VMTUNE Parameters

Special Notices

Information in this document is correct to the best of our knowledge at the time of this writing. Please send feedback by fax to "AIXServ Information" at (512) 823-4009.

Please use this information with care. IBM will not be responsible for damages of any kind resulting from its use. The use of this information is the sole responsibility of the customer and depends on the customer's ability to evaluate and integrate this information into the customer's operational environment.

About This Document

This Document discusses the vmtune command used to modify the VMM parameters that control the behavior of the memory management subsystem. This document is applicable to AIX 4.1.

About vmtune

The command and source for vmtune is found in the /usr/lpp/bos/samples (3.2) or /usr/samples/kernel (4.1) directory. It is installed with the bosadt.lib.obj (3.2) or bos.adt.samples (4.1) fileset.

There may be times when the default VMM parameters will not prove satisfactory. For instance, consider the situation where the aggregate size of the working pages is very large and file system pages are only referenced infrequently. In this case, the pages from the file system are read once and never used again. With the default parameters, the system would try to maintain up to ((RAMSIZE-4MB)*.8) (approx.80%) of real memory with persistent pages leaving only ((RAMSIZE-4MB)*.2) (approx. 20%) for working pages. This imbalance would cause more working pages than necessary to be paged to/from paging space.

In this example you can chose to reduce the maxperm value to 65% in an attempt to maintain a larger set of working pages in memory. This could lead to less paging to page space with no detrimental effect on the persistent storage. That will decrease, though, as the system brings in working pages and replaces some file pages.

If you find that your system frequently requires quick access to many frames from the free list, you may want to increase the minfree and maxfree values. A larger minfree ensures that more frames are readily available when they are needed. The delta between minfree and maxfree determines how many pages are to be stolen for each call to the page replacement code (PRC). A larger

delta usually means longer intervals between calls to the PRC, but each call takes longer to complete.

Significant differences in read times can be achieved by using various minpgahead and maxpgahead values. It is very dependent upon disk activity and access to the individual files. The VMM read ahead algorithms, which these values affect, is designed to improve sequential access read performance.

It takes experience and know-how to set vmtune parameters properly. Therefore, it is highly recommended that vmtune values are only set after consulting the AIX Performance Team. To reach them, call 1-800-CALL-AIX and ask for a performance analysis through consult line. More information can also be obtain in the Performance Tuning Guide (SC23-2365-03).

Here are all of the parameters available for AIX 4.1.3 or higher. Earlier versions of AIX contain some but not all of the following options:

- minperm: -p

  Specifies the point below which file pages are protected from the repage algorithm. This value is a percentage of the total real-memory page frames in the system. The specified value must be greater than or equal to 1. The default value of the minperm percentage is always around 17-19% of memory.

- maxperm: -P

  Specifies the point above which the page stealing algorithm steals only file pages. This value is expressed as a percentage of the total real-memory page frames in the system. The specified value must be greater than or equal to 1. The default value of the maxperm percentage is always around 75-80% of memory. A pure NFS server may obtain better performance by increasing the maxperm value. A system that accesses large files (over 50-75% of the amount of RAM on the system - look at "numperm" to see how much memory is currently used for file mapping) may benefit by increasing the maxperm value. Maxperm can be reduced on systems with large active working storage requirements (AVM column from vmstat compared to total real page frames) to reduce or eliminate page space I/O.

- minpgahead: -r

  Specifies the number of pages with which sequential read-ahead starts. This value can range from 0 through 4096. It should be a power of 2. The default value is 2.

- maxpgahead: -R

  Specifies the maximum number of pages to be read ahead. This value can range from 0 through 4096. It should be a power of 2 and should be greater than or equal to minpgahead. The default value is 8. Increasing this number will help large sequential read performance. Because of other limitations right now in the kernel and the LVM, the maximum value should

not be greater than 128. The delta between minfree and maxfree should always be equal to or greater than maxpgahead.

- minfree: -f

  Specifies the minimum number of frames on the free list. This number can range from 8 to 204800. The default value depends on the amount of RAM on the system. minfree is by default the value of maxfree - 8. The value of maxfree is equal to minimum(#of memory pages/128, 128). The delta between minfree and maxfree should always be equal to or greater than maxpgahead.

- maxfree: -F

  Specifies the number of frames on the free list at which page stealing is to stop. This number can range from 16 to 204800 but must be greater than the number specified by the minfree parameter by at least the value of maxpgahead. The default value of maxfree is determined by this formula: maxfree=minimum(#of memory pages/128, 128). The delta between minfree and maxfree should always be equal to or greater than maxpgahead.

- maxpin: -M

  Specifies the maximum percentage of real memory that can be pinned. The default value is 80%. If this value is changed, the new value should ensure that at least 4M of real memory will be left unpinned for use by the kernel. Maxpin values must be greater than 1 and less than 100. The value under "maxpin" is converted to a percentage at the end of the output of vmtune.

