

How I built a 2.8TB RAID storage array

Source: <http://linux.derkeiler.com/Newsgroups/comp.os.linux.hardware/2005-02/0537.html>

From: Yeechang Lee (ylee_at_pobox.com)

Date: 02/20/05

Date: 20 Feb 2005 03:14:21 GMT

My 2.8TB RAID 5 array is finally up and running. Here I'll discuss my initial intended specifications, what I actually ended up with, and associated commentary. Please see

<URL:<http://groups.google.ca/groups?selm=slrnch28at.j0n.ylee%40pobox.com>>
and

<URL:<http://groups.google.ca/groups?selm=slrncu34ip.55k.ylee%40pobox.com>>
for background material.

STORAGE MEDIUM

Initial: Eight 250GB SATA drives.

Actual: Nine 400GB PATA drives; eight for use, one as a cold spare.

Why: Found a stupendous sale at CompUSA Christmas week; just-released-in-November Seagate Barracuda 7200.8 400GB PATA drives at \$230 each, with no quantity limitation. I'd have loved to have gone with the SATA model, but given that Froogle lists the lowest price for one at \$350 (the PATA model retails at \$250-350), it was an easy choice.

CASE

Initial: Antec tower case.

Actual: Antec 4U rackmount case.

Why: I'd always thought of rackmounts as unsuitable for anyone with an actual rack sitting in their data center, but after realizing that a rackmount case is simply a tower case sitting on its size, it was an easy decision given the space advantages. The Antec case here comes with Antec's True Power 550W EPS12V power supply, and both have great reputations. In practice, I found that the Antec case was remarkably easy to open up (one thumbscrew), work with (all drive cages are removable), and roomy.

MOTHERBOARD

Initial: Unspecified, but probably something Athlon-based and cheap.

Actual: Gigabyte X5DAL-G Intel server motherboard

Why: I became convinced that the sheer volume of the PCI traffic generated by my proposed array under software RAID would overwhelm any non-server motherboard, resulting in errors. In addition, I wanted PCI-X slots for optimal performance. Even though I think AMD in general offers much better bang for the buck, since I didn't want to

comp.os.linux.hardware: How I built a 2.8TB RAID storage array

spend the \$\$\$ for Opteron, a Xeon motherboard with an Intel server chipset was the best compromise.

CONTROLLER CARDS

Initial: Two Highpoint RocketRAID 454 cards.

Actual: Two 3Ware 7506-4LP cards.

Why: I needed PATA cards to go with my PATA drives, and also wanted to put the two PCI-X slots on my motherboard to use. I found exactly two PATA PCI-X controller cards: The 3Ware, and the Acard AEC-6897. Given that the Acard's Linux driver compatibility looked really, really iffy, I went with the 3Ware. I briefly considered the 7506-8 model, which would've saved me about \$120, but figured I'd be better off distributing the bandwidth over two PCI-X slots rather than one.

SOFTWARE

Initial: Linux software RAID 5 and XFS or JFS.

Actual: Linux software RAID 5 and JFS.

Why: Initially I planned on software RAID knowing that the Highpoint (and the equivalent Promise and Adaptec cards) didn't do true hardware RAID. Even after switching over to 3Ware (which *does* do true hardware RAID), everything I saw and read convinced me that software RAID was still the way to go for performance, long-term compatibility, and even 400GB extra space (given I'd be building one large RAID 5 array instead of two smaller ones).

I saw *lots* of conflicting benchmarks on whether XFS or JFS was the way to go. Ultimately
<URL:http://pcbunn.cacr.caltech.edu/gae/3ware_raid_tests.htm> pushed me toward JFS, but I suspect I could have gone XFS with no difficulty whatsoever.

