Oracle Solutions Team
Advanced Technical Skills (ATS) North America

Oracle on AIX – Configuration & Tuning

R Ballough
ballough@us.ibm.com

© 2012 IBM Corporation

Legal Information

The information in this presentation is provided by IBM on an "AS IS" basis without any warranty, guarantee or assurance of any kind. IBM also does not provide any warranty, guarantee or assurance that the information in this paper is free from errors or omissions. Information is believed to be accurate as of the date of publication. You should check with the appropriate vendor to obtain current product information.

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.

IBM, the IBM logo, ibm.com AIX, AIX (logo), AIX 6 (logo), pSeries, xSeries, AIX 5L, Chiphopper, Chipkill, Cloudscape, DB2 Universal Database, DS4000, DS6000, DS8000, EnergyScale, Enterprise Workload Manager, General Purpose File System, GPFS, HACMP, HACMP/6000, IBM Systems Director Active Energy Manager, Micro-Partitioning, POWER, PowerExecutive, PowerVM, PowerVM (logo), PowerHA, Power Architecture, Power Everywhere, Power Family, POWER Hypervisor, Power Systems, Power Systems (logo), Power Systems Software, Power Systems Software (logo), POWER2, POWER3, POWER4, POWER4+, POWER5, POWER5+, POWER6, System p, System p5, System Storage, Workload Partitions Manager and X-Architecture are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Oracle is a registered trademark of Oracle Corporation and/or its affiliates.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

All other products or company names are used for identification purposes only, and may be trademarks of their respective owners.

 © 2012 IBM Corporation

Key Metalink Notes

- **Minimum Software Versions and Patches Required to Support Oracle Products on Power Systems – 282036.1**
- **Rac and Oracle Clusterware Best Practices and Starter Kit (AIX) 811293.1**
- **AIX: Top Things to DO NOW to Stabilize 11gR2 GI/RAC Cluster 1427855.1**
- **Oracle Database on UNIX AIX, HP-UX, etc Unix Operating Systems Installation and Configuration Requirements Quick Reference 169706.1**
- **GPFS and Oracle RAC 1376369.1**

 © 2012 IBM Corporation

Status of Oracle Certification -> My Oracle Support -> Certify

POWERVM :

- **partitioning, micropartitioning for all deployments**
- **VIOS with AIX 5.3-7.1, 10gR2+, 11.2.0.2 minimum for AIX 7.1**
- **LPM with 10.2.0.4+, 11.1.0.7+, 11.2.0.1+, RAC with 11.2.0.2+**
- **WPAR with 10.2.0.4+ or 11.2.0.2+ (ML 889220.1 covers WPAR), Single instance only**
- **AME with 11.2.0.2+**

RAC Cluster interconnects

- VIOS VLAN/SEA
- IVE
- RDS, IP over IB (RDS requires AIX 5.3 TL8+ AND Oracle RAC 11.1.0.6+)
- 10 Gig E, 1 GigE, 100 Mbps

RAC Storage

- ASM
  - raw OCR & Voting Disk, 10.2.0.3+ or 11gR1
  - 11gR2 new installs require OCR & Voting disk on ASM or GPFS
- GPFS 3.1 – 3.4
- RAW with HACMP (Note: Oracle has stated intent to desupport RAW device storage after 11.2 (ML #754305.1)

All Certification status information should be validated with My Oracle Support for latest updates & details

© 2012 IBM Corporation

AIX Configuration for Oracle “starting points”

- **The suggestions presented here are considered to be basic configuration “starting points” for general Oracle workloads**
- **Customer workloads will vary**
- **Ongoing performance monitoring and tuning is recommended to ensure that the configuration is optimal for the particular workload characteristics**

 © 2012 IBM Corporation

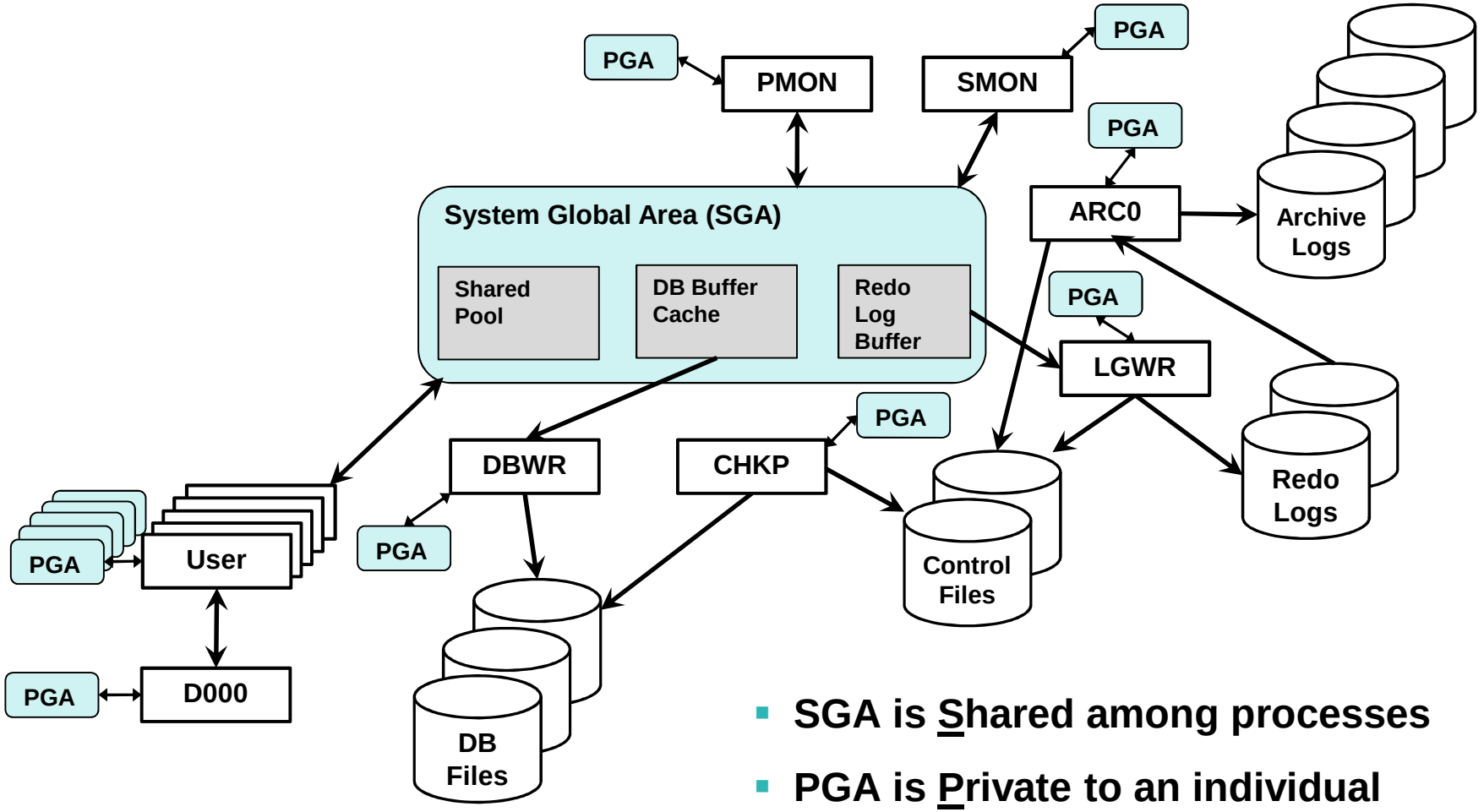
Agenda

AIX Configuration/Tuning for Oracle

- Memory
- CPU
- I/O
- Network
- Miscellaneous

 © 2012 IBM Corporation

Oracle Server Architecture – Memory Structures

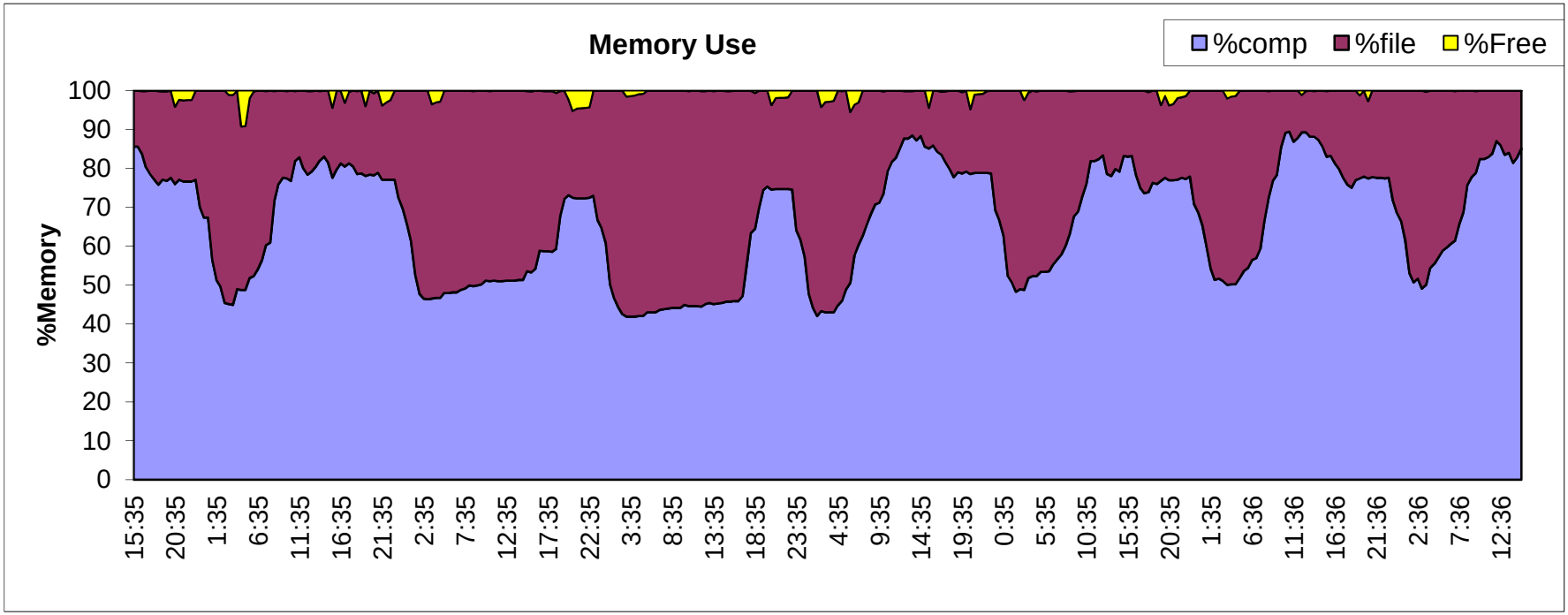


- **SGA is Shared among processes**
- **PGA is Private to an individual server or background process**

 © 2012 IBM Corporation

AIX Memory Management Concepts

- **Two primary categories of memory pages: Computational and File System**
- **AIX tries to utilize all of the physical memory available**
  - What is not required to support computational page demand will tend to be used for filesystem cache
- **Requests for new memory pages are satisfied from the free page list**
  - Small reserve of free pages maintained by “stealing” Computational or File pages
  - AIX uses “demand paging” algorithm – generally not written to swap file until “stolen”



 © 2012 IBM Corporation

AIX System Paging Concepts & Requirements

- **By default, AIX uses a "demand paging" policy**
  - For Oracle DB, the goal is ZERO system paging activity
  - Filesystem pages written back to filesystem disk (if dirty); never to system paging space
  - Unless otherwise specified, computational pages are not written to paging space unless/until they are stolen by lrud
    - **Known Exception:** When Java garbage collection shrinks the JVM heap size, the pages recovered (collected) are written to paging space
      - Recommendation: Set –Xmx = -Xms to prevent heap shinks/pageouts
- **Once written to paging space, pages are not removed from paging space until the process associated with those pages terminates**
  - For long running processes (e.g. Oracle DB), even low levels of system paging can result in significant growth in paging space usage over time
  - Paging space should be considered a fail-safe mechanism for providing sufficient time to identify and correct paging issues, not a license to allow ongoing system paging activity
- **Paging space allocation Rule-of-Thumb:**
  - ½ the physical memory + 4 GB, but not more than capacity of a single drive

Resolve paging quickly issues:

- Reduce effective filesystem cache size (minperm)
- Reduce Oracle SGA or PGA size
- Add physical memory

