V91

z/VM Performance Update

Bill Bitner, IBM, bitnerb@us.ibm.com

Brian Wade, IBM, bkw@us.ibm.com

IBM System z Expo
September 17-21, 2007
San Antonio, TX

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 news**
  - PAV
  - CMM & VMRM
  - OMEGAMON XE
- **z/VM 5.2 limits**
- **z/VM 5.3 performance**
  - Workloads that tend to benefit
  - Line items that have an impact
- **See Performance Report on Web for details**
  - http://www.vm.ibm.com/perf/reports/

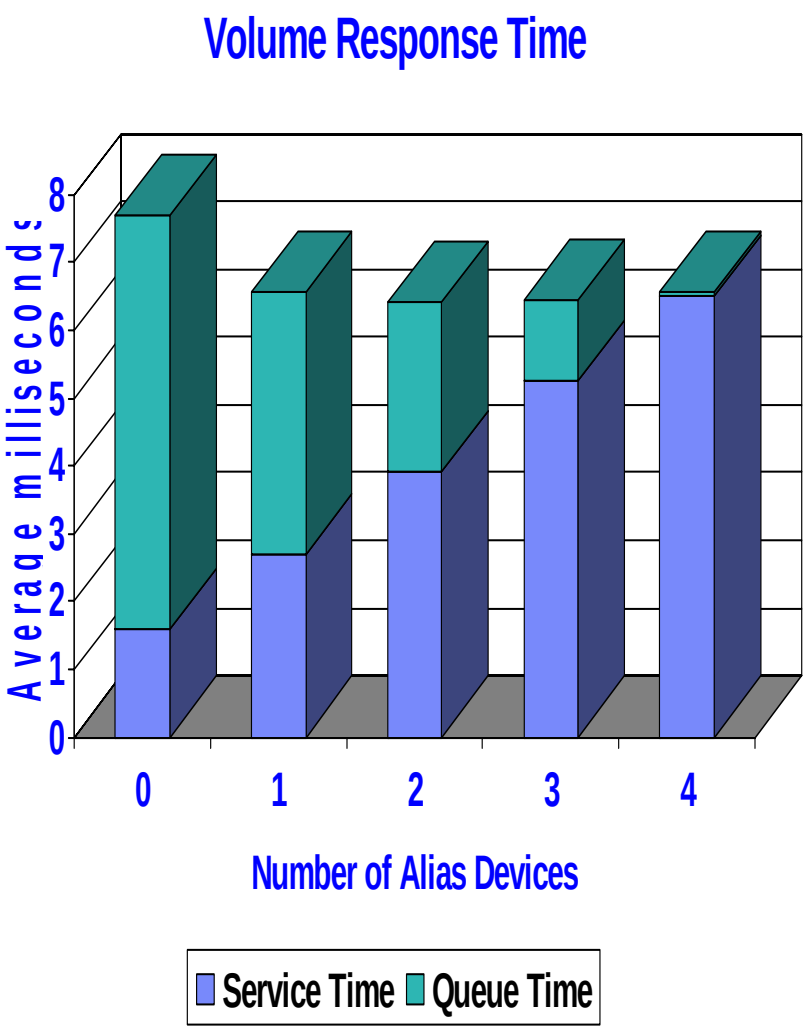
PAV Exploitation for VM Minidisks

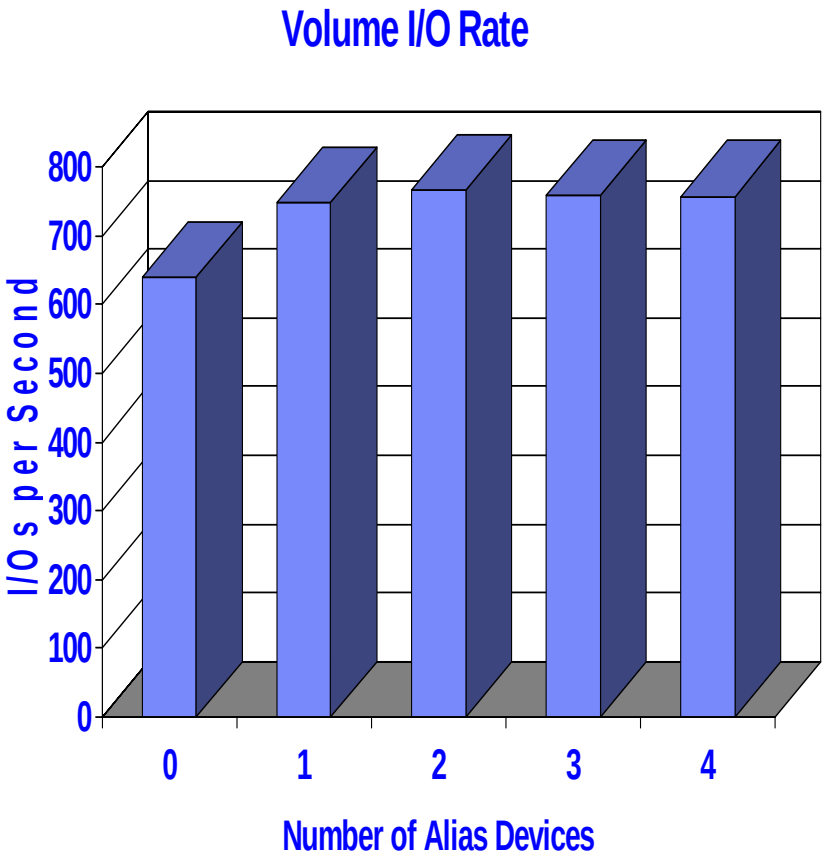
- **Previously supported only as dedicated disks.**
- **VM63855 (for z/VM 5.2, available May 2006):**
  - CP uses PAV to run multiple concurrent I/Os to user extents
  - We tightened the rules about ATTACHing or DEDICATEing PAV devices
- **VM63855 – also virtualizes PAV for minidisks.**
- **Useful for 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 subsystem**
- **Summary of results**
  - Varies depending on model of storage subsystem
  - 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

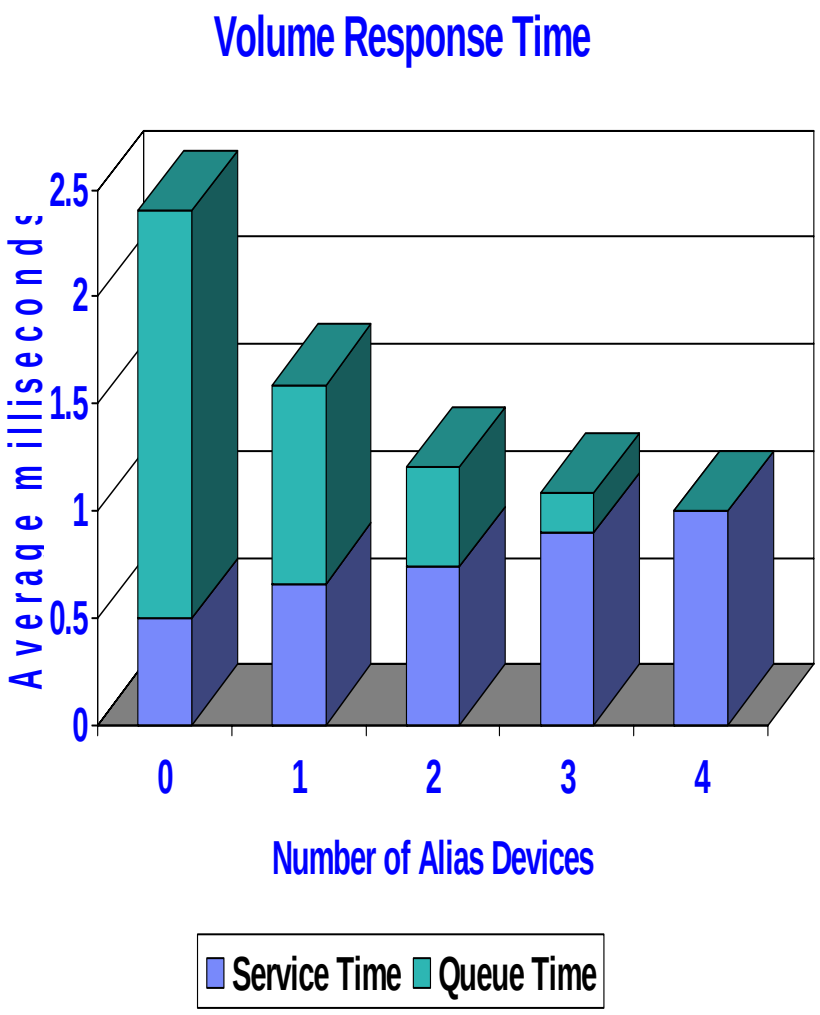
- **Symptom:**
  - I/O wait queue forming at real volume where minidisks are
  - See Performance Toolkit FCX168 reports (or equivalent)
- **Remedy:**
  - Configure a PAV alias device in the storage subsystem
  - Make sure the alias device is varied online
  - Make sure the alias device is ATTACHed to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

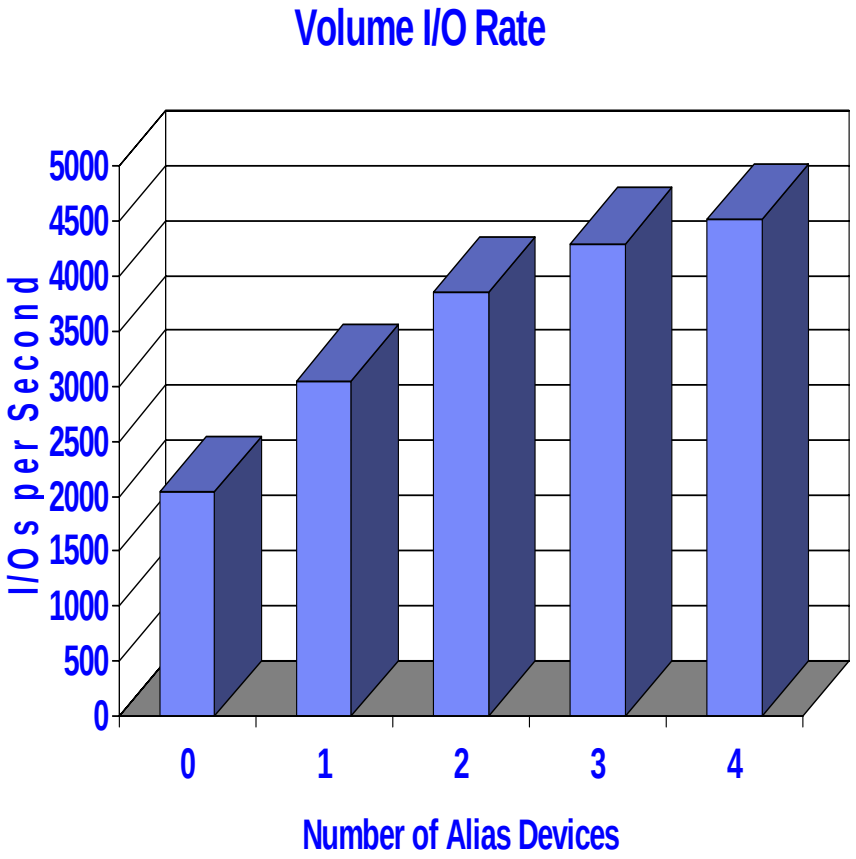
System-Owned PAV Results – DS8100 – 100% Writes



