IBM @server™

VM Performance Update - z/VM 4.4.0

WAVV 2004

Bill Bitner
IBM Endicott
bitnerb@us.ibm.com

Last updated April 24, 2004

Trademarks

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

CICS*
DB2
DB2 Connect
DB2 Universal Database
e-business logo*
FICON
HiperSockets
IBM*
IBM logo
MQSeries*
Multiprise*
OS/390
RISC
S/390
S/390 Parallel Enterprise Server*
Virtual Image Facility
VM/ESA*
VSE/ESA
WebSphere
z/OS
z/VM
zSeries

* Registered trademarks of the IBM Corporation

The following are trademarks or registered trademarks of other companies.

Lotus, Notes, and Domino are trademarks or registered trademarks of Lotus Development Corporation.
Tivoli is a trademark of Tivoli Systems Inc.
Linux is a registered trademark of Linus Torvalds.
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.

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.

IBM considers a product “Year 2000 ready” if the product, when used in accordance with its associated documentation, is capable of correctly processing, providing and/or receiving date data within and between the 20th and 21st centuries, provided that all products (for example, hardware, software and firmware) used with the product properly exchange accurate date data with it. Any statements concerning the Year 2000 readiness of any IBM products contained in this presentation are Year 2000 Readiness Disclosures, subject to the Year 2000 Information and Readiness Disclosure Act of 1998.

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.

Agenda

- z/VM 4.4.0 Performance Related Line Items
- Performance Management Changes
- A few misc tidbits
  - Service

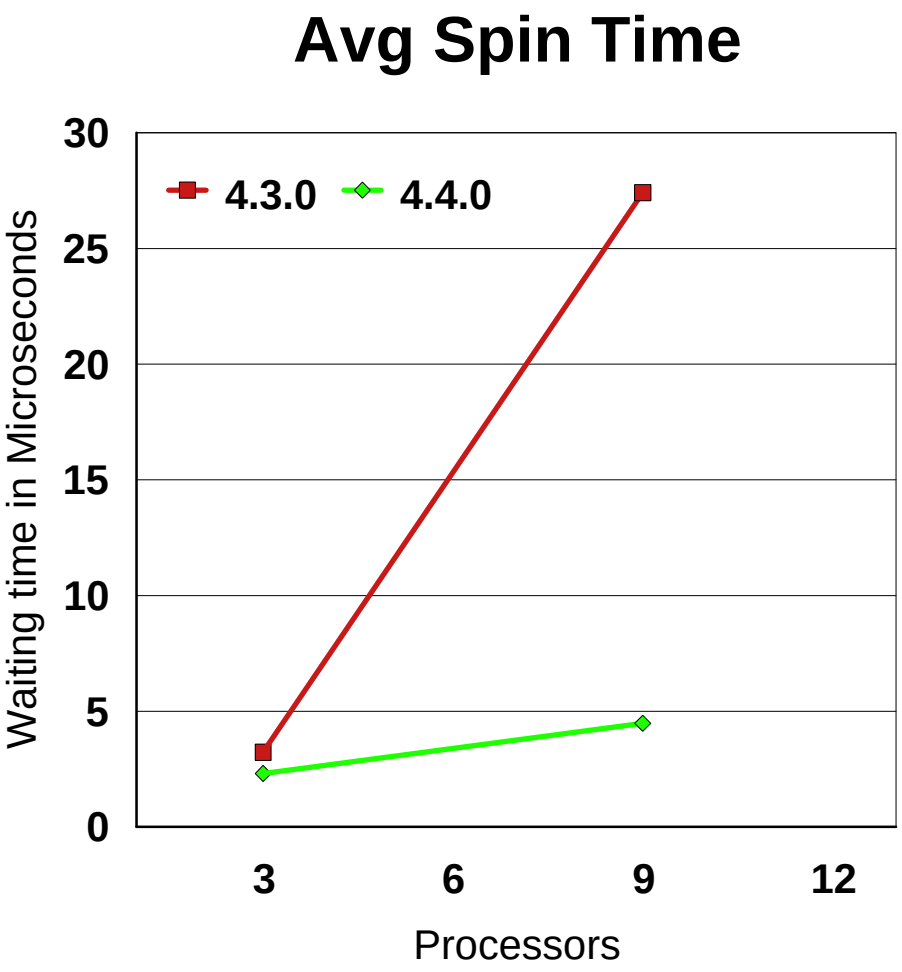
z/VM 4.4.0

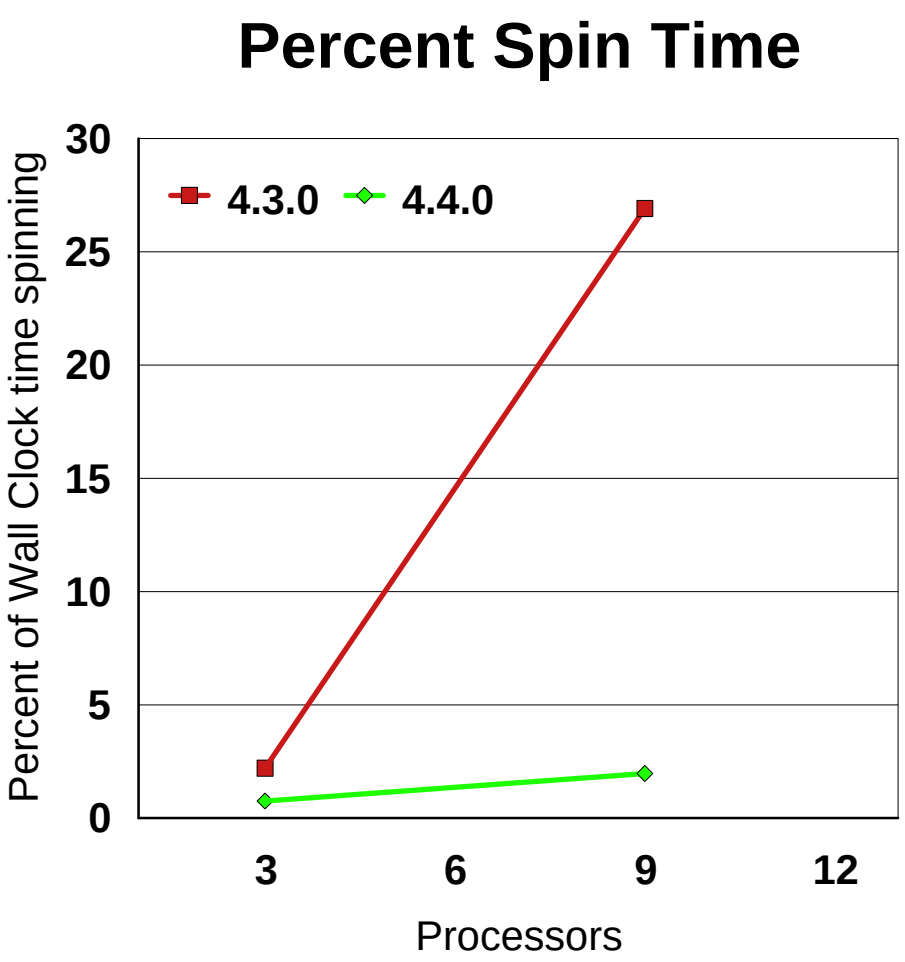
- GA August 15, 2003
- Performance Improvements
  - Scheduler Lock enhancements
  - System Utility Space enhancements
  - TCP/IP enhancements
  - New Virtual Switch Guest LAN
  - Queued I/O assists
- Monitor Changes
- VMRM Changes
- Introduction of Performance Toolkit for VM

Scheduler Lock Enhancements

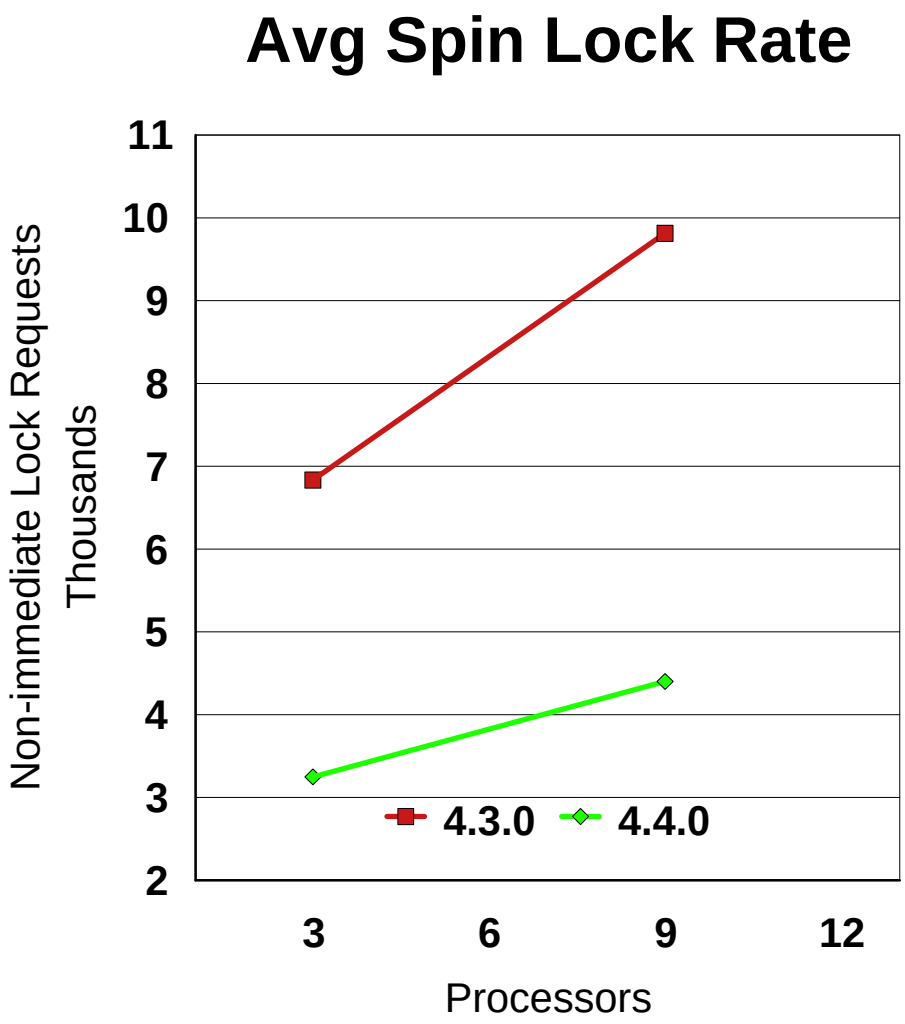
- Scheduler Lock shown to be a bottleneck on large scale testing
- Moved serialization of timer management related functions to a new lock
- Symptoms of scheduler lock contention:
  - High system time (>1 processor worth of time)
    - RTM: %SY on D GENERAL
    - VMPRF: Syst on Processors_By_Time
    - Toolkit: %SYS on CPU Load screen
  - High spin time dominated by scheduler lock
    - RTM: %SP
    - VMPRF: Pct Spin Time on System_Summary2_by_Time
    - Toolkit: %SP on CPU Load screen

