Siegfried Langer – Business Development Manager z/VSE & Linux on System z
April 16, 2011

Performance in a Virtualized Environment

Abstract

Performance in a virtualized environment

Performance tuning is an art. Typically there are no fixed rules to its optimization as many factors are influencing system throughput and resource consumption, as well as service level requirements. In a virtualized environment this becomes even more complex as virtualized systems may compete with the hypervisor for resources. The presentation will cover general performance considerations in a virtualized environment with focus on Linux and z/VM.

Agenda

- Definition of performance
- Performance tuning (what is different in a virtualized environment?)
- z/VM storage hierarchy
- Performance guidelines in a z/VM – Linux on System z environment
  - Paging
  - Memory
  - Processor
  - Disks
- Network co-location
- Performance monitoring
- Information sources

Definition of Performance

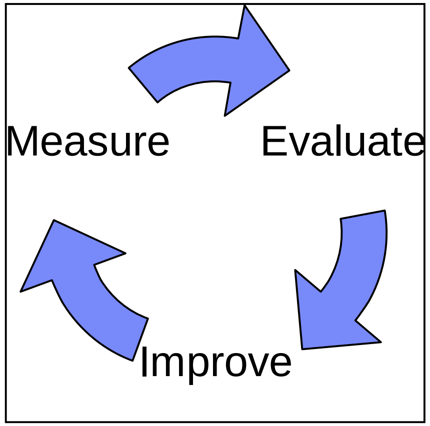
Performance tuning is the improvement of system performance.

- Response time
- Batch elapsed time
- Throughput
- Utilization
- Users supported
- Internal Throughput Rate (ITR)
- External Throughput Rate (ETR)
- Resource consumed per unit of work done

Performance tuning

Systematic tuning follows these steps:

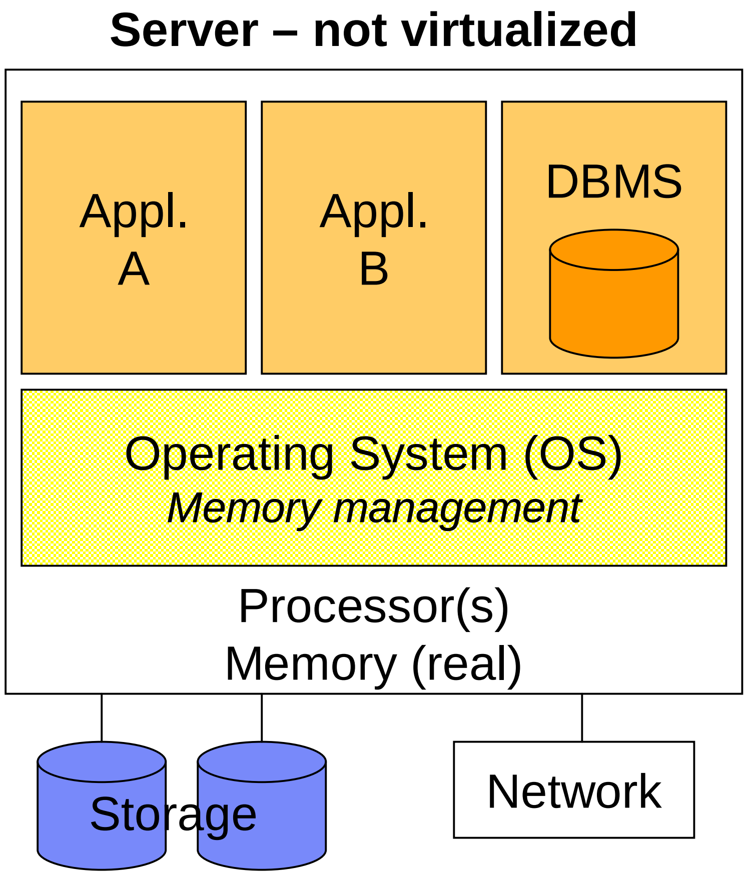
- Assess the problem and establish numeric values that categorize acceptable behavior.
- Measure the performance of the system before modification.
- Identify the part of the system that is critical for improving the performance. This is called the bottleneck.
- Modify the part of the system to remove the bottleneck.
- Measure the performance of the system after modification.



Typically, removing a bottleneck will reveal a new bottleneck in another area!

