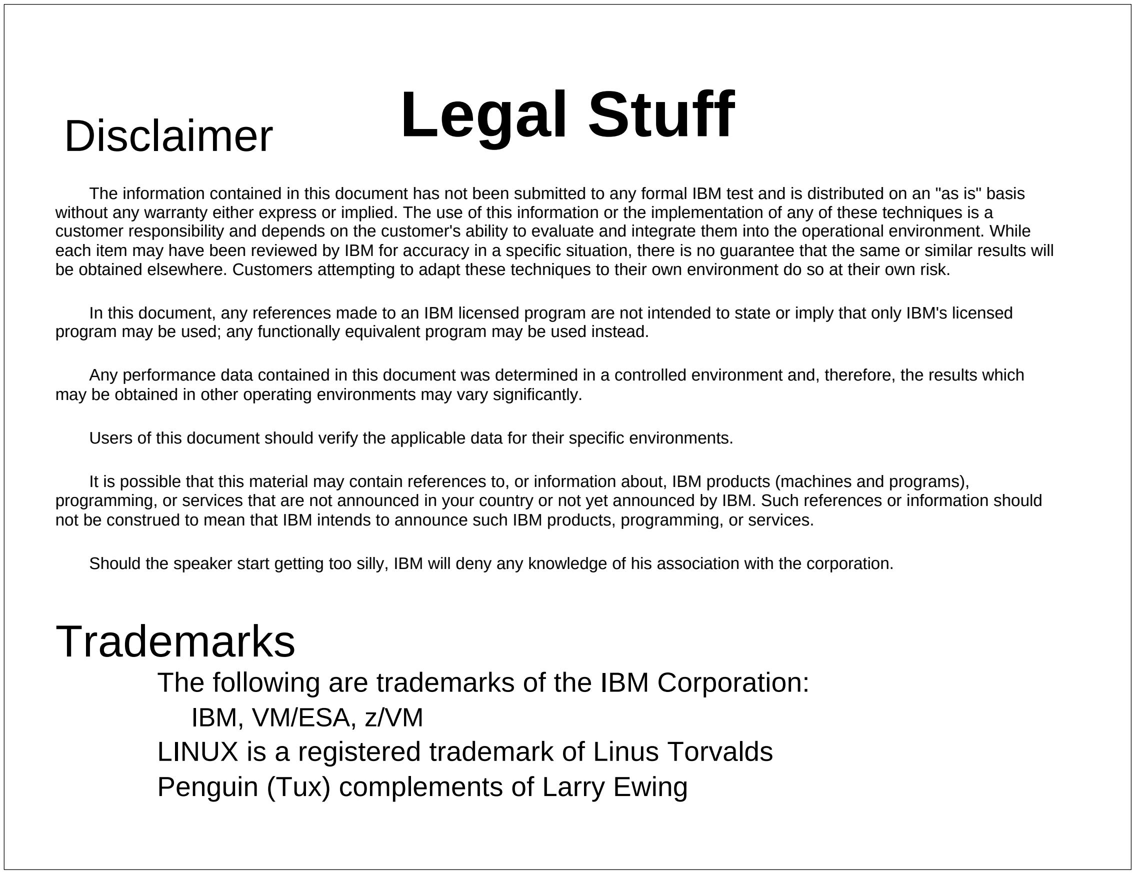
VM Performance Update

Bill Bitner
IBM Endicott
607-752-6022
bitnerb@us.ibm.com

Last Updated: May 29, 2002

►



Legal Stuff

Disclaimer

The information contained in this document has not been submitted to any formal IBM test and is distributed on an "as is" basis without any warranty either express or implied. The use of this information or the implementation of any of these techniques is a customer responsibility and depends on the customer's ability to evaluate and integrate them into the operational environment. While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Customers attempting to adapt these techniques to their own environment do so at their own risk.

In this document, any references made to an IBM licensed program are not intended to state or imply that only IBM's licensed program may be used; any functionally equivalent program may be used instead.

Any performance data contained in this document was determined in a controlled environment and, therefore, the results which may be obtained in other operating environments may vary significantly.

Users of this document should verify the applicable data for their specific environments.

It is possible that this material may contain references to, or information about, IBM products (machines and programs), programming, or services that are not announced in your country or not yet announced by IBM. Such references or information should not be construed to mean that IBM intends to announce such IBM products, programming, or services.

Should the speaker start getting too silly, IBM will deny any knowledge of his association with the corporation.

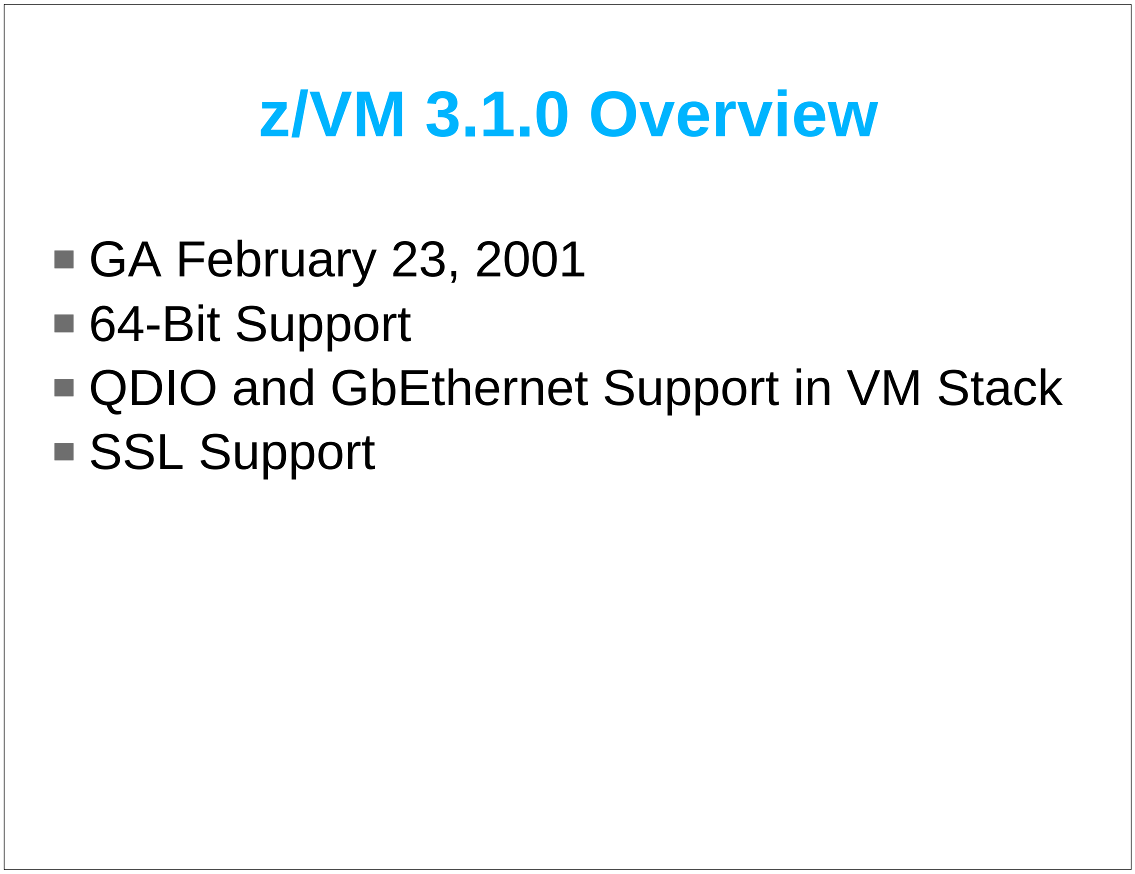
Trademarks

The following are trademarks of the IBM Corporation:
IBM, VM/ESA, z/VM
LINUX is a registered trademark of Linus Torvalds
Penguin (Tux) complements of Larry Ewing

- I will show various examples of reports and data in this presentation. Many of the reports have been slightly edited to allow them to fit on the page and to highlight the important information.
- The speaker notes you are reading are meant as a supplement to the presentation. I can not guarantee that they will have the same impact or accuracy as seeing the presentation first hand. Please excuse grammar and typos. However, any suggestions or corrections are appreciated.

Presentation Contents

- z/VM 3.1.0 (GA Feb 23, 2001)
  - Regression performance
  - Large Storage Considerations
  - TCP/IP - QDIO - Gigabit Ethernet
  - SSL
- z/VM 4.1.0 (GA July 20, 2001)
  - Network CCW Translation Improvements
- z/VM 4.2.0 (GA Oct 26, 2001)
  - Hipersockets
  - Page Fault Processing
  - IMAP
- z/VM 4.3.0 (GA May 31, 2002)
  - CP timer management
  - TCP/IP Stack enhancements
  - Managing contention for storage under 2GB
  - Large volume CMS minidisks



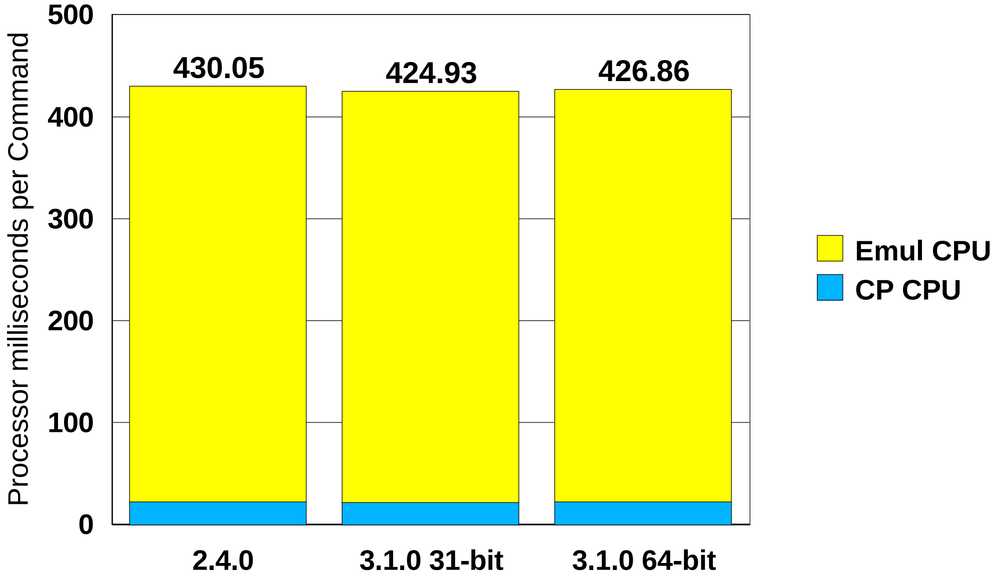
z/VM 3.1.0 Overview

- GA February 23, 2001
- 64-Bit Support
- QDIO and GbEthernet Support in VM Stack
- SSL Support

z/VM 3.1.0 64-bit Background

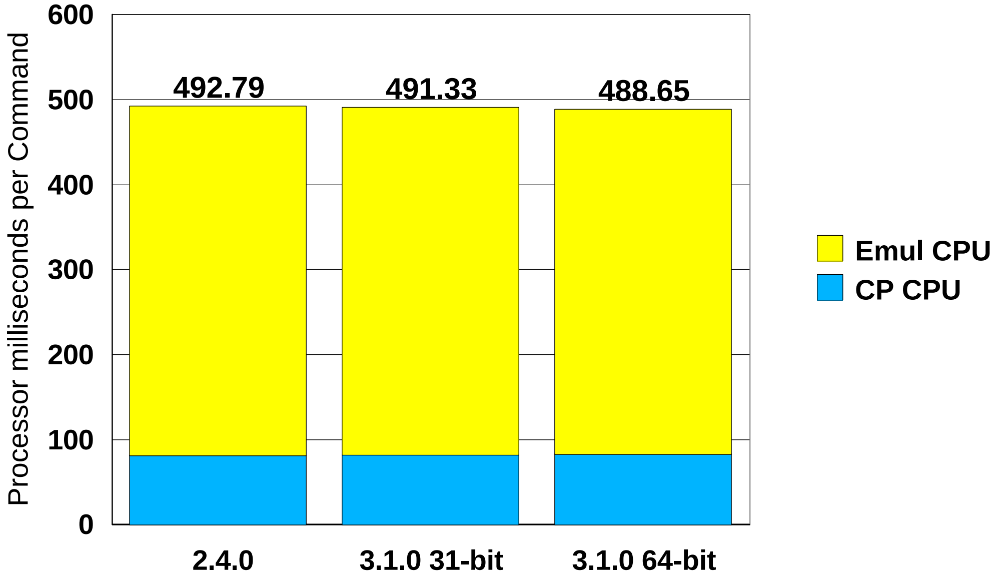
- CP will run in 31-bit or 64-bit
- Much of the code is still common
- RIO370 dropped
- V=R area still must reside below 2GB
- Storage above 2GB used for DPA & MDC

VSE Guest V=R Regression



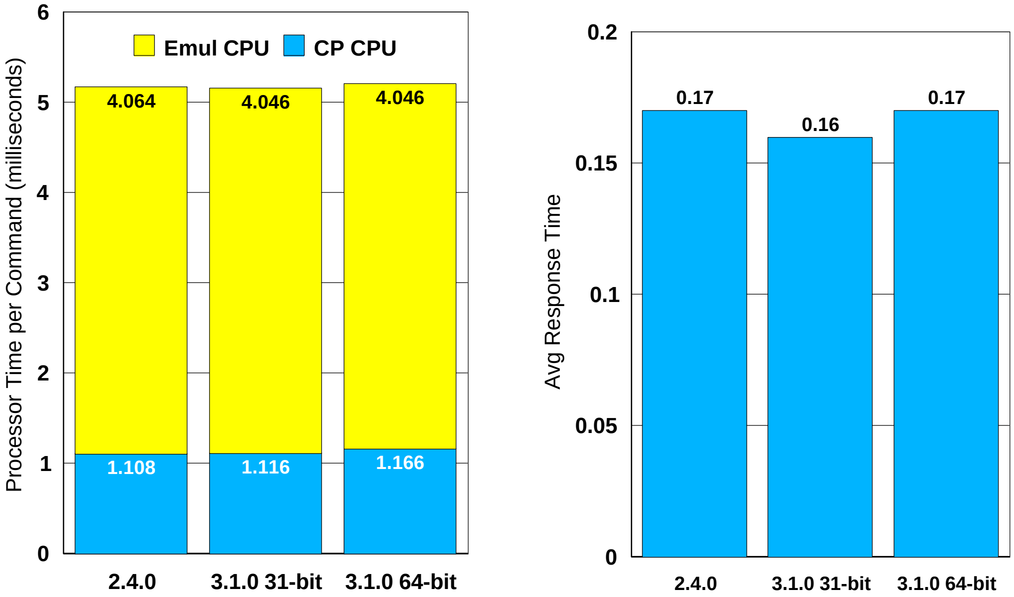
2064-109 2-proceesors online; 2G/2G; V=R VSE Dynapace

VSE Guest V=V Regression



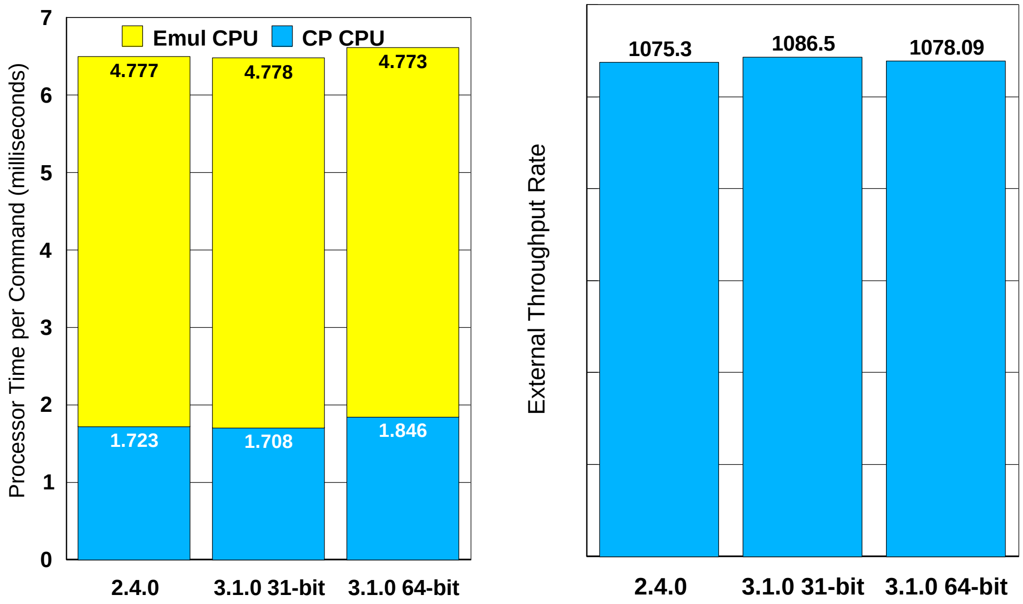
2064-109 2-proceesors online; 2G/2G; V=V VSE Dynapace

