Tips Learned Implementing Oracle Solutions With Linux on IBM System z (Part I & II)

Christian Ehrhardt (ehrhardt@de.ibm.com)

&

David Simpson (simpson.dave@us.ibm.com)

Speakers Company: IBM

Date of Presentation: **1:30 & 3:00 PM, Thursday, August 15, 2013**

Room 310, Hynes Convention Center

Session Number: **13535 + 13556**

Twitter -> @IBMandOracle

http://linuxmain.blogspot.com/

Trademarks

Agenda

- **Hardware Setup**
- z/VM / LPAR
- Linux
- CPU
- Memory
- I/O
- Networking
- Oracle

Hardware setup - network

- Use latest network cards and attachments
  - today: OSA4
  - continuous improvements

- Plan for direct attached OSA cards for performance critical servers

- Define and use Hipersockets for LPAR-LPAR communication
  - Alternative: shared OSA4 10 GbE

OSA4 throughput improvements

LPAR 10 Gbit MTU 1492 - throughput improvement vs OSA3

Complete your sessions evaluation online at SHARE.org/BostonEval

OSA4 CPU savings

LPAR 10 Gbit MTU 1492 - CPU savings server vs OSA3

Legend: OSA4, OSA4 + GRO/TSO
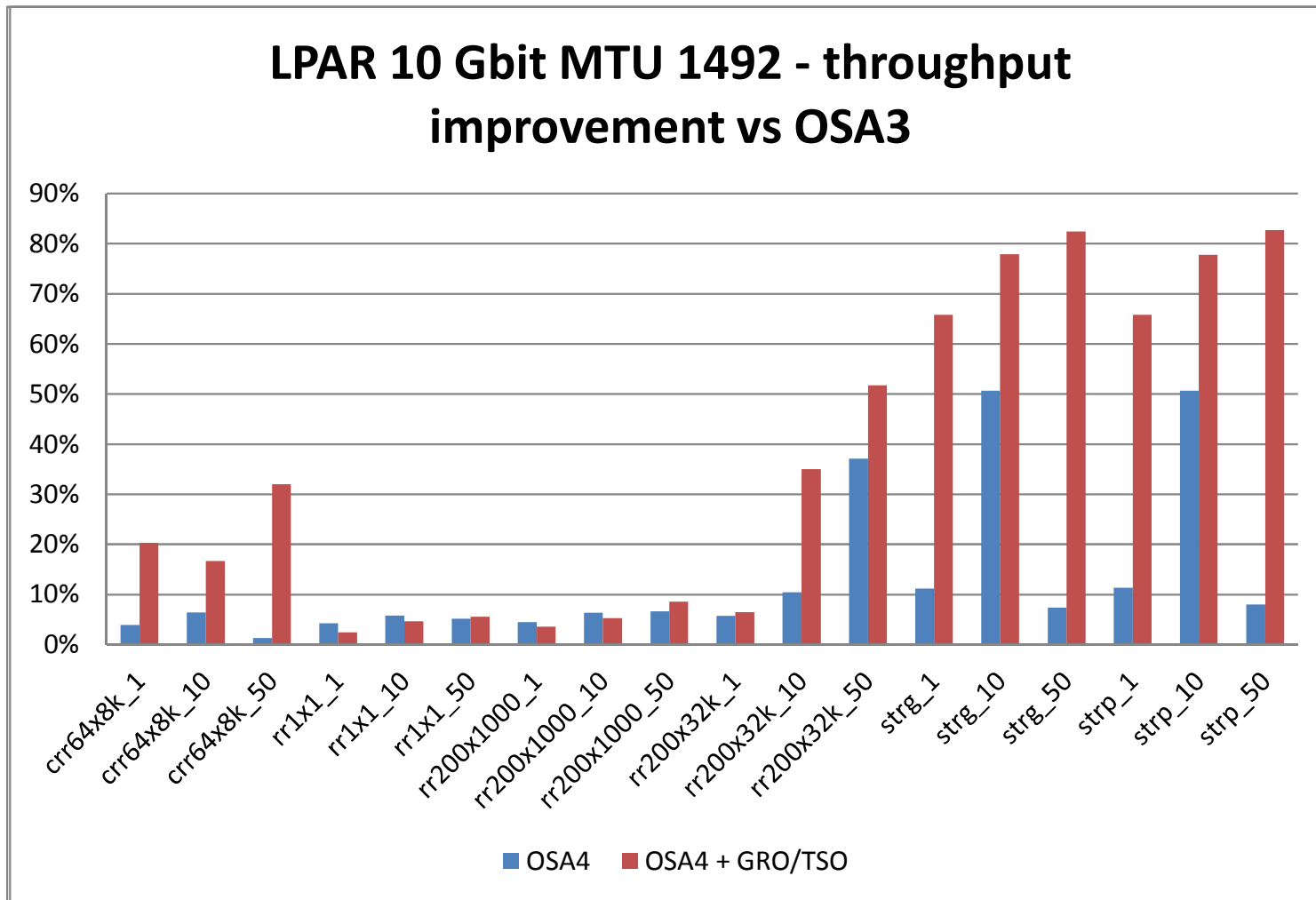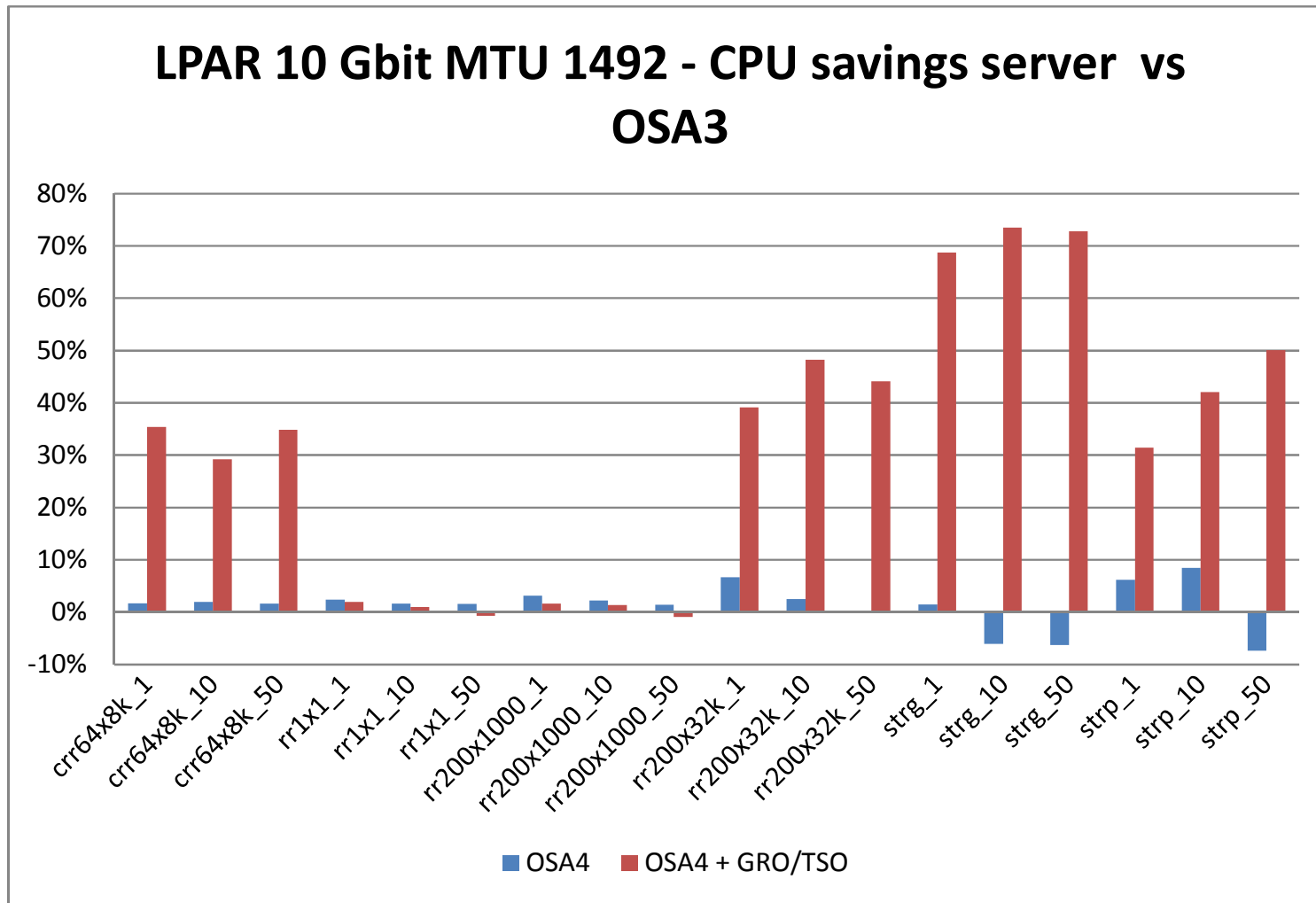
General I/O layout for FICON/ECKD

The dasd driver starts the I/O on a subchannel

Each subchannel connects to all channel paths in the path group

Each channel connects via a switch to a host bus adapter

A host bus adapter connects to both servers

Each server connects to its ranks
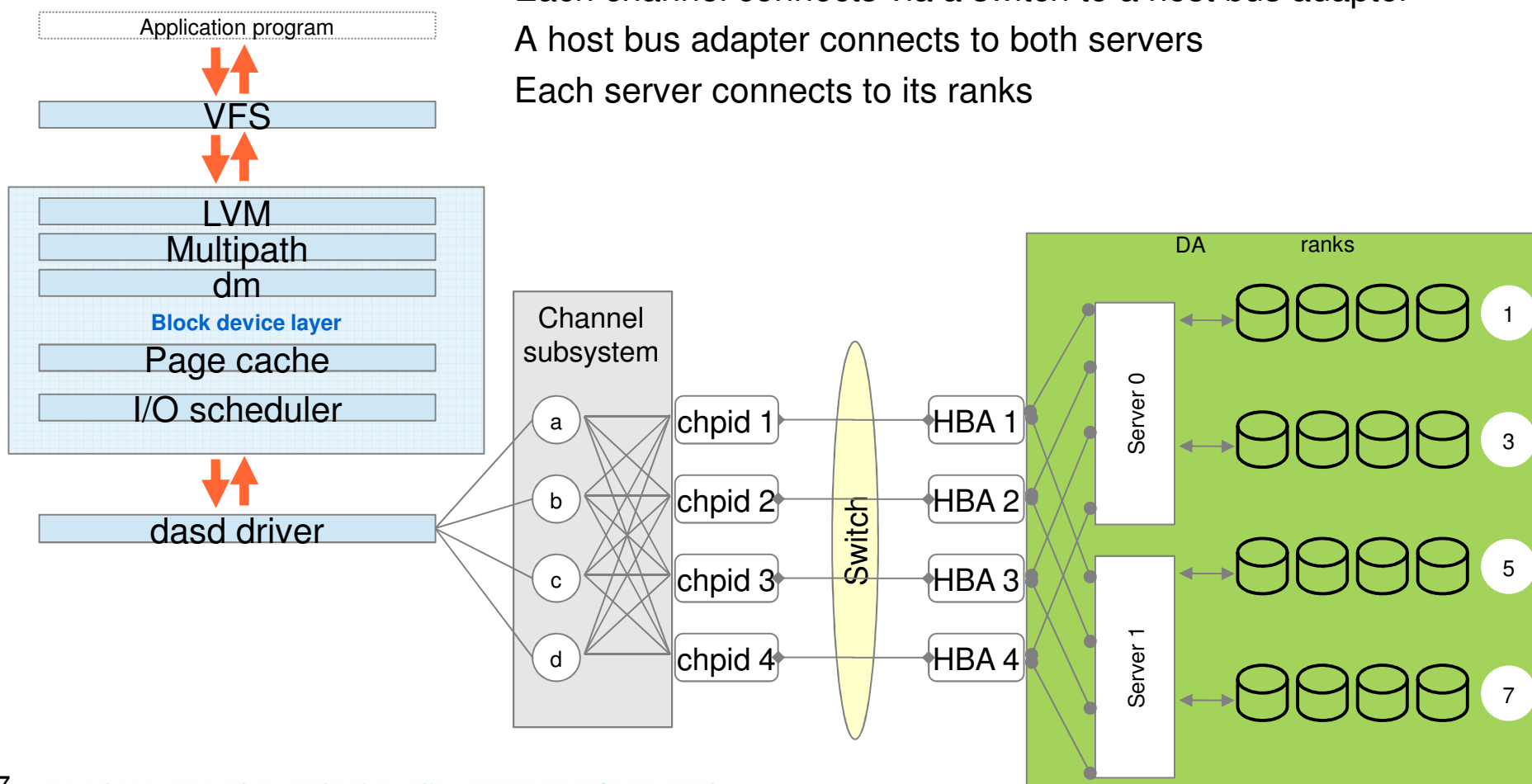
General I/O layout for FCP/SCSI

