IBM Systems & Technology Group

z/VM Performance Update

Session 9106

Bill Bitner, bitnerb@us.ibm.com

z/VM Performance Evaluation

Trademarks

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries. For a complete list of IBM trademarks, see www.ibm.com/legal/copytrade.shtml: AS/400, DBE, e-business logo, ESCON, eServer, FICON, IBM, IBM Logo, iSeries, MVS, OS/390, pSeries, RS/6000, S/390, System z, VM/ESA, VSE/ESA, Websphere, xSeries, z/OS, zSeries, z/VM

The following are trademarks or registered trademarks of other companies

Lotus, Notes, and Domino are trademarks or registered trademarks of Lotus Development Corporation
Java and all Java-related trademarks and logos are trademarks of Sun Microsystems, Inc., in the United States and other countries
LINUX is a registered trademark of Linus Torvalds
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
* 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.

References in this document to IBM products or services do not imply that IBM intends to make them available in every country.

Any proposed use of claims in this presentation outside of the United States must be reviewed by local IBM country counsel prior to such use.

The information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

Acknowledgments

- **VM Performance Evaluation**
  - Bill Bitner
  - Dean DiTommaso
  - Wes Ernsberger (retired) ☹
  - Bill Guzior
  - Virg Meredith
  - Patty Rando
  - Dave Spencer
  - Joe Tingley
  - Xenia Tkatschow
  - Brian Wade

  - Fred Shaheen (manager)

- **VM Organization**
  - Kevin Adams
  - Lori Cramer
  - Karen Gardner
  - Tim Greer
  - Bill Holder
  - Roger Lunsford
  - Hongjie Yang
  - Mike Wilkins

Agenda

- **z/VM performance update**
  - z/VM 5.4 major items
  - z/VM 5.4 minor items
  - Interesting APARs
- **Hardware notes**
- **z10 performance**

z/VM 5.4 – Enhancements and Notes

- **Enhancements**
  - Specialty engines
  - Dynamic memory upgrade
  - Virtual CPU share redistribution
  - DCSS above 2 GB
  - TCP/IP layer 2
  - Telnet IPv6
  - Linux install from HMC
  - Upper DAT tables
  - Other changes
  - Service: VMRM “safety net”
  - Service: MDC
- **Notes**
  - Long eligible list stays
  - Reorder processing
  - VMDUMP

z/VM 5.4: Specialty Engines

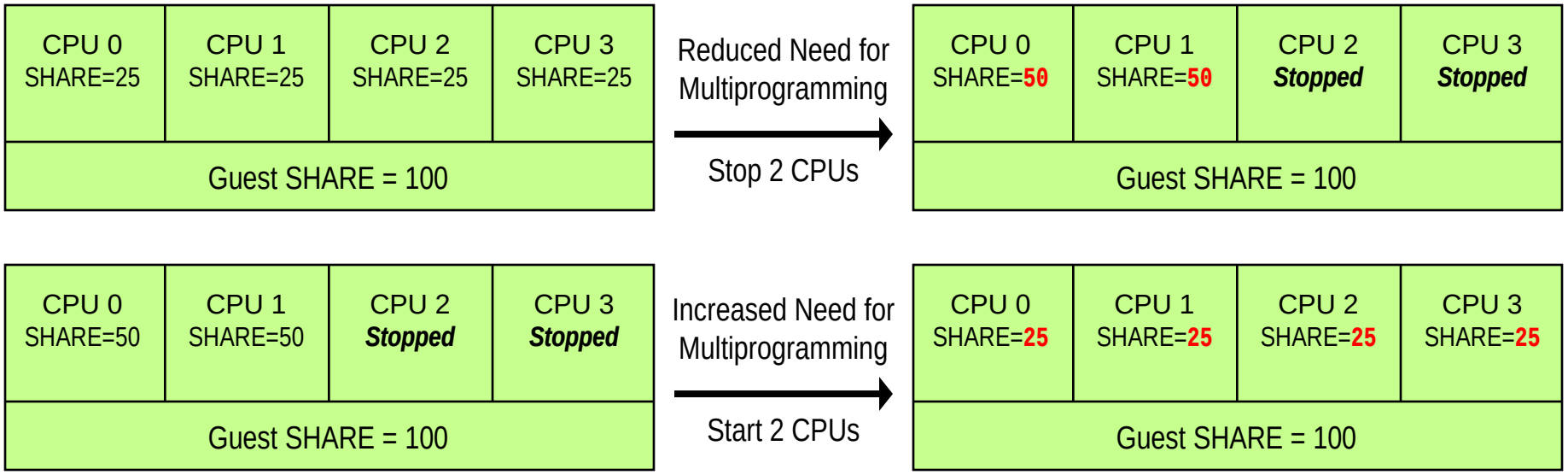
- **Now allow environment with real IFLs and real CPs in same LPAR (as well as other specialty engines)**
  - Called “VM Mode LPAR” – be careful this phrase can be ambiguous at times
- **New options to SET SHARE allow a value for each processor type in a mixed virtual machine.**
- **CPUAFFINITY OFF will result in virtual specialty engines being dispatched on CPs**
- **Merging CP LPAR with IFL LPAR requires thought:**
  - First step, make virtual machines on IFL LPAR have virtual IFLs
  - For duplicated work (RACF, TCP/IP, etc.), need to determine which to use or in some cases which to duplicate
  - Remember that in some environments, the IFLs may be faster than the CPs.
  - You may need to revisit your charge back model.
- **See the z/VM Performance Report**

