

Identifying Gaps in Grid Middleware on Fast Networks with the Advanced Networking Initiative

Dave Dykstra, Gabriele Garzoglio¹, Hyunwoo Kim, Parag Mhashilkar

Scientific Computing Division, Fermi National Accelerator Laboratory

E-mail: {dwd, garzoglio, hyunwoo, parag}@fnal.gov

Abstract. As of 2012, a number of US Department of Energy (DOE) National Laboratories have access to a 100 Gb/s wide-area network backbone. The ESnet Advanced Networking Initiative (ANI) project is intended to develop a prototype network, based on emerging 100 Gb/s Ethernet technology. The ANI network will support DOE's science research programs. A 100 Gb/s network test bed is a key component of the ANI project. The test bed offers the opportunity for early evaluation of 100Gb/s network infrastructure for supporting the high impact data movement typical of science collaborations and experiments. In order to make effective use of this advanced infrastructure, the applications and middleware currently used by the distributed computing systems of large-scale science need to be adapted and tested within the new environment, with gaps in functionality identified and corrected. As a user of the ANI test bed, Fermilab aims to study the issues related to end-to-end integration and use of 100 Gb/s networks for the event simulation and analysis applications of physics experiments. In this paper we discuss our findings from evaluating existing HEP Physics middleware and application components, including GridFTP, Globus Online, etc. in the high-speed environment. These will include possible recommendations to the system administrators, application and middleware developers on changes that would make production use of the 100 Gb/s networks, including data storage, caching and wide area access.

1. Introduction

In 2011 the Energy Sciences Network (ESNet)[1] of the Department of Energy[2] has deployed a 100 Gb/s wide area network backbone and made it available to several National Laboratories. At the same time, ESNet has provided a network test bed for the Advanced Networking Initiative (ANI)[3], to validate the upcoming infrastructure. Fermilab will connect to the backbone in summer 2012, following a decades-long tradition of providing for our stakeholders sustained, high speed, large and wide-scale data distribution and data access. To give a sense of scale, the laboratory hosts 40 PB of data on tape, now mostly coming from offsite for the Compact Muon Solenoid (CMS) experiment, sustaining traffic peaks of 30 Gb/s on the WAN, and supporting regularly connectivity of 140 Gb/s of LAN traffic from archive to local processing farms.

In preparation for Fermilab connecting to the 100 Gb/s backbone, we have been running the High Throughput Data Program (HTDP) for validating the deep stack of software layers and services that make

¹ To whom any correspondence should be addressed.

up the end-to-end analysis systems of our stakeholders. Not only are these the High Energy Physics (HEP) experiments associated with the laboratory, but these also include a variety of large and small multi-disciplinary communities involved with Fermilab through Grid initiatives, such as the Extreme Science and Engineering Discovery Environment (XSEDE)[4] and the Open Science Grid (OSG)[5]. The goal of the HTDP program is to ensure that the stakeholders' analysis systems are functional and effective at the 100 Gb/s scale, by tuning the configuration to ensure full throughput in and across each layer/service, measuring the efficiency of the end-to-end solutions, and monitoring, identifying, and mitigating error conditions. The program includes technological investigations to identify gaps in the middleware components integrated with the analysis systems and the development of system prototypes to adapt to the high-speed network.

In 2011, the program has utilized the 30 Gb/s prototype ANI test bed at the Long Island Metropolitan Area Network (LIMAN). It has then demonstrated high throughput data movement for CMS at SuperComputing 2011, sustaining rates of 70 Gb/s for one hour. Section 2 discusses the results from these early studies. Section 3 presents the current setup of the 100 Gb/s ANI test bed. We discuss our evaluation of GridFTP[6] and Globus Online[7] on the ANI in section 4 and of xrootd[8], a popular data management framework in HEP, in section 5. We present our plans for future work in section 6, before concluding in section 7.

2. Previous ESnet ANI test beds

In 2011, the ANI team has deployed two fast-network test environments, in preparation for the full 100 Gb/s test bed. The first was the LIMAN network, a 30 Gb/s test bed in the Long Island metropolitan area between BNL and New York (section 2.1). The second was showcased at SuperComputing 2011, whereby a full 100 Gb/s network was made available for several communities to show their ability to saturate the network (sec 2.2). In the sections below we discuss our experience in these two environments.

2.1. The LIMAN 30 Gb/s Test Bed

On the LIMAN ANI test bed, we tested GridFTP and Globus Online. GridFTP is a well-established data transfer middleware, widely deployed in the end-to-end analysis environments of the Fermilab stakeholders. Globus Online is an emerging reliable data transfer service, built on top of GridFTP servers. For both technologies, we used this test bed to develop the testing techniques for the 100 Gb/s ANI test bed.

With GridFTP, our goal was to tune the transfer parameters to saturate network bandwidth. In general, on high-latency networks such as the 100 Gb/s ANI test bed (sec 3), achieving full bandwidth may be challenging because of the overhead per file introduced by the control channel of the protocol, in addition to the security overhead. This effect becomes particularly relevant for small size files; therefore, we gathered measurements with 3 datasets of large, medium, and small file sizes. Although the LIMAN test bed was a relatively low-latency network (section 2.1.1), we could not achieve maximum bandwidth for small files (section 2.1.3). Since we observe the same behavior in the 100 Gb/s test bed, we plan to look at operating system tuning parameters, such as process scheduling algorithm, file system caching, etc. to improve the performance.