CMS Regression



2064-109 LPAR 2-way; 1G/2G; CMS1 External TPNS

CMS Regression

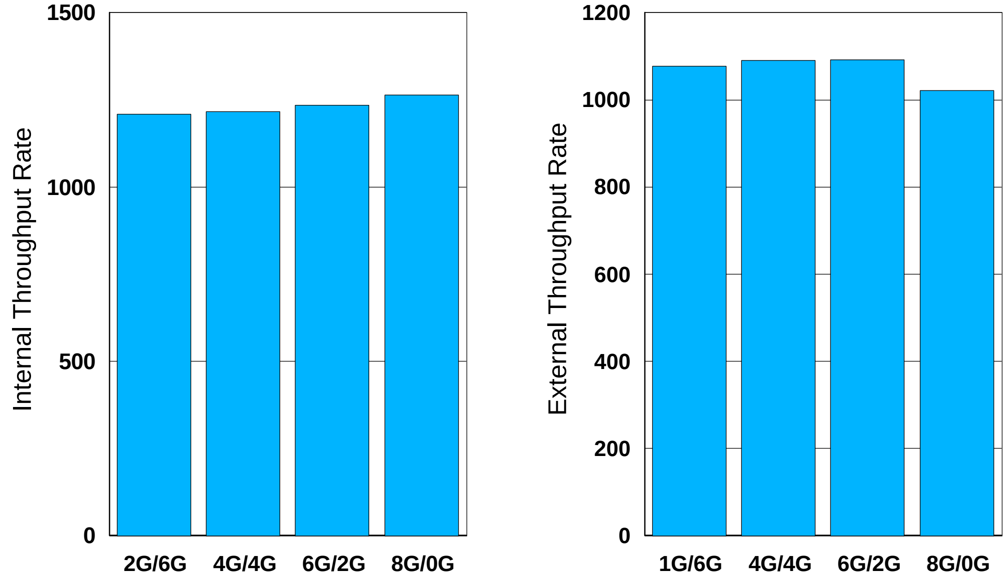


2064-1C8 8-way; 2G/6G; CMS1 Internal TPNS

Storage Allocation Considerations

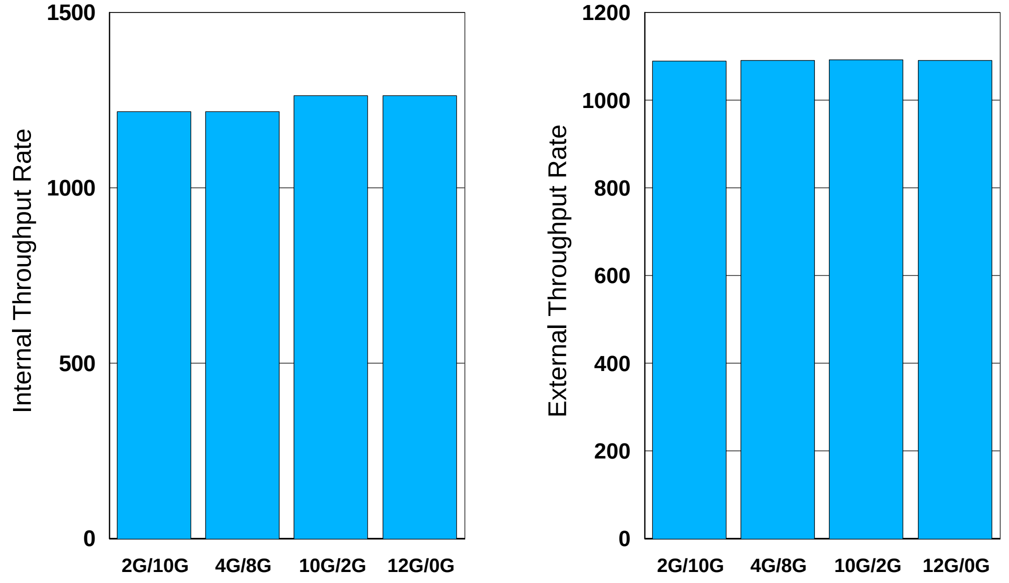
- Can now run VM with greater than 2GB of real storage on 2064 processors.
  - Should there be any expanded storage?
  - How should storage be used for MDC?
    - Real only?
    - Expanded only?
  - How much can I use for the V=R area?

Storage Allocation - 8 GB



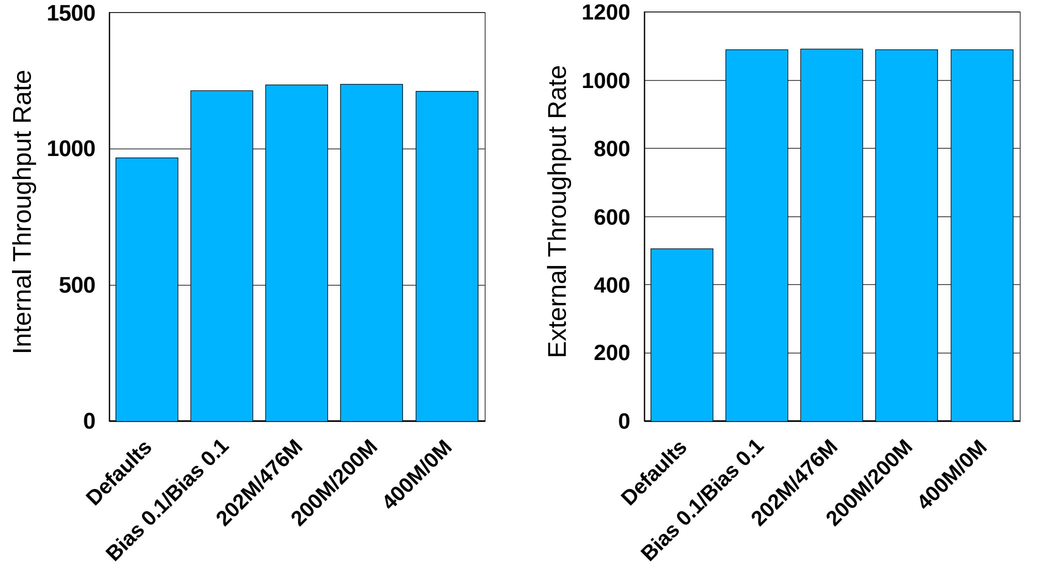
2064-1C8, 8-way; 10800 users; CMS1 Internal TPNS

Storage Allocation - 12 GB



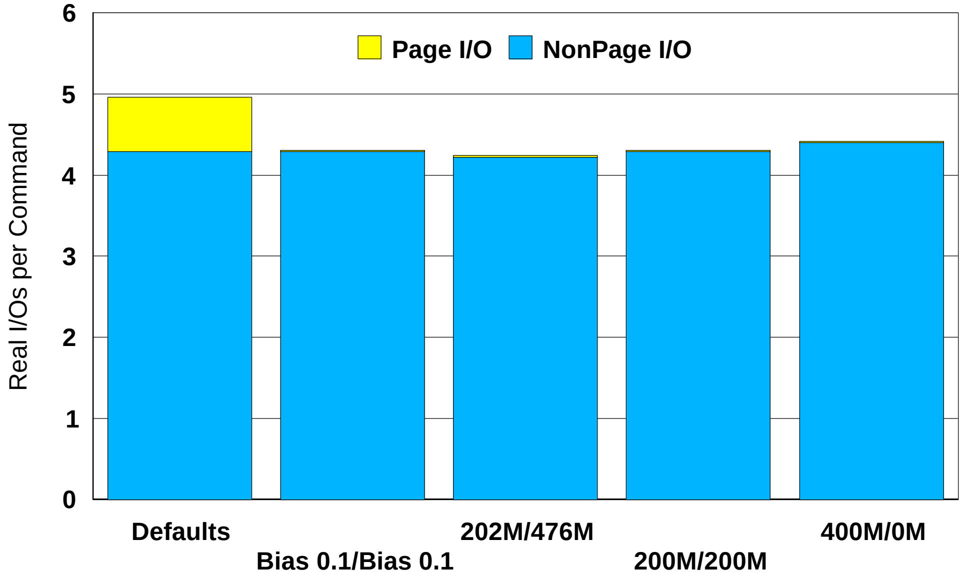
2064-1C8, 8-way; 10800 users; CMS1 Internal TPNS

MDC Tuning 8GB



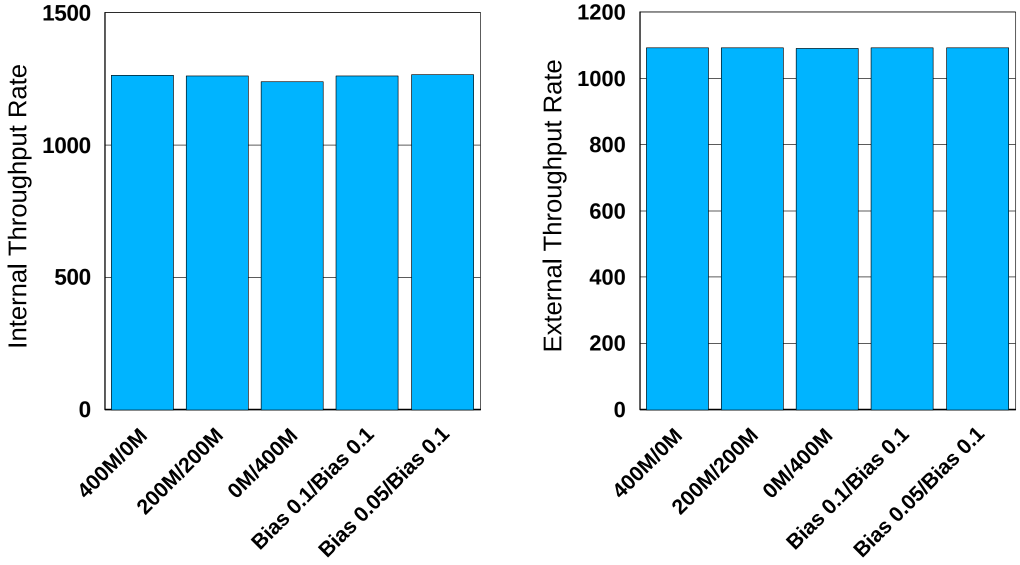
2064-1C8, 8-way; 6G real/2G xstore; 10800 users; CMS1 Internal TPNS

MDC 8GB - I/O per Command



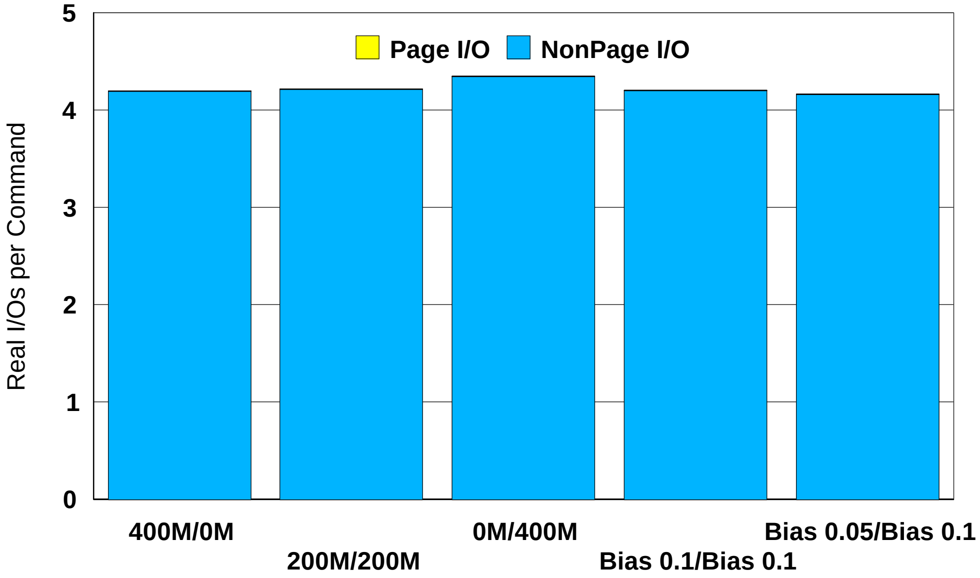
2064-1C8, 8-way; 6G real/2G xstore; 10800 users; CMS1 Internal TPNS

MDC Tuning 12 GB



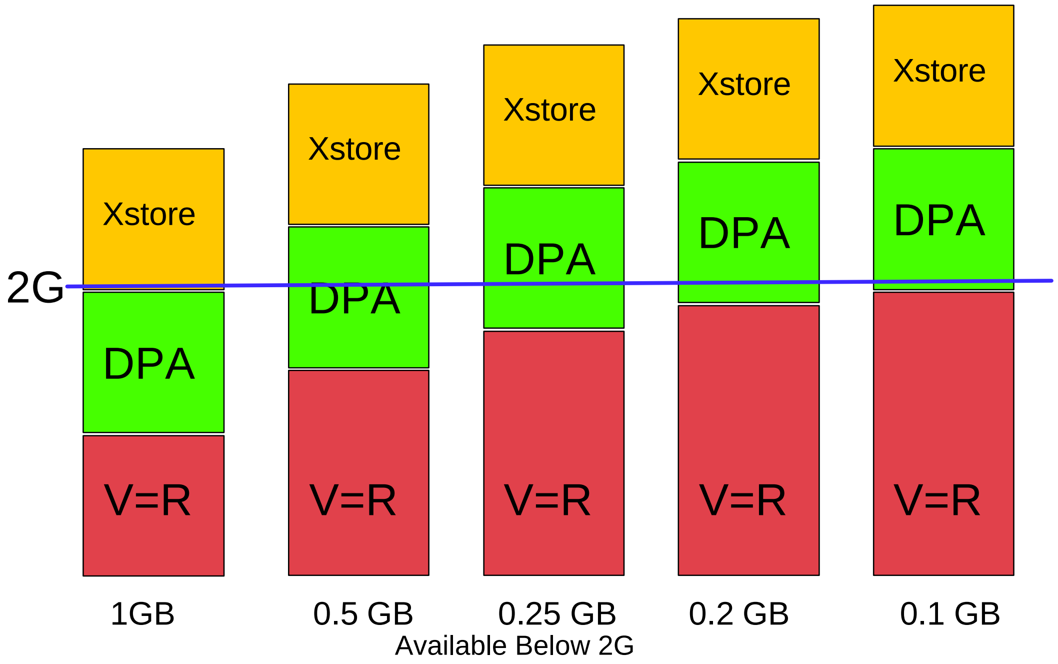
2064-1C8, 8-way; 10G real/2G xstore; 10800 users; CMS1 Internal TPNS