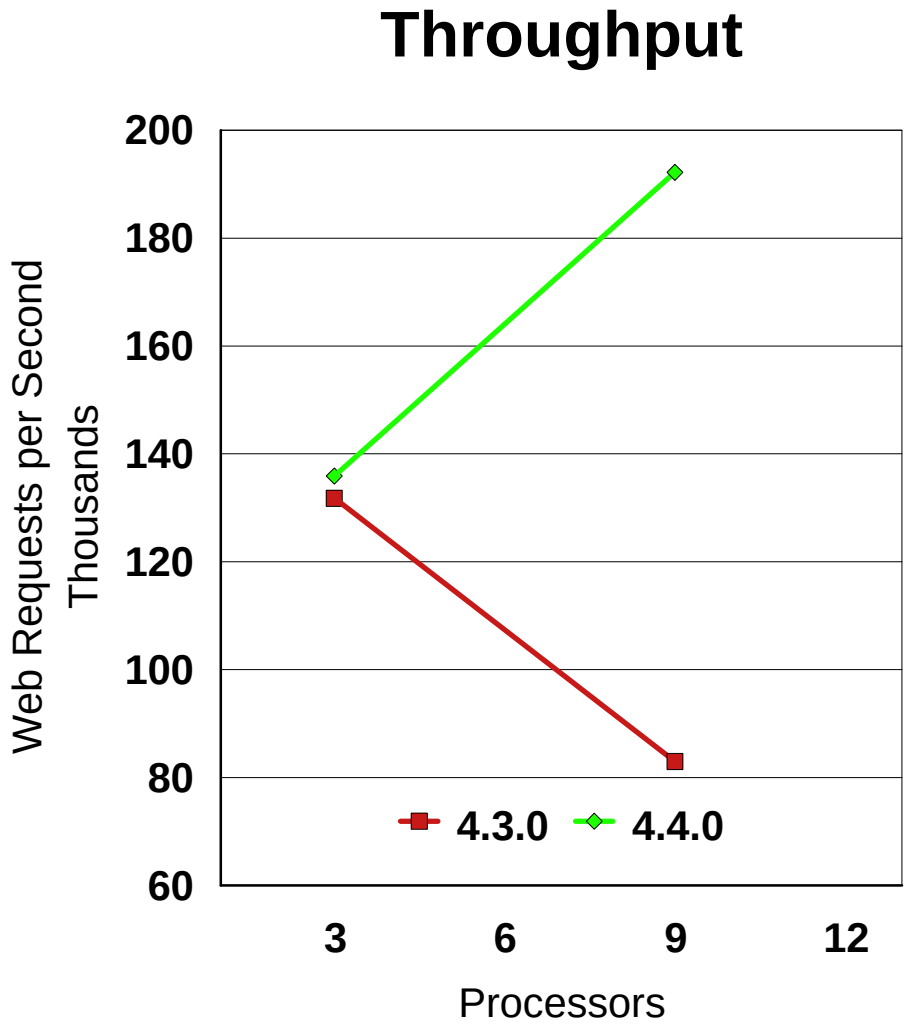
Scheduler Lock - Linux Apache Guest Measurements





Scheduler Lock - Linux Apache Guest Measurements





Scheduler Lock - Linux Crypto Guest Measurements

- 120 Linux Guests with SSL exerciser
  - SLES 8, kernel 2.4.19, 31-bit
  - RC4 MD5 US cypher , No Server SID Cache
  - 128MB
- z990 2084-316
  - 16 processors
  - 12 PCICA Cards
- Performance improvements going from z/VM 4.3.0 to z/VM 4.4.0
  - 73% improvement in ETR
  - 75% improvement in ITR
  - 43% improvement in processor time per command

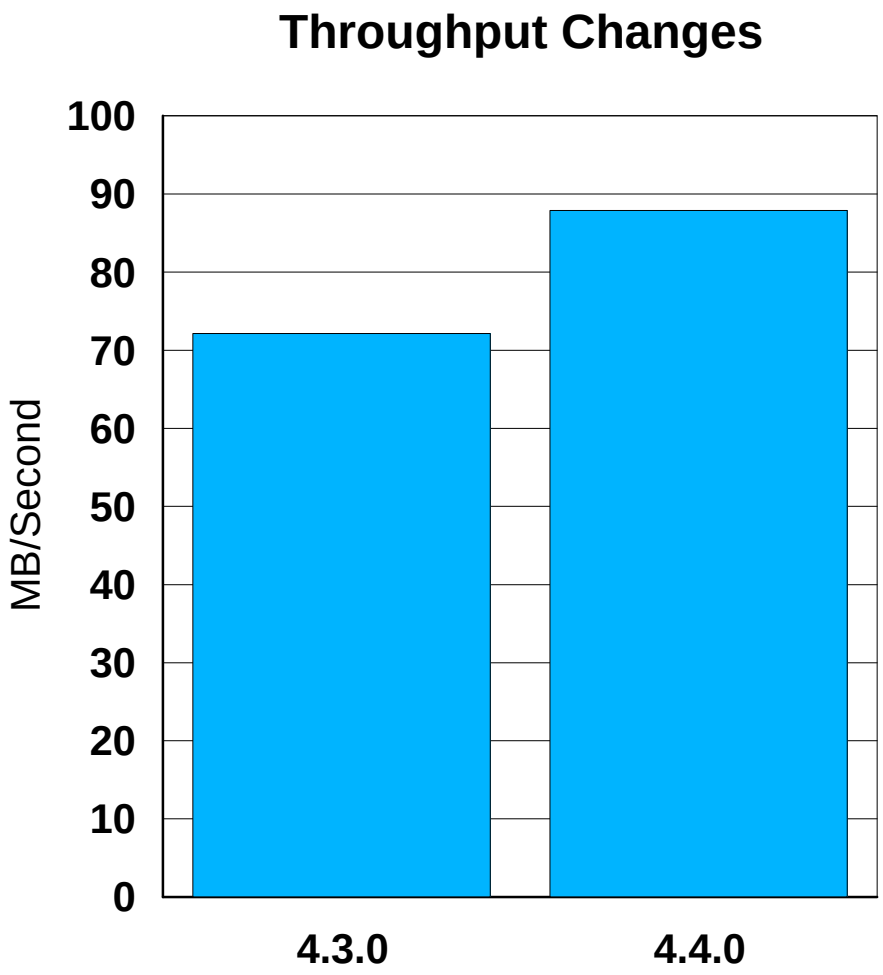
System Utility Space Management Changes

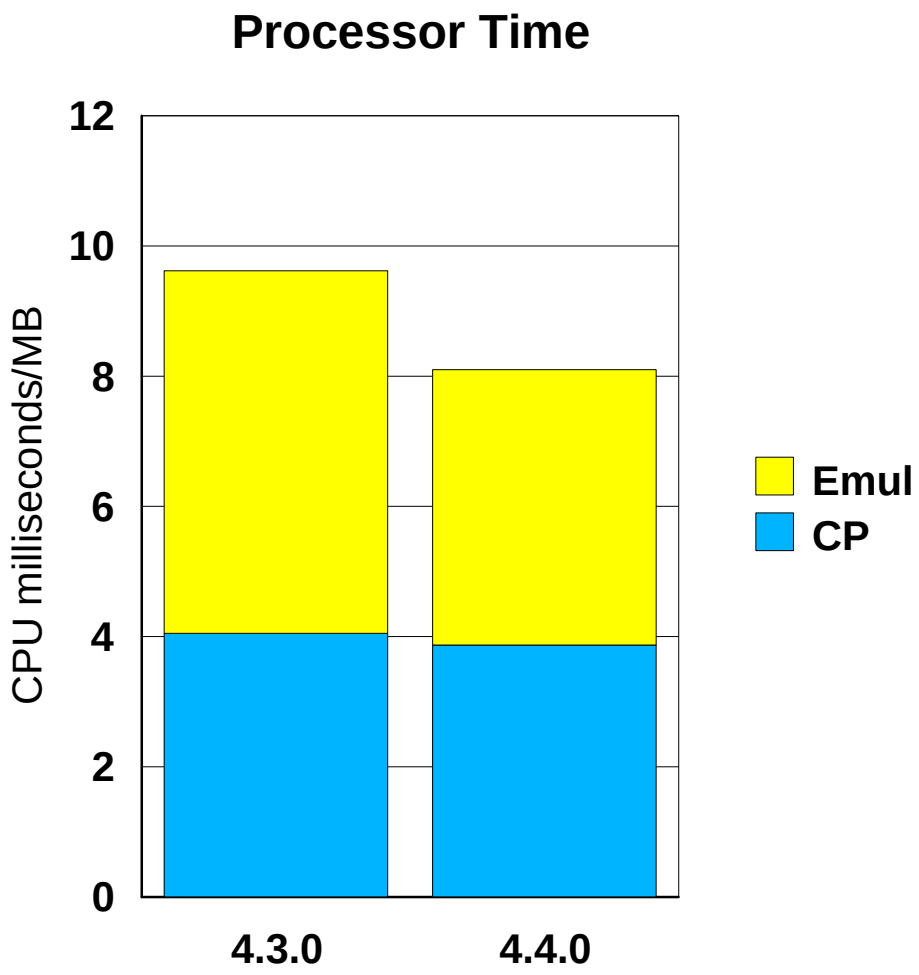
- System owned Utility Spaces
  - address spaces owned by the system
  - Used for virtual disks in storage, virtual free storage, etc.
- Prior to z/VM 4.4.0
  - pages faulted on are brought into storage below 2 GB
  - For virtual disk these are pages that make up the virtual disk blocks
- In z/VM 4.4.0
  - pages are faulted into storage available above 2GB
  - only applies to pages faulted for AR mode access
  - note: Page Management Blocks (PGMBKs) for virtual disks still must reside below 2GB

VM TCP/IP Code Optimization

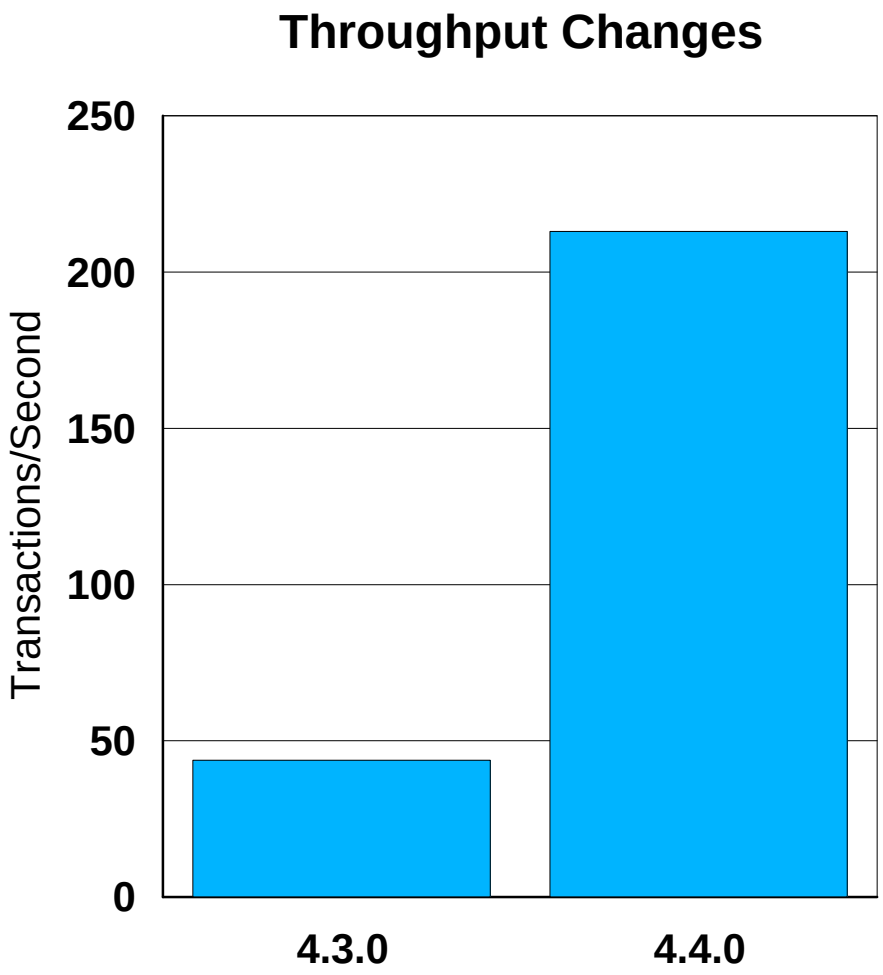
- Follow-on improvements from z/VM 4.3.0
- Hot path analysis in TCP and UDP layers
- Replace some Pascal code segments with Assembler
- Modify algorithms
- Lowers processor time per transaction
- Measurements
  - 2064-109 LPAR with 3 dedicated processors
  - Network Driver
    - Streaming workload: 20 byte request, 20 Mbyte response
    - Connect-Request-Response workload: connect, 64 byte request, 8 Kbyte response
  - GbE results shown
  - Reduced processor timer per transaction 5 to 80%.

