IBM Systems & Technology Group

VM Performance Update

Bill Bitner
VM Performance Evaluation
bitnerb@us.ibm.com

4/14/2008

Trademarks

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, ESCO, eServer, FICON, IBM, IBM Logo, iSeries, MVS, OS/390, pSeries, RS/6000, S/30, 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.

Introduction

- **Some Post z/VM 5.2.0 News**
  - PAV
  - OMEGAMON XE
- **z/VM 5.3.0 Performance**
  - Line Items that have an impact
- **APARs of Interest**
- **z10 Performance**
- **See Performance Report on web for details**
  - http://www.vm.ibm.com/perf/reports/

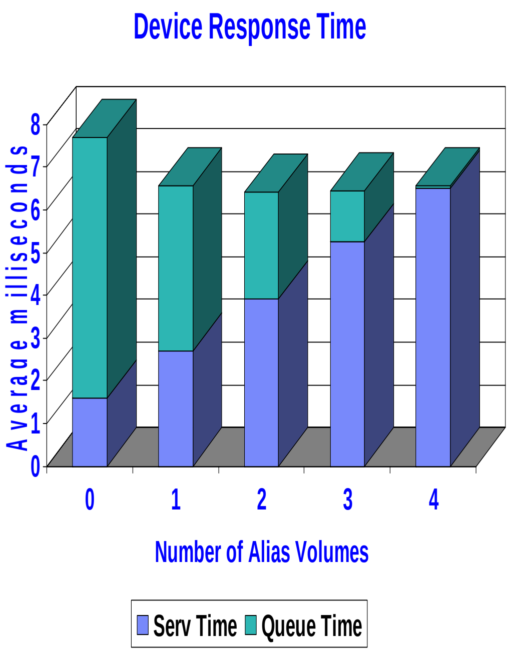
PAV Exploitation for VM Minidisks

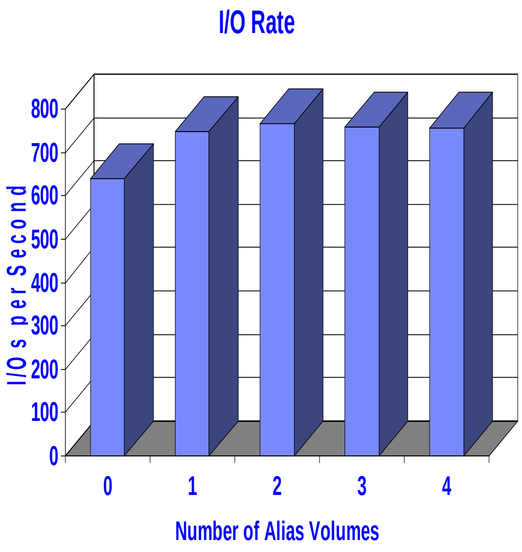
- **Previously only supported as dedicated disks.**
- **VM63855 (for z/VM 5.2.0, available May 2006):**
  - CP uses PAV to potentially decrease response time on minidisk I/O
  - We tightened the rules about ATTACHing or DEDICATEing PAV devices
- **VM63855 – virtualizes PAV for minidisks.**
- **Useful for environments where queuing on I/O occurs for minidisk I/O.**
- **Sometimes referred to as SYSTEM-owned PAV volumes**
- **PAV Base and Alias volumes defined on the Storage CU**
- **Summary of Results**
  - Varies depending on DASD CU Model
  - Varies depending on read-write mix
  - Helpful when I/O queuing occurs
  - Law of diminishing return; that is, defining more Alias than needed can lower performance

PAV – Rules of Thumb

- **Symptom:**
  - I/O wait queue forming at real volume where minidisks are
  - See Performance Toolkit FCX168 reports (or equivalent)
- **Remedy:**
  - Configure a PAV alias device in the storage controller
  - Make sure the alias device is varied online
  - Make sure the alias device is ATTACHed to SYSTEM
- **Measure:**
  - Re-run your workload
  - Look again at those disk performance reports
- **Success criterion:**
  - Response time equals service time (no wait queue)

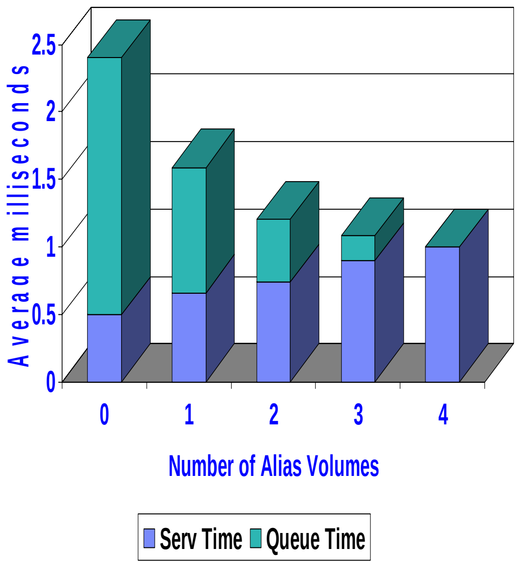
System Owned PAV Results – DS8100 – 100% Writes