The SCSI driver finalizes the I/O requests

The zFCP driver adds the FCP protocol to the requests

The qdio driver transfers the I/O to the channel

A host bus adapter connects to both servers

Each server connects to its ranks
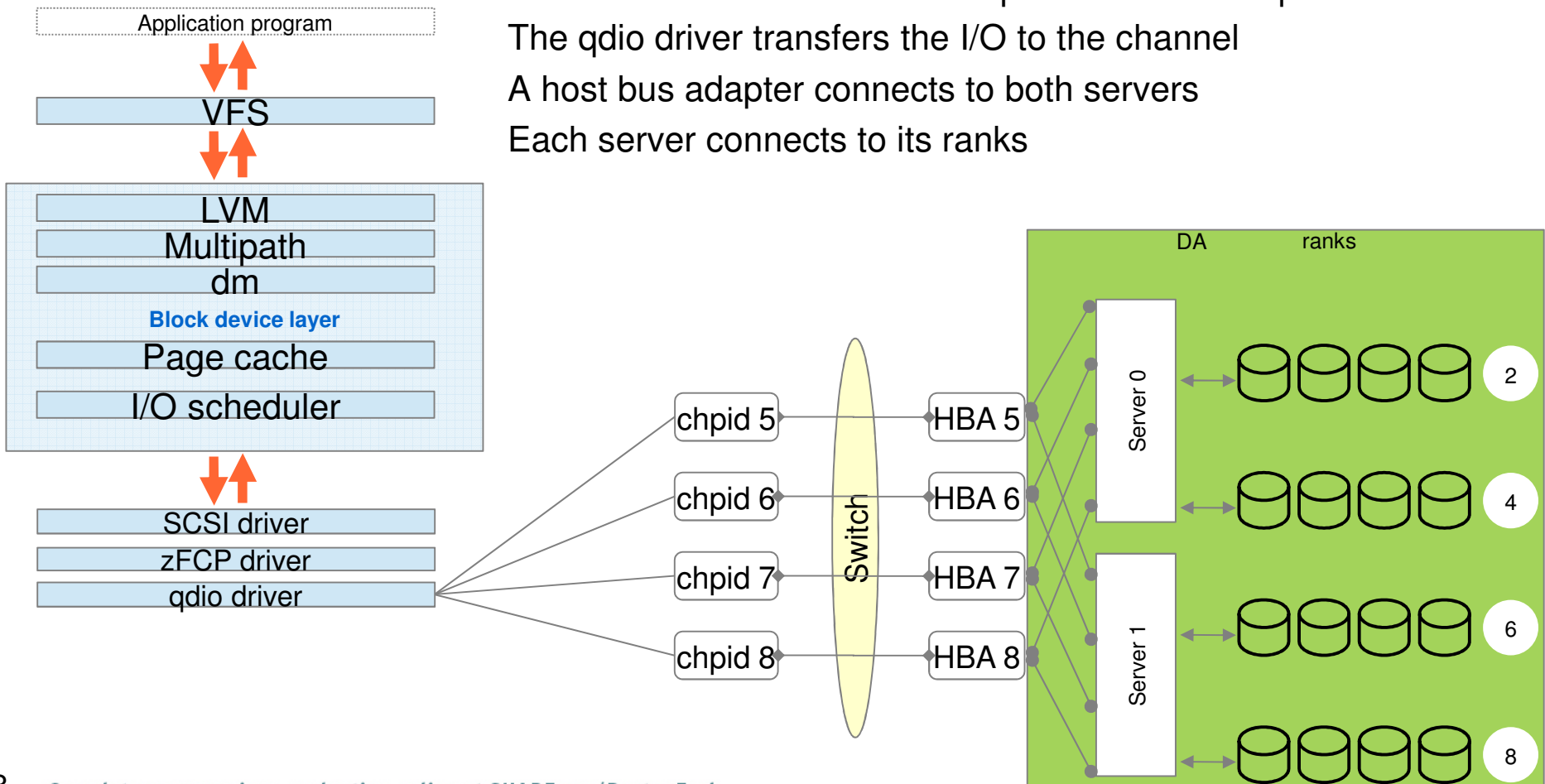
DS8000 storage pool striped volume (1)

A storage pool striped volume (rotate extents) is defined on a extent pool consisting of several ranks

It is striped over the ranks in stripes of 1 GiB

As shown in the example the 1 GiB stripes are rotated over all ranks

volume

1	4	7	rank 1
2	5	8	rank 2
3	6		rank 3
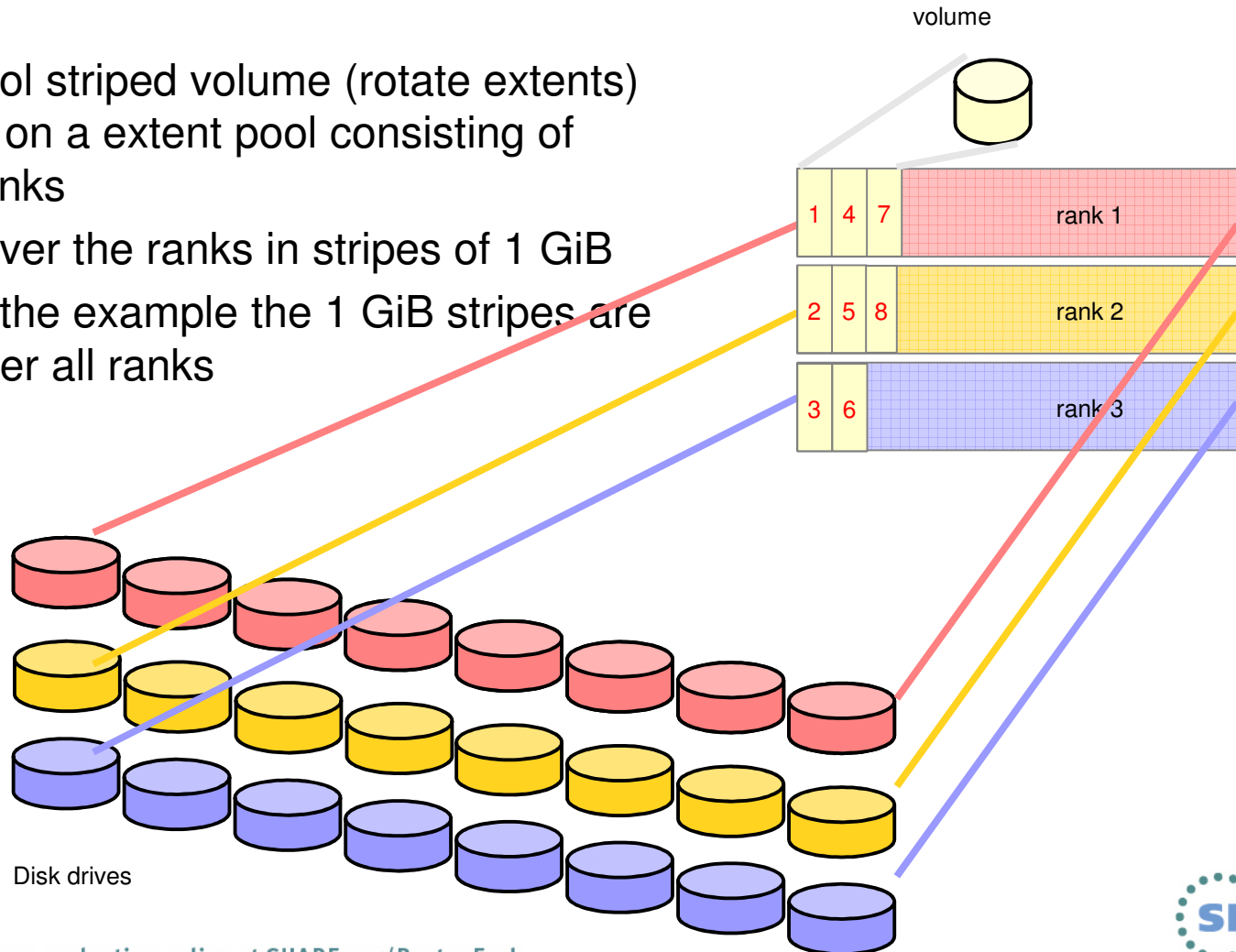
Disk drives

DS8000 storage pool striped volume (2)

A storage pool striped volume

- – uses disk drives of multiple ranks
- – uses several device adapters
- – is bound to one server

Device Adapter

volumes

ranks

Server 0

Server 1

Disk drives
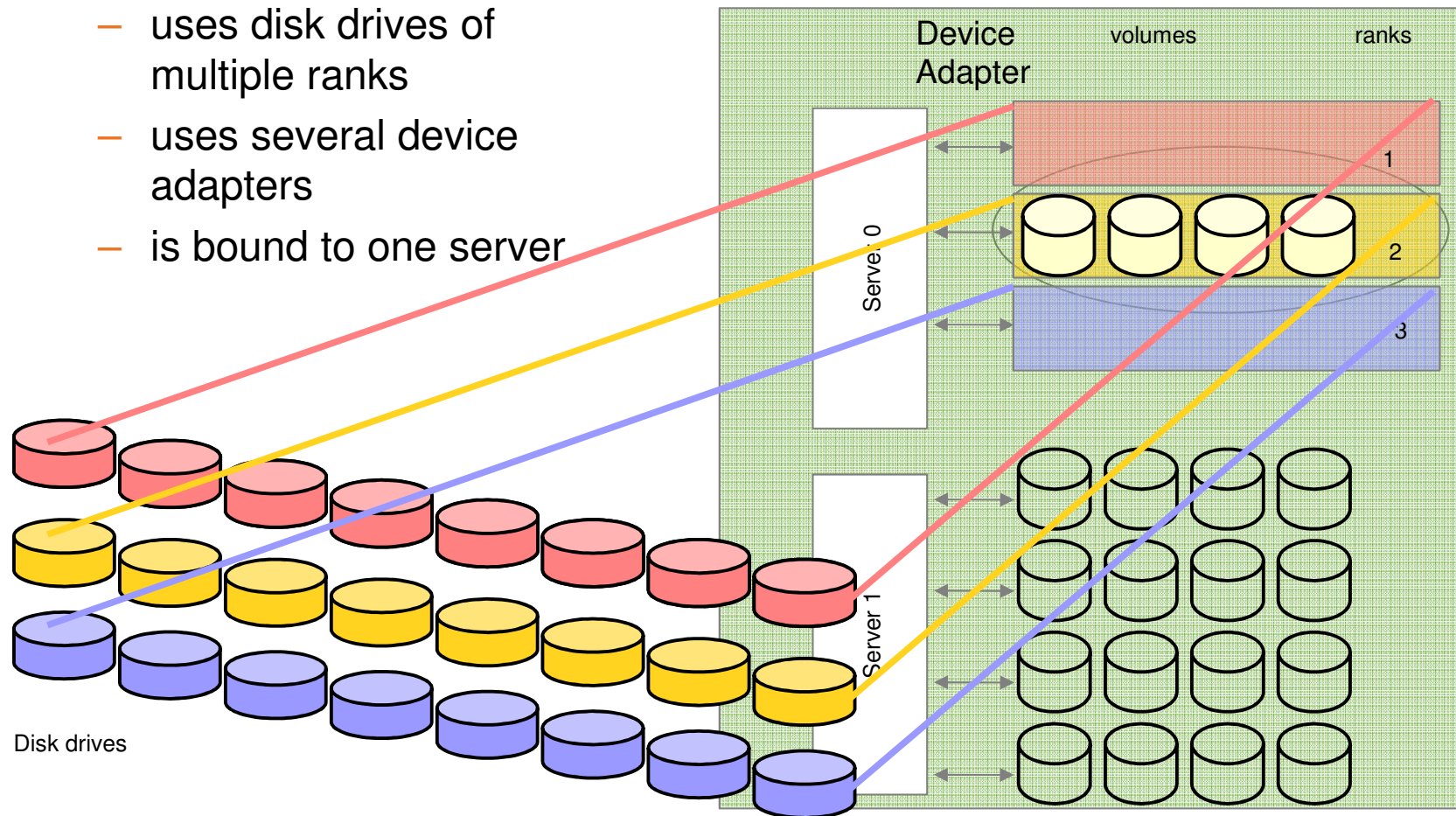
DS8000 storage pool striped volume (3)

	LVM striped logical volumes	DS8000 storage pool striped volumes
Striping is done by...	Linux (device-mapper)	Storage server
Which disks to choose...	plan carefully	don't care
Disks from one extent pool...	per rank, alternating over servers	out of multiple ranks
Administrating disks is...	complex	simple
Extendable...	yes	no "gluing" disks together as linear LV can be a workaround
Stripe size...	variable, to suit your workload (64KiB, default)	1GiB

Complete your sessions evaluation online at SHARE.org/BostonEval

Hardware setup - storage recommendations

- Keep as many parts busy at each level as you can
  - Multiple storage servers, CHPIDs, HBAs, ranks, spindles….

- Plan for capacity on each level!

- Use storage pool striping

Hardware setup – System z machine

- BC or EC?
  - Verify CPU capacity for growth available
  - Verify memory capacity for growth

- For new projects use latest technology!

- Always consider the full application technology stack.

Agenda

- Hardware Setup
- **z/VM / LPAR**
- Linux
- CPU
- Memory
- I/O
- Networking
- Oracle

z/VM reorder processing

- The cost of reorder is proportional to the number of **resident** frames for the virtual machine

- Delay of ~ 1s per 8 GB resident memory, the whole guest is stopped

- For details see: http://www.vm.ibm.com/perf/tips/reorder.html

- Recommendation: Turn reorder off for larger Oracle guests
  - SET REORDER OFF FOR …..