Source: Wikipedia

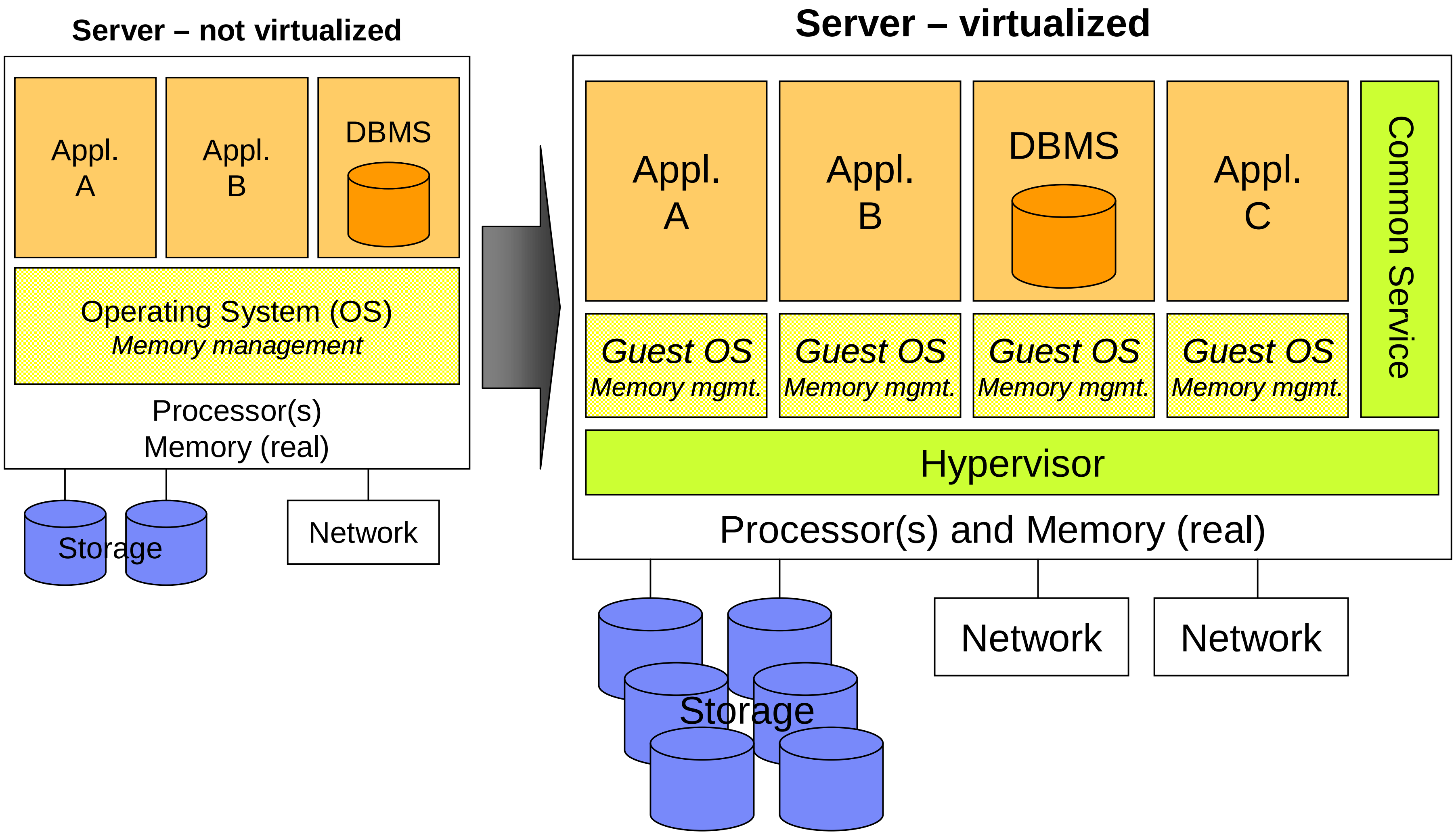
Tuning consideration



- Storage layout
  - Striping
- Memory management
  - Memory layout (heap, etc.)
  - Data in memory
  - Virtual memory
- Priority settings
- Buffers
- Application tuning/optimization
- Database Management System (DBMS)
  - Database physical design
  - DB logical design
  - Buffers/cache size
- Network settings
  - MTU size
  - Buffers
-

Tune/optimize for most critical application(s)

Tuning consideration



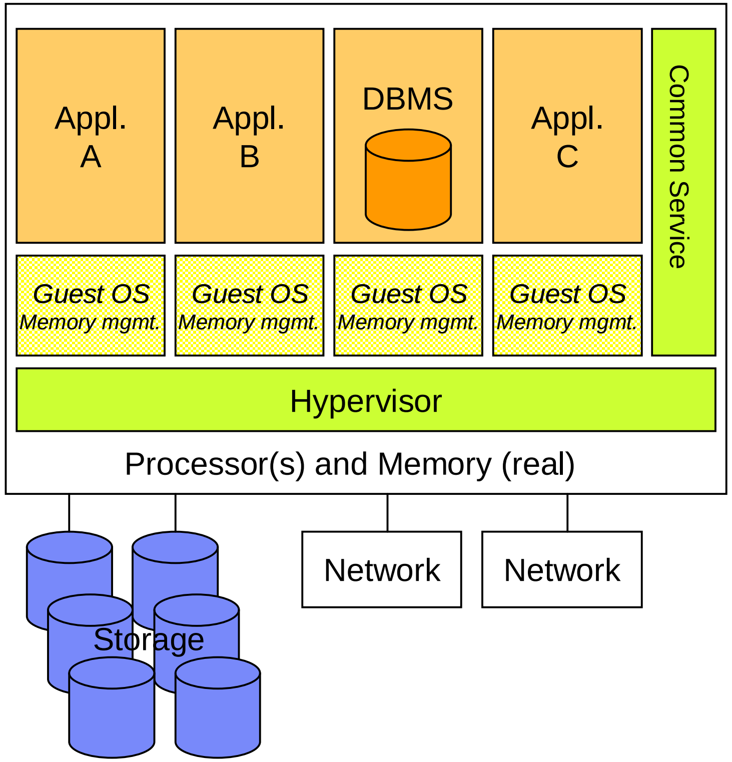
Tuning consideration

- Storage layout
  - Striping
- Memory management
  - Memory layout (heap, etc.)
  - Data in memory
  - Virtual memory
- Priority settings
- Buffers
- Application tuning/optimization
- Database Management System (DBMS)
  - Database physical design
  - DB logical design
  - Buffers/cache size
- Network settings
  - MTU size
  - Buffers
-

Plus:

- Resource allocation (processors, memory, I/O)
- Multi-level memory management
- Internal network
- Virtual I/O
- Common services (e.g., security services)
- more users
-

Server – virtualized



Tune/optimize for balanced system

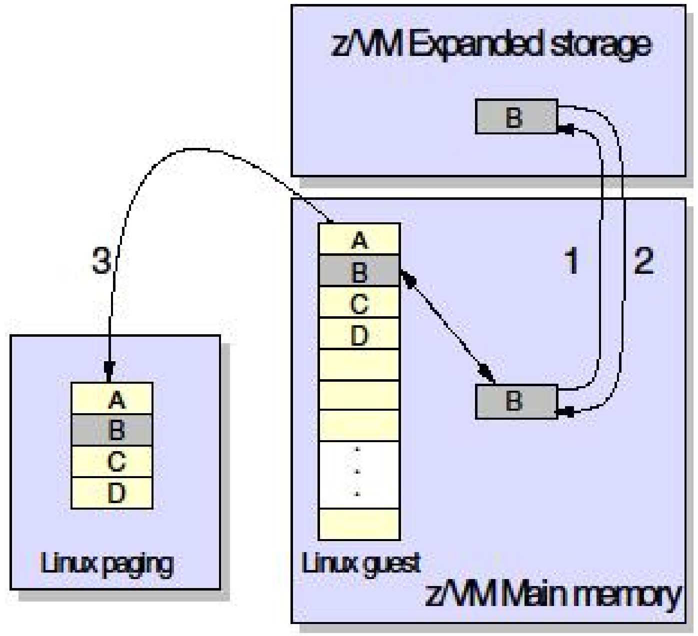
z/VM storage hierarchy

Usually, the term ***storage*** is used by z/VM, while ***memory*** is used by Linux.

- **Main storage**
  - Directly addressable and fast-accessible by user programs
  - Maximum size of main storage is restricted by the amount of physical storage.
- **Expanded storage**
  - Expanded storage also exists in physical storage, but is addressable only as entire pages.
  - Physical storage allocated as expanded storage reduces the amount for main storage.
  - Expanded storage is optional, and its size is configurable.
  - Expanded storage acts as a fast paging device used by z/VM.
- **Paging space**
  - Paging space resides on DASD.
  - When paging demands exceed the capacity of expanded storage, z/VM uses paging space.

Double paging effect

1. z/VM pages out inactive page
2. Page-out attempt from Linux guest moves page into main memory again
3. Linux completes its page-out attempt and moves page B to swap device



Solution:

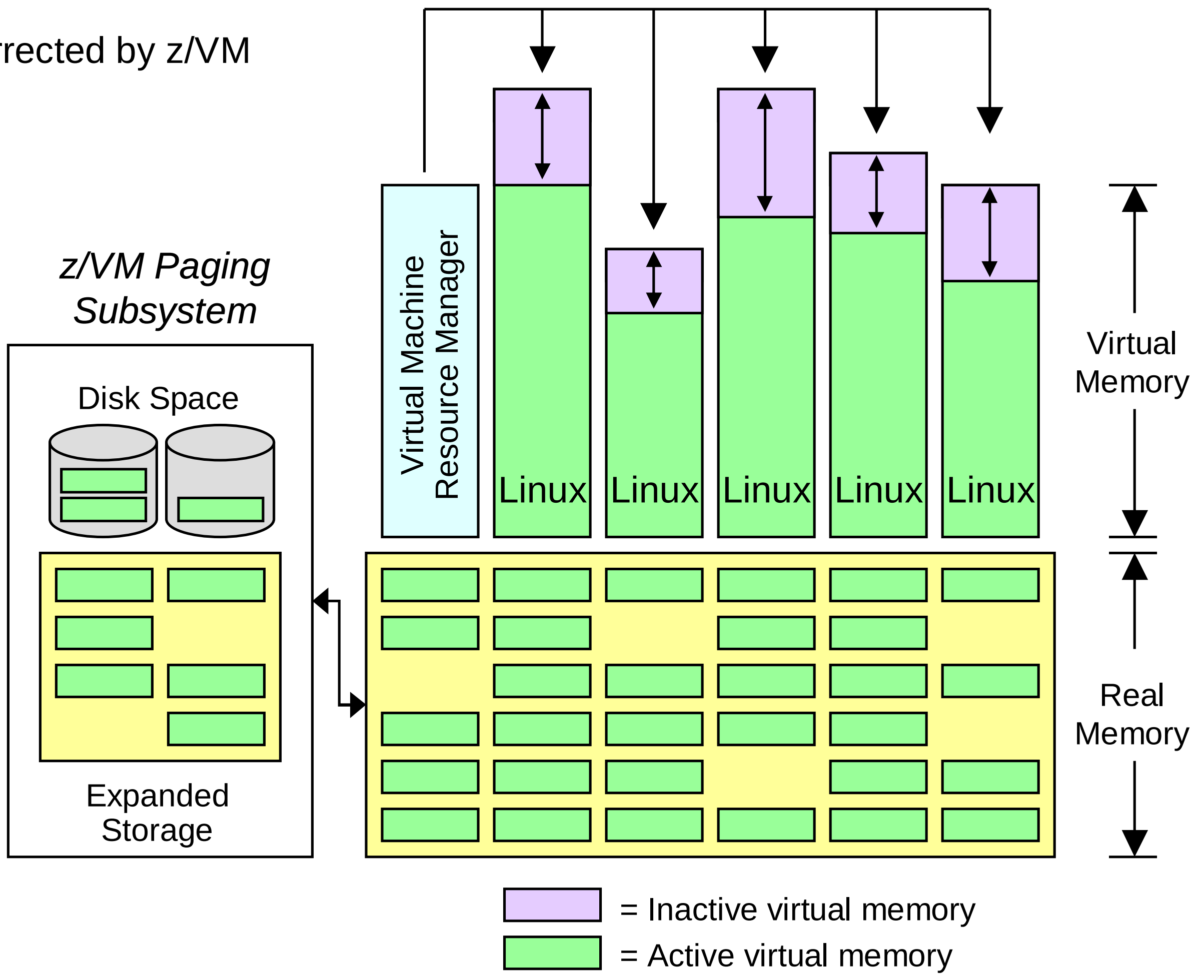
- ***Ensure that one party does not attempt to page!***
- Make the Linux guest virtual machine size small enough for z/VM to keep in main storage.
- Make the virtual machine size large enough that Linux does not attempt to swap.
- Cooperative Memory Management (CMM).
  - Storage usage information is passed from Linux to z/VM.
- Collaborative Memory Management Assist (CMMA).
  - Collaborative memory management assist is completely controlled by the Linux guest

Extreme Virtualization with Linux on z/VM

VMRM Cooperative Memory Management (VMRM-CMM)

- Problem scenario: virtual memory utilization far exceeds real memory availability
- Solution: real memory constraint corrected by z/VM *Virtual Machine Resource Manager*
- Linux images signaled to reduce virtual memory consumption
- Demand on real memory and z/VM paging subsystem is reduced
- Helps improve overall system performance and guest image throughput

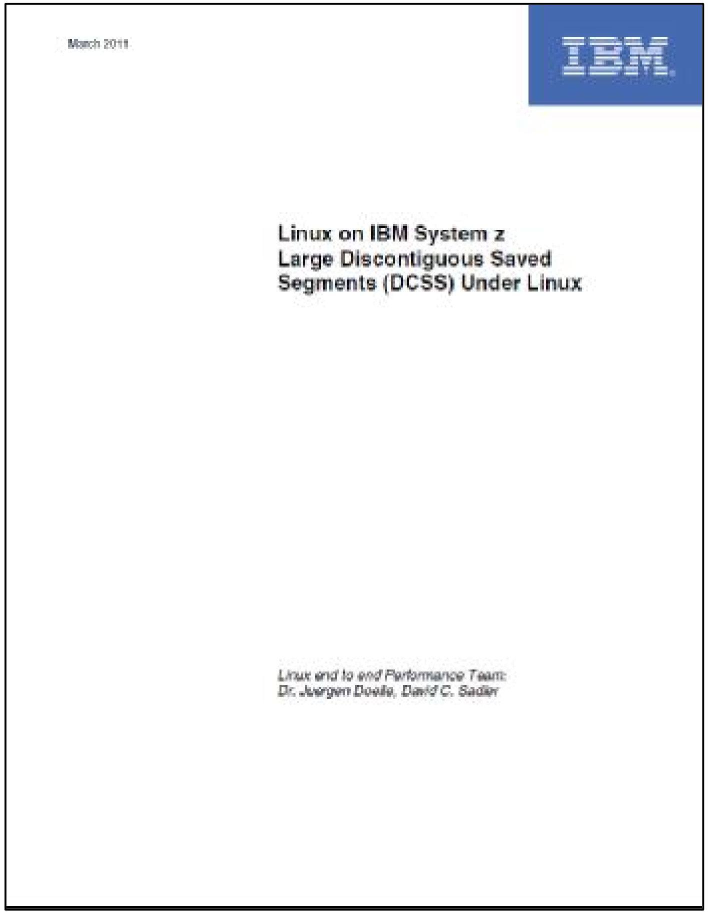
Lab tests have shown up to 50% more throughput using CMM with z/VM 5.3



New white paper - DCSS under Linux on System z

http://www.ibm.com/developerworks/linux/linux390/perf/tuning_vm.html#dcss

- This document provides results for tests run using large Discontiguous Saved Segments under Linux®.
- This paper focuses on three areas of application for a large DCSS: sharing code, sharing read only data, and using a DCSS as a swap device.
- A *saved segment* is a special feature of z/VM that provides a range of virtual storage pages, which are defined to hold data or reentrant code (programs). The administrator can save code or data in saved segments, assign them a name, and dynamically attach or detach them from multiple virtual machines.


March 2011

Linux on IBM System z
Large Discontiguous Saved Segments (DCSS) Under Linux

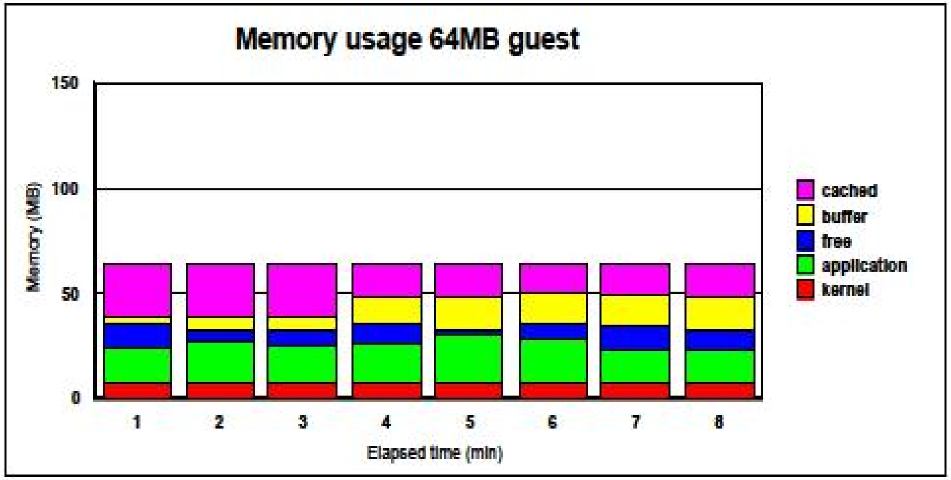
Linux end to end Performance Team: Dr. Juergen Doelle, David C. Sadler