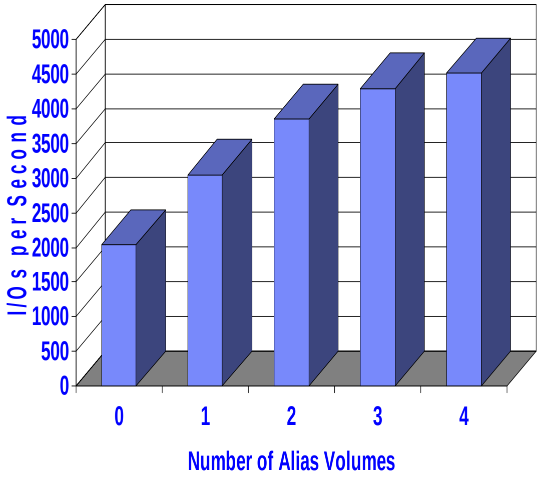


System Owned PAV Results – DS8100 – 100% Reads

Device Response Time

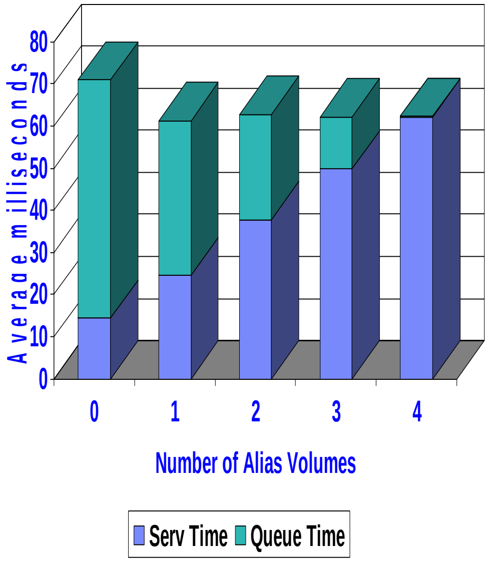


I/O Rate

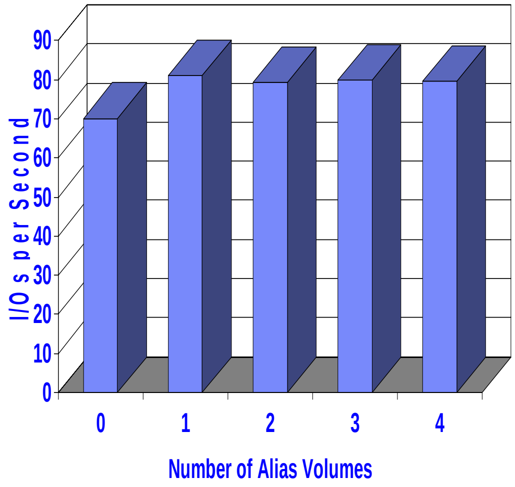


System Owned PAV Results – DS6800 – 100% Writes

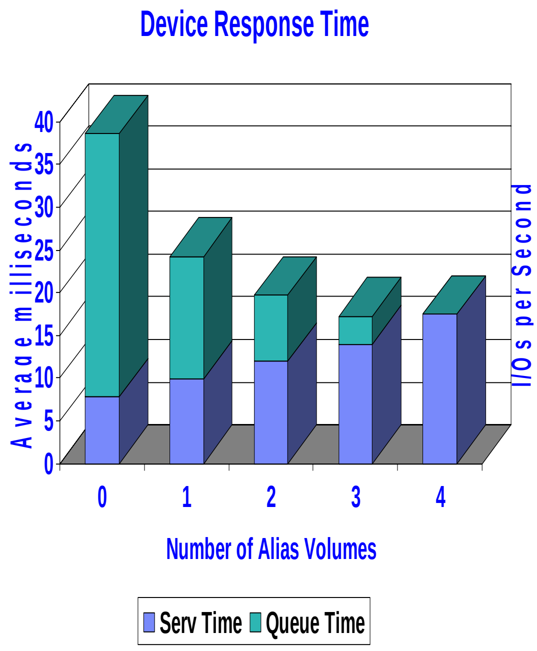
Device Response Time

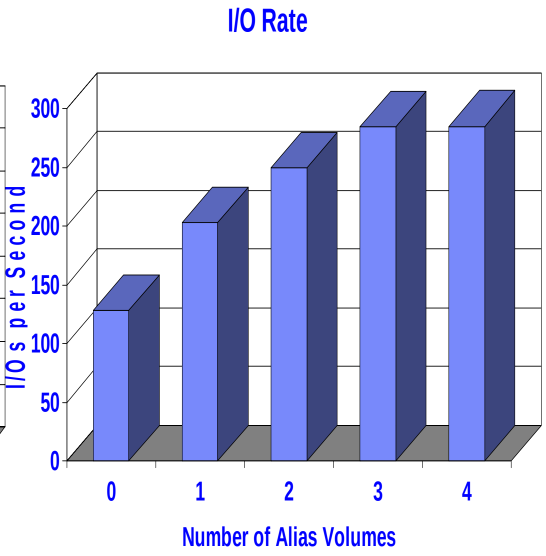


I/O Rate

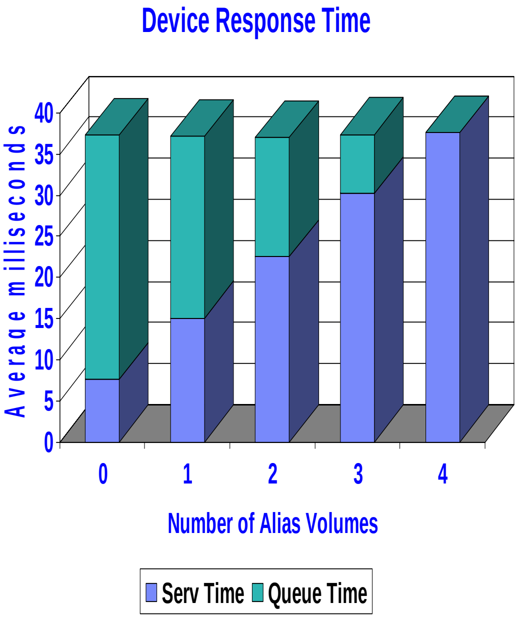


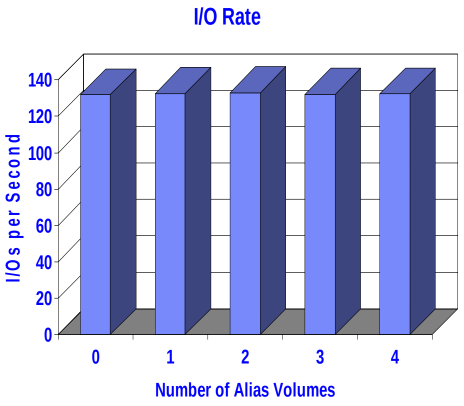
System Owned PAV Results – DS6800 – 100% Reads