- npswarn: -w

  Specifies the number of free paging-space pages at which AIX begins sending the SIGDANGER signal to processes. The default value in AIX 3.2 is 512 pages. The formula to determine the default value in AIX version 4 is:

  ```
  MAX(512, 4*npskill)
  ```

  See npskill description below. The value of npswarn has to be greater than 0 and less than the total number of paging space pages on the system.

- npskill: -k

  Specifies the number of free paging-space pages at which AIX begins killing processes. The default value in AIX version 3 is 128 pages. The formula to determine default value of npskill in version 4 is:

  ```
  MAX(64, number_of_paging_space_pages/128)
  ```

  The npskill value has to be greater than 0 and less than the total number of paging space pages on the system.

- numclust: -c

  Specifies the number of 16K clusters processed by the vmm's sequential write-behind algorithm. The default value is 1. Values can be any integer above 0. Higher number of clusters may result in larger sequential write

performance on devices that support very fast writes (RAID, ...). Setting the value to a very high number like 500000 will essentially defeat the write-behind algorithm. This can be beneficial in the cases such as database index creations where pages that were written to are read a short while later - write-behind could actually cause this process to take longer. One suggestion is to turn off write-behind before building indexes and then turn it back on after indexes have been built. For example, the "mkpasswd" command can run significantly faster when write-behind is disabled.

- numfsbufs: -b

  Specifies the number of file system bufstructs. The default value is 64 in 3.2; the current default in AIX version 4 is 93 since it is dependent on the size of the 'buf' struct. This value must be greater than 0. Increasing this value will help write performance for very large writes sizes (on devices that support very fast writes). In order to enable this value, a file system has to be unmounted and mounted again after changing the value.

- hd_pbuf_cnt: -B

  Controls the number of pbufs available to the LVM device driver. Pbufs are pinned memory buffers used to hold I/O requests related to a Journaled File System. In AIX Version 3, one pbuf is required for each page being read or written. On systems where large amounts of sequential I/O occurs, this can result in I/O's bottlenecking at the LVM layer waiting for pbufs to be freed. In AIX Version 4, a single pbuf is used for each sequential I/O request regardless of the number of pages in that I/O. Therefore, it is harder to encounter this type of bottleneck. The maximum value is 128.

- lvm_bufcnt: -u

  Specifies the number of LVM buffers for raw physical I/Os. The default value is 9. It may be beneficial to increase this value if doing large raw I/Os (ie, not going through the JFS). The possible values can range between 1 and 64. This option is only available in AIX version 4.

- pd_npages: -N

  Specifies the number of pages that should be deleted in one chunk from RAM when a file is deleted. Changing this value may only be beneficial to real-time applications that delete files. By reducing the value of pd_npages, a real-time application can get better response time since few number of pages will be deleted before a process/thread is dispatched. The default value is the largest possible file size divided by the page size (currently 4096); if the largest possible file size is 2GB, then pd_npages is by default 524288.

- maxrandwrt: -W

  Specifies a threshold (in 4KB pages) for random writes to accumulate in RAM before these pages are sync'd to disk via a write-behind algorithm. This threshold is on a per file basis. This option is only available in AIX version 4.1.3 and later. The default value of maxrandwrt is 0 which disables random write-behind. By enabling random write-behind (a typical value

might be 128), applications that make heavy use of random writes can get better performance due to less of a dependence on the sync daemon to force writes out to disk. Note that some applications may degrade their performance due to write-behind (such as database index creations). In these cases, it may be beneficial to disable write-behind before creating database indexes and then re-enabling write-behind after the indexes are created.

The current values are displayed when vmtune is executed without any parameters. Changes to the values go into affect immediately but do not exist after rebooting.

09/27/96, 4FAX# 1777

Reader's Comments

Please fax this form to (512) 823-4009, attention "AIXServ Information". You may also e-mail comments to: elizabet@austin.ibm.com. These comments should include the same customer information requested below.

Use this form to tell us what you think about this document. If you have found errors in it, or if you want to express your opinion about it (such as organization, subject matter, appearance) or make suggestions for improvement, this is the form to use.

If you need technical assistance, contact your local branch office, point of sale, or 1-800-CALL-AIX (for information about support offerings). These services may be billable. Faxes on a variety of subjects may be ordered free of charge from 1-800-IBM-4FAX. Outside the U.S. call 415-855-4329 using a fax machine phone.

When you send comments to IBM, you grant IBM a nonexclusive right to use or distribute your comments in any way it believes appropriate without incurring any obligation to you.

Note: If you have a problem report or item number, supplying that number may help us determine why a procedure did or did not work in your specific situation.

Problem Report or Item #: Branch Office or Customer #:

Be sure to print your name and fax number below if you would like a reply:

Name: Fax Number:

END OF DOCUMENT (vmtune.parameters.lxp, 4fax# 1777)