1 Gbyte/sec NFS, streaming from disk

09 Jan 2009

I originally posted this at http://blogs.sun.com/brendan/entry/1_gbyte_sec_nfs_streaming.

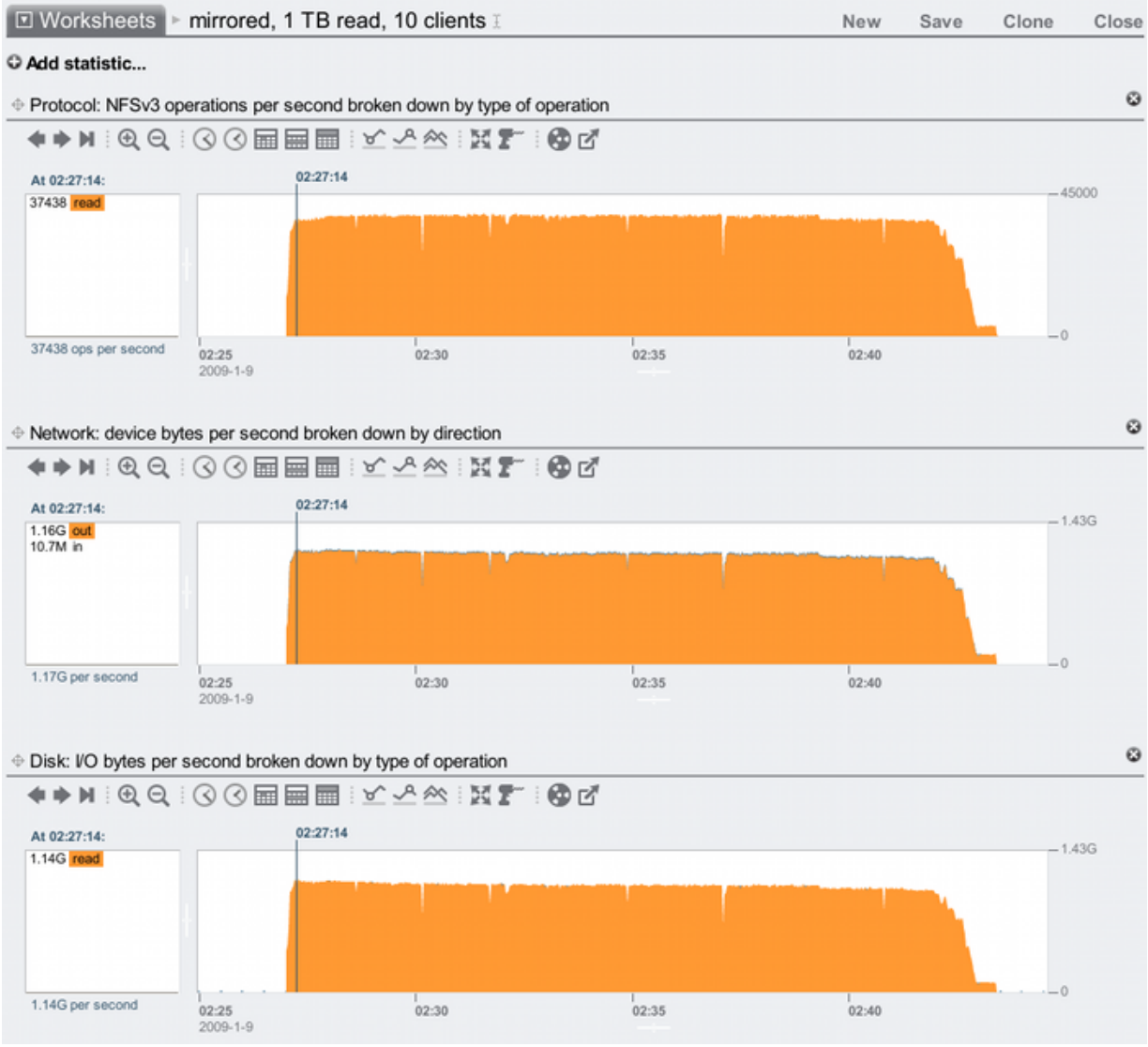
I've previously explored maximum throughput and IOPS that I could reach on a Sun Storage 7410 by caching the working set entirely in DRAM, which may be likely as the 7410 currently scales to 128 Gbytes of DRAM per head node. The more your working set exceeds 128 Gbytes or is shifting, the more I/O will be served from disk instead. Here I'll explore the streaming disk performance of the 7410, and show some of the highest throughputs I've seen.