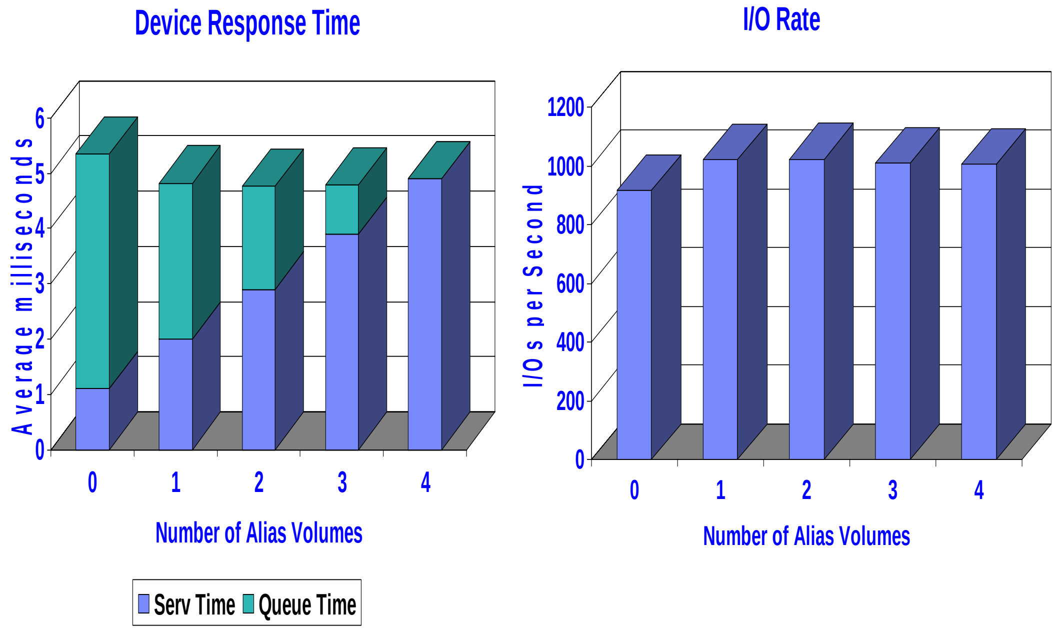


System Owned PAV Results – ESS F20 – 100% Writes



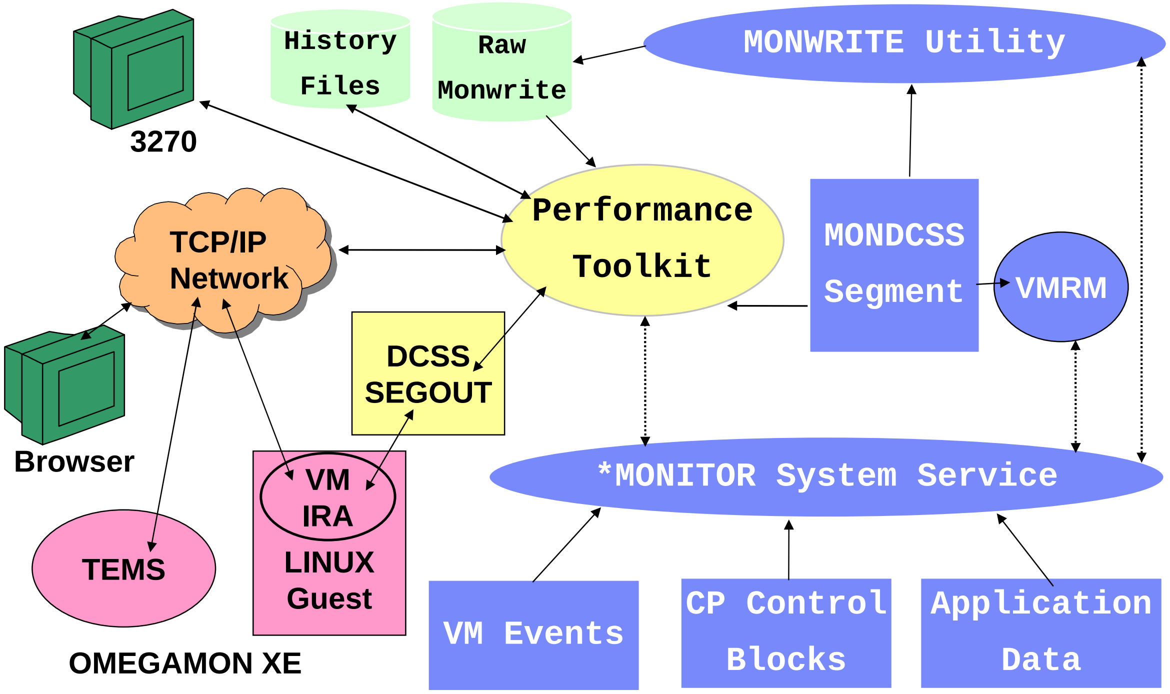


System Owned PAV Results – ESS F20 – 100% Reads


Device Response Time

I/O Rate

5,000 Foot View


3270

History Files

Raw Monwrite

MONWRITE Utility

TCP/IP Network

Performance Toolkit

MONDCSS Segment

VMRM

DCSS SEGOUT

Browser

VM IRA

LINUX Guest

*MONITOR System Service

TEMS

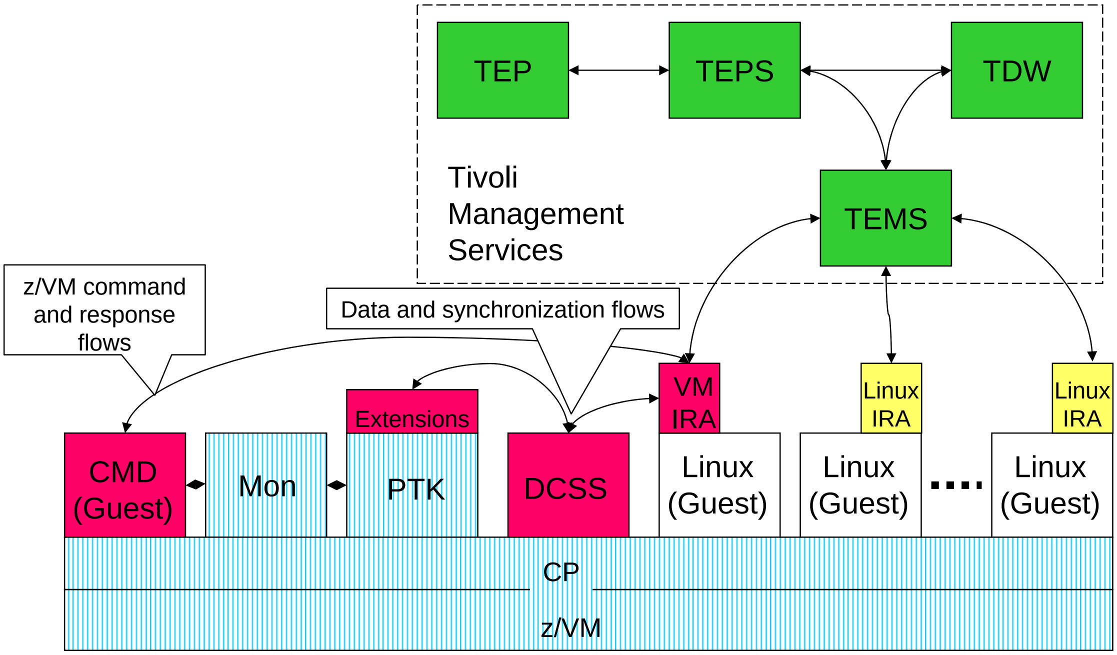
OMEGAMON XE

VM Events

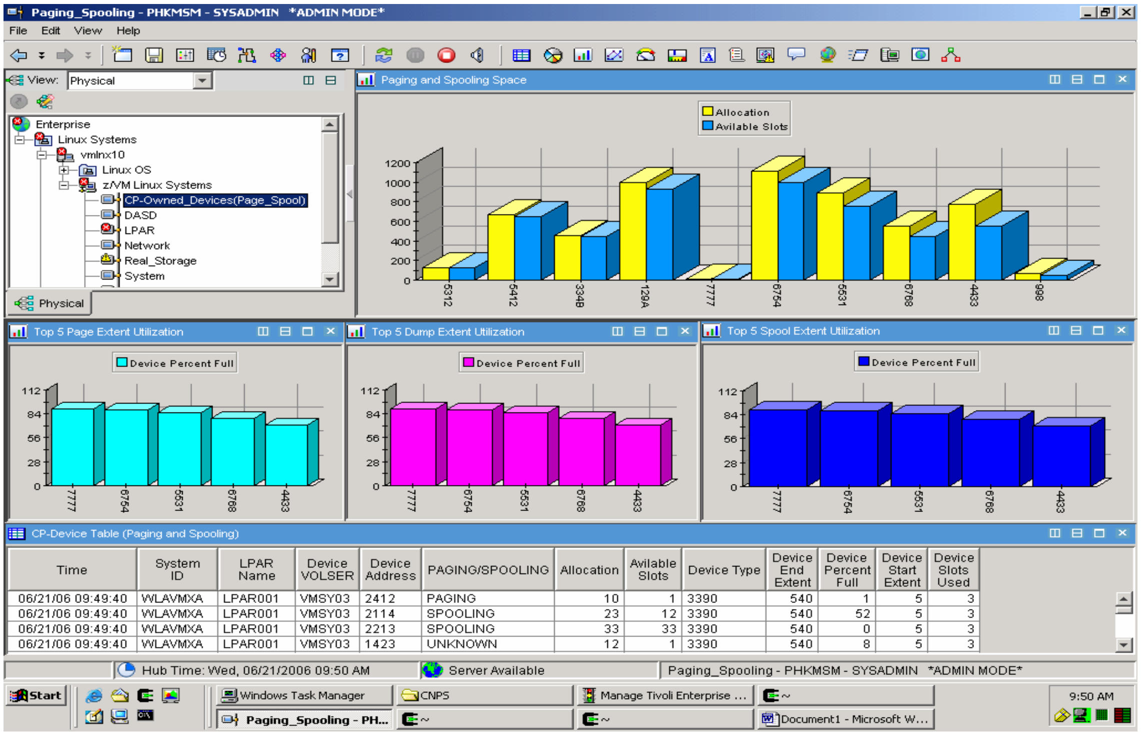
CP Control Blocks

Application Data

Basic Architecture



PAGING and SPOOLING Utilization



Time	System ID	LPAR Name	Device VOLSER	Device Address	PAGING/SPOOLING	Allocation	Avilable Slots	Device Type	Device End Extent	Device Percent Full	Device Start Extent	Device Slots Used