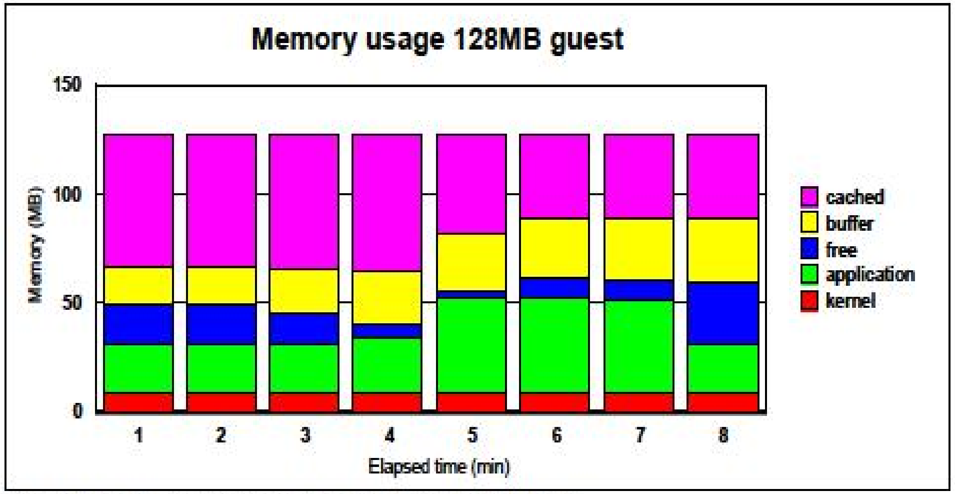
Aggressive caching within Linux

Linux manages its memory without regard to the fact that it is running in a virtual machine.

- § The Linux kernel attempts to load as much information (applications, kernel, cache) into its perceived memory resources as possible and caches it there.
- § Comparing the two Linux guests, we see a similar memory usage pattern: In both cases, additional application memory is obtained at the expense of buffer and cache memory.
- § Reducing the virtual machine size by 50% reduced average caching by 60%.

Note: Although the 64 MB guest required half the amount of memory, no appreciable effect on server response time was noted.



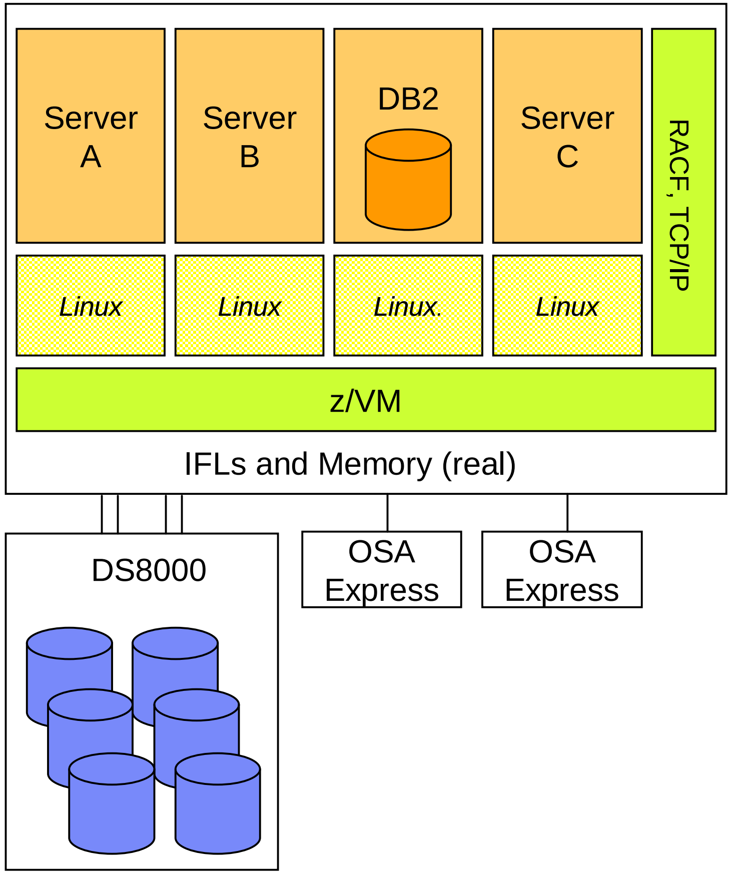


Performance guidelines - Paging

§ Paging
- To determine the smallest memory footprint required, decrease the size of the Linux virtual machine to the point where swapping begins to occur under normal load conditions.
- At that point, slightly increase the virtual machine size to account for some additional load.

§ The general rule does not apply to some sort of servers that have special memory needs.
- Database servers
  - Database servers maintain buffer pools to prevent excessive I/O to disk. These buffer pools should not be downsized. Otherwise, performance suffers.
- Servers running Java workload (that is, WebSphere Application Server)
  - An amount of memory is needed to host the Java heap.
  - Too small heap size degrades the performance even if no swapping occurs.

z/VM and Linux on System z



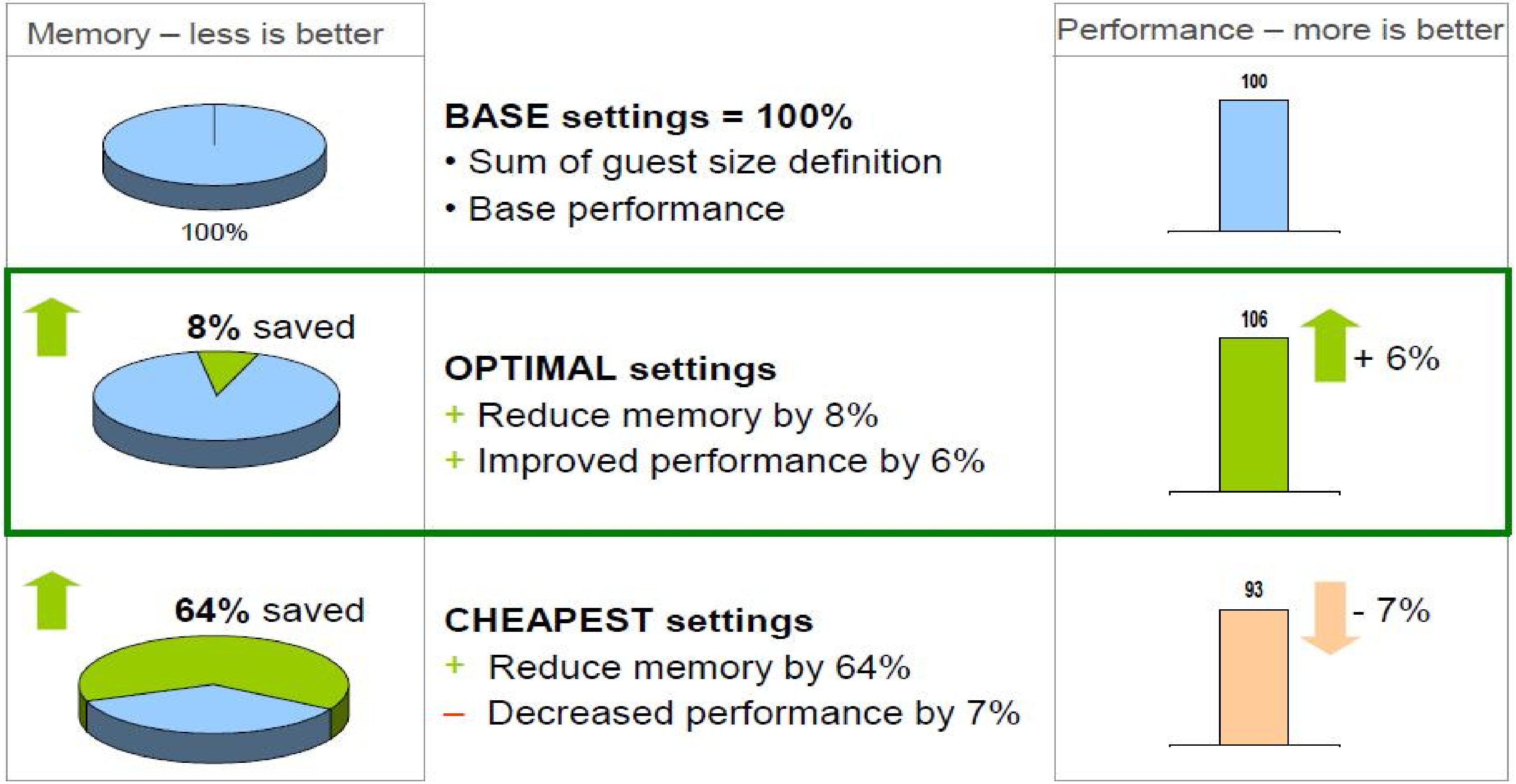
Performance guidelines – Memory & Paging