MDC 12GB - I/O per Command

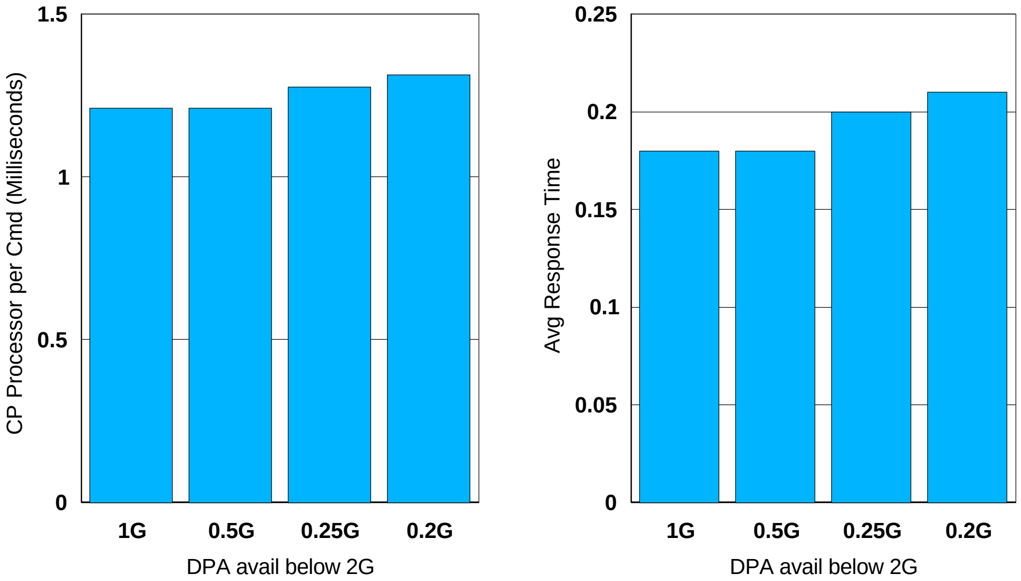


2064-1C8, 8-way; 6G real/2G xstore; 10800 users; CMS1 Internal TPNS

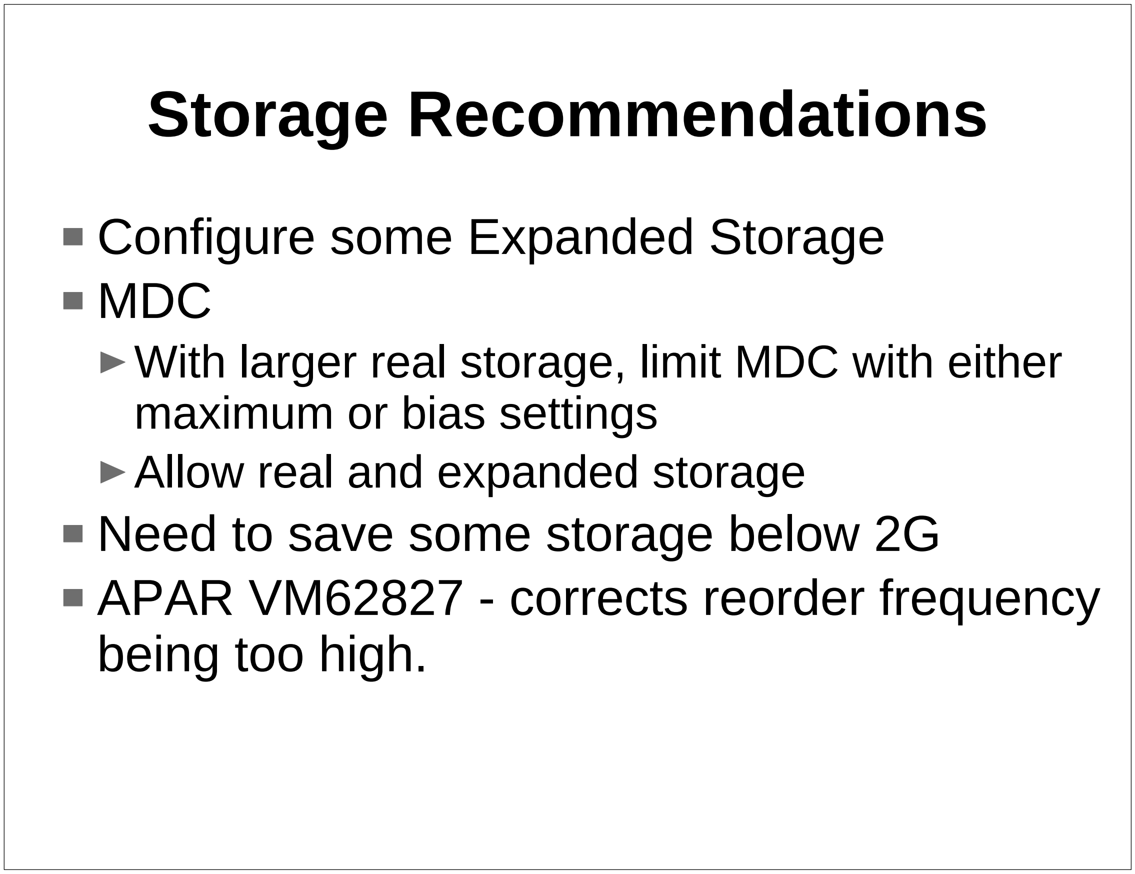
Contention Below 2G



Contention Below 2G



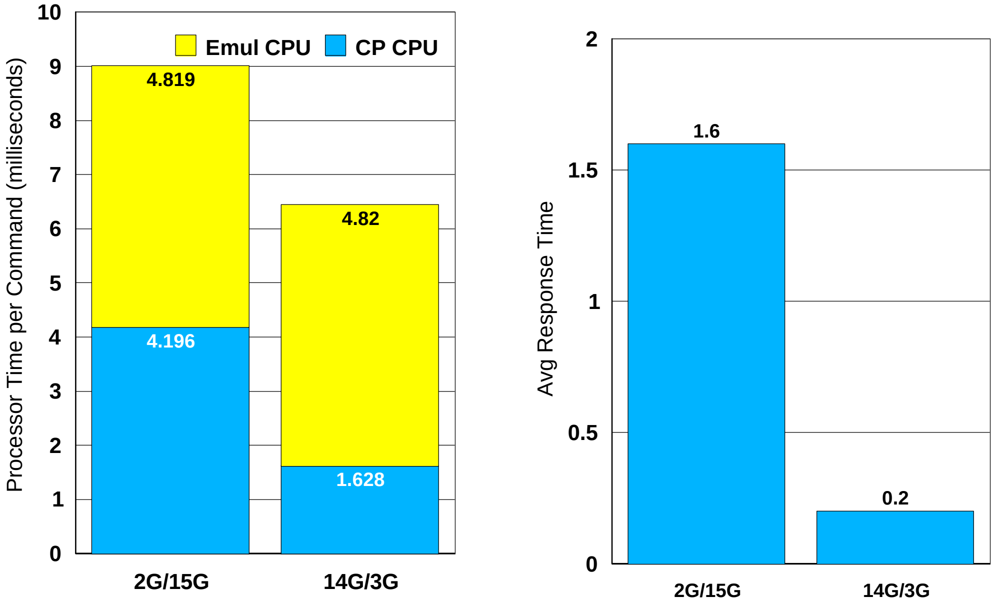
2064-109; LPAR 2-way; 3420 users; CMS1 External TPNS



Storage Recommendations

- Configure some Expanded Storage
- MDC
  - With larger real storage, limit MDC with either maximum or bias settings
  - Allow real and expanded storage
- Need to save some storage below 2G
- APAR VM62827 - corrects reorder frequency being too high.

Storage Exploitation

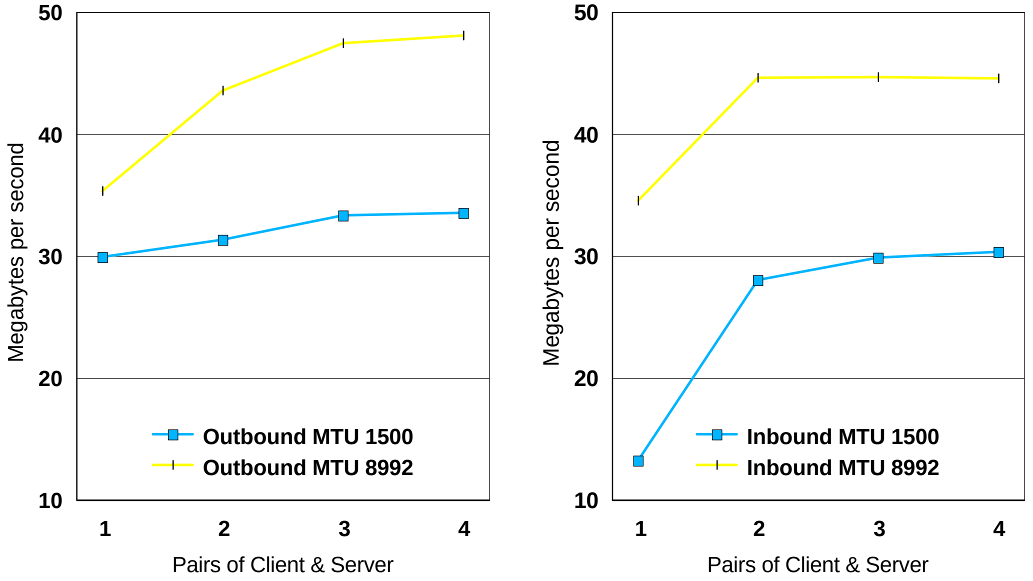


2064-110; 17G total; CMS1 Internal TPNS; zVM 3.1.0

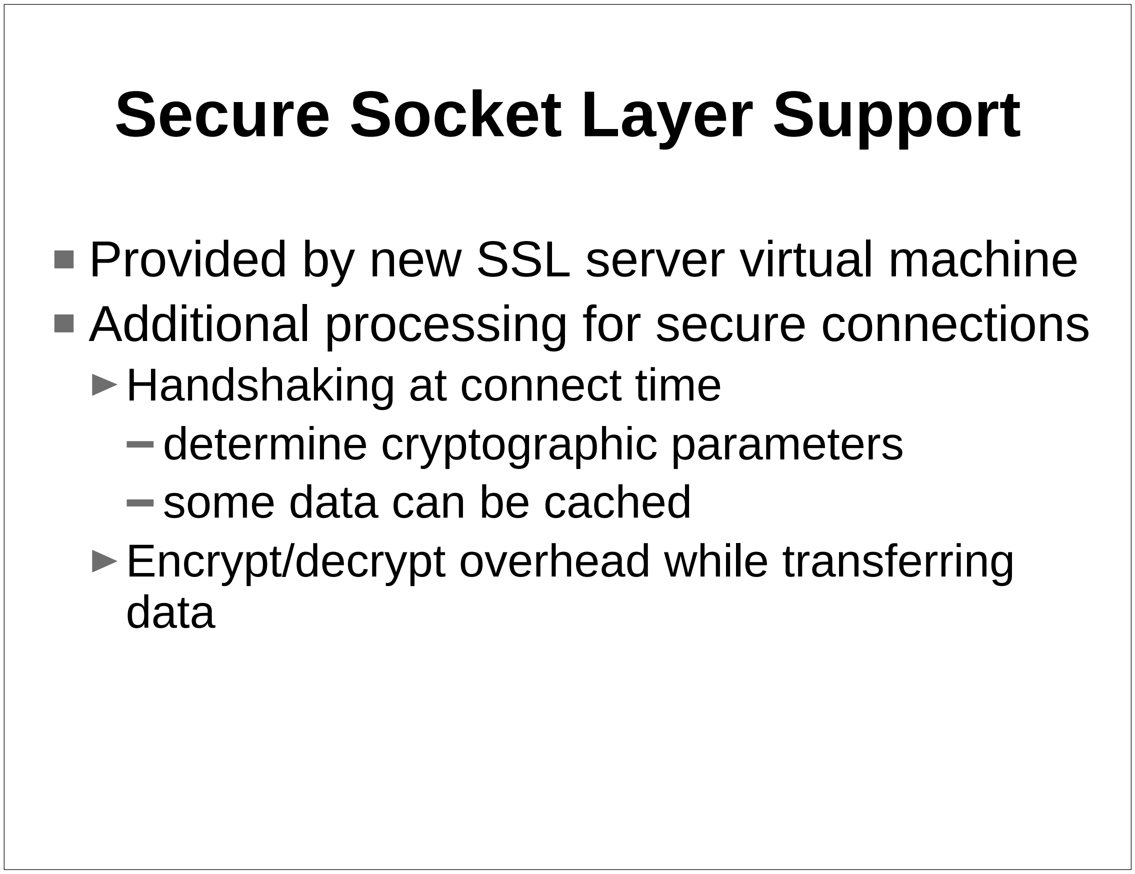
Queued Direct I/O Support

- Previously QDIO available to guests
- TCP/IP Level 3A0 uses for Gigabit Ethernet
- Available on G5, G6, and zSeries processors
- Data transfer via data queues instead of SSCH
- Controlled via state-change-signaling protocol
- Also supports ATM and Fast Ethernet

QDIO Datastream Results



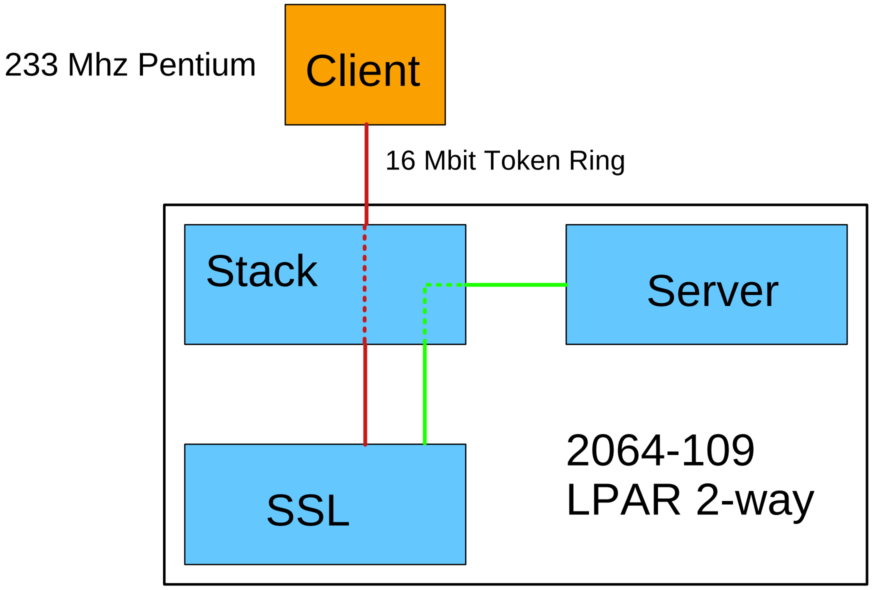
9672-ZZ7 LPAR; z/VM 3.1.0 TCP/IP 3A0



Secure Socket Layer Support

- Provided by new SSL server virtual machine
- Additional processing for secure connections
  - Handshaking at connect time
    - determine cryptographic parameters
    - some data can be cached
  - Encrypt/decrypt overhead while transferring data

