IBM Advanced Technical Support - Americas

AIX Configuration and Tuning for Oracle

Stephen Poon
IBM Advanced Technical Support
February 22, 2007

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.

Revised September 26, 2006

Special notices (cont.)

The following terms are registered trademarks of International Business Machines Corporation in the United States and/or other countries: AIX, AIX/L, AIX/L(logo), alphaWorks, AS/400, BladeCenter, Blue Gene, Blue Lightning, C Set++, CICS, CICS/6000, ClusterProven, CT/2, DataHub, DataJoiner, DB2, DEEP BLUE, developerWorks, DirectTalk, Domino, DYNIX, DYNIX/ptx, e business(logo), e(logo)business, e(logo)server, Enterprise Storage Server, ESCON, FlashCopy, GDDM, i5/OS, IBM, IBM(logo), ibm.com, IBM Business Partner (logo), Informix, IntelliStation, IQ-Link, LANStreamer, LoadLeveler, Lotus, Lotus Notes, Lotusphere, Magstar, MediaStreamer, Micro Channel, MQSeries, Net.Data, Netfinity, NetView, Network Station, Notes, NUMA-Q, Operating System/2, Operating System/400, OS/2, OS/390, OS/400, Parallel Sysplex, PartnerLink, PartnerWorld, Passport Advantage, POWERparallel, Power PC 603, Power PC 604, PowerPC, PowerPC(logo), Predictive Failure Analysis, pSeries, PTX, ptx/ADMIN, RETAIN, RISC System/6000, RS/6000, RT Personal Computer, S/390, Scalable POWERparallel Systems, SecureWay, Sequent, ServerProven, SpaceBall, System/390, The Engines of e-business, THINK, Tivoli, Tivoli(logo), Tivoli Management Environment, Tivoli Ready(logo), TME, TotalStorage, TURBOWAYS, VisualAge, WebSphere, xSeries, z/OS, zSeries.

The following terms are trademarks of International Business Machines Corporation in the United States and/or other countries: Advanced Micro-Partitioning, AIX 5L, AIX PVMe, AS/400e, Chiphopper, Chipkill, Cloudscape, DB2 OLAP Server, DB2 Universal Database, DFDSM, DFSORT, DS4000, DS6000, DS8000, e-business(logo), e-business on demand, eServer, Express Middleware, Express Portfolio, Express Servers, Express Servers and Storage, GigaProcessor, HACMP, HACMP/6000, IBM TotalStorage Proven, IBMLink, IMS, Intelligent Miner, iSeries, Micro-Partitioning, NUMACenter, On Demand Business logo, OpenPower, POWER, PowerExecutive, Power Architecture, Power Everywhere, Power Family, Power PC, PowerPC Architecture, PowerPC 603, PowerPC 603e, PowerPC 604, PowerPC 750, POWER2, POWER2 Architecture, POWER3, POWER4, POWER4+, POWER5, POWER5+, POWER6, POWER6+, Redbooks, Sequent (logo), SequentLINK, Server Advantage, ServeRAID, Service Director, SmoothStart, SP, System i, System i5, System p, System p5, System Storage, System z, System z9, S/390 Parallel Enterprise Server, Tivoli Enterprise, TME 10, TotalStorage Proven, Ultramedia, VideoCharger, Virtualization Engine, Visualization Data Explorer, X-Architecture, z/Architecture, z/9.

A full list of U.S. trademarks owned by IBM may be found at: http://www.ibm.com/legal/copytrade.shtml.
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, Windows NT and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries or both.
Java and all Java-based trademarks and logos are trademarks of Sun Microsystems, Inc. in the United States, other countries or both.
Oracle is a registered trademark of Oracle Corporation in the USA and/or other countries.
MERCURY and Mercury LoadRunner are registered trademarks or trademarks of Mercury.

Other company, product and service names may be trademarks or service marks of others.

Revised September 28, 2006

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 System p5, eServer p5, pSeries, OpenPower, RS/6000 and BladeCenter 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.

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
NotesBench	http://www.notesbench.org
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
Microsoft Exchange	http://www.microsoft.com/exchange/evaluation/performance/default.asp
Veritest	http://www.veritest.com/clients/reports
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

Revised April 27, 2006

Agenda

- Basic AIX Configuration/Tuning for Oracle
  - Memory
  - I/O
  - Network
  - Miscellaneous

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

Agenda

- Basic AIX Configuration/Tuning for Oracle
  - Memory
  - I/O
  - Network
  - Miscellaneous

VMM Tuning Pre AIX 5.2 ML4

MINPERM% – minimum % real memory for fs buffer cache

15-20%: JFS or JFS2 filesystems without DIO or CIO
5%: RAW logical volumes
JFS or JFS2 with DIO or CIO
GPFS

MAXPERM%, MAXCLIENT% - max % real memory for file system buffer cache

40-60%: JFS or JFS2 filesystems without DIO or CIO
<= 20%: Raw logical volumes
JFS or JFS2 with DIO or CIO
GPFS

- Never more than 20 GB prior to AIX 5.3
- To start, set to vmtune "numperm" value
- Reduce until vmstat freed (fr) to scanned (sr) ratio is 4:1

VMM Tuning – AIX 5.2ML4+

MINPERM% =5%

MAXPERM%, MAXCLIENT%=80% or higher
make this a threshold which is > (1-computational memory)

LRU_FILE_REPAGE=0
LRU_POLL_INTERVAL=10ms

LRU_FILE_REPAGE=0 is a “hint” to lrud to ignore repage rates when determining what to page out – effectively favoring paging out file pages (filesystem buffer cache) rather than computational pages

LRU_POLL_INTERVAL indicates the time period after which LRUD pauses and interrupts can be serviced. Default value of “0” means no preemption.