 © 2012 IBM Corporation

VMM Tuning (AIX 7.1,6.1, 5.3, 5.2ML4+)

lru_file_repage=0 **(6.1/7.1 restricted, 5.3 override)**

Tells lrud to page out file pages (filesystem buffer cache) rather than computational pages when numperm > minperm

lru_poll_interval=10 **(6.1/7.1 restricted, 5.3 default)**

Indicates the time period (in milliseconds) after which lrud pauses and interrupts can be serviced. Default value of “0” means no preemption.

minperm%=3 **(6.1/7.1 default, 5.3 override)**

The target minimum % of physical memory desired for filesystem cache

maxperm%, maxclient%=90 **(6.1/7.1 restricted, 5.3 override)**

The target maximum % of physical memory desired for filesystem cache (permanent or client pages)

strict_maxperm=0 **(6.1/7.1 restricted, 5.3 default)**

Use soft limit for maxperm% target

strict_maxclient=1 **(6.1/7.1 restricted, 5.3 default)**

Enforce hard limit for maxclient% target

 © 2012 IBM Corporation

AIX Kernel Locking

- **In AIX 6.1 TL6+ and 7.1, Kernel Memory Locking may be used to avoid unnecessary kernel page faults thereby improving performance**
  - Locked memory is not stolen until no other pages are available, thereby giving preference to kernel pages
- **Controlled via vmo vmm_klock_mode parameter:**
  - 0 = kernel locking is disabled
  - 1 = enabled for some types of kernel memory only when Active Memory Expansion (AME) is also enabled **(6.1 default)**
  - 2 = enabled for all kernel memory types **(7.1 default)**
    - Recommended for Oracle RAC environments, or where EMC storage is used for AIX paging devices
  - 3 = only kernel stacks of processes are locked
- **Before changing vmm_klock_mode value on AIX 6.1, verify the following is installed:**
  - AIX 6.1 TL6 SP5+, plus APAR IZ95744

 © 2012 IBM Corporation

Oracle Memory Structures Allocation

- **9i : Dynamic memory resizing**
  - db_cache_size (dynamic) – size of area for caching database blocks
  - sga_max_size (static) – maximum size of the SGA for the lifetime of the instance.
  - pga_aggregate_target (dynamic) – specifies the target aggregate PGA memory available to all server processes attached to the instance
  - db_cache_advice (dynamic) – enables or disables statistics gathering used for predicting behavior with different cache sizes.
- **10g : Automatic Shared Memory Management (ASMM)**
  - sga_target (dynamic) – if set, the db_cache_size, shared_pool_size, large_pool_size and streams_pool_size dynamically sized
    - Minimum values for these pools may optionally be specified
  - Can be increased up to sga_max_size
  - *To use ASMM, sga_target must be >0*
- **11g : Automatic Memory Management (AMM)**
  - memory_target (dynamic parameter) – specifies the total memory size to be used by the instance SGA and PGA. Exchanges between SGA and PGA are done according to workload requirements
  - If sga_target and pga_aggregate_target are not set, the policy is to give 60% of memory_target to the SGA and 40% to the PGA
  - memory_max_target (static parameter) – specifies the maximum memory size for the instance
  - *To use Automatic Memory Management, memory_target must be >0 and LOCK_SGA=false*

 © 2012 IBM Corporation

AIX Multiple Page Size Support

- **4K (Default)**
  - Always used for filesystem cache
  - Can be paged to paging space
  - Can be coalesced to create 64K pages if required
  - Used system wide if Active Memory Sharing (AMS) or AME is used
  - Typically used on older hardware which does not support 64K pages or with older Oracle versions (< 10.2.0.4)

- **64K available with POWER5+ and later & AIX 5.3 TL4+**
  - Can be paged to paging space
  - Can be converted to 4K pages if not enough 4K pages are available
  - Can be utilized for application code, data and stack as well, if specified
  - Kernel page size used in AIX 5.3 TL4+ and AIX 6.1 (can be configured)
  - In 11g Oracle will automatically use 64k for SGA if supported by system
  - May also be used for program data, text and stack areas:

    # export LDR_CNTRL=DATAPSIZE=64K@TEXTPSIZE=64K@STACKPSIZE=64K oracle

- **16M (Large Pages) available with POWER4 hardware (or later)**
  - Must be explicitly preconfigured and reserved, even if not being used
  - Are pinned in memory
  - Unused 16M pages can be converted to 4K or 64K pages if required
  - Cannot be paged to paging space

 © 2012 IBM Corporation

Large Segment Aliasing (AKA Terabyte Segment)

- **Workloads with large memory footprints and low spatial locality may perform poorly due to Segment Lookaside Buffer (SLB) faults**
  - May consume up to 20% of total execution time for some workloads
- **Architectural trend toward smaller SLB sizes can exacerbate SLB related performance issues:**
  - POWER6 has 64 SLB entries – 20 for kernel, 44 for user processes – allowing 11GB of accessible memory before incurring SLB faults
  - POWER7 has 32 SLB entries – 20 for kernel, 12 for user processes – allowing 3GB of accessible memory before incurring SLB faults
- **Oracle SGA sizes are typically in the 10s to 100s of Gigabytes**
- **With Large Segment Aliasing, each SLB entry may address 1TB of memory**
  - Supports **shared memory** addressability for up to 12TB on POWER7 and up to 44TB on POWER6 without SLB faults
  - Enabled by default on AIX 7.1 – 32 bit processes may need fix for IV11261
  - Disabled by default on AIX 61 – May be enabled by setting vmo esid_allocator=1 **(Recommended)**
  - Unshared memory issue – documented in APAR IV23859 & ML 1467807.1
    - Shm_1tb_unsh_enable **or**
    - **AIX 7.1 TL1 SP5 or AIX 6.1 TL7 SP5**

 © 2012 IBM Corporation

Oracle 11.2.0.2+ USLA Heap (ML 1260095.1)

- **Issue: 11gR2 uses 2 additional linker options (-bexpful and –brtllib) to implement hot patching**
- **Global symbols in the Oracle binary are saved in the USLA (User-Space Loader Assistant) heap region for resolving the new object files in the shared libraries included in a patch.**
- **The initial implementation led to much larger USLA heap areas for Oracle processes**
- **Unpatched – about 6.6 MB**

```
root@lp06 / # svmon -P 8192008 | grep USLA
  9a0e9a  80020014 work USLA heap                    sm   1616     0    0    1616
  990139  9fffffff pers USLA text,/dev/hd2:14423      s     14     0    -       -
```

- **Patched – about 77K**

```
root@lp07 / # svmon -P 18546706 | grep USLA
  8983c8  80020014 work USLA heap                    sm     19     0    0      19
  9a015a  9fffffff pers USLA text,/dev/hd2:20586      s     14     0    -       -
```

- **Recommendation:**
- **Use AIX 6.1 TL7 SP2+ or AIX 7.1 TL1 SP2+ and apply Oracle patches 13443029, 13494030**
- **Or….if for some reason upgrading to this AIX level is not feasible, turn off hot patch functionality following instructions in ML 10190759**

 © 2012 IBM Corporation

Recommended vmo Starting Points Review

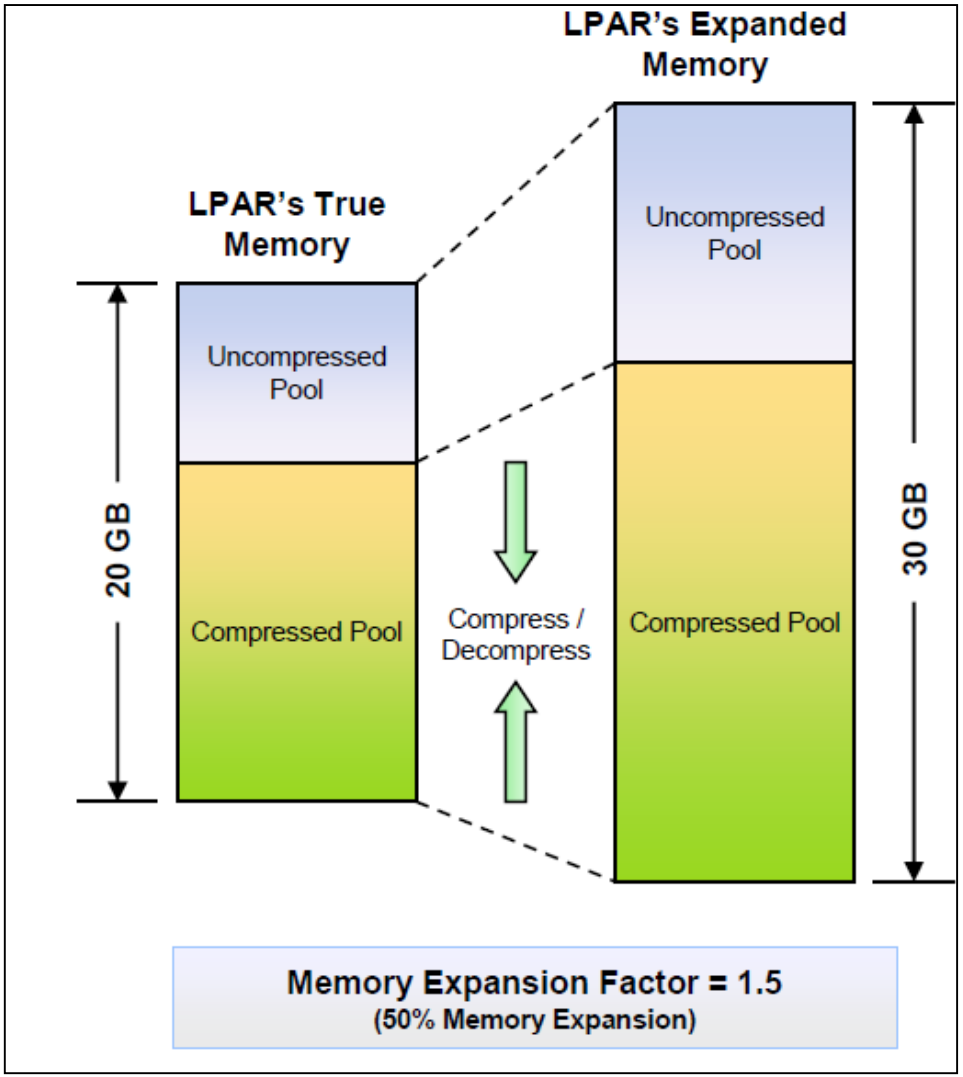
Parameter	Recommend Value	AIX 7.1 Default	AIX 6.1 Default	AIX 6.1/7.1 Restricted	AIX 5.3 Default
esid_allocator	1	1	0	Yes	n/a
vmm_klock_mode	2	2	1	No	n/a
minperm%	3	3	3	No	20
maxperm%	90	90	90	Yes	80
maxclient%	90	90	90	Yes	80
strict_maxclient	1	1	1	Yes	1
strict_maxperm	0	0	0	Yes	0
lru_file_repage	0	0	0	Yes	1 or 0*
lru_poll_interval	10	10	10	Yes	10
minfree	960+	960	960	No	960
maxfree	1088+	1088	1088	No	1088
page_steal_method	1	1	1	Yes	0
memory_affinity	1	1	1	Yes	1
v_pinshm	0	0	0	No	0
lgpg_regions	0	0	0	No	0
lgpg_size	0	0	0	No	0
maxpin%	80	80	80	No	80

*** Depending on AIX 5.3 TL level**

 © 2012 IBM Corporation

Active Memory Expansion (AME)

- **May potentially be used to increase the effective memory capacity of an LPAR without increasing the physical memory**
- **The AME planning tool (amepat) can be used to predict the amount of CPU overhead required to support varying levels of memory expansion**
  - For best accuracy, should be run during multiple heavy workload periods
- **Should be used cautiously, with low compression factors until the workload behavior is well understood**
- **Certified for use with Oracle database**
  - 11.2.0.2 and above
  - AIX 6.1 TL06 SP5 and above
  - Single instance and RAC