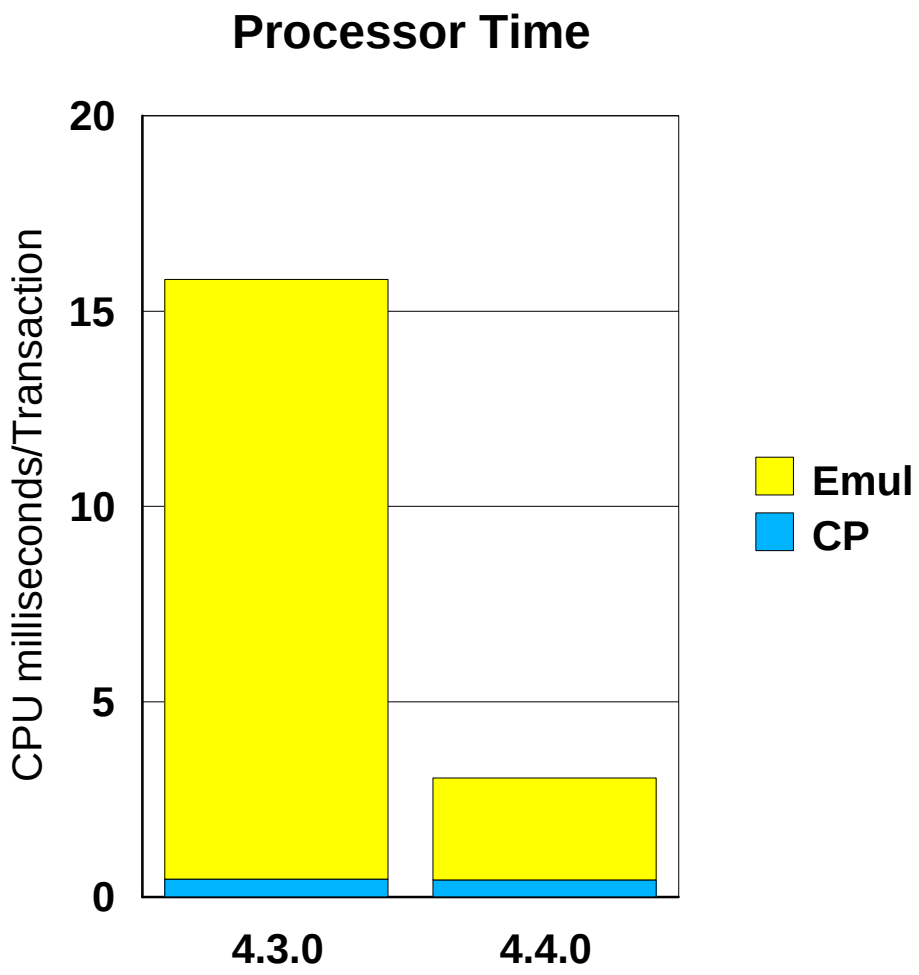
TCP Streaming Measurement Results



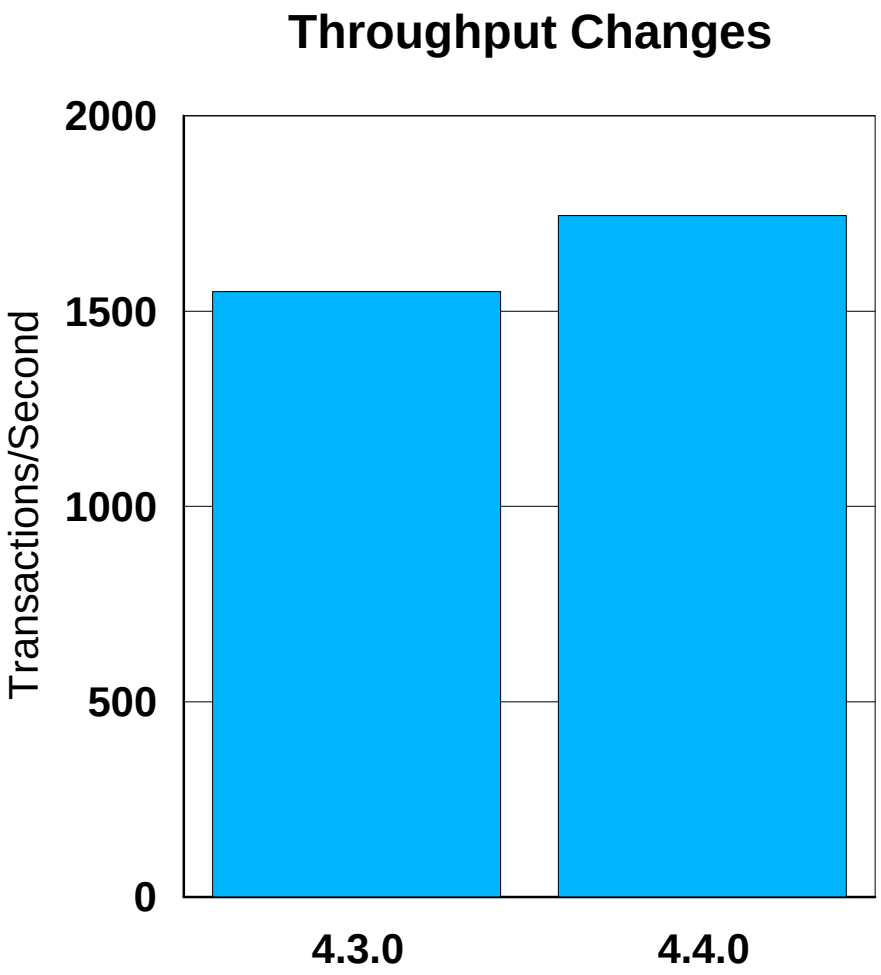


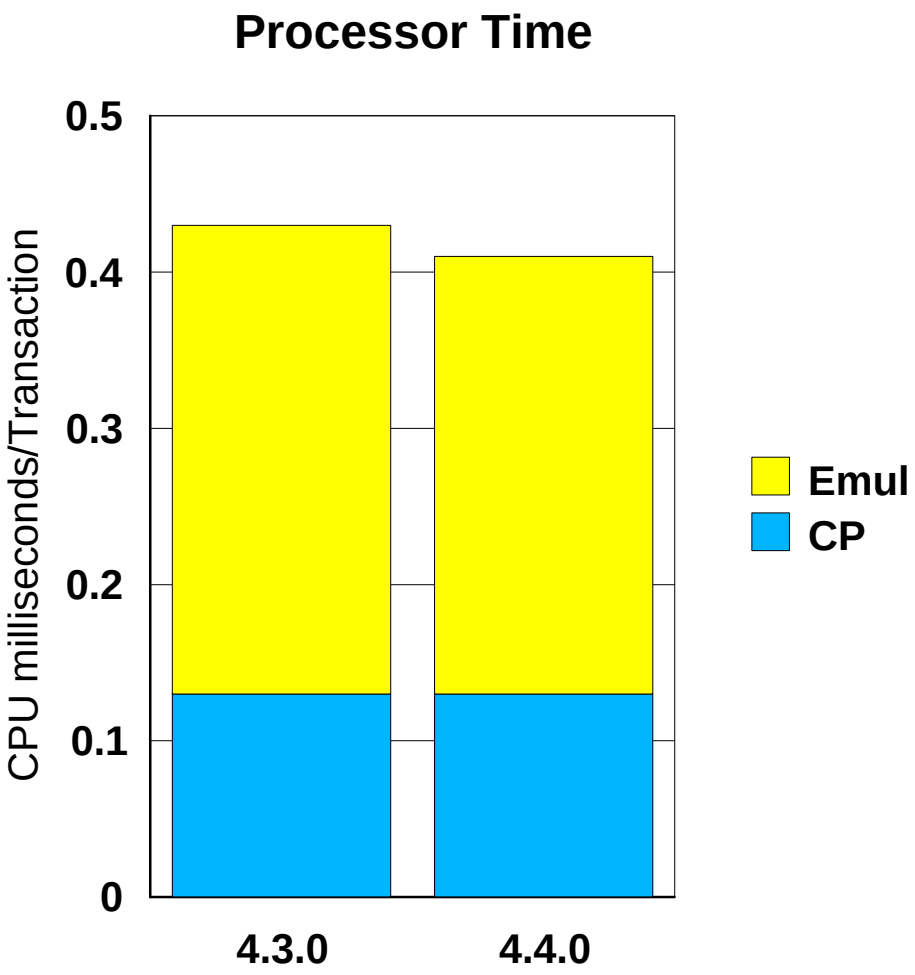
TCP CRR Measurement Results





UDP Measurement Results

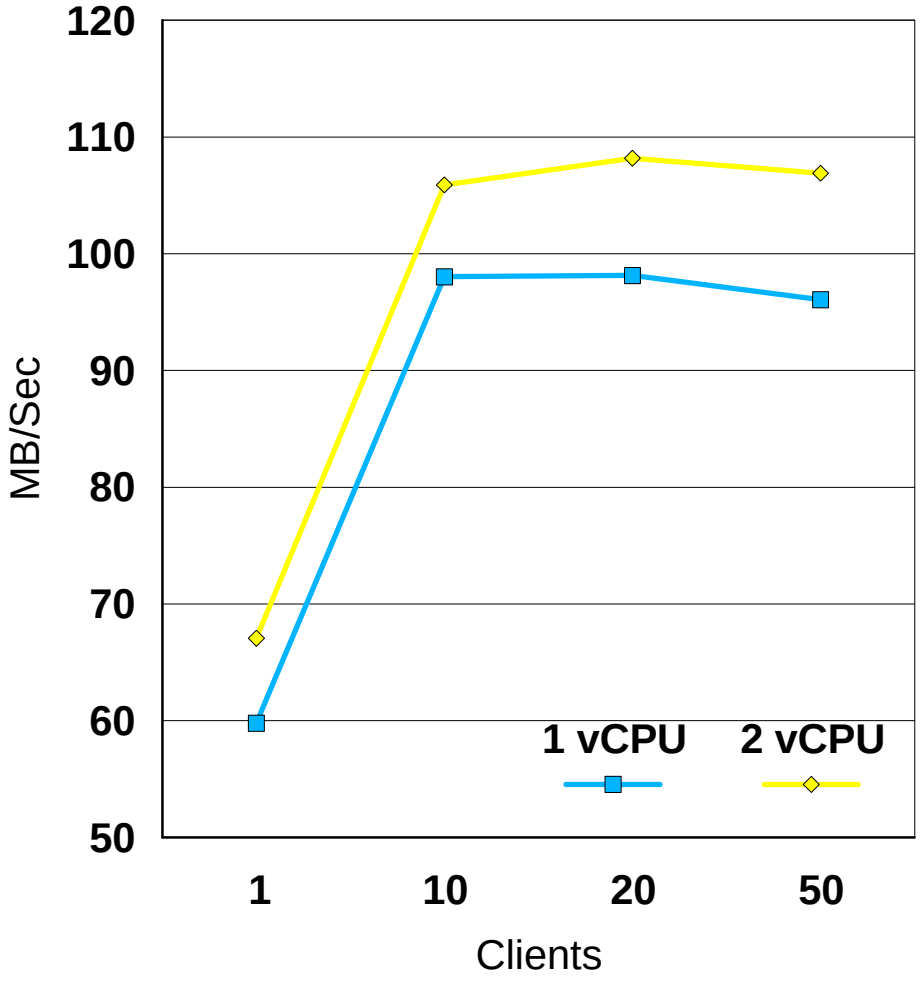


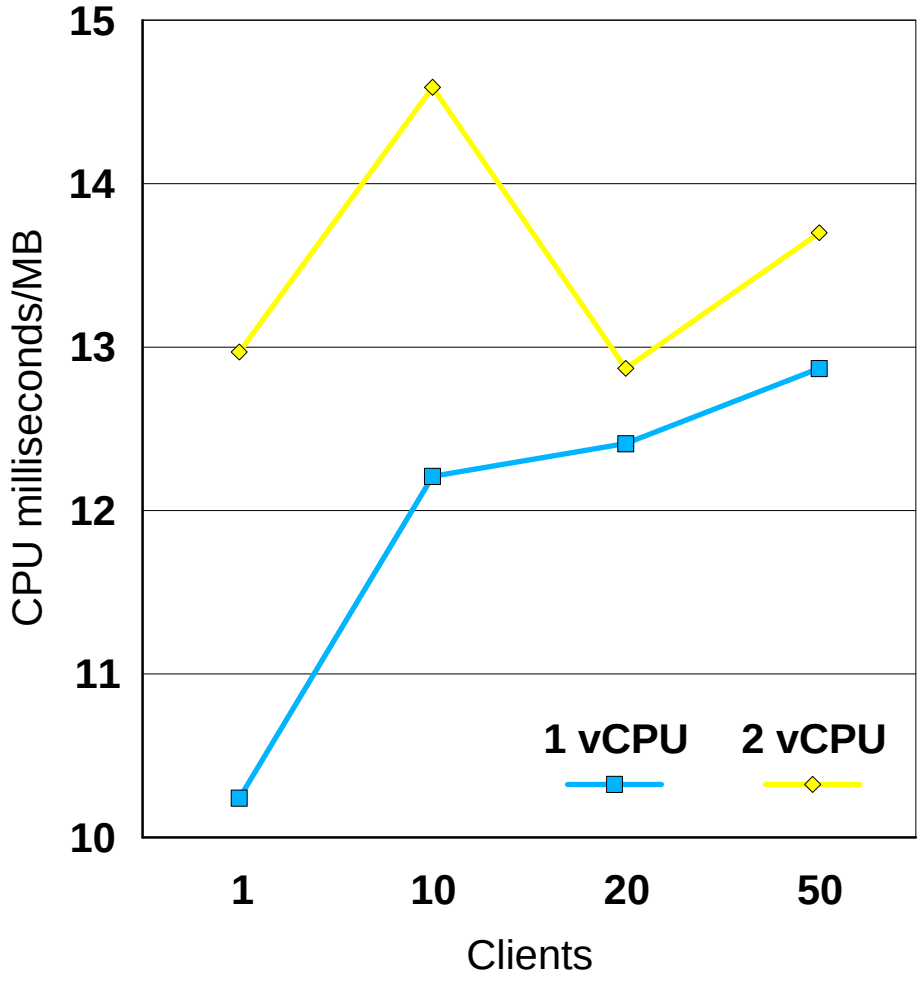


VM TCP/IP Code Optimization

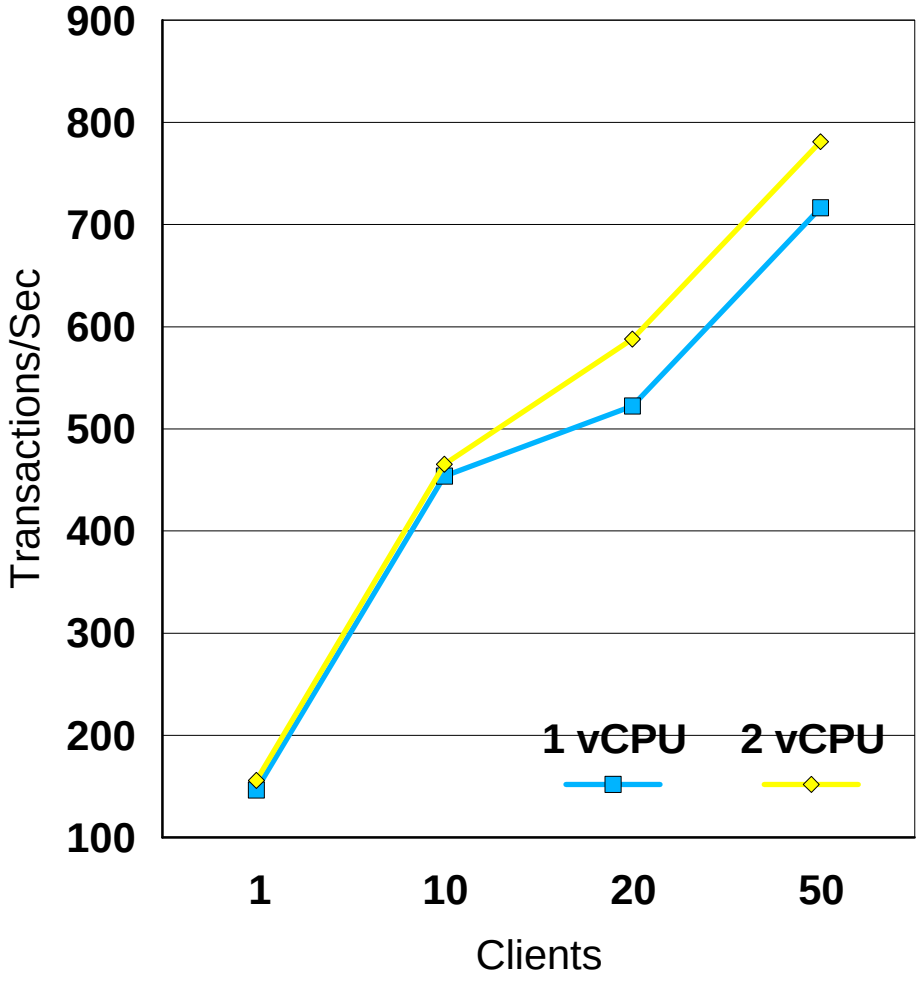
- Code optimization
  - Follow-on improvements from z/VM 4.3.0
  - Hot path analysis in TCP and UDP layers
  - Replace some Pascal code segments with Assembler