- Virtual:Real ratio should be ≤ 3:1 or make sure you have robust paging system
  - To avoid any performance impact for production workloads, you may need to keep ratio to 1:1
  - 1.6:1 might be a good starting point/compromise for many loads
- Use SET RESERVE instead of LOCK to keep users pages in memory
- Define some processor storage as *expanded storage* to provide paging hierarchy
- Exploit shared memory where appropriate
- Size guests appropriately

- Multiple volumes and multiple paths to paging DASD
- Paging volumes should be of the same geometry and performance characteristics
- Paging to FCP SCSI may offer higher paging bandwidth with higher processor requirements
- In a RAID environment, enable cache to mitigate write penalty

Test results

- Running a mix of server types as Linux guests on z/VM
  - LPAR with 28 GB central storage + 2 GB expanded storage
  - Guest workloads: WAS (13.5 GB), DB2 (12.0 GB), Tivoli Directory Server (1.5 GB), idling guest (1.0 GB)
- Leave guest size fixed – decrease LPAR size in predefined steps to scale level of memory over-commitment



Results & recommendations (1 of 2)

- **Virtual memory = Physical memory**
  - Does not provide the best performance (at least not for large LPARs, e.g. 28GB).

- **Optimal memory setting: No z/VM paging !**
  - See PerfKit Panel FXC113 User Paging Activity and Storage Utilization and
  - Panel FCX254 Available List Management

- **Recommendations** (minimum memory requirements):
  - **WebSphere Application Server**: Sum of all active Java heaps
  - **DB2:** Sum of MAX_PARTITION_MEM as reported from:
    "SELECT * FROM TABLE (SYSPROC.ADMIN_GET_DBP_MEM_USAGE()) AS T".
    Value of PEAK_PARTITION_MEM might be used, when highest allocation is captured.

- **Linux Page Cache**:
  - Sum of read/write throughput,
  - e.g. 20 MB/sec read and 10 MB/sec write throughput require 30 MB/sec pages
    → are ignored in our case in regard to the sizes contributed from WebSphere and DB2

- **Idling guests** (no kind of server started!): Can be ignored

Results & recommendations (2 of 2)

- **The minimum memory size defines the lower limit, do not cross!**
- **Be aware of the dynamic of a virtualized environment**
  - New guests are easily created,
  - Java heaps and database pools might be increased without notifying the System z administrator
  - **Monitor paging activity of your system!**
- **Other workload types might follow similar considerations**
- For more information see
  - Chapter 9. Memory overcommitment in

    Tivoli Provisioning Manager Version 7.1.1.1: Sizing and Capacity Planning
    http://public.dhe.ibm.com/software/dw/linux390/perf/ZSW03168USEN.PDF

