Bruce Hayden – IBM Advanced Technical Sales Support
15 April 2014

z/VM and Linux on System z Performance “Best Practices”

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.

Acknowledgements

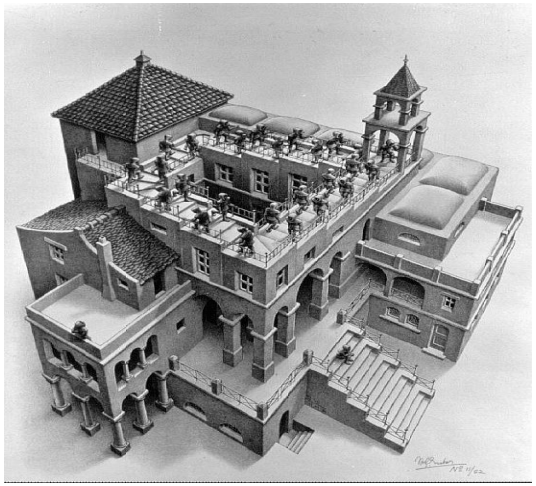
- Original Author (now retired!)
  - Jon vonWolfersdorf a.k.a. **Wolf**
- The following people contributed material to this presentation:
  - Bill Bitner
  - Horst Hartmann
  - Richard Lewis
  - John Schnitzler (retired)
  - Steve Gracin (retired)
  - Stephen Kinder

Agenda

- Installation and Configuration Performance “Best Practices”
  - Introduction
  - Maintenance/Service
  - Processors
  - Memory
  - I/O
  - Virtual Networking
  - z/VM 6.2/6.3 Considerations
- Reference Materials

Programmer's perspective on the “real server” world

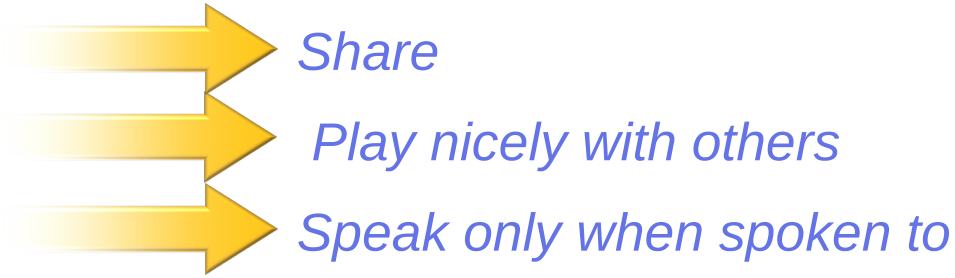
Get programmers
to change their perspective



- On **dedicated** servers it “doesn’t matter” if the application uses all the resources of the server
- “Real” server practices that **hurt** on **virtualized** systems:
  - Extra CPU burned by an idling application, background process, and/or redundant processing isn't wasted because other processes can't use it anyway.
  - Extra memory used because it’s dedicated, can't be shared anyway. If the application doesn’t use it, memory will just sit there unused.
  - No need to debug mem leaks – just reboot the world daily
    - ➢ A twofer!!! extra memory used all day, and CPU burned on restart. Ugh

Programmer's perspective on the virtualized world

Kindergarten principles:



- Gone are the days of unlimited CPU and memory dedicated to a single application.
- Applications need to be concerned with the impact on their hypervisor neighbors.
- Virtualized Systems have cost advantages
  - efficient use of CPU and memory, faster server provisioning, easier disaster recovery, administration, lower power, space,cooling consumption
- SysAdmins will pack as many guests as possible into a single hypervisor instance.
- See “*Java Design and Coding for Virtualized Environments*”
  - http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102089

z/VM and Linux on System z

Performance “Best Practices”

Maintenance and Service

z/VM service level

- Recommend maintaining a policy or schedule for software currency
  - 6 months to a year schedule recommended
- Apply Recommended Service Upgrade (RSU):
  - Released every 3-6 months
  - Contains cumulative service including all pre and co-requisites in a pre-built format
  - Includes service for all integrated components and pre-installed program products
  - Available on 3590 tape, DVD, or electronically (servlink envelope)
  - Includes service required by most customer installations
  - Pre-tested by development
  - Easy to install:
    - SERVICE
    - PUT2PROD
  - Easy to remove or back out
    - SAPL – IPL from CPLOLD MODULE
    - VMSES/E - VMFREM
- Check the following urls for important z/VM service topics:
  - http://www.vm.ibm.com/service/news
  - http://www.vm.ibm.com/service/redalert
  - http://www.vm.ibm.com/service/rsu

Linux kernel level

- Recommend using the most current distribution/version that has been tested and officially supports required hardware, middleware, and/or applications for target workload
- Recommend maintaining current service pack level via:
  - **SLES** > YaST Online Update (YOU)
  - **Red Hat** > Red Hat Network (RHN)
- Distribution service pack updates include:
  - Security Fixes
  - Other Fixes
  - Performance enhancements
  - New function
- Kernel level easily identified by “uname” command

z/VM and Linux on System z

Performance “Best Practices”