  - Modify algorithms
  - lowers processor time per transaction (5 to 80%)
- Virtual MP support for device management
  - Allow additional virtual processors to be assigned to devices (via DEVICE statement)
  - interrupt processing and other CP functions work off additional processors
  - May increase throughput by allowing multiple processors to be used
  - May increase processor costs per transaction

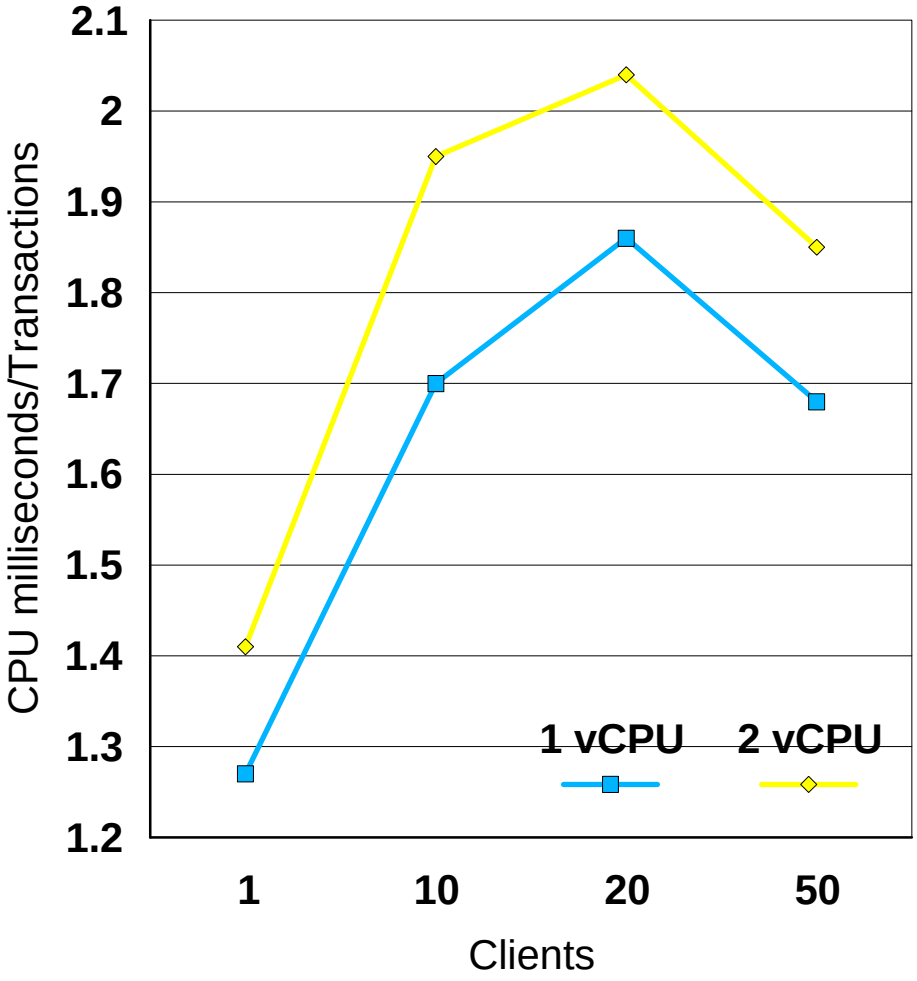
Virtual MP Streaming GbE Measurements (MTU 8992)





Virtual MP CRR GbE Measurement (MTU 8992)



