Brendan Gregg's Blog

home

Up to 2 Gbytes/sec NFS

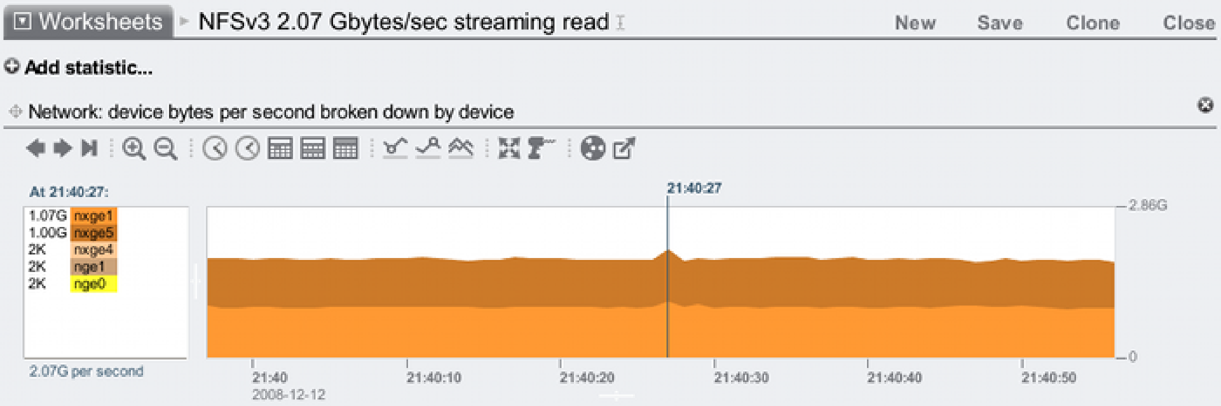
15 Dec 2008

I originally posted this at http://blogs.sun.com/brendan/entry/up_to_2_gbytes_sec.

In a previous post, I showed how many NFS read ops/sec I could drive from a Sun Storage 7410, as a way of investigating its IOPS limits. In this post I'll use a similar approach to investigate streaming throughput, and discuss how to understand throughput numbers. Like before, I'll show a peak value along with a more realistic value, to illustrate why understanding context is important.

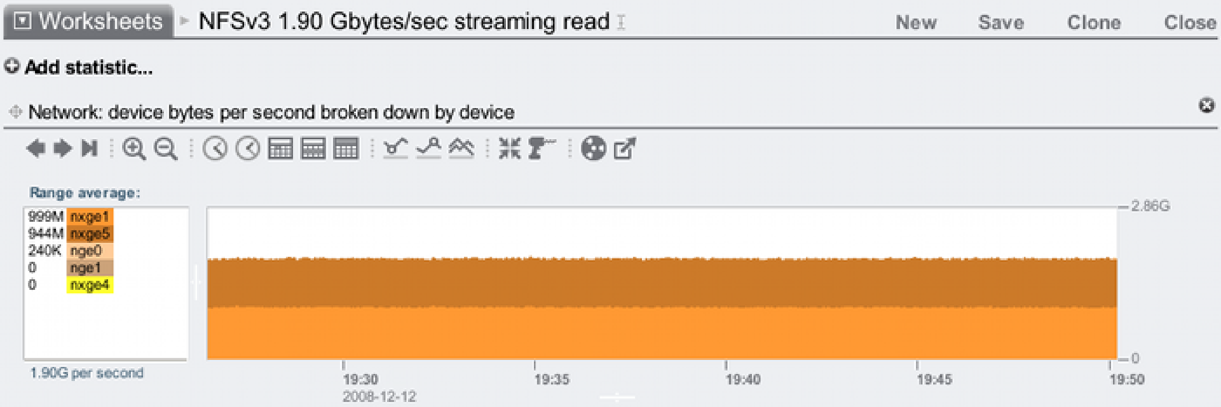
As DRAM scalability is a key feature of the Sun Storage 7410, currently reaching 128 Gbytes per head node, I'll demonstrate streaming throughput when the working set is entirerly cached in DRAM. This will provide some idea of the upper bound: the most throughput that can be driven in ideal conditions.

The screenshots below are from Analytics on a single node Sun Storage 7410, which is serving a streaming read workload over NFSv3 to 20 clients:



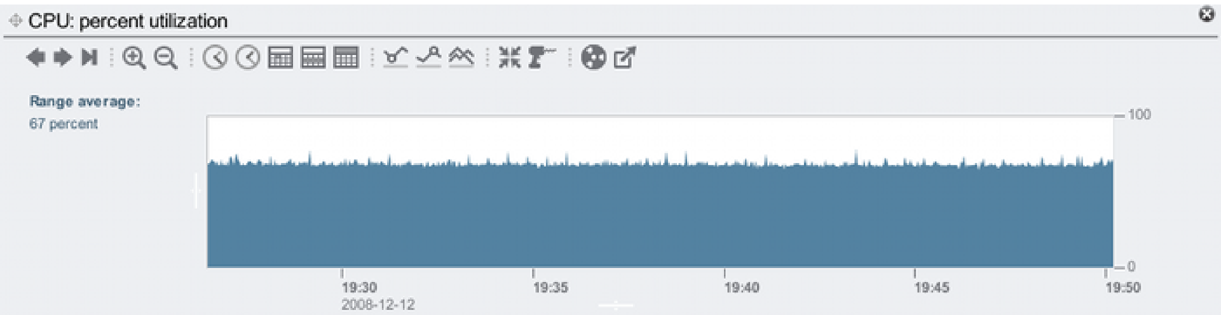
Here I've highlighted a peak of 2.07 Gbytes/sec – see the text above and below the box on the left.