- Non issue with Re-order in z/VM 6.3 as memory management enhanced.

z/VM - qioassist

- Hardware assist to reduce Hipervisor overhead

- Enable for all FCP and OSA / Hipersocket channels

- Reduces the number of SIE exits
  - Shorter path length
  - Less cache pollution

z/VM – stay current and plan ahead

- z/VM 6.3 – "Making room to grow your business"
  - Support for 1 TB memory per LPAR
  - Reordering replaced
  - Support for HiperDispatch
    - Dispatching affinity!
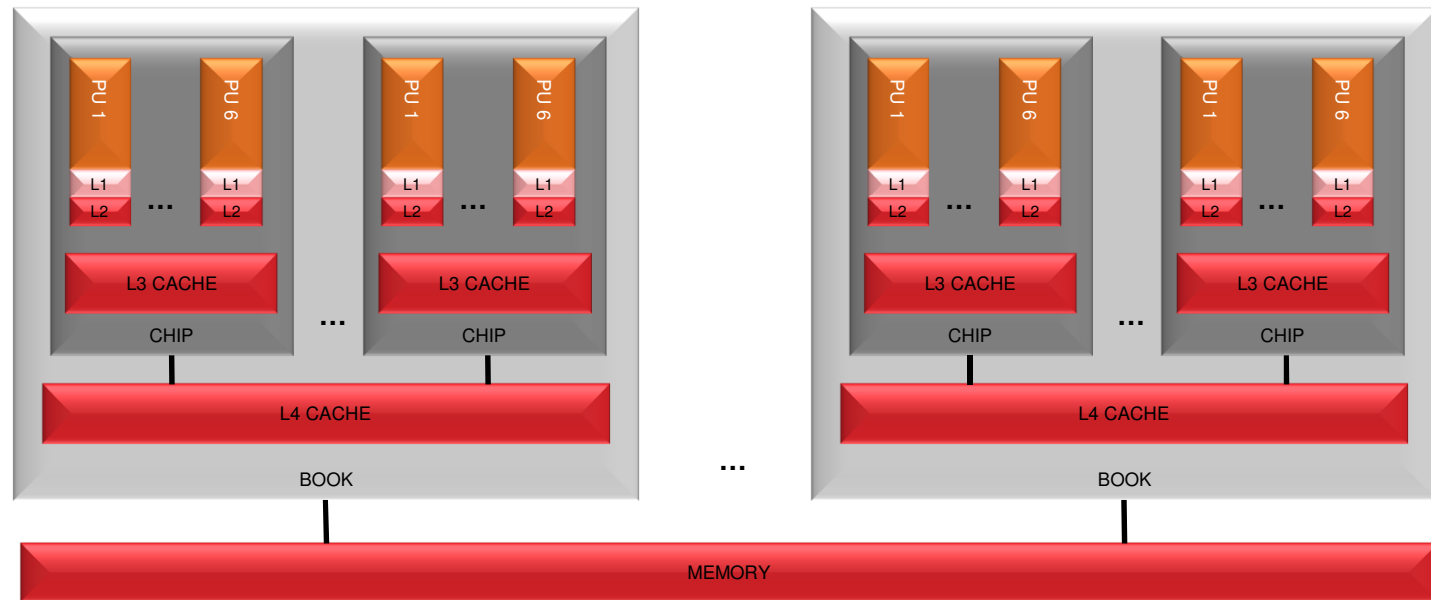
z/VM – monitor your system

- Collect z/VM performance data as default
  - http://www.vm.ibm.com/perf/tips/collect.html

  - Other tooling from ISVs / IBM works as well

- Really needed if debugging performance problems under z/VM

z/VM or LPAR

- Larger guests can monopolize a z/VM

- There is always some overhead with virtualization

- Some high end / high volume production is better placed in separate LPARs
  - Resource sharing still possible with the exception of memory

- However use z/VM for
  - Many low utilized guests
  - Test and development systems
  - Fast changing environments
  - Guests with (planned) peak workloads at different times
  - Memory over commit needed

Complete your sessions evaluation online at SHARE.org/BostonEval

Agenda

- Hardware Setup
- z/VM / LPAR
- **Linux**
- CPU
- Memory
- I/O
- Networking
- Oracle

Complete your sessions evaluation online at SHARE.org/BostonEval

Linux configuration

- Disable all not needed services
  - splash, postfix, nfs, ……

- Disable selinux
  - Kernel parameter selinux=0

- Disable cgroup memory
  - Kernel parameter cgroup_disable=memory
  - Saves 1% of memory per guest.

Oracle RPM checker

- Before you do first Oracle Install – run the Oracle rpm checker!

- Oracle Note -> **Getting Started - 11gR2 Grid Infrastructure, SI(Single Instance), ASM and DB (IBM: Linux on System z)** - (**1306465.1**)

- These rpms are "dummy" rpms that have dependency checks against all the required rpms for both Grid Infrastructure and Database installs.

- Must have an Oracle support ID to download

  RHEL5 - 11.2 Grid Infrastructure, SIHA, DB Install
  RHEL6 - 11.2 Grid Infrastructure, SIHA, DB Install
  SLES 10 - 11.2 Grid Infrastructure, SIHA, DB Install
  SLES 11 - 11.2 Grid Infrastructure, SIHA, DB Install

SLES 11 SP2+ & Red Hat 6.2+ – Oracle Install Warnings for Oracle 11.2.0.3

- You can "Ignore All" the following Oracle Installer Warnings

Some of the minimum requirements for installation are not completed. Review and fix the issues listed in the following table, and recheck the system.

Checks	Status	Fixable
Checks		
Swap Size	Ignored	No
Packages		
Package: libstdc++43-4.3.4_20091019-0.7.35 (s390x)	Ignored	No
Package: libgcc43-4.3.4_20091019-0.7.35	Ignored	No
Package: compat-libstdc++-33-3.2.3-47.3	Ignored	No

- SLES 11 SP1 compat-libstdc++-33.3.2.3-47.3 → rpm libstdc++-33 provides the required files.
- SLES 11 SP2 the libstdc++43 and libgcc43 → libstdc++46– **these are not problems.**

SLES 11 SP2/SP3 – New KVM Service Conflict

- Oracle 11gR2 (ASM Single Instance & RAC) may encounter a conflict with the SuSe KVM service in the " **/etc/inittab** " file for fresh SLES 11 SP2+ installs (Upgrades from SP1 OK):

---

h1:35:respawn:/etc/init.d/init.ohasd run >/dev/null 2>&1 </dev/null **- Installed by Oracle**

h1:2345:respawn:/sbin/ttyrun hvc1 /sbin/agetty -L 9600 %t linux   **- Default KVM service**

- Details see My Oracle Support (MOS) – Note: **1476511.1**

ASM or LVM

- LVM – Logical Volume Manager in Linux
- ASM – Automated Storage Management provided by Oracle
  - Oracle RAC One and Oracle RAC will require ASM

	LVM	ASM
pro	• Direct control on setting and layout • Can choose file system	• Automated, out of the box environment • Very good integration with Oracle
con	• Complex setup	• RMAN required for backup

- Overall recommendation: **ASM**
- **Don't combine both!**

Database files on Filesystem: Disable read ahead & Use Async / direct I/O

- Oracle parameter file systems: **filesystemio_options=**setall
  - Provides asynchronous & direct I/O (avoids linux file system cache)

- Reduce Linux Read-Ahead for LVM file systems.
  - **lvchange -r none <lv device name>**

Scaling users with swingbench 2.4

CompletedTransaction, Runtime 33:32 - 34:33, sales history

CPU utilization — completed Transaction
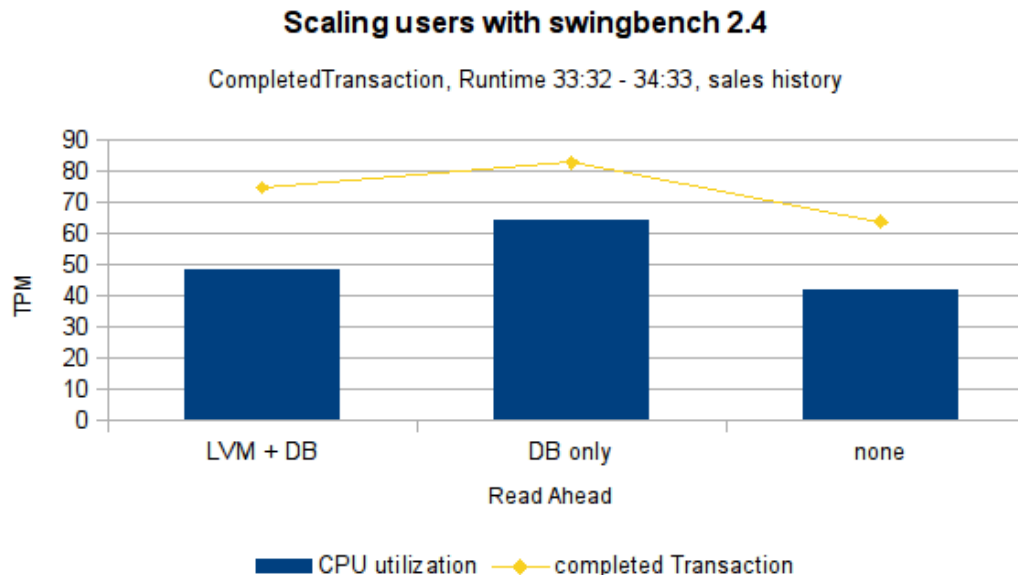
File System Types & Noatime mount option

- Recommended file system types:
  - SuSe – ext3 file system recommended for database files.
  - Red Hat 6.x - ext4 recommended for database files.
  - Reiser (the default) does not perform that well with Oracle databases.

- Linux ext3/4 file systems keep track of different times: access, modify, change time
- Problem is nearly always the access time
  - Updated at every read
  - Only very few applications make use of that
  - Turned on by default
- For any Oracle file systems on ext3/4 mount them with the noatime, (nodiratime) options
  - Reduces CPU usage
  - Reduces # IOs
- See Oracle Note **1561740.1**

Linux paging / swappiness

- With the default swappiness setting of 60 Linux does proactive paging

- Oracle data / code on a Linux (or VM) paging disk has a performance hit when it's needed
  - Observed long (>10s) waits at swap in
  - Guest was sized correctly
  - Guest was using database in the file system without direct I/O

- Recommendation: set swappiness to zero
  - In /etc/syctl.conf add `vm.swappiness=0`

- Largepages are ineligible for swapping.

Collect Linux performance data

- Standalone performance collection in Linux is sysstat
  - http://sebastien.godard.pagesperso-orange.fr

- For standard monitoring use same interval as for your z/VM monitoring

- Always monitor your system

- Include monitoring for disks (default off)

- http://linuxmain.blogspot.com/2011/12/gathering-performance-data-with-sysstat.html

Stay current with your Linux updates

- Check updates for performance enhancements
  - RHEL 5.9
    - VDSO
    - HyperPAV
  - SLES 11 SP2+
    - GRO / TSO

- Security updates need to be considered as well

Agenda

- Hardware Setup
- z/VM / LPAR
- Linux
- **CPU**
- Memory
- I/O
- Networking
- Oracle

Complete your sessions evaluation online at SHARE.org/BostonEval

Sizing Consolidated CPU consumption – equivalent MIPS

October 2012 - equivalent MIPS (wo z/VM)

2937 MIPS

Legend: oralnx01, oralnx02, oralnx03, Total Mips Prime Shift, Total Mips Night
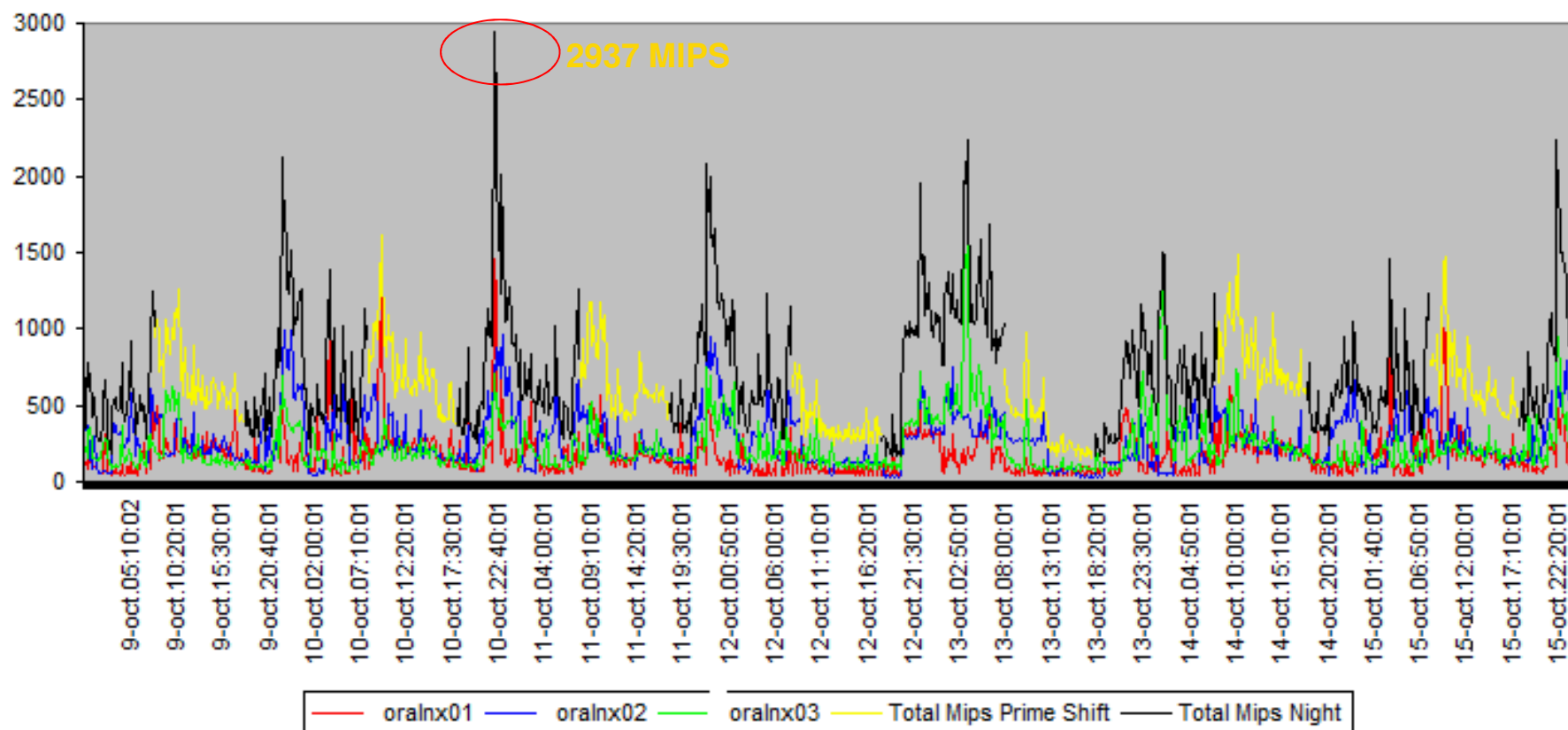
Monitoring CPU Run Levels / Oracle Parallel Query

Watch the run queue!

```
# vmstat 3      (on virtual 2 way machine)
procs -----------memory---------- ---swap-- -----io---- -system-- -----cpu------
 r  b   swpd   free   buff  cache   si    so    bi     bo    in    cs us sy id wa st
 4  0 276900 286468   1164 468472    0     0     5     26     7     8  0  0 100 0  0  ← Typically Ignore 1st
 1  0 276896 284772   1256 468900    0     0   267     76   257   760 43  7  49 1  0
 2  0 276888 272052   1392 470320    0     0   475    107   218   439 47  4  47 1  2
 3  0 275672   8988   1228 464564  277 42971  1224  47888  1332   350 67 11   0 15  6
 2  0 273636   8884    652 489576  524     3   889  20575   397   321 59  4  37 0  1
 1  0 271560   8580    788 536964  599     5   984  29069   470   255 61  3  34 1  1
 1  0 267576   8732   1068 591056 1412     0  3772  31208   796   696 50 11  22 16  1
 6  5 283124   6168    240 586176  299  5451  2148  17865  1220   528 15 24   6 53  1
 0  8 307192   5840    432 614808  437  8451 12868  26735  1249   575 14 21   2 59  4
16 12 307192   6668    136 572948    3    17 46792    701  1744   963  0 87   0 13  1
15 15 307192   7796    120 570384    0     0 13271      0   393   188  0 99   0  0  1
```

- r –run queue –how many processes currently waiting for CPU
  - try to keep < # of Virtual IFLs for Oracle Parallel Query
- b – how many processes waiting in uninterruptible sleep
- Steal time (**st**) is the percentage of time a virtual CPU waits for a real CPU while the hypervisor is servicing another virtual processor.

