# CIFS at 1 Gbyte/sec

03 Mar 2009

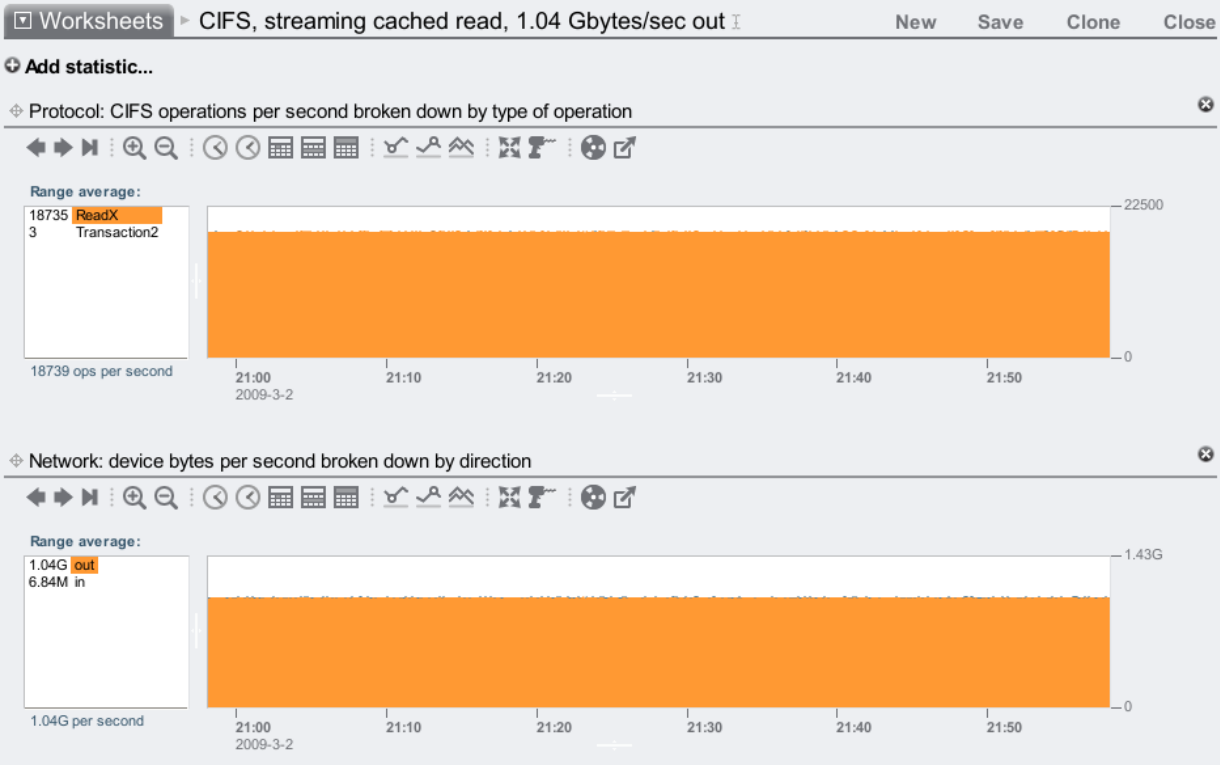
*I originally posted this at http://blogs.sun.com/brendan/entry/cifs_at_1_gbyte_sec.*

I've recently been testing the limits of NFS performance on the Sun Storage 7410. Here I'll test the CIFS (SMB) protocol: the file sharing protocol commonly used by Microsoft Windows, which can be served by the Sun Storage 7000 series products. I'll push the 7410 to the limits I can find, and show screenshots of the results. I'm using 20 clients to test a 7410 which has 6 JBODs and 4 x 10 GbE ports, described in more detail later on.

## CIFS streaming read from DRAM

Since the 7410 has 128 Gbytes of DRAM, most of which is available as the filesystem cache, it is possible that some workloads can be served entirely or almost entirely from DRAM cache, which I've tested before for NFS. Understanding how fast CIFS can serve this data from DRAM is interesting, so to search for a limit I've run the following workload: 100 Gbytes of files (working set), 4 threads per client, each doing streaming reads with a 1 Mbyte I/O size, and looping through their files.