SSL Environment



233 Mhz Pentium

Client

16 Mbit Token Ring

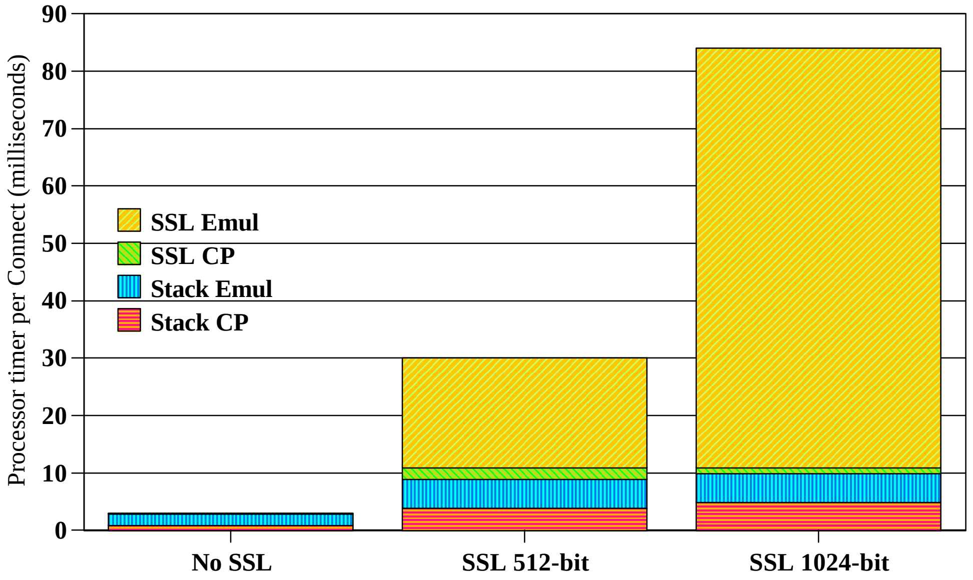
Stack

Server

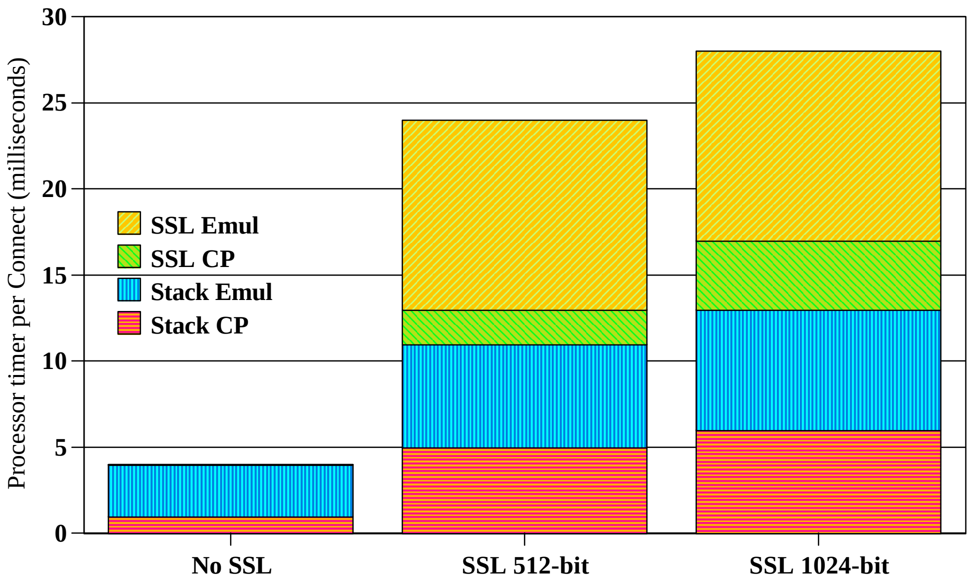
SSL

2064-109
LPAR 2-way

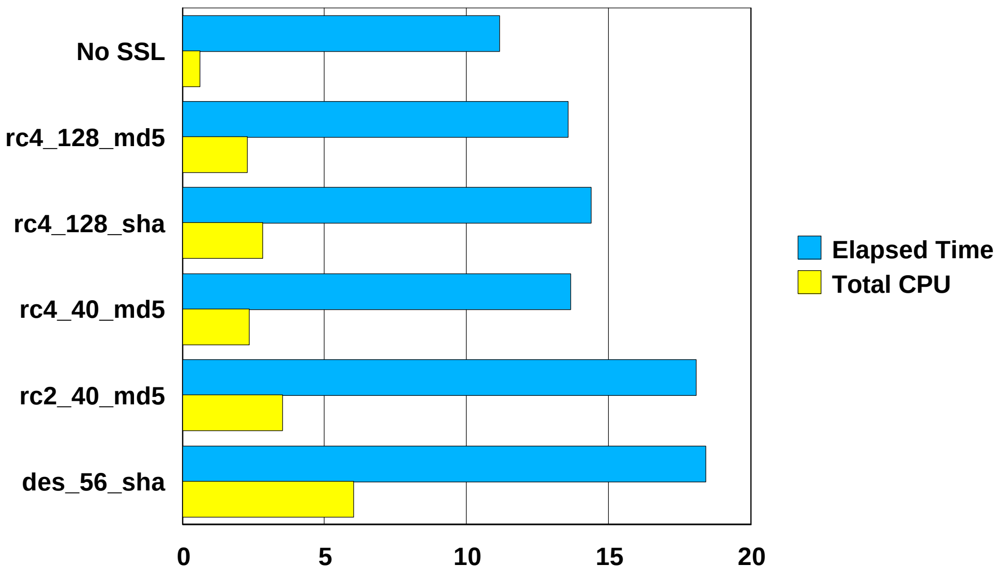
SSL Connect - New Session

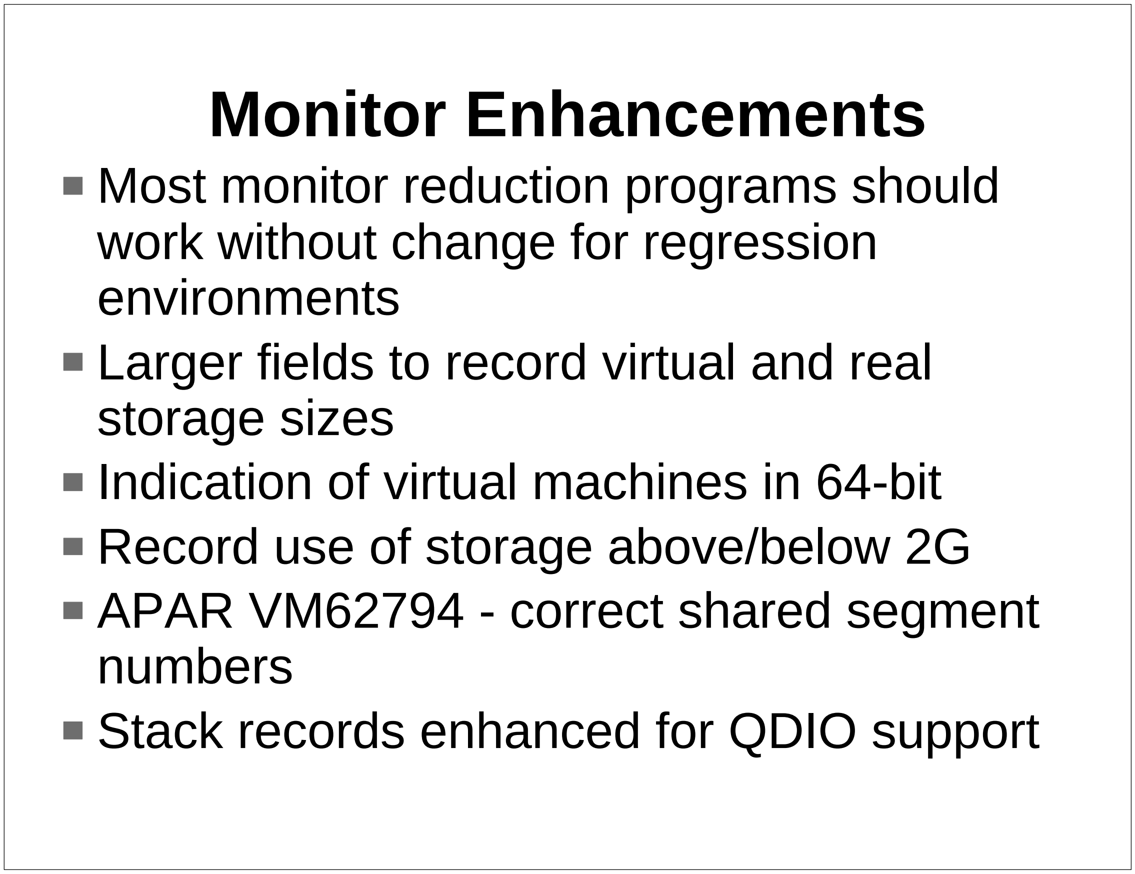


SSL Connect - Resume Session



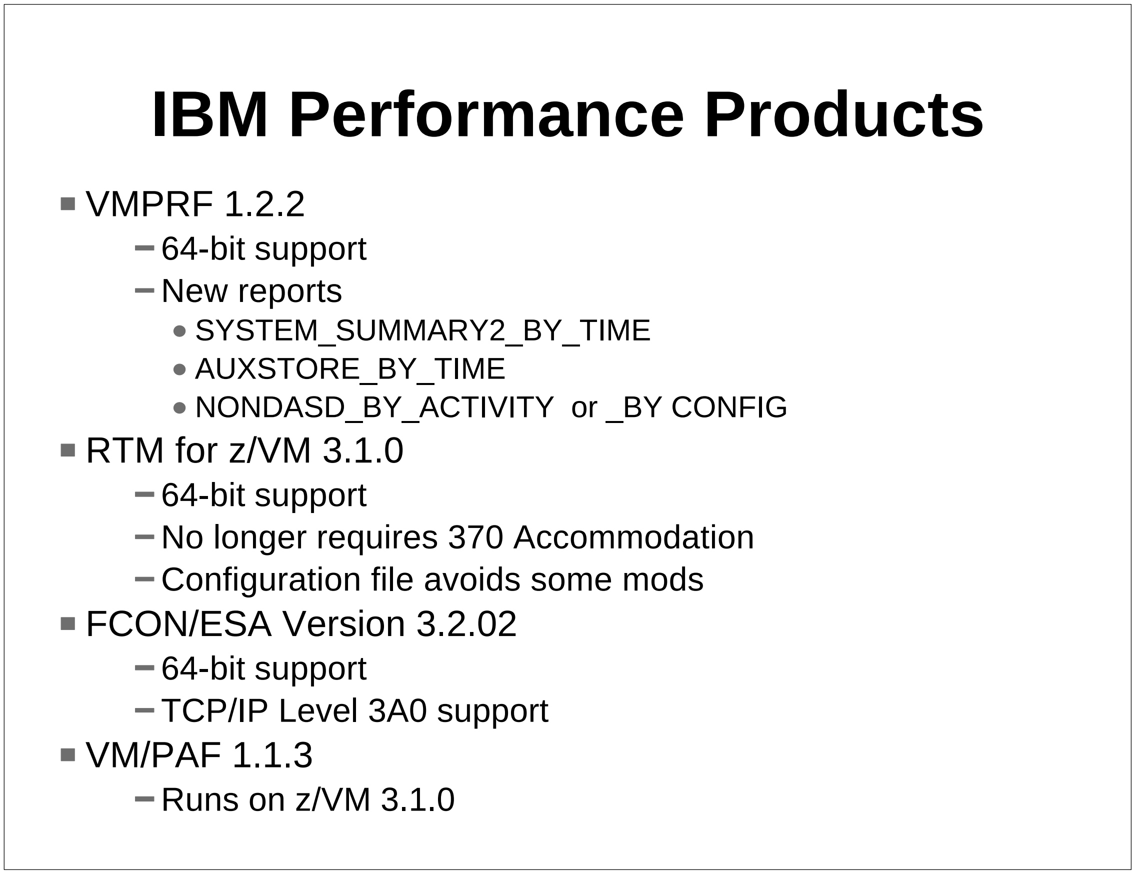
SSL - FTP Binary Get 10M





Monitor Enhancements

- Most monitor reduction programs should work without change for regression environments
- Larger fields to record virtual and real storage sizes
- Indication of virtual machines in 64-bit
- Record use of storage above/below 2G
- APAR VM62794 - correct shared segment numbers
- Stack records enhanced for QDIO support



IBM Performance Products

- VMPRF 1.2.2
  - 64-bit support
  - New reports
    - SYSTEM_SUMMARY2_BY_TIME
    - AUXSTORE_BY_TIME
    - NONDASD_BY_ACTIVITY or _BY CONFIG
- RTM for z/VM 3.1.0
  - 64-bit support
  - No longer requires 370 Accommodation
  - Configuration file avoids some mods
- FCON/ESA Version 3.2.02
  - 64-bit support
  - TCP/IP Level 3A0 support
- VM/PAF 1.1.3
  - Runs on z/VM 3.1.0

Performance Management

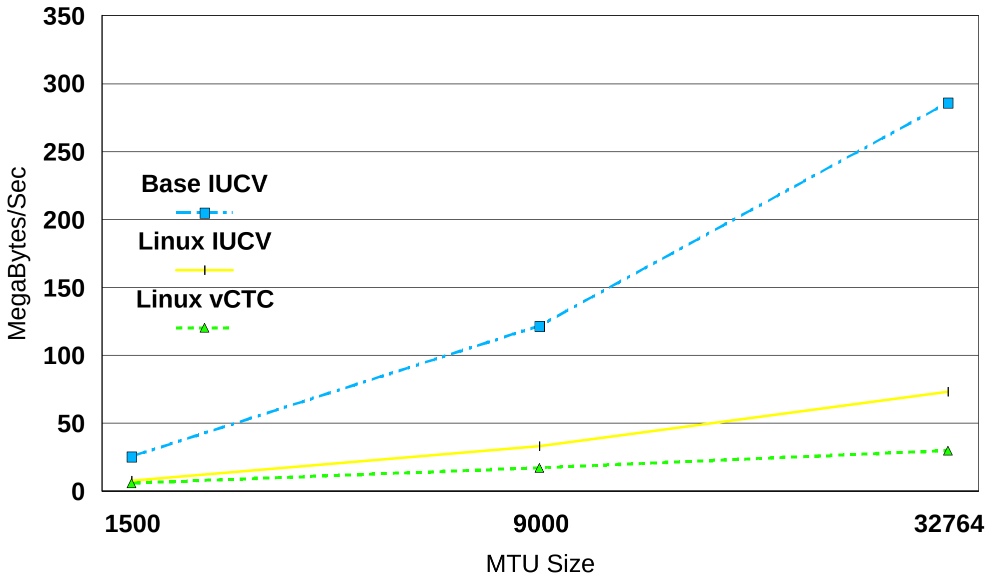
- CP logic and control blocks drastically changed
  - Review CP mods
  - Review tools that pull data from CP control blocks
- CP Trace Table Changes
  - Some entries double in size
- QUERY FRAMES

```
SYSGEN  REAL    USABLE  OFFLINE
524287  524287  524287  000000
V=R     RESNUC  PAGING  TRACE   RIO370
000000  000667  523070  000550  000000
AVAIL   PAGNUC  LOCKRS  LOCKCP  SAVE    FREE    LOCKRIO
506751  009916  000300  000000  000061  006042  000000
Storage >= 2G:
   Online         = 786432      Available List = 58941
   Not init       = 0           Offline        = 0
```

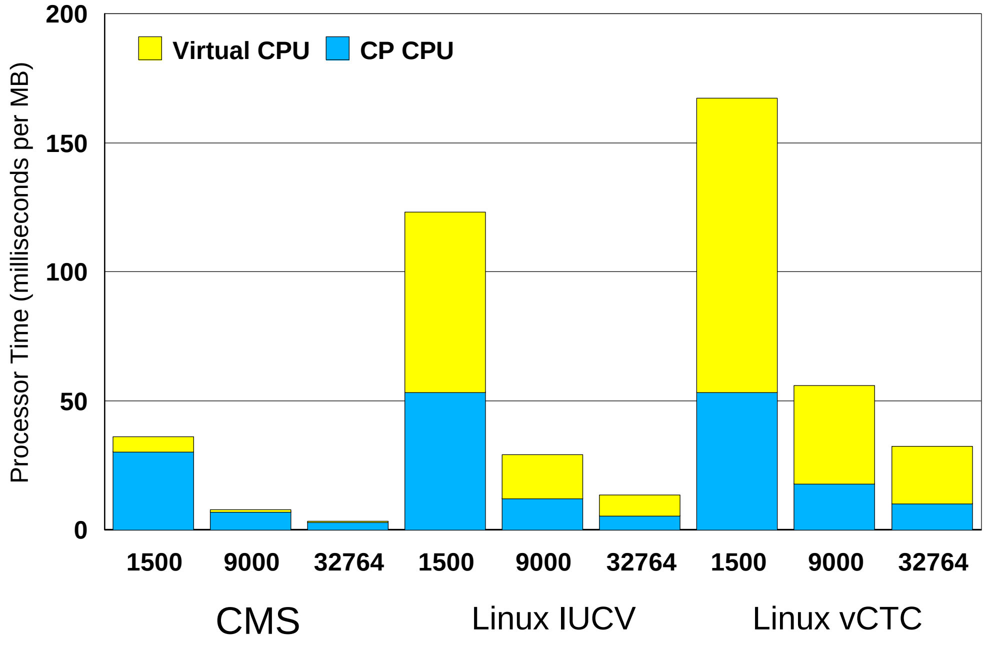
Linux Virtual Connectivity

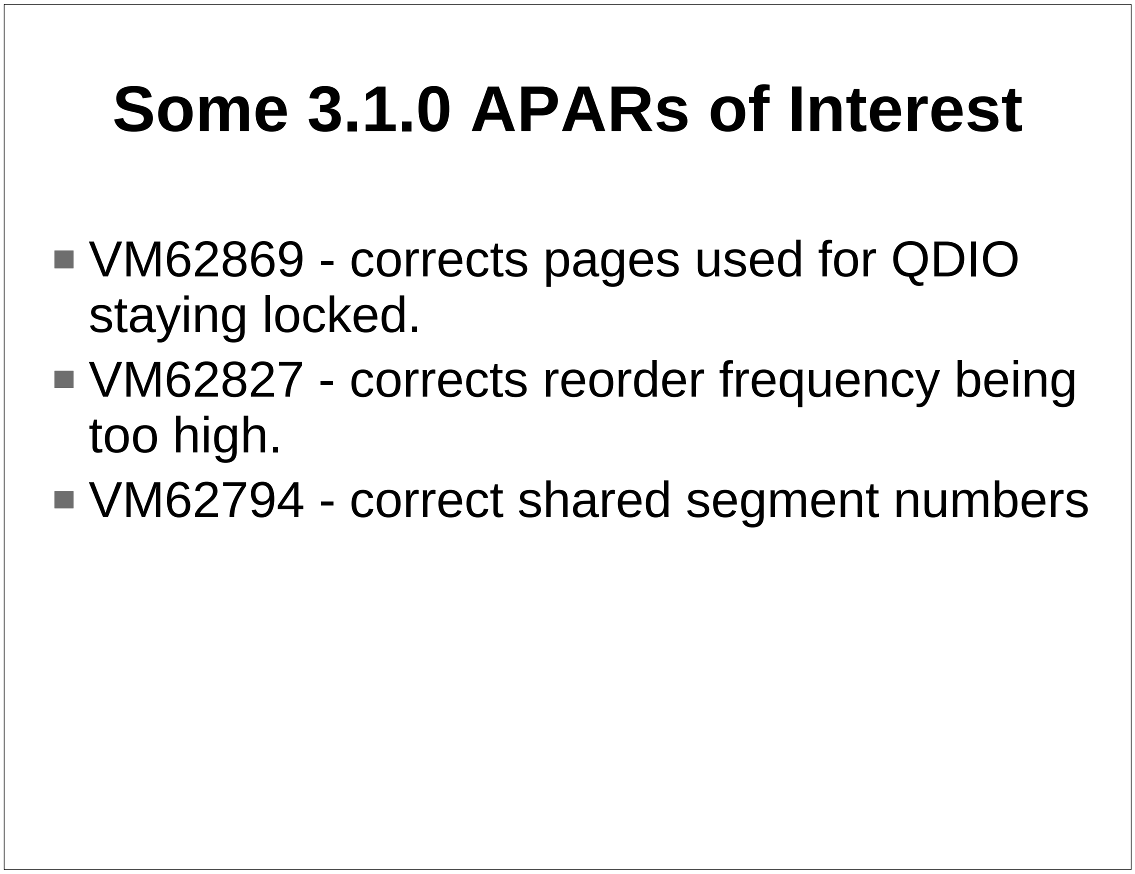
- CMS Driver
  - Really synchronous APPC/VM
  - Very little application/protocol overhead
- Linux 2.2.19 using IUCV
  - Internal Tool to drive networks
  - Application and protocol overhead included
- Linux 2.2.19 using Virtual CTC
  - Internal Tool to drive networks
  - Application and protocol overhead included
- Test Environment
  - 9672-XZ7, Two processor LPAR with 2G/2G
  - z/VM 3.1.0 running 128MB Linux guests

Linux Virtual Communication



Communication Processor Time





Some 3.1.0 APARs of Interest

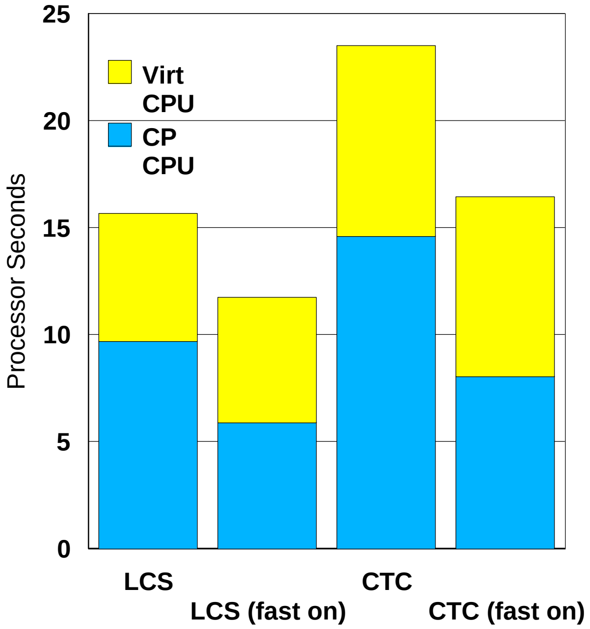
- VM62869 - corrects pages used for QDIO staying locked.
- VM62827 - corrects reorder frequency being too high.
- VM62794 - correct shared segment numbers

z/VM 4.1.0 Overview

- GA July 20, 2001
- New version - new pricing
- Mostly packaging changes
  - CMS Utilities Feature in Base
  - Dirmaint, RTM, VMPRF are now priced features
- Enhanced CCW Translation for Network I/O
- Equivalent regression performance