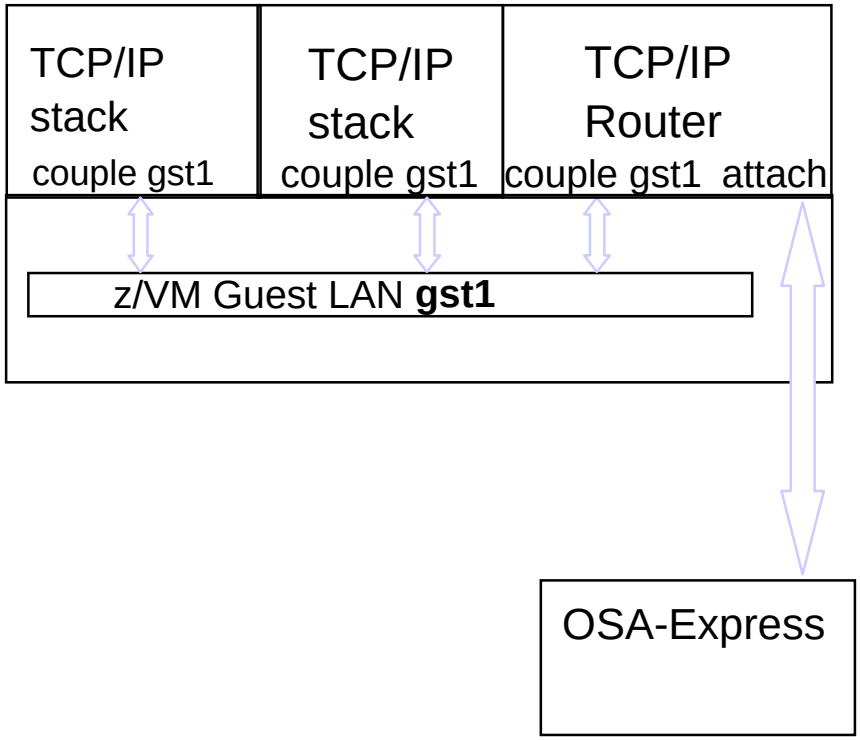


Virtual Switch

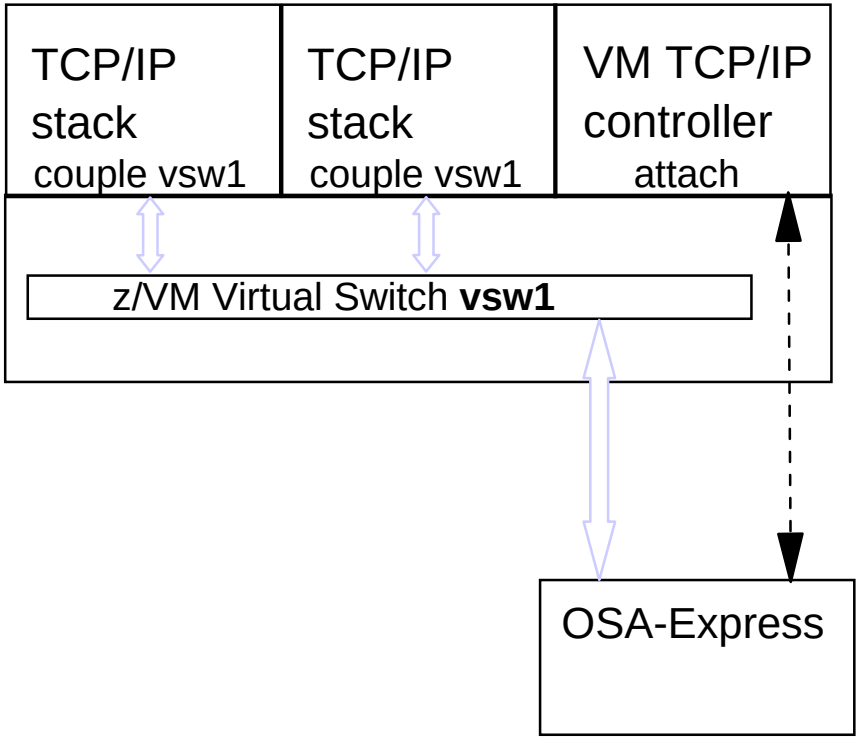
- Layer 3 switch
  - Switches packets between QDIO guest LAN and OSA Express physical network
  - Eliminates need for layer 3 router
  - Supports transparent VLAN specifications for guests connected to Virtual Switch
  - Switching function performed entirely by CP
  - z/VM TCP/IP stack used for setup and control functions
- Provides transparent bridging
  - Learning - automatic configuration of IP addresses
  - Flooding - deliver packets for unknown IP addresses to all stations
  - Aging - forget learned IP addresses after some period of inactivity
- Supports locally-administered MAC addresses