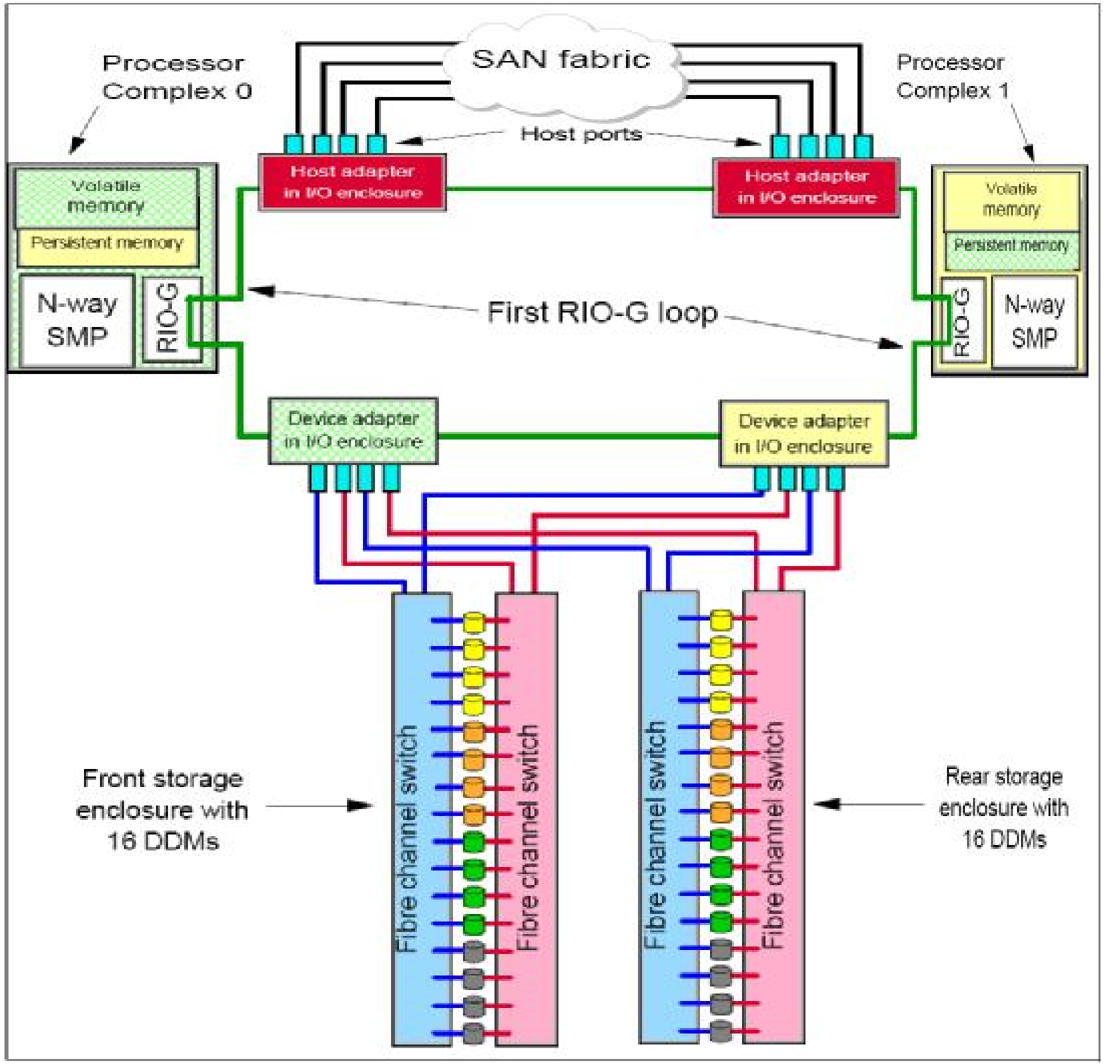
Performance guidelines - Processor

- Dedicated processors – mostly political
  - Absolute share can be almost as effective
  - A virtual machine should have all dedicated or all shared processors
  - Gets wait state assist and 500 ms minor slice time
- Share settings
  - Use absolute if you can judge percentage of resources required
  - Use relative if difficult to judge and if slower share as system load increases is acceptable
  - Do not use LIMITHARD settings unnecessarily
- Do not define more virtual processors than are needed
- Small minor time slice keeps processor reactive

Performance guidelines – Disks

- Don’t treat a storage server as a black box, understand its structure.
- Several conveniently selected disks instead of one single disk can speed up the sequential read/write performance to more than a triple. Use the logical volume manager to set up the disks.
- Avoid using subsequent disk addresses in a storage server (e.g., the addresses 5100, 5101, 5102, … in an IBM Storage Server), because
  - they use the same rank
  - they use the same device adapter.
- If you ask for 16 disks and your system administrator gives you addresses 5100-510F
  - from a performance perspective this is close to the worst case

DS8000 Architecture



- **Structure** is complex
  - disks are connected via two internal FCP switches for higher bandwidth
- DS8000 is still divided into two parts
  - Caches are organized per server
- One **device adapter pair** addresses 4 array sites
- One **array site** is build from 8 disks
  - disks are distributed over the front and rear storage enclosures
- One **RAID array** is defined using one array site
- One **rank** is built using one RAID array
- Ranks are assigned to an **extent pool**
- Extent pools are assigned to **one of the servers**
  - this assigns also the caches

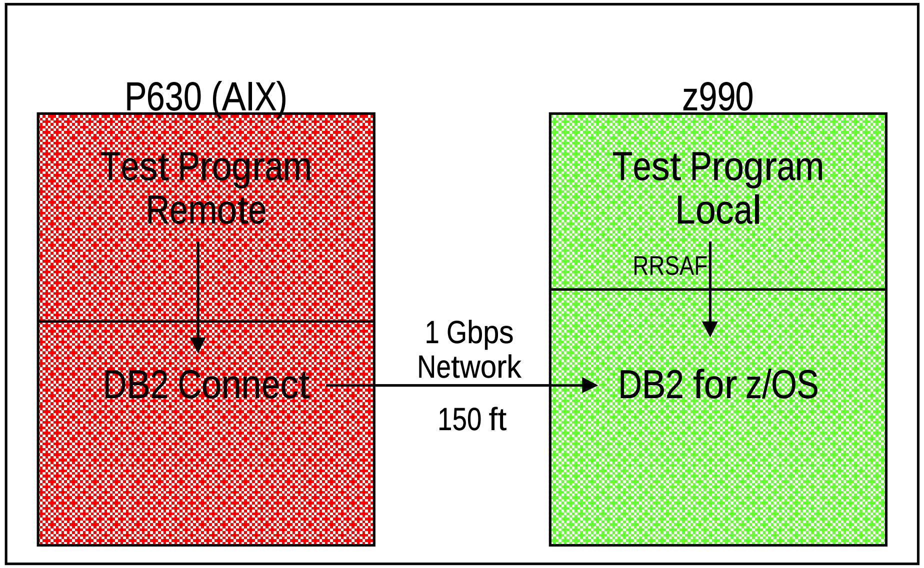
I/O processing characteristics

- FICON/ECKD:
  - 1:1 mapping host subchannel:dasd
  - Serialization of I/Os per subchannel
  - I/O request queue in Linux
  - Disk blocks are 4KB
  - High availability by FICON path groups
  - Load balancing by FICON path groups and Parallel Access Volumes
  -
- FCP/SCSI
  - Several I/Os can be issued against a LUN immediately
  - Queuing in the FICON Express card and/or in the storage server
  - Additional I/O request queue in Linux
  - Disk blocks are 512 bytes
  - High availability by Linux multipathing, type failover
  - Load balancing by Linux multipathing, type multibus

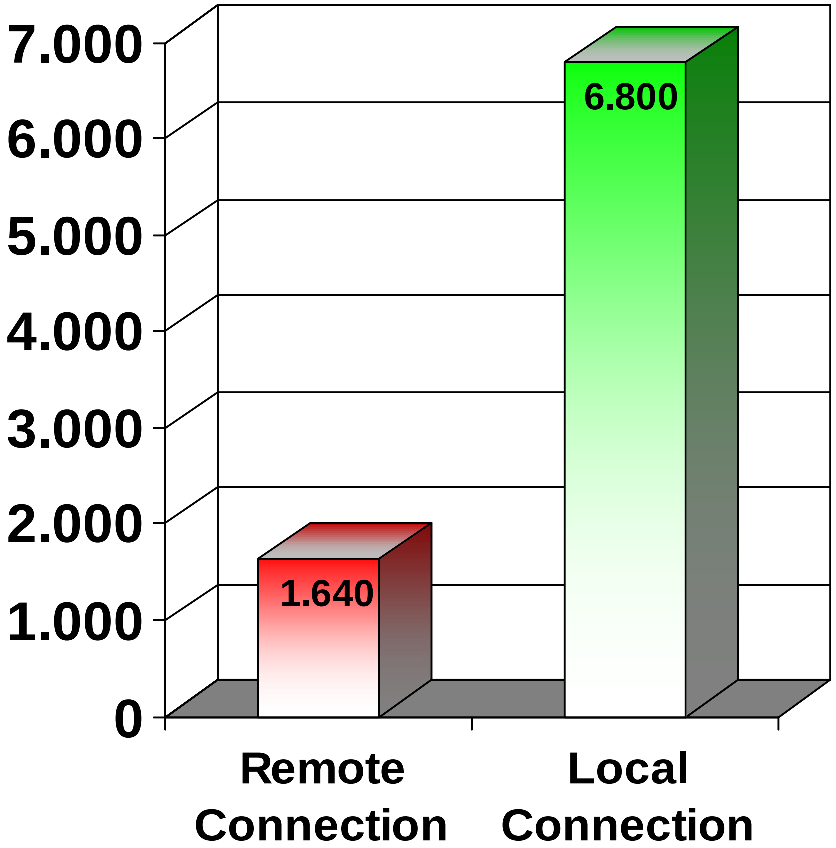
Co-located applications maximize performance