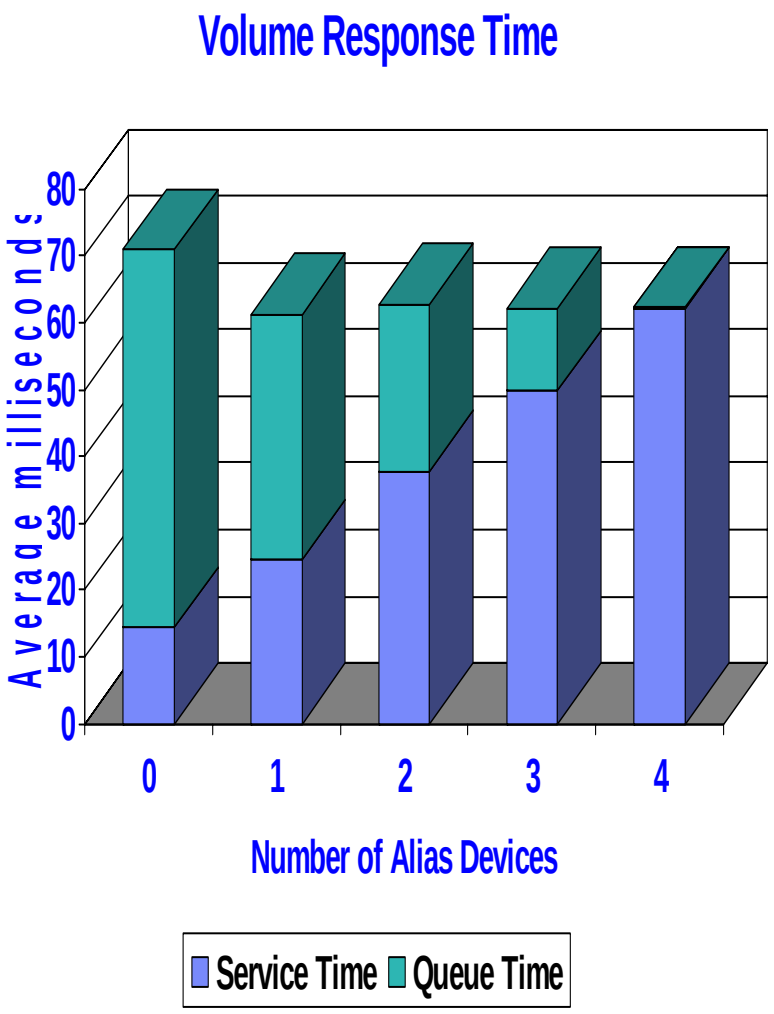


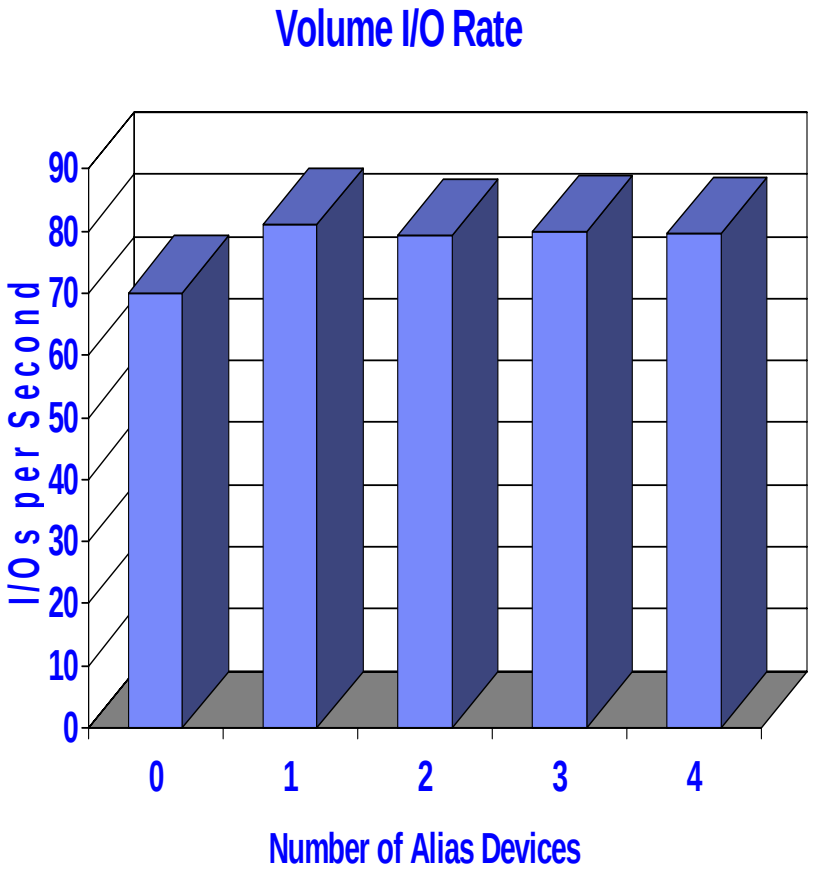
System-Owned PAV Results – DS8100 – 100% Reads