 © 2012 IBM Corporation

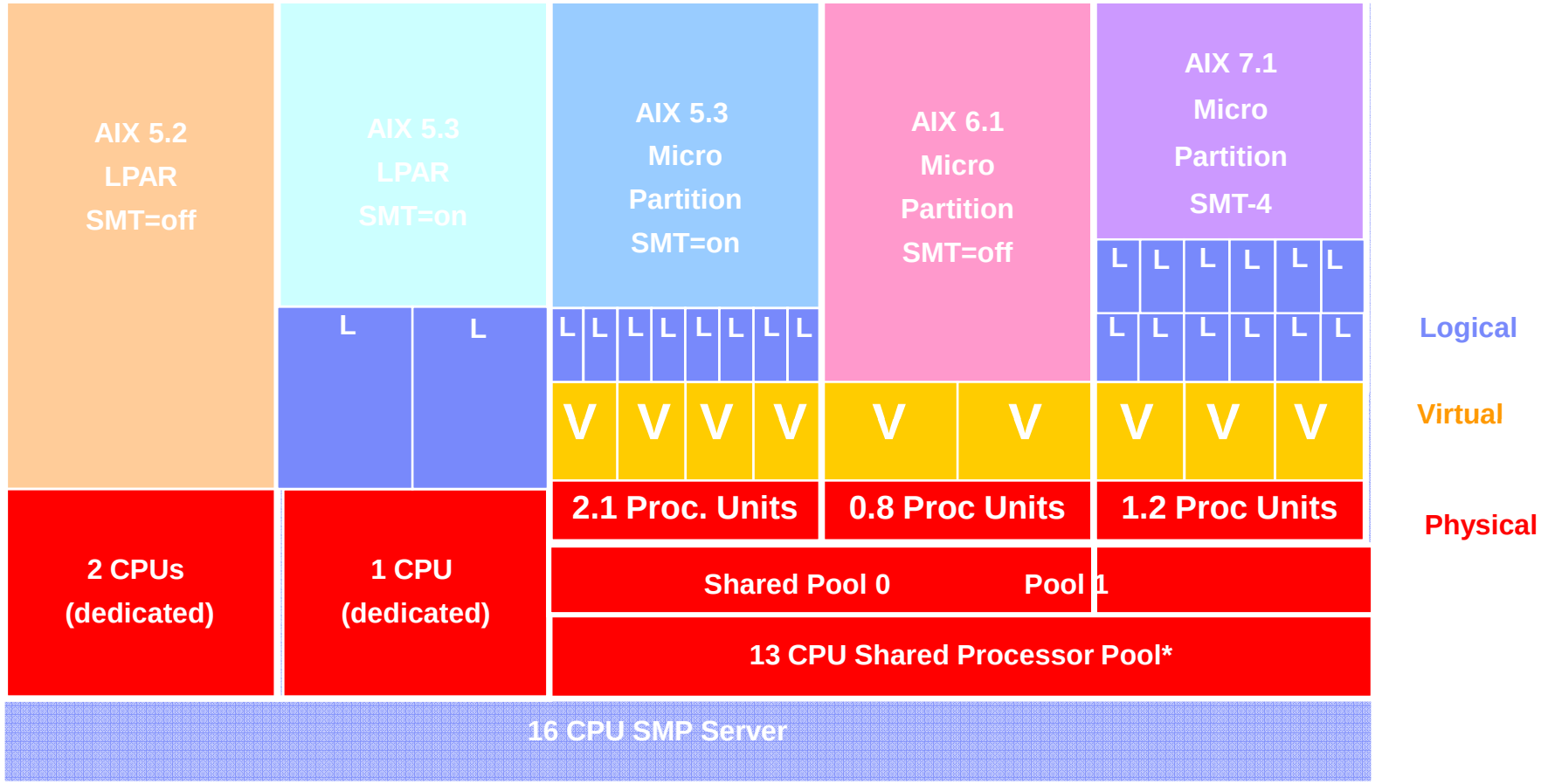
Agenda

AIX Configuration/Tuning for Oracle

- Memory
- CPU
- I/O
- Network
- Miscellaneous

 © 2012 IBM Corporation

Physical, Logical, Virtual Layers



© 2010 IBM Corporation

CPU Considerations

- **Use SMT with AIX 5.3/Power5 (or later) environments**
- **Micropartitioning Guidelines**
  - Virtual CPUs <= physical processors in shared pool

  CAPPED
  - Virtual CPUs should be the nearest integer >= capping limit

  UNCAPPED
  - Virtual CPUs should be set to the max peak demand requirement
  - Preferably, set Entitlement >= Virtual CPUs / 3, not less than minimum demand requirement
- **DLPAR considerations**

  Oracle 9i
  - Oracle CPU_COUNT does not recognize change in # cpus
  - AIX scheduler can still use the added CPUs

  Oracle 10g/11g
  - Oracle CPU_COUNT recognizes change in # cpus
    - Max CPU_COUNT limited to 3x CPU_COUNT at instance startup

 © 2012 IBM Corporation

CPU Related Oracle Parameters

Oracle Parameters based on the # of CPUs

- DB_WRITER_PROCESSES
- CPU_COUNT
- Degree of Parallelism
  - user level
  - table level
  - query level
  - MAX_PARALLEL_SERVERS or AUTOMATIC_PARALLEL_TUNING (CPU_COUNT * PARALLEL_THREADS_PER_CPU)
- FAST_START_PARALLEL_ROLLBACK – should be using UNDO instead
- CBO – execution plan may be affected; check explain plan

 © 2012 IBM Corporation

AIX Active System Optimizer (ASO)

- **Introduced for POWER7**
  - Version 1: AIX 7.1 TL1 SP1 (Fall 2011)
  - Version 2: AIX 7.1 TL2 SP1, AIX 6.1 TL8 SP1 (Fall 2012)

- **The Active System Optimizer (ASO) daemon autonomously tunes the allocation of system resources to improve system performance**

- **Optimization Strategies:**
  - **Cache Affinity** (Fall 2011)
    - Move threads of workloads closer together via affinity domains
  - **Memory Affinity** (Fall 2011)
    - Move process private memory closer to affinity domain
  - **Large Page** (Fall 2012)
    - Promote heavily used (shared memory) regions to 16MB pages
  - **Data Stream Pre-Fetch** (Fall 2012)
    - Modify hardware Data Stream Pre-Fetch (DSCR) behavior based on workload

 © 2012 IBM Corporation

Virtual Processors - Folding

- **Dynamically adjusting active Virtual Processors**
  - System consolidates loads onto a minimal number of VPs
    - Scheduler computes utilization of VPs every second
      - If VPs needed to host physical utilization is less than the current active VP count, a VP is put to sleep
      - If VPs needed are greater than the current active VPs, more are enabled
  - On by default in AIX 5.3 ML3 and later
    - **`vpm_xvcpus` tunable**
    - **`vpm_fold_policy` tunable**
- **Increases processor utilization and affinity**
  - Inactive VPs don't get dispatched and waste physical CPU cycles
  - Fewer VPs can be more accurately dispatched to physical resources by the Hypervisor
- **Rac and Oracle Clusterware Best Practices and Starter Kit (AIX) 811293.1**
  - Older versions of this document recommended disabling
  - Document has been modified to correctly reflect that current TL levels should be used for support of processor folding

© 2012 IBM Corporation

Oracle RAC scheduling changes

Process	Priority	Scheduling Policy
oprocd	**0**	**2**
ocssd.bin	**1**	**-**
LMS/VKTM* (new in 11g)	**39**	**--**

Oracle 10.2.0.4 and above

- Chuser capabilities = CAP_NUMA_ATTACH,CAP_BYPASS_RAC_VMM,CAP_PROPAGATE oracle
  - CAP_NUMA_ATTACH gives authority for non-root processes to increase priority
  - CAP_PROPAGATE allows parent->child capability propagation
  - CAP_BYPASS_RAC_VMM required for oprocd,ocssd.bin to be pinned in memory.

Oracle 10.2.0.3 – 11gR1

- ML #559365.1, – Oracle “diagwait 13” increases OPROCD_DEFAULT_MARGIN from 500 ms to 10s

Oracle 11gR2

- OPROCD no longer exists in 11gR2; Clusterware processes have been rewritten

© 2012 IBM Corporation

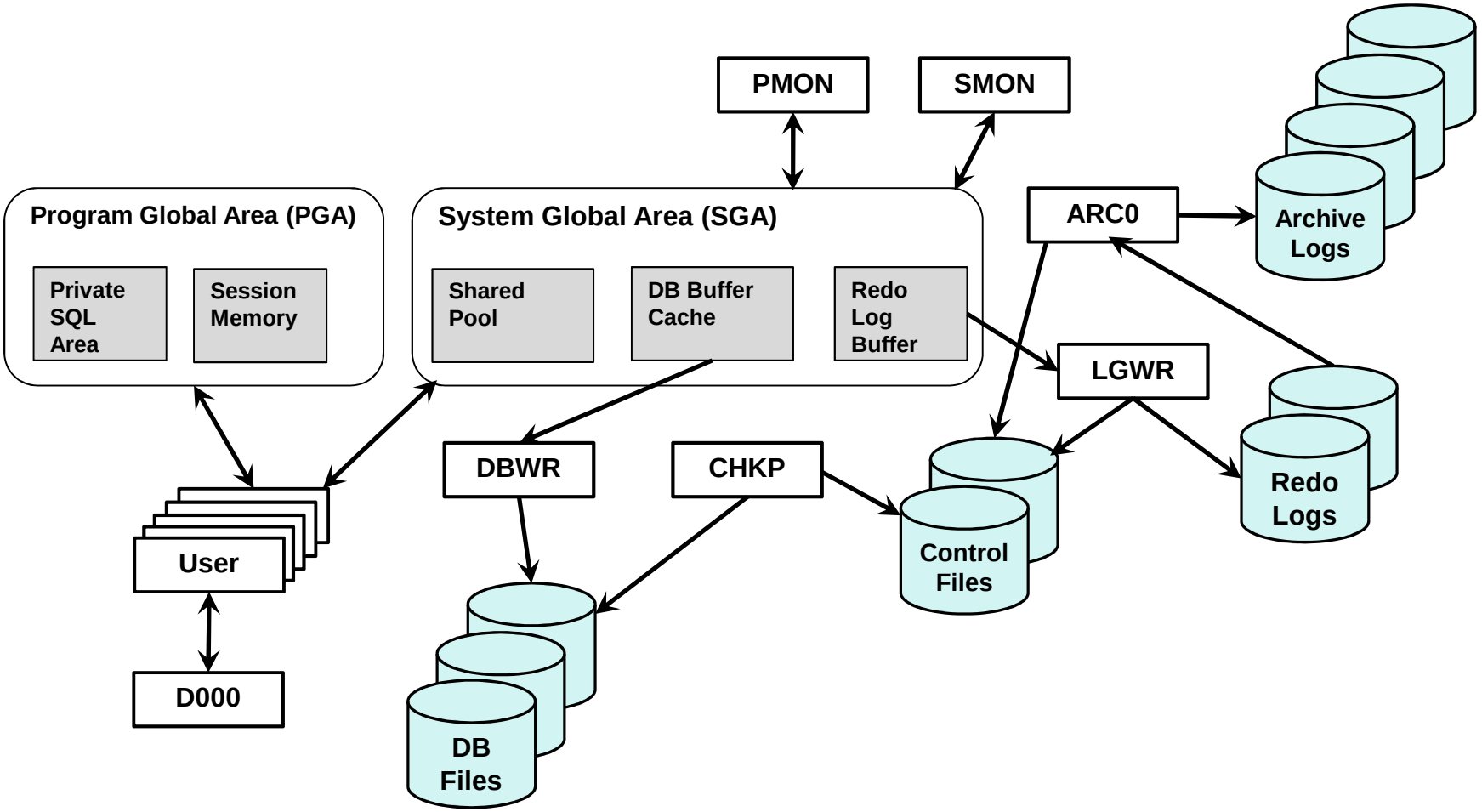
Agenda

AIX Configuration/Tuning for Oracle

- Memory
- CPU
- I/O
- Network
- Miscellaneous

 © 2012 IBM Corporation

Oracle Server Architecture – Files



 © 2012 IBM Corporation

Improving I/O performance

- **Reduce the Amount of Physical I/O**
  - Improve database cache hit ratio or increase PGA
  - Improve execution plans
- **Reduce the Cost of Physical I/O (Improve Service Time)**
  - Use SSD
  - Improve storage subsystem layout
  - Increase the # of physical spindles
  - Increase the disk RPM

 © 2012 IBM Corporation

Data Layout for Optimal I/O Performance

Stripe and mirror everything (SAME) approach:

- **Goal is to balance I/O activity across all disks, loops, adapters, etc...**
- **Avoid/Eliminate I/O hotspots**
- **Manual file-by-file data placement is time consuming, resource intensive and iterative**

The "SAME" methodology was introduced in the "Optimal Storage Configuration Made Easy" Oracle Open World presentation and companion whitepaper in 2001 by Juan Loaiza, Oracle

- **To achieve the benefits of "SAME", not every detail needs to be taken literally, e.g. the same concepts can be adapted for any RAID configuration, not just RAID-1 or RAID-10, or not just host level mirroring (e.g. ASM redundancy).**