CPU and Processors

Physical/logical processors

- Physical processors (hardware limits):
  - zEC12: 101
  - z196: 80
  - z10EC: 64
- LPAR logical processors same as hardware limits
- Logical processors (supported by z/VM LPAR)
  - z/VM 5.4 and newer support up to 32 logical processors
  - POK recommendation, maximum 4:1 logical to real ratio
    - Real life experience, 3:1 is about the maximum
  - Recommend defining reserved processors in the LPAR Activation Profile
    - Permits non-disruptive activation of engines to the z/VM LPAR at a later date

z/VM HiperDispatch

- HiperDispatch is a new feature in z/VM 6.3
  - Uses vertical mode partitions
  - Features topology-aware dispatching
    - Try to keep virtual MP cpus on close together real cpus
    - Be smarter about moving VCPUs for balancing
- Recommended, and enabled by default when running in an LPAR
  - SET SRM POLARIZATION HORIZONTAL to turn it off dynamically
  - Not available when running second level
- Needs Global Performance Data Control enabled in the LPAR profile
- Tuning knobs:
  - SET SRM CPUPAD
    Influence the system predicted future CPU consumption
  - SET SRM EXCESSUSE
    How aggressive in using excess real CPU capacity (high, medium, low)

Guest virtual processors

- Virtual processors
  - Maximum per virtual machine (architected): 64
  - Various guest operating systems and workloads scale differently
  - Recommendations:
    - Configure the number of virtual processors per guest for peak workload, no more
    - Do not define more virtual processors to a guest than logical processors defined to the z/VM LPAR
  - High diagnose x’44’ or x’9C’ rates (spinlock) may be an indication of too many virtual processors
    - Perfkit report - FCX104 (Privileged Operations)
    - Thresholds to watch for:

      x’44’ > 50,000/sec

      x’9C’ > 5,000/sec
    - See the following url for details on diagnose x’9C’ support in z/VM:

      http://www.vm.ibm.com/perf/reports/zvm/html/diag9c.html

Beware of “SHARE” of virtual MP machines

- The default SHARE setting for all virtual machines is “Relative 100”:
  - VM dispatches users by VMDBK
  - There is one VMDBK per virtual processor defined
  - A users SHARE setting is divided among the defined virtual processors
- Recommend the initial SHARE of virtual MP machines:
  - Set SHARE RELATIVE (100 * number of virtual CPUs defined)
  - This maintains an initial “level playing field”
- Adjust SHARE of guest virtual machines from this point, as required:
  - Increase SHARE to prioritize
  - Decrease SHARE to penalize
- A virtual machine's SHARE only comes into play when there is contention for resources, primarily CPU

More on SHARE – surplus SHARE distribution problem

User ID	Share	Normalized	Wants	Should get	Actually gets
LINUX01	100	17%	100%	24.5%	17%
LINUX02	100	17%	100%	24.5%	17%
LINUX03	200	33%	100%	48%	63%
LINUX04	200	33%	3%	3%	3%

- The relative “surplus” share that LINUX04 does not want to use is not distributed proportionally to the remaining user ids that would use it.
- IBM is aware of the problem and is studying ways to correct it.
- Recommendation:
  - Use absolute shares for users with low average usage that require a high share.
  - Examples from the default user directory:
    - TCPIP and its related servers
    - Default SFS servers (VMSERVR, VMSERVS, etc.)

Quick dispatch

- Setting QUICKDSP:
  - Bypasses System Resource Management memory controls
  - Places a virtual machine directly into the dispatch list
  - The virtual machine exempt from being placed in an eligible list
- QUICKDSP should be reserved for:
  - Service Virtual Machines performing critical functions on behalf of other guests (i.e. RACFVM, TCPIP)
  - Select key production (i.e. data base) guests
- SRM values should be used to adjust scheduler/dispatcher behavior for servicing most guests.
- See http://www2.marist.edu/htbin/wlvtype?LINUX-VM.30359 for an excellent detailed explanation by Malcolm Beattie (IBM)

Linux runlevel

- Similar to Microsoft Windows, Linux has different modes of operation or “runlevels”
- When you boot Linux, it will initialize at a predefined default runlevel (this is usually 3 or 5). There are six different runlevels defined by most Linux distributions:
  - 0 - Halt the system
  - 1 - Single-user mode
  - 2 - Multi-user mode (without networking)
  - **3 - Multi-user mode**
  - **5 - Multi-user mode (display manager, GUI)**
  - 6 - Reboot the system
- Most desktop Linux systems boot into runlevel 5 by default and users are presented with a graphical interface
- Most server Linux systems boot into runlevel 3 by default and users are presented with a line mode interface
- Recommend runlevel 3 for Linux guests of VM:
  - X services are costly in terms of cpu cycles
  - Use a lightweight X-server like VNC server, instead of full GUI desktop

