Session L34

Making z/VM and Linux Guests Production Ready…
"Best Practices"

Jon vonWolfersdorf
wolff@us.ibm.com

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

Trademarks

Acknowledgements

- The following people contributed material to this presentation:
  - Bill Bitner
  - Steve Gracin
  - Richard Lewis
  - John Schnitzler

Agenda

- Design and Configuration Planning
- Installation/Configuration “Best Practices”
  - z/VM
  - Linux on zSeries
  - Virtual Networking
- System Health Check
- References

Design & Configuration Planning

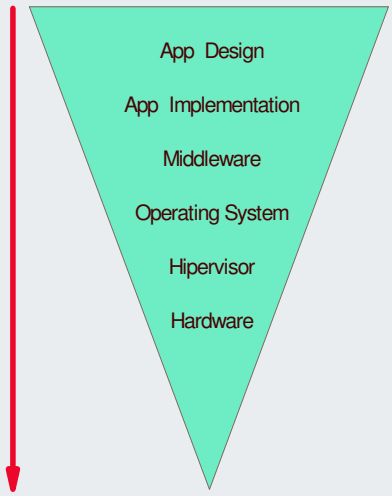
- Planning is important!
- Contact IBM Techline for:
  - Sizing assistance
  - Scheduling a Solution Assurance Review (SAR)
    - http://dalnotes1.dfw.ibm.com/atss/techxpress.nsf/request?OpenForm
- Contact Tim Hayford, zSeries New Workload TSS Mgr. for:
  - Application assessment/selection assistance
  - z/VM and Linux installation assistance
    - thayfor@us.ibm.com
- The following resources can be very helpful:
  - SC24-6083 z/VM CP Planning and Administration
    - http://publibz.boulder.ibm.com/epubs/pdf/hcsg0b11.pdf
  - SC24-6096 Getting Started with Linux on zSeries
    - http://publibz.boulder.ibm.com/epubs/pdf/hcsx0b10.pdf
  - Virtualization Cookbooks by Michael MacIsaac
    - http:/www.redbooks.ibm.com

Monitoring Your System Performance

- Basic CP commands provide only general performance indicators.
- Additional monitoring products are required.
- Multiple levels of monitoring is required:
  - z/VM – Performance ToolKit, Omegamon XE, Velocity ESAMON
  - Linux for zSeries – systat package, top, rmf data gatherer, appldata, Velocity ESATCP
  - Application – Tivoli Performance Viewer, Wily Introscope, Sitraka Jprobe
- Capturing performance data as a base line is a must:
  - General history data – business as usual.
  - Detailed raw monitor data prior to and following any major changes.
- Change management can be critical to avoiding or solving performance problems.

Tuning Your System

- There is an effect of diminishing returns from tuning efforts:
  - Application design
  - Application implementation
  - Middleware
  - Operating system
  - Hipervisor
  - Hardware



z/VM
"Best Practices"

Maintenance Levels

- Recommend maintaining current service levels.
- Apply latest 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 the following pre-installed program products:
    - DirMaint
    - VM/RACF
    - Performance ToolKit
  - Available on tape or electronically.
  - Includes service required by most customer installations and all closed HIPER fixes available at the date of release.
  - Pre-tested by development.
  - Easy to install:
    - SERVICE
    - PUT2PROD

Memory Configuration

- Plan on a virtual to real (V:R) memory ratio in the range of 1.5:1 – 3:1.
- Recommend configuring some processor memory as expanded storage:
  - Increases consistency of response time.
  - See http://www.vm.ibm.com/perf/tips/storconf.html for gory details.
- Rule of Thumb - start with 25% of memory configured as expanded:
  - Typically 2–4GB of expanded storage is sufficient.
  - The lower the paging rate, the lower the amount of expanded storage required.
  - The greater the number of page frames available in central storage above 2GB, the higher the amount of expanded storage required.

Paging Subsystem

- Plan for DASD page space utilization < 50%:
  - Page space tends to get fragmented over time.
  - Large contiguous free page space allows for greater block paging efficiency.
  - Monitor usage with Q ALLOC PAGE command.
  - Block page size is a key performance indicator:
    - Aim for double digits – 10 or more pages per block set.
    - Performance Toolkit report DEV CPOWN (FCX109) “Block Page Size” field.
- 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 and geometry.
- Calculation guidelines are located in the CP Planning and Administration Manual.