Roch posted performance invariants for capacity planning, which included values of 1 Gbyte/sec for streaming read from disk, and 500 Mbytes/sec for streaming write to disk (50% of the read value due to the nature of software mirroring). Amitabha posted Sun Storage 7000 Filebench results on a 7410 with 2 x JBODs, which had reached 924 Mbytes/sec streaming read and 461 Mbytes/sec streaming write – consistent with Roch's values. What I'm about to post here will further reinforce these numbers, and include the screenshots taken from the Sun Storage 7410 browser interface to show this in action, specifically from Analytics.

Streaming read

The 7410 can currently scale to 12 JBODs (each with 24 disks), but when I performed this test I only had 6 available to use, which I configured with mirroring. While this isn't a max config, I don't think throughput will increase much further with more spindles. After about 3 JBODs there is plenty of raw disk throughput capacity. This is the same 7410, clients and network as I've described before, with more configuration notes below.

The following screenshot shows 10 clients reading 1 Tbyte in total over NFS, with 128 Kbyte I/Os and 2 threads per client. Each client is connected using 2 x 1 GbE interfaces, and the server is using 2 x 10 GbE interfaces:



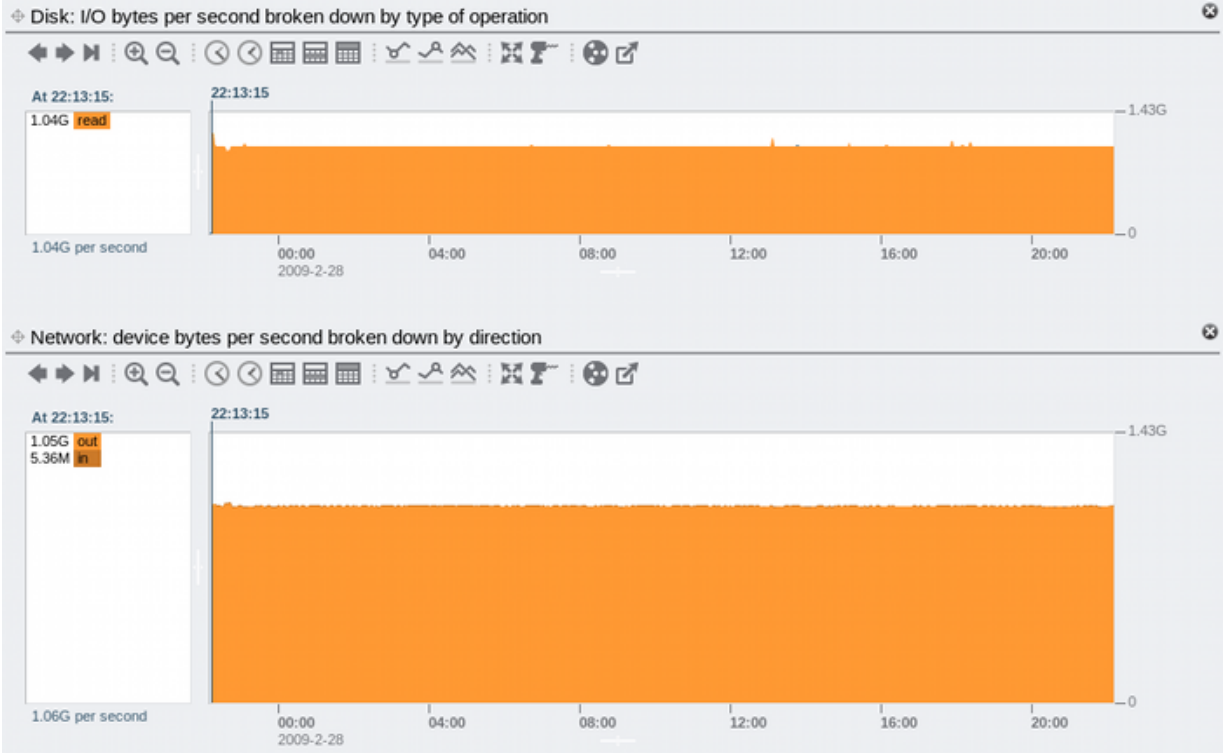
Showing disk I/O bytes along with network bytes is important, as it helps us confirm that this streaming read test did read from disk. If the intent is to go to disk, make sure it does, otherwise it could be hitting from the server or client cache.