Enhanced Network CCW Translation

- SSCH oriented network device I/O
- Lowers the CP CPU time required for CCW translation
- Fast CCW Translation previously only for DASD
- 39 to 45% reduction in CP processor time for workloads measured.
- 2064-109, 2-way LPAR



z/VM 4.2.0 Overview

- GA October 26, 2001
- Network improvements
  - Hipersockets
  - Guest LAN
  - VM TCP/IP Stack
- Linux related enhancements
  - Page Fault Resolution (also APAR to 4.1.0)
  - 64-bit CCW Translation
  - Crypto HW Support
- Regression performance equivalent to 4.1.0

HiperSockets Hardware Elements

- 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 (TCP/IP stacks) across all 4 HiperSockets and 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
    - New "thin interrupt" without use of System Assist Processor
  - optional dispatcher polling mechanism
- HiperSockets Hardware I/O Configuration with new CHPID type=IQD
  - Controlled like a regular CHPID
  - Each CHPID has configurable Maximum Frame Size
- Works with both standard and IFL CPs
- Secure connections
- Both 31 bit and 64 bit operating systems supported
- Pre-req: IBM eServer zSeries 900 LIC Update

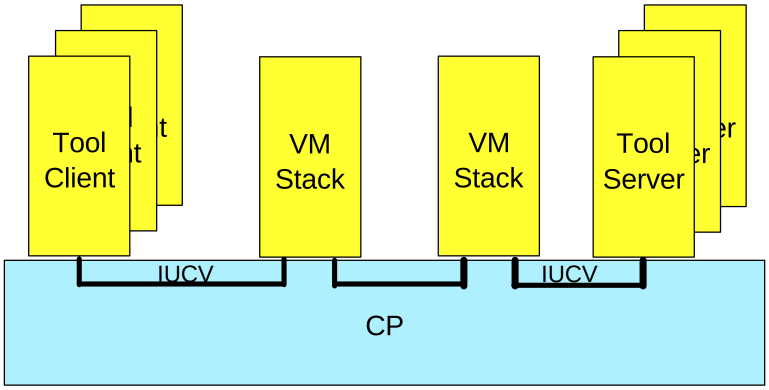
VM and HiperSockets

- VM Support for real HiperSockets
  - VM TCP/IP Stack can use
  - Guests with support (z/OS and Linux)
- Can be used to communicate between guests on same VM system
- Guest LAN is simulated HiperSockets within a VM system. Available on all machines that z/VM 4.2.0 supports.
- Enabled with VM62938 and PQ51738
  - Also recommend VM63034

Network Driver Tool

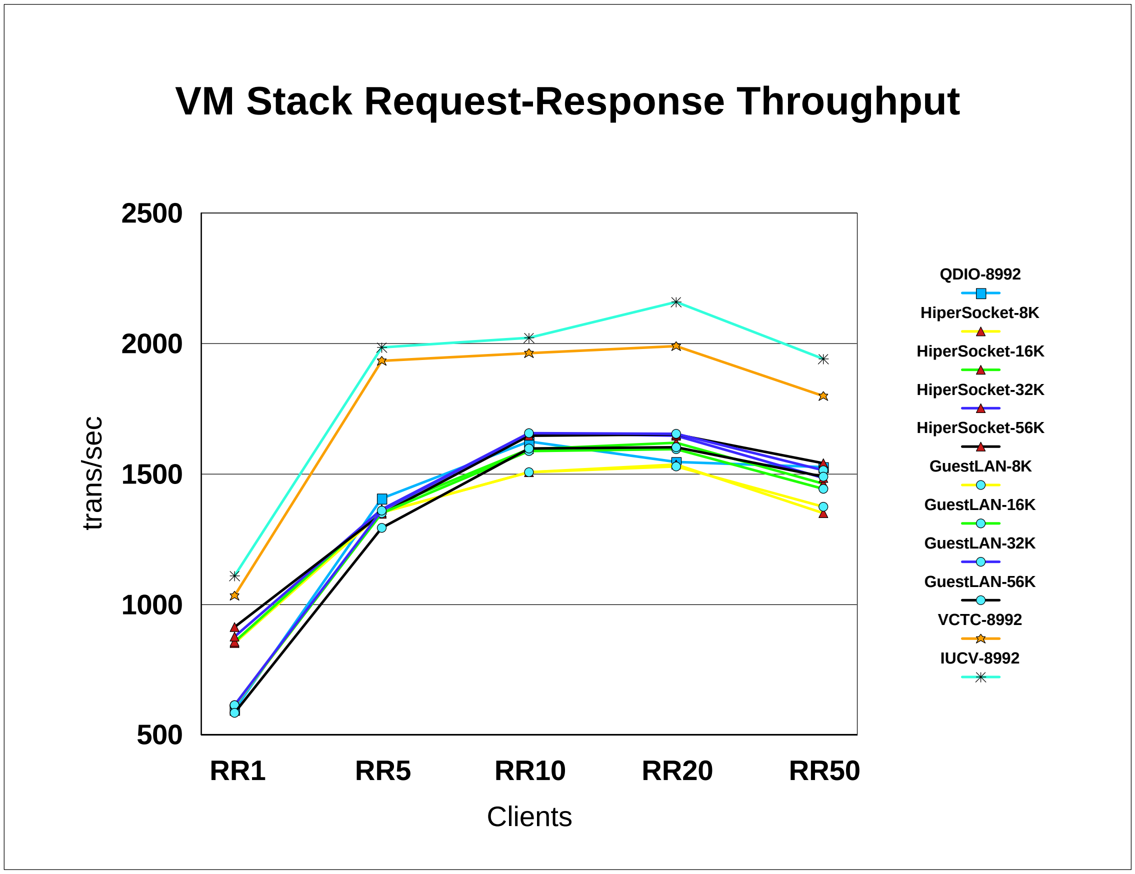
- Request-Response (RR)
  - client sends 200 bytes
  - server responds with 1000 bytes
- Connect-Request-Response (CRR)
  - client connects
  - client sends 64 bytes
  - server responds with 8K bytes
- Streaming (STR)
  - client sends 20 bytes
  - server responds with 20MB
- Various number of clients/users can be used.

VM TCP/IP Measurements

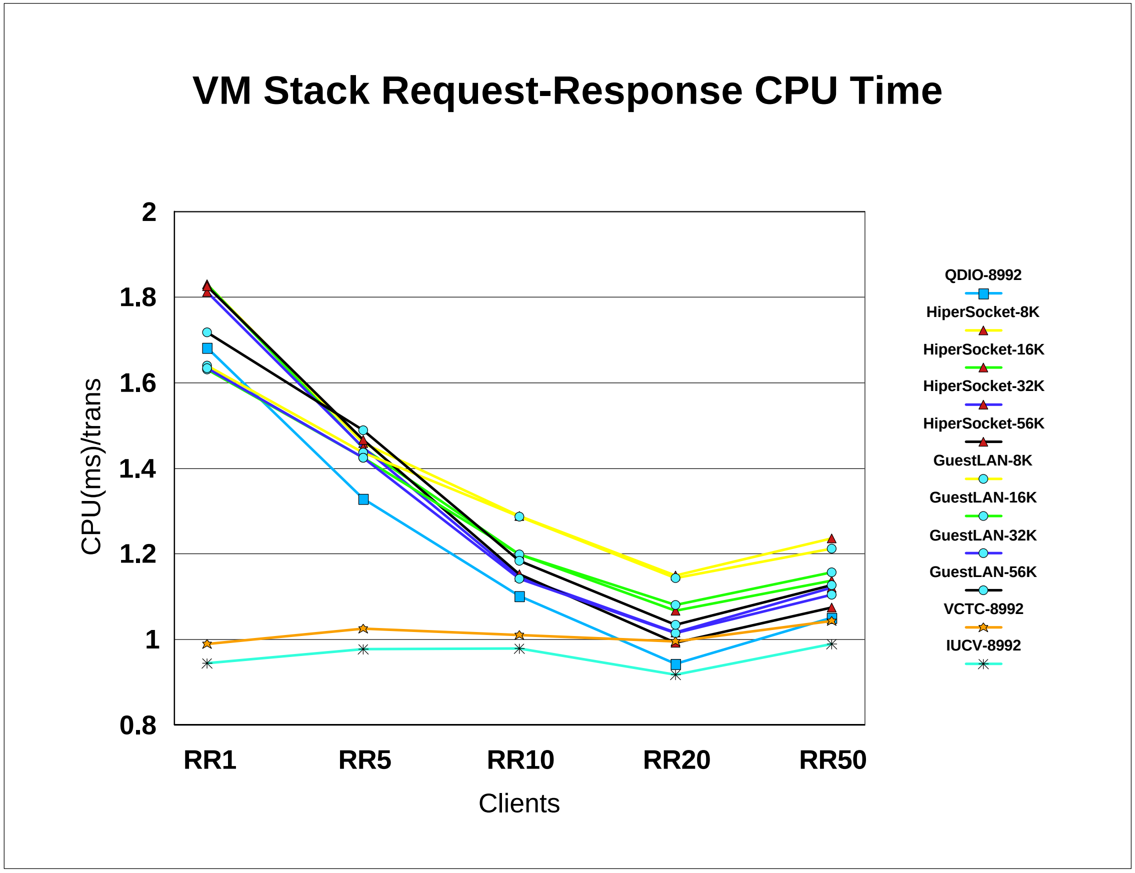


2064-109; dedicated 2-way LPAR

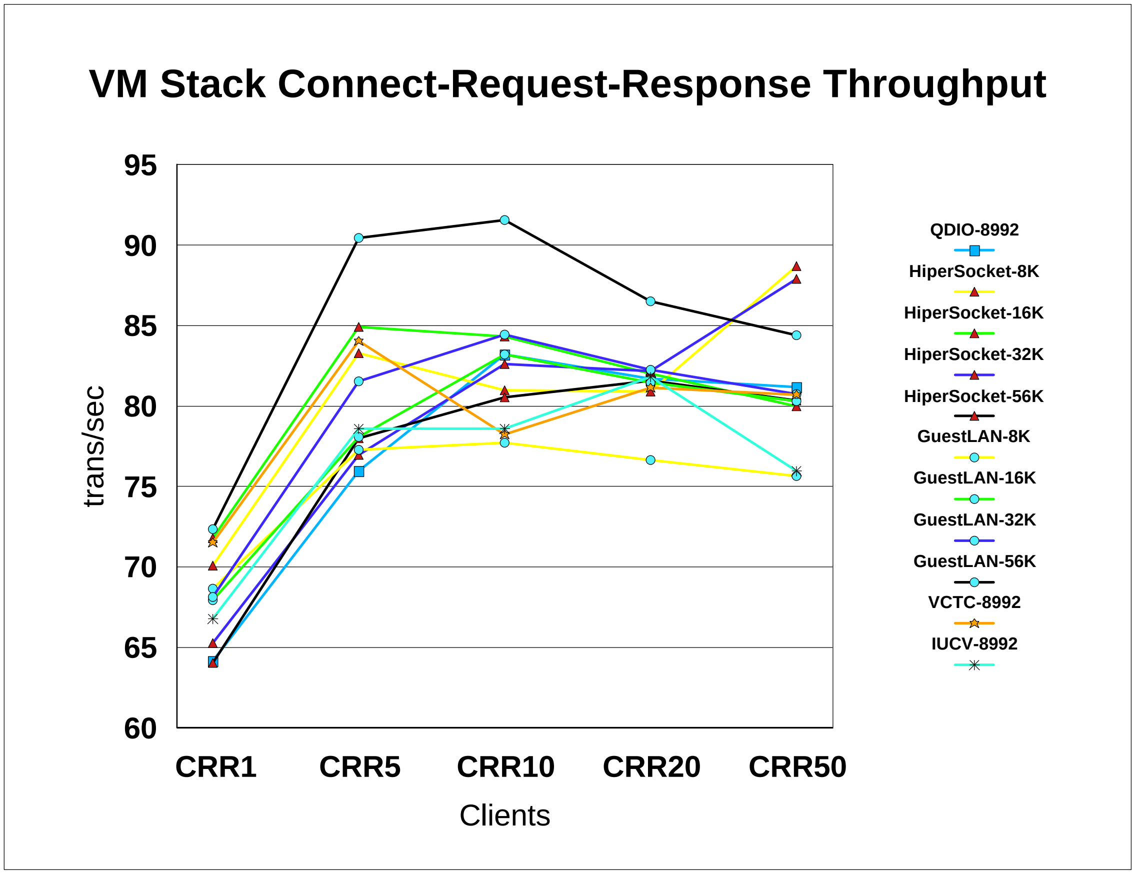
- Internal Network driver is used for these measurements.



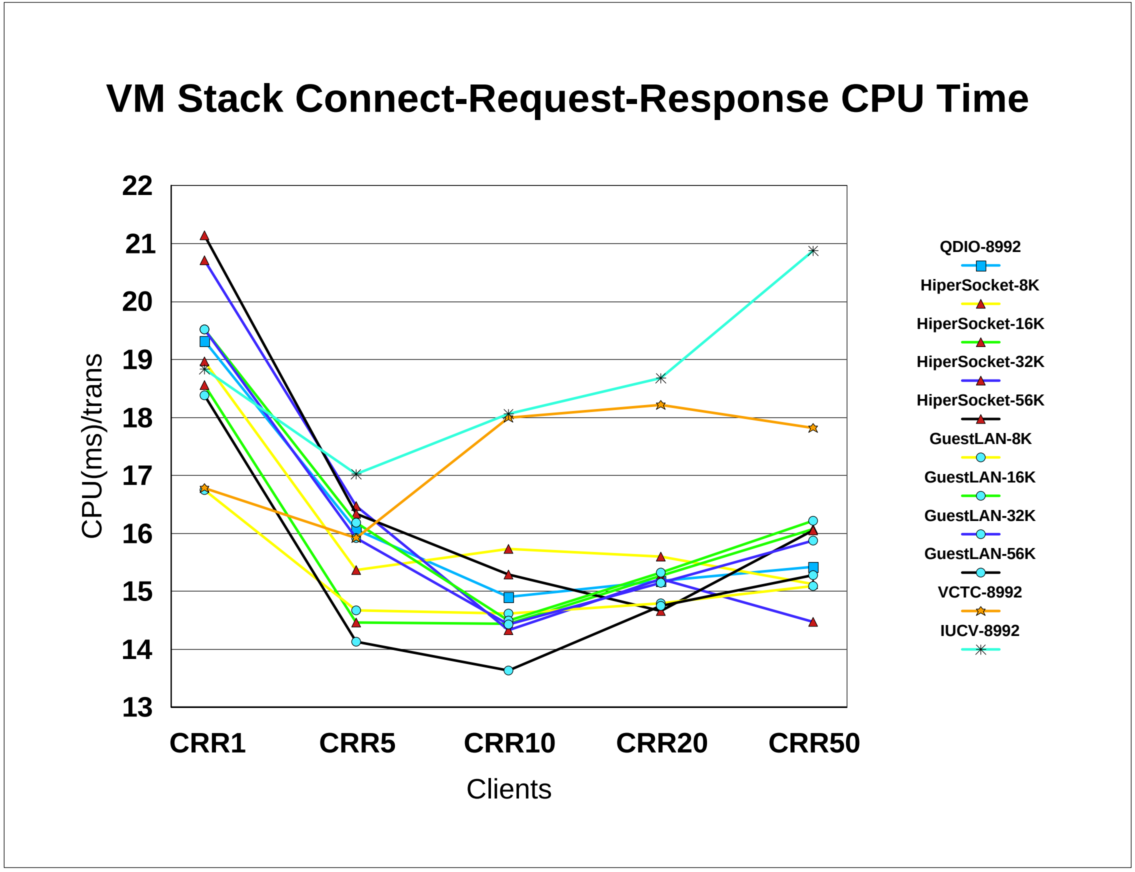
- Legacy connection type IUCV and vCTC have highest throughput rates.
- MTU size does not impact results sufficiently.
- IUCV and vCTC plateau where the total processor time is close to 100%.
- The network driver tool was a significant part of the workload here since the datatransfer and processing by stacks were smaller than other workloads.