Minidisk Cache

- z/VM minidisk cache is a write-through cache:
  - Improves read I/O performance.
  - But it’s not free.
- Not recommended for:
  - Memory constrained systems.
  - Linux swap file disks.
- Default system settings are less than optimal.
- Recommended settings:
  - Eliminate MDC in expanded storage.
  - Limit MDC in central storage – 10% is a good starting point.
  - Monitor with Q MDC command and/or a performance monitor.

System Resource Management Settings

- Influence the z/VM scheduler and dispatcher behavior.
- Default values are an artifact from the past:
  - Interactive CMS virtual machines.
  - Small memory footprint.
- STORBUF
  - Defines amount of memory to be used in scheduler algorithms.
  - Recommend modification to over-commit central storage.
    - Default values - STORBUF 120 105 95
    - Recommended starting values - STORBUF 300 250 200
- LDUBUF
  - Defines amount of 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:
      - Amount of DASD paging capacity defined.
      - Number and size of active Linux guests.
      - Workload characteristics.
- 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 defect support.

Quick Dispatch

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

Guest Privilege Classes

- Most Linux guests do not require anything more than privilege class “G”.
- Not a performance issue.
- More of a security issue:
  - A privileged Linux guest could shutdown the z/VM system.
  - A privileged Linux guest could compromise other guests or the entire z/VM system.
- Not limited to the VM 3215 Linux console session:
  - Neale Ferguson's cpint package (hcp command)
  - Linux vmcp command (October 2005 2.6 kernel stream)
- Recommend restricting privileges to the minimum required.

Processors

- Real Processors
  - z/VM 5.1 and z/VM 5.2 support up to 24 processors.
  - z/VM 5.3 supports up to 32 processors.
  - LPAR recommendation – no greater than a 4:1 logical to real ratio.
- Virtual Processors
  - Various guest systems and workloads scale differently.
  - Virtual Machine recommendation:
    - Configure the number of virtual processors per guest for peak workload, but no more.
    - Never define more virtual processors to a guest than logical processors defined to z/VM.
  - High diagnose x’44’ rates may be an indication of too many virtual processors.
    - Performance Toolkit reports CPU (FCX100) or PRIVOP (FCX104) can be used to monitor diagnose rates.

System Dump & Spool Space

- Dump Space
  - Ensure there is sufficient dump space defined to the system.
  - Dump space requirements varies 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
  - Management:
    - Monitor with Q ALLOC SPOOL command.
    - SFPURGER utility:
      - Rule based tool to clean up spool space.
      - Included in the no charge CMS Utilities Feature (CUF).

Linux on zSeries
“Best Practices”

© IBM Corporation 2007
2007 System z Technical Conference

Kernel/Update Level

- Recommend using the most current distribution/version that has been tested and officially supports required middleware and/or application.
- Recommend maintaining current service via:
  - YaST Online Update (YOU)
  - RedHat Network (RHN)
- Distribution service updates include:
  - Fixes
  - Performance enhancements
  - New function
- Kernel level identified by “uname” command.
- SuSE kernel and package levels can be identified with the “SPident” command.

Virtual Memory Sizing

- The most common mistake made by customers running Linux guests under z/VM is over-configuring Linux memory:
  - In a dedicated server environment, traditional wisdom suggests installing as much memory as possible/feasible. 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.
  - Rightsizing Linux memory requirements on z/VM:
    - Is accomplished by trial and error.
    - Monitored with the “free” command.
  - See http://www.ibm.com/systems/z/os/linux/pdf/avmlinux.pdf for a detailed sizing document by Stephen Wehr, IBM.

On-Demand Timer

- Linux uses a timer tic based interrupt model.
- By default, the timer “pops” 100 times per second.
- Timer interrupts on idle Linux guests:
  - Keep the guests in the dispatch list.
  - Create unnecessary overhead for z/VM.
- All current supported zSeries distributions include the jiffy timer patch and by default it is activated:
  - Novell – SLES9, SLES10
  - RedHat – RHEL4, RHEL5
- Query with the “sysctl kernel.hz_timer” command.
  - “sysctl -w kernel.hz_timer=1” enables the 100 Hz timer. The On-Demand Timer Patch is deactivated.
  - “sysctl -w kernel.hz_timer=0” disables the 100 Hz timer. The On-Demand Timer Patch is activated.