Oracle Parallelism

Default Value:
**PARALLEL_MAX_SERVERS =
(CPU_COUNT x PARALLEL_THREADS_PER_CPU x 10)**

- If too many query server processes, memory contention (paging), I/O contention, or excessive context switching can occur

- Contention can reduce system throughput to a level lower than if parallel execution were not used.

- Can utilize **Oracle Consumer Group** to limit processes for certain types of users/jobs

CPUPLUGD

- CPUPLUGD Daemon can be configured to add or reduce the number of Virtual processors based on the load

- Oracle dynamically changes the Oracle internal parameter "**cpu_count**" based on the number of Virtual processors available.
  - This should be the default!

- Explicitly setting cpu_count will disable the automatic adaption of Oracle DB to cpuplugd changes

- CPUPLUGD configuration recommendations
  - Need fast sampling interval (1s)
  - Create sensitive configuration for CPU add

VDSO – Linux cpu Improvements

- **V**irtual **D**ynamically-linked **S**hared **O**bject (VDSO) is a shared library provided by the kernel. This allows normal programs to do certain system calls without the usual overhead of system calls like switching address spaces.

- On a z196 system for example by using the VDSO implementation **six times** reduction in the function calls are possible.

- Newer Linux distributions (RHEL 5.9 & 6.x, SLES 11) have this feature and it's enabled by default.

- Oracle calls Linux **gettimeofday()** hundreds of times a second for reporting statistics.

- VDSO reduces cpu cost, especially useful in virtualized environments

Java on System z – 15 Years of Innovation

31-bit and 64-bit Java 6
1. J9/Testarossa Technology
2. GA 4Q2007
3. z/OS and Linux on z

31-bit and 64-bit Java6.0.1/626/7.0
1. J9/TR Technology
2. GA 1Q2011/4Q2011
3. z/OS & zLinux (7.0 only)

31-bit and 64-bit Java 5
J9/TR Technology
GA 4Q2005
z/OS and Linux on z

z/OS 64-bit Java 1.4.2
J9/TR Technology (1st product use)
GA 4Q2004

Java 1.4
GA 4Q2002
31-bit z/OS and 31-bit and 64-bit Linux on z

31-bit Java 1.3.1
z/OS and Linux on z
GA 3Q2000

31-bit Java 1.1.8
GA 1999

31-bit Java 1.1.1, then 1.1.4 and 1.1.6
First z/OS Java product – GA 1997

2012
2011
2009
2007
2005
2004
2003
2001
2000
1999
1998

WAS6.0 JEE1.4
WAS6.1 JEE1.4
WAS7.0 JEE5
WAS8.0 JEE6
WAS8.5 JEE6

IBM continues to invest aggressively in Java for System z, demonstrating a rich history of innovation and performance improvements.

Testimonials: http://www-01.ibm.com/software/os/systemz/testimonials/
http://www.centerline.net/review/#/3332_B

Timelines and deliveries are subject to change.

Linux on System z and Java7SR3 on zEC12:
64-Bit Java Multi-threaded Benchmark on 16-Way

Linux on System z - Multi-Threaded 64 bit Java Workload 16-Way
~12x Improvement in Hardware and Software

Chart: Normalized Throughput (y-axis, 0 to 160) vs Threads (x-axis: 1, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32)

Legend:
- zEC12 SDK 7 SR3 Aggressive + LP Code Cache
- zEC12 SDK 7 SR1
- z196 SDK 7 SR1
- z196 SDK 6 SR9
- z10 SDK 6 SR4
- z9 SDK 5 SR4 NO (CR or Heap LP)
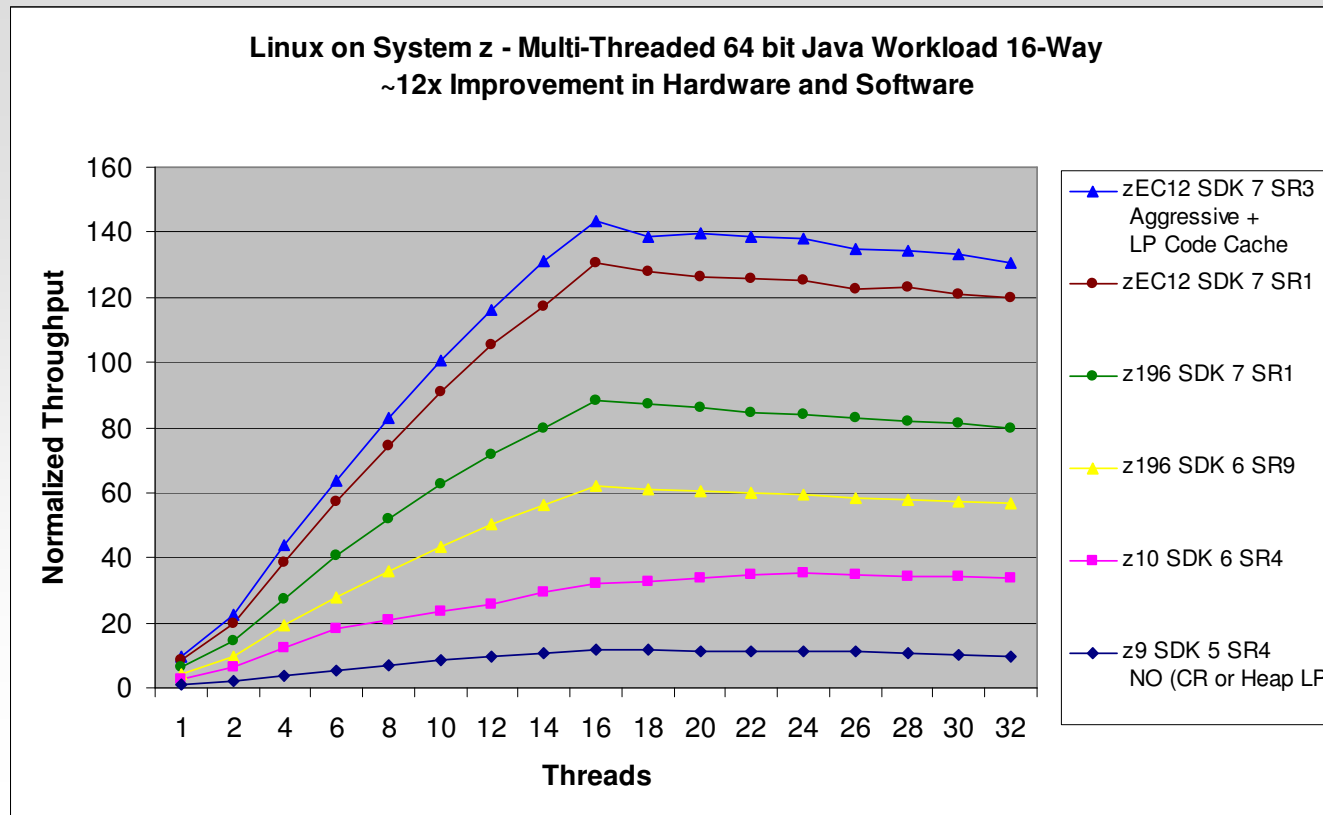
~12x aggregate hardware and software improvement comparing Java5SR4 on z9 to Java7SR3 on zEC12
   LP=Large Pages for Java heap   CR= Java compressed references
   Java7SR3 using -Xaggressive + 1Meg large pages

SHARE in Boston

(Controlled measurement environment, results may vary)

Agenda

- Hardware Setup
- z/VM / LPAR
- Linux
- CPU
- **Memory**
- I/O
- Networking
- Oracle

Complete your sessions evaluation online at SHARE.org/BostonEval

Memory Sizing Oracle on System z Linux and 11gR2

- Customer attempted install 11gR2 with 512mb – **could not re-link on install.**
  - Oracle recommends **4GB** for all Linux Platforms, **smallest we would suggest is 2GB of Virtual Memory for a Single Oracle 11g instance.**

- One customer experienced 200 MB more RAM consumption 10gR2 to 11gR2

- **Right Size** the Virtual Memory based on What is needed:
  - **All SGA's (including ASM)** – consider Large Pages
  - **Oracle PGA's** (not eligible for Large Pages**)**
  - **User Connections** to the database (4.5mb – not eligible)
  - **Linux Page Tables** and **Linux Kernel Memory**
  - Try NOT to oversize the Linux Guest under z/VM, use VDISKs

- Production workloads 1 to 1.5:1 Virtual to Physical Memory, for Test and Dev 2 to 3:1 are possible.

Swap Sizing Oracle on System z Linux and 11gR2

- Example of VDISK for **1st** and or **2nd** Level Swap with higher priority and then DASD as a lower priority swap in case of an unexpected memory pattern

```
# swapon -s
Filename                        Type            Size      Used    Priority
/dev/dasdo1                     partition       131000    0       10
/dev/dasdp1                     partition       524216    0       5
/dev/mapper/u603_swap3          partition       6291448   0       1
```

- You may want to recycle the swap from time to time to free swap slots (check swapcache in /proc/meminfo)
  - Ensure there is enough memory (e.g. at night)
  - drop caches
  - swapoff / swapon

Linux Huge Pages

- **Consider Using Linux Huge Pages for Oracle Database Memory**
    - →In general 10-15% can be gained by the reduction in CPU usage as well as having a lot more memory for applications that would be consumed in Linux Page Tables…

/proc/meminfo – customer example (before)

```
MemTotal: 82371500 kB           Writeback: 0 kB
MemFree: 371220 kB              AnonPages: 2743884 kB
Buffers: 4956 kB               Mapped: 48976112 kB
Cached: 50274732 kB            Slab: 243944 kB
SwapCached: 2248480 kB         PageTables: 26095124 kB
Active: 53106388 kB           NFS_Unstable: 0 kB
Inactive: 2164644 kB          Bounce: 0 kB
HighTotal: 0 kB               CommitLimit: 57594252 kB
HighFree: 0 kB                Committed_AS: 62983256 kB
LowTotal: 82371500 kB         VmallocTotal: 4211073024 kB
LowFree: 371220 kB            VmallocUsed: 12028 kB
SwapTotal: 16408504 kB        VmallocChunk: 4211060796 kB
SwapFree: 9834092 kB          HugePages_Total: 0
Dirty: 468 kB                 HugePages_Free: 0
                              HugePages_Rsvd: 0
                              Hugepagesize: 2048 kB
```

Complete your sessions evaluation online at SHARE.org/BostonEval

/proc/meminfo – customer example (after)

```
MemTotal:        82371500 kB      Writeback:            108 kB
MemFree:          7315160 kB      AnonPages:        3241568 kB
Buffers:           352624 kB      Mapped:            170176 kB
Cached:          12824152 kB      Slab:              439912 kB
SwapCached:             0 kB      PageTables:        318848 kB
Active:           4000920 kB      NFS_Unstable:           0 kB
Inactive:        12309216 kB      Bounce:                 0 kB
HighTotal:              0 kB      CommitLimit:     30802308 kB
HighFree:               0 kB      Committed_AS:     6001276 kB
LowTotal:        82371500 kB      VmallocTotal:  4211073024 kB
LowFree:          7315160 kB      VmallocUsed:        13032 kB
SwapTotal:       18456496 kB      VmallocChunk:  4211059808 kB
SwapFree:        18456496 kB      HugePages_Total: 28164
Dirty:                504 kB      HugePages_Free:   1208
                                  HugePages_Rsvd:   1205
                                  Hugepagesize:      2048 kB
```

HugePage Considerations:

- Can not use **MEMORY_TARGET** with Huge Pages.
  - Set manually to **SGA_TARGET** not including the **PGA_AGGREGATE_TARGET**.

- Not swappable: Huge Pages are not swappable

- General guideline consider when combined Oracle SGA's are greater than **8 GB** (particularly if a lots of connections)

- Decreased page table overhead; more memory can be freed up for other uses. i.e. more Oracle SGA memory, and less physical I/O's (See also Oracle Note: **361468.1**)

Recommendation: Use Huge Pages - even under z/VM

- Under z/VM (which has 4K pages) it's still recommended to use Huge Pages for SGA's > 10GB particularly with many connections

- Saves Memory that would otherwise be used for pagetables

- Stability for user process spikes (avoiding swap)

- Less work to manage smaller number of pagetables

- ~10% improvement (ROT) for memory intensive databases

Agenda

- Hardware Setup
- z/VM / LPAR
- Linux
- CPU
- Memory
- **I/O**
- Networking
- Oracle

Verify I/O Performance with Oracle Orion

- Oracle ORION Simulates Oracle reads and writes, without having to create a database

- No Longer Download from Oracle – it is now included with Oracle Code in $ORACLE_HOME/bin/orion

```
./orion_zlinux –run oltp –testname test –num_disks 2 –duration 30 –simulate raid0
ORION VERSION 11.2.0.0.1
Commandline: –run oltp –testname mytest –num_disks 2 –duration 30 –simulate raid0
This maps to this test: Test: mytest
Small IO size: 8 KB Large IO size: 1024 KB
IO Types: Small Random IOs, Large Random IOs
Simulated Array Type: RAID 0   Stripe Depth: 1024 KB
Write: 0% Cache Size: Not Entered
Duration for each Data Point: 30 seconds
Small Columns:,      2,      4,      6,      8,     10,     12,     14,     16,     18,
   20,     22,     24,     26,     28,     30,     32,     34,     36,     38,     40
Large Columns:,      0 Total Data Points: 22
Name: /dev/dasdq1       Size: 2461679616
Name: /dev/dasdr1       Size: 2461679616
2 FILEs found.
Maximum Small IOPS=5035 @ Small=40 and Large=0
Minimum Small Latency=0.55 @ Small=2 and Large=0
```

Kernel I/O Scheduler

- 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)

- General Linux default is the "cfq" scheduler:
  - Designed to optimize access to physical disks
  - Check in `/sys/block/<device>/queue/scheduler`
    ```
    noop anticipatory [deadline] cfq
    ```
  - Not suitable for typical storage servers
  - Default configurable by setting the "elevator=[…]" boot parameter in /etc/zipl.conf

- Recommendation: **deadline or noop**

HyperPAV (1)

- HyperPAV allows multiple IO operations on the same sub channel
- Very important for random access workload with relative small data transfers
- 10-20 HyperPAV aliases per LCU showed best performance gains in our internal tests.

