Hardware & OS description

Motherboard: Tyan K8SD-Pro with two AMD Opteron 246 (2Ghz)

RAID :

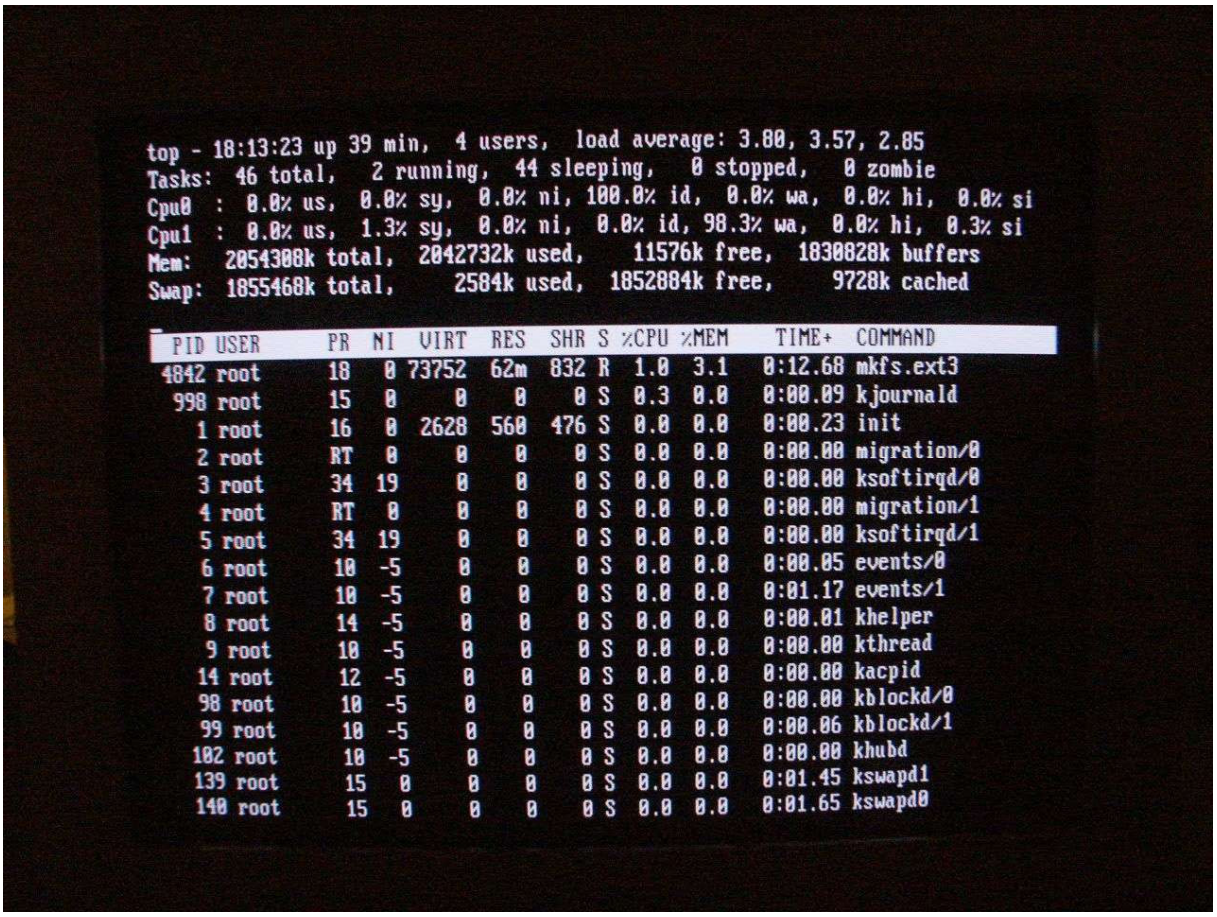
- 3ware 9550SX-4LP, PCI-X 133Mhz
- 4 Seagate Barracuda 7200.9 SATA2 (500Gb/3Gbps/NCQ) HDD.
- The array is setup to use the 4 disk drives in RAID-5 mode (1.5TB)
- Bios and Firmware from 9.3.0.1 suite

OS: Ubuntu linux 5.10.

