Subject: Weird IOwait raising using 028stab085.2 x86_64 and i686 kernel
Posted by AnVir on Sun, 20 Mar 2011 19:17:40 GMT
View Forum Message <> Reply to Message

I did kernel upgrade on two production HN. Each is 8 cores intel 2.5 GHz, one has 16 GB ram with x86_64 CentOS 5.5, another has 8 GB ram with i686 CentOS 5.5 (enterprise kernel).
On the first, 64 bit HN started about 80 VEs, each vith 1-2 java process, total ~15000 threads on HN.
On the second, 32 bit HN started about 40 VEs, each vith 1 proccess of HLDS or SRCDS game server, total ~2500 threads.
Each HN has 2x500 GB SATA HDD without RAID of any kind.
Memory load is about 50%, so IOwait not raises more than 15% even in primetime.
Kernel was upgraded from 028stab081.1 to 028stab085.2.
About 60% of VEs started as always (with high IOwait, but still good disk IO responce time - it is usual during HN boot, while disk cache is filling up), then two of eight cores on each HN show me 100% IOwait state, and other six - 0% load at all. At the same time LA start to raise from usual 50-200 value to ~4000 (maybe, higher, but i reboot HN at this point, ~30 vinutes of uptime).
I tryed to perform some actions, for example, to find processes in "D..." state trough "ps axuww | grep ' D"'. I got some free breath when I killed all syslogd and crond processes, but not for long, after 3-5 minutes system hang again with the same symtomths.
So, I was forced to revert kernels to 028stab081.1, did reboot - and both nodes started as usual, with low IOwait when disk cache was filled.

I have some specific settings on file systems, in sysctl (/proc) and disk scheduler, which I got by tuning up them for about a year on 12 hardware nodes with compareble load. I tryed to change them to much lower/higher values, hope to obtain the reason of this activity, and got no effect.
Some of this settings you can see below:

In /etc/fstab:
/dev/sda3 /vz/private ext3 defaults,noatime,nodiratime,commit=29,data=journal 1 1
Commit interval on each FS is different, about 30 seconds, and it is simple number (divides only by 1 and by self: 23, 29, 31 etc).

In /etc/sysctl.conf:
kernel.vcpu_sched_timeslice = 5
kernel.vcpu_hot_timeslice = 4
kernel.fairsched-max-latency = 20
kernel.sched_interactive = 0
kernel.hung_task_timeout_secs = 60
kernel.max_lock_depth = 1024
kernel.sysrq = 0
kernel.exec-shield = 0
kernel.randomize_va_space = 0
kernel.pid_max = 65534
vm.swappiness = 1
vm.pagecache = 90
vm.vfs_cache_pressure = 1000
vm.flush_mmap_pages = 0

vm.dirty_background_ratio = 10
vm.dirty_writeback_centisecs = 3013
vm.dirty_expire_centisecs = 30031
vm.dirty_ratio = 30
vm.max_writeback_pages = 65536
I repeat: all of them i tried to change while solving the problem. It always worked fine, and more - it gains performance with my tasks.

In /etc/rc.local:
blockdev --setra 2048 /dev/sda
blockdev --setra 2048 /dev/sdb
echo 4096 > /sys/block/sda/queue/nr_requests
echo 4096 > /sys/block/sdb/queue/nr_requests
echo 'cfq' > /sys/block/sda/queue/scheduler
echo 'cfq' > /sys/block/sdb/queue/scheduler
Changing length of queue, RA or cfq scheduler to deadline or noop has no effect.

I'll post any additional information if you require any to help me. Now, I installed 028stab087.1 testing kernel on 64-bit node that I described above, and plan to boot it in next 2-3 hours. I see a lot of fixes in changelog - maybe, someone fixed my issue already... Anyway, thank you for any reply

...and sorry for my ugly english spelling.