Virtual Switch Topology

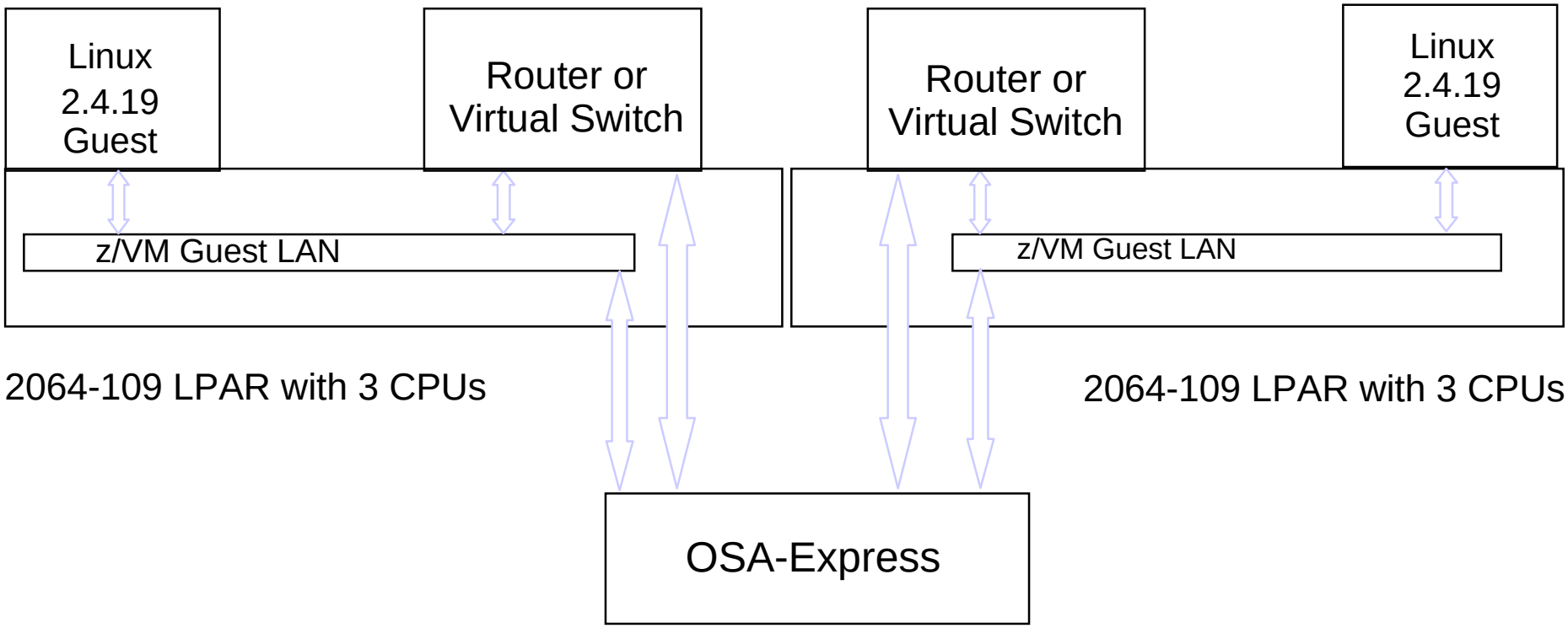
Traditional Guest LAN



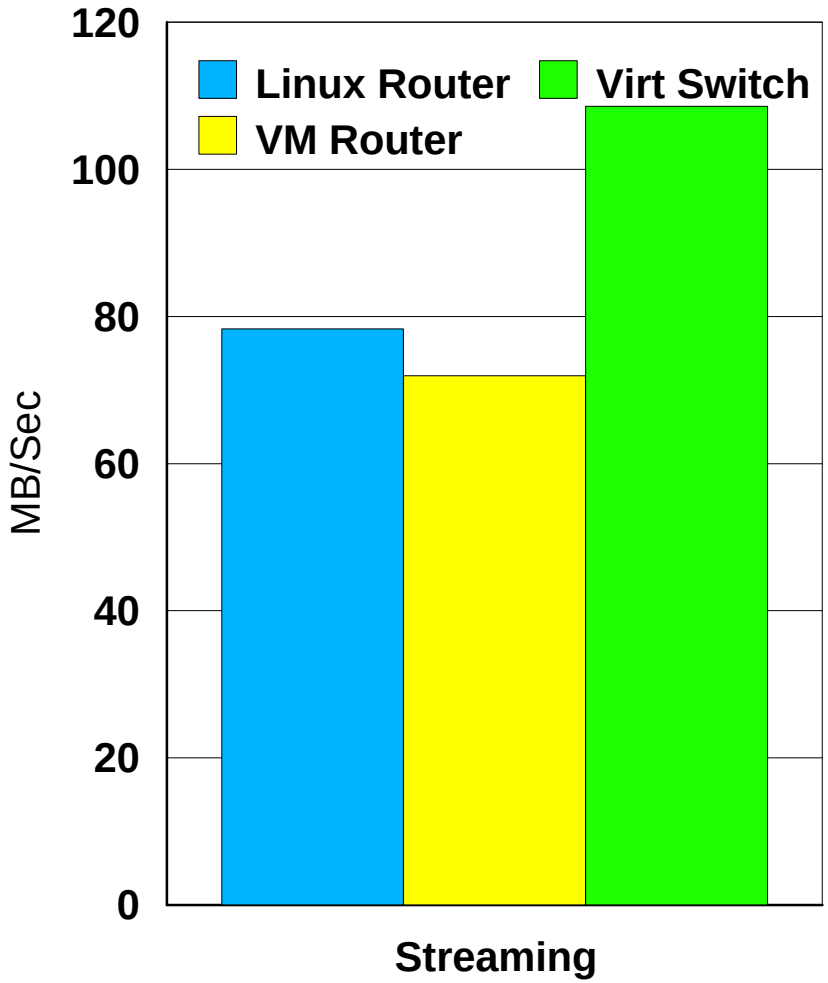
Virtual Switch Guest LAN

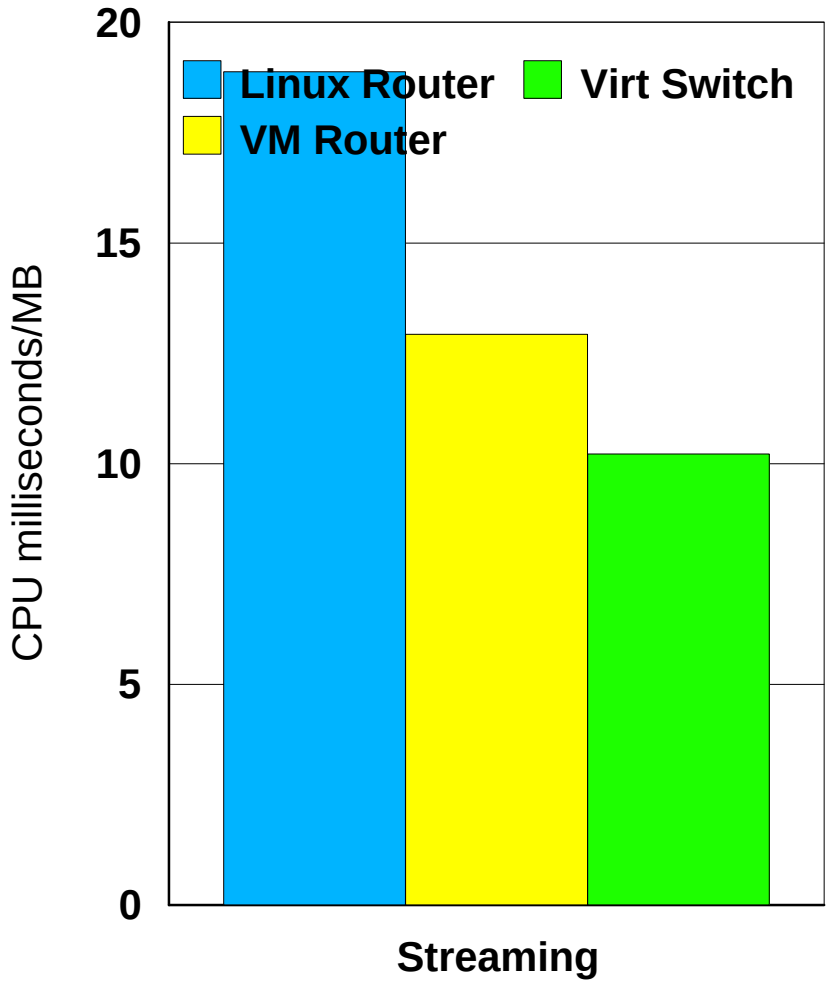


Virtual Switch Test Configuration

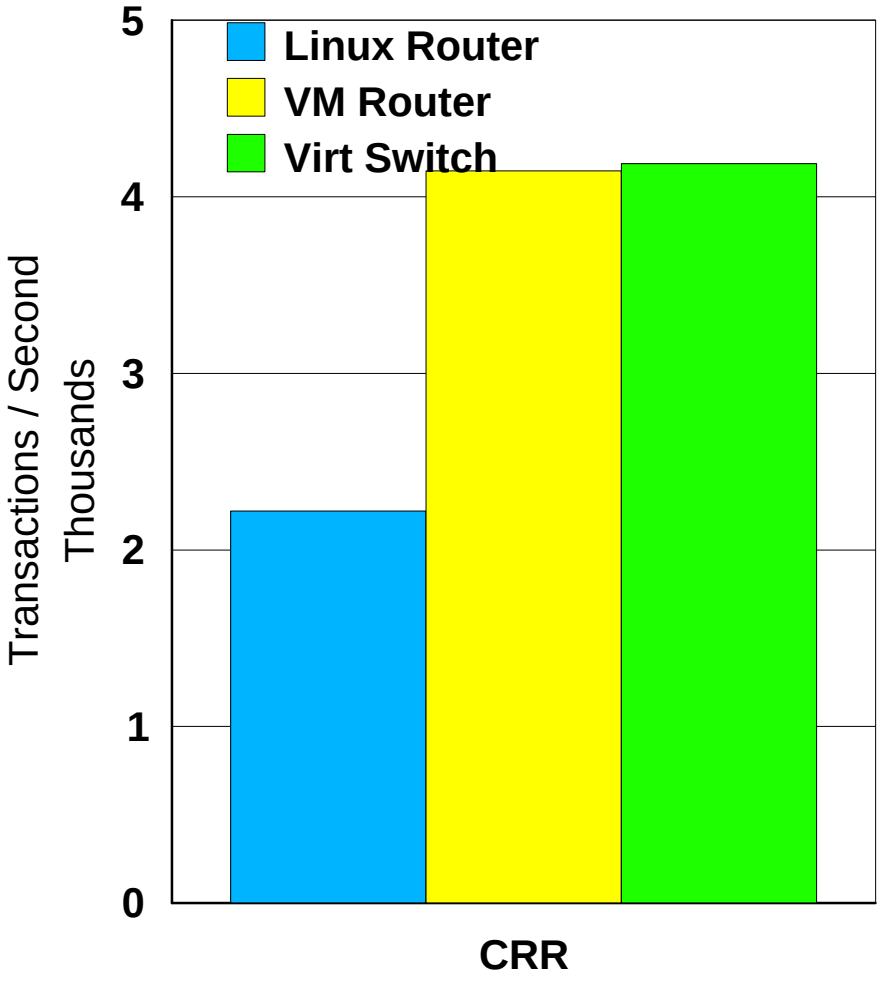


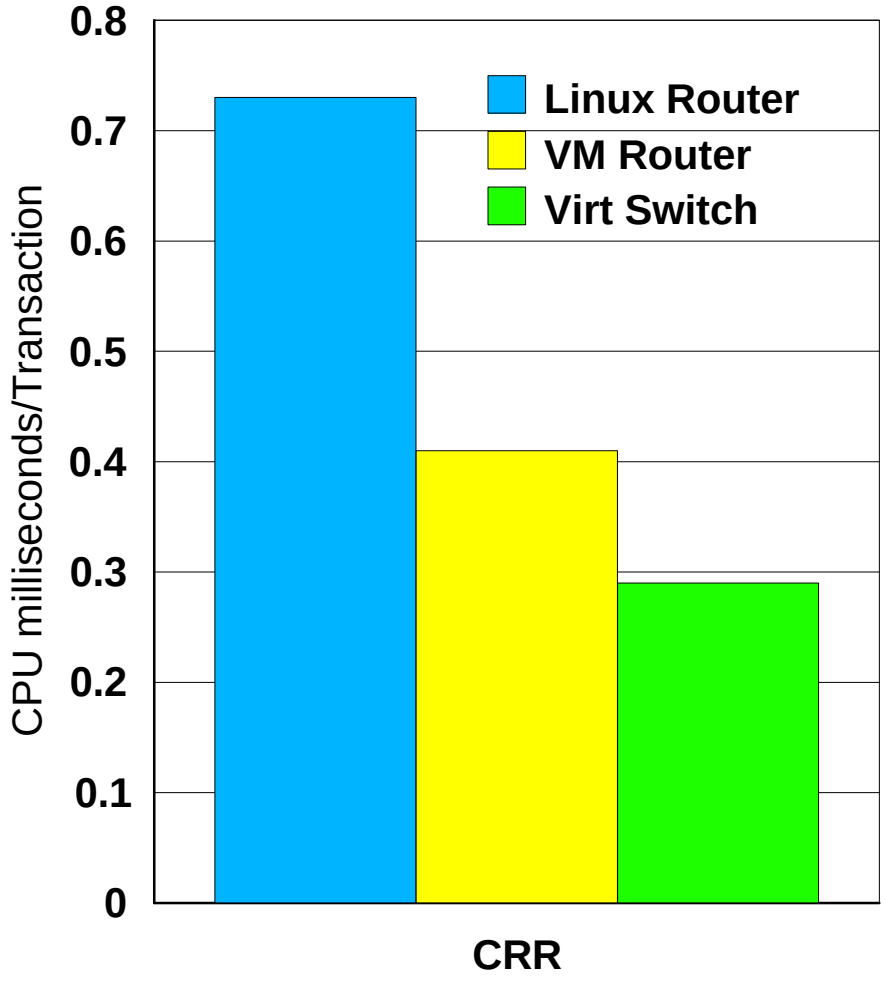
Virtual Switch - Streaming (MTU 8992)



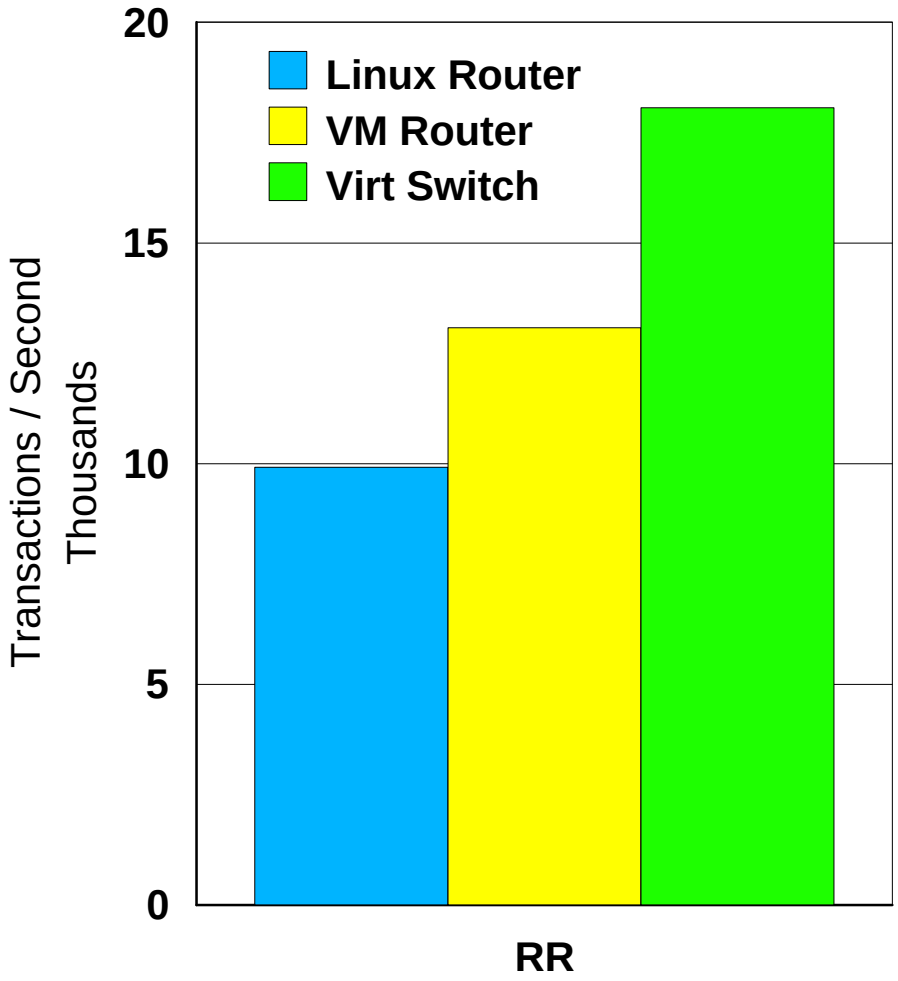


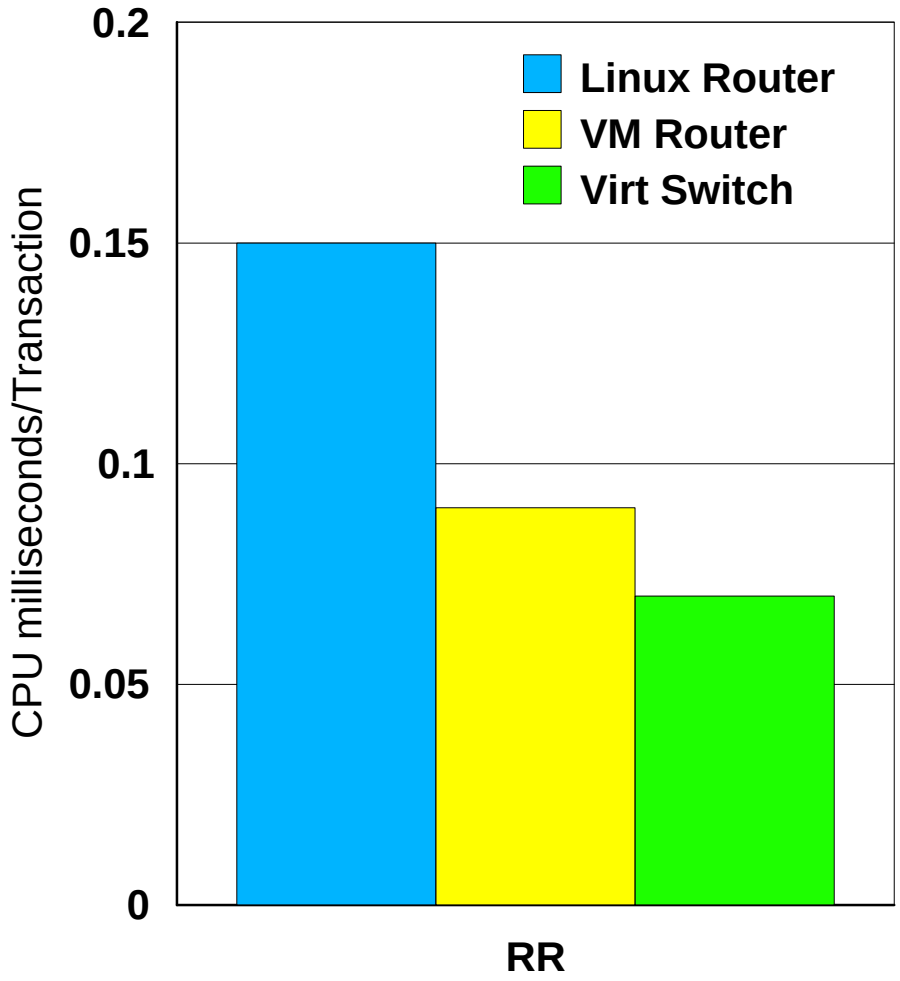
Virtual Switch - CRR (MTU 8992)