I don't have to worry about client caching affecting the observed result – as is the case with other benchmarks – since I'm not measuring the throughput on the client. I'm measuring the actual throughput on the 7410, using Analytics:




I've zoomed out to show the average over 60 minutes – which was 1.04 Gbytes/sec outbound!

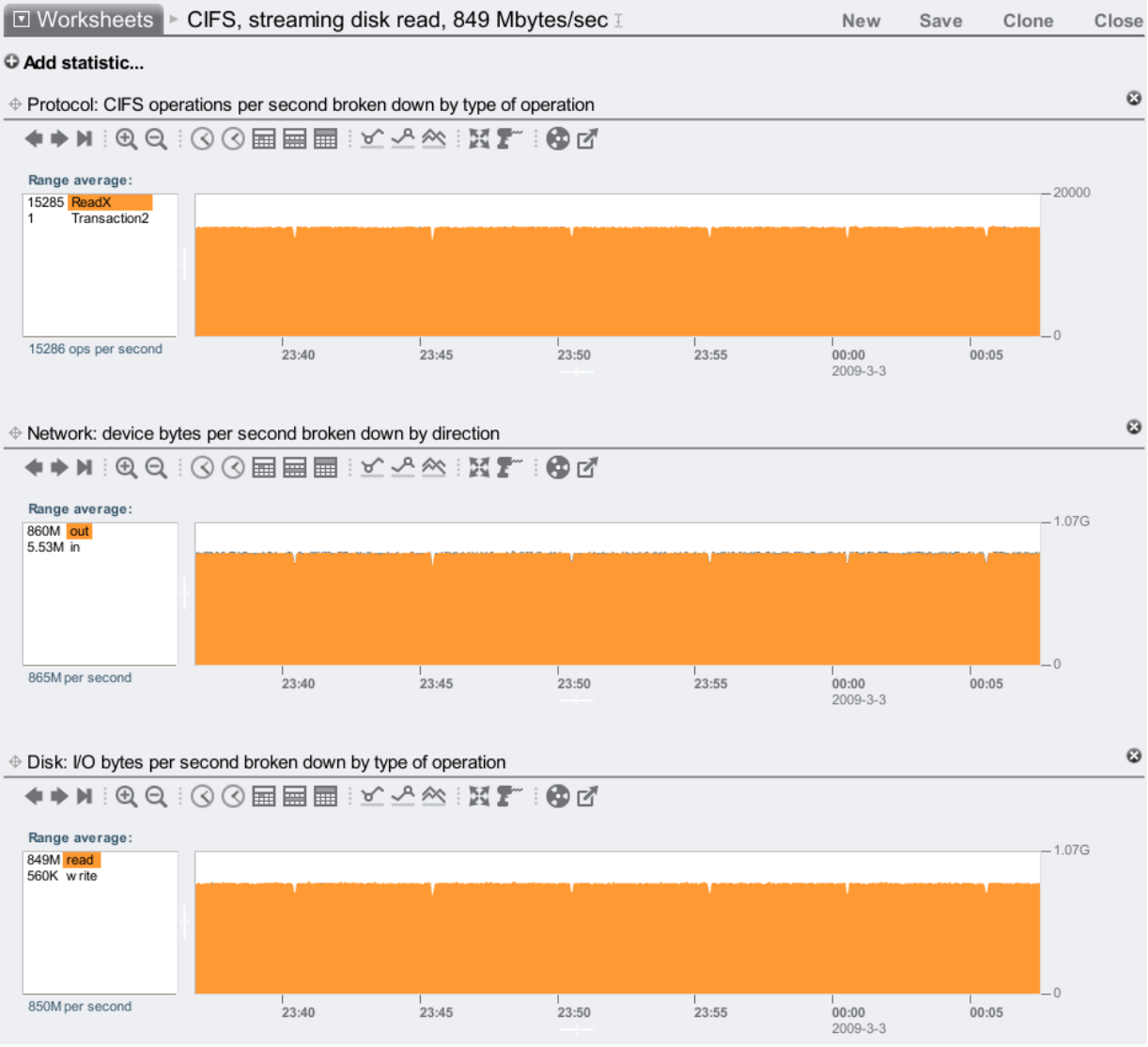
This result was measured as outbound network throughput, so it includes the overhead of Ethernet, IP, TCP and CIFS headers. Since jumbo frames were used, this overhead is going to be about 1%. So the actual data payload moved over CIFS will be closer to 1.03 Gbytes/sec.