While it's great to see 2 Gbytes/sec reached on a single NAS node, the above screenshot should also show that this was a peak only. It is more useful to see a longer term average:



The average for this interval (over 20 mins) is 1.90 Gbytes/sec. This graph shows network bytes by device, which shows the traffic was balanced across two nxge ports (each 10 GbE, and each about 80% utilised).

1.90 Gbytes/sec a good result, but Analytics suggests the 7410 can go higher. Resources which can bound (limit) this workload include the CPU cycles to process instructions, and the CPU cycles to wait for the memory instructions needed to push 1.90 Gbytes/sec -- both of these are reported as "CPU utilization":



For the same time interval, the CPU utilization on the 7410 was only around 67%. Hmm. Be careful here – the temptation is to do some calculations to predict where the limit could be based on that 67% utilization, but there could be other resource limits that prevent us from going much faster. What we do know is that the 7410 has sustained 1.90 Gbytes/sec and peaked at 2.07 Gbytes/sec in my test. The 67% CPU utilization encourages me to do more testing, especially with faster clients (these have 1600 MHz CPUs).

Answers to Questions

To understand these results, I'll describe the test environment by following the questions I posted previously:

- This is not testing a cluster – this is a single head node.
- It is for sale.
- Same target as before: a Sun Storage 7410 with 128 Gbytes of DRAM, 4 sockets of quad-core AMD Opteron 2300 MHz CPU, and 2 x 2x10 GigE cards. It's not a max config since it isn't in a cluster.

- Same clients as before: 20 blades, each with 2 sockets of Intel Xeon quad-core 1600 MHz CPUs, 6 Gbytes of DRAM, and 2 x 1 Gig network ports.
- Client and server ports are connected together using 2 switches, and jumbo frames are enabled. Only 2 ports of 10 GbE are connected from the 7410, one from each of its 2 x 10 GbE cards, to load balance across the cards as well as the ports. Both ports on each client are used.
- The workload is streaming reads over files, with a 1 Mbyte I/O size. Once the client reaches the end of the file, it loops to the start. 10 processes were run on each of the clients. The files are mounted over NFSv3/TCP.
- The total working set size is 100 Gbytes, which cache in the 7410's 128 Gbytes of DRAM.
- The results are screenshots from Analytics on the 7410.
- The target Sun Storage 7410 may not have been fully utilized – see the CPU utilization graph above and comments.
- The clients aren't obviously saturated, although they are processing the workload as fast as they can with 1600 MHz CPUs.
- I've been testing throughput as part of my role as a Fishworks performance engineer.

Traps to watch out for regarding throughput

For throughput results, there are some specific additional questions to consider:

- How many clients were used?

  While hardware manufactures make 10 GbE cards, it doesn't mean that clients can drive them. A mistake to avoid is to try testing 10 GbE (and faster) with only one underpowered client, and end up benchmarking the client by mistake. Apart from CPU horsepower, if your clients only have 1 GbE NICs then you'll need at least 10 of them to test 10 GbE, connected to a dedicated switch with 10 GbE ports.

- What was the payload throughput?

  In the above screenshots I showed network throughput by device, but this isn't showing us how much data was sent to the clients – rather that's how busy the network interfaces were. The value of 1.90 Gbytes/sec includes the inbound NFS requests, not just the outbound NFS replys (which includes the data payload); it also includes the overheads of the Ethernet, IP, TCP and NFS protocol headers. The actual payload bytes moved is going to be a little less than the total throughput. How much exactly wasn't measured above.

- Did client caching occur?

  I mentioned this in my previous post, but it's worth emphasising. Clients will usually cache NFS data in the client's DRAM. This can produce a number of problems for benchmarking. In particular, if you measure throughput from the client, you may see throughput rates much higher than the client has network bandwidth, as the client is reading from its own DRAM rather than testing the target over the network (eg, measuring 3 Gbit/sec on a client with a 1 Gbit/sec NIC). In my test results above, I've avoided this issue by measuring throughput from the target 7410.

- Was the throughput peak or an average?