06/21/06 09:49:40	WLAVMXA	LPAR001	VMSY03	2412	PAGING	10	1	3390	540	1	5	3
06/21/06 09:49:40	WLAVMXA	LPAR001	VMSY03	2114	SPOOLING	23	12	3390	540	52	5	3
06/21/06 09:49:40	WLAVMXA	LPAR001	VMSY03	2213	SPOOLING	33	33	3390	540	0	5	3
06/21/06 09:49:40	WLAVMXA	LPAR001	VMSY03	1423	UNKNOWN	12	1	3390	540	8	5	3

Hub Time: Wed, 06/21/2006 09:50 AM

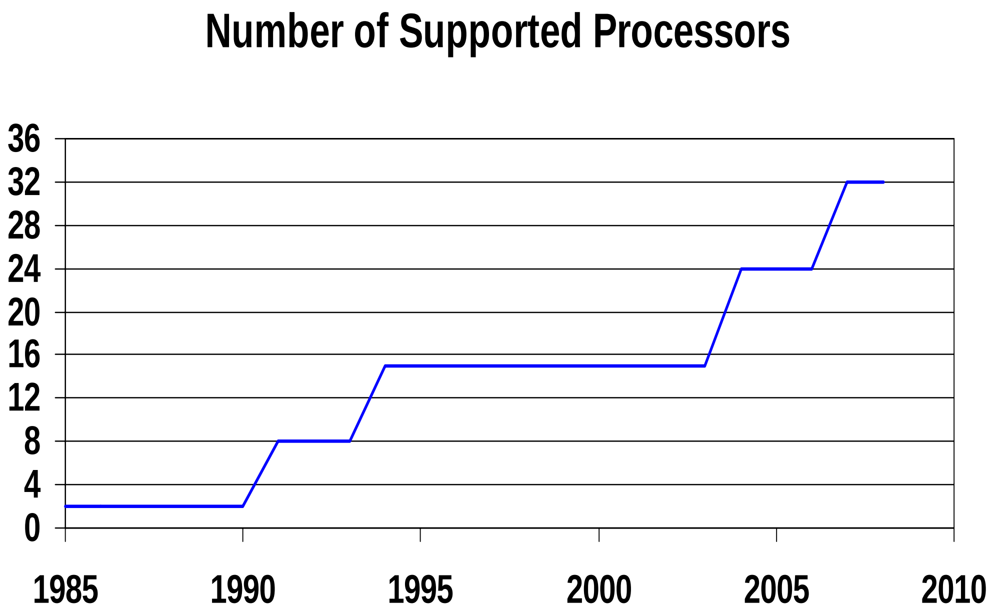
Server Available

Paging_Spooling - PHKMSM - SYSADMIN *ADMIN MODE*

z/VM 5.3.0

- **GA June 29, 2007**
  - Many more details available in the z/VM Performance Report:
    - http://www.vm.ibm.com/perf/reports/
- **Scalability and capability extended in several directions**
  - Processors, Memory, I/O, Network
  - What were the old limits?
  - What are the new limits?
- **Other Performance Enhancements**