z/VM 5.4: Dynamic Memory Upgrade

- **Aka Dynamic Storage Reconfiguration**
- **Dynamically add real central storage**
  - No removal; no expanded storage
- **SET STORAGE command response may be delayed**
  - Request for small increase in comparison to memory available to partition
- **Memory is not all initialized instantaneously**
- **Also virtualizes for guests that support it**
- **Some monitor event record anomalies exist**
  - Not surfaced in Performance Toolkit
  - Query commands are correct
  - Fixed in service stream with VM64483
- **See the z/VM Performance Report**

z/VM 5.4: Virtual CPU SHARE Redistribution

- **Allows z/VM guests to expand or contract the number of virtual processors it uses without affecting the overall CPU capacity it is allowed to consume**
  - Guests can dynamically optimize their multiprogramming capacity based on workload demand
  - Starting and stopping virtual CPUs does not affect the total amount of CPU capacity the guest is authorized to use
  - Linux CPU hotplug daemon starts and stops virtual CPUs based on Linux load average value
- **Helps enhance the overall efficiency of a Linux-on-z/VM environment**
- **Previously, stopped virtual processors were given a portion of the guest share.**



Note: Overall CPU capacity for a guest system can be dynamically adjusted using the SHARE setting

z/VM 5.4: DCSSs Above 2 GB

- **DCSSs can be loaded above 2 GB**
  - Each DCSS is still limited to 2047 MB
  - Continue to allow multiples, thus aggregate > 2 GB
- **Requires additional Linux support:**
  - To use a segment above 2 GB
  - To let Linux use ‘stacked’ contiguous DCSSs as one large block device
- **Benefits of sharing much larger amounts of memory**
  - Varies significantly based on configuration and workload.
  - Example: throughput improvements of ~35% for static web serving observed with prototype Linux code.
- **See the z/VM Performance Report**

z/VM 5.4: DCSS Loading Time & Tradeoffs

- **Support for DCSSs above 2 GB means great amount of memory can be used for DCSSs.**
- **Performs well once DCSS is built and data saved.**
- **Placing at higher than necessary addresses results in more memory required for data structures in guest and in z/VM.**
- **Time required to save or fill a DCSS is non-trivial.**
- **Two primary methods of defining segments for use by Linux with unique pros and cons**
  - SR: Shared read-only access.
  - SN: Shared read/write access, no data saved.

z/VM 5.4: DCSS Tradeoffs, SN vs. SR

DCSS Attribute	SN: Shared R/W; not saved[2]	SR: Shared R/O
Non-Volatile (Saved to spool)	NO	YES
Initial elapsed time to 'make ready' the file system[1]	Tends to be faster (not necessarily writing to DASD)	Tends to be slower (must write to DASD)
Spool processing for DCSS can delay other spool activity	NO	YES

[1] Making ready the file system involves various steps of defining, copying data into the DCSS, and saving as necessary.