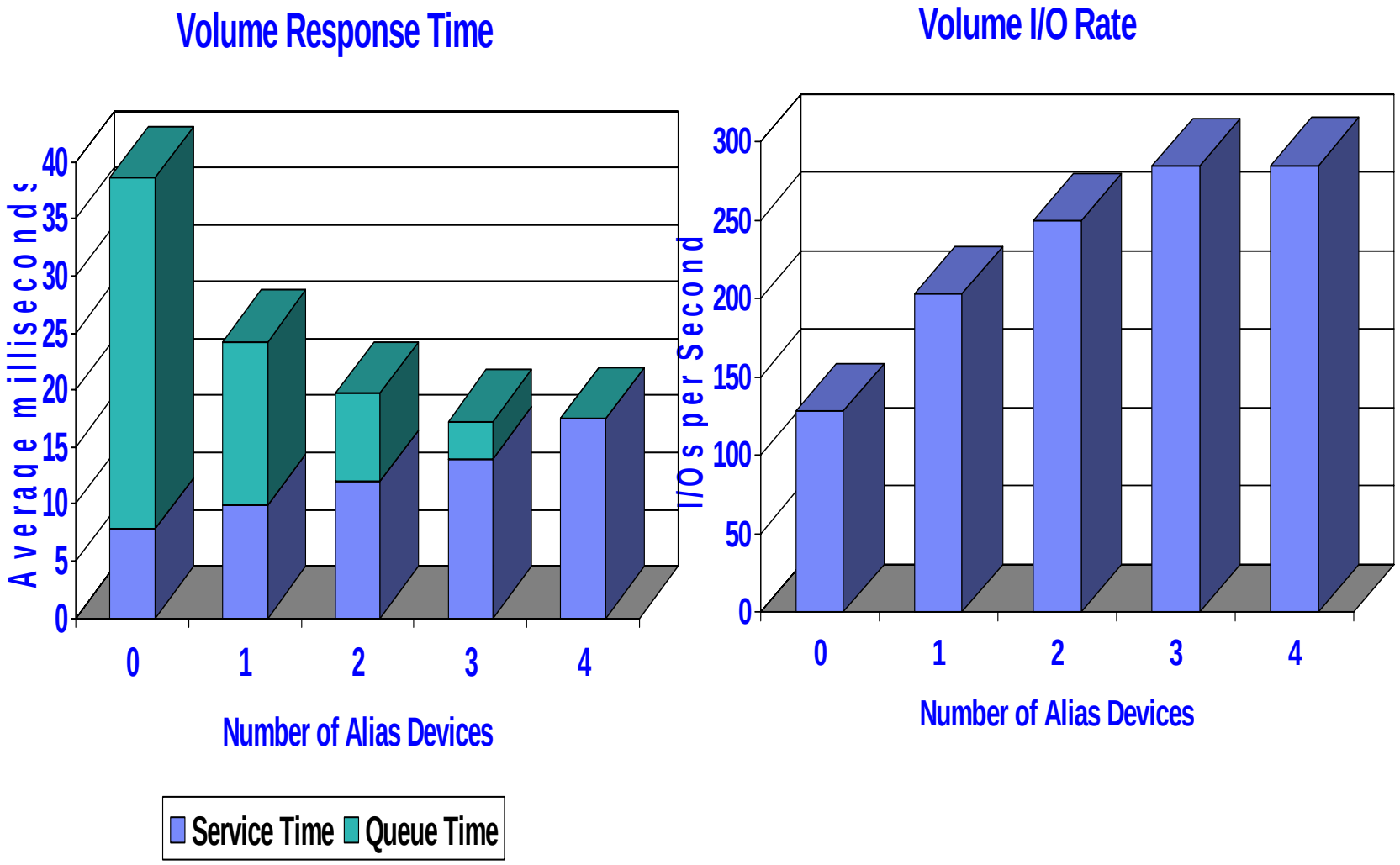


System-Owned PAV Results – DS6800 – 100% Writes

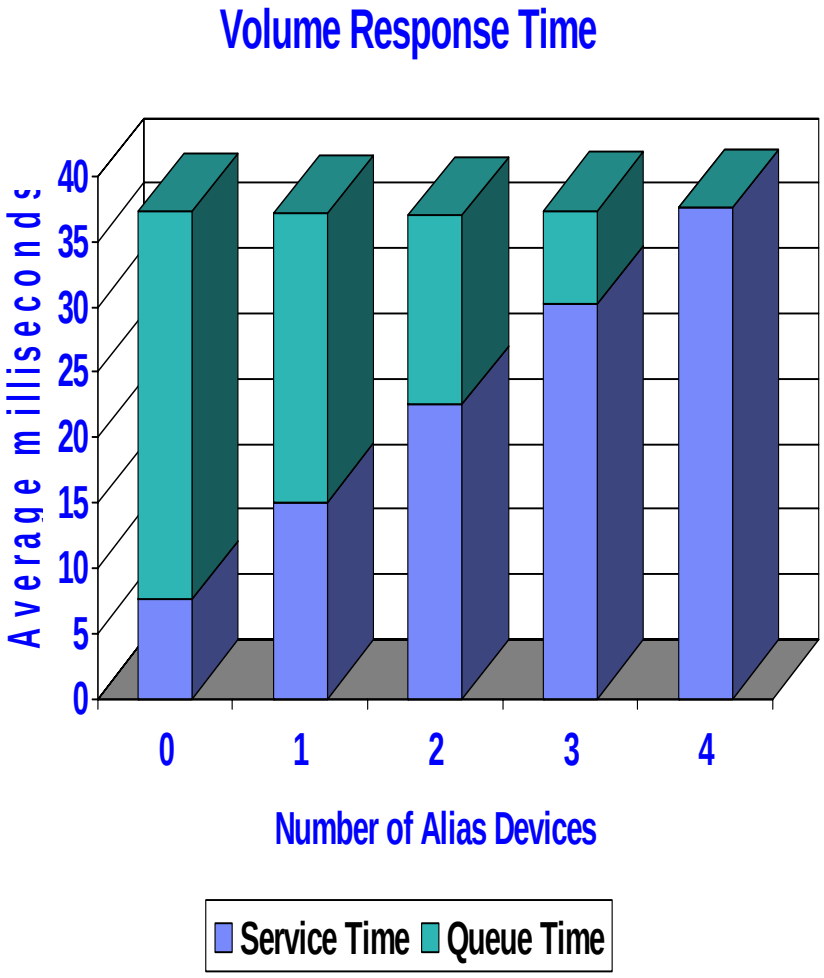


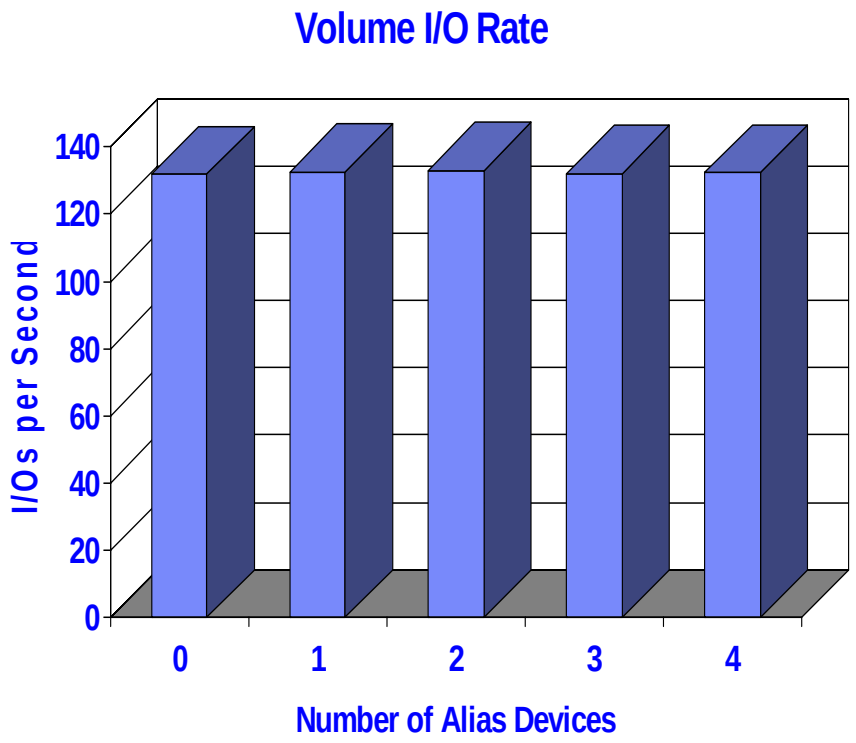


System-Owned PAV Results – DS6800 – 100% Reads

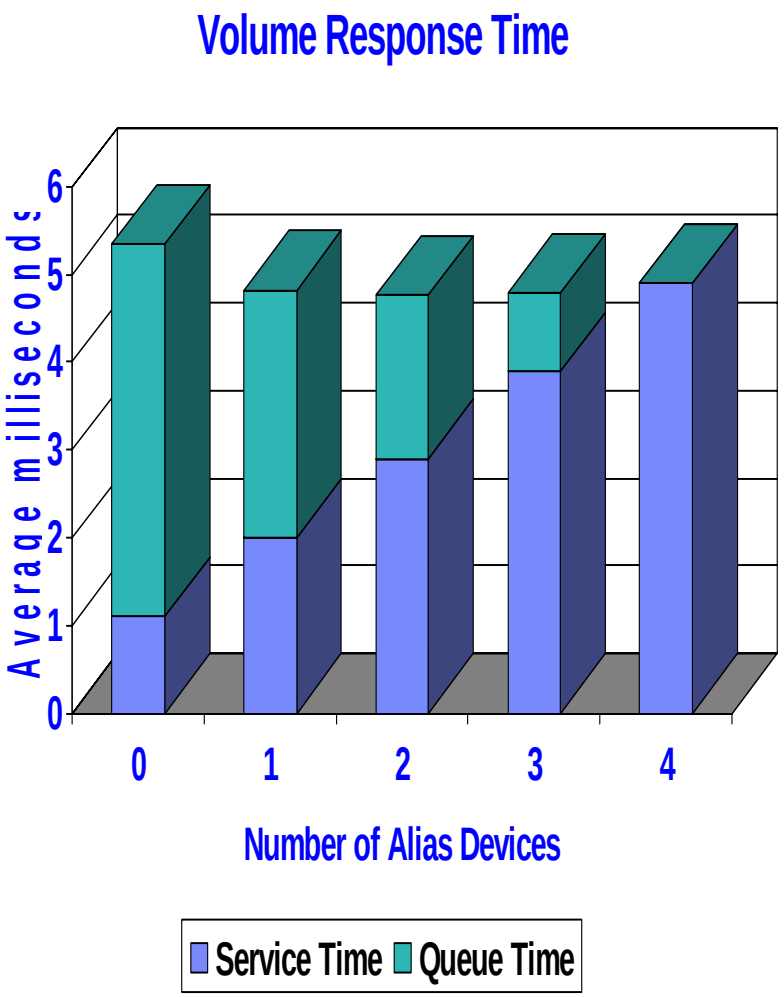


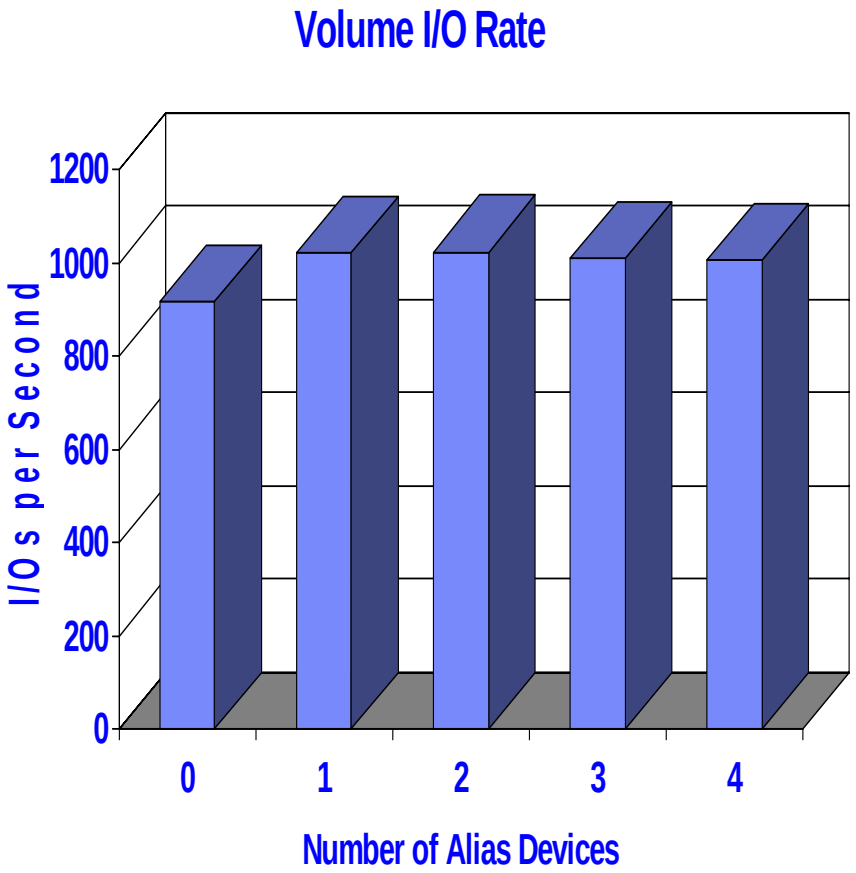
System-Owned PAV Results – ESS F20 – 100% Writes





System-Owned PAV Results – ESS F20 – 100% Reads

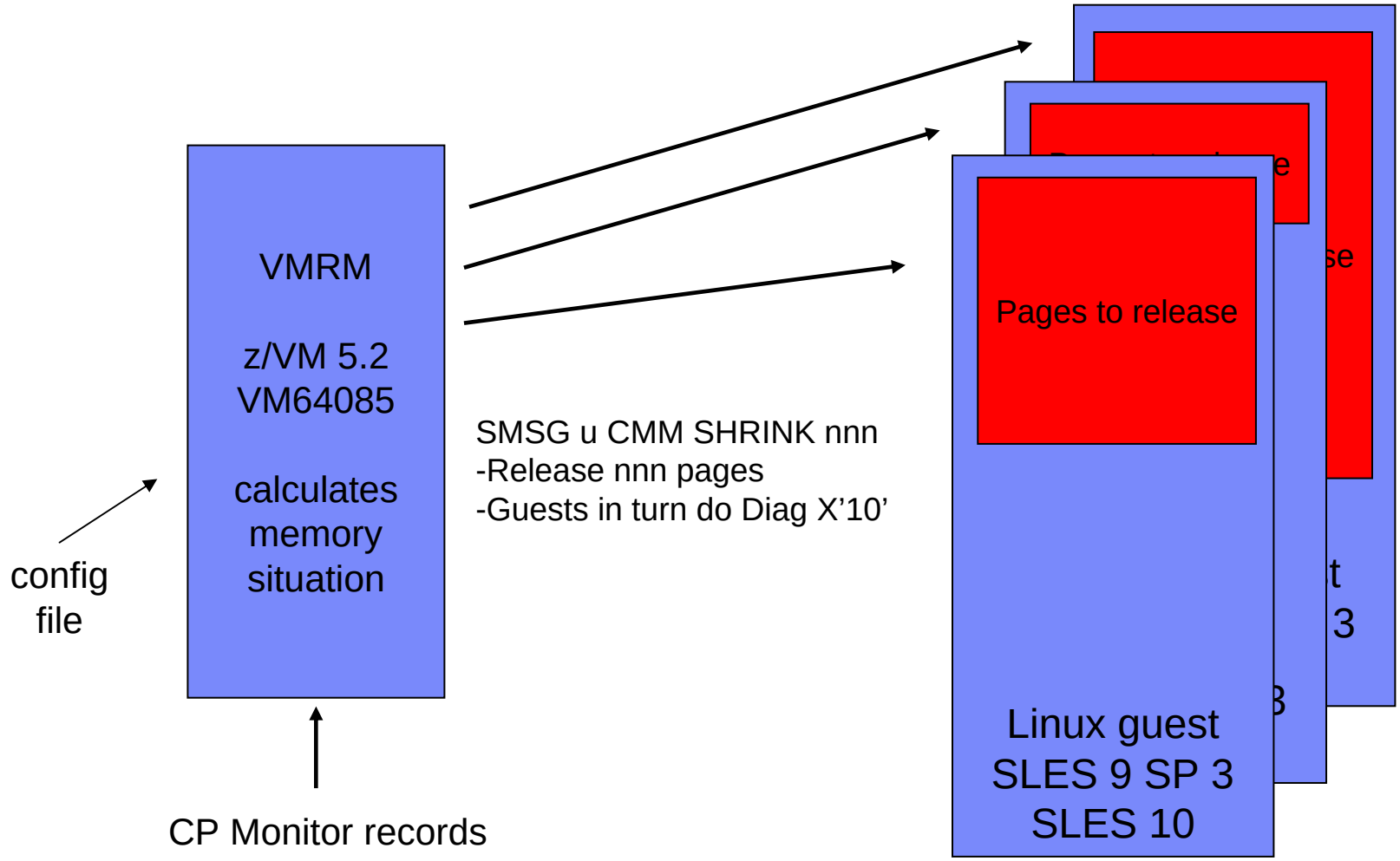




Cooperative Memory Management

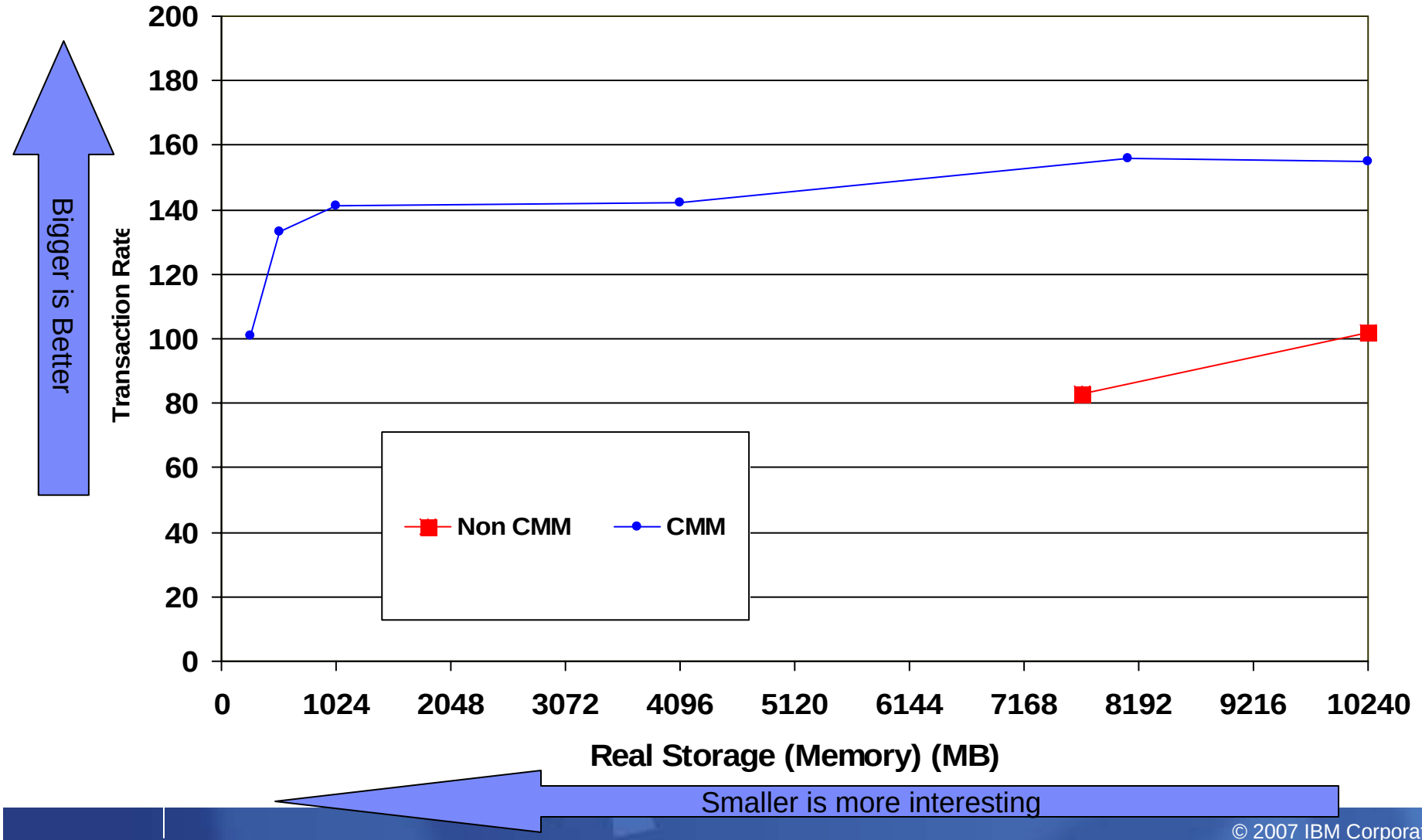
- **VMRM support of Linux CMM API**
- **Evaluates z/VM memory usage and notifies Linux guests to release memory**
- **z/VM 5.2 APAR VM64085**
  - http://www.vm.ibm.com/sysman/vmrm/vmrmcmm.html