Unnecessary guest virtual machines

- Shutting down unnecessary guest virtual machines helps to improve the overall performance of the system:
  - Linux guest virtual machines almost never go dormant
- Logoff:
  - Golden images used for cloning
  - Test machines and “sand boxes”

- Suspend:
  - Production guests not necessary during POC testing or benchmarking of another application or workload
  - See *Methods to pause a z/VM guest: Optimize the resource utilization of idling servers* at http://public.dhe.ibm.com/software/dw/linux390/perf/l0wadp00.pdf
- Reduce “SHARE” setting:
  - For virtual machines running lower priority workloads

Unnecessary services/applications

- There are a number of services in Linux that get started at boot depending on:
  - Distribution
  - Linux kernel level/version
  - Installed software packages
- Shutting down unnecessary services and unused applications helps to improve the overall performance of the system
  - Status of services can be queried/changed with the “chkconfig” command
- The cron daemon is useful for scheduling events to be kicked off automatically at a specific time or at regular intervals
  - Running many guests with identical schedules can cause CPU spikes and stress the z/VM paging subsystem:
    - Remove unnecessary events from cron
    - Stagger scheduled kick-off time of events

z/VM and Linux on System z

Performance “Best Practices”

Memory

z/VM memory configuration

- zEC12, z196, & z10 hardware limits:
  - 1 TB Central Storage in an LPAR
- z/VM 5.4 through z/VM 6.2 support up to:
  - 256 GB Central, 128 GB Expanded Storage
  - z/VM 6.3: 1 TB Central, 128 GB Expanded Storage
  - Plan on a virtual to real (V:R) memory ratio in the range of 1.5:1 to 3:1
    - Production systems will typically be closer to the low end of range
    - Development/Test systems may be able to push the upper end of range
- Recommend configuring some processor memory on 6.2 as expanded storage:
  - Serves as high speed cache
  - See http://www.vm.ibm.com/perf/tips/storconf.html for details
- For z/VM 6.3, recommend no expanded storage
  - Statement of direction issued that xstore will not be supported in a future release
- Rule Of Thumb for z/VM 6.2: 2GB to 4GB xstore is sufficient for most workloads
- As of z/VM5.4, STANDBY memory can be added dynamically to central storage:
  - Storage must be defined as “RESERVED” in the VM LPAR Activation Profile
  - Cannot be removed dynamically, only added
  - Central only, dynamic expansion not supported for expanded storage

Linux virtual memory sizing

- The maximum supported virtual machine memory size is 1 TB (hardware limit)
- The most common mistake made by customers running Linux guests under z/VM is over-configuring virtual memory:
  - In a dedicated server environment, traditional wisdom suggests installing as much memory as possible. Excess memory used as:
    - I/O buffer
    - File system cache
  - In a virtualized environment under z/VM, oversized guests place unnecessary stress on the VM paging subsystem:
    - Real memory is a shared resource, caching pages in a Linux guest reduces memory available to other Linux guests
    - Larger virtual memory requires more kernel memory for address space management
  - Right sizing Linux memory requirements on z/VM:
    - Is accomplished by trial and error
    - Monitored with the “free” or “vmstat” commands along with /proc/meminfo
  - See the tuning hints and tips available on this web page:
    - http://www.ibm.com/developerworks/linux/linux390/perf/index.html

z/VM paging subsystem

- Primary rule: Have enough page space! You WILL abend if you don't!
  - z/VM 6.2 and older: Plan for a DASD page space utilization < 50%
    - Page space tends to get fragmented over time
    - Large contiguous free space allows for greater paging efficiency
    - Block page size is the key performance indicator:
      - Aim for double digits, 10 or more 4K pages per block set
      - Perfkit report - FCX109 (CP Owned Device)
  - z/VM 6.3 may use more page space than previous releases!
    - Plan for what you need. Read “Amount of paging space” in http://www.vm.ibm.com/perf/reports/zvm/html/630con.html
    - It is OK for page space utilization to go above 50%
  - Monitor usage with Q ALLOC PAGE command
- Use multiple channels to spread out I/O to paging devices
- Do not mix page space with any other space on a volume
- Recommend using devices of the same size/geometry
- EDEVs as paging drives are an option:
  - Have observed 1.6 I/Os per emulated FBA volume
  - At slightly higher CPU costs