 © 2012 IBM Corporation

hdisk and Fiber Channel Device attributes

- **hdisknn**
  - **The maximum transfer size should be at least equal to the largest I/O transfer requested by Oracle (typically up to 1 Megabyte, though empirically 256K seems to optimize performance between random & sequential workloads)**

    # chdev –l hdisknn –a max_transfer=
  - **For workloads with a large number of concurrent DB connections and/or parallel query slaves, the queue depth should be large enough to support the anticipated concurrent I/O load (within storage subsystem vendor guidelines)**

    # chdev –l hdisknn –a queue_depth=
    - **Monitor with iostat –D looking for 'sqfull'**
- **fcsn**
  - **The maximum transfer size should be at least equal to the value set at the hdisk level**

    # chdev –l fcsn –a max_xfer_size=
  - **The number of concurrent active I/O requests should be large enough to support the anticipated I/O workload (within storage subsystem vendor guidelines)**

    # chdev –l fcsn –a num_cmd_elems=

    Monitor with fcstat for 'No Command Resource Count'

 © 2012 IBM Corporation

Host Level Striping Options

- **ASM**
  - **Stripes by default when multiple LUNs configured per ASM disk group.**
  - **10gR2 Strip size is 128k (Fine-grained) or Allocation Unit (AU) Size (Coarse-grained)**
  - **11g Strip size = Allocation Unit (AU) Size, default = 1 MB**

- **Single-instance filesystems or Raw Devices**
  - **use AIX Physical Partition (PP) Spreading or LV striping**

    LV Strip sizes: 4k-128M, 1M most common
    PP Strip sizes: use default (depends on LUN size), generally about 256
    Use Scalable Volume Groups (VGs), or use "mklv –T O" with Big VGs

- **GPFS**
  - **Stripes by default when multiple LUNs configured per filesystem.**
  - **Strip size (block size) is configurable**

 © 2012 IBM Corporation

JFS2 Filesystem Mount Options (Non-RAC)

Mount options:

- **Buffer Caching (default): stage data in fs buffer cache**
- **Direct I/O (DIO): no filesystem caching**
- **Concurrent I/O (CIO): DIO + no write serialization (JFS2 only)**
- **Release Behind Read (RBR): memory pages released (available for stealing) after pages copied to internal buffers**
- **Release Behind Write (RBW): memory pages released (available for stealing) after pages written to disk**
- **No Access Time (NOATIME): do not update last accessed time when file is accessed**

  An existing Oracle bug (expected fixed in Oracle 12) causes unnecessary accesses to the ORACLE_HOME, GRID_HOME and root ('/') filesystems

  - Effects may be partially mitigated by by using the 'noatime' option

    # chfs -a options=noatime /
    # bosboot –r (followed by a reboot)
    # umount /u01
    # chfs -a options=noatime /u01
    # mount /u01

 © 2012 IBM Corporation

JFS/JFS2 settings

Data Base Files (DBF)

- I/O size ranges from db_block_size to db_block_size * db_file_multiblock_read_count
- Use CIO (or DIO for JFS) or filesystem cache, depending on I/O characteristics
- If block size is >=4096, use a filesystem block size of 4096, else use 2048

Redo Log/Control Files

- I/O size is always a multiple of 512 bytes
- Use CIO (or DIO for JFS) and set filesystem block size (agblksize) to 512

Archive Log and Backup Files

- Don't use CIO or DIO
- 'rbrw' mount option can be advantageous

Flashback Log Files

- Writes are sequential, sized as a multiple of db_block_size
- By default, dbca will configure a single location for – the flash recovery area - for flashback logs, archive logs, and backup logs
- Flashback Log files should use CIO, DIO, or 'rbrw' mount

Oracle Binaries

- Don't use CIO or DIO
- Use NOATIME to reduce 'getcwd' overhead

System Root (/) Filesystem

- Use NOATIME to reduce 'getcwd' overhead (requires botboot, reboot)

 © 2012 IBM Corporation

Asynchronous I/O

Use of Asynchronous I/O is strongly advised for Oracle workloads

- AIX supports 2 types: POSIX and Legacy – Oracle uses Legacy

AIX parameters:

minservers = minimum # of AIO server processes
maxservers = maximum # of AIO server processes
maxreqs = maximum # of concurrent AIO requests
"enable" at system restart

Oracle parameters:

disk_asynch_io	= TRUE
filesystemio_options	= {ASYNCH | SETALL}
db_writer_processes	(let default)

Raw Devices , ASM (rhdisk), and CIO environments using AIX 6.1+ (or 5.3 + fsfastpath) use kernelized or "fastpath" AIO

- aio routines are dynamically managed by the kernel
- AIO parameters above do not apply

 © 2012 IBM Corporation

Setting Asynchronous I/O Parameters

AIX 6.1/7.1

- Use ioo command to change
- Defaults are generally ok though aio_maxservers may need to be increased:

aio_minservers = 3	(per logical CPU)
aio_maxservers = 30	(per logical CPU)
aio_maxreqs = 65536	

Monitor Oracle usage:

- Watch alert log and *.trc files in BDUMP directory for warning message:
  "Warning "lio_listio returned EAGAIN"
  Usually indicates that maxreqs (or sometimes maxservers) is set too low

Monitor from AIX:

- "pstat –a | grep aios"
- Use "-A" option for NMON
- iostat –Aq (new in AIX 5.3)

 © 2012 IBM Corporation

What's a Physical Volume Identifier (PVID)?

- ID numbers used to track AIX Physical Volumes (PVs)
- Physically written on the disk (LUN) and registered in the ODM
- Must be set before (or when) disks are assigned to a Volume Group
- Preserves hdisk numbering across reboots and storage reconfigurations
- Can be displayed with 'lspv' command:

```
# lspv
hdisk0 00cb0e8fcc7ab6d2                  rootvg   active
hdisk1      002c41afc70de886             rootvg   active
hdisk2      00ca00aeff4f4d42             oravg    active
hdisk3      00ca00aeff5475ef             None
hdisk4      none                         None
hdisk5      none                         None
```

- If PVIDs haven't been system generated, they can be set or cleared with the 'chdev' command:

```
# chdev -l hdisk4 –a pv=yes
# chdev –l hdisk4 –a pv=clear
```


© 2010 IBM Corporation

PVIDs and Oracle ASM

- **ASM Managed Disks**
  - **ASM preserves its own mapping between LUNs and assignments in ASM disk**
  - **Some Oracle installation documentation recommends temporarily setting PVIDs during the install process to identify hdisks. This is not a good habit to start!!**
  - **Assigning or clearing a PVID on an existing ASM managed disk will overwrite the ASM header, making data unrecoverable without the use of KFED (See Metalink Note # 353761.1)**

- **OCR and Voting devices (for RAC,11gR1 and earlier)**
  - **Can't have a PVID associated with them either**
  - **Should be given a logical device name (assigned by major/minor #) because numbering for hdisks not having PVIDs may change**

```
# ls –l /dev/hdisk4
brw------  1 root   system 14,   3  Jan  31 2006  /dev/hdisk4
# mknod /dev/ocr1 c 14  3
# ls –l /dev/ocr1
crw-r–r--  1  root   system 14,   3  May 03 12:30 /dev/ocr1
```


© 2010 IBM Corporation

Identifying ASM disks

```
# lquerypv -h /dev/hdisk7

00000000   00820101 00000000 8000000A BCA9C9A9  |................|
00000010   00000000 00000000 00000000 00000000  |................|
00000020   4F52434C 4449534B 00000000 00000000  |ORCLDISK........|
00000030   00000000 00000000 00000000 00000000  |................|
00000040   0A100000 000A0103 414C4C44 41544131  |........ALLDATA1|
00000050   5F303031 30000000 00000000 00000000  |_0010...........|
00000060   00000000 00000000 414C4C44 41544131  |........ALLDATA1|
00000070   00000000 00000000 00000000 00000000  |................|
00000080   00000000 00000000 414C4C44 41544131  |........ALLDATA1|
00000090   5F303031 30000000 00000000 00000000  |_0010...........|
000000A0   00000000 00000000 00000000 00000000  |................|
000000B0   00000000 00000000 00000000 00000000  |................|
000000C0   00000000 00000000 01F61B31 C21B2C00  |...........1..,.|
000000D0   01F61C52 34C96000 02001000 00100000  |...R4.`.........|
000000E0   0001BC80 00016800 00000002 00000001  |......h.........|
000000F0   00000002 00000000 00000000 00000000  |................|
```

NOTE: This doesn't work for OCR, Voting disk outside ASM control

© 2010 IBM Corporation

ASM Related Enhancements

- **IY95599: LVM SHOULD DETECT ORACLE RAW DISKS**
  - Included in AIX 5.3. TL07 and AIX 6.1
  - 'mkvg and 'extendvg' commands check for presence of ASM header before writing PVID information on disk
    - chdev pv=yes or pv=clear operations still do not check for ASM signature before overwriting PVID area

- **rendev command in AIX 6.1 TL6+ allows disks to be dynamically renamed – example:**
  - **`# rendev -l hdisk5 -n hdiskASM5`**
- **lkdev command locks device attributes from being changed with 'chdev'**
  - **`# lkdev -l hdisk5 -a`**
- Oracle 11.1.0.7+ makes an automatic backup copy of the ASM header and stores it in the 2nd ASM allocation unit. KFED can be used by Oracle support to recover the header.
- **....And one Setback.**
  - **IV09021 Using CSPOC with ASM devices can cause data loss**
  - **http://www-01.ibm.com/support/docview.wss?crawler=1&uid=isg1IV09021**

 © 2012 IBM Corporation

Other ASM Considerations

- **Disk Size**
- 2 TB storage per ASM disk *
  - Bugs in 11.2.0.1, 11.2.0.1 with disks > 1.1 TB, ML note 1095202.1
- **Disk Groups**
- **OCR and Voting Disks can now be stored in ASM as of 11gR2**
  - OCR/Voting Disks can NO LONGER be created on raw LVs. Upgrades supported.
- **Oracle Best Practice is to use the same diskgroup as Database**
- SAP best practice is to use a separate diskgroup, +OCR
  - COMPATIBLE.ASM must be 11.2.0.0
- **Can't stop ASM (unless stop the cluster) if OCR/Voting Disk in ASM**
  - Stopping ASM will cause the database instance to get restarted, ASM will not stop
  - Can't unmount diskgroup housing OCR/Voting disk

© 2010 IBM Corporation

Looking for Buffer Structure Shortages

```
# vmstat -v |tail -5        ← we only need to look at the last 5 lines
              0 pending disk I/Os blocked with no pbuf      ← if blocked on pbuf, increase
                                                              pv_min_pbuf (ioo restricted) and
                                                              varyoff/varyon VG
              0 paging space I/Os blocked with no psbuf     ← if blocked on psbuf, stop paging or
                                                              add more paging spaces
           2484 filesystem I/Os blocked with no fsbuf       ← if blocked on fsbuf (JFS), increase
                                                              numfsbufs (ioo restricted) to 1568
              0 client filesystem I/Os blocked with no fsbuf ← if blocked on client fsbuf (NFS) ,
                                                              increase nfso nfs_vX_pdts and
                                                              nfs_vX_vm_bufs values ("X" = 2,3, or 4)

              0 external pager filesystem I/Os blocked with no fsbuf ← if blocked on JFS2 fsbuf,
                        1) increase j2_dynamicBufferPreallocation (ioo) to 32 (or higher)
                        2) If that is not sufficient, increase  j2_nBufferPerPagerDevice (ioo restricted)
                           to 2048 and unmount/remount JFS2 filesystems
```

Collect "vmstat –v" output at multiple times and compare statistics –

We're only concerned about values that increase significantly over time

 © 2012 IBM Corporation

Agenda

AIX Configuration/Tuning for Oracle

- Memory
- CPU
- I/O
- Network
- Miscellaneous

 © 2012 IBM Corporation

Grid Interconnect Redundancy – Strongly Advised

- **Etherchannel (802.3ad)**
  - Up to 8 primary adapters per etherchannel
  - All adapters in the etherchannel should be the same, and should be configured identically (eg gigabit full duplex)
  - All adapters constituting an etherchannel must be connected to the same switch
  - Switch support for etherchannel (may be called “aggregation” or “trunking”) required
- **Network Interface Backup**
  - One per etherchannel
  - No switch requirements
  - Provides backup capability only, no increased throughput like etherchannel
- **Shared Ethernet Adapter (SEA) Failover**
  - VIOS based failover and load balancing
- **New in 11.2.0.2 – Oracle HAIP**
  - Up to 4 primary adapters can be aggregated for redundancy and bandwidth
  - Requires multicast communication on 230.0.1.0 network, although 224.0.0.251 will work with patch 9974223
  - 11.2.0.3 HAIP will work with broadcast communication