VMRM: Helping Linux Guests "Right-Size" Since 2006



Transaction Rate vs. Real Storage

32 Apache servers each with a virtual memory size of 1.5GB



z/VM 5.2 PGMBK Limit

- **PGMBKs (page management blocks) must all fit below 2 GB.**
  - This affects how much guest storage can be resident at any one instant.
  - Perfect world: 256 GB of resident guest pages
    - Below-2-GB storage is completely filled with PGMBKs
    - Each PGMBK is maximally filled with descriptors of guest pages
  - Issues:
    - What else is below 2 GB besides PGMBKs
    - Fragmentation below 2 GB (a PGMBK requires two contiguous pages)
    - How densely filled the PGMBKs are
  - z/VM 5.2 operation guidelines:
    - Yellow light at 1 GB of PGMBKs (262k pages)
    - Red light at 1.5 GB of PGMBKs (393k pages)
- **z/VM 5.3 removes this limit**

FCX134, DSPACESH

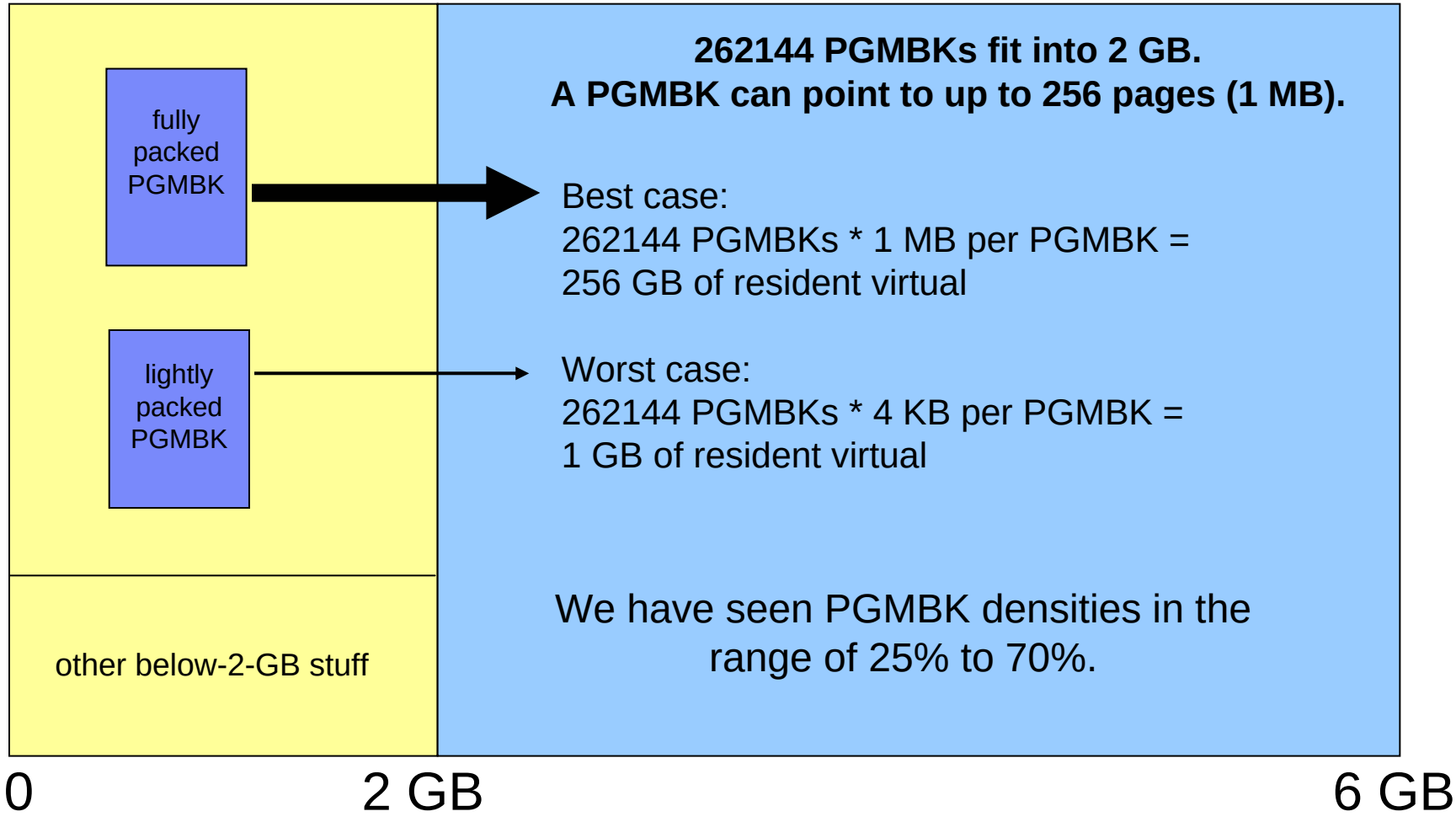
```
FCX134  Run 2007/06/04 12:39:06         DSPACESH                                                    Page    3
                                        Shared Data Spaces Paging Activity
From 2007/06/04 12:15:45                                                                GDLSPRF1
To   2007/06/04 12:38:45                                                                CPU 2094-733   SN 46A8D
For   1380 Secs 00:23:00                This is a performance report for GDLSPRF1       z/VM   V.5.2.0 SLU 0000
```

Owning Userid	Data Space Name	Users Permt	Rate per Sec. Pgstl	Pgrds	Pgwrt	X-rds	X-wrt	X-mig	Number of Pages Total	Resid	Resid R<2GB	Locked Lock	Locked L<2GB	Aliases Count	Aliases Lockd	XSTOR	DASD
>System<	--------	0	4.921	.805	1.022	4.441	4.417	.016	4194k	28840	28840	0	0	0	0	169	394
SYSTEM	FULL$TRACK$CACHE$1	0	.000	.000	.000	.000	.000	.000	524k	0	0	0	0	0	0	0	0
SYSTEM	FULL$TRACK$CACHE$2	0	.000	.000	.000	.000	.000	.000	524k	0	0	0	0	0	0	0	0
SYSTEM	FULL$TRACK$CACHE$3	0	.000	.000	.000	.000	.000	.000	524k	0	0	0	0	0	0	0	0
SYSTEM	FULL$TRACK$CACHE$4	0	.000	.000	.000	.000	.000	.000	524k	0	0	0	0	0	0	0	0
SYSTEM	ISFCDATASPACE	0	.000	.000	.000	.000	.000	.000	524k	0	0	0	0	0	0	0	0
SYSTEM	PTRM0000	0	44.24	7.243	9.171	39.95	39.70	.093	524k	260k	260k	0	0	0	0	1492	170
SYSTEM	REAL	0	.000	.000	.000	.000	.000	.000	34M	0	0	0	0	0	0	0	0
SYSTEM	SYSTEM	0	.041	.000	.026	.020	.041	.045	524k	0	0	0	0	0	0	24	3351
SYSTEM	VIRTUAL$FREE$STORAGE	0	.002	.001	.002	.001	.002	.002	524k	0	0	0	0	0	0	1	24

PGMBKs live in the PTRM0000 address space.
This system has 260,000 PGMBK pages resident, out of a possible 524,288.

Notes:
- On z/VM 5.1 the reported value is *half* of the actual value (it's a bug)
- On z/VM 5.2 the reported value is *somewhat higher* than the actual value (it's a bug)
- On z/VM 5.3 the reported value is correct to within a small tolerance

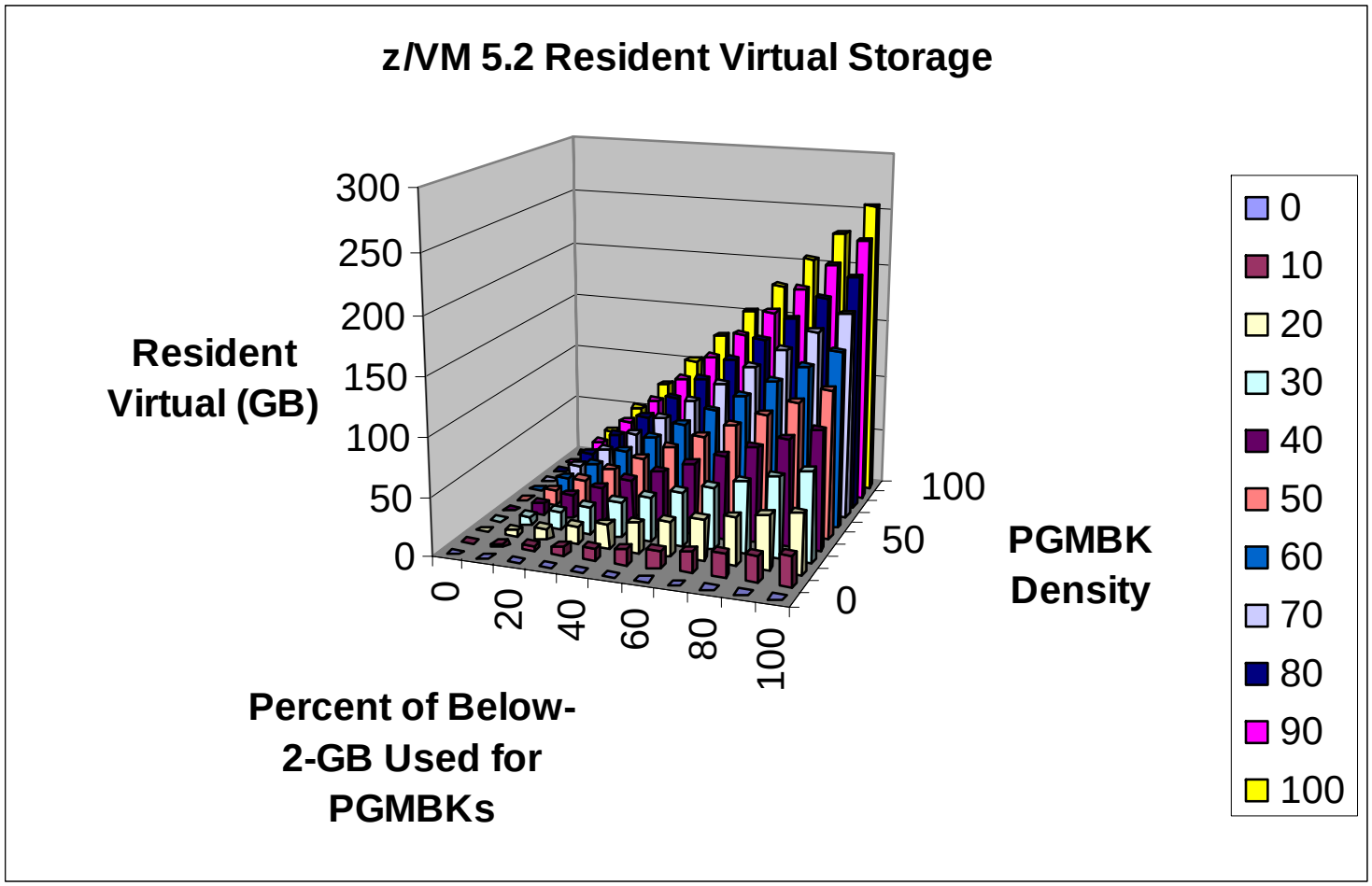
z/VM 5.2 PGMBK Density –- Best and Worst Cases



So What Determines **Your** Limit on Resident Virtual?

- **How many PGMBKs happen to fit into your below-2-GB storage**
  - This is largely out of your guests' control
- **How densely filled those PGMBKs are**
  - This is largely a function of your guests' page reference and page release habits
- **See graph next page**

z/VM 5.2 PGMBKs and Resident Virtual



How Much Resident Virtual Do I Have?

- **FCX113 UPAGE gives, for each user:**
  - Resident < 2 GB
  - Resident > 2 GB
  - XSTOR
  - DASD
- **>SYSTEM< line gives average values**
- **Multiply averages by number of users to get totals**

FCX113, UPAGE

```
FCX113  Run 2007/06/04 12:39:06         UPAGE                                                                  Page
67
                                        User Paging Activity and Storage Utilization
From 2007/06/04 12:15:45                                                                        GDLSPRF1
To   2007/06/04 12:38:45                                                                        CPU 2094-733   SN
46A8D
For   1380 Secs 00:23:00                This is a performance report for GDLSPRF1               z/VM   V.5.2.0 SLU
0000
```

Userid	Data Spaces Owned	Paging Activity/s <Page Rate> Reads	Page Rate> Write	Page Steals	<--Page Migration--> >2GB>	X>MS	MS>X	X>DS	Number of Pages WSS	Resrvd	<-Resident-> R<2GB	R>2GB	<--Locked--> L<2GB	L>2GB	XSTOR	DASD	Stor Size	Nr of Users
>System<	.0	34.8	49.4	77.1	.0	27.1	76.1	50.5	1082k	0	8574	1073k	40	67	13315	42989	4786M	29
User Class Data:																		
CMS1_Use	.0	.0	.3	.3	.0	.1	.3	1.2	100	0	0	1	0	1	415	3063	512M	1
LNX_Clie	.0	17.2	30.2	53.9	.0	28.2	53.1	30.7	73554	0	7934	65639	1	11	6165	16547	1024M	4
LX9_Serv	.0	71.2	94.0	120	.0	23.6	118	95.4	2388k	0	15608	2372k	0	13	23820	84504	10240M	13
User Data:																		
FTPSERVE	.0	.0	.0	.0	.0	.3	.3	.0	19	0	1	2	0	1	18	334	32M	
GCSXA	.0	.0	.0	.0	.0	.0	.0	.0	51	0	0	1	0	1	0	51	16M	

This system has [(8574 + 1073k) * 29] = 119.7 GB of resident virtual.