Linux swap space

- The traditional recommendation in a dedicated server environment is that swap space should be twice the memory size of a Linux machine
- This should not apply to a z/VM Linux guest:
  - Some swap space should be defined to prevent Linux from hanging and/or a kernel panic during unexpected memory demands
  - Properly sized Linux guests should have minimal swapping under normal load
- z/VM offers multiple options for swap devices
- Recommendation:
  - One or two small V-disks (256MB - 512MB)
  - If necessary, additional minidisk(s) or dedicated volume(s)
    - Set priorities in fstab so that the V-disk(s) are used first
- See http://www.redbooks.com/abstracts/sg246926.html for more details and test results for various swap device options

z/VM reorder processing – z/VM 6.2 and earlier

- *Page reorder* is the process in z/VM for managing user owned frame lists as input to demand scan processing:
  - It serializes the virtual machine
  - It includes checking/resetting the hardware reference bit
  - It is done periodically on a virtual machine basis
- The cost of reorder processing is proportional to the number of resident frames owned by a virtual machine:
  - Roughly 130ms per GB of resident memory

- Recommendation:
  - Turn reorder processing “off” for Linux guests >= 8GB
  - Things to watch for:
    - Resident page fields (R<2GB & R>2GB) on Perfkit report - FCX113 (User Page Data)
    - Console Function Wait field (%CFW) on Perfkit report - FCX114 (User Wait States)
  - See the following url for additional details on Reorder Processing:
    - http://www.vm.ibm.com/perf/tips/reorder.html
- Page reorder has been eliminated in z/VM 6.3.

z/VM SRM settings

- Tuning knobs that influence the z/VM scheduler and dispatcher behavior
- Default values in z/VM V5 and 6.1 were an artifact from the past:
  - Interactive CMS virtual machines
  - Small memory footprints
- **STORBUF**
  - Defines amount of memory to be used in scheduler algorithms
  - Systems today are configured to over-commit central storage
    - Recommended starting values (and the default in z/VM 6.2 and later):
      STORBUF 300 250 200
- **LDUBUF**
  - Defines amount of parallel paging “capacity” to be used in scheduler algorithms
  - There are conflicting opinions on a recommended setting:
    - Default values - LDUBUF 100 75 60
    - Default values may be “OK” as a starting point depending on:
      Number of page volumes defined
      Number and size of active Linux guests
- **DSPBUF**
  - Defines number of guests allowed in the dispatch list:
    - Default values - DSPBUF 32767 32767 32767
    - Not recommended to adjust these settings unless directed by development

z/VM minidisk cache

- z/VM minidisk cache is a write-through cache:
  - Improves read I/O performance, but It’s not free
- Not recommended for:
  - Highly memory constrained systems
  - Linux swap file disks
  - Linux file systems for logging, etc.
- Default system settings are less than optimal and can consume too much storage
- Recommended settings:
  - Disable MDC when usage is not recommended
    - Add **MINIOPT NOMDC** after the MDISK directory statement
  - Eliminate MDC in expanded storage (code this in AUTOLOG1's PROFILE EXEC)
    - **CP SET MDC XSTORE 0M 0M**
  - Limit MDC in central storage, amount depends on storage size and usage
    - **CP SET MDC STORE 0M 256M**
  - Monitor with Q MDC command and/or a performance monitor
    - Perfkit report - FCX103 (Storage Utilization)

z/VM and Linux on System z

Performance “Best Practices”