[2] The DCSS itself is read-write. After data is loaded into the DCSSs, one mounts the file system read-only.

z/VM 5.4: TCP/IP Ethernet Mode

- **TCP/IP stack can now run a real OSA in Ethernet mode (aka "layer 2" mode)**
- **It can also [of course] run a QDIO virtual NIC in Ethernet mode**
  - … thereby letting it couple to an Ethernet-mode guest LAN or Ethernet-mode VSWITCH
- **Key findings: for VSWITCH case, according to workload,**
  - 0% to 13% improvement in throughput
  - 0% to 7% decrease in CPU time per unit of work
- **See the z/VM Performance Report**

z/VM 5.4: Telnet IPv6

- **Our Telnet server can now support IPv6 clients**
  - New Pascal APIs for IPv6
  - Telnet server calls *only* these new IPv6 APIs
    - Uses IPv6-mapped IPv4 addresses if necessary
- **Key findings:**
  - IPv4 regression: impact is “around zero”
  - From IPv4 to IPv6:
    - 12% to 23% increase in throughput
    - 3% to 13% decrease in CPU per unit of work
- **See the z/VM Performance Report**

z/VM 5.4: HMC Linux Install

- **The HMC DVD can now be used to install Linux**
  - z/VM FTP server will communicate with it
  - Alternative to LAN-based server
- **Elapsed time for install is much greater than LAN-based server install**
  - 11 to 12 times slower
  - Up to 3 hours to do base install
- **Do apply the service from the GA RSU**
  - APAR PK69228

z/VM 5.4: Upper DAT Structures Above 2 GB

- **z/VM 5.3 let page tables (PGMBKs) reside above 2 GB**
- **z/VM 5.4 further lets segment tables and region tables reside above 2 GB**
  - These structures tend to require multiple contiguous frames.
  - Helps reduce fragmentation and long searches
    - Resulting in better performance for some configurations
  - Prereq for increasing supported real memory (which we did not do this release)

z/VM 5.4: CMS Based SSL Server

- **Previously, z/VM SSL Server was based on Linux**
  - Required customers to install Linux in a virtual machine
  - Now a CMS based server is used instead of Linux
- **APARs shipped in December for this support:**
  - PK65850 & PK73085
- **Performance**
  - Now exploits the CP Assist for Cryptographic Functions (CPACF) feature where available.
  - Very sensitive to the “maximum session” parameter. Do not set this higher than necessary.
  - Future performance enhancements are planned.

