AIX5 Initial Settings for Databases Servers

Introduction

Here are the AIX 5L settings I automatically change when installing a pSeries database server. They are provided here as a reference point for tuning an AIX system. As always, all settings should be reviewed once the system is running production workloads.

Disclaimer: The settings are provided "as is", without guarantees. Every system is different, so your tuning requirements may be different.

AIX 5L

In a previous "AIX Tip", I listed the AIX 4.3 settings I used for an Oracle benchmark. This tip extends those settings to AIX 5L. However, the settings in this case are based on general field experience, rather than a specific benchmark.

Specifying settings for AIX 5L is more complex than for AIX 4.3. The reason is that AIX 5 has more configuration alternatives, each with different tuning requirements. Two important configuration alternatives are the 32/64 bit kernel and the JFS/JFS2 file systems. Another consideration is the differences between AIX 5.1 and 5.2. AIX 5.2 defines some tuning settings very differently (asynch I/O), uses different tuning commands, and has new tuning capabilities (concurrent and direct I/O). In addition, there are subtle tuning considerations when using AIX 5.2 Dynamic LPAR. However, DLPAR is outside the scope of this document. Check with IBM Supportline for more information.

Whenever possible, use the tuning recommendations from your database vendor. In absence of this information, this document provides a reasonable starting point. For the initial setup, I automatically change several memory, I/O, JFS and network settings. However, I use default CPU and JFS2 settings, as these settings can cause problems if set incorrectly. The settings listed here are not comprehensive and should be reviewed after the system is under load.

Although system tuning is a customer responsibility, IBM can help. IBM Supportline (800-CALLAIX) can help with tuning questions, and IBM Global Services is available for a more in-depth study of your system.

Tuning Methodology

The tuning objective is to meet the Service Level Agreement (SLA). The SLA is an agreement between IT and their users that clearly states the level of service users can expect. The SLA is most effective when the objectives are quantified, such as batch or query response times. Quantified objectives help avoid endless tuning loops.

AIX5 Initial Settings for Databases Servers

My strategy is to get the AIX tuning in the "ballpark", and then focus on database tuning. Database tuning often provides an order of magnitude more benefit than AIX tuning. However, AIX must be tuned first to provide a solid base. The AIX settings only need to be in the "ballpark" to be effective.

System tuning is an iterative process. As soon as you eliminate one bottleneck, another takes its place. However, it usually isn't beneficial to try to eliminate all bottlenecks. I consider AIX tuning done when:

1) The SLA is being met at peak loads, or
2) There are no system bottlenecks
    a. CPU has idle time (sar –P ALL)
    b. Memory is not paging (vmstat => pi/po column)
    c. I/O response times average below 10-15 ms (filemon)
    d. Network response times are within acceptable limits
3) The system is healthy
    a. No hardware errors (errpt)
    b. No AIX issues (http://www-1.ibm.com/servers/eserver/support/pseries/fixes/ or call IBM SupportLine.)

To summarize, the decision matrix for tuning is as follows:

	SLA Is Not Being Met	**SLA Is Being Met**
System Bottlenecks Exist	Tune the system	Done
No System Bottlenecks	Tune the database	Done

If bottlenecks persist after tuning AIX, the system may be undersized. Consider adding hardware (CPU, memory, storage). In some cases, poorly written queries or applications can cause the bottlenecks. The only alternative is to fix the application.

Bruce Spencer
baspence@us.ibm.com

AIX5 Initial Settings for Databases Servers

CPU/System Settings

Setting	Command	Default	Guideline	Comments
Max number of user processes	smit chgsys	128	5000+	
Maximum file size	/etc/security/limits	2 GB	fsize=-1	For the database ID
Maintain I/O history	smit chgsys	No	Yes	A “nice to have” for monitoring disk activity

The CPU requires almost no initial tuning.

AIX5 Initial Settings for Databases Servers

Memory Settings

Setting	AIX 5.1 Command	AIX 5.2 Command	Guideline	Comments
All				
Paging space	smit pgsp	smit pgsp	1-5 GB	
maxclient	vmtune -t	vmo: maxclient%	=maxperm	maxclient must be ≤ maxperm
minfree	vmtune -f	vmo: minfree		Use default
maxfree	vmtune –F	vmo: maxfree	maxfree = minfree + maxpgahead	See I/O settings for maxpgahead. Use the larger of the two: JFS **maxpgahead** Or JFS2 **j2_max_Read_Ahead**
JFS				
minperm	vmtune –p	vmo: minperm%	15%	
maxperm	vmtune -P	vmo: maxperm%	30%	Be sure maxclient is ≤ maxperm (see above)
JFS2				
minperm	vmtune -p	vmo: minperm%	Use default	
maxperm	vmtune -P	vmo: maxperm%	Use default	

Page space: Page space can be relatively small, assuming memory is sized correctly and there are no memory leaks. AIX 5 uses "demand paging", which means page space is used only as a "spill over" in the event AIX runs out of memory. However, if memory is undersized or if there is a memory leak, you will need to increase the page space to accommodate the "spillover".