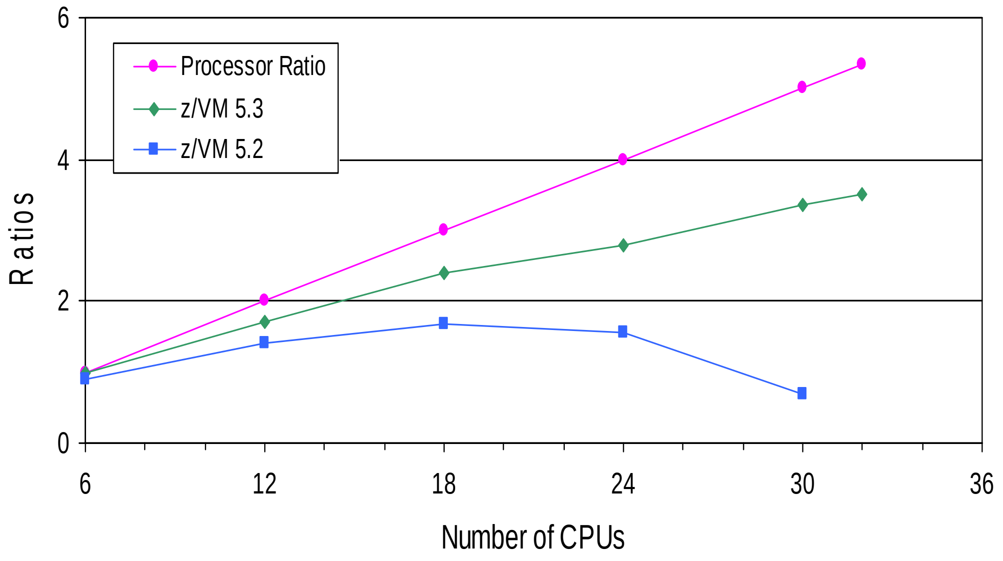
Processor Scaling



Greater than 24 CPU Support

- **While z/VM 5.2.0 would run on up to 31 processors, it only supported 24 due to performance limitations**
- **z/VM 5.3.0 supports 32 processors**
- **Serialization Changes**
  - General support for exclusive and shared formal spin locks
  - First to exploit is the Scheduler Lock (SRMSLOCK)
  - New lock associated with each Processor Local Dispatch Vector (PLDV) for dispatching (DSVLOCK)
- **Performance is Workload Dependent**
  - Watch for Master Processor Limitations
    - Tend to be more traditional workloads, not Linux environments
  - Single non-MP virtual machine limits
    - Example: DB2 for z/VM & VSE can only use 1 processor

Large N-Way Effects on ITR Ratio



Linux Guests with Apache Webserving

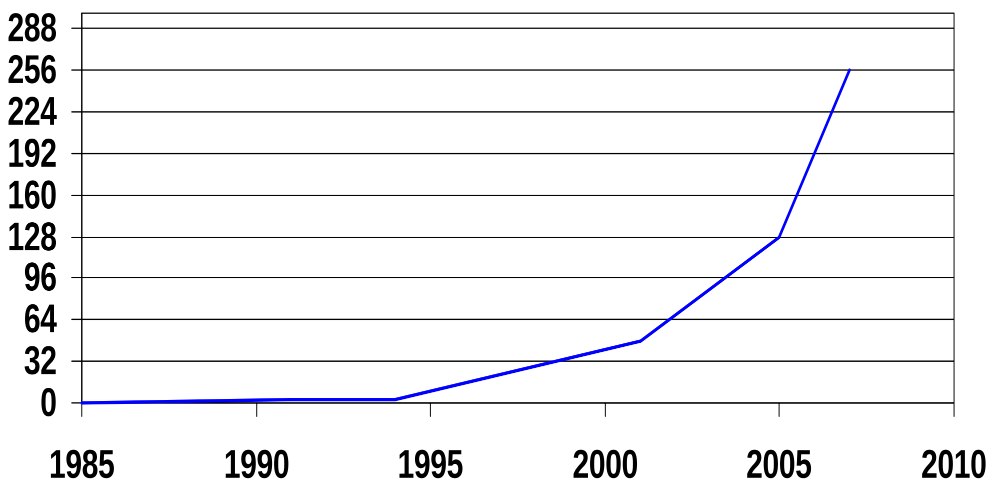
Metrics for Formal Spin Locks

```
FCX265      CPU 2094  SER 19B9E  Interval 02:31:51 - 12:34:01    GDLVM7

                     <------------------- Spin Lock Activity ------------------->
                     <----- Total -----> <--- Exclusive ---> <----- Shared ---->
Interval             Locks Average  Pct  Locks Average  Pct  Locks Average  Pct
End Time LockName     /sec    usec Spin   /sec    usec Spin   /sec    usec Spin
>>Mean>> SRMATDLK      1.9    .539 .000    1.9    .539 .000     .0    .000 .000
>>Mean>> RSAAVCLK       .0   2.015 .000     .0   2.015 .000     .0    .000 .000
>>Mean>> FSDVMLK        .0   24.97 .000     .0   24.97 .000     .0    .000 .000
>>Mean>> SRMALOCK       .0    .000 .000     .0    .000 .000     .0    .000 .000
>>Mean>> HCPTRQLK      4.1    .195 .000    4.1    .195 .000     .0    .000 .000
>>Mean>> SRMSLOCK     34.0   1.096 .001   32.7   1.037 .001    1.3    .001 .000
```

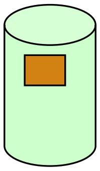
Memory Scaling

Effective Real Memory Use Limits

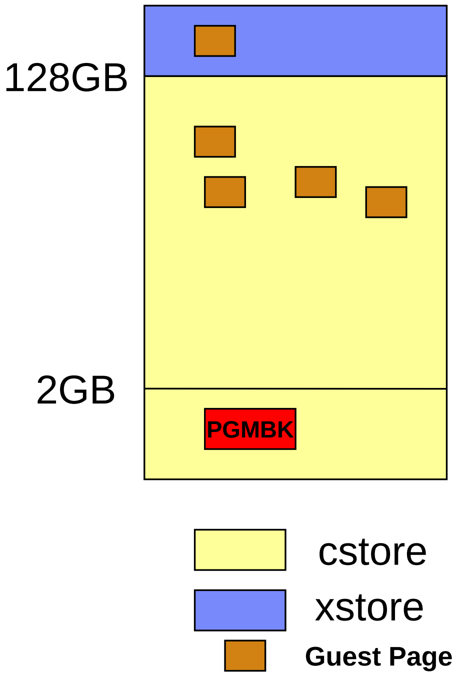


Greater than 128GB Memory Support (256GB)