Calculating PGMBK Density of Our System (If You Care)

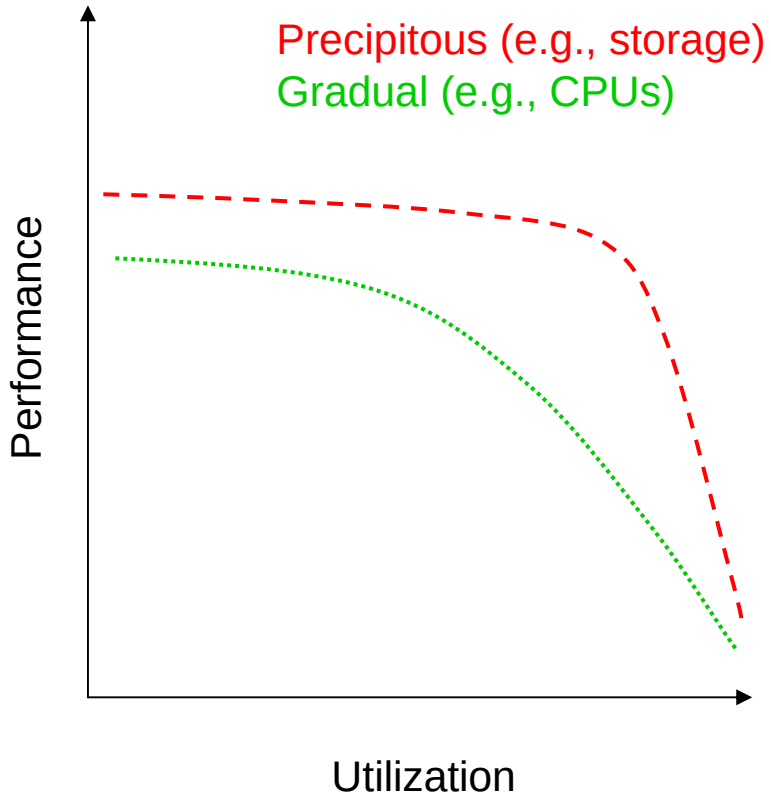
- **119.6 GB resident virtual * 1024 = 122,470 MB**
- **But we have (260,000 / 2) = 130,000 PGMBKs**
- **So we have 122,470 / 130,000 = 0.94 MB / PGMBK**
  - It's probably a little denser because the z/VM 5.2 reported PGMBK count is a little higher than it should be
- **This is a very dense situation**
- **This workload was contrived by our measurement experts**
- **You will probably never see a density this high**

z/VM 5.2 PGMBK Limit Recommendations

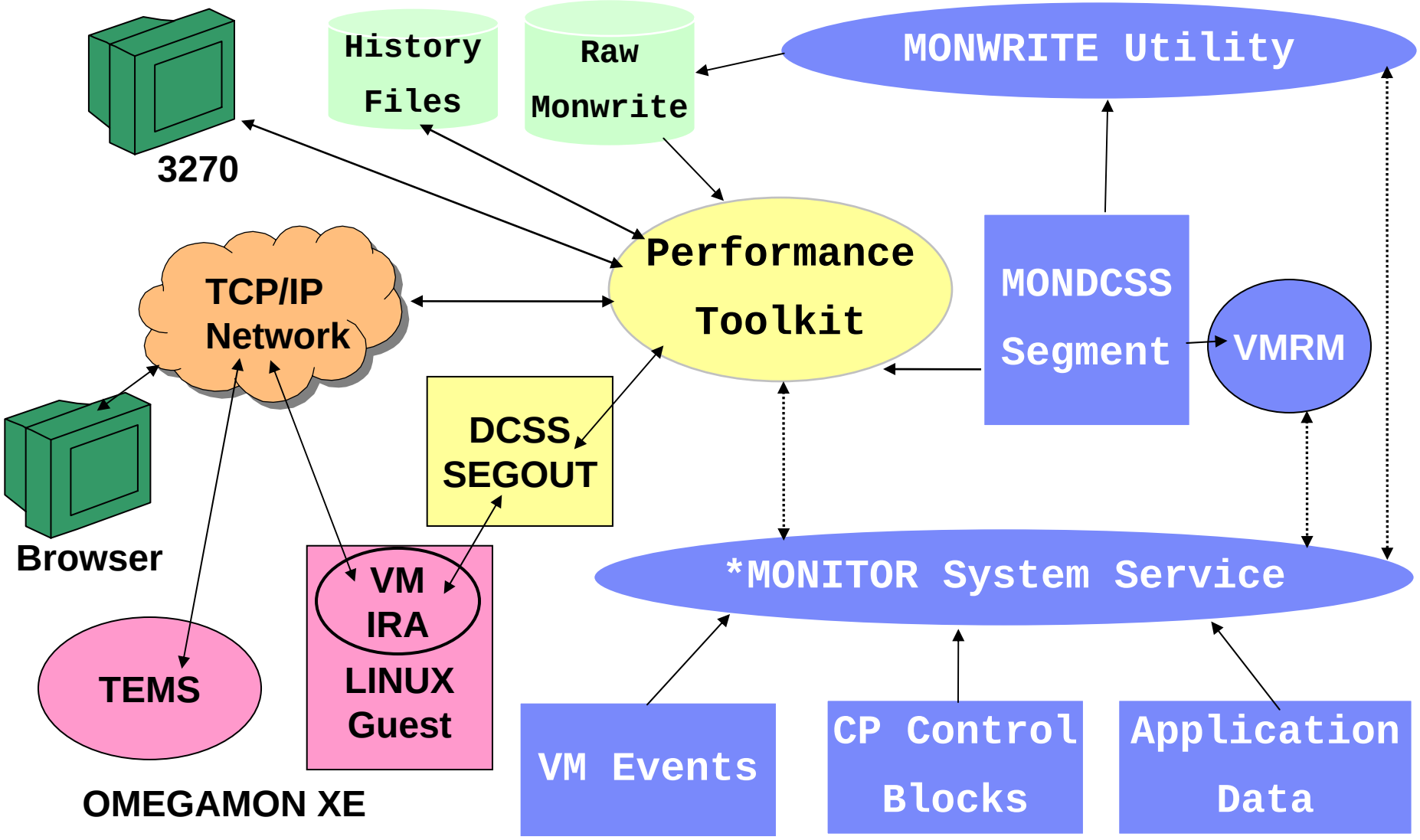
- **Look at FCX134 DSPACESH periodically**
  - Watch those PTRM0000 pages
    - Yellow at 262k
    - Red at 393k
- **Migrate to z/VM 5.3 if needed**
  - PTRM0000, …1, …2, …, …F (16 4-GB spaces)
  - PGMBKs can live above the real 2 GB bar
  - z/VM 5.3 is more proactive about removing PGMBKs
    - When last page goes, the PGMBK can go immediately
    - Thus on z/VM 5.3 you might see your resident PGMBK count decrease

Other Notes on z/VM Limits

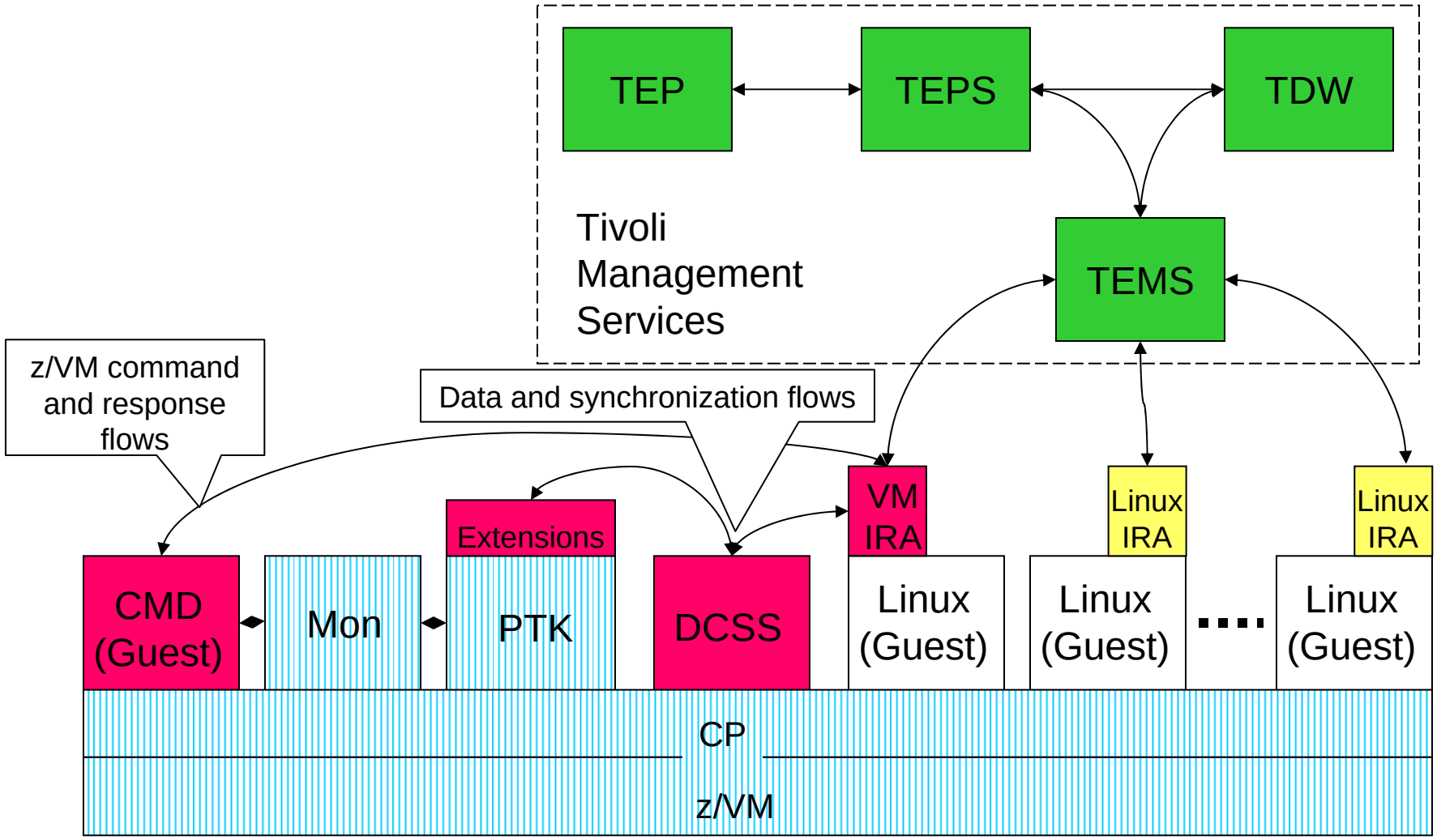
- **Sheer hardware:**
  - z/VM 5.2: 24 engines, 128 GB real
  - z/VM 5.3: 32 engines, 256 GB real
  - zSeries: 65,000 I/O devices
- **Workloads we've run in test have included:**
  - 54 engines
  - 408 GB real storage
  - 128 GB XSTORE
  - 240 1-GB Linux guests
  - 8 1-TB guests
- **Utilizations we routinely see in customer environments**
  - 85% to 95% CPU utilization without worry
  - Tens of thousands of pages per second without worry
- **Our limits tend to have two distinct shapes**
  - Performance drops off slowly with utilization (CPUs)
  - Performance drops off rapidly when wall is hit (storage)



