

swapping and the value of /proc/sys/vm/swappiness

Source: <http://linux.derkeiler.com/Mailing-Lists/Kernel/2004-09/2077.html>

From: Ray Bryant (raybry_at_sgi.com)

Date: 09/06/04

Date: Mon, 06 Sep 2004 14:11:29 -0500

To: Andrew Morton <akpm@osdl.org>

Andrew (et al),

The attached results started as an exercise to try to understand what value of "swappiness" we should be recommending to our Altix customers when they start running Linux 2.6 kernels. The benchmark is very simple — a task first mallocs around 90% of memory, touches all of the memory, then sleeps forever. After the task begins to sleep, we start up a bunch of "dd" copies. When the dd's all complete, we record the amount of swap used, the size of the page cache, and the data rates for the dd's. (Exact details are given in the attachment.) The benchmark was repeated for swappiness values of 0, 20, 40, 60, 80, 100, for a number of recent 2.6 kernels.

What is unexpected is that the amount of swap space used at a particular swappiness setting varies dramatically with the kernel version being tested, in spite of the fact that the basic `swap_tendency` calculation in `refile_inactive_zone()` is unchanged. (Other, subtle changes in the vm as a whole and this routine in particular clearly effect the impact of that computation.)

For example, at a swappiness value of 0, Kernel 2.6.5 swapped out 0 bytes, whereas Kernel 2.6.9-rc1-mm3 swapped out 10 GB. Similarly, most kernels have a significant change in behavior for swappiness values near 100, but for SLES9 the change point occurs at `swappiness=60`.

A scan of the change logs for swappiness related changes shows nothing that might explain these changes. My question is: "Is this change in behavior deliberate, or just a side effect of other changes that were made in the vm?" and "What kind of swappiness behavior might I expect to find in future kernels?".

Lots more detail is in the attachment.

--

Best Regards,

Ray

Ray Bryant

Linux–Kernel: swapping and the value of /proc/sys/vm/swappiness

- (2) 2.6.5–7.97–default (SLES9) Swaps nothing until swappiness is ≥ 60 at which point all of the malloc'd area is swapped out. I/O rate is high until swapping starts. This kernel is the only one tested for which the swappiness change point was swappiness=60 instead of swappiness=100.
- (3) 2.6.7: Avg swapout is 1600–1900 MB for all values of swappiness below 100, but there is wide variation in the trials (see the max and min values). Max I/O rate is significantly better than for previous kernels. At swappiness of 100, entire malloc'd area is swapped out.
- (4) 2.6.8.1–mm4: much like 2.6.7, except the average swap is smaller below swappiness of 100. I/O rate is similar to that of 2.6.7. Still significant variations among the trials, but not quite as severe as 2.6.7.
- (5) 2.6.9–rc1–mm3: Well, it almost looks as if the swappiness code is broken in this version. Even for swappiness of 0, it swaps out 10 GB worth of data. Can this be right?

A comparison of the `refill_inactive_zone()` routines from the above kernel versions shows that there are subtle differences in the scanning loop, but that the base calculation for `swap_tendency` (and hence for the influence of swappiness on the decision to set `reclaim_mapped`) are the same.

Scanning the changelogs for swappiness doesn't bring up any changes, so I wonder if this change was deliberate or inadvertent?

So my question is, is all of this intended, and which variation of swappiness behavior is the one I should expect in future kernels?