As a screenshot it's clear that I'm not showing a peak value – rather a sustained average over a long interval – to show how well the 7410 can serve CIFS at this throughput.

## CIFS streaming read from disk

While 128 Gbytes of DRAM can cache large working sets, it's still interesting to see what happens when we fall out of that, as I've previously shown for NFS. The 7410 I'm testing has 6 JBODs (from a max of 12), which I've configured with mirroring for max performance. To test out disk throughput, my workload is: 2 Tbytes of files (working set), 2 threads per client, each performing streaming reads with a 1 Mbyte I/O size, and looping through their files.

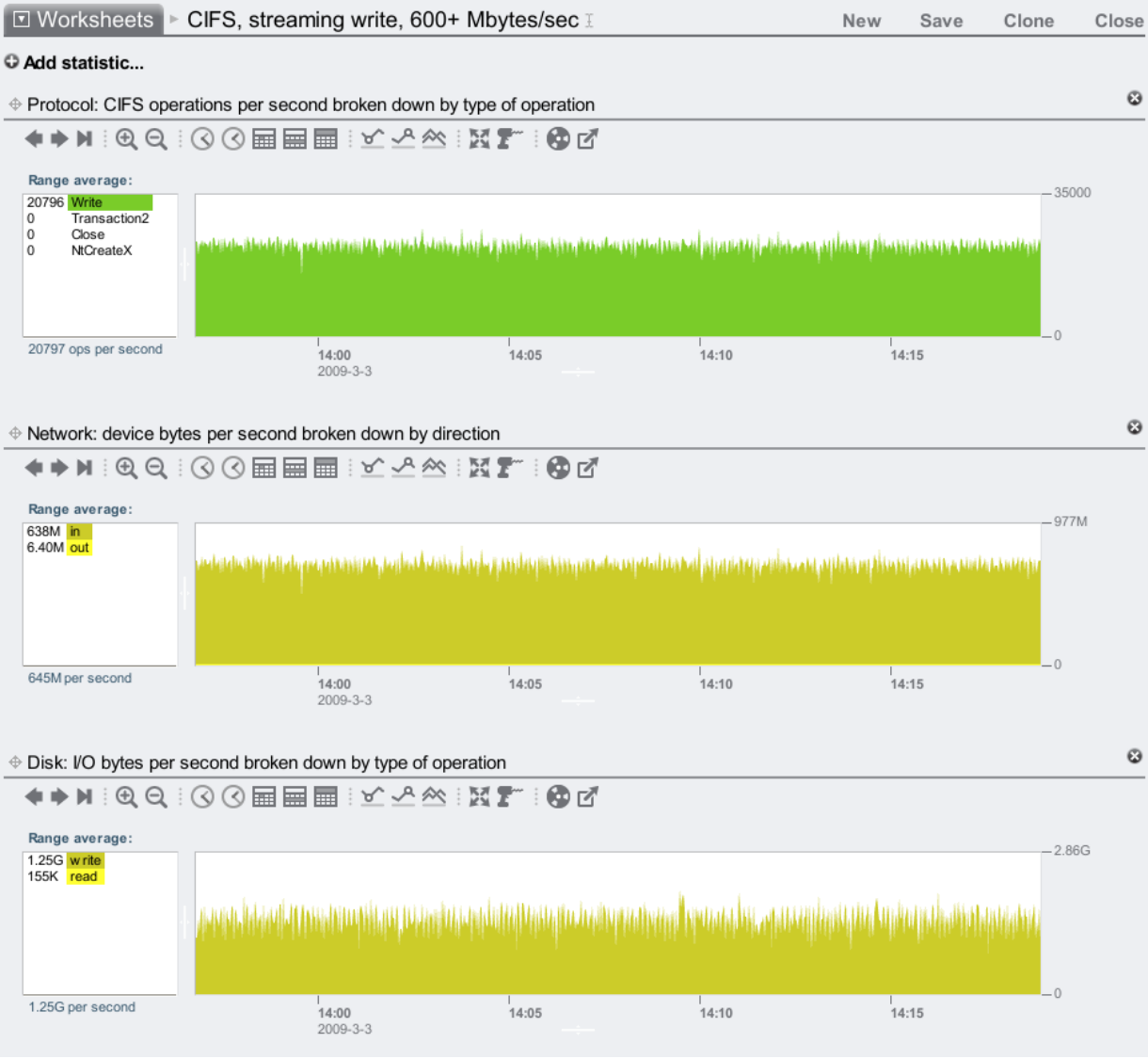
As before, I'm taking a screenshot from Analytics to show what the 7410 is really doing:




Here I've shown read throughput from disk at 849 Mbytes/sec, and network outbound at 860 Mbytes/sec (includes the headers). 849 Mbytes/sec from disk (which will be close to our data payload) is very solid performance.

## CIFS streaming write to disk

Writes are a different code path to reads, and need to be tested separately. The workload I've used to test write throughput is: Writing 1+ Tbytes of files, 4 threads per client, each performing streaming writes with a 1 Mbyte I/O size. The result:

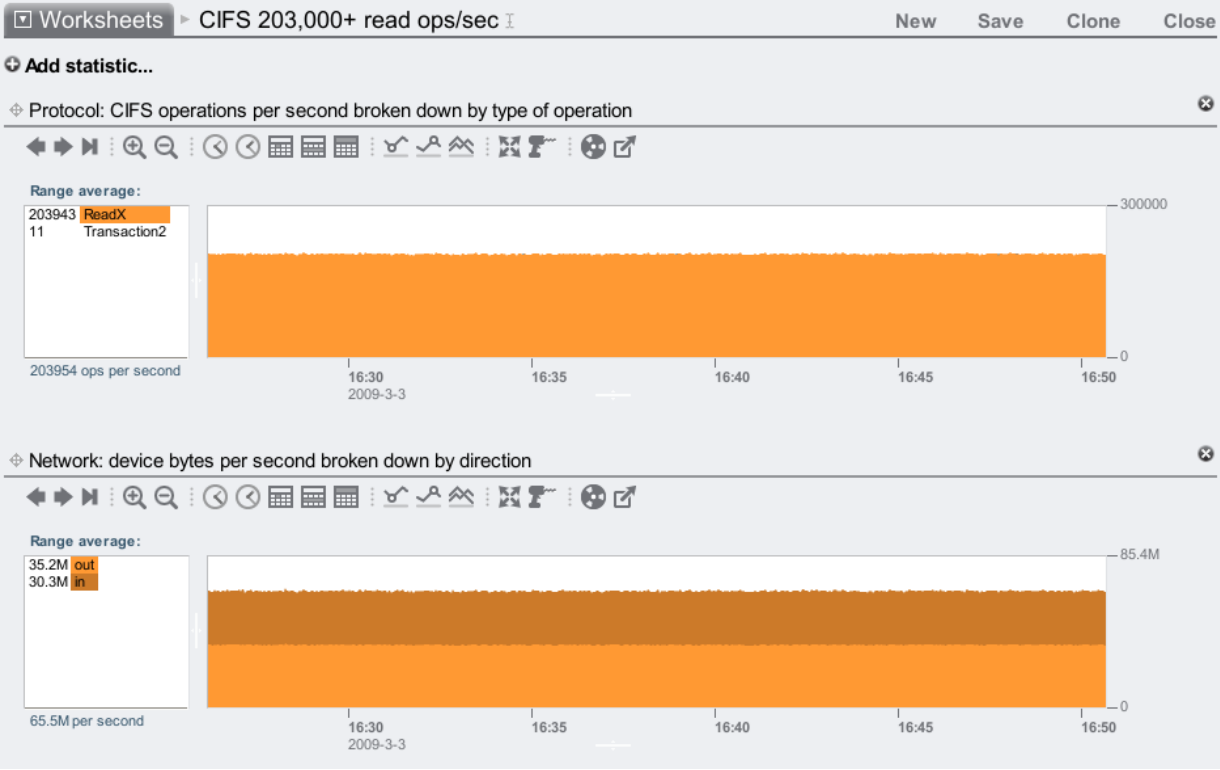



The network inbound throughput was 638 Mbytes/sec, which includes protocol overheads. Our data payload rate will be a little less than that due to the CIFS protocol overheads, but should still be at least 620 Mbytes/sec, which is very good indeed (even beat my NFSv3 write throughput result!)