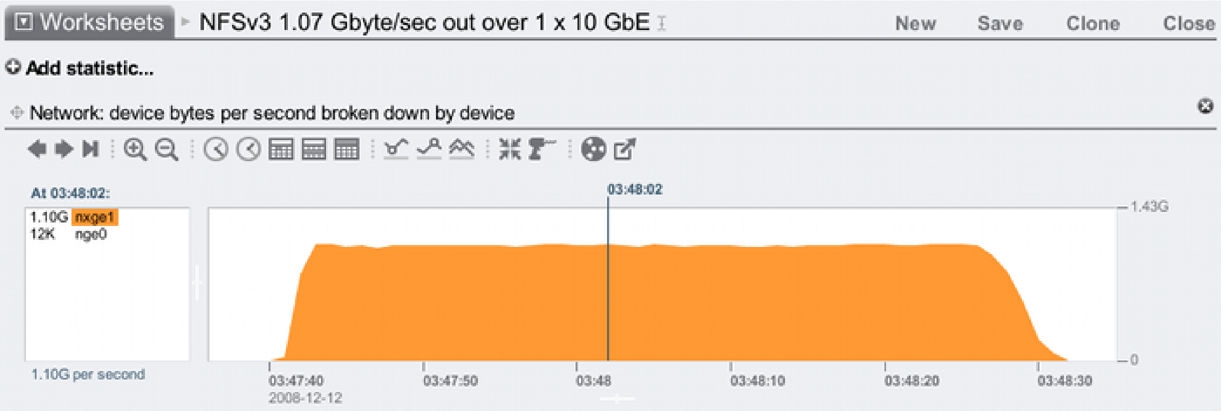
  Again, this was mentioned in my previous post and worth repeating here. The first two screenshots in this post show the difference. The average throughput over a long interval is more interesting, as this is more likely to be repeatable.

Being more realistic: 1 x 10 GbE, 10 clients

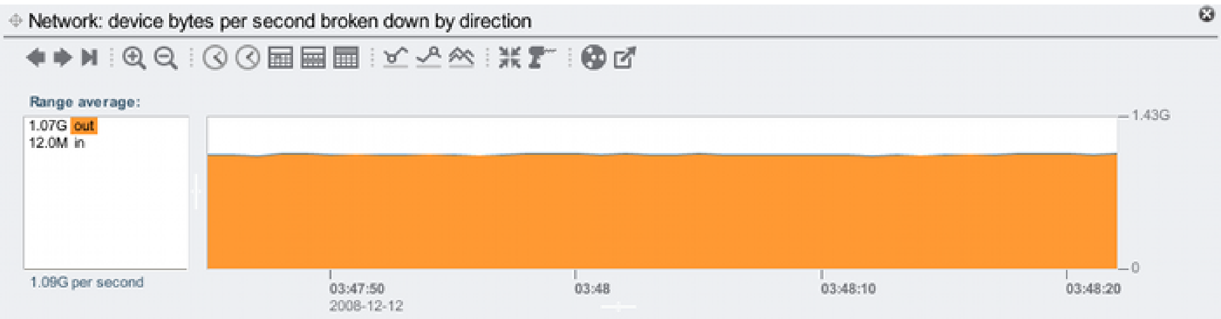
The above test showed the limits I could find, although to do so required running many processes on each of the 20 clients, using both of the 1 GbE ports on each client (40 x 1 GbE in total), balancing the load across 2 x

10 GbE cards on the target – not just 2 x 10 GbE ports – and using a 1 Mbyte I/O size. The workload the clients are applying is as extreme as they can handle.

I'll now test a lighter (and perhaps more realistic) workload: 1 x 10 GbE port on the 7410, 10 clients using one of their 1 GbE ports each, and running a single process on each client to perform the streaming reads. The client process is /usr/bin/sum (file checksum tool, which sequentially reads through files), which is run on a 5 x 1 Gbyte files for each client, so a 50 Gbyte working set in total:



This time the network traffic is on the nxge1 interface only, peaking at 1.10 Gbytes/sec for both inbound and outbound. The average outbound throughput can be shown in Analytics by zooming in a little and breaking down by direction:



That's 1.07 Gbytes/sec outbound. This includes the network headers, so the NFS payload throughput will be a little less. As a sanity check, we can see from the first screenshot x-axis that the test ran from 03:47:40 to about 03:48:30. We know that 50 Gbytes of total payload was moved over NFS (the shares were mounted before the run, so no client caching), so if this took 50 seconds, our average payload throughput would be about 1 Gbyte/sec. This fits.

10 GbE should peak at about 1.164 Gbyte/sec (converting gigabits to gibibytes) per direction, so this test reaching 1.07 Gbytes/sec outbound is a 92% utilization for the 7410's 10 GbE interface. Each of the 10 client's 1 GbE interface would be equally busy. This is a great result for such a simple test – everything is doing what it is supposed to. (While this might seem obvious, it did take much engineering work during the year to make this work so smoothly; see posts here and here for some details.)

In summary, I've shown that the Sun Storage 7410 can drive NFS requests from DRAM at 1 Gbyte/sec, and even up to about 2 Gbytes/sec with extreme load – which pushed 2 x 10 GbE interfaces to high levels of utilization. With a current max of 128 Gbytes of DRAM in the Sun Storage 7410, entirely cached working set workloads are a real possibility. While these results are great, it is always important to understand the context of such numbers to avoid the common traps, which I've discussed in this post.