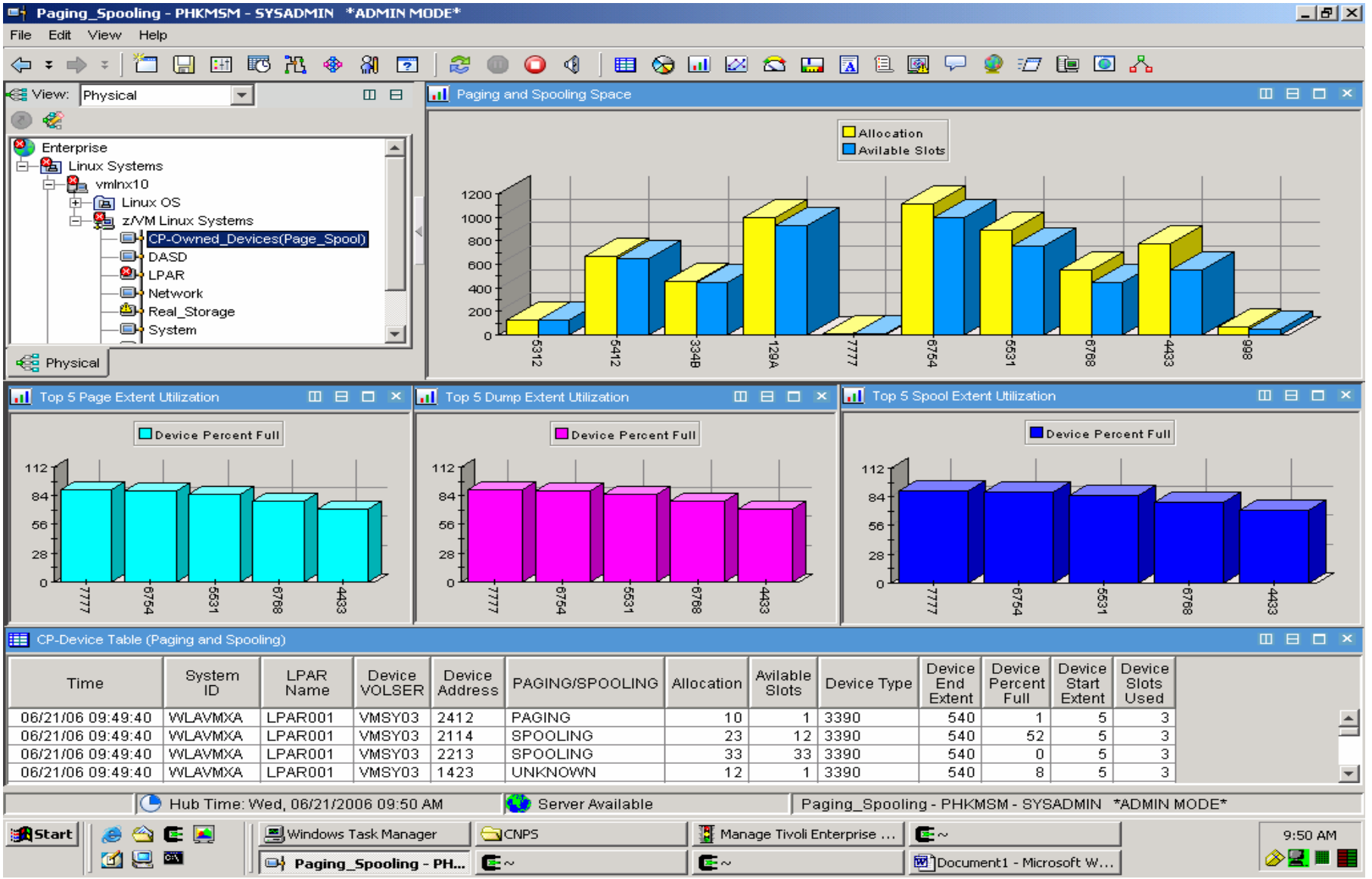
5,000 Foot View



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

More on OMEGAMON XE This Week

- **V96**
- **Tivoli OMEGAMON XE on z/VM and Linux**
- **Raymond Sun, IBM**
- **Thursday 1:00 PM, 206A**

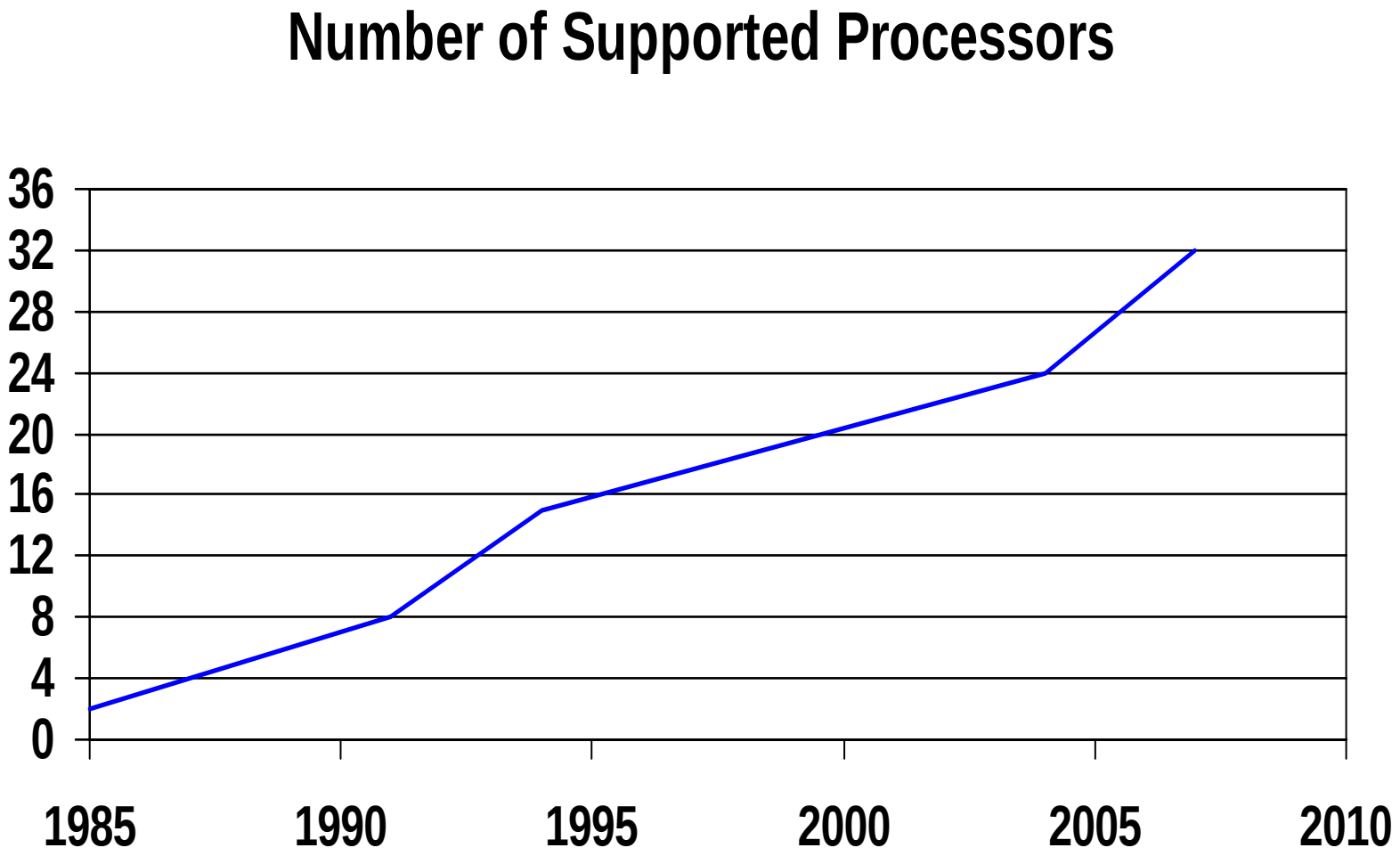
z/VM 5.3

- **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**

Workloads That Tend to Improve on z/VM 5.3

- **Systems with > 2 GB real and which do paging**
- **Systems that heavily use 6 or more real processors**
- **Workloads that make extensive use of Emulated FBA on SCSI (aka EDEVs)**

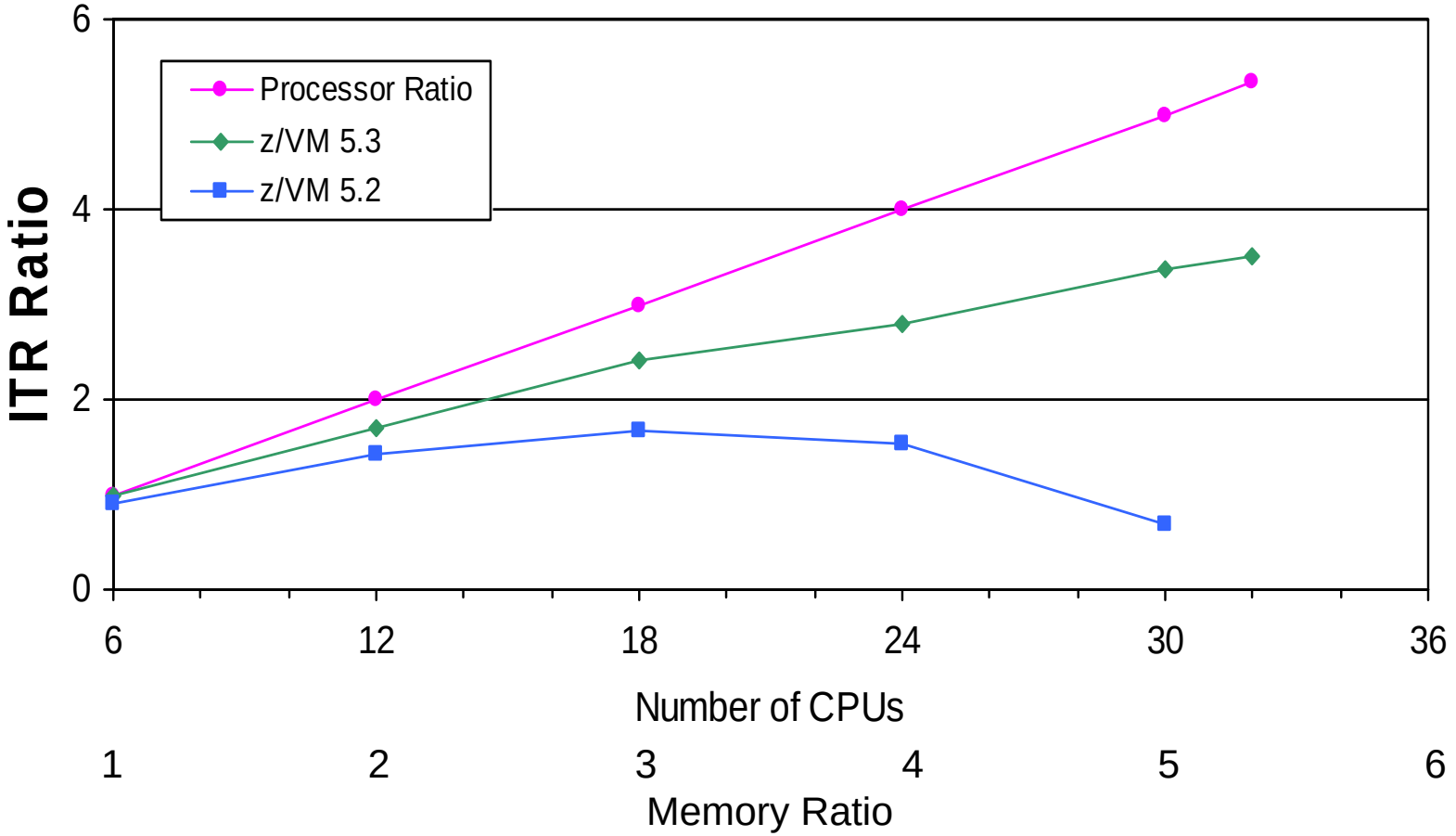
Processor Scaling



Support for 32 CPUs

- **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 exploiter 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 use only 1 processor