Note that the disk I/O bytes were at 1.25 Gbytes/sec write: the 7410 is using 6 JBODs configured with software mirroring, so the back end write throughput is doubled. If I picked other storage profiles like RAID-Z2, the back end thoughput would be less (as I showed in the NFS post).

## CIFS read IOPS from DRAM

Apart from throughput, it's also interesting to test the limits of IOPS, in particular read ops/sec. To do this, I'll use a workload which is: 100 Gbytes of files (working set) which caches in DRAM, 20 threads per client, each performing reads with a 1 byte I/O size. The results:

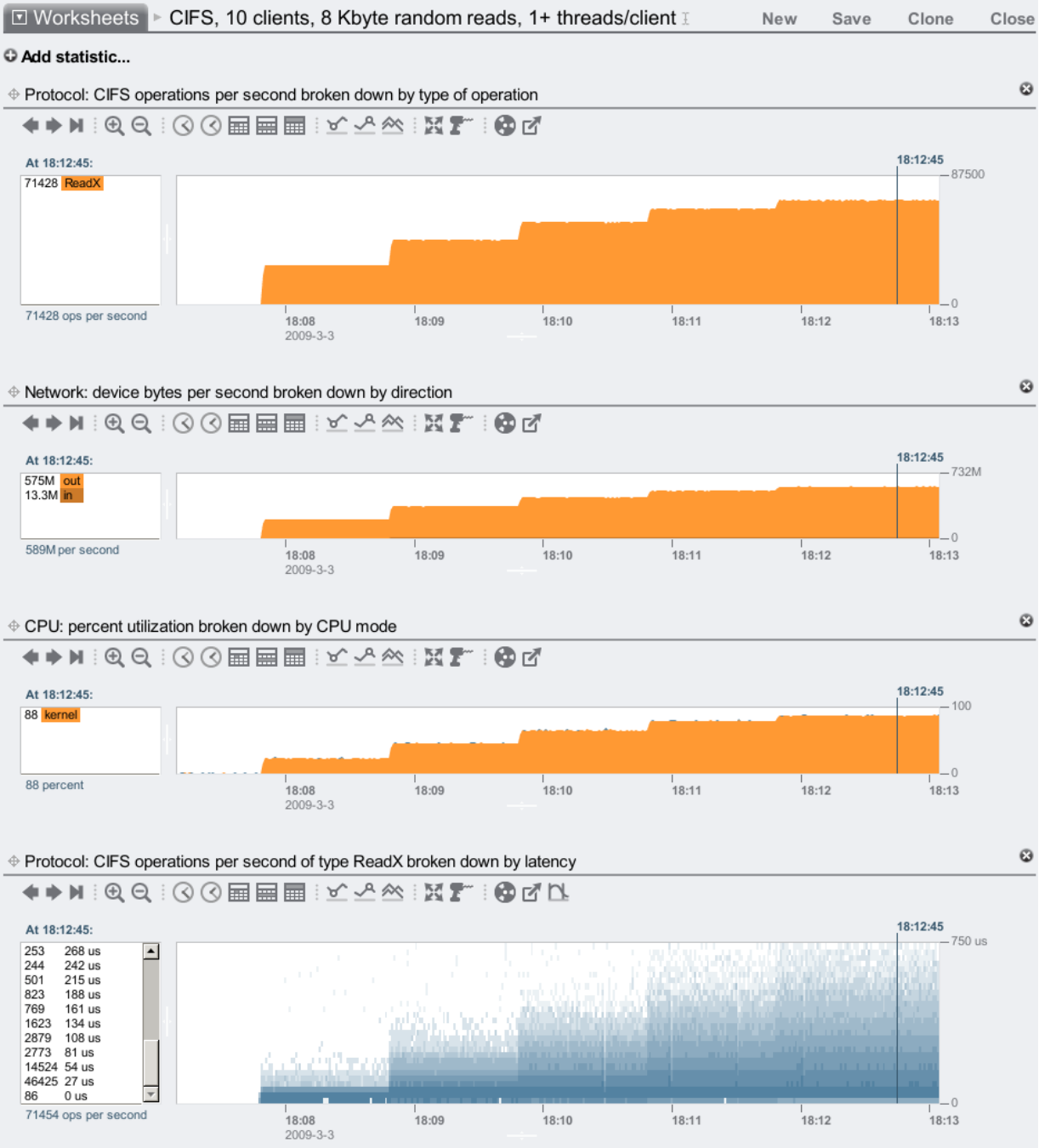



203,000+ is awesome; while a realistic workload is unlikely to call 1 byte I/Os, still, it's interesting to see what the 7410 can do (I'll test 8 Kbyte I/O next.)

Note the network in and out bytes is about the same. The 1 byte of data payload doesn't make much difference beyond the network headers.

## Modest read IOPS

While the limits I can reach on the 7410 are great for heavy workloads, these don't demonstrate how well the 7410 responds under modest conditions. Here I'll test a lighter read ops/sec workload by cutting it to: 10 clients, 1 x 1 GbE port per client, 1 x 10 GbE port on the 7410, 100 Gbytes of files (working set), and *8 Kbyte* random I/O. I'll step up the threads per client every minute (by 1), starting at 1 thread/client (so 10 in total to begin with):




We reached 71,428 read ops/sec – a good result for 8 Kbyte random I/O from cache and only 10 clients.

It's more difficult to generate client load (involves context switching to userland) than to serve it (kernel only), so you generally need more CPU grunt on the clients than on the target. At one thread per client on 10 clients, the clients are using 10 x 1600 MHz cores to test a 7410 with 16 x 2300 MHz cores, so the clients themselves will limit the throughput achieved. Even at 5 threads per client there is still headroom (%CPU) on this 7410.