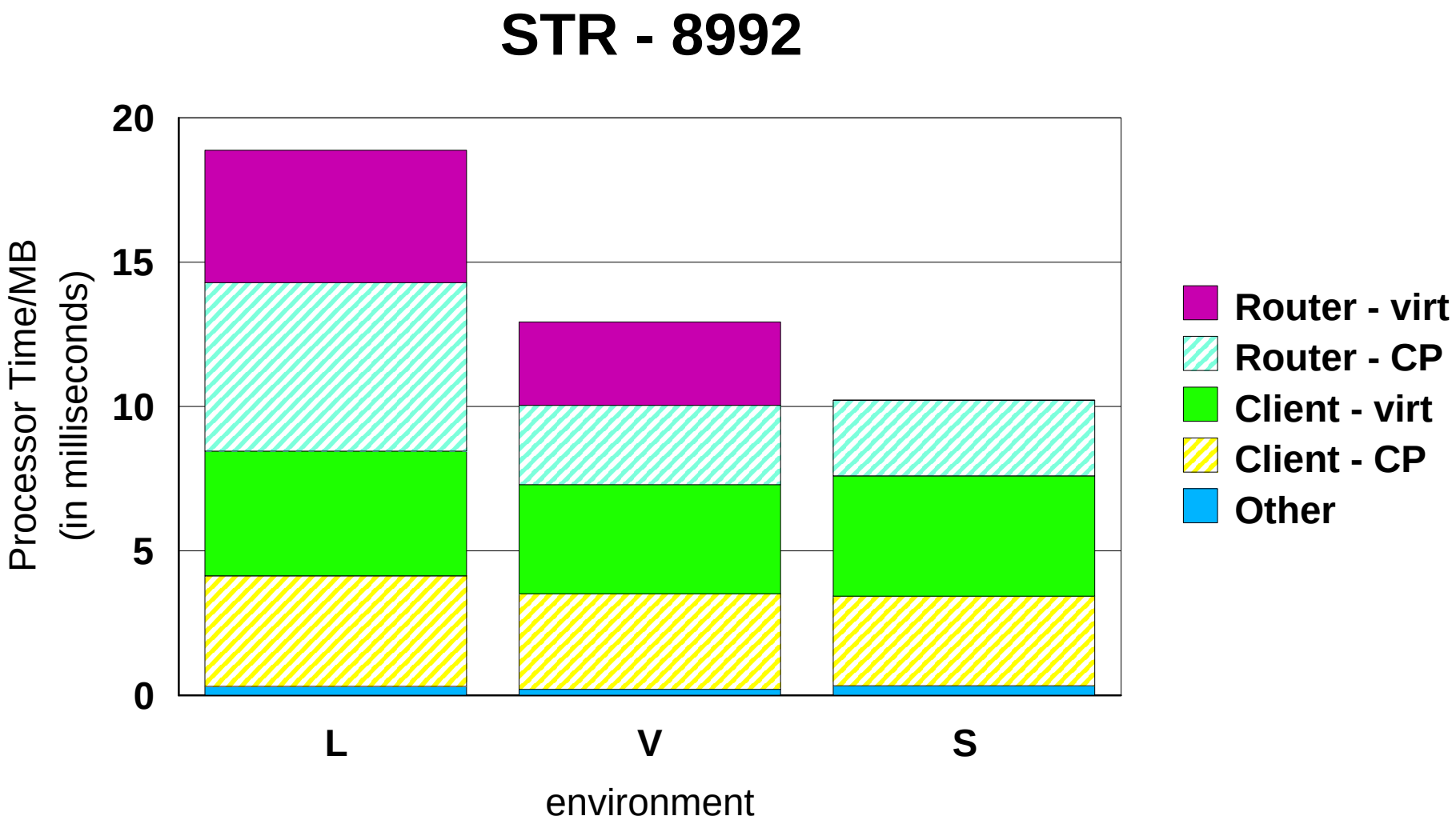


Virtual Switch - RR (MTU 1492)

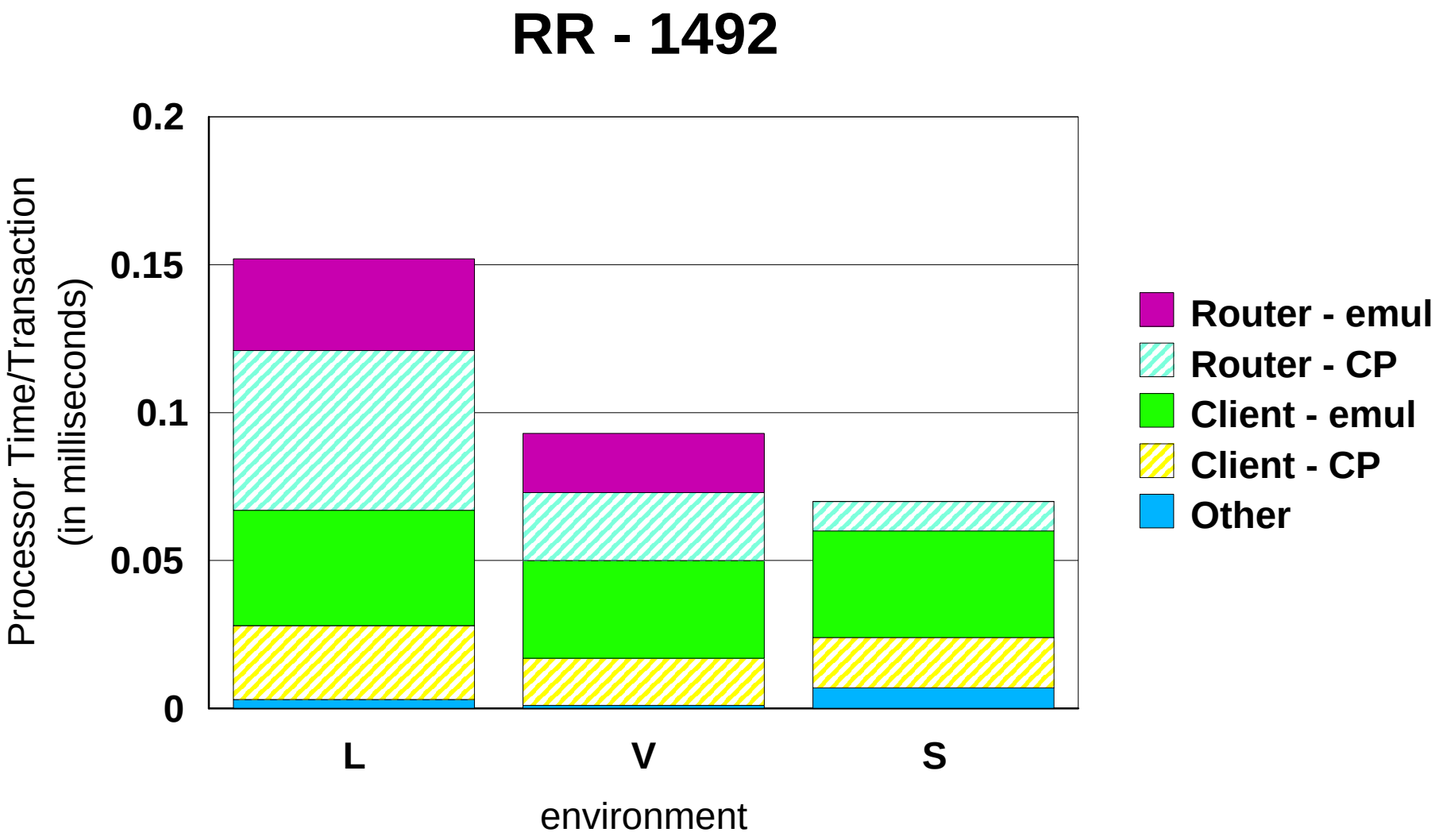




Detailed Processor Time Breakdown - Streaming



Detailed Processor Time Breakdown - RR

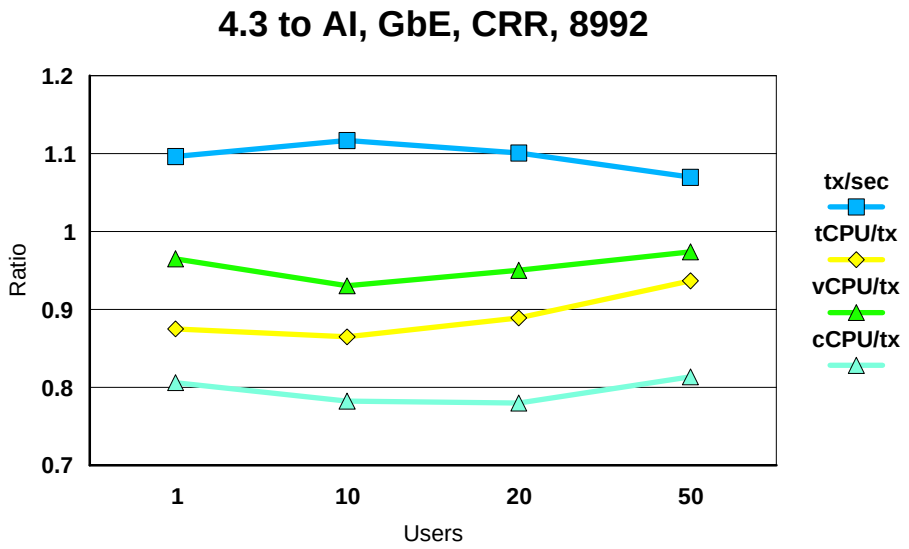


Queued I/O Assist

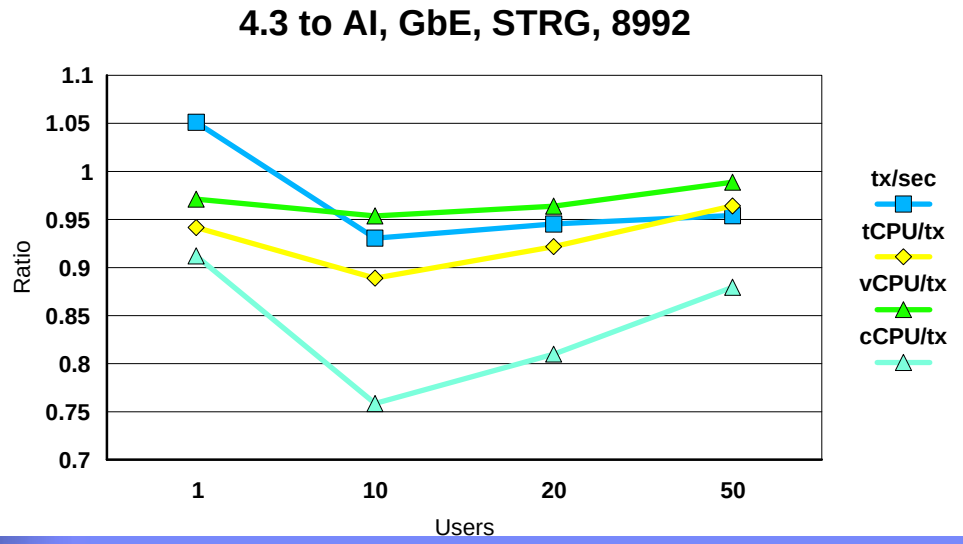
- QDIO devices (FCP, OSA Express, HiperSockets) induce overhead due to high interruption rates
  - z/VM Control Program has to mediate between hardware interruptions and guests
  - As interruption rates go up, this overhead increases
- New hardware facility designed to address this problem
  - Allows interruptions to be presented directly by hardware for active guest
  - Delivers "thin" signal to CP when interruption is for idle guest
  - Extends "thin interrupts" from iQDIO to QDIO and FCP
  - New feature limited to z990.
- Changes in z/VM and Linux to take advantage of assist
  - QUERY/SET QIOASSIST
- See http://www.vm.ibm.com/perf/aip.html for more information.

PCI to AI, Linux, GbE

Usually it's great!