- Recommendation:
    - **Enable** HyperPAV whenever using ECKD devices
    - Don't use too many aliases

  lsdasd –u

```
0.0.75bc  dasda     IBM.750000000NA461.1555.bc.000000000000eac50000000000000000
0.0.75ca  dasdk     IBM.750000000NA461.1555.ca.000000000000eac50000000000000000
0.0.75d0  alias     IBM.750000000NA461.1555.xx.000000000000000000000000000000
0.0.75d1  alias     IBM.750000000NA461.1555.xx.000000000000000000000000000000
```

HyperPAV (2)

ECKD Devices: Scaling HyperPAV aliases

Normalized Transactional throughput and total Disk I/O (read + write)
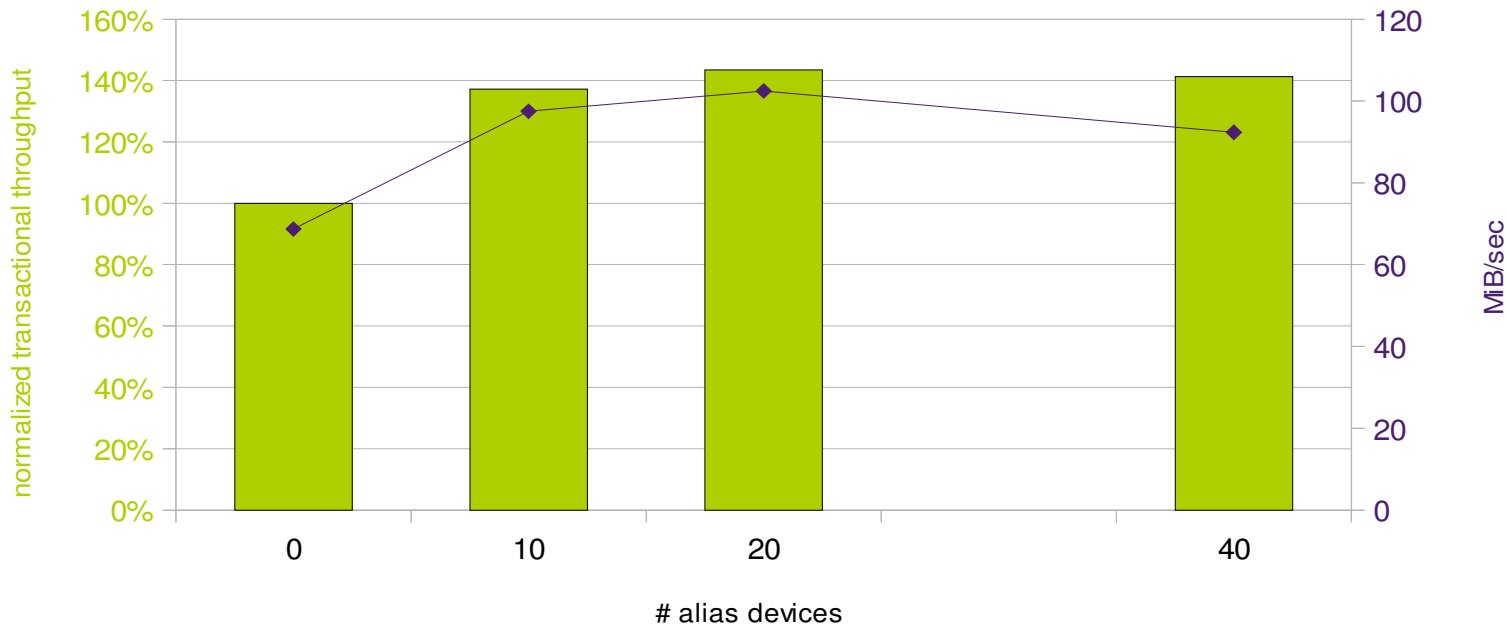
FICON / FCP comparison (1)

normalized transactional throughput

Legend: ■ FICON (20 aliases)   ■ FCP (rr_min_io=100)

Chart axis labels: 160%, 140%, 120%, 100%, 80%, 60%, 40%, 20%, 0%

X-axis categories: 4 CPUs, 6 CPUs
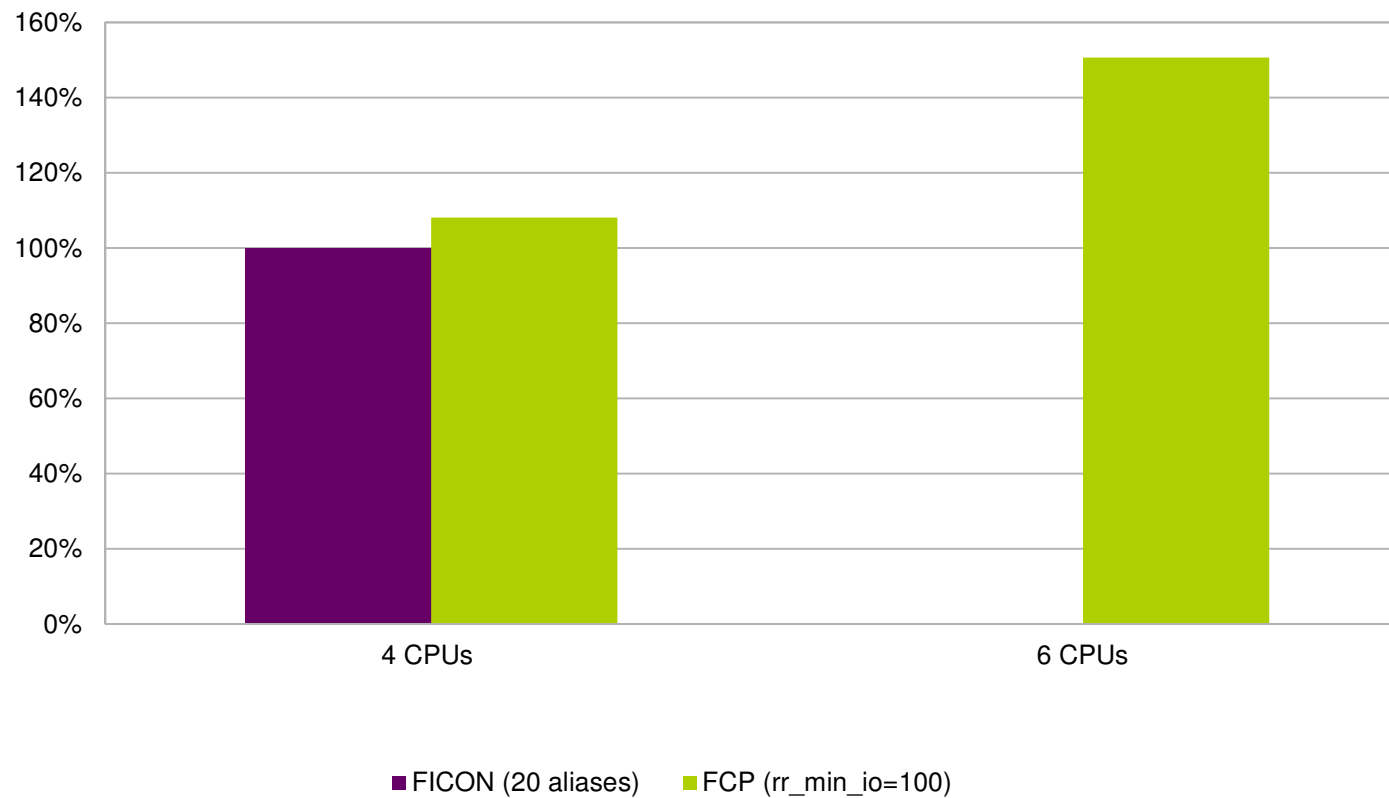
FICON / FCP comparison (2)

Normalized CPU cost per transaction
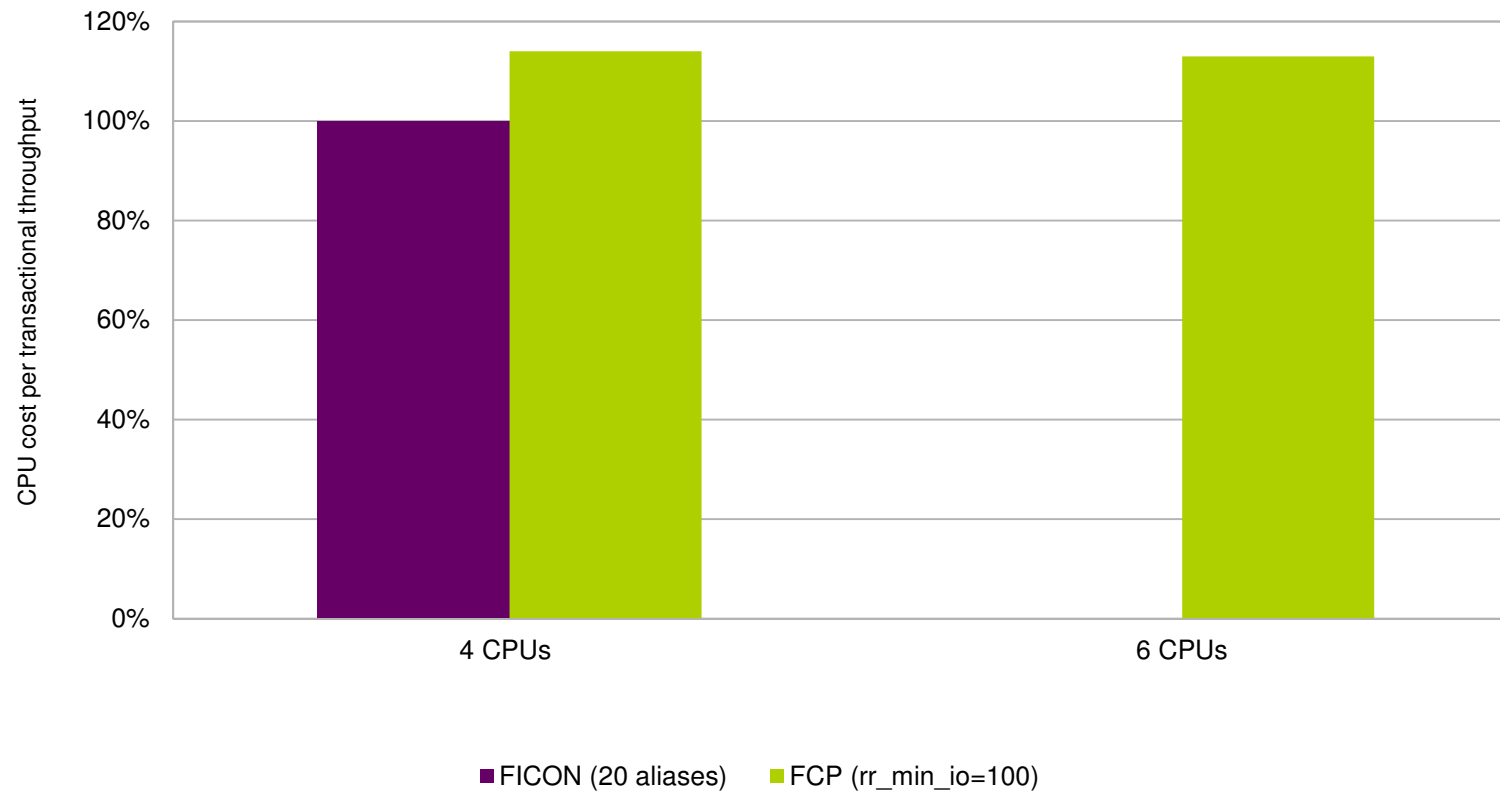
FICON / FCP comparison (3)

- FCP offers better throughput and performance
- FICON uses less CPU per transaction
- You have to tune both environments

- Recommendation: it depends

Linux multipathing – rr_min_io / rr_min_io_rq

- For FCP attached devices multipathing is needed for availability
  - Guidance for SLES11 + RHEL6 is to use multibus option in multipath.conf.