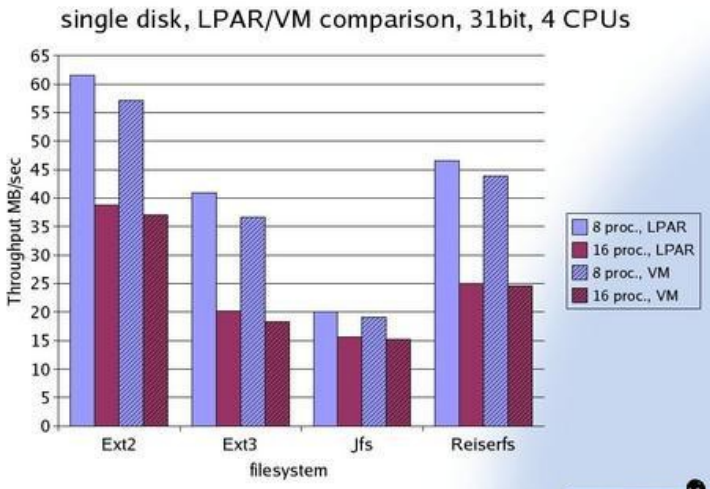
DASD I/O

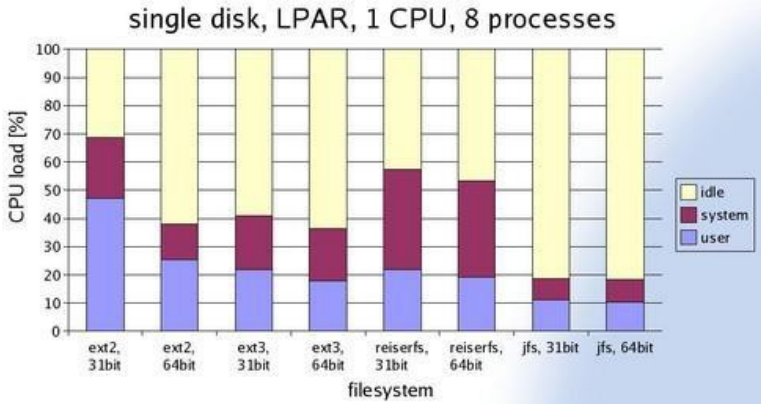
Disk performance

- Hardware connectivity choices:
  - ESCON 17 Mb/sec (!)
  - FICON available in 1 Gb, 2 Gb, 4 Gb, & 8 Gb channel speeds
- SCSI verses ECKD/FBA recommendations:
  - ECKD or FBA for z/VM and Linux “/” file system
  - SCSI LUNs for application data and databases
- Maximize hardware performance:
  - Use maximum speed channels
  - Configure maximum number of channel paths
  - Spread disks over multiple ranks within a storage subsystem
  - Use logical volumes with striping
  - Exploit PAV or HyperPAV to prevent queuing
- References:
  - http://www.vm.ibm.com/perf/reports/zvm/html/scsi.html
  - http://www.ibm.com/developerworks/linux/linux390/perf/tuning_diskio.html

Linux file systems

- EXT2 - most widespread Linux file system.
- EXT3 - evolved from ext2, adds journaling features.
- JFS - a port of OS/2 Warp Server jfs to Linux.
- Reiserfs – journaling behavior is comparable to ext3 in order mode.
- XFS - the IRIX file system, which was released in 2000 as open source.
- Recommend using xfs or ext3 because of their journaling capabilities and reduced cpu load compared to other journaling file systems.
- EXT4 – Now available…
- See performance reports at:
  - http://www.ibm.com/developerworks/linux/linux390/perf/tuning_filesystems.html





z/VM dump & spool space

- Dump space
  - Ensure there is sufficient dump space defined to the system
  - Recommend defining dedicated dump volumes
  - Dump space requirements vary according to memory usage
    - Q DUMP – identifies allocated dump space.
    - Calculation guidelines are located in the CP Planning and Administration Manual
- Spool space
  - Various uses:
    - User printer, punch, reader files (console logs)
    - DCSS, NSS
    - System files
    - Page space overflow
  - Spool management:
    - Monitor with Q ALLOC SPOOL command
    - Use the SFPURGER utility:
      - Rule based tool to clean up spool space
      - Included in z/VM

z/VM and Linux on System z

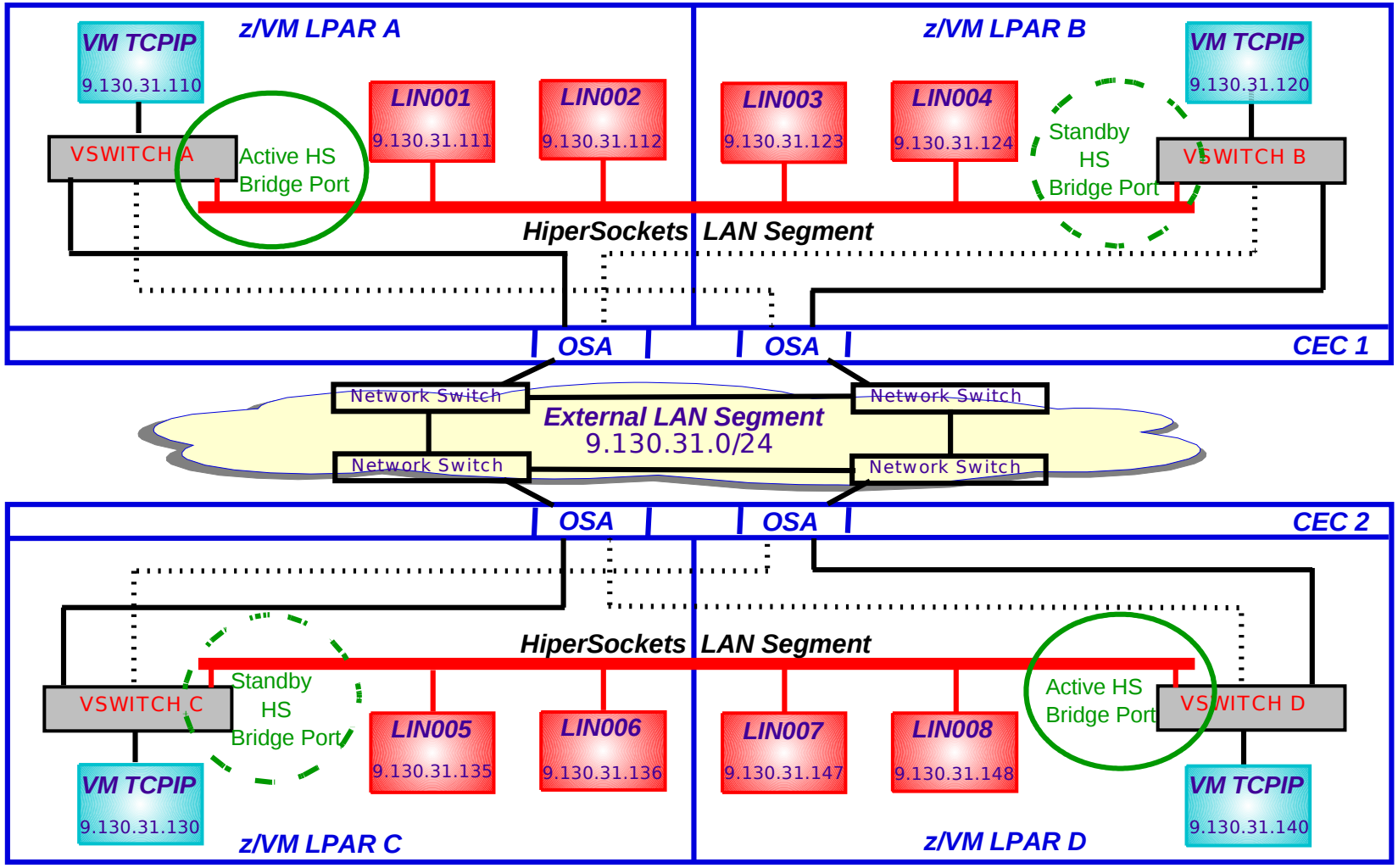
Performance “Best Practices”