Maxperm: maxperm influences paging behavior. Paging frees memory by copying the "least recently used" memory to disk for temporary storage. Tune maxperm to keep as

much of the database in memory as possible when the system pages. The default setting is usually not optimum if the system pages as it favors paging out the database first.

The initial maxperm setting depends on whether the database filesystem resides on JFS or JFS2. For JFS, I automatically change the default "maxperm/maxclient" as above. For JFS2, I recommend using the defaults, and defer tuning to when the system is running a representative workload. The reason is that AIX classifies JFS and JFS2 memory differently. JFS is controlled by "maxperm", which is a "soft" limit and more forgiving if set too low. JFS2, on the other hand, is controlled by "maxclient", which is a "hard" limit. If set incorrectly, performance can be degraded. See "maxclient" section for JFS2.

When the system is under load, you can optimize the "maxperm" setting as follows. But first a little background. AIX classifies memory either as persistent or working storage. Persistent storage includes file cache, executables, and metadata. Working storage includes the database. Target amount of memory for persistent storage when the system is paging is the "maxperm" setting. (Note if the system is not paging, the maxperm setting is mostly ignored.) The default "maxperm" value of 80% favors paging out the database first, which degrades performance. In addition, the actual persistent memory requirements are typically much lower (20-50%).

The tuning methodology involves finding the actual persistent memory used when the system is under load (numperm). Then set "maxperm" to 5% below that value.

maxperm = numperm% - 5%.

(Note: be sure to set maxclient=maxperm. See "maxclient" below for more info.)

There are multiple commands for displaying memory usage. Here are two that I use.

AIX 5.1: "/**usr/samples/kernel/vmtune | grep numperm"**
AIX 5.2: **"vmstat –v | grep numperm"**

If you don't have root access to run vmtune, you can use a different approach. Use vmstat "avm" column to estimate the size of the working storage. Calculate numperm% as follows

numperm% = 100% - "working storage%" = 100% - (avm *100 / total memory)

Note "avm" is in measured in 4k pages, so you'll have to convert this number to the same units as total memory.

maxclient: A subgrouping of persistent memory. As such, "maxclient" is always less than or equal to "maxperm" (I set "maxclient=maxperm").

The “maxclient” value sets the maximum amount of memory used by its members: JFS2, NFS, CDROM and Veritas file systems. The “maxclient” value is a “hard” limit, which is always enforced.

The importance is that JFS2 has a “hard” limit, while JFS does not. It’s better to leave the JFS2 as default until the system can be tuned while running a representative load. Setting “maxclient” too low can degrade JFS2 performance.

Tune the”maxclient” with the system under load. Use the “maxperm” tuning procedure for setting “maxclient”, and set “maxclient” equal to “maxperm”. If you want to be more precise, you can use the “svmon –G” command to view the amount of memory used by client storage. It’s shown under the “clnt” column (“in use”).

```
# svmon -G
               size       inuse        free         pin     virtual
memory       131072      129422        1650       11704       50091
pg space     131072        4279

               work        pers        clnt       lpage
pin           11704           0           0           0
in use        47052       76146        6224           0
```

The svmon command also shows the working storage under the “work” column. It all comes together as follows:

Total memory = free+”working”+”persistent”
= free+work+(pers+clnt)
= 1650+47052+(76146+6224) = 1331072

minfree: Paging starts when free memory reaches “minfree”.

maxfree: Specifies when paging should stop.

maxfree = minfree + “MaxPageAhead”

The “MaxPageAhead” value is *maxpgahead* for JFS, and *j2_maxPageReadAhead* for JFS2.