- rr_min_io defines the number of I/O operations that are send to path before switching to the next (round robin)
  - Defined in multipath.conf
  - In RHEL6.2+ / SLES11 SP2+ → now called `rr_min_io_rq`
    - Device mapper does accept old rr_min_io – but does nothing with it
    - Watch for upgrades!
- The **rr_min_io** value is storage dependent
  - For DS8K rr_min_io=100 provided good results
  - XIV recommends rr_min_io=15

Linux queue_depth

- Default of 32 generally pretty good
  - Set in /sys/bus/scsi/devices/*<SCSI device>*/queue_depth

- Reasons to decrease value:
  - Latency problems (pretty rare)
  - Storage subsystem overload

- Reasons to increase value:
  - System with heavy I/O load
  - Storage vendor suggestion / recommendation

- Use with care, due to the overload problem

Separate Redo log files from database (1)

- Conflicting kind of I/O
  - Logs are large sequential writes (good to optimize)
  - Normal database workloads are many small random read / writes

- Storage subsystem can't optimize if everything put together

- Watch Oracle events "log file sync" and "log file parallel write"

- **Recommendation:** put logs and data in different ASM disk groups

Separate Redo log files from database (2)

Data and Logs - Disk Setup

Complete your sessions evaluation online at SHARE.org/BostonEval
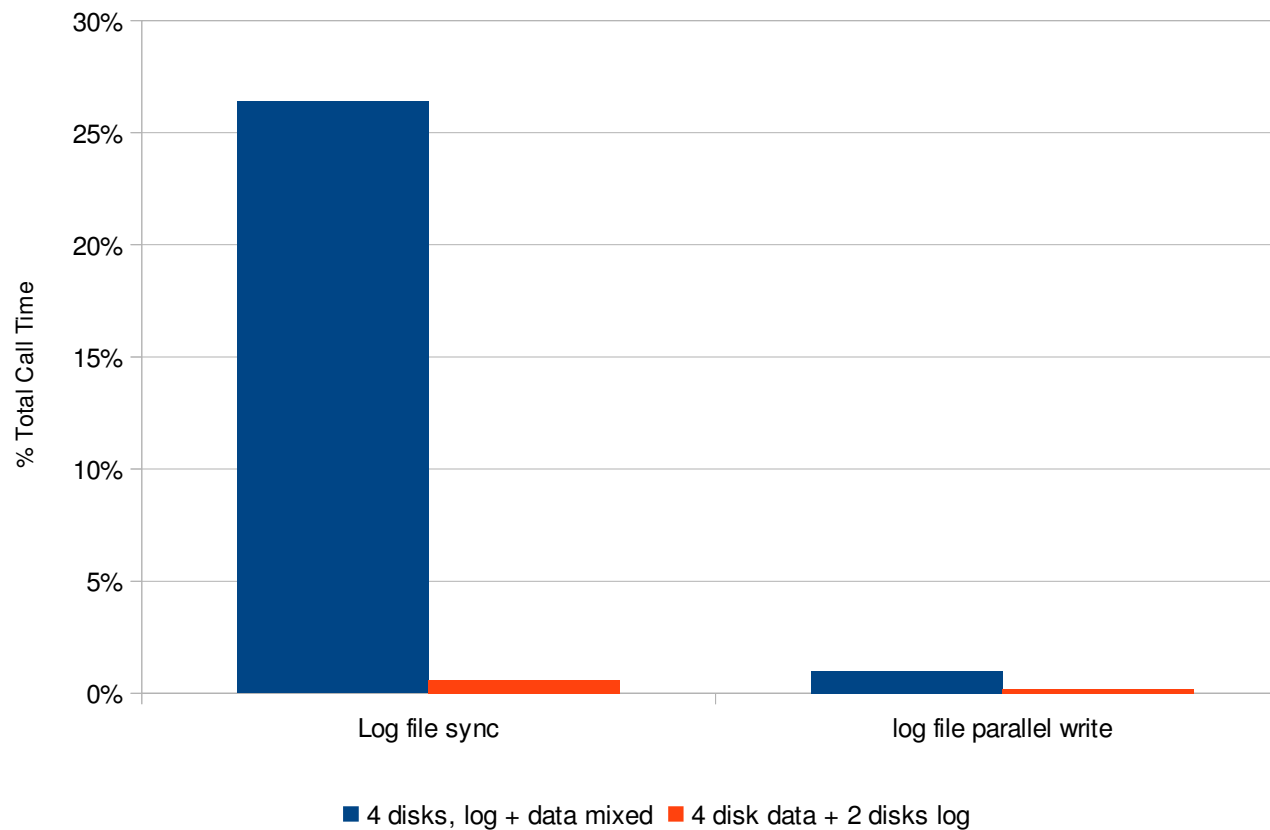
AWR Analysis for an I/O Bound System - Recommendations

- ## Summary

  - I/O bound AWR's can benefit from IBM FlashSystem in the following way:
    - **21x** reduction in response times
    - **272%** improvement in CPU utilization
    - **957%** improvement in IO wait time

Top 5 Timed Foreground Events

Event	Waits	Time(s)	Avg wait (ms)	% DB time	Wait Class
db file sequential read	16,543,046	126,308	8	47.26	User I/O
log file sync	453,034	81,895	181	30.64	Commit
DB CPU		42,771		16.00	
local write wait	42,921	2,858	67	1.07	User I/O
db file parallel read	94,662	1,847	20	0.69	User I/O

- ## Other Recommendations

  - Rerun AWR reports after implementing IBM FlashSystem to validate outcomes
  - Use application partitioning on the app-tier to ensure data files and log files are located on the IBM FlashSystem

Top 5 Timed Foreground Events

Event	Waits	Time(s)	Avg wait (ms)	% DB time	Wait Class
DB CPU		4,497		78.46	
log file sync	2,047,256	982	0	17.14	Commit
library cache: mutex X	212,427	105	0	1.82	Concurrency
db file scattered read	58,832	44	1	0.77	User I/O
db file sequential read	43,311	18	0	0.31	User I/O

- ## Top 5 Foreground User I/O Wait Events
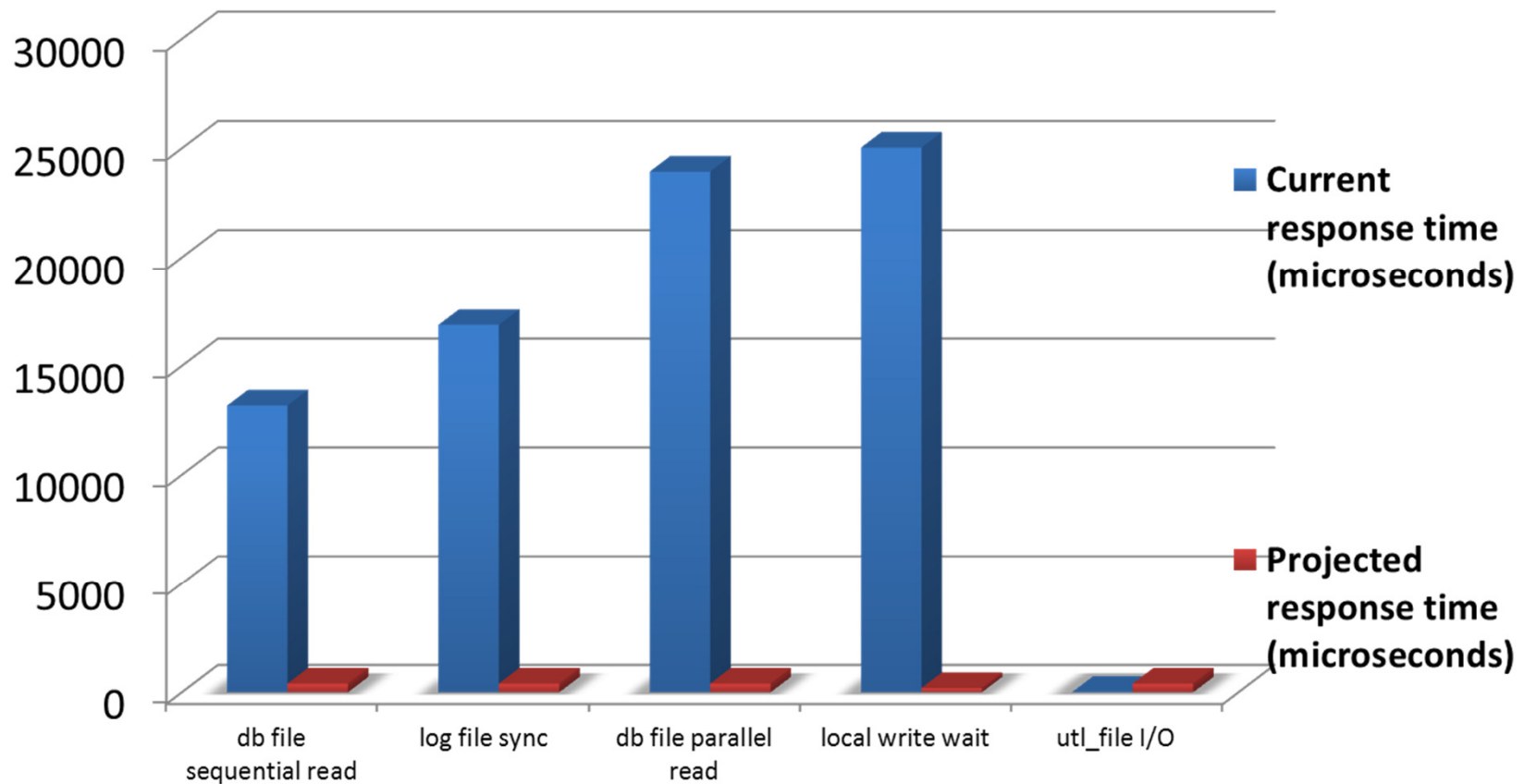
The Top 5 User IO foreground Wait Events indicate substantive improvements in IO response are possible with a FlashSystem implementation. The IO wait time would go down from 53.6% of overall wait time to 5% , an improvement of **957%.**

AWR Analysis - Latency Impact

Latency Improvements

	Latency
■ Average Latency:	4.2262776
■ Adjusted Average Latency	0.268136581
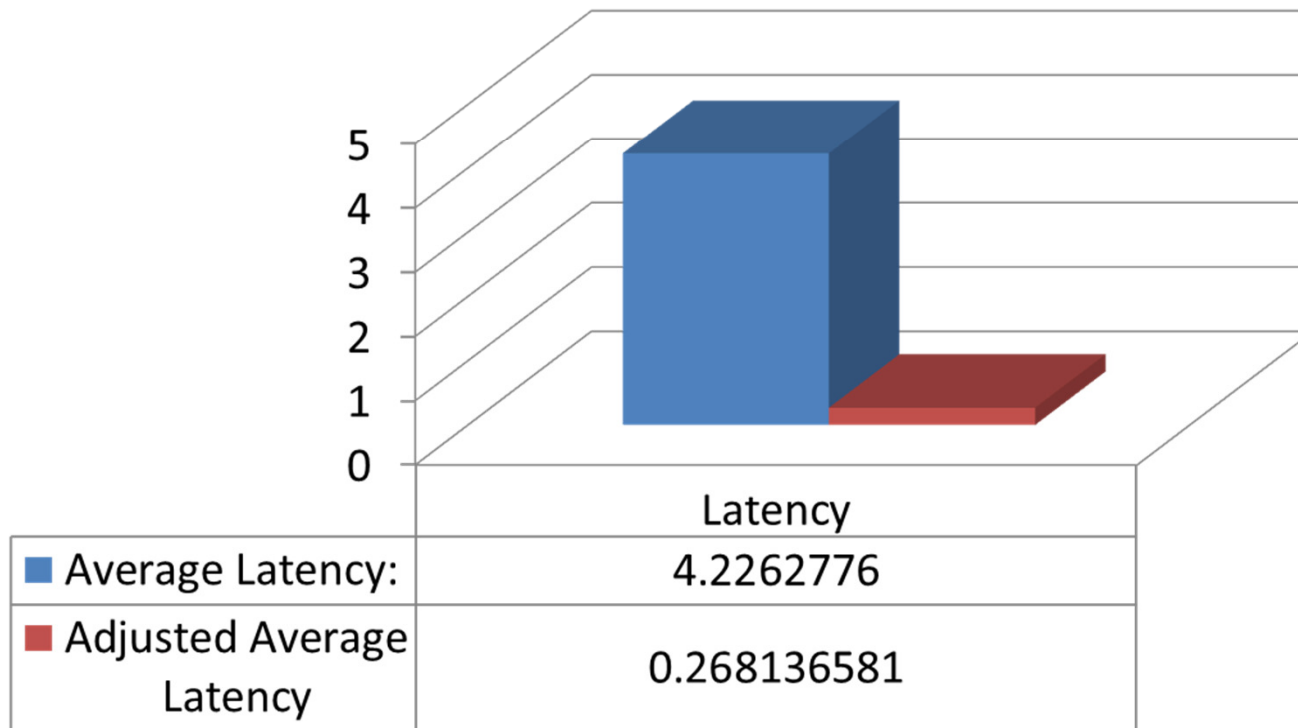
The microsecond response times of the FlashSystem would significantly reduce latency, more than 21x , driving higher utilization at the server and application level. Latency would go down from 4.2ms to .28ms.

Agenda

- Hardware Setup
- z/VM / LPAR
- Linux
- CPU
- Memory
- I/O
- **Networking**
- Oracle

Complete your sessions evaluation online at SHARE.org/BostonEval

Networking

- Choose correct MTU size (should match with App Servers)
- Network queue length
- SHARE session 13523 - Oracle Networking Alternatives

Complete your sessions evaluation online at SHARE.org/BostonEval
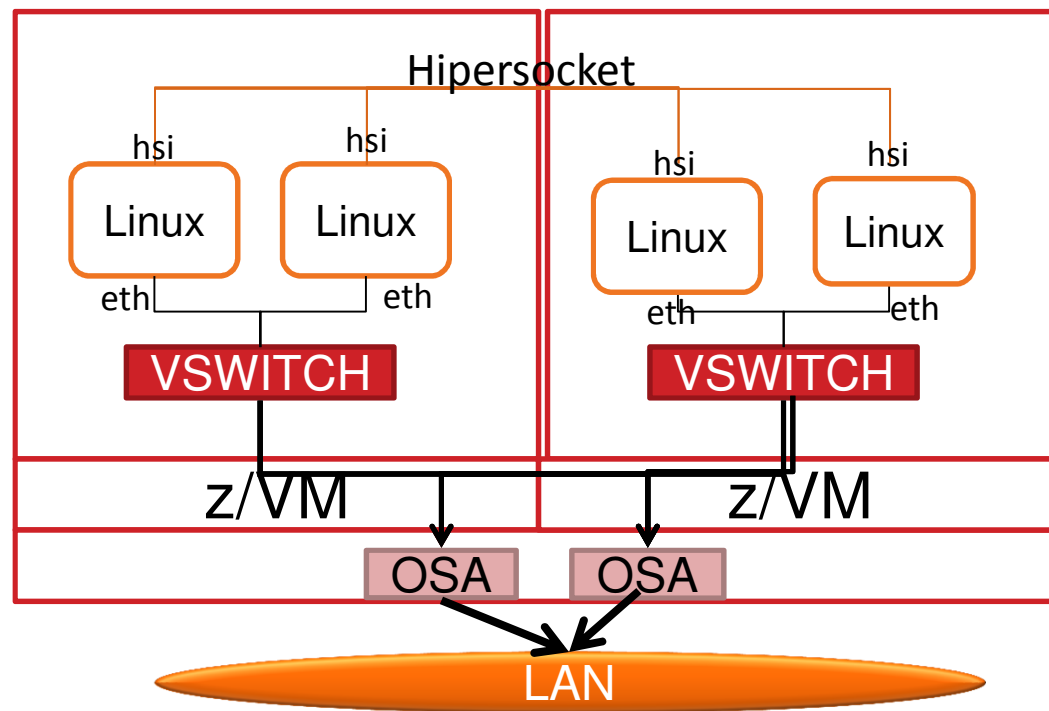
Choose the Correct Network MTU size

netstat –s of Interconnect	MTU Size of 1492 (default)	MTU Size of 8992 (with 8K DB block size, especially RAC)
Before reassemblies	43,530,572	1,563,179
After reassemblies	54,281,987	1,565,071
Delta assemblies	10,751,415	1,892

Network Queue Length

- The device queue length should be increased from the default size of 1000 to at least 2000 using sysctl:

  sysctl -w net.core.netdev_max_backlog =2000

Oracle RAC - Scaling device queue length

Networking: Hipersockets Checksumming Disable

- **HiperSockets does not require network checksum since it is a memory to-memory operation.**

– To save CPU cycles, switch checksumming off:

SuSe SLES10: in **/etc/sysconfig/hardware/hwcfg-qeth-bus-ccw-0.0.F200 add QETH_OPTIONS="checksumming=no_checksumming"**

SuSe SLES11: in **/etc/udev/rules.d/51-qeth-0.0.f200.rules add ACTION=="add", SUBSYSTEM=="ccwgroup", KERNEL=="0.0.f200", ATTR{checksumming}="no_checksumming"**

Red Hat: in /etc/sysconfig/network-scripts/ifcfg-eth0 add OPTIONS="checksumming=no_checksumming"

Oracle Network Configuration Testing

- VSwitch (Active / Passive), Linux Bonding, VSwitch Link Aggregation and Oracle's HAIP

- Tests included shared OSA cards across multiple System z machines

- Separation of Interconnect traffic (application server as well) including VLANs improves performance and stability.