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 does 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 not swap or should have minimal swapping under normal load.
  - z/VM offer multiple options for swap devices:
    - Dedicated DASD
    - Minidisk
    - T-disk
    - V-disk
    - Expanded storage (XPRAM driver)
- Recommendation:
  - One or two small V-disks (200-300MB).
  - One medium sized minidisk or small dedicated volume (1 -2G).
  - Set priorities so that the V-disk(s) are used first.
- See http://www.redbooks.ibm.com/abstracts/sg246926.html?Open for more details and test results for various swap device options.

Runlevel

- Similar to Microsoft Windows “safe” and “command prompt only” modes, 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 most Linux distributions use:
  - 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 the user is presented with a graphical login prompt.
- Most server Linux systems boot into runlevel 3 by default and the user is presented with a text-mode login prompt.
- Recommend runlevel 3 for Linux guests of VM:
  - X services consume system resources.
  - Use a lightweight X-server like VNC server, instead of KDE/GNOME desktop.

Unnecessary Services/Applications

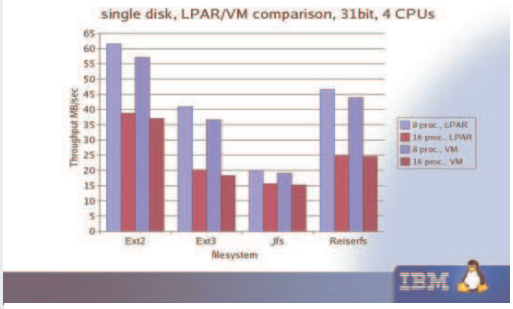
- There are a number of services in Linux that get started at boot, depending on:
  - Distribution
  - Version
  - Software selection at installation
- Shutting down unnecessary services and applications helps to improve the overall performance of the system.
  - Status of services can be queried/changed 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.
- If the security package is selected during install, SuSE Linux configures its cron daemon to run seccheck daily, weekly, and monthly at midnight.
  - Running many guests can cause high demand for cpu and stress the z/VM paging subsystem. Choices are to:
    - Remove from cron
    - Stager scheduled kick-off times

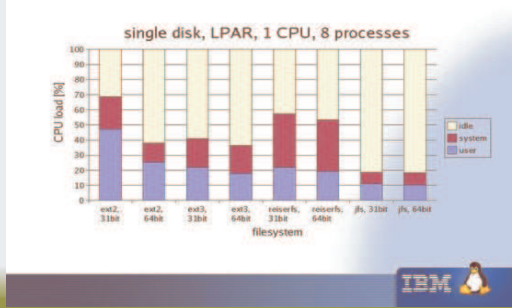
Disk Performance

- Hardware choices:
  - FICON verses ESCON
    - No comparison
      - ESCON 17MB
      - FICON available in 1Gb, 2Gb, and 4Gb channel speeds
  - SCSI verses ECKD
    - ECKD for z/VM and Linux "/" file system
    - SCSI for application data and databases
- Maximize hardware performance:
  - Configure maximum number of channel paths
  - Spread disks over different ranks within a storage server
  - Use logical volumes with striping
  - Consider exploiting PAV
- References:
  - http://www.vm.ibm.com/perf/reports/zvm/html/520lxd.html
  - http://www.ibm.com/developerworks/linux/linux390/perf/tuning_more_dasd_optimizedisk.html

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.
- Recommend using ext3 because of its journaling capabilities and reduced cpu load compared to other journaling file systems.