© 2012 IBM Corporation

Network parameters (no)

- **use_isno = 1 means any parameters set at the interface level override parameters set with 'no'**
  - DEFAULT (restricted) in AIX 6.1
- **If use_isno = 0, any parameters set with 'no' override interface-specific parameters**
- **If use_isno = 1, set parameters for each interface using 'ifconfig' or 'chdev'**
- **Refer to the following URL for a chart on appropriate interface-specific parameters:**
  - http://publib.boulder.ibm.com/infocenter/systems/topic/com.ibm.aix.prftungd/doc/prftungd/prftungd.pdf
- **Generally appropriate parameters for 1 or 10 Gigabit Ethernet Oracle public network interfaces:**
  - tcp_sendspace = 262144
  - tcp_recvspace = 262144
  - rfc1323 = 1
- **Examples:**

```
# no -p -o tcp_sendspace=262144
# no -p -o tcp_recvspace=262144
# no -p -o rfc1323=1
```

 © 2012 IBM Corporation

TCP/IP Ephemeral Ports

- **Oracle 11gR2 checks the current TCP/IP ephemeral port range at binary install time**
  - If the current tcp and udp ephemeral port ranges are not set to 9000 – 65500, the Oracle installer will generate warning messages
- **If the default range (32768 – 65535) is sufficient to support the anticipated server workload, the warnings may be ignored**
- **If the workload will require a high number of ephemeral ports, such as high node counts or heavy use of Parallel Query – or to avoid the install time warning messages, the ephemeral port ranges may be re-configured**
- **Examples:**

```
# no -p -o tcp_ephemeral_low=9000
# no -p -o tcp_ephemeral_high=65500
# no -p -o udp_ephemeral_low=9000
# no -p -o udp_ephemeral_high=65500
```

 © 2012 IBM Corporation

Additional Network Parameters for RAC:

The following parameters should be configured for RAC private interconnect:

- **udp_sendspace = db_block_size * db_file_multiblock_read_count +4k**
  - (not less than 65536)
- **udp_recvspace = 10 * udp_sendspace**
  - Must be < sb_max
  - Increase if buffer overflows occur
- **Use Jumbo Frames if supported at the switch layer**

- **CTSS (Cluster Time Synchronization Service) provided for environments where NTP is not available**
  - CTSS is an 'observer' where NTP/XNTP is present (recommended implementation on AIX)
  - Note that NTP/XNTP does NOT have to be actually working for CTSS to operate in observer mode

 © 2012 IBM Corporation

Miscellaneous parameters

- **/etc/security/limits**
  - Set to “-1” for everything except core for Oracle, grid and root users
    - Soft FILE size
    - Soft CPU time
    - Soft DATA segment
    - Soft STACK size
    - Soft Real Memory size
    - Processes (per user)
- **Sys0 maxuproc attribute**
  - Should be >= 16384
  - For workloads with a large number of concurrent connections an/or parallel servers, should be at least 128 plus the sum of PROCESSES and PARALLEL_MAXSERVERS for all instances in the LPAR
- **Environment variables:**

  AIXTHREAD_SCOPE=S

  LDR_CNTRL=DATAPSIZE=64K@TEXTPSIZE=64K@STACKPSIZE=64K
- **Use 64-bit AIX kernel**
- **Time synchronization – For RAC environments, use the xntpd “-x” flag**

 © 2012 IBM Corporation

References

- **Oracle Architecture and Tuning on AIX v2.20 (WP100883)**

  http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP100883
- **Diagnosing Oracle Database Performance on AIX Using IBM NMON and Oracle Statspack Reports (WP101720)**

  http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101720
- **IBM System p Advanced POWER Virtualization (PowerVM) Best Practices**

  http://www.redbooks.ibm.com/abstracts/redp4194.html?Open
- **Oracle Database Reference – 11g Release 2 (11.2) (E25513-01)**

  http://docs.oracle.com/cd/E11882_01/server.112/e25513.pdf
- **Oracle Database Installation Guide – 11g Release 2 (11.2) for IBM AIX on POWER Systems (64-bit) (E24332-01)**

  http://docs.oracle.com/cd/E11882_01/install.112/e24332.pdf
- **Power Systems Enterprise Servers with PowerVM Virtualization and RAS**

  http://www.redbooks.ibm.com/abstracts/sg247965.html?Open

 © 2012 IBM Corporation

References…

- **AIXpert Blog on Local, Near and Far Memory**

  https://www.ibm.com/developerworks/mydeveloperworks/blogs/aixpert/entry/local_near_far_memory_part_1_large_power7_boxes_more_local_memory26?lang=en

- **Oracle Database and 1 TB Segment Aliasing (TD105761)**

  http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/TD105761

- **IBM EnergyScale for POWER7 Processor-Based Systems**

  ftp://public.dhe.ibm.com/common/ssi/ecm/en/pow03039usen/POW03039USEN.PDF

- **Active Memory Expansion: Overview and Usage Guide**

  ftp://ftp.software.ibm.com/common/ssi/sa/wh/n/pow03037usen/POW03037USEN.PDF

- **IBM PowerVM Virtualization Active Memory Sharing**

  http://www.redbooks.ibm.com/abstracts/redp4470.html?Open

 © 2012 IBM Corporation

 © 2012 IBM Corporation

Special Notices

This document was developed for IBM offerings in the United States as of the date of publication. IBM may not make these offerings available in other countries, and the information is subject to change without notice. Consult your local IBM business contact for information on the IBM offerings available in your area.

Information in this document concerning non-IBM products was obtained from the suppliers of these products or other public sources. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

IBM may have patents or pending patent applications covering subject matter in this document. The furnishing of this document does not give you any license to these patents. Send license inquires, in writing, to IBM Director of Licensing, IBM Corporation, New Castle Drive, Armonk, NY 10504-1785 USA.

All statements regarding IBM future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

The information contained in this document has not been submitted to any formal IBM test and is provided "AS IS" with no warranties or guarantees either expressed or implied.

All examples cited or described in this document are presented as illustrations of the manner in which some IBM products can be used and the results that may be achieved. Actual environmental costs and performance characteristics will vary depending on individual client configurations and conditions.

IBM Global Financing offerings are provided through IBM Credit Corporation in the United States and other IBM subsidiaries and divisions worldwide to qualified commercial and government clients. Rates are based on a client's credit rating, financing terms, offering type, equipment type and options, and may vary by country. Other restrictions may apply. Rates and offerings are subject to change, extension or withdrawal without notice.

IBM is not responsible for printing errors in this document that result in pricing or information inaccuracies.

All prices shown are IBM's United States suggested list prices and are subject to change without notice; reseller prices may vary.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

Any performance data contained in this document was determined in a controlled environment. Actual results may vary significantly and are dependent on many factors including system hardware configuration and software design and configuration. Some measurements quoted in this document may have been made on development-level systems. There is no guarantee these measurements will be the same on generally-available systems. Some measurements quoted in this document may have been estimated through extrapolation. Users of this document should verify the applicable data for their specific environment.

 © 2012 IBM Corporation

Enabling Large Pages for Oracle SGA

If you MUST do it, the following is required to implement:

- **Oracle**
  - LOCK_SGA = TRUE
- **AIX**
  - *Calculate required # of large pages* = INT [(SGA size – 1) / 16 MB)] + 1
  - # vmo –r –o lgpg_regions = <*no_of_large_pages*> -o lgpg_size=16777216
  - # vmo –p –o v_pinshm = 1
  - # chuser capabilities = CAP_NUMA_ATTACH, CAP_BYPASS_RAC_VMM, CAP_PROPAGATE oracle
    leave maxpin% at 80 (Default)
- **To verify that Oracle is using large pages:**
  - svmon -P $(ps -elf | egrep " ora_smon_${ORACLE_SID} " | grep -v egrep | awk '{print $4}') | grep shmat
  - Part of output will look similar to the output below (Note the L) which indicates you are using large pages:

```
5390db9 70000019 work default shmat/mmap L 16 16 0 16
 4670ce7 7000000c work default shmat/mmap L 16 16 0 16
 821002 7000001e work default shmat/mmap L 16 16 0 16
 7e80f68 70000005 work default shmat/mmap L 16 16 0 16
 7e50f65 7000001b work default shmat/mmap L 16 16 0 16
 7d90f59 7000001a work default shmat/mmap L 16 16 0 16
 7d30f53 70000014 work default shmat/mmap L 16 16 0 16
```

 © 2012 IBM Corporation

Verifying Oracle Large Page Usage

Use ‘svmon’ to verify that Oracle is using large pages:

```
# svmon -P $(ps -elf | egrep " ora_smon_${ORACLE_SID} " | grep -v egrep |
  awk '{print $4}') | grep shmat
5390db9 70000019 work default shmat/mmap L 16 16 0 16
4670ce7 7000000c work default shmat/mmap L 16 16 0 16
 821002 7000001e work default shmat/mmap L 16 16 0 16
7e80f68 70000005 work default shmat/mmap L 16 16 0 16
7e50f65 7000001b work default shmat/mmap L 16 16 0 16
7d90f59 7000001a work default shmat/mmap L 16 16 0 16
7d30f53 70000014 work default shmat/mmap L 16 16 0 16
```

The “L” indicates that the shared memory segment is using Large pages

 © 2012 IBM Corporation

Special Notices (Cont'd)

IBM, the IBM logo, ibm.com AIX, AIX (logo), AIX 6 (logo), AS/400, BladeCenter, Blue Gene, ClusterProven, DB2, ESCON, i5/OS, i5/OS (logo), IBM Business Partner (logo), IntelliStation, LoadLeveler, Lotus, Lotus Notes, Notes, Operating System/400, OS/400, PartnerLink, PartnerWorld, PowerPC, pSeries, Rational, RISC System/6000, RS/6000, THINK, Tivoli, Tivoli (logo), Tivoli Management Environment, WebSphere, xSeries, z/OS, zSeries, AIX 5L, Chiphopper, Chipkill, Cloudscape, DB2 Universal Database, DS4000, DS6000, DS8000, EnergyScale, Enterprise Workload Manager, General Purpose File System, , GPFS, HACMP, HACMP/6000, HASM, IBM Systems Director Active Energy Manager, iSeries, Micro-Partitioning, POWER, PowerExecutive, PowerVM, PowerVM (logo), PowerHA, Power Architecture, Power Everywhere, Power Family, POWER Hypervisor, Power Systems, Power Systems (logo), Power Systems Software, Power Systems Software (logo), POWER2, POWER3, POWER4, POWER4+, POWER5, POWER5+, POWER6, System i, System p, System p5, System Storage, System x, System z, Tivoli Enterprise, TME 10, Workload Partitions Manager and X-Architecture are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. If these and other IBM trademarked terms are marked on their first occurrence in this information with a trademark symbol (® or ™), these symbols indicate U.S. registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at www.ibm.com/legal/copytrade.shtml

The Power Architecture and Power.org wordmarks and the Power and Power.org logos and related marks are trademarks and service marks licensed by Power.org.

UNIX is a registered trademark of The Open Group in the United States, other countries or both.

Linux is a registered trademark of Linus Torvalds in the United States, other countries or both.

Microsoft, Windows and the Windows logo are registered trademarks of Microsoft Corporation in the United States, other countries or both.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

AMD Opteron is a trademark of Advanced Micro Devices, Inc.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

TPC-C and TPC-H are trademarks of the Transaction Performance Processing Council (TPPC).

SPECint, SPECfp, SPECjbb, SPECweb, SPECjAppServer, SPEC OMP, SPECviewperf, SPECapc, SPEChpc, SPECjvm, SPECmail, SPECimap and SPECsfs are trademarks of the Standard Performance Evaluation Corp (SPEC).

NetBench is a registered trademark of Ziff Davis Media in the United States, other countries or both.

AltiVec is a trademark of Freescale Semiconductor, Inc.

Cell Broadband Engine is a trademark of Sony Computer Entertainment Inc.

InfiniBand, InfiniBand Trade Association and the InfiniBand design marks are trademarks and/or service marks of the InfiniBand Trade Association.

Other product and service names might be trademarks of IBM or other companies.

 © 2012 IBM Corporation

Notes on benchmarks and values

The IBM benchmarks results shown herein were derived using particular, well configured, development-level and generally-available computer systems. Buyers should consult other sources of information to evaluate the performance of systems they are considering buying and should consider conducting application oriented testing. For additional information about the benchmarks, values and systems tested, contact your local IBM office or IBM authorized reseller or access the Web site of the benchmark consortium or benchmark vendor.