- Multiple Write/Write intensive databases performed best with Link Aggregation or HAIP
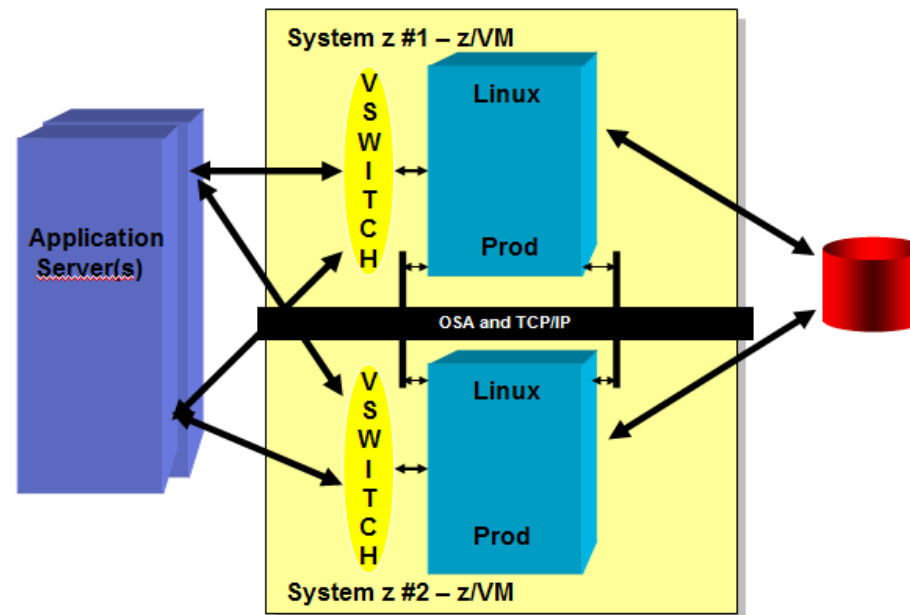
Agenda

- Hardware Setup
- z/VM / LPAR
- Linux
- CPU
- Memory
- I/O
- Networking
- **Oracle**

Complete your sessions evaluation online at SHARE.org/BostonEval

Oracle 11g OLTP improvements

Comparison Oracle 10g vs Oracle 11g Database

User scaling - transactional throughput
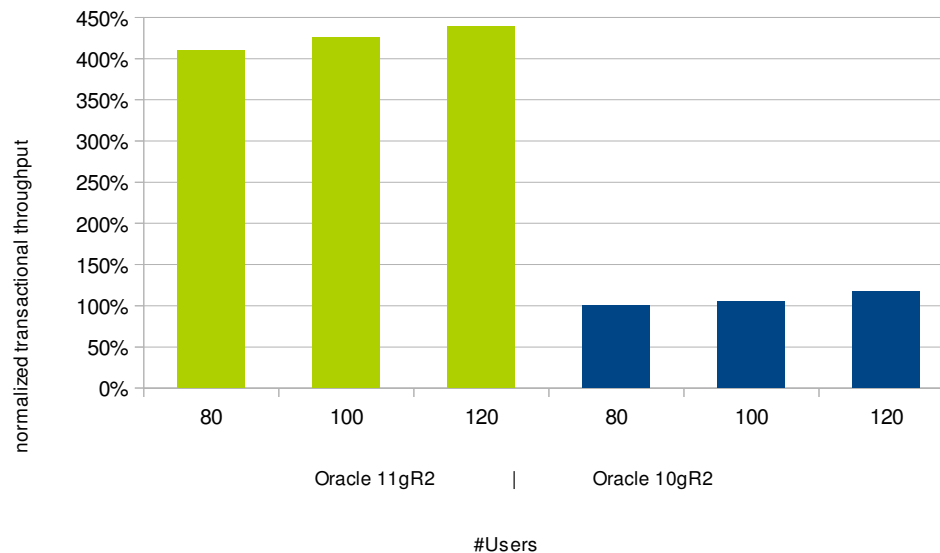
- **Better Caching in Oracle, better CPU efficiency**
- **Recommendation:** Upgrade if not already done!

Oracle 11.2.0.3 Improvements

- Oracle's **VKTM** process uses slightly less CPU minutes
  - (about **0.08** vs. 0.09 with 11.2.0.2)

- Great improvements with **ora_dia0** process.
  - (about **0.07** sec cpu/minute vs. **0.28** with 11.2.0.2)

- Only Install the database modules that are needed
  - DB installed with **NO** options
    The "gettimeofday" function is called 300 times every 15 seconds.

  - DB installed with **all** options : (java, xml, Text, spatial, APEX, etc ....)
    The "gettimeofday" function is called 1500 times every 15 seconds.

Changing the database blocksize in a RAC environment

Changing the database blocksizes 8K ► 4K
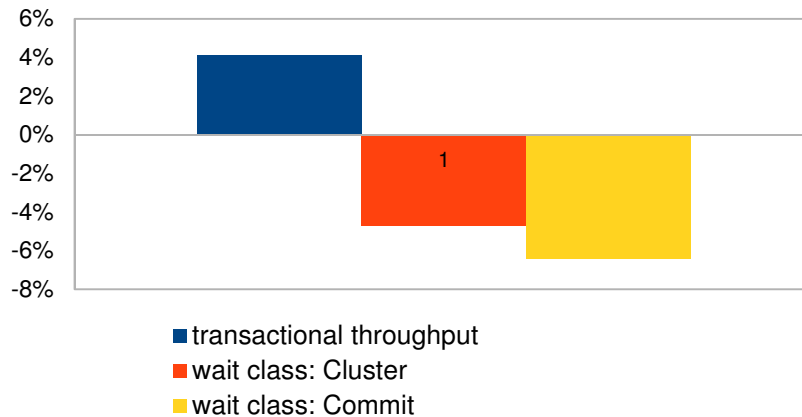
Changing the database blocksizes 8K ► 4 K
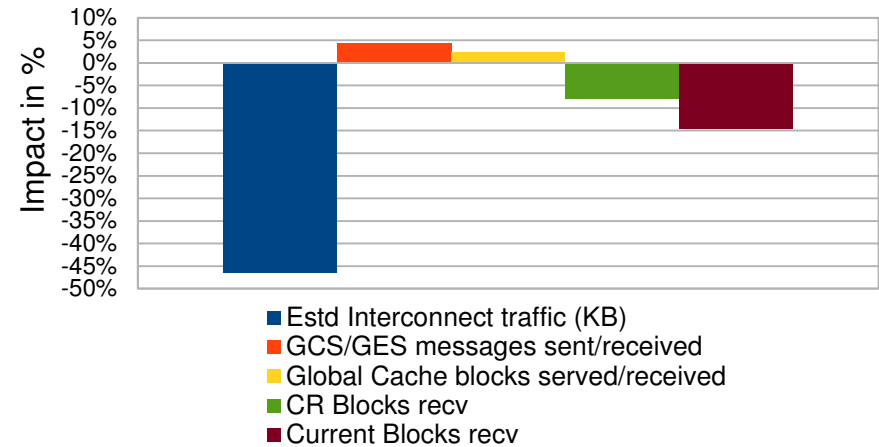
- Comparing environment using 4KB database blocks instead of 8KB
  - improves the throughput about 4%
  - contention based on cluster events is reduced by more than 4%

- Cluster events
  - Locking: one node has to wait until the other node releases a lock, lock size is database page!
  - Waits for datablocks, smaller blocks reduce the load on the Interconnect approximately by the half

- Smaller database page size
  - reduces the chance of a lock collision
  - reduces the latencies

Choose the Best Oracle Audit Options

- Problem: substantial additional CPU load depending on where the data is being stored

- Details see: Oracle Database Auditing: Performance Guidelines

- Investigate if creating an OS audit file is an option for your organization (**audit_trail** oracle parameter = OS or NONE)

- Oracle will create an audit file in the Oracle file system for system operations anyway

Oracle RMAN Backup Compression

Backup Compression	Backup Time	Compression Size Source DB - 1.29 GB	% Compression / Input MB/s
'Basic' 10gR2 (**BZIP2**) Compression	02:48 (168 s)	278.95 MB	78.9 % 7.89 MB/s
'High' 11gR2 (**BZIP2**) Compression	08:41 (521 s)	224.82 MB	83.0 % 2.54 MB/s
'Medium' (**ZLIB**) Compression	01:08 (68 s)	295.53 MB	77.6 % 19.46 MB/s
'Low' (**LZO**) Compression	00:28 (28 s)	357.03 MB	73.0 % 47.26 MB/s

- **RMAN Command -> CONFIGURE COMPRESSION ALGORITHM 'Low'**
- **Oracle Advanced Compression Feature required for Low, Medium, High**
- **Very High CPU observed with BZIP2**

Oracle Optimizer Hints

- Oracle calculates the cpu cost for a sql query plan with:
  - number cores (cpu_count)
  - optimizer_mode (all_rows, first_rows etc) and
  - the number of rows and Bytes in table.

Before updating System Statistics

```
SQL> select * from sys.aux_stats$ where sname='SYSSTATS_MAIN':

SNAME                    PNAME              PVAL1      PVAL2
------------------------ ------------------ ---------- ----------
SYSSTATS_MAIN            CPUSPEEDNW                    1866.16702
SYSSTATS_MAIN            IOSEEKTIM                     10
SYSSTATS_MAIN            IOTFRSPEED                       4096
SYSSTATS_MAIN            SREADTIM
SYSSTATS_MAIN            MREADTIM
SYSSTATS_MAIN            CPUSPEED
SYSSTATS_MAIN            MBRC
SYSSTATS_MAIN            MAXTHR
SYSSTATS_MAIN            SLAVETHR

SQL> execute dbms_stats.gather_system_stats('stop'):
```

run some workload....

```
SQL> execute dbms_stats.gather_system_stats('stop'):
```

After updating System Statistics

```
SQL> select * from sys.aux_stats$ where sname='SYSSTATS_MAIN':

SNAME                    PNAME              PVAL1      PVAL2
------------------------ ------------------ ---------- ----------
SYSSTATS_MAIN            CPUSPEEDNW                    1866.16702
SYSSTATS_MAIN            IOSEEKTIM                     10
SYSSTATS_MAIN            IOTFRSPEED                    4096
SYSSTATS_MAIN            SREADTIM                      .238
SYSSTATS_MAIN            MREADTIM
SYSSTATS_MAIN            CPUSPEED                      2701
SYSSTATS_MAIN            MBRC
SYSSTATS_MAIN            MAXTHR                        885868544
SYSSTATS_MAIN            SLAVETHR                      52770816
```

Oracle Optimize – Running Statics

exec DBMS_STATS.GATHER_SYSTEM_STATS('NOWORKLOAD');

```
z9:
SNAME                           PNAME                           PVAL1
------------------------------- ------------------------------- ----------
SYSSTATS_MAIN                   CPUSPEEDNW                            533
Linux bogomips per cpu: 6510.00

z196:
SNAME                           PNAME                           PVAL1
------------------------------- ------------------------------- ----------
SYSSTATS_MAIN                   CPUSPEEDNW                           2335
Linux bogomips per cpu: 14367.00

zEC12:
SNAME                           PNAME                           PVAL1
------------------------------- ------------------------------- ----------
SYSSTATS_MAIN                   CPUSPEEDNW                           2613
Linux bogomips per cpu: 18115.00
```

Should be done for every hardware upgrade on an idle system

Locking Table Statistics for Large Tables

DBMS_STATS.UNLOCK_TABLE_STATS(ownname => 'USERS', tabname => 'XXX');

DBMS_STATS.GATHER_TABLE_STATS(ownname => 'USERS ', tabname => ' XXX', estimate_percent=>1, cascade =>TRUE, degree =>4);

DBMS_STATS.LOCK_TABLE_STATS(ownname => 'USERS', tabname => 'XXX');
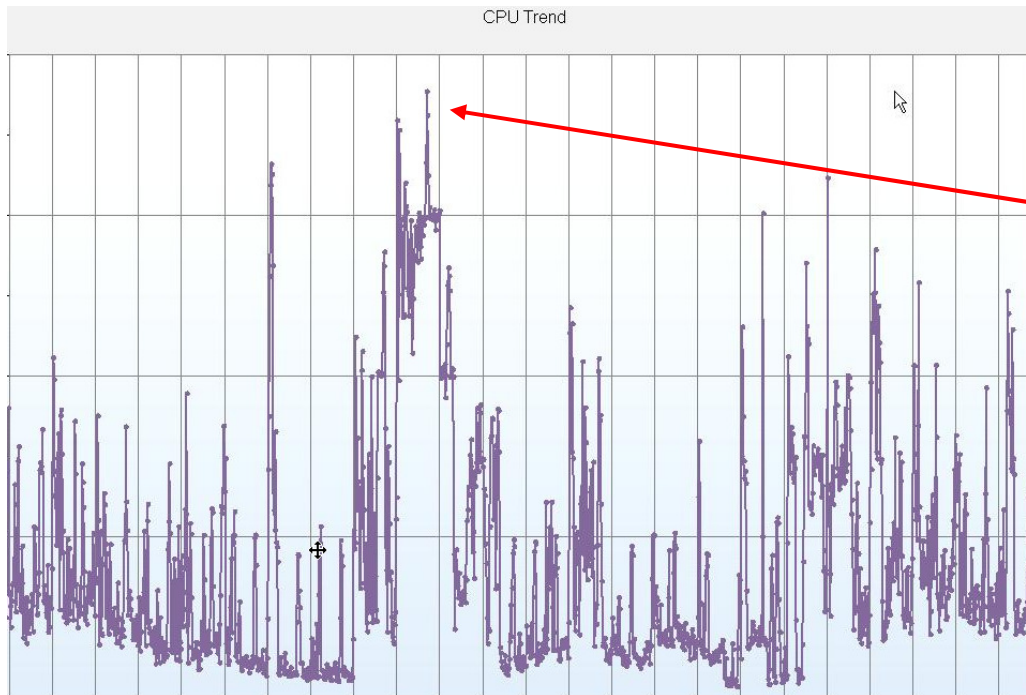
Reduces Unnecessary Statistics Collection