z/VM 5.4: Other Changes

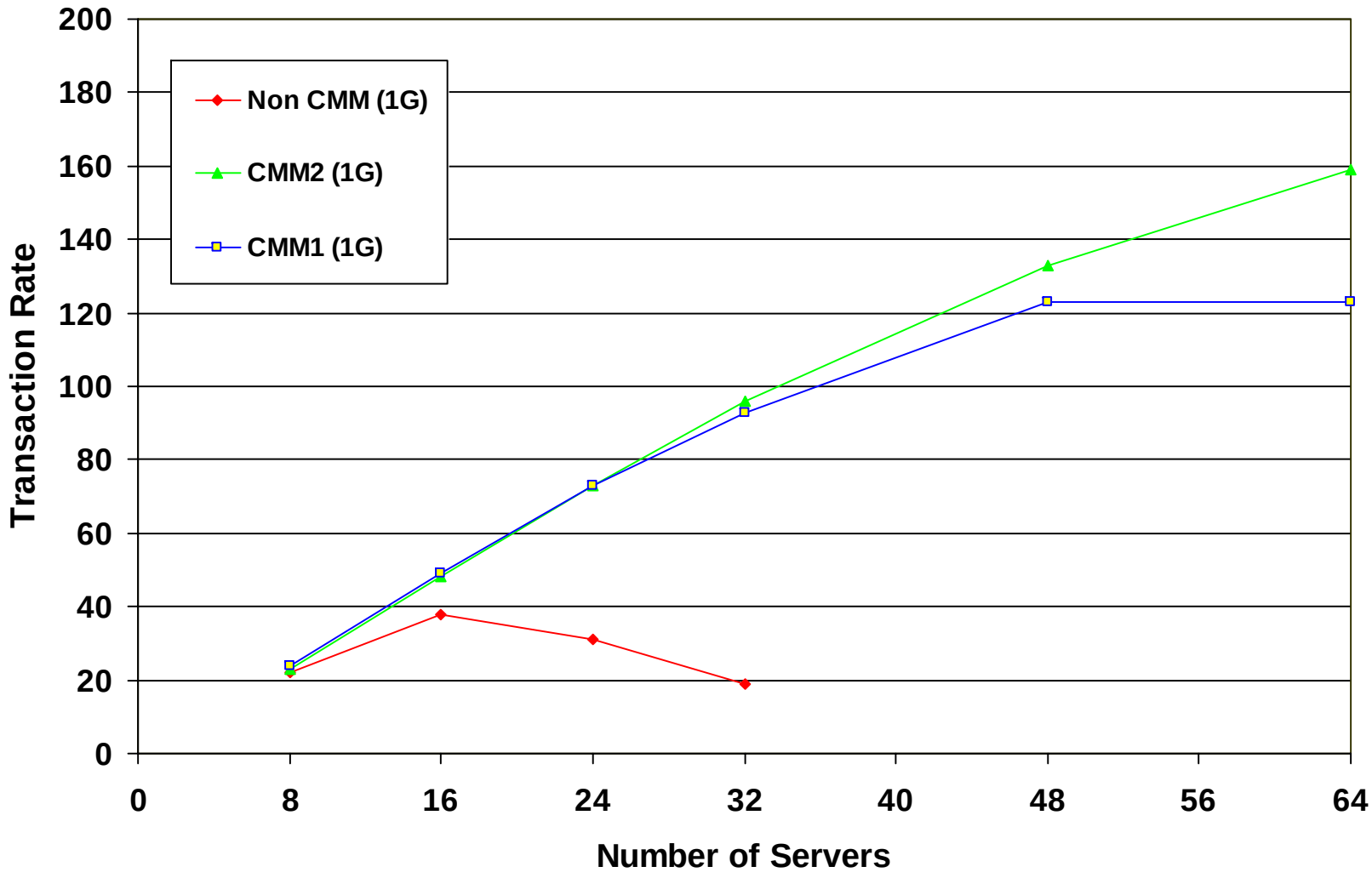
- **PLDV push-through stack: improved efficiency in dequeueing from PLDVs**
- **Virtual CTC: buffering changes help workloads that do > 32 KB writes to VCTCs**
- **VSWITCH changes: improved dispatching, suppressed unnecessary load rebalance calculations, increased packet queuing limits**
- **MDC 8 GB: once MDC reached 8 GB, it stopped doing inserts. Not anymore. ☺**
- **Contiguous available lists: moved low-water and high-water marks closer together – seemed like the right thing for the workloads we tried**

VM64439: CMM–VMRM Change

- **CMM1 aka CMM & VMRM aka ballooning aka Cooperative Memory Management**
- **CMM2 aka MEMASSIST aka CMMA aka Collaborative Memory Management**
- **New CMM-VMRM 64 MB "safety net" –**
  - In base of z/VM 5.4.0
  - As APAR VM64439 for z/VM 5.2.0 & z/VM 5.3.0
    - z/VM 5.2.0 PTF UM32427
    - z/VM 5.3.0 PTF UM32428
- **For additional details see:**
  http://www.vm.ibm.com/perf/reports/zvm/html/530cmm.html