Problems : Heavy load and CPUs in high "wa"

Making the filesystem structure

I first created some partitions on the array and managed to create a filesystem on them using usual ext3. The filesystem structure begins to create normally but at some level the process slows down until it is rather unusable and uninterruptible. I managed to understand what was going on and had a look at the top output :



"top" output while doing mkfs.ext3 on 1.5TB partition

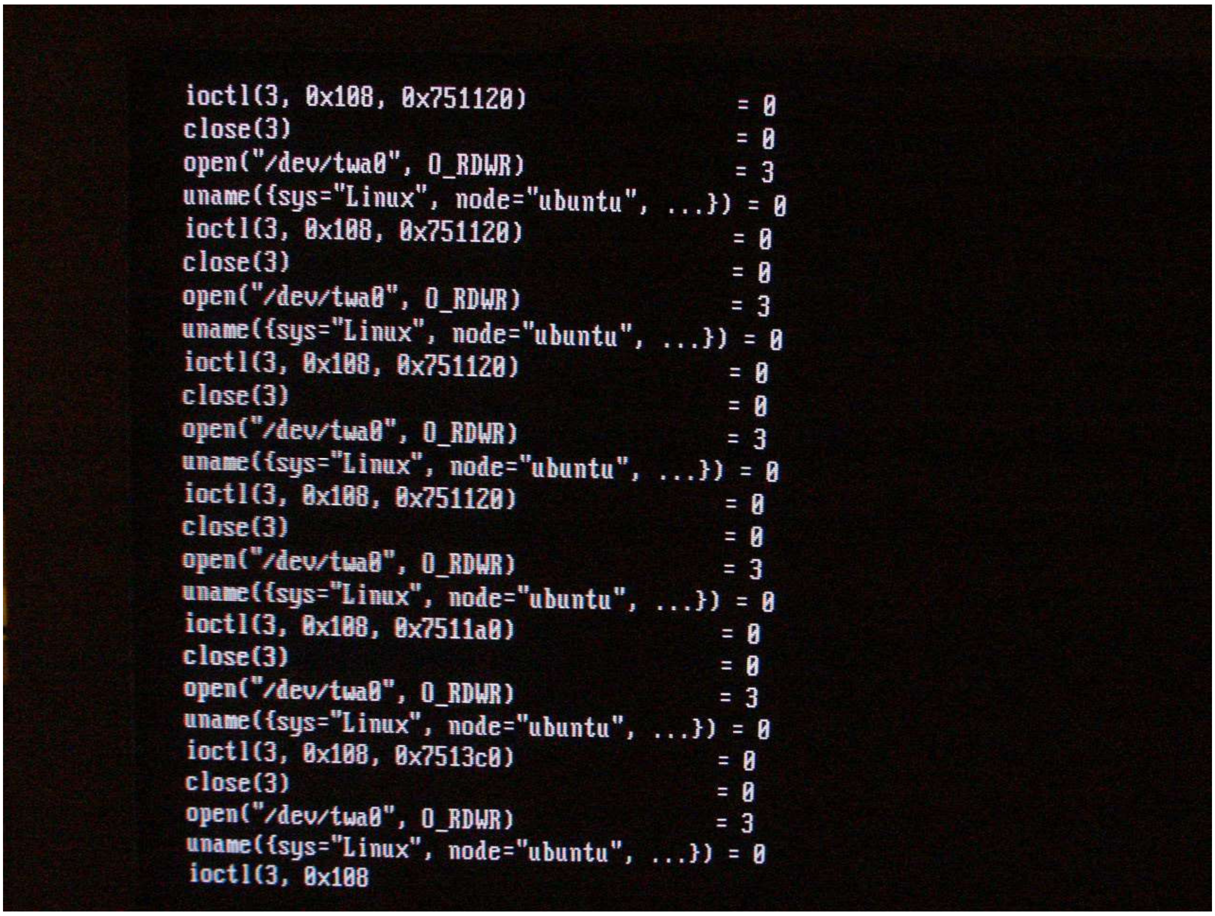
The system load reaches 4.50 (oops) Both CPUs are totally idle (in term of computation) and top reports a "wa" of 100% (for the CPU executing mkfs.ext3 which was currently accessing the array). When doing this with a non SMP kernel (thus with only one CPU available), the computer is totally unworkable. No interactive mode is possible. In dual CPU mode, as the second CPU is not locked in IOs, you can type commands and have some results if you're patient.

Rsycing data from one server to another

After the formatting was done, and the partition mounted, I managed to rsync over ssh all the data from my dying SCSI server to the fresh new bi-amd64 server. It led me to even worse results. While the SCSI server reads and encrypts data trough the ssh channel with a load of 0.7 or so, the bi amd-64 decrypts and writes the data on the array with a load near 7.50 and both CPUs are idle but locked with "wa" at 100%.