I've highlighted the network peak in this screenshot, which was 1.16 Gbytes/sec network outbound. This is just a peak. The average throughput can be seen by zooming in on Analytics (not shown here), which found the network outbound throughput to average 1.10 Gbytes/sec, and disk bytes at 1.07 Gbyte/sec. The difference is that the network result includes data payload plus network protocol headers, whereas the disk result is data payload plus ZFS metadata, which is adding less than the protocol headers. So 1.07 Gbytes/sec is closer to the average data payload read over NFS from disk.

It's always worth sanity checking results however possible, and we can do this here using the times. This run took 16:23 to complete, and 1 Tbyte was moved – that's an average of 1066 Mbytes/sec of data payload – 1.04 Gbytes/sec. This time includes the little step at the end of the run, and when I zoomed in to see the average it didn't include that. The step is from a slow client completing (I found out who using the Analytics statistic "NFS operations broken down by client", and I need to now check why I have one client that is a little slower than the others!) Even including the slow client, 1.04 Gbytes/sec is a great result for delivered NFS reads from disk, on a single head node.

Update: 2-Mar-09

In a comment posted on this blog, Rex noticed a slight throughput decay over time in the above screenshots, and wondered if this continued if left longer. To test this out, I had the clients loop over their input files (files which were much too big to fit in the 7410's DRAM cache), and mounted the clients with forcedirectio (to avoid them caching); this kept the reads being served from disk, which I could leave running for hours. The result:



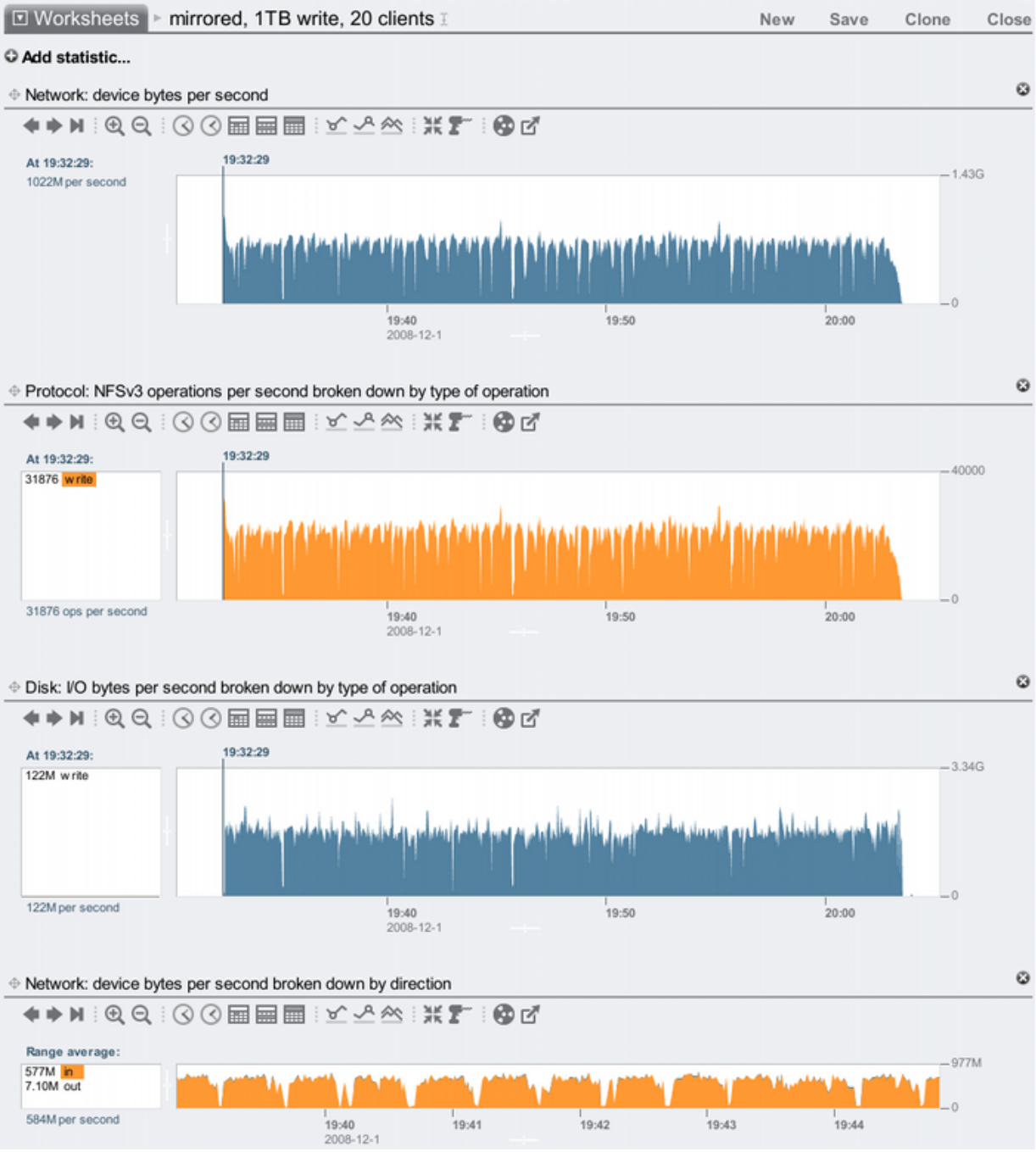
The throughput is rock steady over this 24 hour interval.

