The schedtune Command

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 schedtune command used to modify CPU scheduler and Virtual Memory Manager processing parameters. This document is applicable to AIX 4.1 and AIX 3.2.

About schedtune

The command and source for schedtune 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.

Warning: Schedtune is in the samples directory because it is VMM-implementation dependent. The schedtune code that accompanies each release of AIX was tailored specifically to the VMM in that release. Schedtune is not supported under SMIT, nor has it been tested with all possible combinations of parameters. Misuse of this command can cause performance degradation or operating-system failure. If execution of schedtune causes a system to crash with flashing 888's, then the appropriate level of schedtune for that release of AIX needs to be installed.

It takes experience and know-how to set schedtune parameters properly. Therefore, it is highly recommended that schedtune values be set only after consulting the AIX Performance team. To reach them, call 1-800-CALL-AIX or your support center and ask for a performance analysis through consult line.

Priority-Calculation

The priority of most user processes varies with the amount of CPU time the process has used recently. The formula for calculating a process' or thread's priority value is:

```
priority value = base value + nice value +
(CPU penalty based on recent CPU usage)
```

The scheduler's CPU penalty calculations are based on two parameters that can be set with schedtune: -r and -d.

The formula used by the scheduler to calculate the amount to be added to a process' or thread's priority value as a penalty for recent CPU use is:

```
CPU penalty = (recent CPU use value) * (-r value/32)
```

Note: The default value for "r" is 16.

The once per second recalculation of the recently used CPU value of each process/thread is:

```
(old recent CPU use value) * (-d value/32)
```

Note: The default value for "d" is 16.

The formula for calculating a process' or thread's priority value is:

```
priority value = base priority + nice value +
(CPU penalty based on recent CPU usage)
```

The recent CPU usage value is displayed as the "C" column in the ps command output. The maximum value of recent CPU usage is 120. You can view the current nice value (NI), priority (PRI) and CPU usage value (C) for user processes by running

```
ps -ef | more
```

Memory-Load Control

The AIX scheduler performs memory load control by suspending processes when memory is overcommitted. The system does not swap out processes; instead pages are "stolen" as they are needed to fulfill the current memory requirements. Pages are stolen based on a least recently used algorithm, hence pages from suspended processes will most likely be stolen first.

Memory-load control is intended to smooth out infrequent peaks in load that might otherwise cause a system to thrash. It is NOT intended to act continuously in a configuration that has too little RAM to handle its NORMAL workload.

Memory is considered overcommitted when the number of pages written to paging space in the last second times the value of the -h parameter is greater than the number of page steals in the last second. A process is suspended when memory is overcommitted and the number of repages that the process has accumulated in the last second times the value of the -p parameter is greater than the number of page faults that the process has accumulated in the last second. The term "repages" refers to the number of pages belonging to

the process, which were reclaimed and are soon after referenced again by the process.

No "interactive" vs. "non-interactive" process notion exists, hence suspended processes may well be those that interact directly with end-users.

You should not change the memory-load control parameter settings unless your workload is consistent and you believe the default parameters are not appropriate for your workload.

Here are all of the parameters available for AIX 3.2 and 4.1: **Note**: The "-d" and "-r" flags require AIX version 3.2.5 and APARs IX39738 and IX42347, or AIX version 4.1.

- SYS: -h

  PURPOSE: Sets the system-wide criteria used to determine when process suspension begins and ends (System is thrashing).

  DEFAULT: 6 for AIX version 3. In version 4, the default value is zero if a system has 128MB or more of RAM; otherwise, it is 6. For example:

  ```
  schedtune -h 0
  ```

  This will turn off memory load control.

- PROC: -p

  PURPOSE: Sets the per-process criterion used to determine which processes to suspend.

  DEFAULT: 4 For example:

  ```
  schedtune -p 2
  ```

  This requires a higher level of repaging by a given process before it is a candidate for suspension by memory load control.

- MULTI: -m

  PURPOSE: Sets the minimum number of processes that are exempt from suspension. This number is in addition to kernel processes, processes with fixed priority less than 60, processes with pinned memory, or processes awaiting events.

  DEFAULT: 2

  While m=2 is appropriate for a desktop, single-user configuration, it is frequently too small for larger, multiuser or server configurations with large amounts of RAM. On those systems, increasing the value of m may result in better performance.

  ```
  schedtune -m 10
  ```

  Requires that the memory load control always have at least 10 user processes running when it is suspending pocesses.