Large N-Way Effects on ITR Ratio



Linux Guests with Apache Webserving

Spin Lock Metrics, z/VM 5.2 to z/VM 5.3

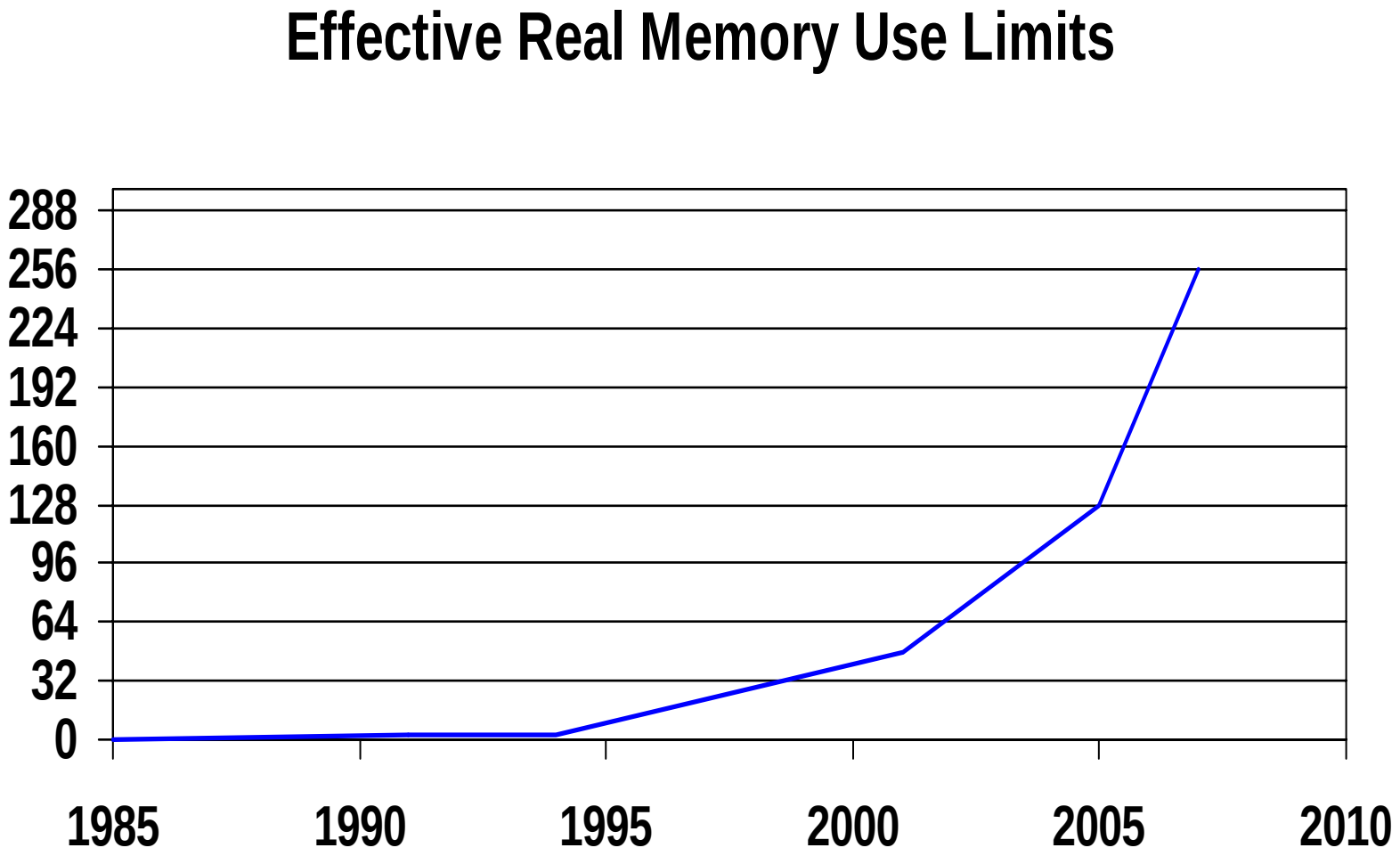
```
FCX239  Run 2006/03/28 15:19:36         PROCSUM                                                    Page   70
                                        Processor Performance Summary by Time
From 2006/03/26 13:18:17                                                                 D320A000
To   2006/03/26 13:28:17                                                                 CPU 2084-324   SN 96F5A
For    600 Secs 00:10:00                Result of D320A000 Run                           z/VM   V.5.2.0 SLU 0000
```

```
           <------- CPU --------> <------------------- Spin Lock Activity -------------------->
                 <--Ratio-->      <----- Total -----> <--- Scheduler ---> <- Timer Request -> <--SIE/sec-->
 Interval    Pct        Cap- On-  Locks Average   Pct  Locks Average   Pct  Locks Average   Pct     In- Inter-
 End Time   Busy   T/V  ture line  /sec    usec  Spin   /sec    usec  Spin   /sec    usec  Spin  struct   cept
 >>Mean>>    2.3 13.46 .7650  2.0   2.9    .714  .000    2.8    .717  .000     .0    .620  .000   1377   1370
 13:19:17    2.2 14.70 .7633  2.0   2.6    .685  .000    2.6    .685  .000     .0     ...     0   1299   1296
 13:20:17    2.4 14.79 .7667  2.0   2.6    .757  .000    2.6    .759  .000     .0    .422  .000   1400   1396
 13:21:17    2.3 14.78 .7653  2.0   3.0    .553  .000    3.0    .556  .000     .0    .094  .000   1375   1370
```

```
FCX265  Run 2007/09/14 12:36:15         LOCKLOG                                                    Page   56
                                        Spin Lock Log, by Time
From 2007/08/29 16:10:51                                                                 Y000284H
To   2007/08/29 16:20:51                                                                 CPU 2084-320   SN 37F2A
For    600 Secs 00:10:00                Result of Y000284H Run                           z/VM   V.5.3.0 SLU 0000
```

```
                    <------------------- 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    7.6    .392  .000    7.6    .392  .000     .0    .000  .000
 >>Mean>> RSACALLK     .0    .078  .000     .0    .078  .000     .0    .000  .000
 >>Mean>> RSAAVLLK    1.2    .245  .000    1.2    .245  .000     .0    .000  .000
 >>Mean>> SRMALOCK     .0    .000  .000     .0    .000  .000     .0    .000  .000
 >>Mean>> HCPTRQLK     .1    .642  .000     .1    .642  .000     .0    .000  .000
 >>Mean>> SRMSLOCK  617.6    .594  .012  617.6    .594  .012     .0    .862  .000
```

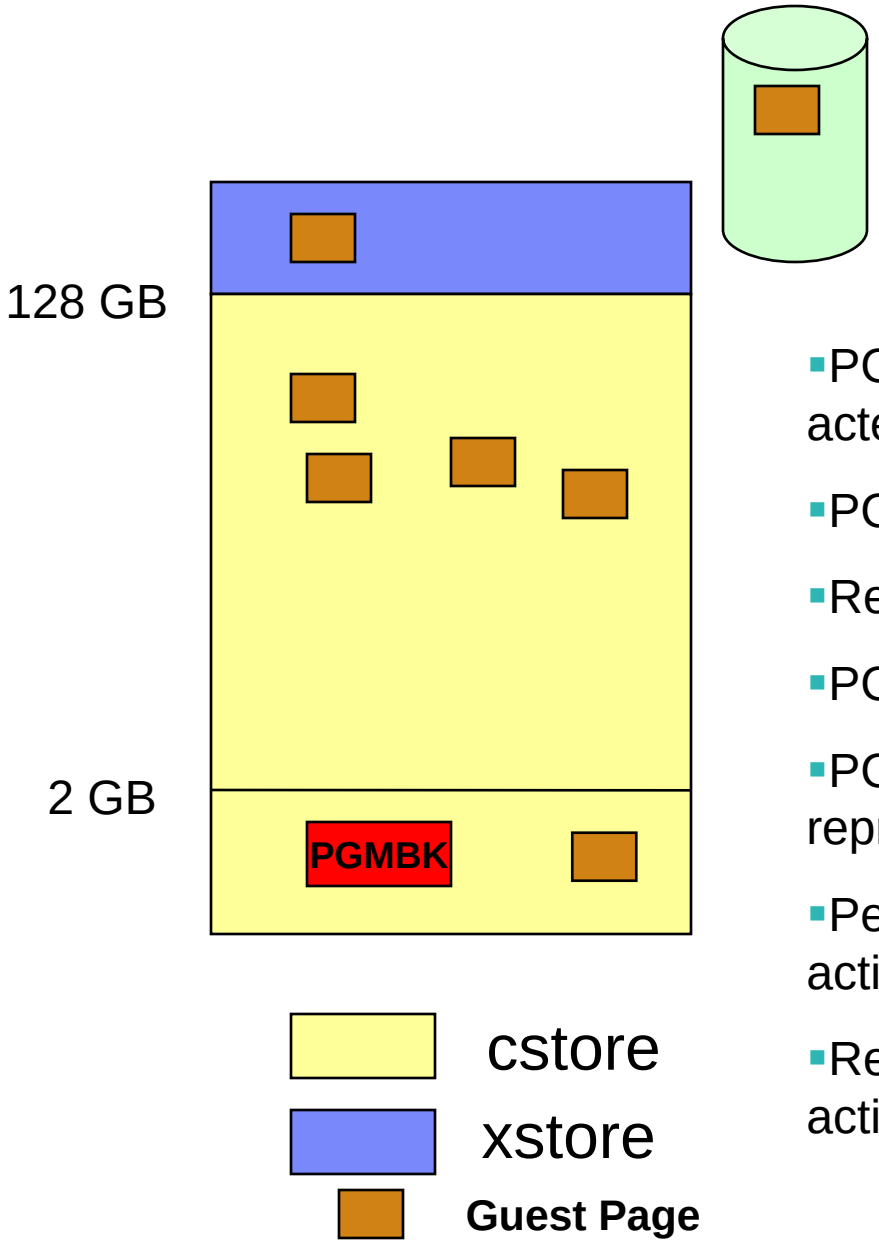
Memory Scaling



Support for 256 GB Real Storage

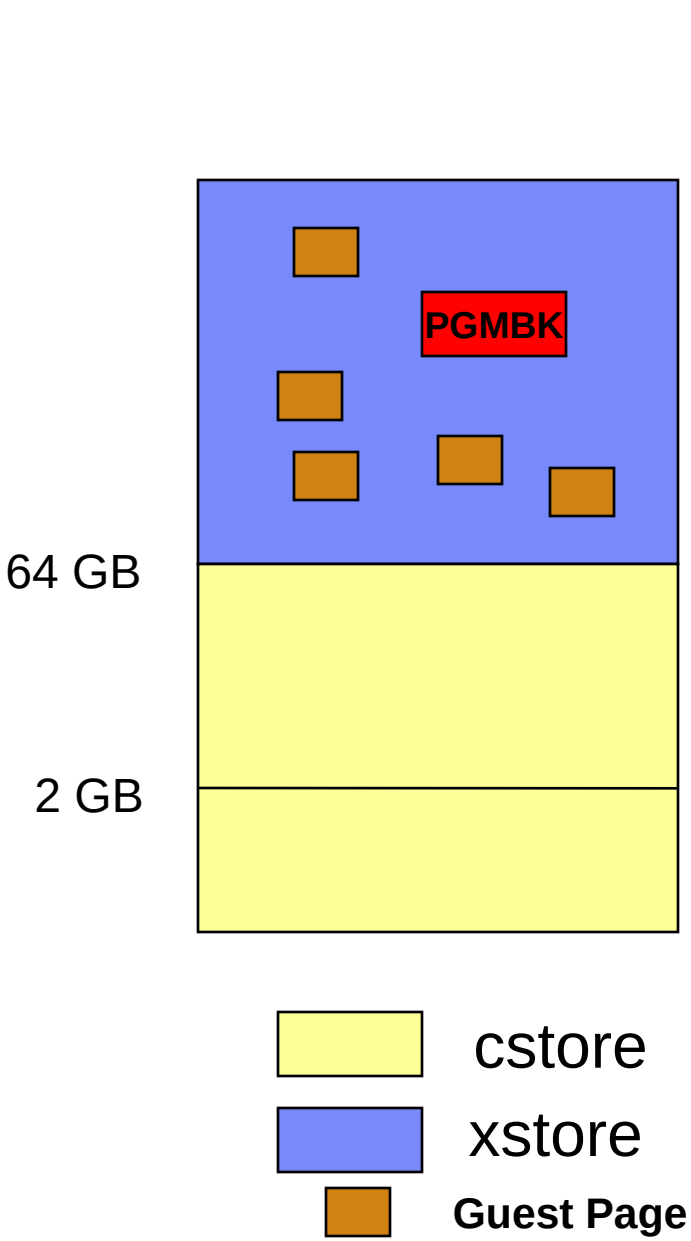
- **z/VM 5.3 improvements:**
  - PGMBKs can now reside above the 2 GB bar
    - Each PGMBK is 8 KB (2 contiguous frames)
  - Enhanced contiguous frame management
- **Also seeing improvements in smaller configurations that are memory-constrained**
- **Be careful with (guest) memory terminology**
  - Try 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