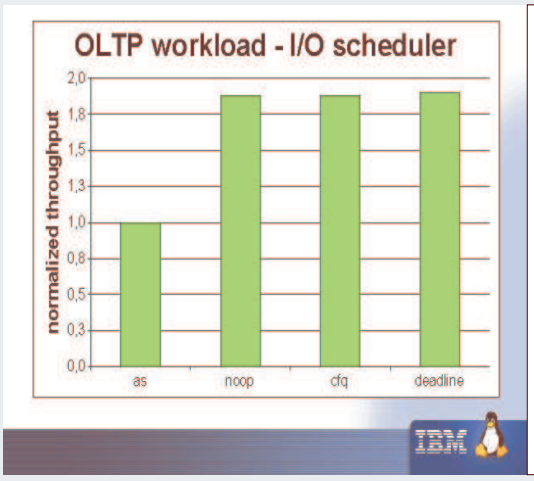


Kernel I/O Scheduler

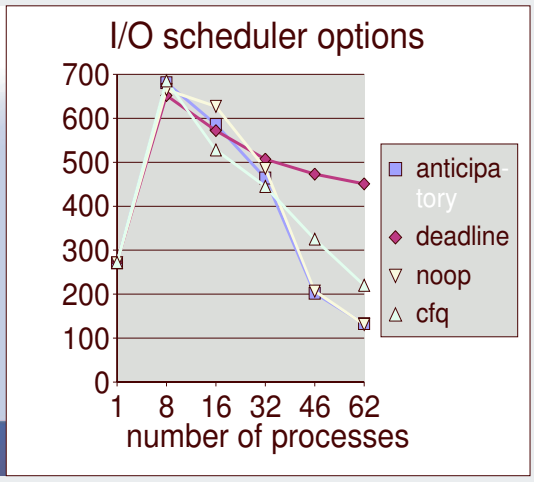
- The I/O scheduler optimizes disk access, the strategy for optimization aims to minimize the number of I/O operations and disk head movements.
- The Linux 2.6 kernel offers a choice of four different I/O schedulers:
  - Noop Scheduler (noop)
  - Deadline Scheduler (deadline)
  - Anticipatory Scheduler (as)
  - Complete Fair Queuing Scheduler (cfq)
- Linux default is the “as” scheduler:
  - Designed to optimize access to physical disks.
  - Not suitable for typical storage servers used in the zSeries environment, like the IBM ESS.
- Both Novell and RedHat zSeries distributions use the “cfq” scheduler by default.
- Selected by setting the “elevator” boot parameter in /etc/zipl.conf .
- Recommended I/O scheduler – deadline scheduler.

Kernel I/O Scheduler Measurements

Informix OLTP benchmark throughput relative to “as”



Dbench throughput MB/sec



Virtual Networking
“Best Practices”

Networking Configuration Options

- Three basic strategies for external network connectivity:
  - Dedicate OSA devices to Linux guests.
    - Can complicate network configurations.
    - Higher memory and cpu requirements.
  - Attach OSA devices to a virtual router.
    - Virtual router can be a bottleneck.
    - Higher cpu requirements.
  - Implement Virtual Switch (recommended).
    - Administration benefits.
    - Lower cpu costs.
    - Layer 2 or Layer 3 switching.
    - Link aggregation capabilities.
- Internal network connectivity:
  - HiperSockets for LPAR-to-LPAR communications.
- References:
  - z/VM Connectivity Manual - (SC24-6080)
    - http://publibz.boulder.ibm.com/epubs/pdf/hcsc9b20.pdf
  - Networking Overview for Linux on zSeries Red paper
    - http://www.redbooks.ibm.com/redpapers/abstracts/redp3901.html

System Health Check

- The z/VM and Linux on zSeries Advanced Technical Support Team offers a System Health Check for new Linux on zSeries accounts.
- Health Check Objectives:
  - Evaluate the basic system configuration of an installed z/VM system and its Linux guest(s).
  - Insure all recommended “ROT” and “Best Practices” are in place prior to running a POC benchmark or production workload.
- To schedule a System Health Check:
  - The client team FTSS should contact the z/VM and Linux on zSeries ATS Team
    - The FTSS will be provided with data collection documentation
    - The FTSS should work with the customer to collect the data
    - The FTSS should forward the data to ATS for evaluation
- After evaluation, the FTSS is provided with documentation containing:
  - Detailed analysis of current system configuration
  - Any recommended changes
  - Instructions for implementing changes

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 zSeries Performance Web Site
- Redbooks
  - http://www.redbooks.ibm.com/
    - Linux on IBM @server zSeries and S/390: Performance Toolkit for VM - (SG24-6059)
    - Linux on IBM @server zSeries and S/390: Performance Measurement and Tuning - (SG24-6926)
- z/VM Library
  - http://www.vm.ibm.com/library/
    - z/VM Performance - (SC24-6109)
    - z/VM V5R2 Performance Toolkit - (SC24-6136)
    - z/VM V5R3 Performance Toolkit Guide - (SC24-6156)
    - z/VM V5R3 Performance Toolkit Reference - (SC24-6157)