The bottom graph is a heat map of CIFS read latency, as measured on the 7410 (from when the I/O request was received, to when the response was sent). As load increases, so does I/O latency, but they are still mostly less than 100 us (fast!). This may be the most interesting from all the results – as this modest load is increased, the latency remains low while the 7410 scales to meet the workload.

## Configuration

As the **filer** I was using a single Sun Storage 7410, with the following config:

- 128 Gbytes DRAM
- 6 JBODs, each with 24 x 1 Tbyte disks, configured with mirroring
- 4 sockets of quad-core AMD Opteron 2300 MHz CPUs
- 2 x 2x10 GbE cards (4 x 10 GbE ports total), jumbo frames
- 2 x HBA cards
- noatime on shares, and database size left at 128 Kbytes

It's not a max config system: the 7410 can currently scale to 12 JBODs, 3 x HBA cards, and have flash based SSD as read cache and intent log – which I'm not using for these tests. The CPU and DRAM size is the current max: 4 sockets of quad-core driving 128 Gbytes of DRAM is a heavyweight for workloads that cache well, as shown earlier.

The **clients** were 20 blades, each:

- 2 sockets of Intel Xeon quad-core 1600 MHz CPUs
- 6 Gbytes of DRAM
- 2 x 1 GbE network ports
- Running Solaris, and mounting CIFS using the smbfs driver

These are great, apart from the CPU clock speed – which at 1600 MHz is a little low.

The **network** consists of multiple 10 GbE switches to connect the client 1 GbE ports to the filer 10 GbE ports.

## Conclusion

A single head node 7410 has very solid CIFS performance, as shown in the screenshots. I should note that I've shown what the 7410 *can do* given the clients I have, but it may perform even faster given faster clients for testing. What I have been able to show is 1 Gbyte/sec of CIFS read throughput from cache, and up to 200,000 read ops/sec. Tremendous performance.

Before the Sun Storage 7000 products were released, there was intensive performance work on this by the CIFS team and PAE (as Roch describes), and they have delivered great performance for CIFS, and continue to improve it further. I've updated the summary page with these CIFS results.