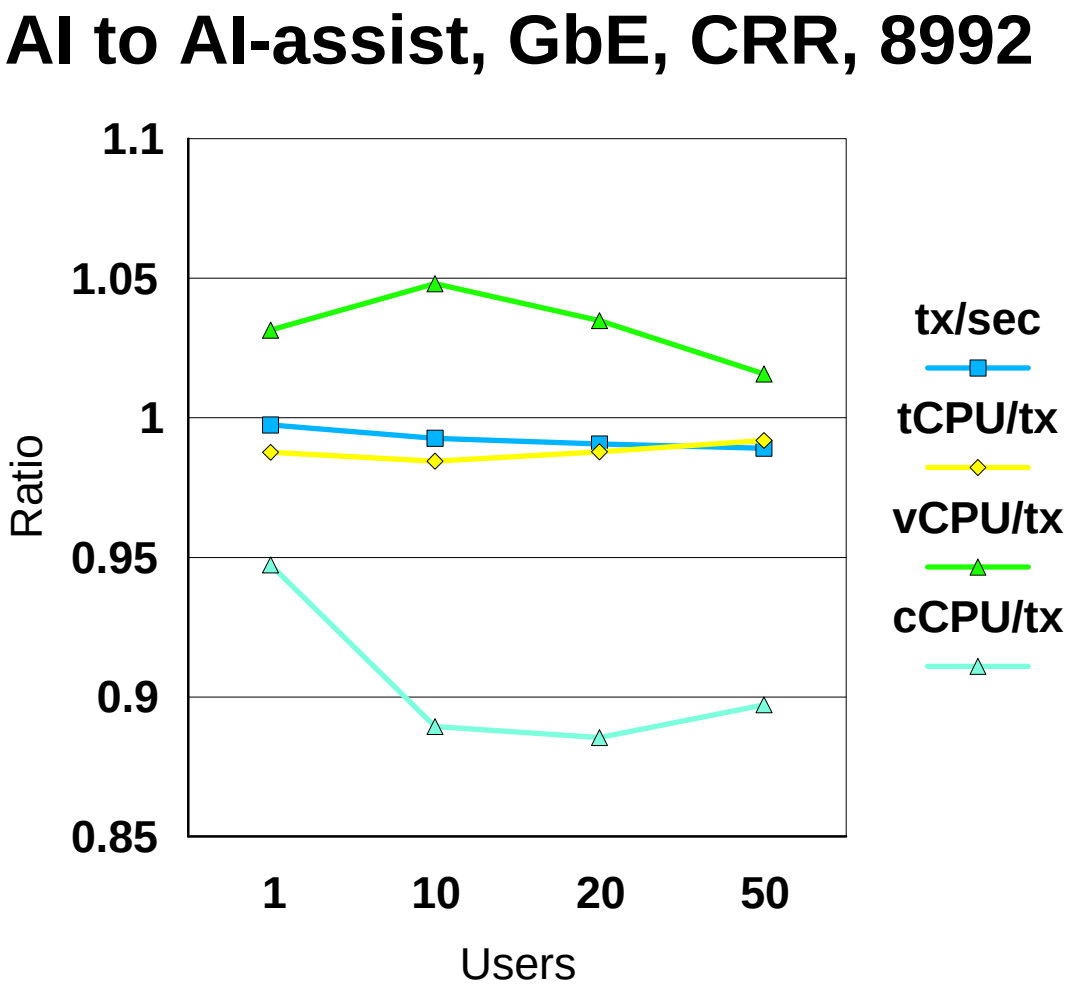
Rarely, it's marginal.



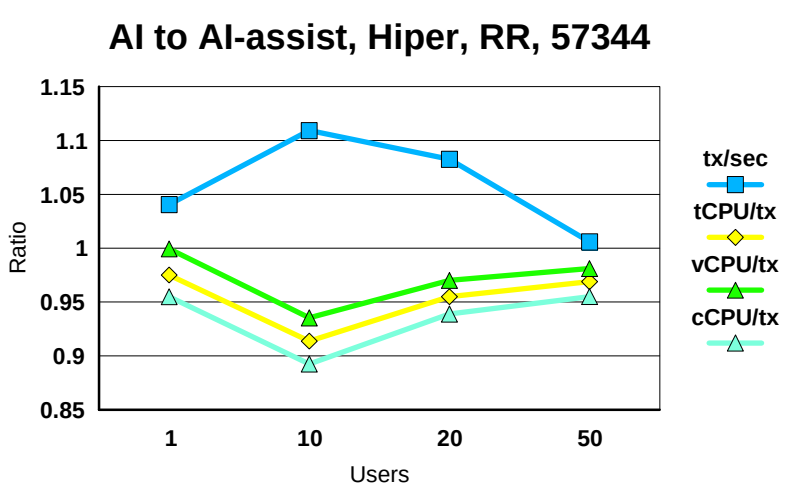
AI to AI-Assist, Linux, GbE

Generally, we see this:

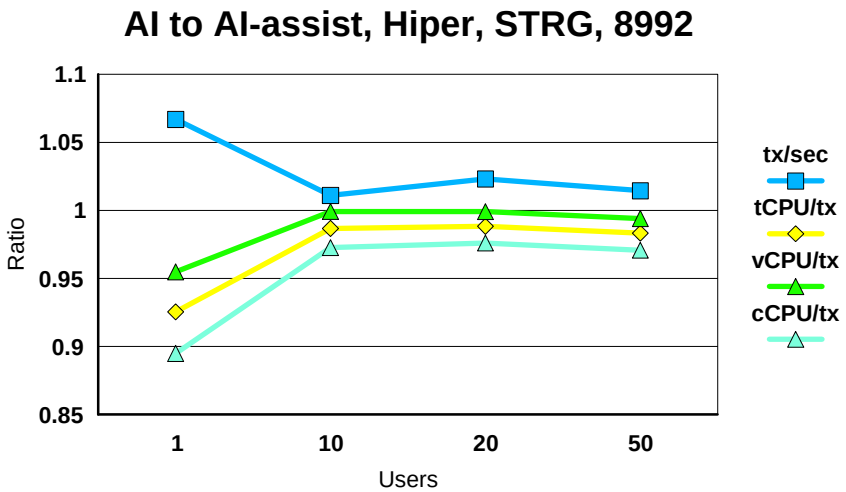
- Tx/sec flat
- Small rise in virtual/tx
- Good drop in CP/tx



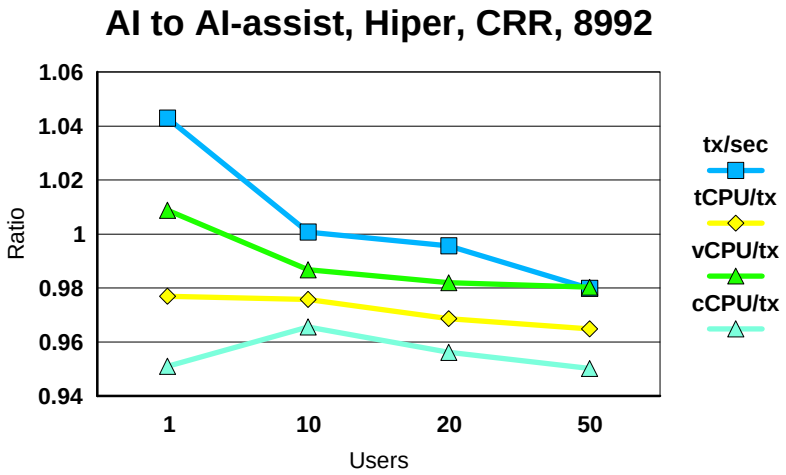
AI to AI-Assist, Linux, HiperSockets



Nice!



Ho-hum.

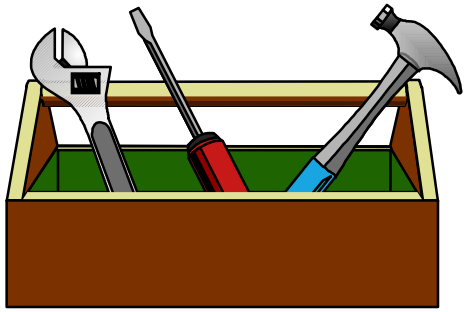


There are only a couple of "oops" cases.

Oops!

Performance Toolkit for VM

- A new z/VM 4.4.0 feature
  - Priced on a one-time-charge, per-processor basis (IPLA Ts&Cs)
  - Based on the FCON/ESA product
  - Will eventually replace RTM and PRF
  - Can be licensed for standard and IFL processors
- Functional highlights
  - Provides an immediate view of system performance
  - Post processes it own history files or CP Monitor data
  - Threshold monitoring
  - User loop detection
  - Can monitor remote systems
  - Results can be graphically viewed by a web browser
  - Processes Linux data provided by the RMF PM data collector
  - Combines and displays both VM and Linux data



Performance Toolkit Linux-Related Functions

- Ability to monitor Linux systems
  - Retrieval based on RMF DDS Interface
  - Originally developed for use with RMF PM
  - Permanent data collection in Linux
  - History data saved in Linux
  - Selective 'ad hoc' retrieval via TCP/IP
  - XML data retrieval requests
  - Linux systems not necessarily under same VM
  - Only data for requested report are retrieved

LPAR Monitor Enhancements

- Prior to z/VM 4.4.0
  - Special diagnose used to acquire data from LPAR
  - Limit of page of logical processor information
  - Anomalies because IFL engines show up as ICF and are not included in physical stats
- z/VM 4.4.0
  - Use of new LPAR interfaces
  - Allows for data collection of a greater number of logical processors
  - Distinguish VM LPARs running on IFLs
- New data reported on by Performance Toolkit for VM

z990 & z890 Monitor Related Notes

- Support for Extended-I/O-Measurement Facility
  - changes to Measurement Block architecture
  - no longer needs to be preallocated
  - larger fields to avoid wrapping scenarios
- Channel Measurement
  - STCPS (STORE CHANNEL PATH STATUS) no longer valid
    - on other processors, STCPS is not valid for CP
  - Used in monitor record Domain 0 Record 9
  - Data for Domain 0 Record 20 (Extended Channel Measurement Data) is valid for channels on the z990

VM Resource Manager

- VMRM introduced in z/VM 4.3.0
  - Manages performance of selected virtual machines based on customer-defined goals for CPU and I/O performance
  - VMRM service virtual machine accepts:
    - Workload definitions (a single workload can include multiple virtual machines)
    - Goal specifications
    - Importance of achieving defined goals
  - VMRM adjusts user CPU shares or I/O performance based on:
    - Velocity goals set for the user's workload class
    - Virtual machine CPU and/or I/O achievement levels
- z/VM 4.4.0 Enhancements
  - Improved support for managing multiple users
  - Improved performance of VMRM service virtual machine
  - Serviceability improvements
  - Monitor data provides information on workloads and their achievements

z/VM Service - VM63282

- Problem: Guests not dropped from dispatch list when idle.
  - Outstanding long term I/O for network devices increments a count field which when non-zero prevents guest from dropping from dispatch list.
  - Guests appeared runnable with high I/O Active wait state percentages.
- Solution:
  - Counter no longer updated for network devices.
  - Shift in user state sampling from I/O Active wait state to Idle state or Test Idle state.
  - Users appropriately dropping from dispatch list allows more effective storage management steal algorithms.

z/VM Service - VM63457

- Problem: V=R and V=F Guests run slower on z/VM 4.4.0.
  - This is most obvious during IPL
  - Guests appeared runnable with high I/O Active wait state percentages.
- Solution:
  - Corrections in HCPALE to handle interruption subclasses correctly.

Summary

- See VM Home page for full report and details:
  - http://www.vm.ibm.com/perf/docs/zvmperf.html
- Equivalent regression performance
  - Except for improvements with VM TCP/IP related workloads
- Significant improvements for Linux environments
  - Scheduler Lock Enhancement
  - Various Network Improvements
  - Queued I/O Assist
- Performance Toolkit for VM