I hoped to post some screenshots showing the decay and drilling down with Analytics to explain the cause. Problem is, it doesn't happen anymore, throughput is always steady. I have upgraded our 10 GbE network since the original tests. Our older switches were getting congested and slowing throughput a little, which may have been happening earlier.

Streaming write, mirroring

These write tests were with 5 JBODs (from a max of 12), but I don't think the full 12 would improve write throughput much further. This is the same 7410, clients and network as I've described before, with more configuration notes below.

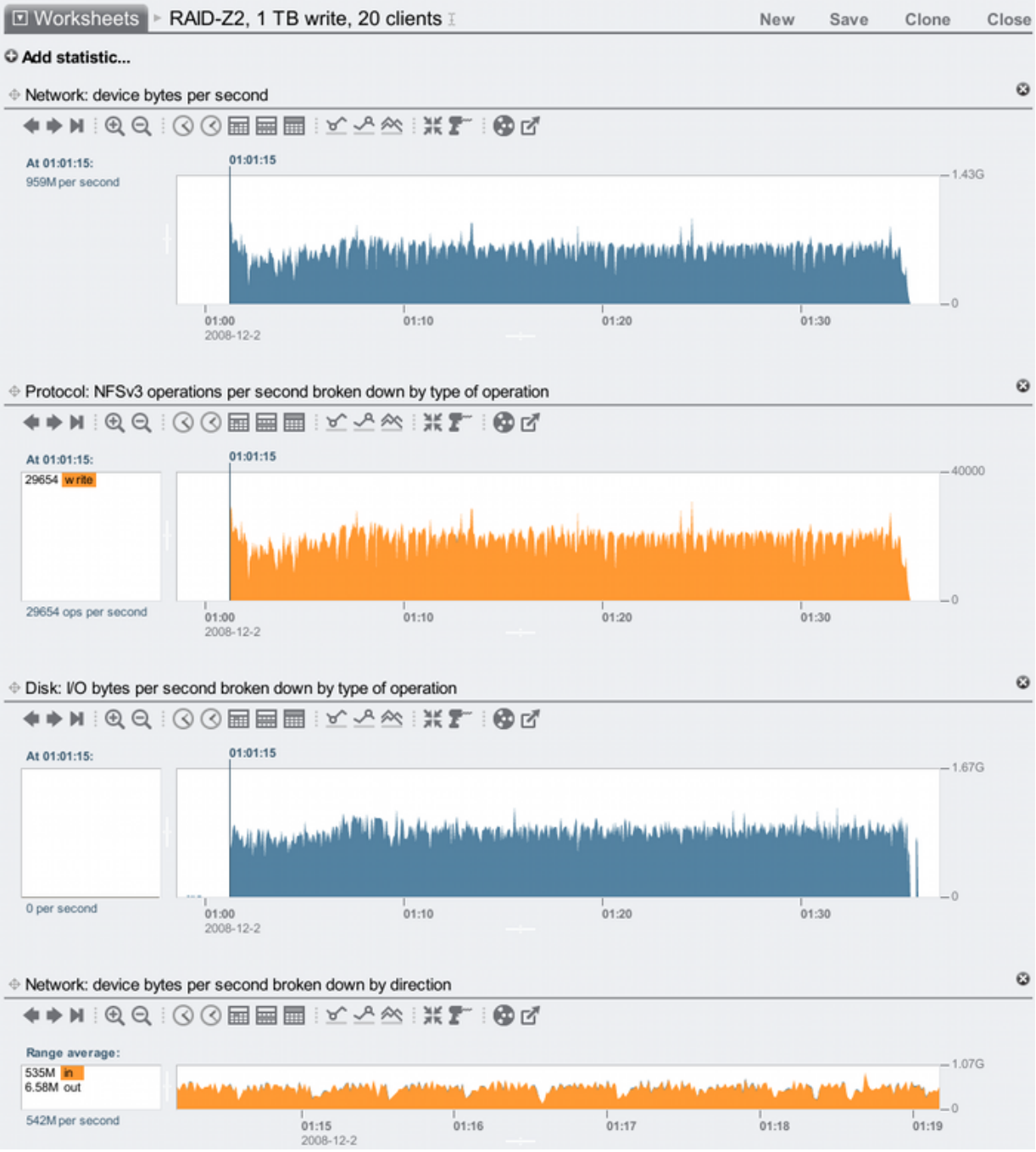
The following screenshot shows 20 clients writing 1 Tbyte in total over NFS, using 2 threads per client and 32 Kbyte I/Os. The disk pool has been configured with mirroring:



The statistics shown tell the story: the network throughput has averaged about 577 Mbytes/sec inbound, shown in the bottom zoomed graph. This network throughput includes protocol overheads, so the actual data throughput is a little less. The time for the 1 Tbyte transfer to complete is about 31 minutes (top graphs), which indicates the data throughput was about 563 Mbytes/sec. The Disk I/O bytes graph confirms that this is being delivered to disk, at a rate of 1.38 Gbytes/sec (measured by zooming in and reading the range average, as with the bottom graph). The rate of 1.38 Gbytes/sec is due to software mirroring ((563 + metadata) x 2).

Streaming write, RAID-Z2

For comparison, the following shows the same test as above, but with double parity raid instead of mirroring:



The network throughput has dropped to 535 Mbytes/sec inbound, as there is more work for the 7410 to calculate parity during writes. As this took 34 minutes to write 1 Tbyte, our data rate is about 514 Mbytes/sec. The disk I/O bytes is much lower (notice the vertical range), averaging about 720 Mbytes/sec – a big difference to the previous test, due to RAID-Z2 vs mirroring.

Configuration Notes

To get the maximum streaming performance, jumbo frames were used and the ZFS record size was left at 128 Kbytes. 3 x SAS HBA cards were used, with dual pathing configured to improve performance.

This isn't a maximum config for the 7410, since I'm testing a single head node (not a cluster), and I don't have the full 12 JBODs.

In these tests, ZFS compression wasn't enabled (the 7000 series provides 5 choices: off, LZJB, GZIP-2, GZIP, GZIP-9). Enabling compression may improve performance as there is less disk I/O, or it may not due to the extra CPU cycles to compress/uncompress the data. This would need to be tested with the workload in mind.

Note that the read and write optimized SSD devices known as Readzilla and Logzilla wern't used in this test. They currently help with random read workloads and synchronous write workloads, neither of which I was testing here.

Conclusion

As shown in the screenshots above, the streaming performance to disk for the Sun Storage 7410 matches what Roch posted, which rounding down is about 1 Gbyte/sec for streaming disk read, and 500 Mbytes/sec for streaming disk write. This is the delivered throughput over NFS, the throughput to the disk I/O subsystem was measured to confirm that this really was performing disk I/O. The disk write throughput was also shown to be higher due to software mirroring or RAID-Z2, averaging up to 1.38 Gbytes/sec. The real limits of the 7410 may be higher than these – this is just the highest I've been able to find with my farm of test clients.

About this blog
Colony Graphs
perf Examples
eBPF Tools
DTrace Tools
DTraceToolkit
DtkshDemos
Guessing Game
Specials
Books
Other Sites