Virtual Networking

Networking configuration options

- Three basic configurations for external network connectivity:
  - Dedicated OSA
  - Routed LAN
  - VSWITCH (recommended)
    - Lower cpu costs
    - Built-in failover
    - Operates in Ethernet or IP modes
    - Supports 802.1q VLANs (by port or by user)
    - Supports port isolation
    - Supports 802.3ad link aggregation
- Cross LPAR network connectivity:
  - Shared OSA Express
  - HiperSockets
  - HiperSockets Bridge (z/VM 6.2 & zEnterprise)
- References:
  - z/VM Connectivity Manual (SC24-6174)
    - http://publib.boulder.ibm.com/cgi-bin/bookmgr/download/HCSC9C20.pdf
  - Linux on System z Tuning Hints and Tips for Networking
    - http://www.ibm.com/developerworks/linux/linux390/perf/tuning_networking.html
  - Advanced Networking Concepts Applied Using Linux on IBM System z
    - http://www.redbooks.ibm.com/Redbooks.nsf/RedbookAbstracts/sg247995.html?Open

HiperSockets bridge - cross CEC



MTU – size matters!

- Set MTU to the maximum size supported by all hops on the path to the final destination to avoid fragmentation:
  - Use tracepath destination to verify the path MTU size
  - If the application data is <=1400 bytes, use an MTU size of 1492
  - If the application is able to send larger chunks of data, use an MTU size of 8992
- TCP uses the MTU for the window size calculation, not the actual application send size
- For VSWITCH, an MTU size of 8992 is recommended:
  - OSA card is optimized for use with an 8992 MTU
  - Synchronous operation, SIGA required for every packet
  - No packing like a dedicated OSA card
- For HiperSockets, select an MTU size to suit the workload:
  - If the application is capable of sending large packets, a larger MTU size will increase throughput and decrease CPU cycles
  - An MTU size of 56K is recommended only for data streaming workloads with packets >32KB

Inbound QDIO buffer

- The QDIO inbound buffer queue can be increased for high volume workloads:
  - The default is 16
  - Valid range is 8–128
  - QDIO OSA buffer size is 64K
  - IQDIO HiperSockets buffer size is equal to the HiperSockets MFS (16K, 24K, 40K, 64K)
- Current buffer count can be displayed with the **lsqeth -p** command
- A QDIO OSA buffer count of 128 equates to 8MB locked memory:
  - 128 x 64KB = 8MB
- Set the inbound buffer queue size in the appropriate config files (for example, virtual NIC F200):

  - SUSE SLES 10: **/etc/sysconfig/hardware/hwcfg-qeth-bus-ccw-0.0.f200** add:

    QETH_OPTIONS="buffer_count=128"
  - SUSE SLES 11: **/etc/udev/rules.d/51-qeth-0.0.f200.rules** add:

    ACTION=="add", SUBSYSTEM=="ccwgroup", KERNEL=="0.0.f200", ATTR{buffer_count}="128“
  - RedHat RHEL 5&6: **/etc/sysconfig/network-scripts/ifcfg-eth0** add:

    OPTIONS="buffer_count=128"

Checksumming

- HiperSockets doesn't require checksumming because it is a memory-to-memory operation protected by ECC:
  - The default setting is sw_checksumming
  - Turning off checksumming for HiperSockets can save between 7%-13% in CPU costs