IBM benchmark results can be found in the IBM Power Systems Performance Report at http://www.ibm.com/systems/p/hardware/system_perf.html .

All performance measurements were made with AIX or AIX 5L operating systems unless otherwise indicated to have used Linux. For new and upgraded systems, AIX Version 4.3, AIX 5L or AIX 6 were used. All other systems used previous versions of AIX. The SPEC CPU2006, SPEC2000, LINPACK, and Technical Computing benchmarks were compiled using IBM's high performance C, C++, and FORTRAN compilers for AIX 5L and Linux. For new and upgraded systems, the latest versions of these compilers were used: XL C Enterprise Edition V7.0 for AIX, XL C/C++ Enterprise Edition V7.0 for AIX, XL FORTRAN Enterprise Edition V9.1 for AIX, XL C/C++ Advanced Edition V7.0 for Linux, and XL FORTRAN Advanced Edition V9.1 for Linux. The SPEC CPU95 (retired in 2000) tests used preprocessors, KAP 3.2 for FORTRAN and KAP/C 1.4.2 from Kuck & Associates and VAST-2 v4.01X8 from Pacific-Sierra Research. The preprocessors were purchased separately from these vendors. Other software packages like IBM ESSL for AIX, MASS for AIX and Kazushige Goto's BLAS Library for Linux were also used in some benchmarks.

For a definition/explanation of each benchmark and the full list of detailed results, visit the Web site of the benchmark consortium or benchmark vendor.

TPC	**http://www.tpc.org**
SPEC	**http://www.spec.org**
LINPACK	**http://www.netlib.org/benchmark/performance.pdf**
Pro/E	**http://www.proe.com**
GPC	**http://www.spec.org/gpc**
VolanoMark	**http://www.volano.com**
STREAM	**http://www.cs.virginia.edu/stream/**
SAP	**http://www.sap.com/benchmark/**
Oracle Applications	**http://www.oracle.com/apps_benchmark/**
PeopleSoft - To get information on PeopleSoft benchmarks, contact PeopleSoft directly	
Siebel	**http://www.siebel.com/crm/performance_benchmark/index.shtm**
Baan	**http://www.ssaglobal.com**
Fluent	**http://www.fluent.com/software/fluent/index.htm**
TOP500 Supercomputers	**http://www.top500.org/**
Ideas International	**http://www.ideasinternational.com/benchmark/bench.html**
Storage Performance Council	**http://www.storageperformance.org/results**

 © 2012 IBM Corporation

Notes on HPC benchmarks and values

The IBM benchmarks results shown herein were derived using particular, well configured, development-level and generally-available computer systems. Buyers should consult other sources of information to evaluate the performance of systems they are considering buying and should consider conducting application oriented testing. For additional information about the benchmarks, values and systems tested, contact your local IBM office or IBM authorized reseller or access the Web site of the benchmark consortium or benchmark vendor.

IBM benchmark results can be found in the IBM Power Systems Performance Report at http://www.ibm.com/systems/p/hardware/system_perf.html .

All performance measurements were made with AIX or AIX 5L operating systems unless otherwise indicated to have used Linux. For new and upgraded systems, AIX Version 4.3 or AIX 5L were used. All other systems used previous versions of AIX. The SPEC CPU2000, LINPACK, and Technical Computing benchmarks were compiled using IBM's high performance C, C++, and FORTRAN compilers for AIX 5L and Linux. For new and upgraded systems, the latest versions of these compilers were used: XL C Enterprise Edition V7.0 for AIX, XL C/C++ Enterprise Edition V7.0 for AIX, XL FORTRAN Enterprise Edition V9.1 for AIX, XL C/C++ Advanced Edition V7.0 for Linux, and XL FORTRAN Advanced Edition V9.1 for Linux. The SPEC CPU95 (retired in 2000) tests used preprocessors, KAP 3.2 for FORTRAN and KAP/C 1.4.2 from Kuck & Associates and VAST-2 v4.01X8 from Pacific-Sierra Research. The preprocessors were purchased separately from these vendors. Other software packages like IBM ESSL for AIX, MASS for AIX and Kazushige Goto's BLAS Library for Linux were also used in some benchmarks.

For a definition/explanation of each benchmark and the full list of detailed results, visit the Web site of the benchmark consortium or benchmark vendor.

SPEC	**http://www.spec.org**
LINPACK	**http://www.netlib.org/benchmark/performance.pdf**
Pro/E	**http://www.proe.com**
GPC	**http://www.spec.org/gpc**
STREAM	**http://www.cs.virginia.edu/stream/**
Fluent	**http://www.fluent.com/software/fluent/index.htm**
TOP500 Supercomputers	**http://www.top500.org/**
AMBER	**http://amber.scripps.edu/**
FLUENT	**http://www.fluent.com/software/fluent/fl5bench/index.htm**
GAMESS	**http://www.msg.chem.iastate.edu/gamess**
GAUSSIAN	**http://www.gaussian.com**
ANSYS	**http://www.ansys.com/services/hardware-support-db.htm**
	Click on the "Benchmarks" icon on the left hand side frame to expand. Click on "Benchmark Results in a Table" icon for benchmark
results.	
ABAQUS	**http://www.simulia.com/support/v68/v68_performance.php**
ECLIPSE	**http://www.sis.slb.com/content/software/simulation/index.asp?seg=geoquest&**
MM5	**http://www.mmm.ucar.edu/mm5/**
MSC.NASTRAN	**http://www.mscsoftware.com/support/prod%5Fsupport/nastran/performance/v04_sngl.cfm**
STAR-CD	**www.cd-adapco.com/products/STAR-CD/performance/320/index/html**
NAMD	**http://www.ks.uiuc.edu/Research/namd**
HMMER	**http://hmmer.janelia.org/**
	http://powerdev.osuosl.org/project/hmmerAltivecGen2mod

 © 2012 IBM Corporation

Notes on performance estimates

rPerf for AIX

rPerf (Relative Performance) is an estimate of commercial processing performance relative to other IBM UNIX systems. It is derived from an IBM analytical model which uses characteristics from IBM internal workloads, TPC and SPEC benchmarks. The rPerf model is not intended to represent any specific public benchmark results and should not be reasonably used in that way. The model simulates some of the system operations such as CPU, cache and memory. However, the model does not simulate disk or network I/O operations.

rPerf estimates are calculated based on systems with the latest levels of AIX and other pertinent software at the time of system announcement. Actual performance will vary based on application and configuration specifics. The IBM eServer pSeries 640 is the baseline reference system and has a value of 1.0. Although rPerf may be used to approximate relative IBM UNIX commercial processing performance, actual system performance may vary and is dependent upon many factors including system hardware configuration and software design and configuration. Note that the rPerf methodology used for the POWER6 systems is identical to that used for the POWER5 systems. Variations in incremental system performance may be observed in commercial workloads due to changes in the underlying system architecture.

All performance estimates are provided "AS IS" and no warranties or guarantees are expressed or implied by IBM. Buyers should consult other sources of information, including system benchmarks, and application sizing guides to evaluate the performance of a system they are considering buying. For additional information about rPerf, contact your local IBM office or IBM authorized reseller.

==========

CPW for IBM i

Commercial Processing Workload (CPW) is a relative measure of performance of processors running the IBM i operating system. Performance in customer environments may vary. The value is based on maximum configurations. More performance information is available in the Performance Capabilities Reference at:
www.ibm.com/systems/i/solutions/perfmgmt/resource.html

 © 2012 IBM Corporation

IOPS/bandwidth examples

Adapter	*IOPS (4K)*	*Sustained Sequential Bandwidth*
2 Gbps FC adapter	38,461	198 MB/s simplex
4 Gbps PCI-E FC adapter		400 MB/s simplex
8 Gbps FC dual port PCI-e	142,000	750 MB/s per port simplex

Disk Drive	*Speed*	*Rotational Latency*	*Avg. Seek Time*	*IOPS*
15K 3.5” FC	15K rpm	2 ms	3.5 ms	182
10K 3.5” FC	10K rpm	3 ms	4.5 ms	133
15K 2.5” SAS	15K rpm	2 ms	3.1 ms	196
7.2K SATA2	7200 rpm	4.2 ms	9 ms	76

 © 2012 IBM Corporation

Displaying Memory Usage Statistics

The 'svmon -G' command provides information on current memory usage per page size: (general numbers are reported in 4K pages)

```
# svmon -G
               size       inuse        free         pin      virtual
memory      1179648      926225      290287      493246       262007
pg space    1572864        5215

               work       pers       clnt       other
pin           91390          0          0       74176
in use       258573        4316     335656

PageSize   PoolSize       inuse        pgsp         pin      virtual
s    4 KB         -      477713        5215       94606       141175
m   64 KB         -        7552           0        4435         7552
L   16 MB        80           0           0          80            0
```

Free 16M pages = PoolSize - inuse

 © 2012 IBM Corporation

Local, Near and Far Memory

- **Power Systems use a "shared memory" model**
  - any processor has access to part of memory
- **High-end Power Systems (e.g. p770, p780, p795) use multiple building blocks (CECs) to scale capacity**
  - Each building block has its own set of processor and memory chips
  - Building blocks are interconnected via a switched communications fabric
- **The closer the memory is to the processor accessing it, the faster the memory access**
  - Local Memory: Directly attached to the chip's memory controller
  - Near Memory: On an adjacent chip, accessed via intra-node communication paths
  - Far Memory: On a different CEC drawer, accessed via inter-node communication paths

Model	Local	Near	Far
Power 710/730	Same Chip	Other Chip	n/a
Power 720/740	Same Chip	Other Chip	n/a
Power 750	Same Chip	Other Chip	n/a
Power 770/780	Same Chip	Other Chip, Same CEC	Different CEC
Power 795	Same Chip	Other Chip, Same CEC	Different CEC

 © 2012 IBM Corporation

Oracle Memory and Memory Affinity

- **Oracle SGA is "striped" across all the available memory in the LPAR**
  - If the LPAR configuration has a combination of near, local and far memory allocated to it, SGA will be (more or less) evenly spread across all of it
  - The greater the number of CECs involved, the greater the likelihood of remote memory accesses
- **Oracle PGA for a given process tends to be allocated in the near memory of the processor that process was running on when the memory was allocated**
  - The AIX dispatcher will attempt to maintain affinity between a given process and the processor that process gets scheduled on
  - rsets may (optionally) be used to force affinity to a subset of available processors (e.g. those on a given chip, or within a given CEC), although this could potentially cause dispatching delays in heavily loaded environments
- **vmo enhanced_affinity_private (vmo restricted parameter)**
  - The percentage of application data that is to be allocated local, with the remaining memory to be striped across all available memory in the LPAR
  - Default value is 20% in AIX 6.1 TL5 and 40% in AIX 6.1 TL6+ and AIX 7.1

 © 2012 IBM Corporation

Displaying the LPAR CPU & Memory Configuration

- **The 'lssrad -va' command displays a summary of the way physical processors and memory is allocated for a given LPAR:**

```
# lssrad -va
REF1        SRAD        MEM         CPU
0
                0       110785.25   0-31
                1       125665.00   32-63
1
                2       17430.00    64-95
                3       0.00        96-127
2
                4       0.00        128-159
                5       0.00        160-191
```

Note the extremely poor distribution of memory in this p780 LPAR example:

- **3 of 6 SRADs have no local memory at all and 2 of them only have access to far memory in other CECs**

  - REF1: The CEC number. e.g. 0 to 3 on p770/p80, 0 to 7 on p795
  - SRAD: A Scheduler Resource Affinity Domain, i.e. an individual group of processors that all reside on the same chip
  - MEM: The amount of local memory (in Megabytes) allocated to the SRAD
  - CPU: The logical CPUs within the SRAD, e.g. with SMT4 enabled, 0-3 would be for the first physical CPU, 4-7 would be for the second physical CPU, etc…

 © 2012 IBM Corporation

Displaying the Local, Near and Far Memory Access Profile

- **The 'mpstat –d' command displays statistics on local, near and far memory accesses for every logical CPU in an LPAR:**

```
# mpstat -d 1 60
cpu . . .      S3hrd      S4hrd      S5hrd  . . .
  0 . . .      99.5       0.1        0.4    . . .
  1 . . .      99.5       0.1        0.4    . . .
  2 . . .      98.1       0.5        1.4    . . .
  3 . . .      98.0       0.6        1.5    . . .
  4 . . .      99.5       0.1        0.4    . . .
  5 . . .      99.5       0.1        0.4    . . .
  6 . . .      98.1       0.5        1.4    . . .
  7 . . .      98.0       0.6        1.4    . . .
...
184 . . .      99.2       0.2        0.6    . . .
185 . . .      98.7       0.1        1.3    . . .
186 . . .      48.3       0.0        51.7   . . .
187 . . .      45.5       0.0        54.5   . . .
188 . . .      99.2       0.2        0.6    . . .
189 . . .      98.7       0.1        1.2    . . .
190 . . .      50.9       0.0        49.1   . . .
191 . . .      47.4       0.0        52.6   . . .
```

Excellent affinity (high percentage of local memory accesses) for the logical CPUs in the first CEC/SRAD

Poor affinity (high percentage of remote memory accesses) for the 3rd and 4th SMT threads for the physical CPUs in the last CEC/SRAD

- S3hrd: The percentage of local thread dispatches on this logical CPU
- S4hrd: The percentage of near thread dispatches on this logical CPU
- S5hrd: The percentage of far thread dispatches on this logical CPU
- '-': Indicates an SMT thread that is not currently active

 © 2012 IBM Corporation

Help the Hypervisor do its job

- **Stay current on Firmware** (e.g. AM720_101 or later) to avoid any known CPU/memory allocation or virtual processor dispatching issues
- Where appropriate, consider LPAR boot order to ensure high priority get optimal choice of the available CPUs and memory
- Consider the allocation of CPUs and memory.