Using 3ware programs during IOs

I tried to launch and use tw_cli (the CLI program to manage array) or 3dm2 (the web based version). The response time was as low as other commands such "find" or "ls" for uncached dirs. I tried to attach strace and observed that the CLI program was issuing the same ioctl() call each second until the array accept it. This ioctl() call is done 5 to 15 times before it completes. Doing another query leads to the same type of comportment with suites of 5 to 15 groups of the same ioctl() call, until all are accepted and done.

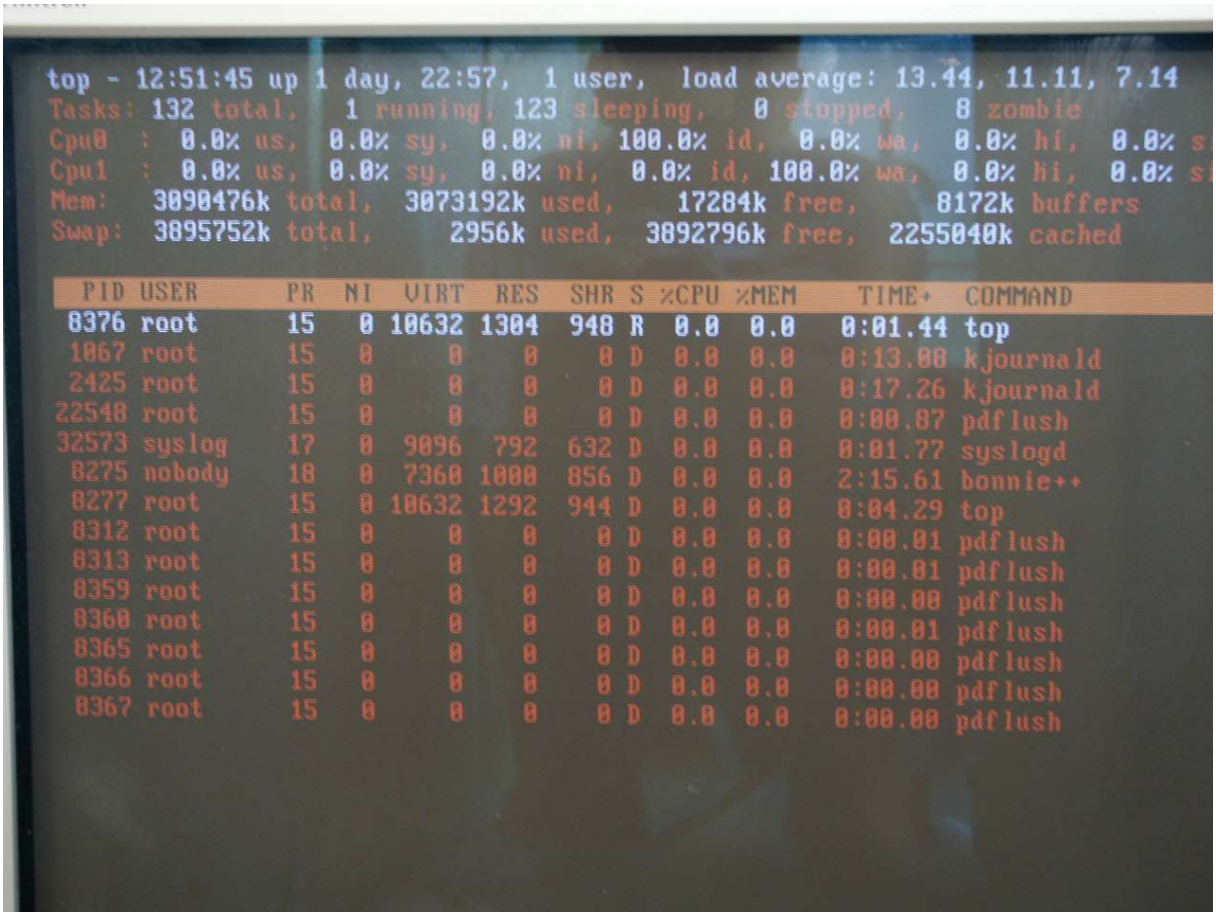


```
ioctl(3, 0x108, 0x751120)                         = 0
close(3)                                          = 0
open("/dev/twa0", O_RDWR)                         = 3
uname({sys="Linux", node="ubuntu", ...}) = 0
ioctl(3, 0x108, 0x751120)                         = 0
close(3)                                          = 0
open("/dev/twa0", O_RDWR)                         = 3
uname({sys="Linux", node="ubuntu", ...}) = 0
ioctl(3, 0x108, 0x751120)                         = 0
close(3)                                          = 0
open("/dev/twa0", O_RDWR)                         = 3
uname({sys="Linux", node="ubuntu", ...}) = 0
ioctl(3, 0x108, 0x751120)                         = 0
close(3)                                          = 0
open("/dev/twa0", O_RDWR)                         = 3
uname({sys="Linux", node="ubuntu", ...}) = 0
ioctl(3, 0x108, 0x7511a0)                         = 0
close(3)                                          = 0
open("/dev/twa0", O_RDWR)                         = 3
uname({sys="Linux", node="ubuntu", ...}) = 0
ioctl(3, 0x108, 0x7513c0)                         = 0
close(3)                                          = 0
open("/dev/twa0", O_RDWR)                         = 3
uname({sys="Linux", node="ubuntu", ...}) = 0
ioctl(3, 0x108
```