Collect Oracle AWR Data

- **Instance Efficiency Percentages**

Buffer Hit% = 98.89

Buffer Nowait %:	99.97	Redo NoWait %:	100.00
Buffer Hit %:	98.89	In-memory Sort %:	100.00
Library Hit %:	70.53	Soft Parse %:	26.01
Execute to Parse %:	28.44	Latch Hit %:	99.96
Parse CPU to Parse Elapsd %:	30.81	% Non-Parse CPU:	89.14

- **Oracle SGA Buffer Pool Advisory**

P	Size for Est (M)	Size Factor	Buffers for Estimate	Est Phys Read Factor	Estimated Physical Reads
D	256	0.64	16,080	1.11	97,368,882
D	288	0.72	18,090	1.11	96,868,286
D	320	0.80	20,100	1.08	94,323,210
D	352	0.88	22,110	1.05	91,776,695
D	384	0.96	24,120	1.02	89,228,794
D	400	1.00	25,125	1.00	87,480,193
D	416	1.04	26,130	0.98	85,731,549
D	448	1.12	28,140	0.94	82,232,582
D	480	1.20	30,150	0.90	78,731,330
D	512	1.28	32,160	0.86	75,225,110
D	544	1.36	34,170	0.82	71,715,825
D	576	1.44	36,180	0.78	68,209,778
D	608	1.52	38,190	0.72	63,357,042
D	640	1.60	40,200	0.67	58,494,659

- **Predicts 29 (of 87) million block reads could be eliminated over 30 minute period by adding 240 MB of buffer pool cache:**

  - **2,000 read IOs /second**

  - **16,000 blocks /second**

  - **125 MB/second**

  - **A 33% savings**

Log Buffer Size & Redo Log File Size

- Oracle10gR2+ best to let Oracle automatically set the optimal **log_buffer** size. (i.e. leave unset in the init.ora).

- Check AWR Report - ideally log switches every 15 – 20 minutes.

- If log switches more frequent you should increase size of logs.

- If using **fast_start_mttr_target** then can use:

  select optimal_logfile_size from v$instance_recovery;
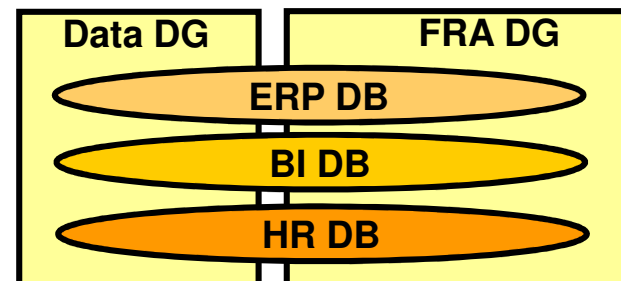
Oracle Resource Manager-
(resmgr:cpu quantum Wait Event)

1) Modify Oracle Initialization parameter - resource_manager_plan = ''

2) Additionally You need disable the Maintenance Window Resource Plan

select window_name,RESOURCE_PLAN
from DBA_SCHEDULER_WINDOWS;

```
WINDOW_NAME                        RESOURCE_PLAN
------------------------------     ------------------------------
MONDAY_WINDOW                      DEFAULT_MAINTENANCE_PLAN
```

execute dbms_scheduler.set_attribute('MONDAY_WINDOW','RESOURCE_PLAN','');

```
WINDOW_NAME                        RESOURCE_PLAN
------------------------------     ------------------------------
MONDAY_WINDOW
```

Oracle's Remote Diagnostic Agent (RDA) Reports – Note: 314422.1

SHARE
Technology · Connections · Results

RDA HTML Menu

- Overview
- Operating System Setup
- User Profile
- Performance
- Network
- Oracle Net
- Oracle Installation
- RDBMS
  - RDBMS Memory
  - RDBMS Log/Trace Files
  - Backup and Recovery
  - SQL*Plus/iSQL*Plus
- IBM WebSphere (Offline)
- J2EE/OC4J
  - Generic
  - J2EE Miscellaneous
- Oracle JDBC
- Cluster
  - Hang Analysis
- ASM
- Data Guard
- Enterprise Manager Server
- Database Control
- External Data Collection

List of Diagnostic Problems

Using: SHOW PROBLEM -ALL -ORDERBY LASTINC_TIME DSC

From: /opt/oracle/diag/rdbms/edpsprd/edpsprd

Problem ID	Problem Key	Last Incident	Last Incident Time
4	ORA 4031	516429	2013-01-12 12:33:39.529000 -05:00
6	ORA 445	411813	2013-01-08 20:06:34.734000 -05:00
7	ORA 240	381339	2012-12-19 19:59:01.195000 -05:00
5	ORA 600 [15709]	246899	2012-08-25 05:41:55.184000 -04:00
2	ORA 7445 [kggmd5Process()+26]	13410	2011-12-12 18:16:11.498000 -05:00
3	ORA 600 [SKGMHASH]	13209	2011-12-12 11:39:00.697000 -05:00
1	ORA 7445 [kglgob()+8490]	9169	2011-12-06 12:57:10.293000 -05:00

Summarized Errors

Current CPU Hogs / Top 15 by CPU Time

```
F S UID        PID  PPID C PRI  NI ADDR SZ WCHAN  STIME TTY          TIME CMD
0 R oracle    23639     1 65  79   0 - 21093142 stext 13:15 ?      04:59:23 ora_j000_edpsprd
0 R oracle    24814     1 47  78   0 - 21089063 stext 16:13 ?      02:12:07 oracleedpsprd (LOCAL=NO)
0 S oracle    17293     1  7  75   0 - 21088031 sk_wai Jan14 ?     02:02:05 oracleedpsprd (LOCAL=NO)
0 S oracle    31422     1  8  75   0 - 21088013 sk_wai Jan14 ?     01:45:42 oracleedpsprd (LOCAL=NO)
0 S oracle     1879     1  3  75   0 - 21090269 sk_wai Jan13 ?     01:42:19 oracleedpsprd (LOCAL=NO)
0 S oracle    29474     1  3  75   0 - 21092455 semtim Jan13 ?     01:39:25 ora_dbw0_edpsprd
0 S oracle    29478     1  2  75   0 - 21090149 semtim Jan13 ?     01:26:40 ora_dbw1_edpsprd
0 S oracle    29482     1  1  75   0 - 21095330 semtim Jan13 ?     00:54:31 ora_lgwr_edpsprd
0 R oracle     1349     1 54  85   0 - 21097455 stext 20:00 ?      00:28:37 oracleedpsprd (LOCAL=NO)
4 S root      27853     1  0  79   0 - 43180 rt_sig Jan13 ?        00:24:34 /opt/tivoli/tsm/StorageAgent/bin/dsmsta
0 S oracle     7960  7933  0  75   0 - 230979 futex Jan13 ?        00:19:24 /opt/oracle/product/11.2.0.3/db/jdk/bin/java
0 R oracle    16863     1 13  75   0 - 21089235 stext 18:43 ?      00:17:18 oracleedpsprd (LOCAL=NO)
0 S oracle    16879     1 13  75   0 - 21089235 sk_wai 18:43 ?     00:17:14 oracleedpsprd (LOCAL=NO)
0 S oracle    16855     1 13  75   0 - 21089235 sk_wai 18:43 ?     00:16:59 oracleedpsprd (LOCAL=NO)
0 S oracle    16897     1 13  75   0 - 21089235 sk_wai 18:43 ?     00:16:50 oracleedpsprd (LOCAL=NO)
```
Back to top

Performance Reports

Root CPU Hogs / Top 5 by CPU Time

```
F S UID        PID  PPID C PRI  NI ADDR SZ WCHAN  STIME TTY          TIME CMD
4 S root      27853     1  0  79   0 - 43180 rt_sig Jan13 ?        00:24:34 /opt/tivoli/tsm/StorageAgent/bin/dsmsta
5 S root      25436     1  0 -40   - - 34880 futex Jan13 ?         00:05:56 /sbin/multipathd
4 S root      21726 20943  0  76   0 -   797 select 13:03 pts/2    00:02:34 top
4 S root      27841     1  0  75   0 - 17482 compat Jan13 ?        00:02:17 /opt/tivoli/tsm/client/ba/bin/dsmc sched
1 S root         24     1  0  70  -5 -     0 worker Jan13 ?        00:00:29 [events/0]
```

Complete your sessions evaluation online at SHARE.org/BostonEval

Oracle's OS Watcher Reports – Pro-Active Problem Avoidance

```
###############################################################################
Section 3: Other General Findings

WARNING : Disk high service time observed.

WARNING : Network TCP segments retrans observed.
(Advise: if retransmitted is over 15% of total packets sent, then TCP experiencing timeouts )
(Check: bottleneck may be on the receiving node )
(Check: general network problems can cause TCP retransmissions (too much network traffic) )


###############################################################################
```

```
###########################
Section 1: Overall Status

Subsystem        Status
-----------------------
CPU              OK
MEMORY           OK
I/O              WARNING
NET              WARNING

###########################
```

```
TCP Errors > 0% Packet Retransmitted:

PARAMETER                    VALUE
-----------------------------------------
segments received            134713581
segments send out            139241863
segments retransmited        6062
connection resets received   3156
resets sent                  3721
failed connection attempts   2426
```

SHARE
in Boston

References (1) – Key Oracle Notes

Note 1306465.1 Getting Started - 11gR2 Grid Infrastructure, SI(Single Instance), ASM and DB (IBM: Linux on System z)
Note 1470834.1 - Requirements for Installing Oracle 11gR2 on RHEL 6 on IBM: Linux on System z (s390x)
Note 1290644.1 - Requirements for Installing Oracle 11gR2 on SLES11 on IBM: Linux on System z (s390x)

Note:1476511.1 OHASD fails to start on SuSE 11 SP2 on IBM: Linux on System z
Note 1308859.1 Requirements for Installing Oracle 11gR2 on SLES 10 on IBM: Linux on System z   (s390x)
Note 1306889.1 Requirements for Installing Oracle 11gR2 on RHEL 5 on IBM: Linux on System z (s390x)
Note 1086769.1 Ensure you have prerequisite rpms to install Oracle Database & AS10g(midtier) IBM: Linux on System z
Note 1377392.1   How to Manually Configure Disk Storage devices for use with Oracle ASM 11.2 on IBM: Linux on System z)
Note 1400185.1 How to Upgrade Oracle Restart i.e. Single Node Grid Infrastructure/ASM from 11.2.0.2 to 11.2.0.3
Note 1276058.1 Oracle GoldenGate Best Practices: Instantiation from an Oracle Source Database
Note 1413787.1 How to completely remove 11.2 Grid Infrastructure, CRS and/or Oracle Restart - IBM: Linux on System z

Note 259301.1 CRS and 10g Real Application Clusters
Note 268937.1 Repairing or Restoring an Inconsistent OCR in RAC
Note 239998.1 10g RAC How to clean up after a failed CRS Install
Note 220970.1 RAC Frequently Asked Questions Topic

Note 1082253 Requirements for Installing Oracle 10gR2 RDBMS on SLES 10 zLinux (s390x)
Note 741646.1 Requirements for Installing Oracle 10gR2 RDBMS on RHEL 5 on zLinux (s390x).
Note 415182.1 DB Install Requirements Quick Reference - zSeries based Linux .

Note 741146.1 Installing Standalone Agent 10.2 on Linux on z

Note 1561740.1 Improve I/O Performance On ext3/ext4 File Systems With The "noatime" Mount Option

References (2)

- White Papers / Presentations
  - Oracle Database on Linux on System z - Disk I/O Connectivity Study
  - Oracle Real Application Clusters on Linux on IBM System z: Set up and network performance tuning
  - Performance of an Oracle 10g R2 Database Import Environment
  - Using the Linux cpuplugd Daemon to manage CPU and memory resources from z/VM Linux guests
  - Oracle Database Auditing: Performance Guidelines
  - Analyzing BI Oracle Workloads Performance Tuning Results – Real Customer Examples
  - Oracle RAC Networking Alternatives on Linux on System z and Red Hat 6 Oracle DB Support
- Redbooks
  - Experiences with Oracle 11gR2 on Linux on System z
  - Installing Oracle 11gR2 RAC on Linux on System z
  - Silent Installation Experiences with Oracle Database 11gR2 Real Application Clusters on Linux on System z
- Other Resources
  - z/VM 6.3 pre-announce
  - International zSeries Oracle Special Interest Group
  - Linux on System z Tuning hints & tips

Thank you!

?

Previous versions published

- SHARE Spring 2013

Special thanks to:

- Dr. Juergen Doelle – IBM Lab Boeblingen
- Dr. Eberhard Pasch -

- http://linuxmain.blogspot.de/
- #IBMandOracle

SHARE in Boston

Tips Learned Implementing Oracle Solutions With Linux on IBM System z (Part I & II)

Christian Ehrhardt(ehrhardt@de.ibm.com)

&

David Simpson (simpson.dave@us.ibm.com)

Speakers Company: IBM

Date of Presentation: **1:30 & 3:00 PM, Thursday, August 15, 2013**

Room 310, Hynes Convention Center

Session Number: **13535 + 13556**

Twitter -> @IBMandOracle

http://linuxmain.blogspot.com/