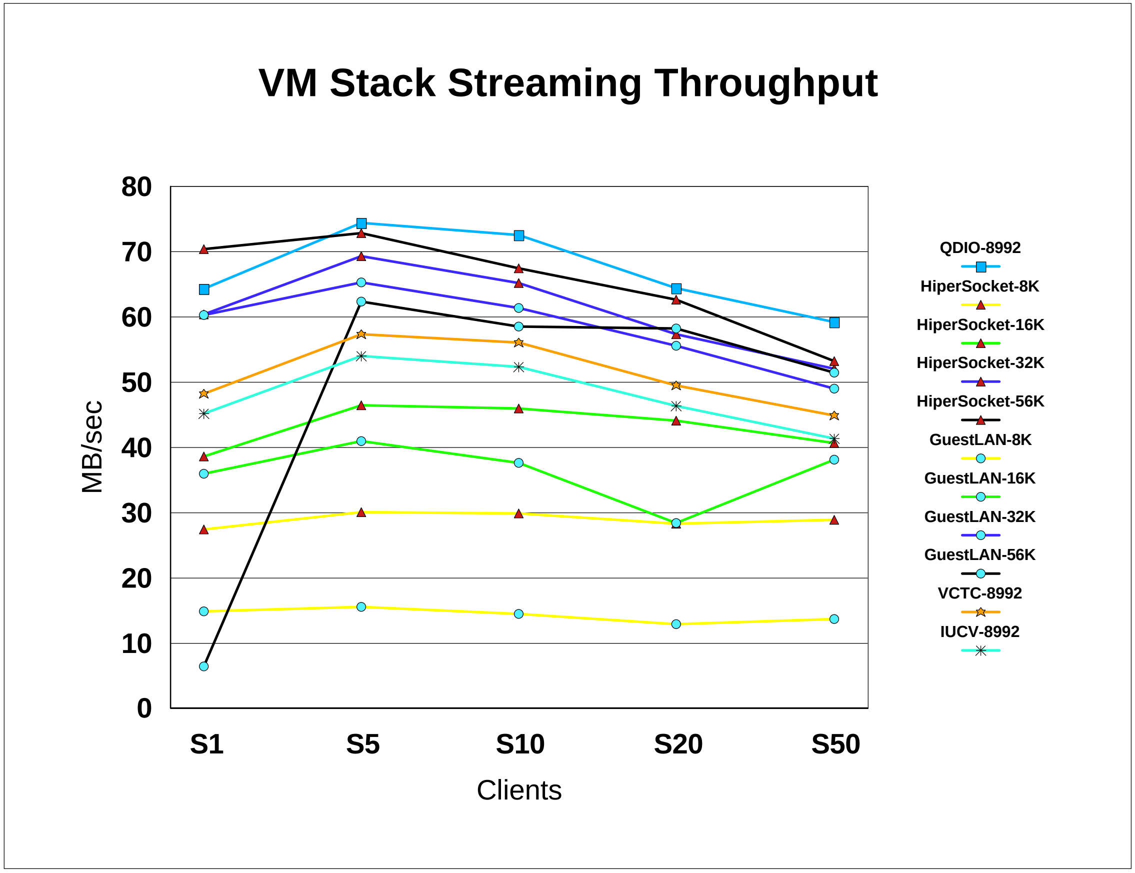
- The legacy connection types have flat processor costs, while the many others become more efficient as load increases. This increased efficiency leveled off around 20 clients.
-



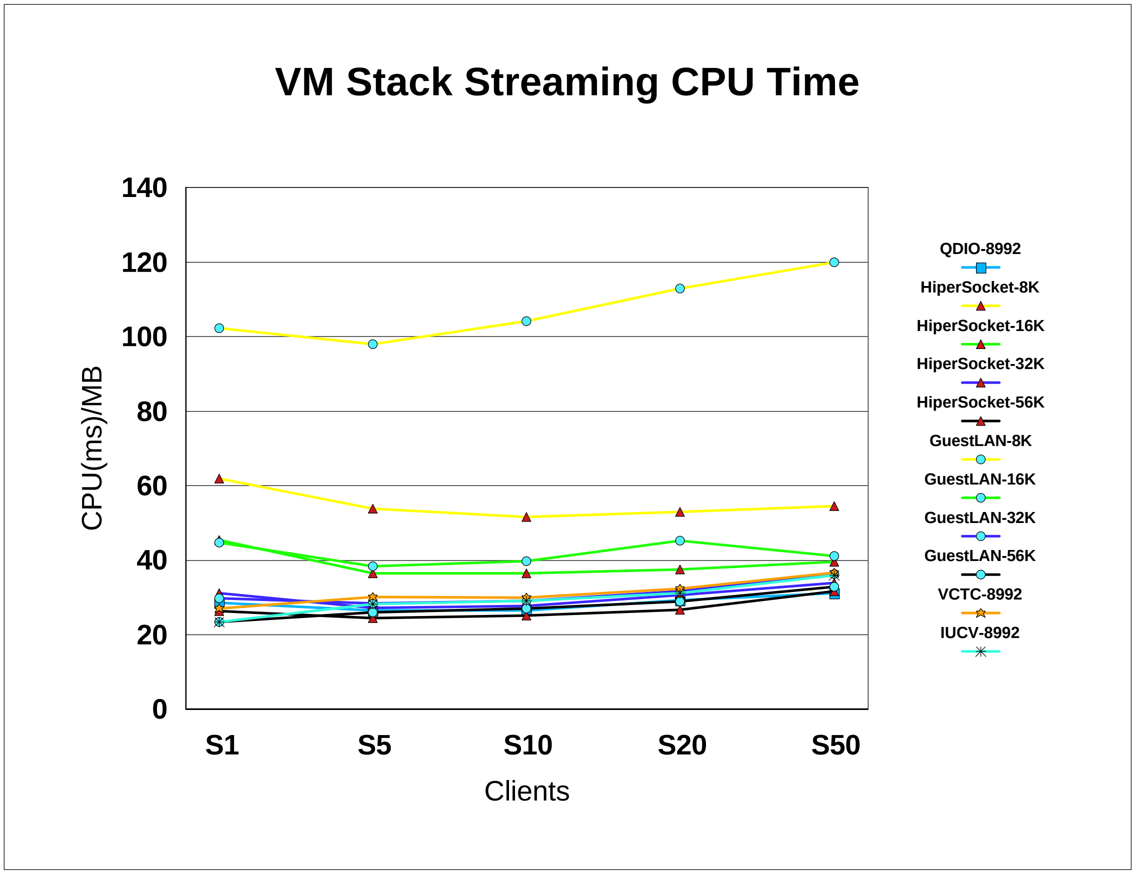
- Throughput in the CRR environment was less than RR due to the high costs of connect processing.
- IUCV and vCTC have lost their edge in the CRR environment since most of the optimization over the years had been on data movement, not on connection processing.
- Guest LAN has some advantages in CRR, and Hipersockets also does well.
- For CRR measurements, the stack closest to the clients tended to be the bottleneck and used significantly more resources than the other stack.



- As discussed earlier, the connection overhead makes the cost of CRR higher than for RR workloads.
- Some efficiencies can be gained with multiple clients.

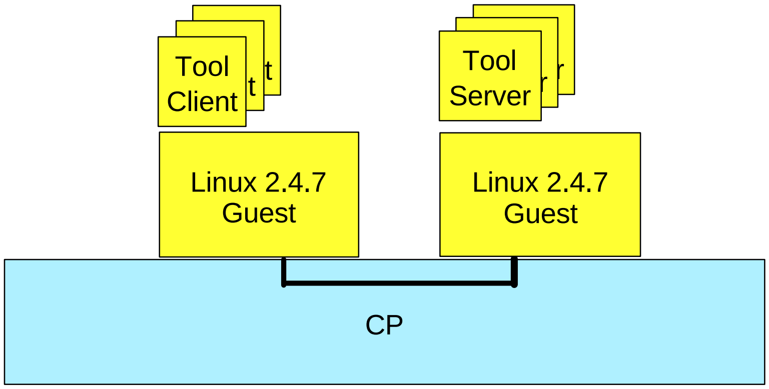


- Obviously, there is some anomaly with the Guest LAN run with 56K MTU at the 1 client level. Other than this one case, multiple connections do not significantly change throughput.
- QDIO GbE and Hipersockets do well.
- MTU size has a large impact in streaming workloads than in the earlier RR and CRR measurements.
- The Streaming workloads tended to be limited by the stacks as often more than 95% of their time was spent running or waiting on the CPU.



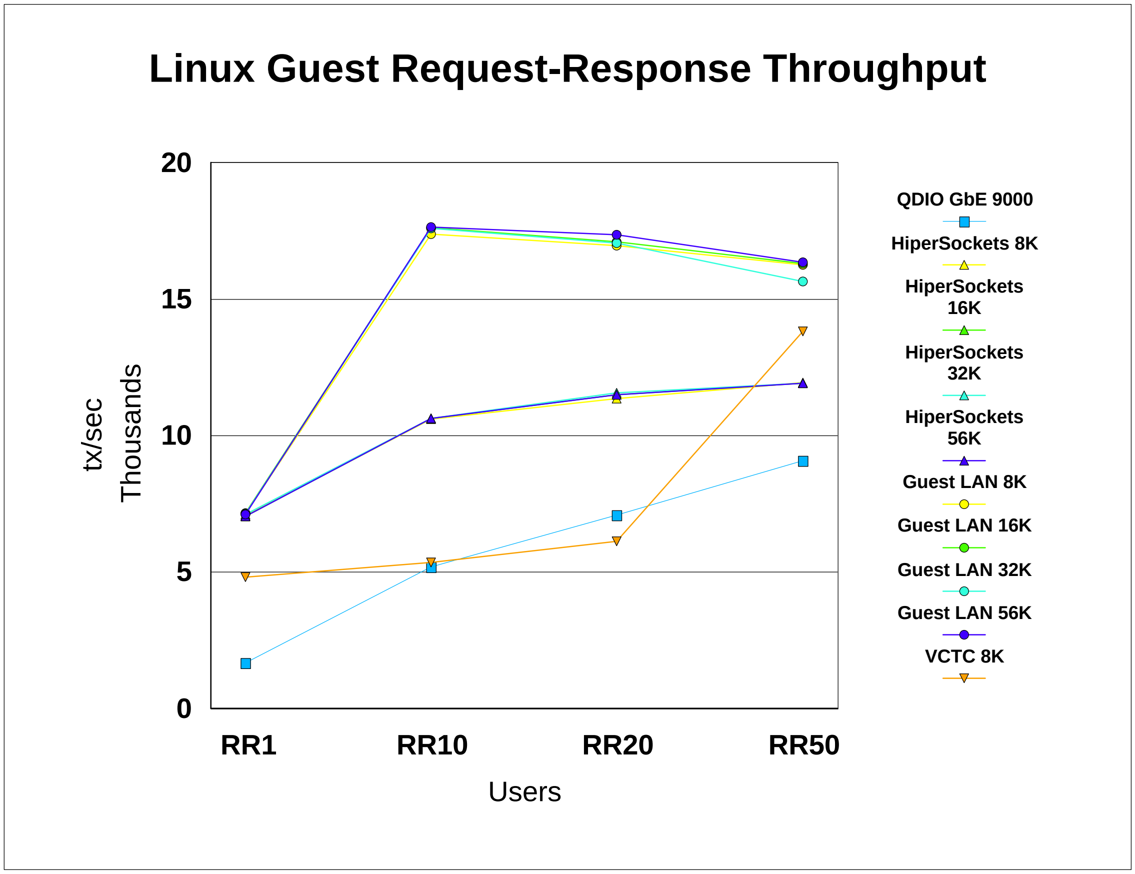
- Some unusual extra overhead exists for the Guest LAN with 8K MTU size.
- The strange 1 client Guest LAN 56K anomaly does not appear to be related to efficiency changes as the processor time for the various number of clients on Guest LAN does not differ.
- Since MTU size is a factor for streaming workloads, the vCTC and IUCV may do better with higher MTU sizes.

Linux Measurements

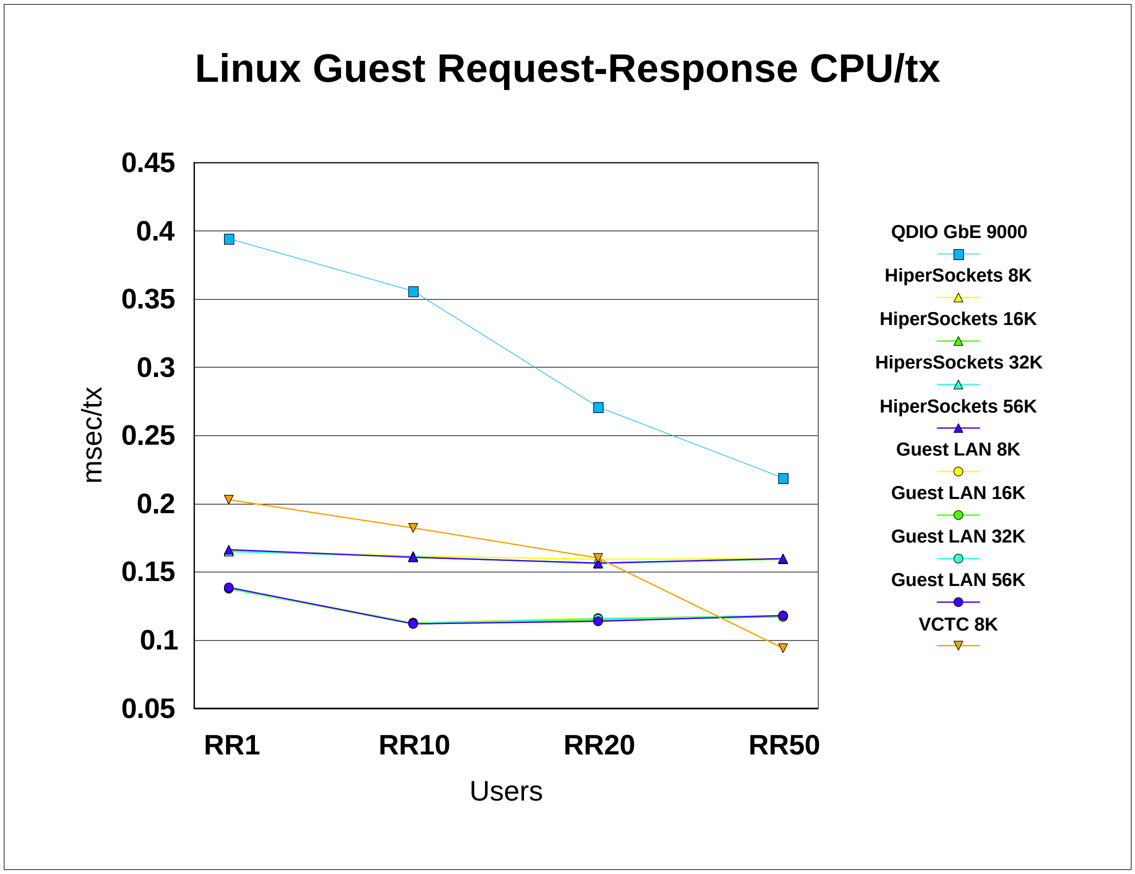


2064-109; dedicated 2-way LPAR

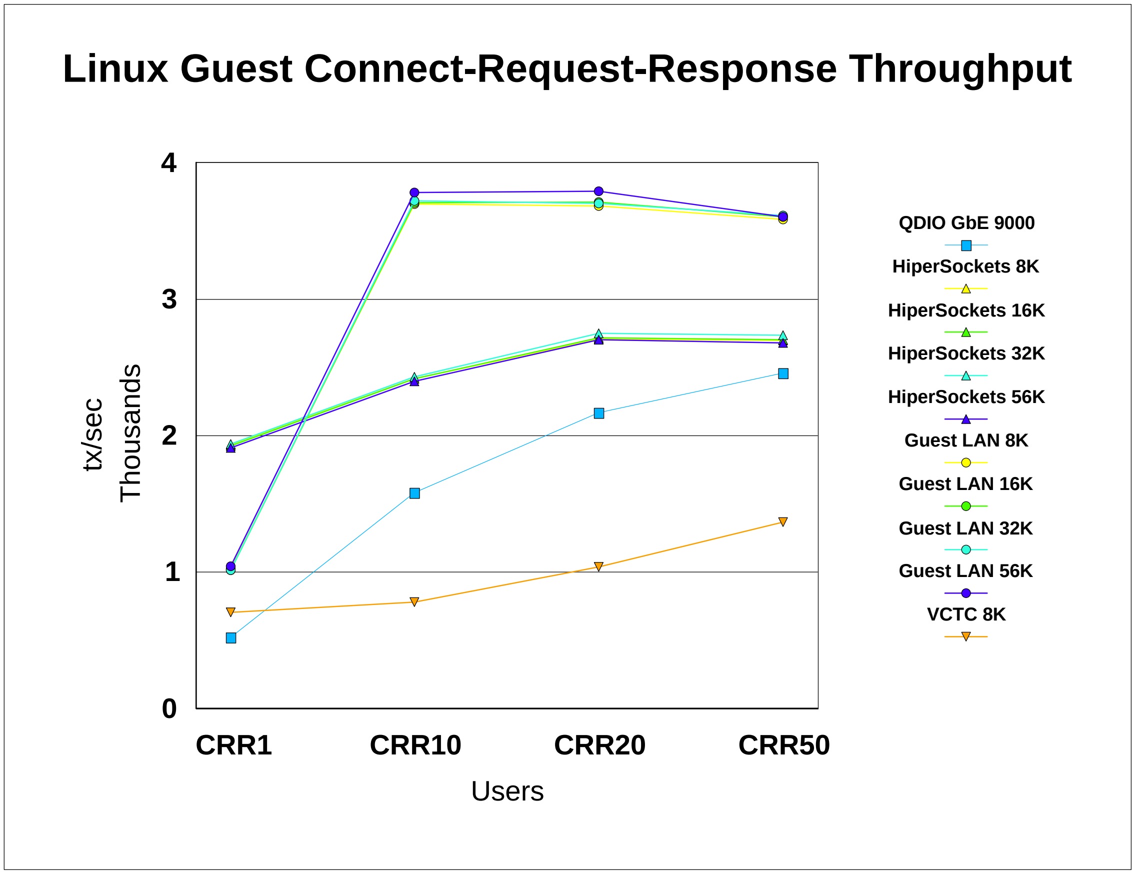
- A key difference between the measurements with the VM stack and the Linux measurements is the number of virtual machines. For the VM stack measurements each client and server were unique pairs of virtual machines. In the Linux measurements, the clients were multiple processes inside a single Linux guest virtual machine. Likewise for the server machines.