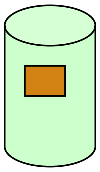
- **z/VM 5.3 Improvements:**
  - PGMBKs allowed to be allocated above 2GB
    - Each PGMBK is 8KB (2 contiguous frames)
  - Enhanced contiguous frame management
- **Also seeing improvements to smaller configurations that are memory constrained**
- **Be careful with memory terminology**
  - Try to define various terms when you use them or hear them
  - Examples:
    - Defined
    - Resident
    - Backed
    - Active
    - Actively Referenced
    - Addressable



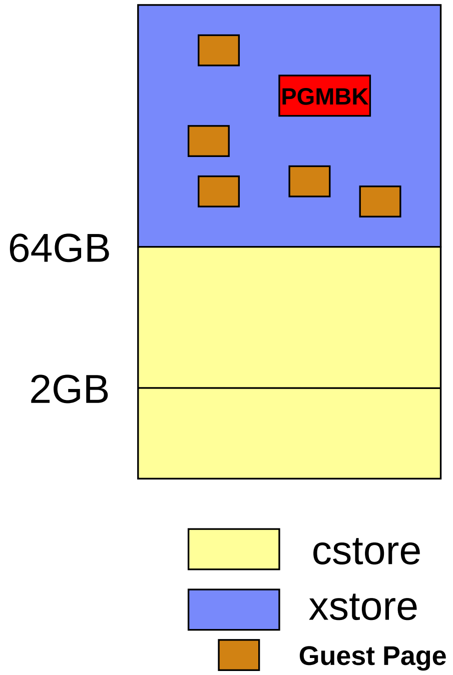
z/VM 5.2.0 Traditional Xstore:Cstore Ratio



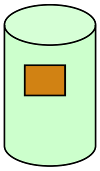
- PGMBK is required for any guest page that is acted on
- PGMBK is 8KB (2 contiguous frames)
- Resident PGMBK located below 2GB
- PGMBKs are pageable
- PGMBK resident if <u>any</u> guest pages it represents are resident
- Perfect world: ~256GB of virtual memory actively being referenced
- Realistic world: ~128GB of virtual memory actively being referenced



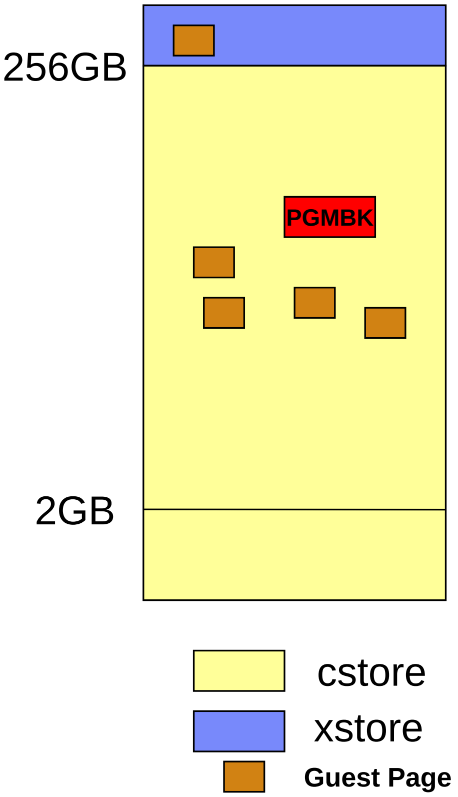
z/VM 5.2.0 Larger Xstore:Cstore Ratio



- More expanded storage can increase the likely hood that guest pages are moved out of central storage.
- This allows PGMBKs to also be moved.



z/VM 5.3.0 PGMBKs above 2GB Bar



- PGMBKs can reside anywhere but still must be better in the hierarchy than the guest pages they represent
- Next Limit: amount of PTRM space
  - Each space 4GB in size mapping 500GB of memory
  - Limit of 16 Spaces
  - Totals: 8TB of virtual machine memory

Performance Toolkit DSPACESH Report

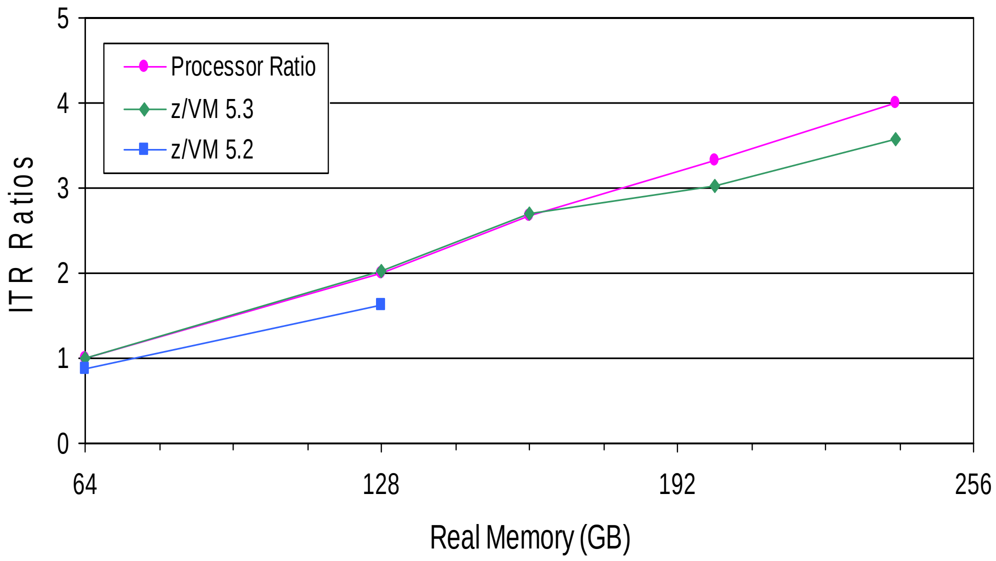
```
 FCX134      CPU 2094  SER 19B9E  Interval 13:04:01 - 13:09:01    GDLVM7
                             <------------------Number of Pages---------------->
Owning                            <--Resid--> <-Locked--> <-Aliases->
Userid    Data Space Name   Total Resid R<2GB  Lock L<2GB Count Lockd XSTOR DASD
SYSTEM    PTRM0000          1049k 35602  1104     0     0     0     0   980 7502
```

- Slightly edited FCX134 report
- PGMBKs live in PTRM0000, PTRM0001, … PTRM000F
- Most systems will just have a PTRM0000