COST

As implied above, I paid \$2070 plus sales tax for the drives. I lucked out and found a terrific eBay deal for a prebuilt system containing the above-mentioned case and motherboard, two Xeon 2.8GHz CPUs, a DVD drive, and 2GB memory for \$1260 including shipping labor aside, I'd have paid *much* more to build an equivalent system myself. The 3Ware cards were \$240 each, no shipping or tax, from Monarch Computer. With miscellaneous costs (such as a Cooler Master 4-in-3 drive cage and an 80GB boot drive from Best Buy for \$40 after rebates), I paid under \$4100, tax and shipping included, for everything. At \$1.46/GB *plus* a powerful dual-CPU system, boatloads of memory, and a spare drive, I am quite satisfied with the overall bang for the buck.

ASSEMBLY: HARDWARE

I spent most of the assembly time on the physical assembly part; it's astonishing just how long the simple tasks of opening up each retail-boxed drive, screwing the drive into the drive cage, putting the cage into the case, removing the cage and the drive when you realize you've put the drive in with the wrong mounting holes, reinstalling the drive and cage, etc., etc. take! My studio apartment

still looks like a computer store exploded inside it.

3Ware wisely provides PATA master-only cables with its cards, which saved some room, but my formerly-roomy case nonetheless looks like the rat's nest to end all rat's nests inside.

ASSEMBLY: SOFTWARE

I'd gone ahead and installed Fedora Core 3 with the boot drive only before the controller cards arrived. The 3Ware cards present each PATA drive as a SCSI device (/dev/sd[a-h]). Once booted, I used mdadm to create the RAID array (no partitions; just whole drives). While the array chugged along to create the parity information (about four hours), I then created one large LVM2 volume group and logical volume on top of the array, then created one large JFS file system.

By the way, I found a RAID-related bug with Fedora Core's bootscripts; see <URL:https://bugzilla.redhat.com/beta/show_bug.cgi?id=129633>).

RESULTS

'df -h':

```
/dev/mapper/VolGroup01-LogVol00
    2.6T 221G 2.4T 9% /mnt/newspace
```

'mdadm --detail /dev/md0':

```
Version : 00.90.01
Creation Time : Wed Feb 16 01:53:33 2005
Raid Level : raid5
Array Size : 2734979072 (2608.28 GiB 2800.62 GB)
Device Size : 390711296 (372.61 GiB 400.09 GB)
Raid Devices : 8
Total Devices : 8
Preferred Minor : 0
Persistence : Superblock is persistent

Update Time : Sat Feb 19 16:26:34 2005
State : clean
Active Devices : 8
Working Devices : 8
Failed Devices : 0
Spare Devices : 0
```

```
Layout : left-symmetric
Chunk Size : 512K
```

```
Number Major Minor RaidDevice State
 0  8  0  0 active sync /dev/sda
 1  8 16  1 active sync /dev/sdb
 2  8 32  2 active sync /dev/sdc
 3  8 48  3 active sync /dev/sdd
 4  8 64  4 active sync /dev/sde
 5  8 80  5 active sync /dev/sdf
```

comp.os.linux.hardware: How I built a 2.8TB RAID storage array

```
6 8 96 6 active sync /dev/sdg
7 8 112 7 active sync /dev/sdh
Events : 0.319006
```

```
'bonnie++ -s 4G -m 3ware-swraid5-type -p 3 ; \
bonnie++ -s 4G -m 3ware-swraid5-type-c1 -y & \
bonnie++ -s 4G -m 3ware-swraid5-type-c2 -y & \
bonnie++ -s 4G -m 3ware-swraid5-type-c3 -y &'
```

(To be honest these results are just a bunch of numbers to me, so any interpretations of them are welcome. I should mention that these were done with three distributed computing [BOINC, mprime, and Folding@Home] projects running in the background. Although 'nice -n 19' each, they surely impacted CPU and perhaps disk performance somewhat.)

```
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
3ware-swraid5-ty 4G 15749 50 15897 8 7791 6 10431 49 20245 11 138.1 2
-----Sequential Create----- -----Random Create-----
--Create-- --Read-- --Delete-- --Create-- --Read-- --Delete--
files /sec %CP /sec %CP /sec %CP /sec %CP /sec %CP /sec %CP /sec %CP
16 381 6 ++++++ +++ 208 3 165 7 ++++++ +++ 192 4
```