 © 2012 IBM Corporation

Oracle Memory Structures Allocation

- **11g : Automatic Memory Management (AMM)**
  - memory_target (dynamic parameter) – specifies the total memory size to be used by the instance SGA and PGA. Exchanges between SGA and PGA are done according to workload requirements
  - If sga_target and pga_aggregate_target are not set, the policy is to give 60% of memory_target to the SGA and 40% to the PGA
  - memory_max_target (static parameter) – specifies the maximum memory size for the instance
  - *To use Automatic Memory Management, memory_target must be >0 and LOCK_SGA=false*
- ***See Metalink notes 443746.1 and 452512.1 explaining AMM and these new parameters***

AMM dynamic resizing of the shared pool can cause a fair amount of "cursor: pin s" wait time. One strategy to minimize this is to set minimum sizes for memory areas you particularly care about.

In addition you can change the frequency how often AMM analyzes and adjusts the memory distribution. See: Metalink note: 742599.1 (_memory_broker_stat_interval)

 © 2012 IBM Corporation

lparstat Command

root@lpn1 / # lparstat -i

```
Node Name                                  : lpn1
Partition Name                             : RACp2_node1
Partition Number                           : 4
Type                                       : Shared-SMT
Mode                                       : Uncapped
Entitled Capacity                          : 0.40
Partition Group-ID                         : 32772
Shared Pool ID                             : 0
Online Virtual CPUs                        : 4
Maximum Virtual CPUs                       : 4
Minimum Virtual CPUs                       : 1
Online Memory                              : 4608 MB
Maximum Memory                             : 5120 MB
Minimum Memory                             : 128 MB
Variable Capacity Weight                   : 128
Minimum Capacity                           : 0.10
Maximum Capacity                           : 4.00
Capacity Increment                         : 0.01
Maximum Physical CPUs in system            : 4
Active Physical CPUs in system             : 4
Active CPUs in Pool                        : 4
Shared Physical CPUs in system             : 4
Maximum Capacity of Pool                   : 400
Entitled Capacity of Pool                  : 280
Unallocated Capacity                       : 0.00
Physical CPU Percentage                    : 10.00%
Unallocated Weight                         : 0
```

 © 2012 IBM Corporation

I/O Options (ioo) Command

- **The AIX "ioo" command provides for the display and/or update of parameters which influence the way AIX manages physical memory**
  - The "-a" option displays parameter settings
    - ioo *[-p] [-r]* -a display permanent, reboot or current settings
    - ioo *[-p] [-r]* -F -a include restricted parameter settings
    - ioo *[-F]* -L display in tabular format
  - The "-o" option is used to change parameter values

    # ioo -o j2_maxPageReadAhead=256
  - The "-p" option is used to make changes persist across a reboot

    # ioo -p -o j2_maxPageReadAhead=256

 © 2012 IBM Corporation

JFS/JFS2 environments - Cached vs. non-Cached (Direct) I/O

File System caching tends to benefit heavily sequential workloads with low write content due to sequential read ahead. To enable caching for JFS2:

- Use default filesystem mount options
- Set Oracle filesystemio_options=ASYNCH (default)

DIO tends to benefit heavily random access workloads and CIO tends to benefit heavy update workloads. To disable JFS2 caching:

- In 9i, set filesystemio_options=ASYNCH and use dio (JFS) or cio (JFS2) mount
- In 10g/11g
  - If Oracle files do not need to be concurrently accessed by external utilities, set filesystemio_options=SETALL
  - Otherwise set filesystemio_options=ASYNCH and use dio (JFS) or cio (JFS2) mount
  - Starting with 11.2.0.2 (and AIX 6.1), an O_CIOR call is used . Use filesystemio_options=SETALL and do NOT use dio or cio mount options

When using DIO/CIO, fs buffer cache isn't used. Consider the following Oracle DB changes:

- Increase db_cache_size
- Increase db_file_multiblock_read_count

- **Read Metalink Note #s 272520.1, 257338.1, 360287.1**

 © 2012 IBM Corporation

Routing Table Entry Locking

- **There are 2 alternative locking strategies for Routing Table entries (rtentry) – simple and complex**
  - The current default locking strategy is “simple”
- **The Simple Performance Lock Analysis Tool (splat) may be used to monitor rtentry lock performance**
- **The complex locking strategy can improve performance when there is a lot of activity on Routing Table entries**
  - Can be enabled by setting rtentry_lock_complex=1 (recommended)
- **Example:**

```
# no -p -o rtentry_lock_complex=1
```

 © 2012 IBM Corporation

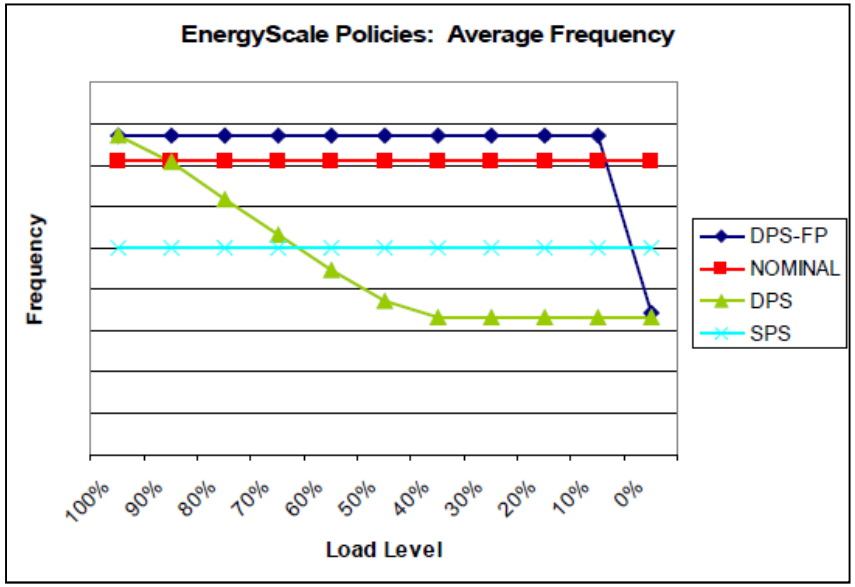
Oracle RAC OPROCD Reboots

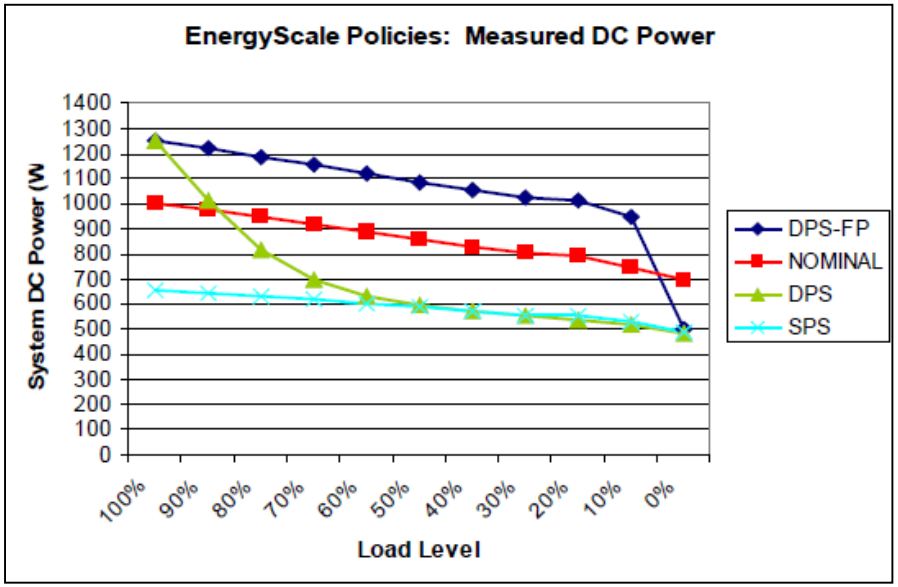
- **/etc/init.cssd**
  - OPROCD_DEFAULT_INTERVAL=1000 (milliseconds)
  - OPROCD_DEFAULT_MARGIN=500 (milliseconds)
  - Implementing “diagwait 13” will increase margin to 10 seconds
- **351374.1**: CRS Keep Rebooting Node after 10.1.0.4 patchset
  - Bug # 4502494
- **360497.1**: OPROCD Reboots Node when Time Is Set Back by XNTPD
  - Bug # 5015469 (affects 10.1.0.5, 10.2.0.2)
- **IY84564**: CPU BRINGUP IS SLOW
- **I/O pacing**
  - chgsys -l sys0 -a maxpout=8193 minpout=4096 (AIX 6.1 defaults)
  - nfso –o nfs_iopace_pages=1024

 © 2012 IBM Corporation

Active Energy Manager (AEM)

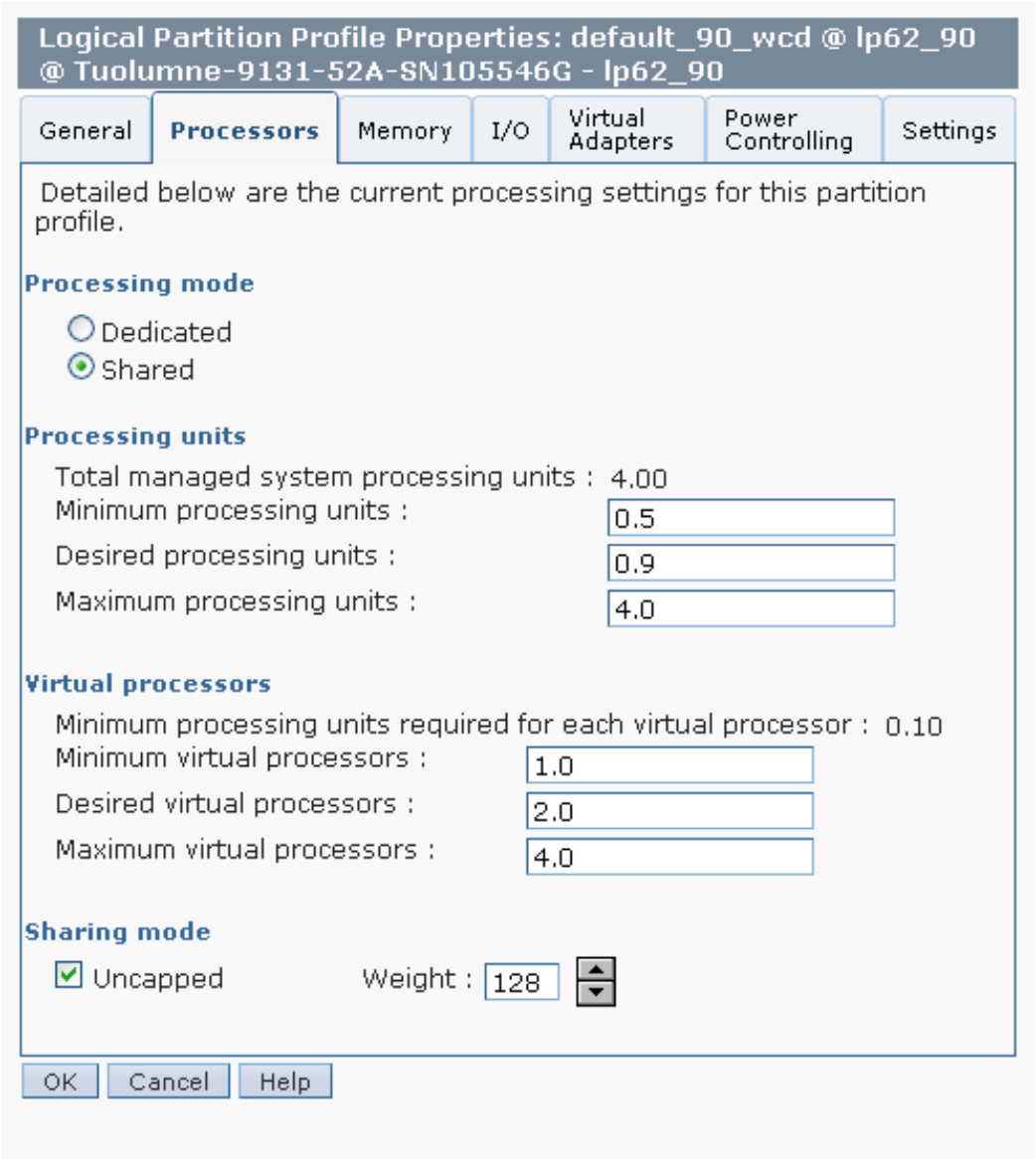
- **The IBM Systems Director Active Energy Manager (AEM) may be used to monitor and configure energy management features on IBM servers and storage**
- **If AEM is available, enabling the Dynamic Power Saver – Favor Performance (DPS-FP) can be used to balance power usage and processor performance**
  - Provides for CPU overclocking for periods of moderate to high CPU demand
  - Provides for significant energy savings during relatively idle periods
- **Much of the performance benefit of "TurboCore" while running in "MaxCore" mode**