Kernel Version 2.6.5:

Total I/O Avg Swap min max pg cache min max

```
0 279.56 MB/s 0 MB ( 0, 0) 3121 MB ( 3009, 3233)
20 273.99 MB/s 0 MB ( 0, 0) 3060 MB ( 3024, 3104)
40 285.12 MB/s 0 MB ( 0, 0) 2996 MB ( 2945, 3057)
60 289.46 MB/s 115 MB ( 62, 185) 3120 MB ( 3056, 3167)
80 288.58 MB/s 103 MB ( 68, 212) 3181 MB ( 3104, 3265)
100 151.31 MB/s 25513 MB ( 25380, 25699) 27760 MB ( 27569, 28093)
```

Kernel Version 2.6.5–7.97–default (SLES9):

Total I/O Avg Swap min max pg cache min max

```
0 273.57 MB/s 0 MB ( 0, 0) 3191 MB ( 3168, 3229)
20 273.75 MB/s 0 MB ( 0, 0) 3151 MB ( 3088, 3180)
40 273.52 MB/s 0 MB ( 0, 0) 3096 MB ( 3076, 3124)
60 229.01 MB/s 23068 MB ( 22042, 24195) 25564 MB ( 24578, 26689)
80 195.63 MB/s 25587 MB ( 25227, 25815) 28046 MB ( 27681, 28260)
```

Linux–Kernel: swapping and the value of /proc/sys/vm/swappiness

100 184.30 MB/s 26006 MB (26006, 26006) 28388 MB (28349, 28434)

Kernel Version 2.6.6:

Total I/O Avg Swap min max pg cache min max

0 242.47 MB/s 0 MB (0, 0) 3195 MB (3138, 3266)
20 256.06 MB/s 0 MB (0, 0) 3170 MB (3074, 3234)
40 267.29 MB/s 0 MB (0, 0) 3189 MB (3137, 3234)
60 289.43 MB/s 666 MB (72, 1680) 3847 MB (3296, 4817)
80 286.49 MB/s 170 MB (86, 393) 3211 MB (2897, 3618)
100 154.87 MB/s 24663 MB (24320, 25054) 26708 MB (26274, 27154)

Kernel Version 2.6.7:

Total I/O Avg Swap min max pg cache min max

0 287.52 MB/s 1688 MB (1590, 2069) 4966 MB (4819, 5363)
20 289.80 MB/s 1838 MB (25, 3741) 5081 MB (3265, 6932)
40 290.39 MB/s 1970 MB (1593, 3453) 5270 MB (4899, 6660)
60 290.25 MB/s 1271 MB (4, 1591) 4559 MB (3297, 4915)
80 288.89 MB/s 1599 MB (6, 3220) 4876 MB (3345, 6436)
100 158.67 MB/s 25768 MB (25474, 26004) 28363 MB (27968, 28753)

Kernel Version 2.6.8.1–mm4:

Total I/O Avg Swap min max pg cache min max

0 287.28 MB/s 710 MB (46, 3060) 4082 MB (3426, 6308)
20 288.05 MB/s 508 MB (94, 1417) 3848 MB (3442, 4739)
40 287.03 MB/s 588 MB (199, 1251) 3909 MB (3570, 4515)
60 290.08 MB/s 640 MB (210, 1190) 3976 MB (3538, 4531)
80 287.73 MB/s 693 MB (316, 1195) 4049 MB (3713, 4545)
100 166.17 MB/s 26001 MB (26001, 26002) 28798 MB (28740, 28852)

Kernel Version 2.6.9–rc1–mm3:

Total I/O Avg Swap min max pg cache min max

0 274.80 MB/s 10511 MB (5644, 14492) 13293 MB (8596, 17156)
20 267.02 MB/s 12624 MB (5578, 16287) 15298 MB (8468, 18889)
40 267.66 MB/s 13541 MB (6619, 17461) 16199 MB (9393, 20044)
60 233.73 MB/s 18094 MB (16550, 19676) 20629 MB (19103, 22192)
80 213.64 MB/s 20950 MB (15844, 22977) 23450 MB (18496, 25440)
100 164.58 MB/s 26004 MB (26004, 26004) 28410 MB (28327, 28455)

–

To unsubscribe from this list: send the line "unsubscribe linux–kernel" in the body of a message to majordomo@vger.kernel.org

More majordomo info at <http://vger.kernel.org/majordomo–info.html>

Please read the FAQ at <http://www.tux.org/lkml/>