AIX5 Initial Settings for Databases Servers

I/O Settings

Setting	AIX 5.1 Command	AIX 5.2 Command	Guideline	Comments
All				
Disk layout	smit lvm	smit lvm	Spread ALL data over ALL physical disks	
Queue depth for fibre channel adapter	smit fcsa	smit fcsa		See comments
Asynch I/O (AIX 5.1)	smit chgaio		Min: 1 Max: 1000 Req: 8192 Enabled	In AIX 5.1, AIO min/max is a system total. Requires reboot
Asynch I/O (AIX 5.2)		smit chgaio	Min: 1 Max: 1000/#CPUs Req: 8192 Enabled	AIX 5.2 AIO min/max is per CPU Requires reboot
Direct & Concurrent I/O	N/A			See comments
JFS				
Filesystem buffers	vmtune -b	ioo: numfsbufs	500	Must be run before mounting filesystems
minpgahead	vmtune –r	ioo: minpgahead	2	
maxpgahead	vmtune –R	ioo: maxpgahead	16	
JFS2				
Filesystem buffers	vmtune -Z	ioo: j2_nBufferPerPagerDevice	Default=512	Must be run before mounting filesystems
minpgahead	vmtune –q	ioo: j2_minPageReadAhead	2	
maxpgahead	vmtune –Q	ioo: j2_maxPageReadAhead	16	

AIX5 Initial Settings for Databases Servers

Disk Layout: The most important I/O tuning step is to spread all of the data over all of the physical disk drives*. If you have a SAN, work closely with the SAN administrator to understand the logical to physical disk layout. In a SAN, two or more hdisks may reside on the same physical disk. (*-The one exception is when you back up to disk. Be sure the backup disks are on a separate storage system to avoid having a single point of failure.)

Queue depth for fibre channel adapter: This setting depends on the storage vendor. For IBM Shark storage, I set this around 100. If using non-IBM storage, check with your vendor for their queue depth recommendation. High queue depth settings have been known to cause data corruption on some non-IBM storage. If unsure, use the default value.

Asynch I/O: Improves write performance for JFS and JFS2 file systems. It does not apply to raw partitions. AIO is implemented differently in AIX 5.1 and 5.2. In 5.1, the min/max AIO settings are for the entire system. In 5.2, the AIO settings are per CPU.

In AIX 5.1, I set the "max server" to 1000. On a 5.2 system, divide 1000 by the number of CPUs. I tend to over configure AIO, as it requires a reboot. Over configuring "max server" doesn't use any extra resources, as AIO servers are only created when needed. The "max server" just sets the maximum, not the actual number used. If you plan to use DLPAR and dynamically add CPU's, contact Supportline to discuss the implications.

Direct and Concurrent I/O: I've included these as placeholders, as they look promising. However, I don't have the experience to recommend them yet. For more information see,

Direct I/O: http://www-106.ibm.com/developerworks/eserver/articles/DirectIO.html
Concurrent I/O: http://www-1.ibm.com/servers/aix/whitepapers/db_perf_aix.pdf

Filesystem buffers: Insufficient buffers will degrade I/O performance. The default AIX setting for these buffers is typically too low for database servers. JFS/JFS2 are tuned separately. JFS uses "vmtune –b", while JFS2 uses "vmtune –Z".

Be careful not to set the value too high if running a 32 bit kernel with a large number of filesystems (50+). The buffer setting is per filesystem, and you can run out of kernel memory if this is set too high. (This does not apply to the 64 bit kernel, which supports larger kernel memory sizes.)

Tune the filesystem buffers when the system is under peak load. Run the following command multiple times:

AIX 5.1: /usr/samples/kernel/vmtune –a | grep fsbufwaitcnt
AIX 5.2: vmstat –v | grep "filesystem I/Os blocked with no fsbuf"

If the value increases with time, this means there are insufficient buffers. (The absolute value isn't important….just the rate of change.). If so, increase the number of filesystem buffers. The filesystem has to be unmounted/mounted for the new setting to take effect.

Network Settings

Setting	AIX 5.1 Command	AIX 5.2 Command	Guideline	Comments
tcp_sendspace	no	smit TunNet	262144	
tcp_recvspace	no	"	262144	
rfc1323	no	"	1	

Implementation

The tuning concepts in AIX 5.1 and 5.2 are the same, but the implementation is different.

In AIX 5.1, I put the vmtune command in /etc/inittab. I locate it just above the "/etc/rc" entry because some of the tuning commands must be run before /etc/rc mounts the fielsystems. You can combine all the vmtune options with the same command. Here's an example of a /etc/inittab entry for vmtune.

```
……..
vmtune:2:wait:/usr/samples/kernel/vmtune –R16 –b465 –p10% -P30%
rc:23456789:wait:/etc/rc…..
……..
```

In AIX 5.2, the vmtune has been split into two commands: vmo and ioo. It is best to configure via "smit tuning".

References

"Database Performance Tuning on AIX",
http://www.ibm.com/redbooks. Search for document SG24-5511.

"AIX Performance Tools Handbook",
http://www.ibm.com/redbooks Search for document SG24-6039

"Understanding IBM eServer pSeries Performance and Sizing"
http://www.ibm.com/redbooks Search for SG24-4810

"Performance Management Guide"
http://publib16.boulder.ibm.com/pseries/en_US/infocenter/base/aix52.htm