3ware-swraid5-type-c1,4G,15749,50,15897,8,7791,6,10431,49,20245,11,138.1,2,16,381,6,+++++,+++,208,3,165,7,+ done.

```
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
3ware-swraid5-ty 4G 13739 46 17265 9 7930 6 10569 50 20196 11 146.7 2
-----Sequential Create----- -----Random Create-----
--Create-- --Read-- --Delete-- --Create-- --Read-- --Delete--
files /sec %CP /sec %CP /sec %CP /sec %CP /sec %CP /sec %CP /sec %CP
16 383 7 ++++++ +++ 207 3 162 7 ++++++ +++ 191 4
```

3ware-swraid5-type-c2,4G,13739,46,17265,9,7930,6,10569,50,20196,11,146.7,2,16,383,7,+++++,+++,207,3,162,7,+ done.

```
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
3ware-swraid5-ty 4G 13288 43 16143 8 7863 6 10695 50 20231 12 149.6 2
-----Sequential Create----- -----Random Create-----
--Create-- --Read-- --Delete-- --Create-- --Read-- --Delete--
files /sec %CP /sec %CP /sec %CP /sec %CP /sec %CP /sec %CP /sec %CP
16 537 9 ++++++ +++ 207 3 161 7 ++++++ +++ 188 4
```

3ware-swraid5-type-c3,4G,13288,43,16143,8,7863,6,10695,50,20231,12,149.6,2,16,537,9,+++++,+++,207,3,161,7,+

FINAL NOTES, THOUGHTS, AND QUESTIONS

I've noticed that over sync NFS, initiating a file copy from my older Athlon 1.4GHz system to the RAID array system is *much, much, much* (seconds as opposed to many minutes) slower than if I initiate the copy in the same direction but from the array system. Why is this?

comp.os.linux.hardware: How I built a 2.8TB RAID storage array

I almost went with the SATA (8506) version of the 3Ware cards and a bunch of PATA-SATA adapters in order to maintain compatibility with future drives, likely to be SATA only. However, a colleague pointed out the foolishness of paying \$200 extra (\$120 for eight adapters plus \$80 for the extra cost of the SATA cards) in order to (possibly) futureproof a \$480 investment.

I was concerned that the drives (and the PATA cables) would cause horrible heat and noise issues. These, surprisingly, didn't occur; according to 'sensors', internal temperatures only rose by a few degrees, and the server is just as (very) noisy now as pre-RAID drives. I think I'll be able to get away with stuffing the array inside my hall closet after all.

The server, before I put the cards and RAID drives into the system but with the distributed-computing projects putting the CPU at 100% utilization, took the power output on my Best Fortress 750VA/450W UPS from about 55% to about 76%. With the RAID up and running and again with 100% CPU utilization, output is 87-101% with the median at perhaps 93%. I realize I really ought to invest in another UPS, but with these figures I'm tempted to get by on what I currently have.

Yes, I could've saved a considerable amount of money had I gone with, say, a used dual PIII server system with regular PCI slots (and, thus, \$80 Highpoint RAID cards, again for the four PATA channels and not for their RAID functionality per se) and 512MB. And I suspect that for a home user like me performance wouldn't have been too much less. But I like to buy and build systems I can use for years and years without having to bother with upgrading, and figure I've made a long-term (at least 4-5 years, which is long term in the computer world) investment that provides me with much more than just storage functionality. And again, \$1.46/GB is hard to beat.

```
--  
Read my Deep Thoughts @ <URL:http://www.vlee.org/blog/>      PERTH ----> *  
Cpu(s):  6.7% us,  3.7% sy,  0.4% ni, 75.4% id, 12.3% wa,  1.4% hi,  0.0% si  
Mem:    515800k total,  511628k used,    4172k free,    5812k buffers  
Swap:   2101032k total,  13152k used,  2087880k free,  163928k cache
```