- Recommendation:
  - Switch checksumming off for HiperSockets:
    - SUSE SLES 10: **/etc/sysconfig/hardware/hwcfg-qeth-bus-ccw-0.0.f200** add: **QETH_OPTIONS="checksumming=no_checksumming"**
    - SUSE SLES 11: **/etc/udev/rules.d/51-qeth-0.0.f200.rules** add: **ACTION=="add", SUBSYSTEM=="ccwgroup", KERNEL=="0.0.f200", ATTR{checksumming}="no_checksumming“**
    - RedHat RHEL 5&6: **/etc/sysconfig/network-scripts/ifcfg-eth0** add: **OPTIONS="checksumming=no_checksumming"**

SYSCTL settings

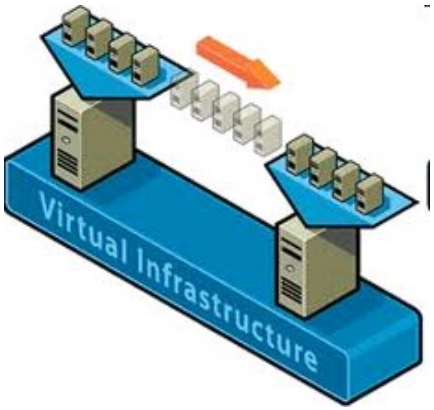
- The following system wide sysctl settings can be changed temporarily by the sysctl command or permanently in the config file:
  - **/etc/sysctl.conf**
- The processor input queue length should be increased from the default size of 1000 to at least 2500 using sysctl:
  - **sysctl -w net.core.netdev_max_backlog =2500**

- Adapt the inbound and outbound window size to suit the workload
  - The following values are recommended for OSA devices:
    - **sysctl -w net.ipv4.tcp_wmem=“4096 16384 131072”**
    - **sysctl -w net.ipv4.tcp_rmem=“4096 87380 174760”**
  - System wide window size applies to all network devices
  - Applications can still use setsockopt to adjust the window size
    - Has no impact on other network devices
- As a general rule of thumb, the default send/receive window size should be at least twice the MTU size
  - The SAP Enqueue Server requires a default send/receive window size of four times the MTU size

z/VM and Linux on System z

Performance “Best Practices”

SSI and LGR Considerations



SSI cluster configuration

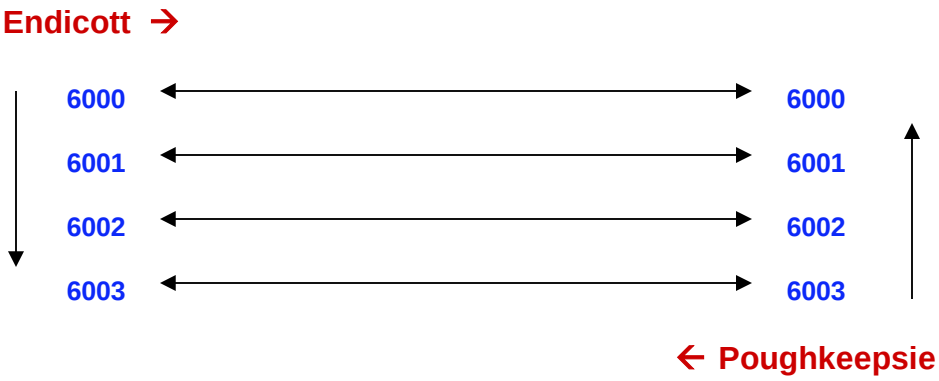
- Suggested configuration for 4-member cluster is 2 LPARs on each of 2 CECs

- Guest relocation time can be impacted by several key factors:
  - Number of ISFC Links (1 – 16)
  - Speed of ISFC Links (1Gb – 8Gb)
  - Size of guest virtual machine
  - How active the guest virtual machine is
  - Resource contention/availability on destination member

- Recommendation:
  - Minimum 4 CTCs between each cluster member
    - 2 FICON CHPIDs, 2 CTCs per CHPID
  - Maximum 16 CTC’s between each cluster member
    - 4 FICON CHPIDs, 4 CTCs per CHPID
  - Testing has shown that 4 CTCs per CHPID provides the most efficient data transfer rates
    - Performance begins to degrade as the number of CTCs are increased beyond 4 per CHPID

CTC subchannel addressing

- Recommended practice: Use the same real device number for the same CTC on each SSI cluster member
  - Potential performance impact
    - Algorithm does not use Round Robin
    - The more CHPIDs the greater the impact
  - ISFC communications between two cluster members is done by:
    - Member name first in alphabet uses lowest subchannel address to highest
    - Member name second in alphabet uses highest subchannel address to lowest