Limitations for Memory

- **Memory limitations dependent on workload & configuration**
  - 256GB Real memory
  - 8TB of ‘addressable’ virtual machine memory – Limit of Page Tables
  - Paging Space (optimal when <50% full)
    - 11.2TB for ECKD
    - 15.9TB for Emulated FBA on FCP SCSI
  - Virtual Machine Size (HW Dependent)
    - 1TB on z9

Scaling Memory Results - Apache Webserving



Improvement in Memory Constrained Environments

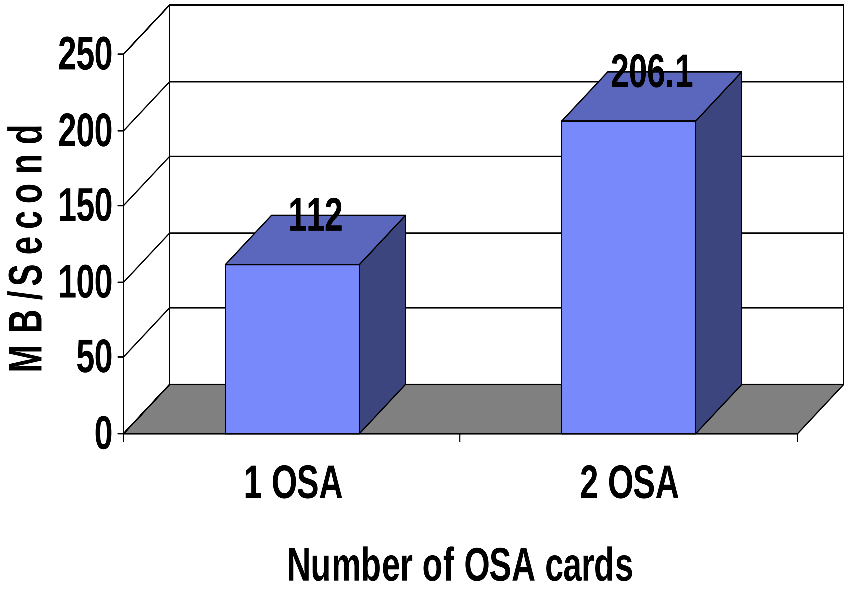
Scenario	Contention (Page Reqs per CPU-second)	Delta Thruput	Delta Total CPU/Tx
3G/4G 2084 3-way	2159	+10.3%	-9.5%
64G/2G 2094 3-way	0	+1.0%	-0.9%
64G/2G 2094 3-way	352	+15.5%	-12.7%
128G/2G 2094 6-way	291	+21.6%	-19.4%

Results of various Linux Apache measurements comparing z/VM 5.3 to z/VM 5.2

Virtual Switch Link Aggregation

- **Ability to attach multiple OSAs to a single virtual Switch**
  - Aggregate bandwidth
  - Failover
- **Requires:**
  - z9 OSA-Express2 Support
  - Running in Layer 2 Mode
- **Dynamic Load Balancing**
  - Influenced by distribution of MAC addresses
  - Influenced by Physical Switch for inbound traffic
  - Cannot balance a single connection.
    - Example: a single data streaming connection will not get split across OSAs.

Streaming Throughput Results



I/O Improvements

- **PAV support for minidisks was provided in z/VM 5.2 via APAR VM63855 (May 2006)**
  - Should also apply VM64199 if using Minidisk Cache
- **z/VM 5.3 adds support for the HyperPAV feature of IBM System Storage DS8000**
  - Requires VM64248 (UM32072)
  - Allows for the creation of pools of Alias devices which the control unit will associate with different Bases as needed.
  - Performance characteristics similar to previous PAV support.

Improved SCSI Disk Performance

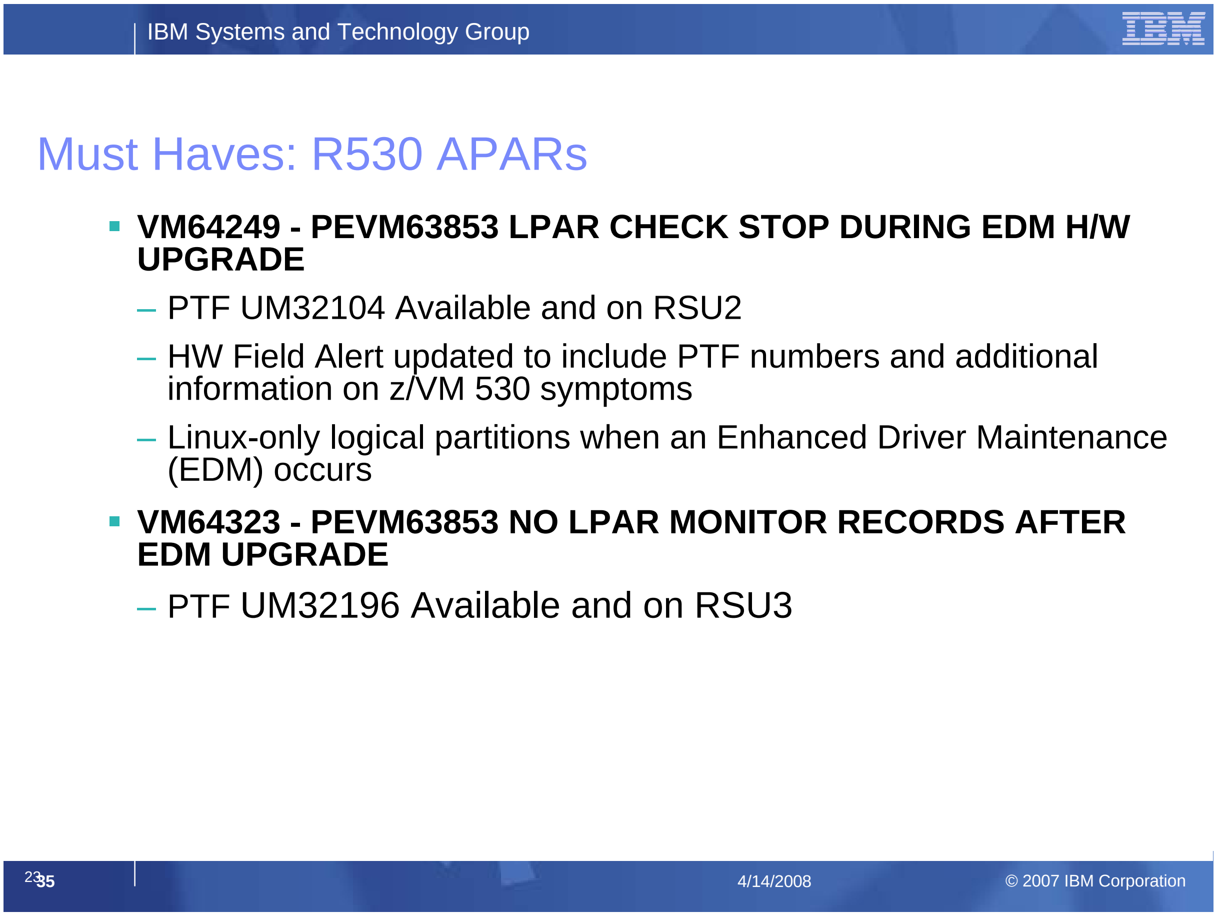
- **Exploitation of SCSI write-same function of 2105 & 2107 improves CMS FORMAT of minidisks on SCSI volumes**
- **Additional pathlength reductions**
- **CP Paging to SCSI volumes now bypasses the FBA emulation, reducing processor resource requirements**

