Blog Export: c0t0d0s0.org, http://www.c0t0d0s0.org/

Wednesday, August 9. 2006

Comparing ZFS RAID and HW RAID

It´s one of the long lasting myths, that RAID needs hardware to be fast. It´s like "Disable autoneg" and "1 MHz per MBit". Those rules were true in the past but obsoleted by the time. But the RAID myth seems to be more persistant than all other before. Whenever i talk about ZFS RAID, the people stare at me and and i see the though "Oh, no, the guy talks about slow SW RAID". Okay, when you don´t believe me perhaps you believe a user our products: Robert Milkowski gives some numbers on RAID performance here and here.

Posted by Joerg Moellenkamp in English, Solaris at 17:45

Let me first say that I think ZFS is brilliant and that I wish we could use it for any of our production stuff. This would, for example, let me get rid of that awful LSI Logic RAID controller in our X4200. Brr.

That aside, RAID 0 and 1 never (ok not since 2000) really needed hardware to be fast. These RAID levels just need the parallel I/O channels to be fast.

For RAID 5 or even 6, until recently, you did need hardware to be fast. I really hope ZFS does change that, but I also see that not all questions have been answered yet. For example, I am sure that some of ZFS's impressive performance is due to ZFS handling the complete I/O chain. How will it perform with database raw i/o when it doesn't have the benefit of knowing filesystem semantics? Or with block i/o from virtual machines?

As for the cited benchmarks, who also show cases where RAID Z is even faster than RAID 5 in a 3510 (btw no criticism of the tester, I'm thankful for any good performance data):

The workload of these cases involves lots of random i/o on small blocks, resulting in a relatively low overall throughput of 30-40 MB/sec.

I wonder what would have happened if the V440 used for testing would have had to do RAID Z writes on 200 MB/sec...

Even our old T3+ units can do sustained 90MB/sec RAID 5 writes on large blocks, and these are units with 1GBit FC only, and older and slower 73 GB FC disks.

Anyway, at this time I would advise anyone to run tests with their own real-world workload to see which solution actually performs better.

One more aspect: if you get a 3510 array with 2 controllers, you have a) spent a lot of money (probably too much), but b) you also have redundant raw storage that's accessible from a number of servers in the SAN.

Let's discuss this again if/when a ZFS raw device can become a FC target.
Anonymous on Aug 9 2006, 21:52