Study shows benefits of local vs. remote connection to data

Test Configuration



Results: SQL Statements / Second

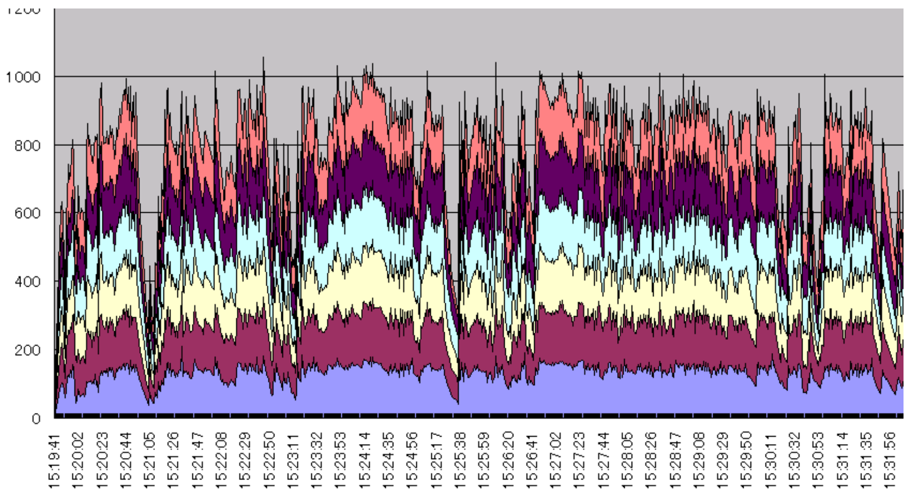


Why the big difference in SQL throughput?

- Elimination of network latency incurred by remote database connections increased SQL throughput **4x!**
- Hipersockets provide this benefit for consolidated applications on zLinux

IBM Study: "Local versus Remote Database Access: A Performance Test", 2005
http://publib-b.boulder.ibm.com/abstracts/redp4113.html

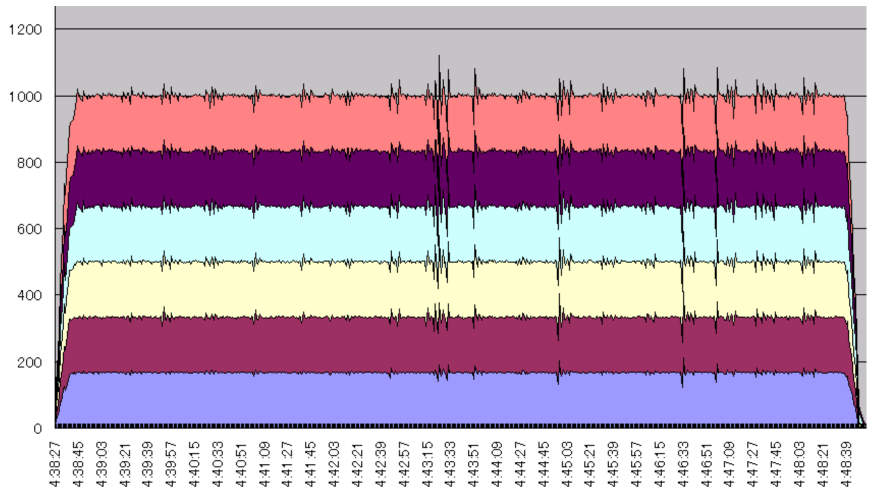
Workload Enablement
HiperSocket versus Blade-based network interconnect



- *Oracle RAC nodes exchange lock information for the shared database*
- *Given high transaction stress, this architecture forces TCP/IP to become bottleneck – exemplified in the Blade Center benchmark.*
- *HiperSockets provides relief to this architecture bottleneck, resulting in stable response time and throughput – making System z the server of choice for high transaction Oracle DBs.*

(Test Case: 1000 Inserts/Sec high workload processing case)

Linux on z（dedicate 1CP+2GB)
& HiperSockets



Networking – HiperSockets or OSA

- HiperSockets work in synchronous mode
- Communications via OSA works asynchronously

- In general, HiperSockets will be the best choice for cross-LPAR communications

- If the CPU speed of two LPAR environments is very different, use OSA
  - Sub Capacity CP communicating with IFL (large difference in MIPS between processors)
    e.g. z/VSE running on Sub Capacity CP and zLinux running on (uncapped) IFL
  - Capping (limiting the MIPS consumption in LPAR) is not affected

z/VM Performance Toolkit

§ The z/VM Performance Toolkit is a z/VM licensed product

```
FCX124        Performance Screen Selection  (FL540 25Feb08)    Perf. Monitor

General System Data          I/O Data                     History Data (by Time)
1. CPU load and trans.       11. Channel load             31. Graphics selection
2. Storage utilization       12. Control units            32. History data files*
3. Reserved                  13. I/O device load*         33. Benchmark displays*
4. Priv. operations          14. CP owned disks*          34. Correlation coeff.
5. System counters           15. Cache extend. func.*     35. System summary*
6. CP IUCV services          16. DASD I/O assist          36. Auxiliary storage
7. SPOOL file display*       17. DASD seek distance*      37. CP communications*
8. LPAR data                 18. I/O prior. queueing*     38. DASD load
9. Shared segments           19. I/O configuration        39. Minidisk cache*
A. Shared data spaces        1A. I/O config. changes      3A. Storage mgmt. data*
B. Virt. disks in stor.                                   3B. Proc. load & config*
C. Transact. statistics      User Data                    3C. Logical part. load
D. Monitor data              21. User resource usage*     3D. Response time (all)*
E. Monitor settings          22. User paging load*        3E. RSK data menu*
F. System settings           23. User wait states*        3F. Scheduler queues
G. System configuration      24. User response time*      3G. Scheduler data
H. VM Resource Manager       25. Resources/transact.*     3H. SFS/BFS logs menu*
                             26. User communication*      3I. System log
I. Exceptions                27. Multitasking users*      3K. TCP/IP data menu*
                             28. User configuration*      3L. User communication
K. User defined data*        29. Linux systems*           3M. User wait states
```

oprofile – the Open Source sampling tool

- oprofile offers profiling of all running code on Linux systems, providing a variety of statistics
  - By default, kernel mode and user mode information is gathered for configurable events
- System z hardware currently does not have support for hardware performance counters, instead timer interrupt is used
  - Enable the hz_timer(!)
- The timer is set to whatever the jiffy rate is and is not user-settable
- Novell / SUSE: OProfile is on the SDK CDs
- More info at:
  - http://oprofile.sourceforge.net/docs/
  - http://www.redhat.com/docs/manuals/enterprise/RHEL-4-Manual/sysadmin-guide/ch-oprofile.html