STRICT_MAXPERM=0 (default)
STRICT_MAXCLIENT=1 (default)

VMM Page Stealing Thresholds

The following define thresholds for the VMM page stealing process (lrud):

- minfree
  - Set minfree = 120 x # logical CPUs / #mem pools
  - Consider increasing if vmstat “fre” column frequently approaches zero or if “vmstat –s” shows significant “free frame waits”

- maxfree
  - Set maxfree = minfree + (MAX(maxpgahead, j2_maxPageReadAhead) * # logical CPUs)/ # mem pools

Example:

- For a 6-way LPAR with SMT enabled, maxpgahead=8, j2_maxPageReadAhead=128, and 2 memory pools:
  - **minfree = 720** = 120 x 6 x 2 / 2
  - **maxfree = 1248** = 480 + (max(128,8) x 6 x 2 / 2)

➢ vmo –o minfree=720 –o maxfree=1248 -p

AIX Paging Space

- DO NOT OVERCOMMIT REAL MEMORY
  - Server should be configured with enough physical memory to satisfy memory requirements
- Allocate Paging Space:
  - With AIX demand paging, paging space does not have to be large
    Provides safety net to prevent system crashes when memory overcommitted
  - ½ memory + 4GB
- Monitor paging activity:
  - vmstat -s
  - sar -r
  - nmon
- Resolve paging issues:
  - Reduce file system cache size (MAXPERM, MAXCLIENT)
  - Reduce Oracle SGA or PGA (9*i* or later) size
  - Add physical memory

Agenda

- Basic AIX Configuration/Tuning for Oracle
  - Memory
  - I/O
  - Network
  - Miscellaneous

Options for storing Oracle data files

- File systems
  - Single-instance:

    JFS, JFS2
  - Clustered:

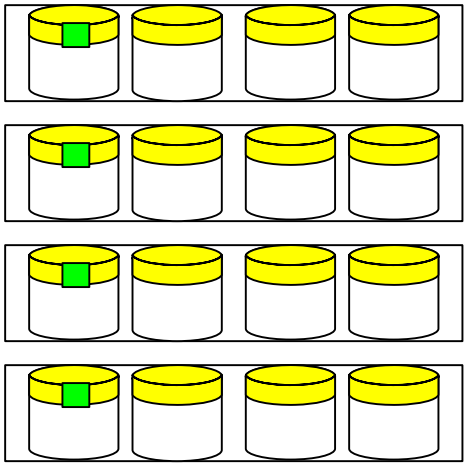
    GPFS
- Raw
- Automatic Storage Management (ASM)
  - new in 10*g*

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

- Use RAID-5 or RAID-10 to create striped LUNs (hdisks)
- Create AIX Volume Group(s) (VG) w/ LUNs from multiple arrays, striping on the front end as well for maximum distribution
  - Physical Partition Spreading (mklv –e x)

    –or-
  - Large Grained LVM striping (>= 1MB stripe size)



http://www-1.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP100319

Single Instance Environments - Filesystems

- JFS – **no longer being enhanced**
  - ***Better for lots of small file creates & deletes***

- JFS2 – **generally the preferred single-instance filesystem**
  - ***Better for large files/filesystems***

Mount options:

- Buffer Caching (default)– stage data in fs cache
- Direct I/O (DIO)– bypasses fs cache
- Concurrent I/O (CIO) – DIO + no write lock (JFS2 only)

  Use CIO for Oracle .dbf, control files and online redo logs only!!!
- Release Behind Read (RBR)– During sequential reads, memory pages released after pages copied to internal buffers
- Release Behind Write (RBW) – During sequential writes, memory pages released after pages written to disk

- In 9*i*, DIO and CIO must be specified at the *filesystem* level
- In 10*g*, Oracle issues o_cio and o_dio calls as appropriate

Single-instance environments
Cached vs. non-Cached (Direct) I/O

File System caching tends to benefit heavily sequential workloads with low write content. To enable caching for JFS/JFS2:

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

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

- In 9*i*, set filesystemio_options=SETALL and use dio or cio mount option
- In 10*g*, set filesystemio_options=SETALL

