

Tuesday, April 24. 2007

ZFS benchmarks

Interesting comment to ZFS performance at ZFS and caching for performance: Watching iostat again also showed that during the read tests, I was using far less of the SAN bandwidth. While performing "read" iotest tests on UFS, I was nearly maxing out the fabric bandwidth which could have lead to resource starvation issues, both for the host running the tests and for other hosts sharing the SAN. Using ZFS, the bandwidth dropped down to 50MB/s or so at peak, and much lower for the remainder of the tests - presumably due to the caching and prefetch behaviour of ZFS. andIt can also give you a staggering performance advantage and conserve IO bandwidth, just so long as you're careful you don't get misled into believing that your storage is faster than it actually is and have enough memory to handle the caching! (via: StorageMojo)

Posted by Joerg Moellenkamp at 06:31

Very interesting!

Since either the spam-preventer or my mind's image recognition has a bug, I can't comment the original blog. So I'm doing it here.

Some remarks:

1. For sustained throughput, we won't see the same behaviour, I think. At some point ZFS just has to write the data. If ZFS then doesn't go higher than 50 MB/Sec bandwidth utilization, it will slow down the throughput. If it does go higher, then the perceived advantage no longer exists - regarding writes.
2. The lozone file size used in the benchmark is too small for that kind of system, I think. For lozone, the file size should at least exceed the RAID controller's memory and also the main memory available to the benchmark. Otherwise, you measure lots of cache performance instead of getting a comprehensive overview. The 3510 has 1GB cache on each controller. We run a single controller 3510 with 12 36GB 15K disks, we did try lozone, and the difference between using 512MB and 2GB file size was staggering.
3. As for the slow write speed: the 3510 with dual controllers has an overhead for the active-active configuration. During writes, the cache needs to be synchronized, and that takes time. In general, we too were a little "underwhelmed" by the 3510s performance. Write speed compared to our dated T3+ units is actually slower, although the T3+ use 9 disks instead of 12 and slower disks, too.
4. As always, I find it rather difficult to apply findings from a benchmark or performance test to a concrete usage scenario.

General comment regarding ZFS:

We will shortly be installing two new machines, and we still won't use ZFS. Reasons? Still no ZFS boot and still not convinced that ZFS is a good match for providing block devices for virtualisation and databases. So we continue looking at ZFS with the I-wish-we-could attitude.

Anonymous on Apr 24 2007, 12:49

1. Of course ZFS can go higher than 50 MB/Second ... generally ZFS makes better use of the resources, by transforming random writes into sequential ones.
2. With Update 4 the performance of an tuned ZFS is within striking distance of a tuned UFS for example with Oracle, while providing all advantages of ZFS.
3. I think the advantages of ZFS to provide block devices are really compelling.

Anonymous on Apr 24 2007, 12:59

I wouldn't compare ZFS with UFS for Oracle or DB2 containers.

The competition is in raw devices on RAID1 or RAID10 volumes, and I'm not convinced that ZFS can compete with these from a performance point of view (for databases!).

The ideal solution: an Oracle and/or DB2 version that hooks into ZFS at a lower level instead of being a regular Joe Vanilla filesystem user.

ZFS for raw devices: ok, for provisioning iSCSI targets, sure, that's a very promising track.

Let's see what happens.

Right now, it would be very nice if I could have a Solaris version (Sparc and x86) where I can just say "use ZFS as filesystem".

Anonymous on Apr 24 2007, 16:51