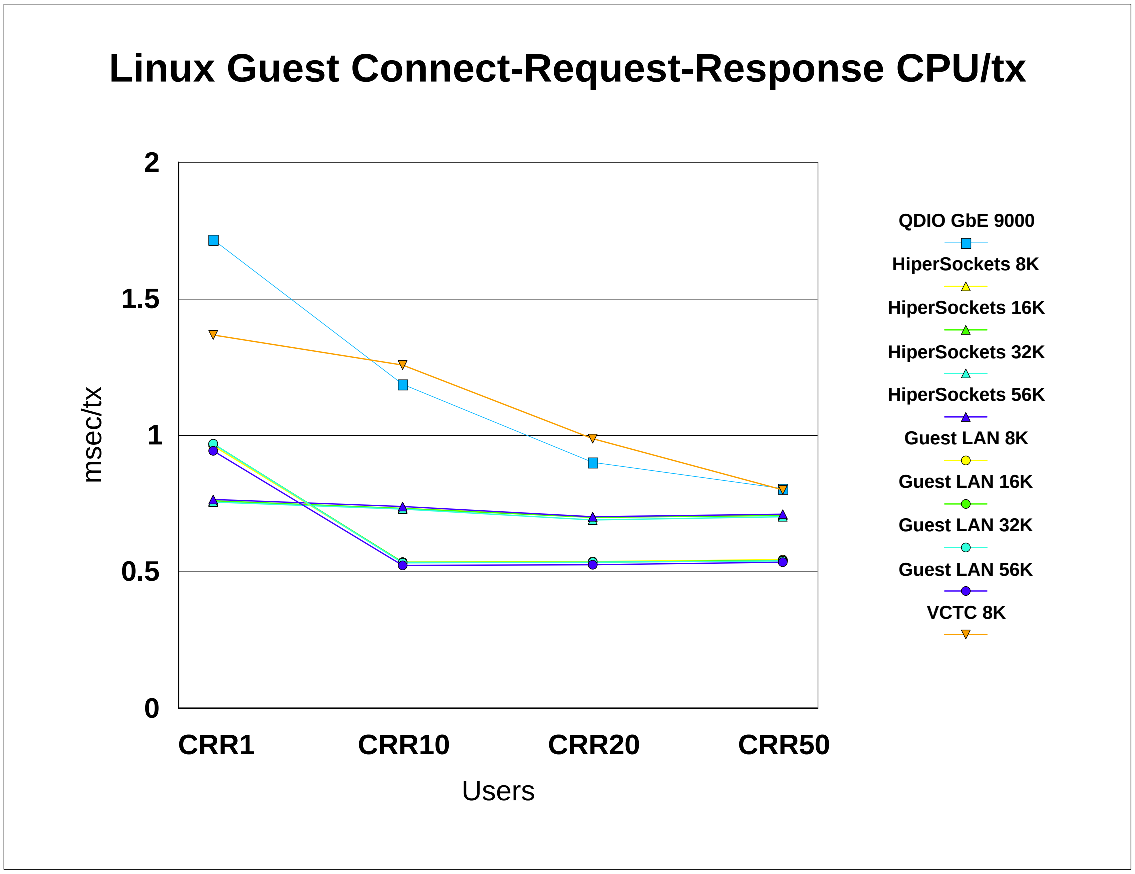
- Unlike the VM stack scenarios, the Guest LAN and Hipersockets connectivity options have far better performance than the QDIO GbE option.
- The throughput increases with multiple connections.
- vCTC appears to have an advantage for very large number of connections.



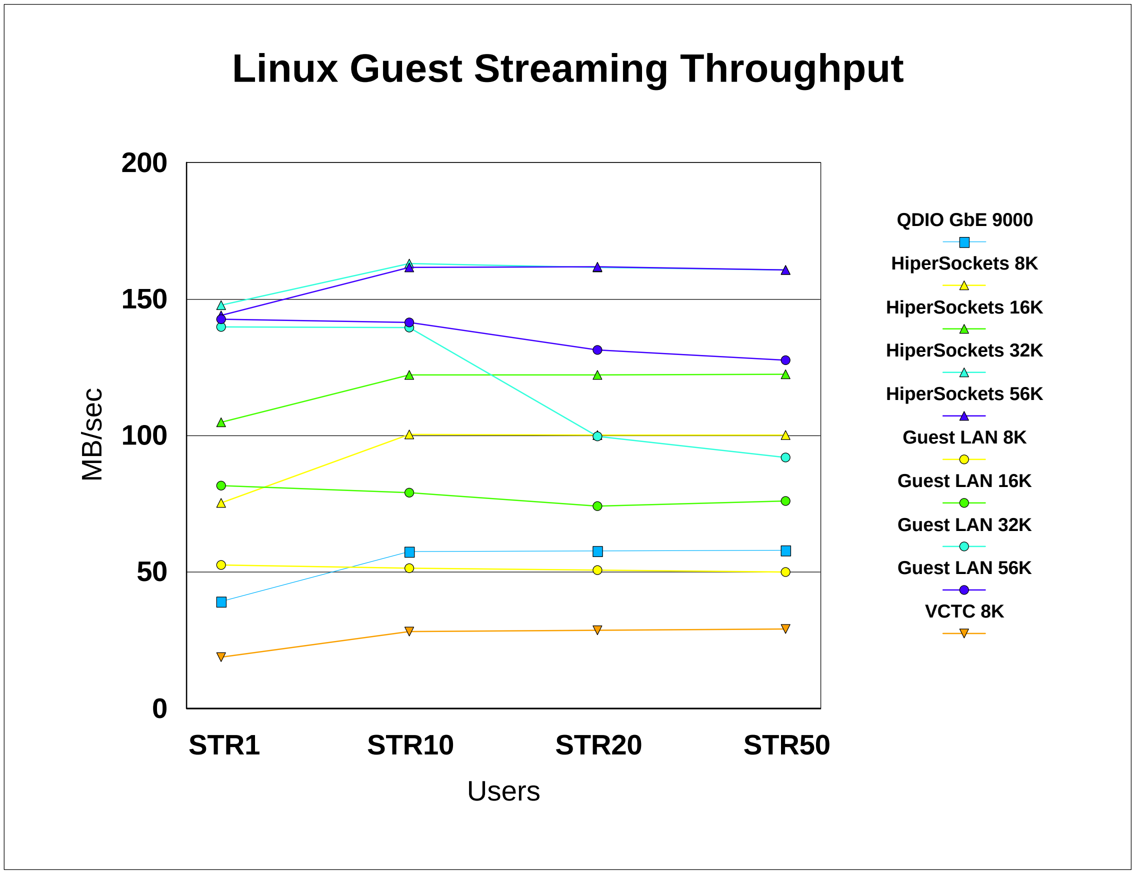
- The efficiency of QDIO GbE improves significantly as the load increases. This also applies to Hipersockets and Guest LAN, but at a far lesser degree.
- Efficiency gains in vCTC vary on number of connections.
- MTU size again does not impact the results significantly.



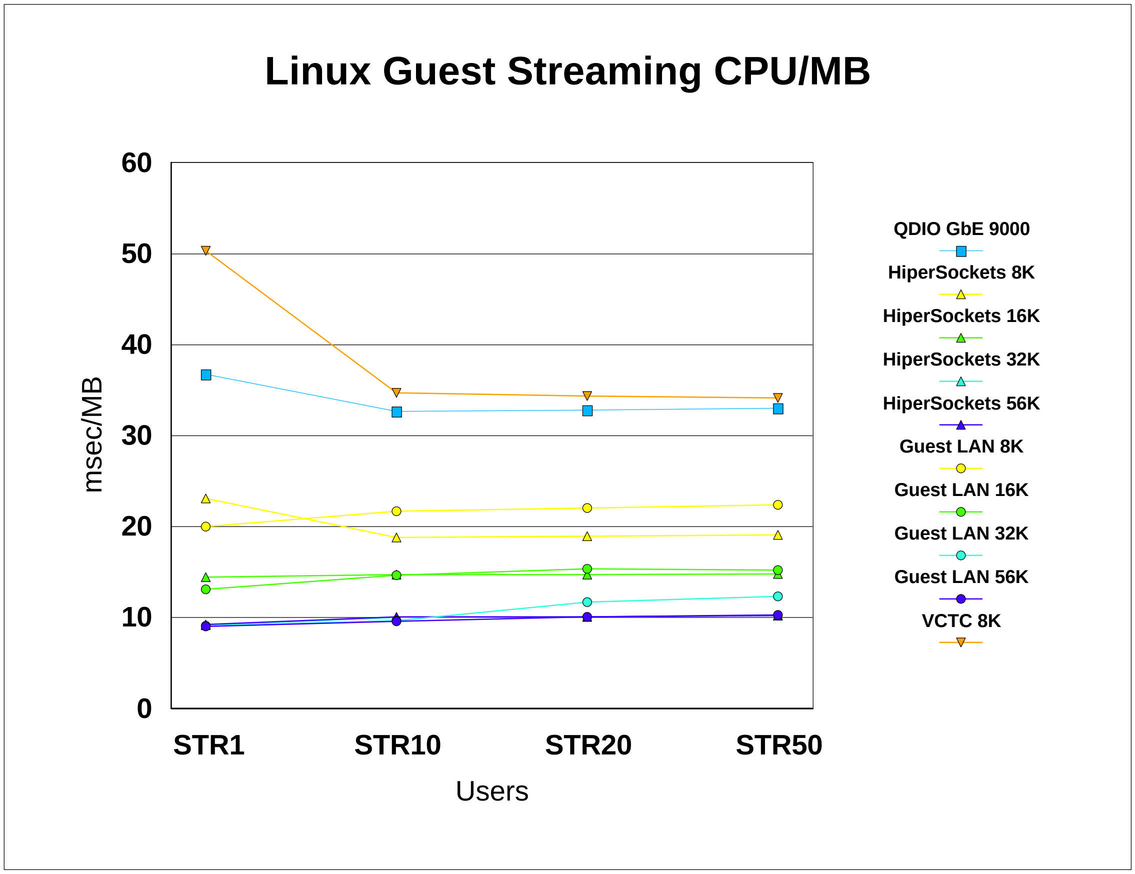
- Guest LAN again does better than Hipersockets in a CRR environment.
- The vCTC connection costs are high and do not do well in the CRR workload. It is interesting that the curve appears to be inverse of some of the other connectivity types.
- Multiple users increase throughput through efficiencies gained as seen in the next chart.



- The greatest improvements in processor efficiency are seen in QDIO GbE and Guest LAN.



- MTU size becomes important in the streaming workloads.
- Here Hipersockets turns in the best results, with Guest LAN following. The vCTC turns in the worse performance for streaming, though if larger MTU sizes were used, the throughput would increase.



- This chart shows the efficiencies gained with larger MTU size.
- The change in efficiency with greater number of clients is not as significant as seen in RR and CRR

Hipersockets- Final Thoughts

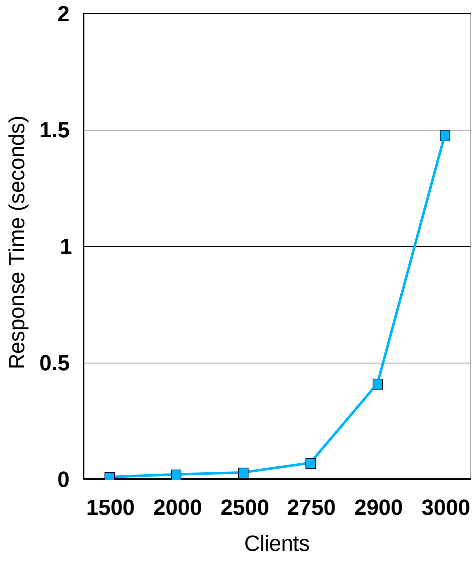
- Guest LAN and Hipersockets are improvements over QDIO GbE.
- Guest LAN
  - + Configuration limits
  - + Storage Requirements
  - - does not work between LPARs
- More efficient as load increases
- IUCV and vCTC become less exciting with the addition of Guest LAN

VM IMAP Server

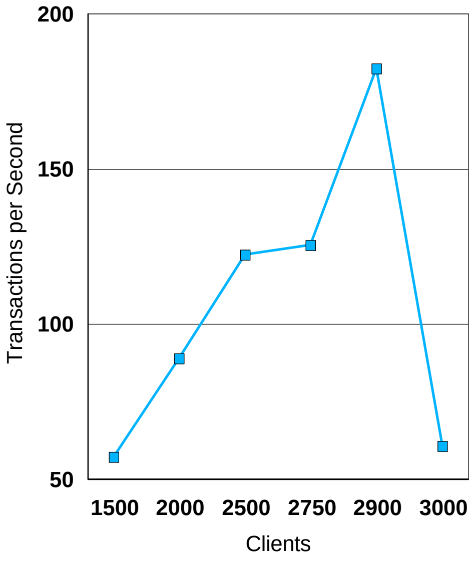
- Internet Message Access Protocol (IMAP)
- Internal CMS tool used to simulate user load against an IMAP server.
- Configuration
  - 2064-109 with Dedicated 2-way LPAR
  - APAR PQ54859 applied - thread priority fix
  - SFS configured with USERS value of 4000
- Response time captured by internal tool

IMAP Results

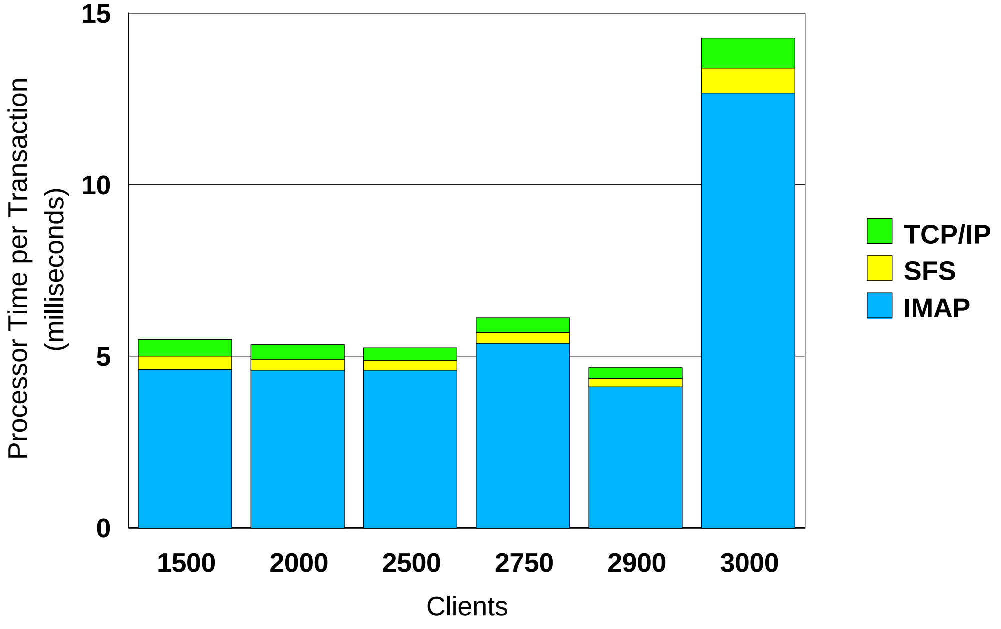
Response Time



Throughput



IMAP Processor Time

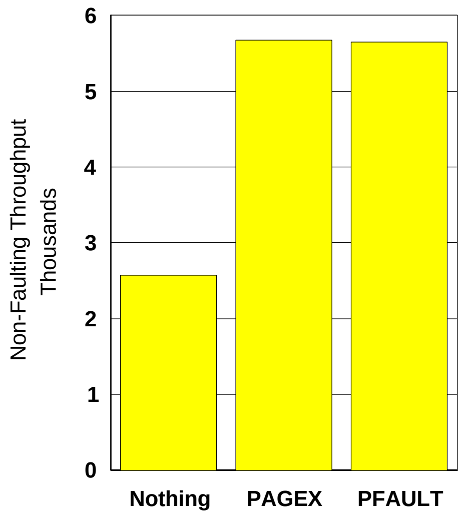


Asynchronous Page Fault Faciltiy

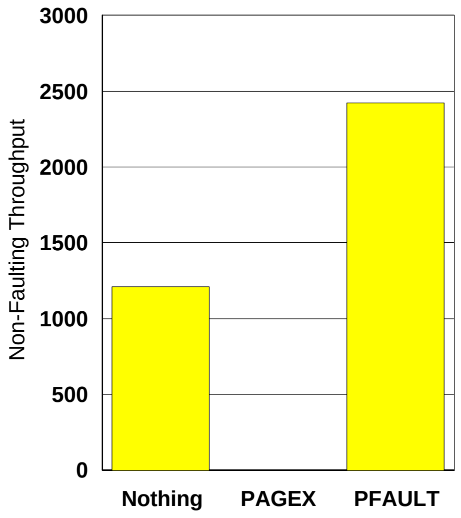
- Ordinarily, page faults serialize the virtual machine. This can be a throughput and response time problem for guest systems
- Enhancements designed for Linux
- PFAULT macro
  - Accepts 64-bit inputs
  - Provides 64-bit PSW masks
- Diagnose x'258'
- Older PAGEX interface limited to 31-bit