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

- Increase db_cache_size
- Increase db_file_multiblock_read_count

Asynchronous I/O

- AIX parameters (smit aio) -- applicable to file system based configs
  - minservers = maxservers / 2
  - maxservers = 10 * # disks
  - AIX 5.2 or later, min/maxservers value is per CPU, for 5.1 it is system wide
  - maxreqs = a multiple of 4096 > 5 * #disks * queue_depth
  - “enable” at system restart
  - Typical settings: minservers=100, maxservers=200, maxreqs=16384
- Oracle parameters (init.ora)
  - disk_asynch_io = TRUE
  - filesystemio_options = {ASYNCH | SETALL}
  - db_writer_processes (normally let default)
- Monitor usage:
  - Watch for Oracle alert log or trace file messages:
    - Warning “lio_listo returned EAGAIN”
  - AIX Monitoring
    - “pstat –a | grep aios”
    - “iostat –Aq” (AIX 5.3)
    - Use “-A” option for NMON

Single-instance environments - Oracle Database Files

Data Base Files (DBF)

- I/O size is db_block_size or db_block_size * db_file_multiblock_read_count
- Use CIO or no mount options for extremely sequential I/O
- 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 and set filesystem block size to 512

Archive Log Files

- Do not 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 use a single location – the flash recovery area - for flashback logs, archive logs, and backup logs
- Flashback Log files should use CIO, DIO, or rbrw

Oracle Binaries

- Do not use CIO or DIO

I/O Tuning (ioo)

- READ-AHEAD (Only applicable to JFS/JFS2 with caching enabled)
  MINPGAHEAD (JFS) or j2_minPageReadAhead (JFS2)
  - Default: 2
  - Starting value: MAX(2,DB_BLOCK_SIZE / 4096)

  MAXPGAHEAD (JFS) or j2_maxPageReadAhead (JFS2)
  - Default: 8 (JFS), 128 (JFS2)
  - Set equal to (or multiple of) size of largest Oracle I/O request
    DB_BLOCK_SIZE * DB_FILE_MULTI_BLOCK_READ_COUNT

- Number of buffer structures per filesystem:
  NUMFSBUFS:
  - Default: 186, Starting Value: 1568

  j2_nBufferPerPagerDevice
  - Default: 512, Starting Value: 2048
  - Monitor with “vmstat –v”

ASM configuration

AIX parameters

- Async I/O needs to be enabled (smitty aio, State of Fast Path), but default values may be used

ASM instance parameters

- asm_power_limit = 1
  - Makes ASM rebalancing a low-priority operation
  - May be changed dynamically.
- PROCESSES = 25 + 15n, where n=# of instances using ASM

DB instance parameters

- disk_asynch_io=TRUE
- filesystemio_options=ASYNCH
- Increase Processes by 16
- Increase Large_Pool by 600k
- Increase Shared_Pool by [(1M per 100GB of usable space) + 2M]

Agenda

- Basic AIX Configuration/Tuning for Oracle
  - Memory
  - I/O
  - Network
  - Miscellaneous

Network Options (no) Parameters

- Set sb_max >= 1 MB (1048576) (generally ok by default)
- Set tcp_sendspace >= 262144
- Set tcp_recvspace >= 262144
- Set rfc1323=1
- Confirm these attributes are set at network interface level also

Additional Network (no) Parameters for RAC:

- Set udp_sendspace = db_block_size * db_file_multiblock_read_count
  (not less than 65536)

- Set udp_recvspace = 4 * udp_sendspace
  - Must be < sb_max
  - Increase if buffer overflows occur

- Use Jumbo Frames

Examples:

- no -a |grep udp_sendspace
- no –o -p udp_sendspace=65536
- netstat -s |grep "socket buffer overflows"

Agenda

- Basic AIX Configuration/Tuning for Oracle
  - Memory
  - I/O
  - Network
  - Miscellaneous

Miscellaneous parameters

- User Limits (smit chuser)
  - Soft FILE size = -1 (Unlimited)
  - Soft CPU time = -1 (Unlimited)
  - Soft DATA segment = -1 (Unlimited)
  - Soft STACK size -1 (Unlimited)
  - /etc/security/limits
- Maximum number of PROCESSES allowed per user (smit chgsys)
  - maxuproc >= 4096
- Memory Related Environment variables:
  - AIXTHREAD_SCOPE=S
  - NUM_SPAREVP=1 (AIX 5.1 only)

Agenda

- Basic AIX Configuration/Tuning for Oracle
  - Memory
  - I/O
  - Network
  - Miscellaneous

Information Sources

- See Oracle Metalink Note # 282036.1 for required and recommended patches
- Oracle Product Certification information:
  http://otn.oracle.com/support/metalink/index.html
- Oracle Technology Network
  http://otn.oracle.com
- IBM Redbooks:
  http://www.ibm.com/redbooks
- IBM Techdocs – Technical Sales Library
  http://www.ibm.com/support/techdocs
- Tuning IBM AIX 5L for an Oracle Database
  http://www-03.ibm.com/servers/enable/site/peducation/wp/9a46/9a46.pdf
- Oracle Database 10g Release 2 Automatic Storage Management Overview and Technical Best Practices
  http://www.oracle.com/technology/products/database/asm/pdf/asm_10gr2_bptwp_sept05.pdf