Monitoring Enhancements

- **Lots of new fields in Monitor for new function:**
  - specialty engines
  - Scheduler changes
  - HyperPAV support
  - Memory management
- **New monitor Domain for Virtual Network Devices**
- **Additional flexibility in MONWRITE utility for starting/stopping**
- **Various changes in Performance Toolkit for VM**

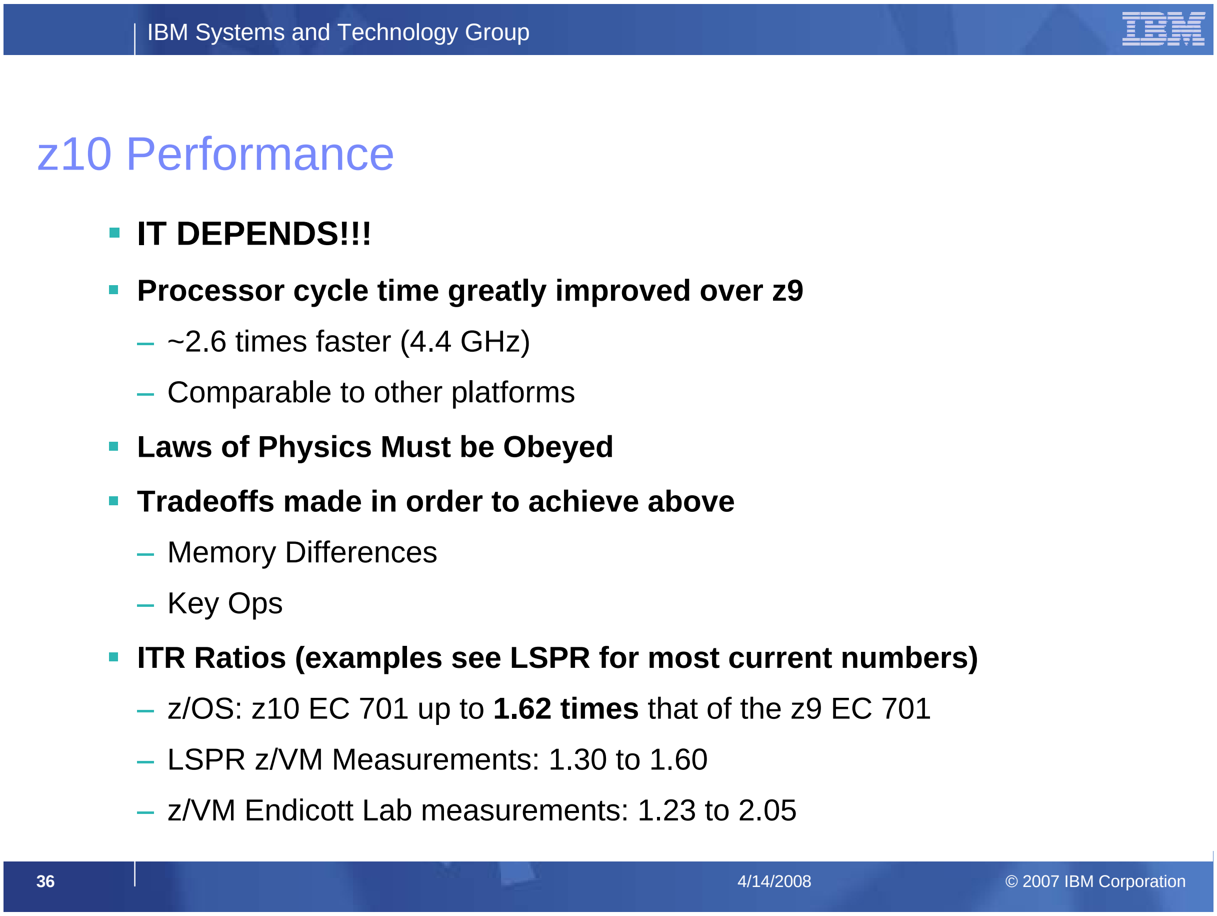
Service Must Haves: R530 APARs

- **VM64297 - PAGING SPIKE DEMAND SCAN SHUTS DOWN SYSTEM HANG**
  - PTF UM32197 Available and on RSU3
- **VM64269 - EXCESSIVE PAGING ACTIVITY DURING DEMAND SCAN**
  - PTF UM32133 Available and on RSU2
- **VM64287 - SLOW PING RESPONSE TIMES OVER QDIO WITH MORE THAN ONE VIRTUAL PROCESSOR**
  - PTF UM32158 available 10/18/07



Must Haves: R530 APARs

- **VM64249 - PEVM63853 LPAR CHECK STOP DURING EDM H/W UPGRADE**
  - PTF UM32104 Available and on RSU2
  - HW Field Alert updated to include PTF numbers and additional information on z/VM 530 symptoms
  - Linux-only logical partitions when an Enhanced Driver Maintenance (EDM) occurs
- **VM64323 - PEVM63853 NO LPAR MONITOR RECORDS AFTER EDM UPGRADE**
  - PTF UM32196 Available and on RSU3



z10 Performance

- **IT DEPENDS!!!**
- **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

It 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

Setting the 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

- **z/VM 5.2 Improvements via Service**
- **z/VM 5.3 significantly extends the capacity of:**
  - Processor
  - Memory
  - I/O
- **See z/VM Performance Report for more details**
  - http://www.vm.ibm.com/perf/reports/
- **Learn more about z/VM**
  - http://www.vm.ibm.com/events/