Live guest relocation

- To qualify for relocation, a guest virtual machine must meet eligibility requirements, including:
  - It must be logged on, but in a disconnected state
  - The architecture and functional environment on destination member must be comparable to origin member
    - A relocation domain defines a set of members among which virtual machines can relocate freely
  - Destination member must have the capacity to accommodate the guest
    - CPU
    - Memory
    - Paging subsystem
  - Devices and resources needed by guest must be shared and available on destination member
    - Network connections
    - DASD
- Use VMRELOCATE command with TEST operand
- Recommend relocating guests serially (do not use asynchronous option)
  - Quiesce time is shorter

Virtual MAC addressing in an SSI cluster

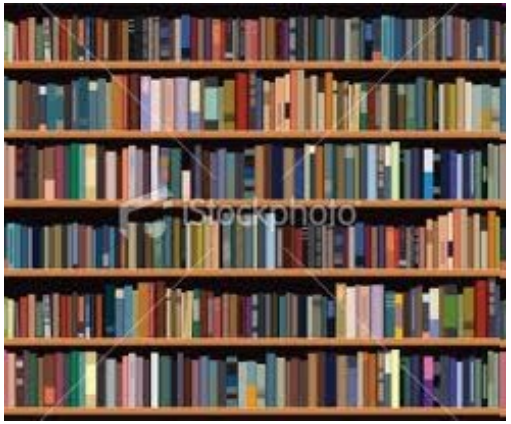
- MAC address assignments are coordinated across an SSI cluster
  - VMLAN statement
    - MACPREFIX must be set to different value for each member
      - Default is 02-00-00 for each member
      - Recommend last two bytes be replaced with the "system number" of each member
    - USERPREFIX must be set for SSI members
      - Must be identical for all members
      - Must not be equal to any member's MACPREFIX value
      - Default is 02-00-00
    - MACIDRANGE is ignored in an SSI cluster
      - Because MAC assignment is coordinated among members
  - Example:
    - **VMSYS01: VMLAN MACPREFIX 021111 USERPREFIX 02AAAA**
    - **VMSYS02: VMLAN MACPREFIX 022222 USERPREFIX 02AAAA**
    - **VMSYS03: VMLAN MACPREFIX 023333 USERPREFIX 02AAAA**
    - **VMSYS04: VMLAN MACPREFIX 024444 USERPREFIX 02AAAA**

Questions?

z/VM and Linux on System z

Performance “Best Practices”

Reference Material



Technical assessments, sizing, & capacity planning

- **Techline Services - http://w3-03.ibm.com/support/techline/global/index.html/**

Installation, planning, & administration

- The following documentation can be extremely helpful for Installation, Planning and Administration:
  - z/VM CP Planning and Administration (SC24-6178)
    - **http://publibz.boulder.ibm.com/epubs/pdf/hcsg0c11.pdf**
  - Getting Started with Linux on System z (SC24-6194)
    - **http://publibz.boulder.ibm.com/epubs/pdf/hcsx0c11.pdf**
  - Virtualization Cookbooks by Michael MacIsaac and friends:
    - **http://www.redbooks.ibm.com/**

z/VM and Linux on IBM System z: The Virtualization Cookbook for SLES 10 SP2 (SG24-7493-00)

z/VM and Linux on IBM System z: The Virtualization Cookbook for RHEL 5.2 (SG24-7492-00)

z/VM and Linux on IBM System z: The Virtualization Cookbook for SLES 11 SP1 (SG24-7931-00)

z/VM and Linux on IBM System z: The Virtualization Cookbook for RHEL 6 (SG24-7932-00)

- **http://www.vm.ibm.com/devpages/mikemac/**

z/VM and Linux on IBM System z: The Virtualization Cookbook for z/VM 6.2, RHEL 6.2, and SLES 11 SP2

z/VM and Linux on IBM System z: The Cloud Computing Cookbook for z/VM 6.2, RHEL 6.2, and SLES 11 SP2

Additional references

- Web Sites
  - **http://www.vm.ibm.com/perf/**
    - z/VM Performance Web Site
  - **http://www.ibm.com/developerworks/linux/linux390/perf/index.html/**
    - Linux on System z Performance Web Site
  - **http://www.linuxvm.org/**
    - Linux on System z interest site maintained by Mark Post of SUSE
- Redbooks
  - **http://www.redbooks.ibm.com/**
    - An Introduction to z/VM SSI and LGR (SG24-8006)
    - Linux on IBM eserver zSeries and S/390: Performance Toolkit for VM (SG24-6059)
    - Linux on IBM eserver zSeries and S/390: Performance Measurement and Tuning (SG24-6926)
- z/VM Library
  - **http://www.vm.ibm.com/library/**
    - z/VM Performance (SC24-6208)
    - z/VM Performance Toolkit Guide (SC24-6209)
    - z/VM Performance Toolkit Reference (SC24-6210)
    - z/VM Connectivity (SC24-6174)