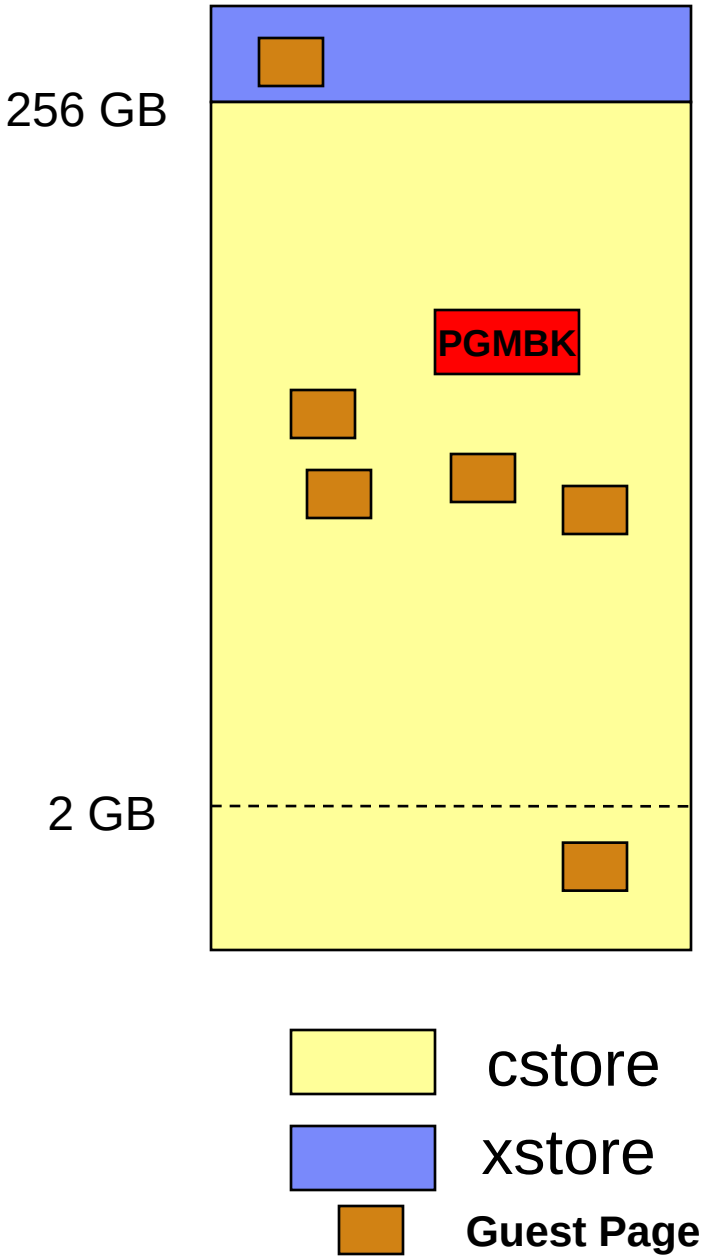


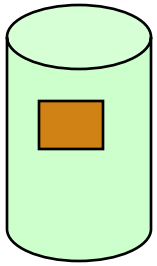
- PGMBK is required for any guest page that is acted on
- PGMBK is 8 KB (2 contiguous frames)
- Resident PGMBK located below 2 GB
- PGMBKs are pageable
- PGMBK resident if any guest pages it represents are resident
- Perfect world: ~256 GB of virtual memory actively being referenced
- Realistic world: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

z/VM 5.2.0 Larger Xstore:Cstore Ratio



- More expanded storage can increase the likelihood that guest pages are moved out of central storage.
- This lets CP move PGMBKs too.





z/VM 5.3.0 PGMBKs above 2 GB Bar

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

Performance Toolkit DSPACESH Report

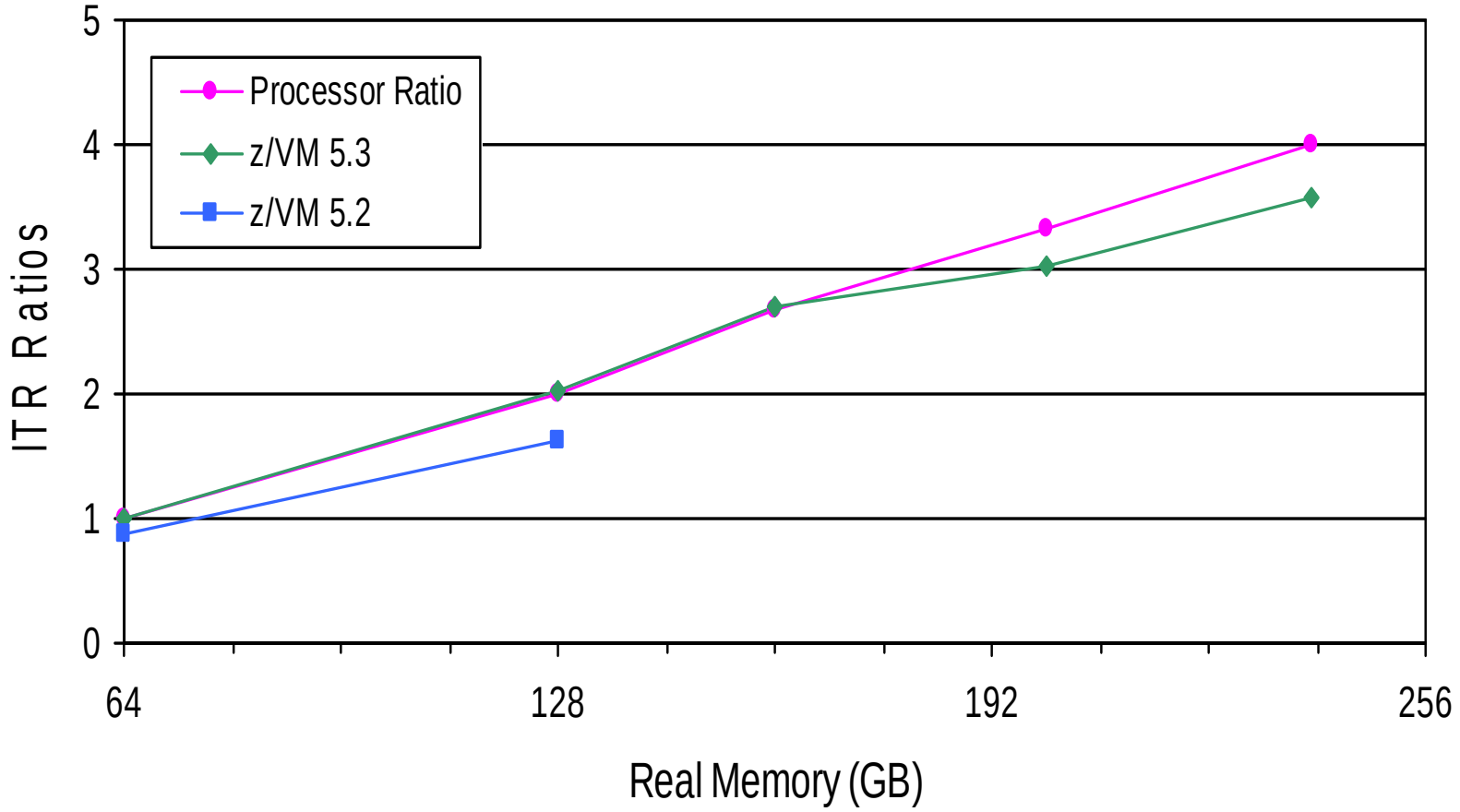
```
 FCX134      CPU 2094  SER 19B9E  Interval 13:04:01 - 13:09:01    GDLVM7
                              <------------------Number of Pages----------------->
Owning                             <--Resid--> <-Locked--> <-Aliases->
Userid    Data Space Name   Total Resid R<2GB  Lock L<2GB Count Lockd 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 are dependent on workload and configuration**
  - 256 GB real memory
  - 2:1 or 3:1 V:R ratio “before things get interesting” suggests about 768 GB virtual as an operational threshold
  - Completely-full paging space (optimal when <50% full)
    - 11.2 TB for ECKD
    - 15.9 TB for Emulated FBA on FCP SCSI
  - Virtual machine size (hardware-dependent)
    - 1 TB on z9
  - 8 TB of “addressable” virtual machine memory
    - Limit of page tables

Scaling Memory Results - Apache Webserving



Improvement in Memory-Constrained Environments

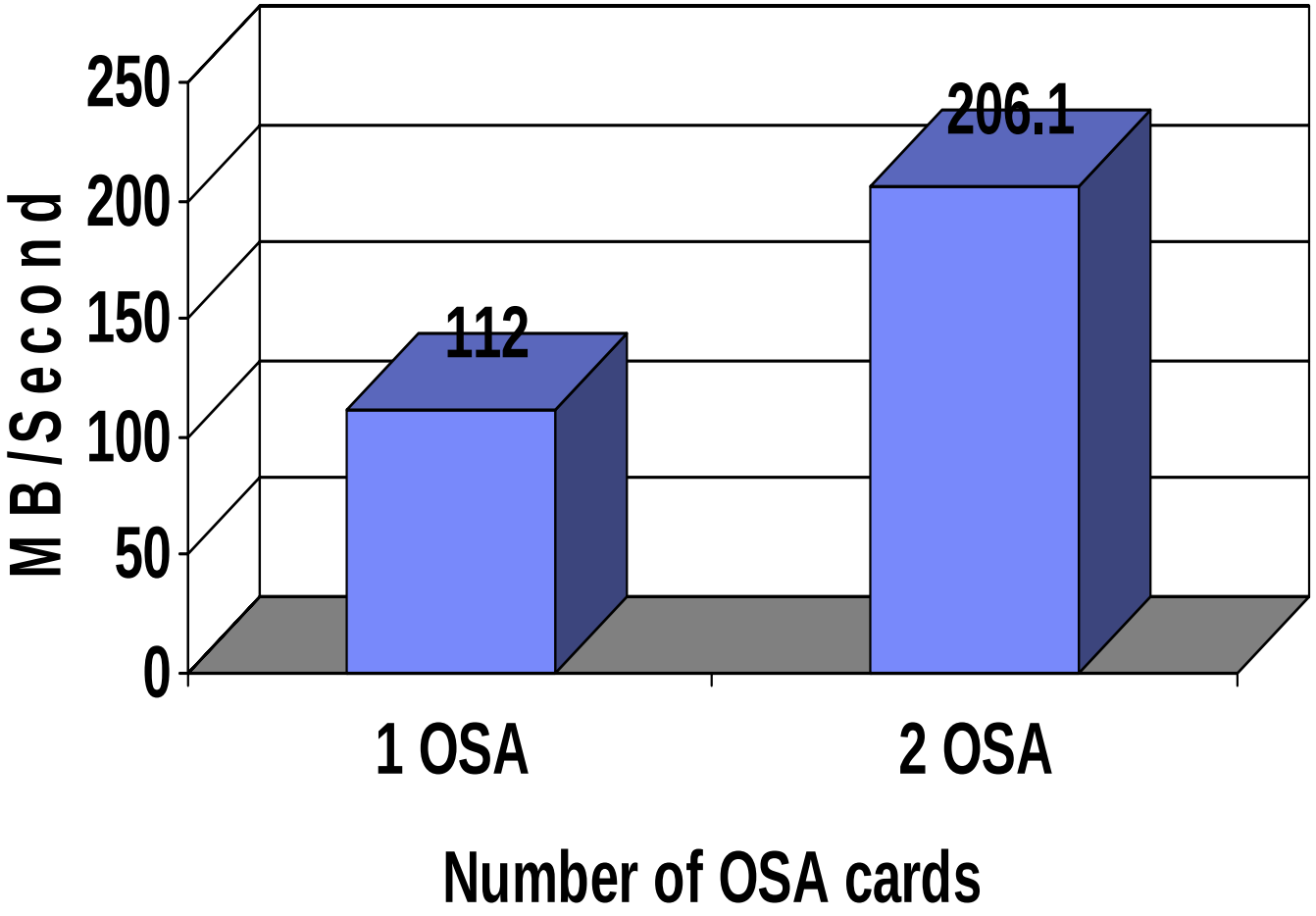
Scenario	Contention (Page Reqs per CPU-second)	Delta Thruput	Delta Total CPU/Tx
3G/4G 2084 3-way	2159	+10.3%	-9.5%
64G/2G 2094 3-way	0	+1.0%	-0.9%
64G/2G 2094 3-way	352	+15.5%	-12.7%
128G/2G 2094 6-way	291	+21.6%	-19.4%

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 TCP 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)**
  - Definitely apply VM64199 if you are 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 DS8000 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**
  - 41% reduction in elapsed time
  - Better than 97% reduction in CP CPU time
- **Additional CP pathlength reductions**
  - 5% to 15% reductions in CP CPU time, depending on workload
- **CP paging to SCSI volumes now bypasses the FBA emulation, reducing processor resource requirements**
  - 24% reduction in CP CPU per page moved, depending on workload

Monitoring Enhancements

- **Lots of new fields in Monitor for new function:**
  - Specialty engines
  - Scheduler changes
  - PAV and HyperPAV support (MRIODDEV extensions)
  - Memory management
- **New monitor domain for virtual network devices**
- **Additional flexibility in MONWRITE utility for starting and stopping**
- **Various changes in Performance Toolkit for VM**

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/
  - Will include updated sections in near future on:
    - Diagnoses x'44' and x'9C'
    - Cooperative Memory Management and Collaborative Memory Management Assist
- **Learn more about z/VM**
  - http://www.vm.ibm.com/events/