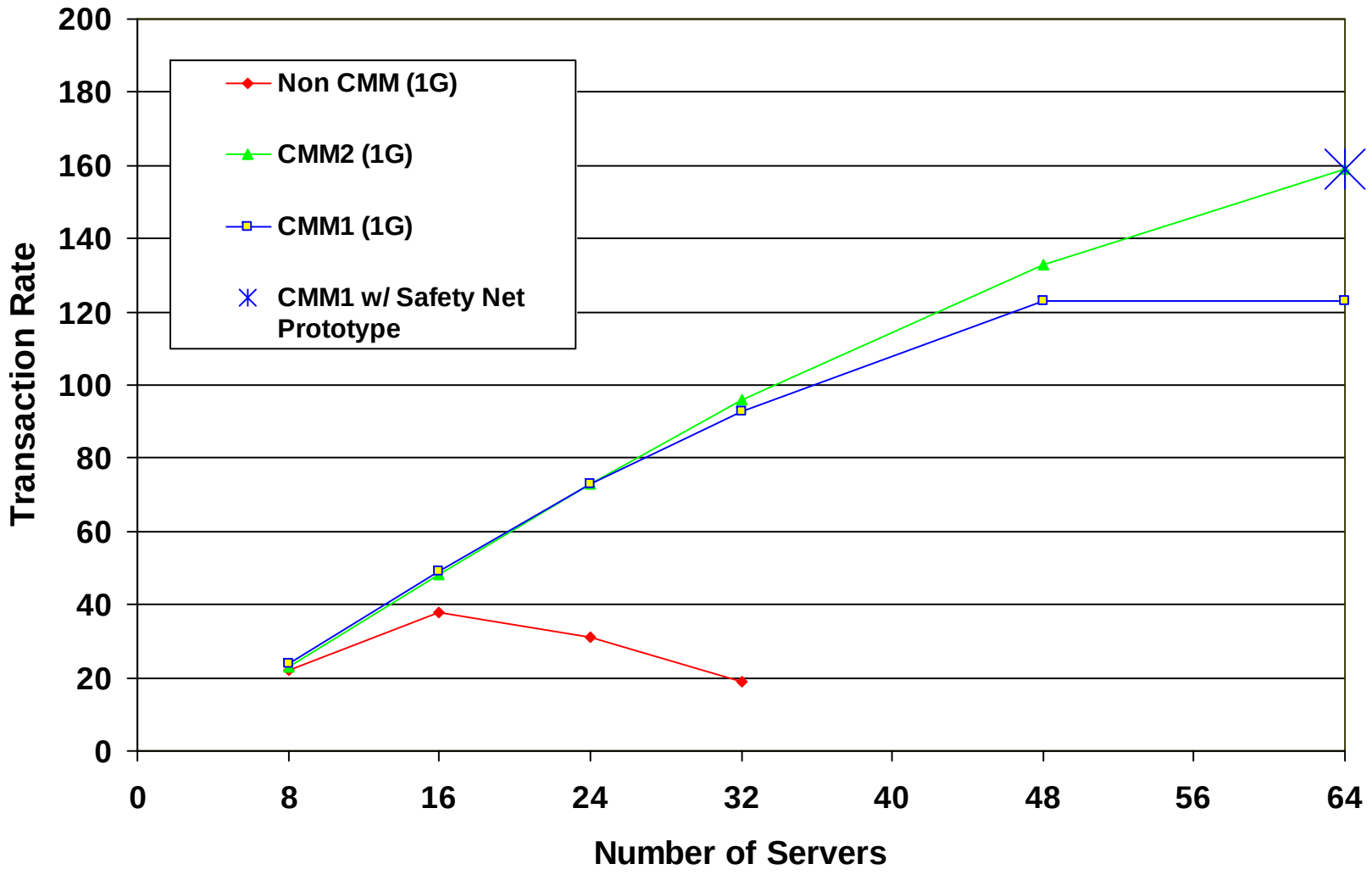
Transaction Rate vs. Number of Servers

for various Storage Management Products using Apache servers with a virtual storage size as shown in parentheses in the legend; z9 6 GB / 2 GB



Transaction Rate vs. Number of Servers

for various Storage Management Products using Apache servers with a virtual storage size as shown in parentheses in the legend



VM64082 and VM64510: MDC Changes

- **VM64082**
  - (z/VM 5.2.0 & 5.3.0). Not on any RSU.
  - Rolled into base of z/VM 5.4.0
  - This change improves MDC Arbiter choices in some environments (doesn't overbias for MDC).
  - Also, makes other environments worse by biasing too much against MDC.
  - Use SET MDC command to bound the range of memory to avoid problems.
- **VM64510: VM64082 had problems on z/VM 5.3 and 5.4**
  - Caused too-frequent steal of pages from MDC
  - VM64510 fixes it