Suites of the same Ioctl() calls in tw_cli, while the array is accessed by rsync

Bonnie++

Launching bonnie++ remotely via some ssh session leads to the total congestion of the machine. My terminal is frozen. Trying to connect using a new ssh session fails. The only thing I can do is accessing apache on port 80, which serves pages at the speed of a RTC connection (on a 100Mbps network …). So I went downstairs logging onto the console. After 5 minutes of patience before my shell prompt and 5 minutes more before top display I took this capture. The results are clear: Total load near 14 and 100% wa, a lot of pdflush and some kjournald, all in "D" status.



Bonnie++ suicides the server

Write cache disabled

```
Version  1.03       ------Sequential Output------ --Sequential Input- --Random-
                    -Per Chr- --Block-- -Rewrite- -Per Chr- --Block-- --Seeks--
Machine        Size K/sec %CP K/sec %CP K/sec %CP K/sec %CP K/sec %CP  /sec %CP
twinpeaks        6G  7202  14  6672   2  4831   1 16445  27 115780  13 286.0   0
                    ------Sequential Create------ --------Random Create--------
                    -Create-- --Read--- -Delete-- -Create-- --Read--- -Delete--
               files  /sec %CP  /sec %CP  /sec %CP  /sec %CP  /sec %CP  /sec %CP
                  16   909  89 +++++ +++ +++++ +++   880  94 +++++ +++  8438  99
```

Write cache enabled

```
Version  1.03       ------Sequential Output------ --Sequential Input- --Random-
                    -Per Chr- --Block-- -Rewrite- -Per Chr- --Block-- --Seeks--
Machine        Size K/sec %CP K/sec %CP K/sec %CP K/sec %CP K/sec %CP  /sec %CP
twinpeaks        6G 43477  86 52946  23 26239   6 36262  59 130698  15 280.6   0
                    ------Sequential Create------ --------Random Create--------
                    -Create-- --Read--- -Delete-- -Create-- --Read--- -Delete--
               files  /sec %CP  /sec %CP  /sec %CP  /sec %CP  /sec %CP  /sec %CP
                  16  1567  99 +++++ +++ +++++ +++  2017  99 +++++ +++  9030  99
```

Summary of tests done

WinXP (32 bits)
- Works perfectly

Linux (64bits): fails
- Kernel 2.6.14.2 amd64-smp failed
- Kernel 2.6.14.2 amd64-smp failed (gcc-4.0.2)
- Kernel 2.6.12 amd64-smp failed
- Kernel 2.6.12 amd64-generic failed

Linux (32bits): fails
- Kernel 2.6.12 i386-generic failed

Conclusions

I also did some dd to the array before creating the ext3 fs. The results were the same, high load and 100% wa. It's almost certain that the problems are not in the ext3 layer and the journal handling.

I tried the array under Windows XP. The formatting was in minutes instead of hours. Copying files from some samba export to the array, while doing other copies from the boot hdd to the array haven't led to any load at all. IOZone ran flawlessly. Everything was perfect (if we leave apart the fact that it's XP and I will never, even under torture, build a XP server). Thus we can conclude that it is not a hardware problem.

I suspect the driver and/or something in the IO chain to be faulty

I opened some calls to 3ware but the responses are very short and no useful information was returned back ("please activate the write cache"), of course I have tested all the possibilities between "write cache", "NCQ" and whatever before doing the call.