- WAIT: -w

  PURPOSE: Sets the number of seconds to wait after thrashing ends before making suspended processes runable.

  DEFAULT: 1

  ```
  schedtune -w 2
  ```

  Would allow processes to be added back after 2 seconds based, first, on their priority and, second, on the length of their suspension period.

- GRACE: -e

  PURPOSE: Sets the number of seconds that a recently resumed process which was previously suspended is exempt from suspension.

  DEFAULT: 2

- TICKS: -f

  PURPOSE: The number of clock ticks to wait before retrying a failed fork call. The system will retry a failed fork call five times. For example, if a fork() subroutine call fails because there is not enough paging space available to create a new process, the system retries the call after waiting the specified number of clock ticks.

  DEFAULT: 10

- SCHED_D: -d

  PURPOSE: Sets the short term CPU usage decay rate. The valid values are 0 to 32. The default is to decay short term CPU usage by 1/2 (16/32) every second. Decreasing this value enables foreground processes to avoid competition with background processes for a longer time.

  DEFAULT: 16

  ```
  schedtune -r 6 -d 16
  ```

  If a background process were started with nice -20, it would be at least one second before the background process began to receive any CPU time. Foreground processes, however, would still be distinguishable on the basis of CPU usage. Long-running foreground processes that should probably be in the background would ultimately accumulate enough CPU usage to keep them from interfering with the true foreground. For Example:

  ```
  schedtune -r 32 -d 32
  ```

  Long-running processes would reach a C value of 120 and stay there, contending on the basis of their nice values. New processes would have priority, regardless of their nice value, until they had accumulated enough time slices to bring them within the priority value range of the existing processes.

- SCHED_R: -r

  PURPOSE: Set the weighting factor for short term CPU usage in priority calculations.

DEFAULT: 16 The valid values are 0 to 32. The default is to include 1/2 (16/32) of the short term CPU usage in the priority calculation. Decreasing this value makes it easier for foreground processes to compete. For example:

```
schedtune -r 0
```

This would mean that the CPU penalty was always 0, making priority absolute. No background process would get any CPU time unless there were no dispatchable foreground processes at all. The priority values of the processes would effectively be constant, although they would not technically be fixed-priority processes. For example:

```
schedtune -r 2
```

would ensure that any new foreground process would receive at least .5 seconds of CPU time before it had to compete with any process with a nice value equal to or greater than 24.

- TIME_SLICE: -t

  PURPOSE: The number of clock interrupts that must occur before the dispatcher is called. The CPU time slice is the period between recalculations of the priority value. Normally, recalculation is done at each tick (10 milliseconds) of the system clock. The number of clock ticks between recalculations (length of the time slice) can be increased by 10 millisecond (one clock tick) increments.

  However, the time slice is not a guaranteed amount of processor time. It is the longest time that a process/thread can be in control before it is replaced by another process or thread. A process or thread can lose control of the CPU before its full time slice when the following occurs:

  – A process/thread with a higher priority returns from a system call.

  – A process/thread with a higher priority completes an I/O request.

  – The current process/thread issues an I/O request.

  – The current process/thread is suspended.

  DEFAULT: The default values are 0 in AIX Version 3.2 and 1 in Version 4.1 but they are equivalent (call the dispatcher after one time slice) The beginning value for the parameter has been incrementalized in AIX Version 4.1. It has been found that different values of "t" have NOT caused significant changes in the execution times of processes or threads.

  In AIX Version 4.1, this parameter only applies to threads with the SCHED_RR scheduling policy.

  If the workload consists almost entirely of very long-running, CPU-intensive programs, increasing this parameter may have some positive effect.

- DEFAULT: -D

  PURPOSE: Restore ALL the default values.

The current values are displayed when schedtune is executed without any parameters. Changes to the values go into affect immediately but do not exist after rebooting. If a permanent parameter change is needed, an appropriate entry should be put in one of the scripts called by /etc/inittab or an inittab entry created for schedtune itself. For example, an appropriate /etc/inittab line for raising the minimum level of multiprogramming to 5 using AIX version 4.1 would be:

```
schedtune:2:wait:/usr/samples/kernel/schedtune -m 5
```

References

AIX Versions 3.2 and 4.1 Performance Tuning Guide (SC23-2365) 1995/AIXtra POWER Technical Conference, Vol. II InfoExplorer

03/07/96, 4FAX# 3239

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 (schedtune.lxp, 4FAX# 3239)