z/VM 5.4: Long Eligible List Stays

- ***Eligible list*** **is where the scheduler keeps virtual machines to prevent them from running and causing thrashing.**
- **SET SRM and SET QUICKDSP tuning options can be used to keep the proper guests from being trapped in eligible list.**
- **There have been known cases of virtual machines trapped in the eligible list for longer than acceptable times in past releases.**
- **There is evidence that in z/VM 5.4 there is a greater tendency for virtual machines to get trapped when system is not properly tuned.**

z/VM: Reorder Processing & Large Resident Counts

- ***Page reorder* is the process of managing user frame owned lists as input to demand scan processing.**
  - It includes resetting the HW reference bit.
  - Serializes the virtual machine (all virtual processors).
  - In all releases of z/VM
- **It is done periodically on a virtual machine basis.**
- **The cost of reorder is proportional to the number of resident frames for the virtual machine.**
  - Roughly 130 ms/GB resident
  - Delays of ~1 second for guest having 8 GB resident
- **Development is investigating improvements**

z/VM: VMDUMP Processing

- **VMDUMP is a very helpful command for problem determination.**
- **Some weaknesses:**
  - Does not scale well, can take up to 40 minutes per GB.
  - It is not interruptible
- **Linux provides a disk dump utility which is much faster relative to VMDUMP.**
  - It is disruptive
  - Does not include segments outside the normal virtual machine.

Hardware Concerns

- **ECPMF and FCP chpids**
- **LPAR monitor records**
- **z10 performance compared to z9**

HW: Monitor – Some FCP channels have incorrect data

- **Domain 0 Record 20 reports Extended Channel Path Measurement Facility information**
- **This data appears incorrect for some FCP channels.**
- **Reference z/VM internal problem number FC01043**
- **Resolved in hardware:**
  - z9 with Driver 67L + MCLs G40938.004 and G40939.004
  - z10 with Driver 76D + MCLs available 2/23/2009

HW: LPAR Monitor Records, z10 + z/VM 5.2

- **z/VM 5.2 customers running on z10 will not get LPAR monitor records for other partitions on the CEC.**
- **z/VM 5.3 and z/VM 5.4 are fine.**
- **Other processors are fine.**
- **A hardware problem ODT S5619 is open to correct this.**
- **Corrected with:**
  - EC Level: F85901
  - MCL No: 006
  - Driver: D73G

HW: z10 Performance

- **Processor cycle time greatly improved over z9**
  - ~2.6 times faster (4.4 GHz)
  - Comparable to other platforms
- **Laws of physics must be obeyed**
- **Tradeoffs made in order to achieve above**
  - Memory differences
  - Key ops
- **ITR ratios (examples see LSPR for most current numbers)**
  - z/OS: z10 EC 701 up to **1.62 times** that of the z9 EC 701
  - LSPR z/VM measurements: 1.30 to 1.60
  - z/VM Endicott lab measurements: 1.23 to 2.05

HW: z10 Performance Depends On….

- **Number of processors**
  - Fewer processors, better ITRR
- **Storage references**
  - Smaller memory footprints, better ITRR
- **Data movement**
  - Less data movement, better ITRR
- **Virtual I/O to real devices**
  - Less virtual I/O, better ITRR
- **Storage overcommitment**
  - Less over commitment, better ITRR
- **Amount of memory involved in long searches**
  - Shorter & less frequent searches, better ITRR
- **Exploitation of new features**
  - More exploitation of features, better ITRR

HW: z10 Performance: Setting Proper Expectations

- **z10 is a great machine, with a number of excellent attributes.**
- **Care must be taken when sizing migrations from z9 to z10.**
- **Additional Information:**
  - LSPR Q & A (complete)
    - Discuss range and factors affecting
    - Pointer to z/VM web page
  - z/VM web page
    - http://www.vm.ibm.com/perf/z10.html
  - "To MIPS or Not to MIPS, That is the Question!" by Gary King
    - http://shareew.prod.web.sba.com/proceedingmod/abstract.cfm?abstract_id=17583

Summary

- **Specialty engines and DMU help z/VM keep up with hardware's capabilities**
- **Several other enhancements in z/VM 5.4**
- **z10 is a good machine, but attention to detail is required**
- **Check out the z/VM performance report: http://www.vm.ibm.com/perf/reports/zvm/html/**
  - z/VM 5.4 topics published on September 19, 2008