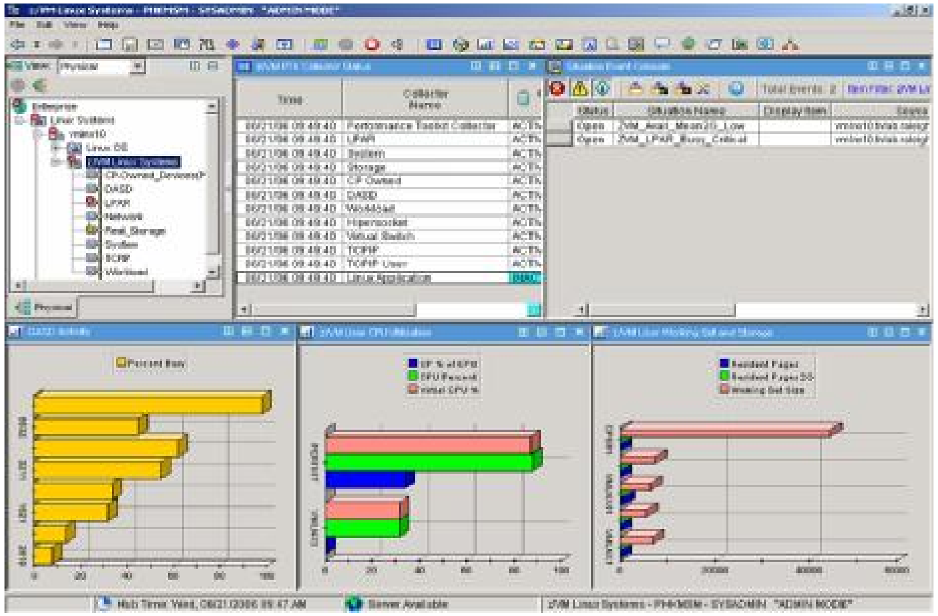
opreport

```
>opreport
CPU: CPU with timer interrupt, speed 0 MHz (estimated)
Profiling through timer interrupt
          TIMER:0|
  samples|      %|
------------------
   140642 94.0617 vmlinux-2.6.16.46-0.4-default
     3071  2.0539 libc-2.4.so
     1925  1.2874 dbench
     1922  1.2854 ext3
     1442  0.9644 jbd
      349  0.2334 dasd_mod
      152  0.1017 apparmor
        6  0.0040 oprofiled
        5  0.0033 bash
        5  0.0033 ld-2.4.so
        1 6.7e-04 dasd_eckd_mod
        1 6.7e-04 oprofile
```

- Kernel
- glibc
- application
- file system
- journaling
- dasd driver
- security
- ...

OMEGAMON XE on z/VM and Linux

A New Solution for the New Needs of z/VM and Linux on System z



- Single solution for managing VM and Linux on System z
- Reflects most common implementation in marketplace
- Leverages value of z/VM Performance Toolkit

Provides workspaces that display:

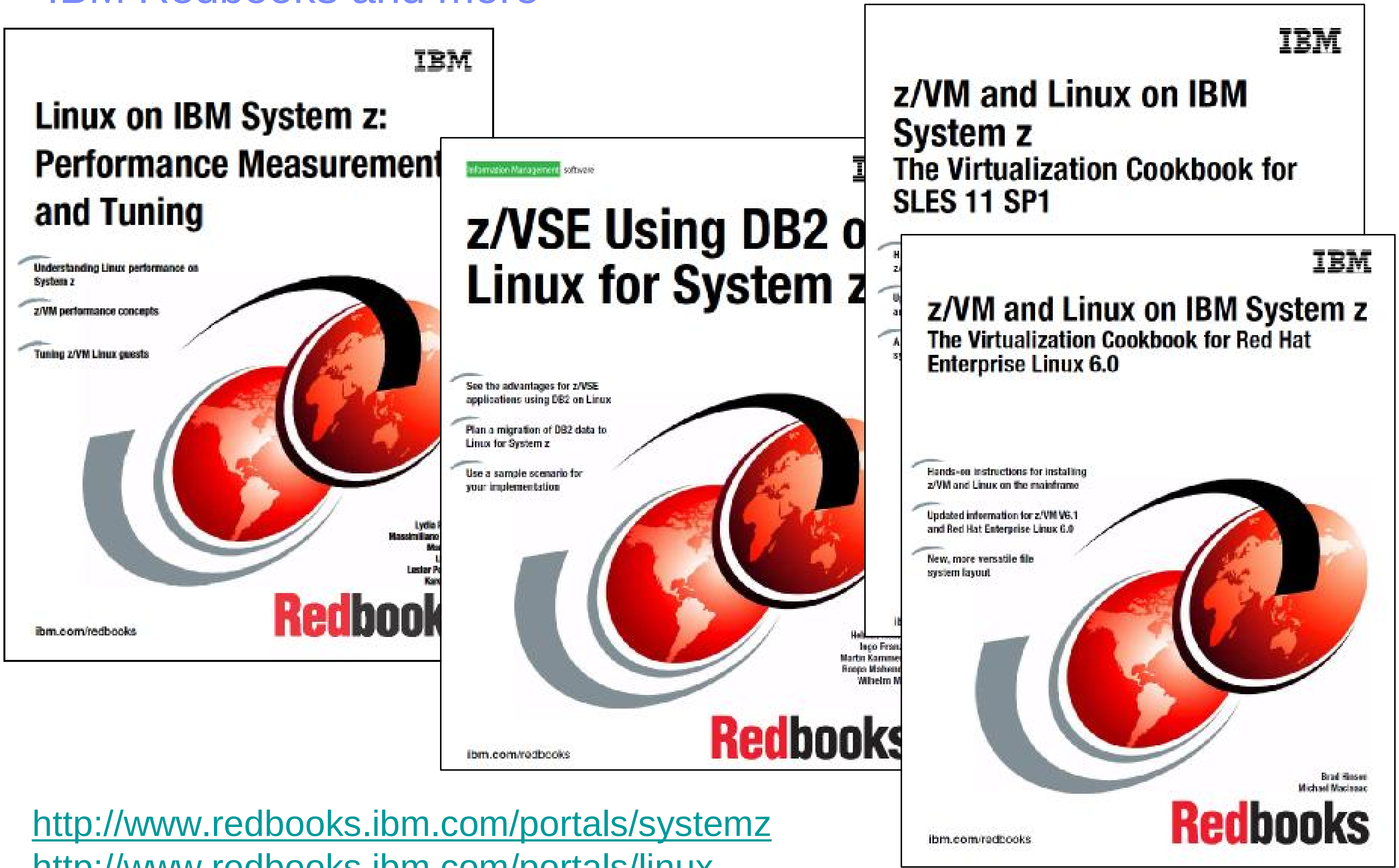
- Overall System Health
- Workload metrics for logged-in users
- Individual device metrics
- LPAR Data

- Composite views of Linux running on z/VM

More info sources on performance

- z/VM performance
  - http://www.vm.ibm.com/perf/
  - http://www.vm.ibm.com/perf/tips/linuxper.html
- Linux on System z
  - http://www-03.ibm.com/systems/z/os/linux/resources/doc_pp.html
  - http://www.ibm.com/developerworks/linux/linux390/perf/index.html
- Linux – VM Organization
  - http://www.linuxvm.org/
- IBM Redbooks
  - http://www.redbooks.ibm.com/
- IBM Techdocs
  - http://www.ibm.com/support/techdocs/atsmastr.nsf/Web/Techdocs

IBM Redbooks and more



http://www.redbooks.ibm.com/portals/systemz
http://www.redbooks.ibm.com/portals/linux

Some final thoughts

- Collect data for a base line of good performance.
- Implement change management process.
- Make as few changes as possible at a time.
- Performance is often only as good as the weakest component.
- Relieving one bottleneck will reveal another. As attributes of one resource change, expect at least one other to change as well.
- Latent demand is real.

Questions?

Trademarks

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

AIX*	FICON*	Parallel Sysplex*	System z10
BladeCenter*	GDPS*	POWER*	WebSphere*
CICS*	IMS	PR/SM	z/OS*
Cognos*	IBM*	System z*	z/VM*
DataPower*	IBM (logo)*	System z9*	z/VSE
DB2*			

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.
Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license there from.
Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.
Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.
InfiniBand is a trademark and service mark of the InfiniBand Trade Association.
Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
UNIX is a registered trademark of The Open Group in the United States and other countries.
Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.
IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency, which is now part of the Office of Government Commerce.

* 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.