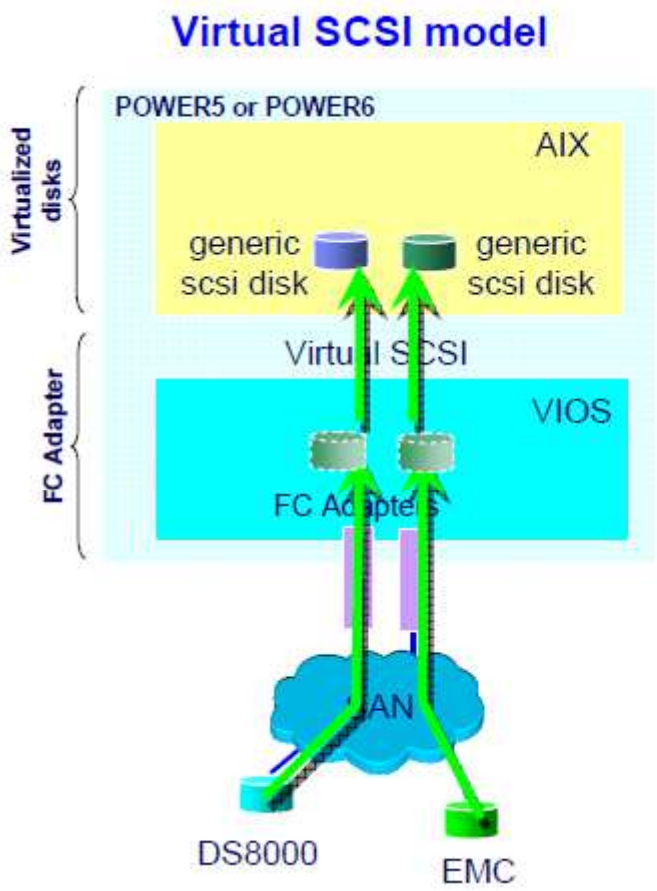


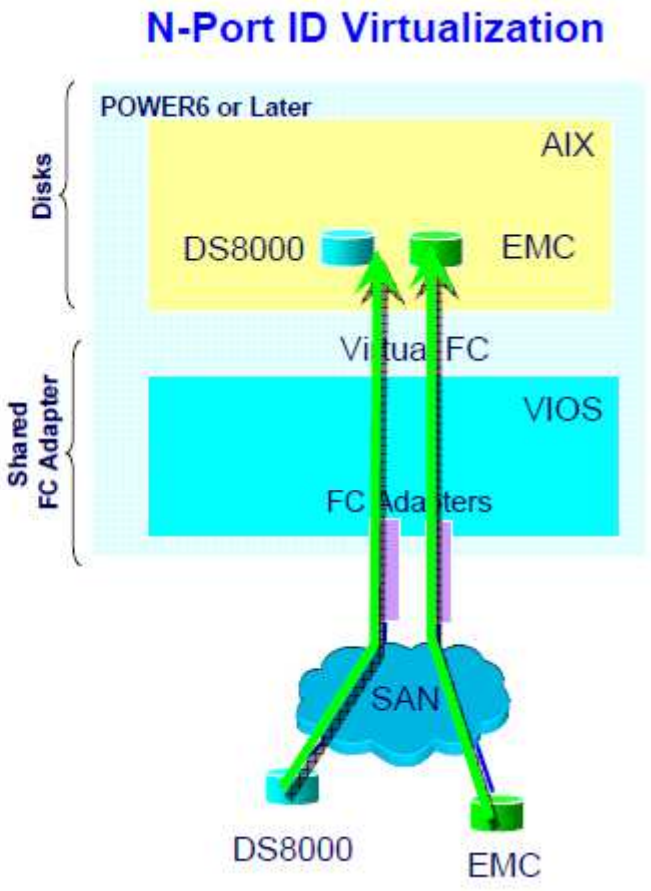
 © 2012 IBM Corporation

Configuring virtual processors - HMC



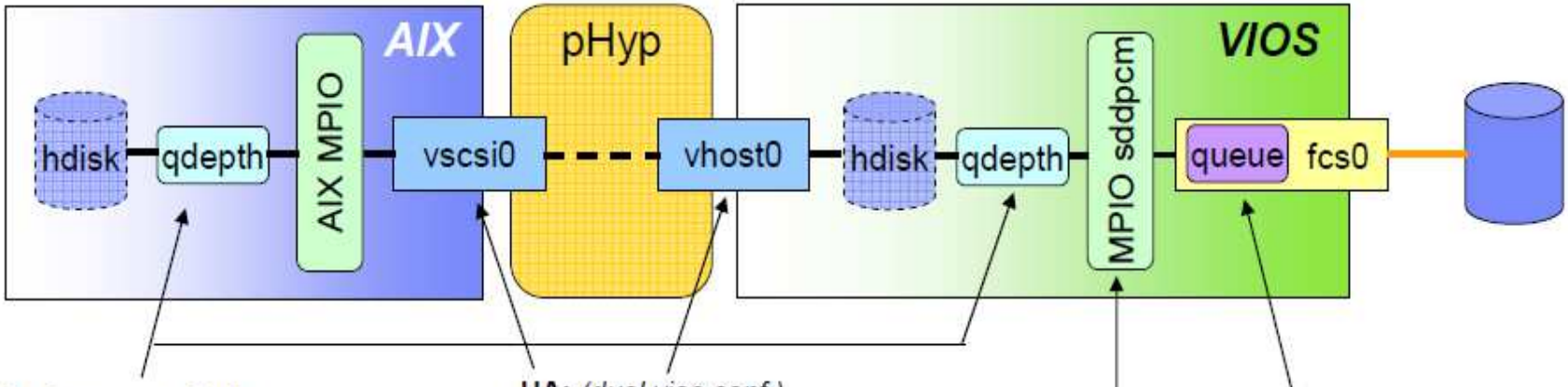
 © 2012 IBM Corporation





 © 2012 IBM Corporation

VIOS : VSCSI IO tuning




```
# lsdev -Cc disk
hdisk0 Available Virtual SCSI Disk Drive
```

Storage driver cannot be installed on the lpar
⇒ Default qdepth=3 !!! **Bad performance**

⇒ **`# chdev -l hdisk0 -a queue_depth=20`**
+ monitor svctime/wait time with nmon to adjust queue depth

* You have to set same queue_depth for the source hdisk on the VIOS

HA: *(dual vios conf.)*
change vscsi_err_recov to fast_fail
`# chdev -l vscsi0 -a vscsi_err_recov=fast_fail`

Performance:
No perf tuning can be made;

We just know that each vscsi can handle
512 cmd_elems. *(2 are reserved for the adapter and 3 reserved for each vdisk)*

So, use the following formula to find the number of disk you can attache behind a vscsi adapter.

 # luns = (512-2)/(q+3)
Where q=qdepth of each disk

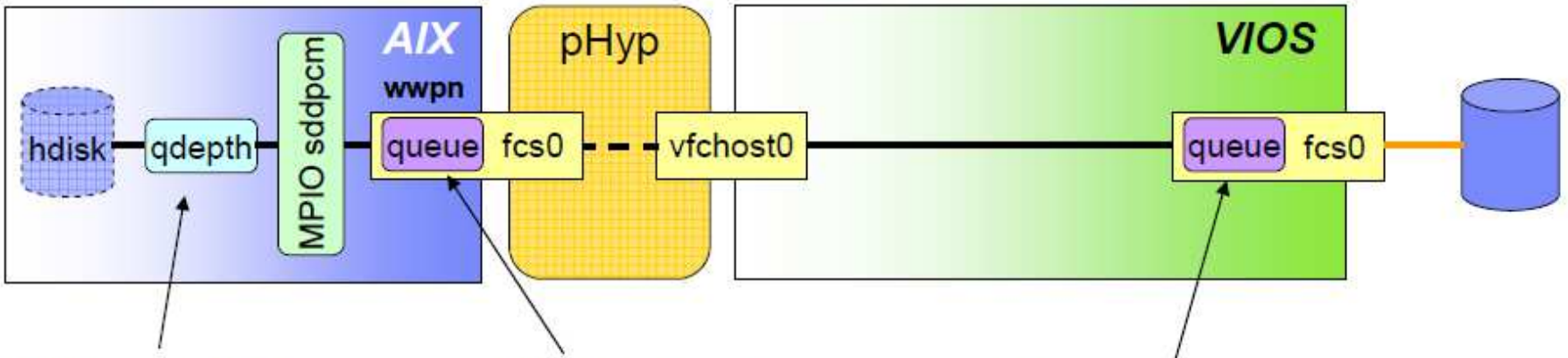
Install Storage Subsystem Driver on the VIO

Monitor VIO FC activity with nmon
(interactive: press "a" or "^")
(reccord with "-^" option)

Adapt num_cmd_elems accordingly with sum of qdepth.
check with: **`# fcstat fcsX`**

© Copyright IBM Corporation 2011
© 2012 IBM Corporation

VIOS: NPIV IO tuning




```
# lsdev -Cc disk
```
hdisk0 Available MPIO FC 2145

Storage driver must be installed on the lpar

⇒Default qdepth is set by the drivers

⇒Monitor "svctime" / "wait time" with nmon or iostat to tune the queue depth

HA: *(dual vios conf.)*
Change the following parameters:

```
# chdev -l fscsi0 -a
fc_err_recov=fast_fail
# chdev -l fscsi0 -a
dyntrk=yes
```

Performance:
Monitor fc activity with nmon
(interactive: option "a" or "^")
(reccording : option "-^")

Adapt num_cmd_elems
Check fcstat fcsX

HA:
Change the following parameters:

```
# chdev -l fscsi0 -a
fc_err_recov=fast_fail
# chdev -l fscsi0 -a
dyntrk=yes
```

Performance:
Monitor fc activity with nmon
(interactive: **option "^" only**)
(reccording : option "-^")

Adapt num_cmd_elems
Check fcstat fcsX

Should = sum of vfcs num_cmd_elems connected to the backend device

© Copyright IBM Corporation 2011
© 2012 IBM Corporation

Hardware Prefetch (POWER7)

- **The Data Stream Control Register (DSCR) controls the hardware streams behavior on the system**
- **The 'dscrctl' command may be used to display the current DSCR settings:**

```
# dscrctl -q
    Current DSCR settings:
            Data Streams Version  = V2.06
            number_of_streams  = 16
            platform_default_pd  = 0x5 (DPFD_DEEP)
            os_default_pd        = 0x5 (DPFD_DEEP)
```

- **For Oracle workloads, it may be beneficial to disable hardware prefetch:**

```
    # dscrctl -b -n -s 1
    # dscrctl -q
    Current DSCR settings:
            Data Streams Version  = V2.06
            number_of_streams  = 16
            platform_default_pd  = 0x5 (DPFD_DEEP)
            os_default_pd        = 0x1 (DPFD_NONE)
```

 © 2012 IBM Corporation