sun.com My Sun | Regional Sites | Site Index | How To Buy

SELECT A TOPIC Go Search

Home › Sun on the Net

-Sun on the Net

Priority Paging

Richard Mc Dougall, Triet Vo, Tom Pothier,
August 1998

What is Priority Paging?

Priority paging is a new paging algorithm which can significantly enhance system response when the file system is being used.

The current Solaris behavior is that response of applications can suffer significantly when the file system is used heavily. On a workstation, this can be noticed as poor interactive response and trashing of the swap disk. On servers, the effect is degraded application response time, and low CPU utilization due to heavy paging. This happens because Solaris allows the file system cache to grow dynamically to the point where it steals memory pages from important applications.

The priority paging algorithm allows the system to place a boundary around the file cache, so that file system I/O does not cause paging of applications.

What systems does priority paging effect?

Priority paging benefits a wide range of systems and recent testing has showed significant benefits on HPC codes, MCAD, Desktop systems, and OLTP workloads. Performance gains between 10-300% have been identified on customer workloads.

HPC codes which write large amounts of I/O to the file system almost always benefit from priority paging. HPC benchmarks have shown as much as 300% better performance.

Desktop systems with 64MB upwards of memory 'feel' significantly more responsive, and most of the swap device activity is avoided.

OLTP applications that use databases on file systems that have users with sleep times between transactions will benefit because their applications are no longer being paged out between transactions.

How do I enable priority paging?

Be default, priority paging is disabled. It is likely to be the basis of the default algorithm in future OS releases, once cusomer feedback has been collected for systems with priority paging enabled.

To use priority paging, you will need either Solaris 2.7 or Solaris 2.6 with kernel patch 105181-09. Work is underway to include priiority paging in a future revision of the Solaris 2.5.1 kernel patch.

To enable priority paging, set the following in /etc/system:

```
 set priority_paging=1
```

Setting priority_paging=1 sets the a new memory tunable, cachefree, high 2 x lotsfree. The cachefree memory tunable is a new parameter which scales with minfree, desfree and lotsfree. The system now attempts to keep cachefree pages of memory on the freelist, but will only free file system pages while free memory is between cachefree and lotsfree.

IMPORTANT NOTE: Ensure that data files do not have the executable bit set. This can fool the VM into thinking that these are really executables, and will not engage priority paging on these files.

To enable priority paging on a live 32 bit system, set the following with adb:

```
# adb -kw /dev/ksyms /dev/mem

lotsfree/D

lotsfree: 730 <- value of lotsfree

cachefree/W 0t1460 <- 2 x value of lotsfree
```

To enable priority paging on a live 64 bit system, set the following with adb:

```
# adb -kw /dev/ksyms /dev/mem

lotsfree/E

lotsfree: 730 <- value of lotsfree

cachefree/Z 0t1460 <- 2 x value of lotsfree
```

How priority paging works

Currently, the VM system uses all free memory as a file cache for random I/O, which means that doing large random I/O on a file system will floodmemory with file pages. The system attempts to cope with this by starting up the scanner when free memory falls to the low water mark, lotsfree (set by default to 1/64th of memory). The scanner frees or swaps out any pages that haven't been reference recently (recently can be as little as one cycle of the scanner, which is measured in seconds).

The net result of the current VM is that heavy file system I/O will pageout significant portions of important applications heap and stack address space, even though there may be ample memory in the system.

The new priority paging algorithm introduces a new additional water mark, cachefree. The paging parameters are now:

minfree < desfree < lotsfree < cachefree

The scanner by default starts when free memory falls below lotsfree. With priority paging, the scanner starts when free memory falls below cachefree. Rather than identifying and freeing any pages, the new algorithm will only free file system pages. Files system pages do not include those which are shared libraries and executables.

A system with enough memory to support the working set size of the processes it is hosting should not need to page it's applications out, and the swap device should be mostly idle.

Additional paging statistics

In the Solaris 2.7 kernel, there are new statistics to help identify system behavior. To access these you will need the memstat utility. Note that memstat is currently unsupported and uncommitted, but plans are in progress to integrate this functionality into vmstat. These statistics are not provided in previous releases as they increase the size of a data structure that some other tools may depend upon.

The memstat command is like vmstat, but breaks out pagein/pageout/page free into three different categories:

```
epi - Executable page in's

epo - Executable page out's

epf - Executable page free's

api - Anonymous page in's

apo - Anonymous page out's

apf - Anonymous page free's

fpi - File page in's

fpo - File page out's

fpf - File page free's
```

Systems without a memory shortage should see little or no activity in the epf and apo fields. Significant consistent activity in these fields indicate a memory shortage on the system.

Note that without priority paging enabled, executables and anonymous memory will be paged with even the smallest amount of I/o on the file system, once memory falls to lotsfree.

```
# memstat 3

executable - - anonymous - -- filesys -- --- cpu ---

epi  epo epf api apo apf fpi fpo fpf  us  sy  wt  id

   0   0   0   0   0   0   0   0   0   2   1   0  97

   0   0   0   0   0   0   0   0   0   3   1   0  96

   0   0   0   0   0   0   0   0   0   3   0   0  97

   0   0   0   0   0   0   0   0   0   3   1   0  96

   0   0   0   0   0   0   0   0   0   3   1   0  96

   0   0   0   0   0   0  10 768 768   2  12   0  86

   0   0   0   0   0   0   0 970 970   3   3   0  94

   0   0   0   0   0   0   0 970 970   3   3   0  94
```

```
 0   0   0   0   0   0   0 952 952   3   4   0  92

 0   0   0   0   0   0   0 970 970   2   5   0  93

 0   0   0   0   0   0   0 746 746   5  20   0  75
```

Home : Sun on the Net