Page Fault Tests

31-bit Scenarios



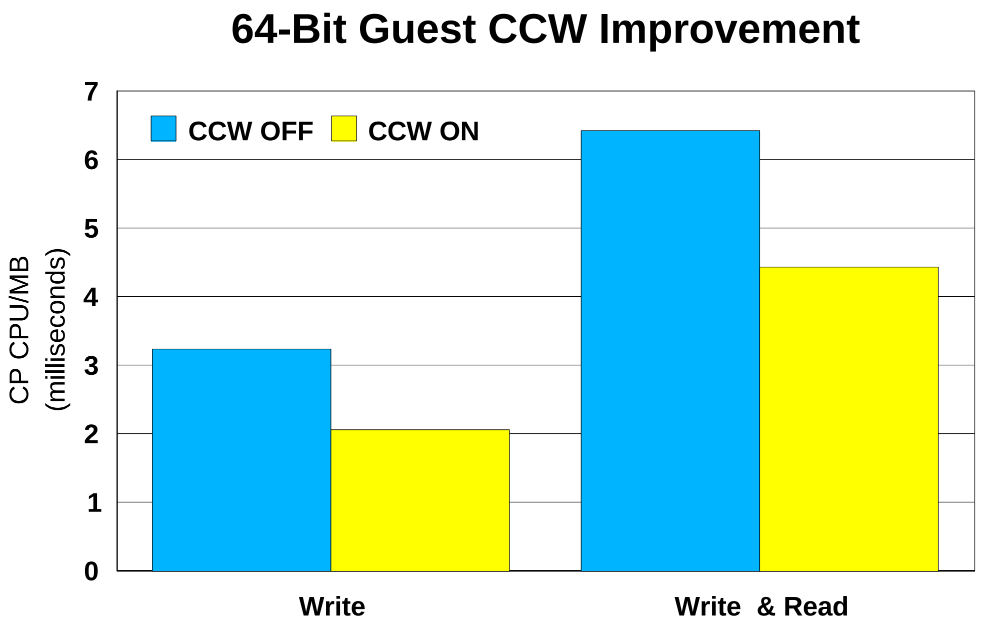
64-bit Scenarios



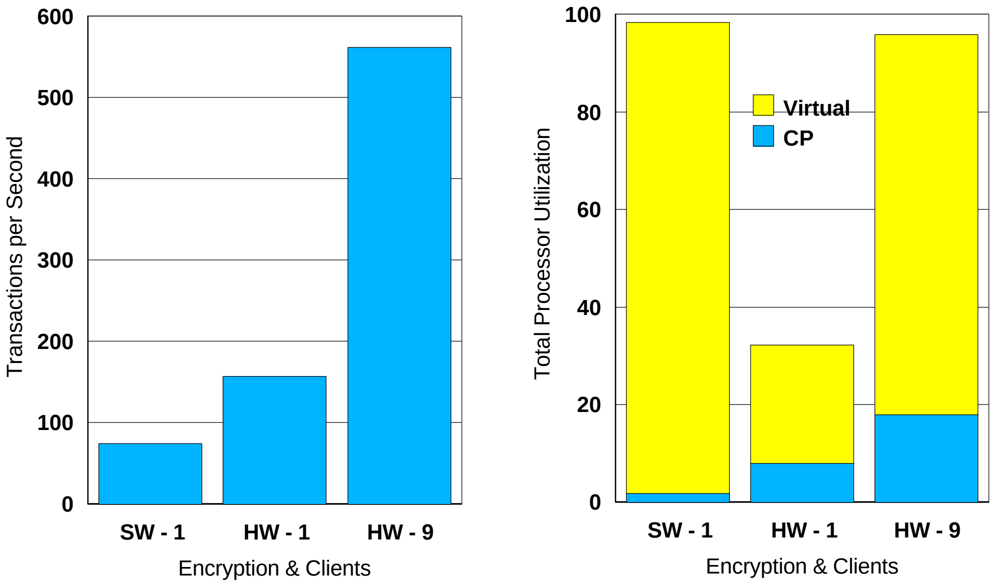
Fast CCW Extensions

- Channel programs with format 2 (64-bit) IDAW previously ineligible for fast path CCW translation and MDC
- This item extends support to cover this.
- Limitations
  - FBA devices with format 2 IDAWs are eligible for fast translation, but not MDC
  - format 2 IDAW that works on 4K boundary supported, but not those for 2K boundary
- Good results

Format 2 Fast CCW Translation

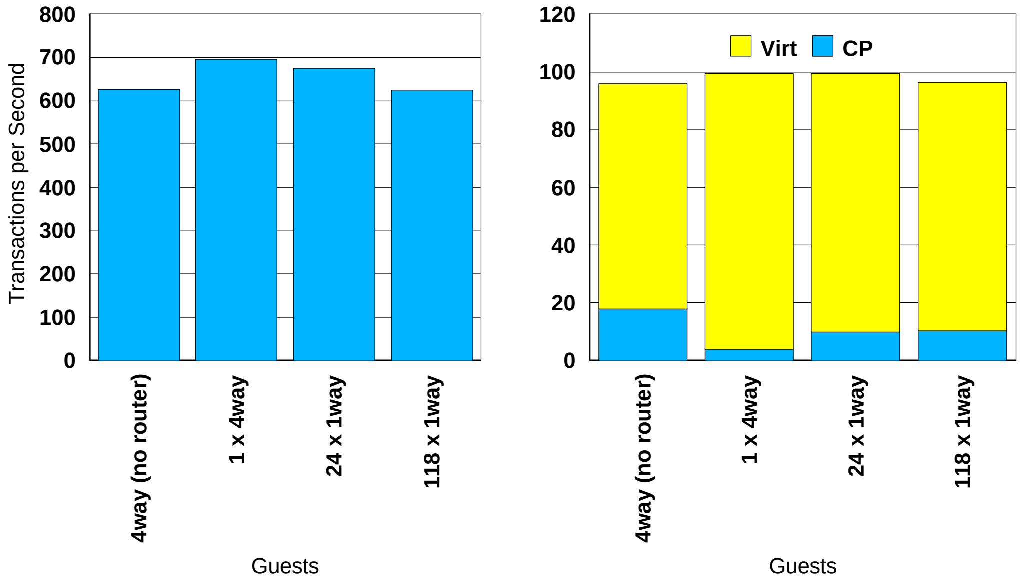


Linux Guest use of Crypto Support



- Processor: 2064-109, 4-way LPAR
- Crypto hardware: 1 PCICA card
- Workload: SSL Exerciser; RC4 MD5, Non-cached 1024 bit keys.
- The chart on the left shows the rate of transactions of uncached SSL transactions. The first column is with encryption being done in software with 1 client (20 threads). The second column shows the performance benefit of the hardware encryption. The last column is when the number of clients is increased to bring the total processor utilizaiton closer to 100%.

Multiple Linux Guests with Crypto



- For a single Linux guest, it owned the network GbE, this is shown as the first column labeled "no router". The other three columns deal with using a VM stack as a router and changing the number of VM guests and the virtual number of processors.

z/VM 4.3.0 Overview

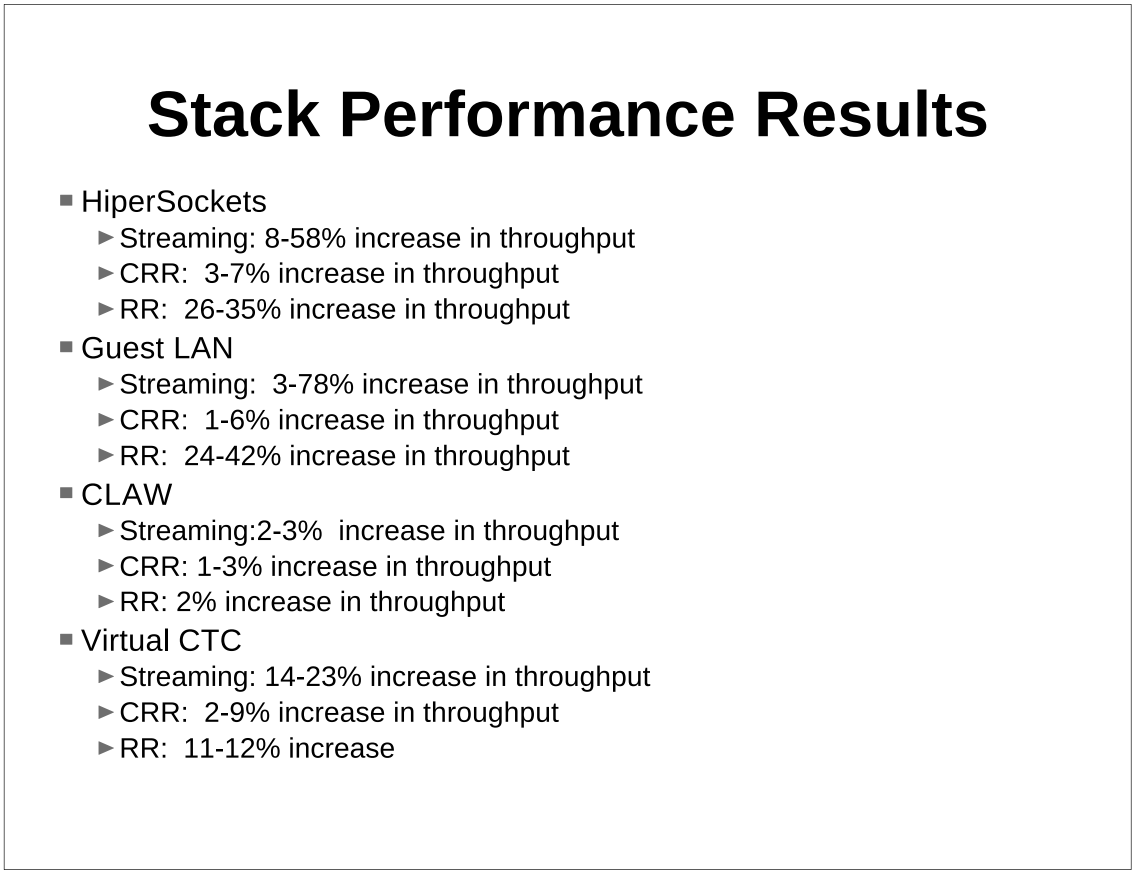
- GA May 31, 2002
- Network improvements
  - VM TCP/IP Stack enhancements
  - Guest LAN use via QDIO
- Linux related enhancements
  - CP timer management
- Other improvements
  - Managing contention for storage under 2GB
  - Large volume CMS minidisks
- Regression performance equivalent to 4.2.0

CP Timer Management

- Improvements for environments where a large number of guests are using the clock comparator interrupts at high frequencies (Linux guests)
- Prior to z/VM 4.3.0, much of this processing was tied to the master processor.
- These ties to the master processor have been removed.

Stack Performance Improvements

- Improvements to various device drivers:
  - HiperSockets, Guest LAN, CLAW, vCTC
- Decrease in processor time requirements depending on workload often resulted in greater throughputs.
  - Streaming: 14-45% decrease in cpu/MB
  - CRR: 2-9% decrease in cpu/transaction
  - RR: 10-26% decrease in cpu/transaction
- In general, configurations with smaller MTU sizes showed greater improvement.



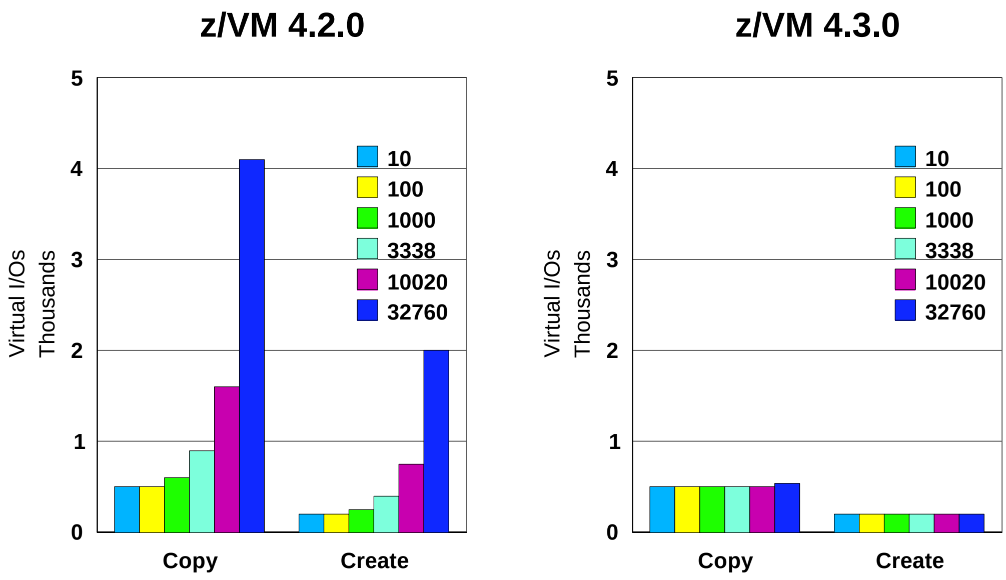
Stack Performance Results

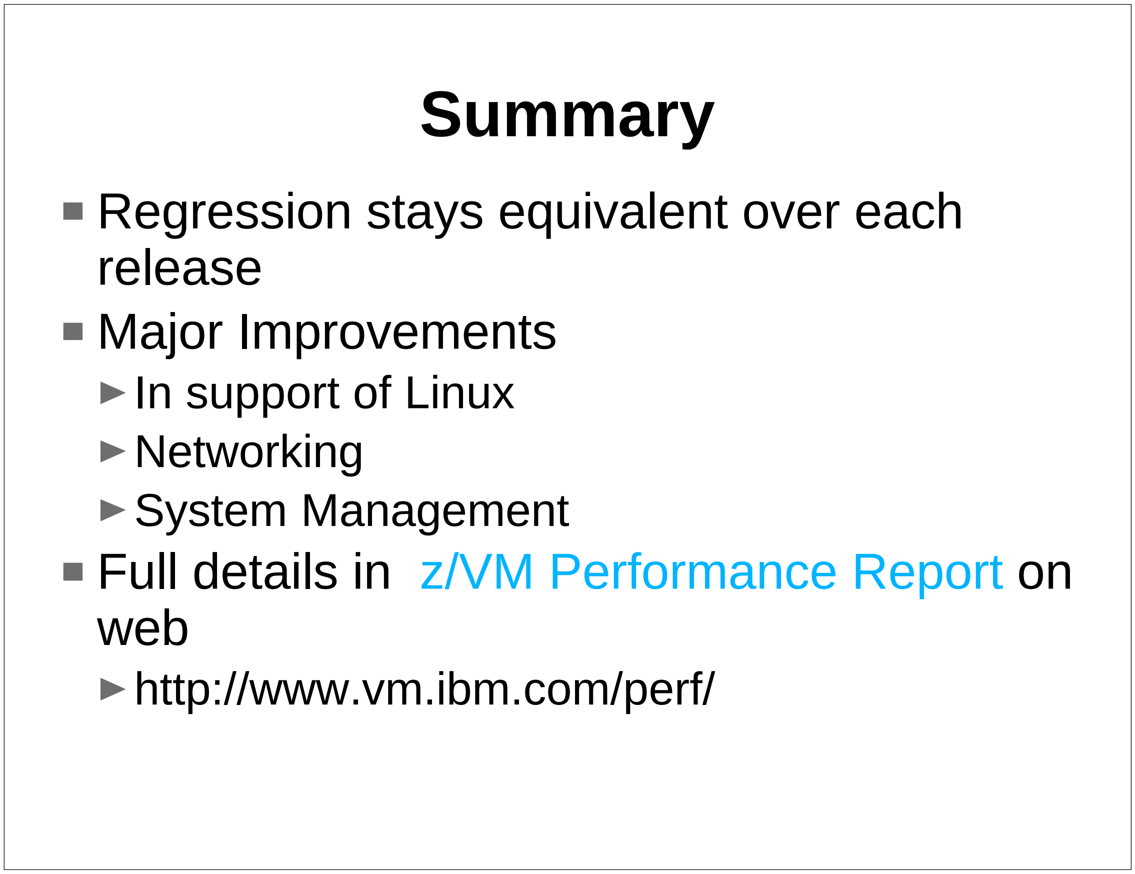
- HiperSockets
  - Streaming: 8-58% increase in throughput
  - CRR: 3-7% increase in throughput
  - RR: 26-35% increase in throughput
- Guest LAN
  - Streaming: 3-78% increase in throughput
  - CRR: 1-6% increase in throughput
  - RR: 24-42% increase in throughput
- CLAW
  - Streaming:2-3% increase in throughput
  - CRR: 1-3% increase in throughput
  - RR: 2% increase in throughput
- Virtual CTC
  - Streaming: 14-23% increase in throughput
  - CRR: 2-9% increase in throughput
  - RR: 11-12% increase

Contention for under 2GB

- Prior to z/VM 4.3.0, contention for storage under 2GB was managed by stealing and paging out appropriate pages from that area to expanded storage (if it existed) or to paging DASD.
- This could result in significant paging to DASD if no, or insufficient, expanded storage was configured.
- In z/VM 4.3.0, CP will attempt to move the selected page from under 2GB to central storage above 2GB if there is room.
- This mitigates problems caused by insufficient expanded storage.
- It is still recommended that some processor storage be configured as expanded storage.

Large Volume Support - CMS





Summary

- Regression stays equivalent over each release
- Major Improvements
  - In support of Linux
  - Networking
  - System Management
- Full details in z/VM Performance Report on web
  - http://www.vm.ibm.com/perf/

Thanks!

- Customers & Vendors
- VM Performance Team
  - Cherie Barnes
  - Dean DiTommaso
  - Wes Ernsberger
  - Bill Guzior
  - Patty Rando
  - Brian Wade
- VM Team
- Worry Coverage
  - Jim McCormick
  - Fred Shaheen

- Linux Performance
  - Martin Kammerer & Team
  - Chris Panetta
  - Eberhard Pasch
  - Donna Von Dehsen
  - Don Corbett
- Crypto Performance
  - Mark Bidwell
  - Virg Meredith
  - Dave Spencer
  - Dave Thornley
  - Joe Tingley
- Hipersockets
  - Bob Perrone