With Globus Online (GO), we also attempted to saturate the available network. This was generally more challenging because the latencies on the GO control channel were larger than for the previous GridFTP tests. In fact, the ANI test beds are only accessible through a Virtual Private Network (VPN). In order for GO to control the GridFTP servers inside the ANI test bed, we implemented a port forwarding mechanism through a VPN gateway machine (section 2.1.1). For a fair comparison with Globus Online, we also made measurements in the same conditions with a GridFTP client (client outside the network, control handled with port-forwarding) (section 2.1.3).

2.1.1. The LIMAN hardware and network

We ran a GridFTP server (on node *bnl-1*) at the Brookhaven National Laboratory (BNL) and clients at two nodes. One client node (*bnl-2*) was also in BNL; the other (*newy-1*) elsewhere in the Long Island Metropolitan Area. The round-trip-time (RTT) between *bnl-1* and *bnl-2* is 0.1 ms, and between *bnl-1* and *newy-1* the RTT is 2 ms. Figure 1 shows the network diagram for the LIMAN test bed. *bnl-1* had four 10 Gb/s interfaces, two connecting to *bnl-2* and two connecting to *newy-1*. The four interfaces were not completely independent, however, and only three at a time could be used in order to get the best throughput. All three computers had relatively fast disk arrays, benchmarked at 14 Gb/s write speed, but they were still not adequate to keep up with 30 Gb/s. For this reason, transfers were directed to memory (/dev/null). The files were small enough to be held in the memory cache of the file system, so disk read speed did not affect these tests.

The test bed hardware was available only via a Virtual Private Network (VPN). To make the test bed servers accessible to Globus Online, a machine at Fermilab (“VPN gateway”) ran the VPN software, accepted GridFTP control port connections from the Internet, and forwarded them to the network control interfaces (1 Gb/s) of the three server machines, using xinetd port forwarding. In turn, again using xinetd, the three server machines forwarded those connections to their own GridFTP control ports, bound to the 10 Gb/s interfaces. The port forwarding mechanism is described in detail (for the 100 Gb/s testbed) in section 4.1. The RTT between Fermilab and BNL is 36 ms.

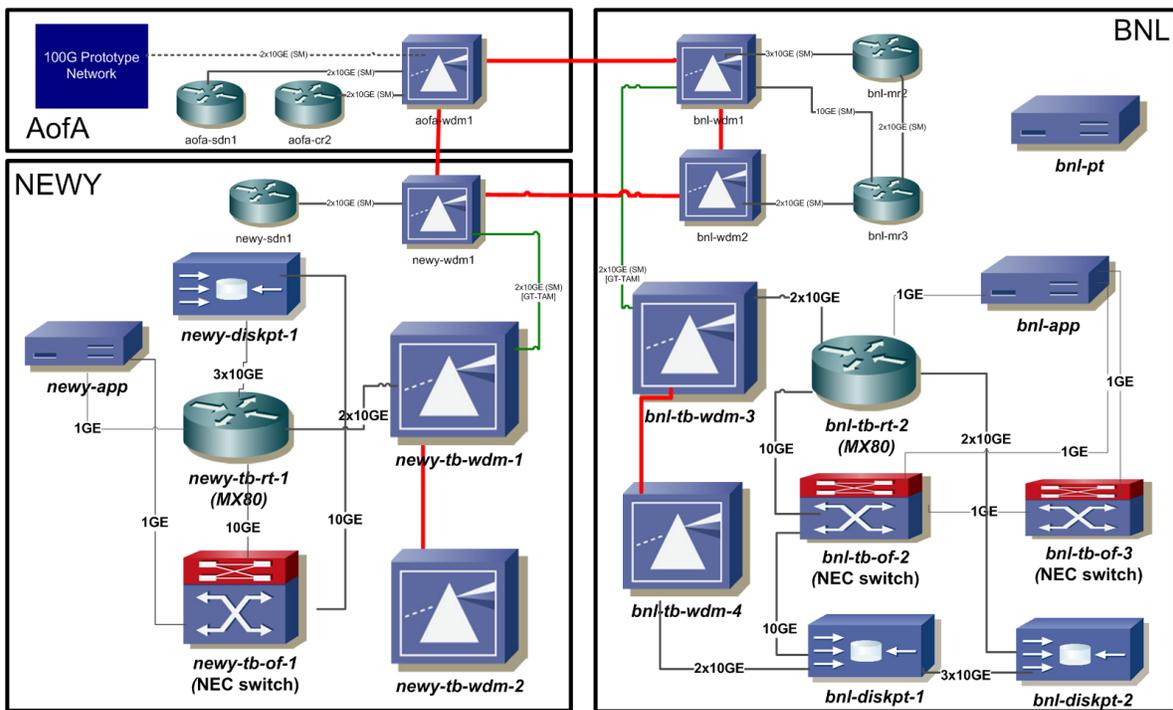


Figure 1: Architectural diagram of the Advanced Network Initiative test bed at the Long Island Metropolitan Area Network.

2.1.2. Tests details

The test dataset of 21 files with sizes from 8 KB to 8 GB, incrementing in power of 2, was divided into three datasets: large, medium, and small. The large dataset consists of 3 files (total size: almost 14 GB), from 2 GB to 8 GB; the medium dataset (total size: almost 2 GB) consists of 8 files, from 8 MB to 1 GB;

the small dataset (total size: almost 8 MB) consists of 10 files, from 8 KB to 4 MB. Files in each dataset were repeated as often as needed in order to have a total of 100 GB transferred: 30 GB in 7 file transfers for the large file dataset, 40 GB in 180 file transfers for the medium file dataset, and 30 GB in 42240 file transfers for the small file dataset.

As shown in figure 2, all the GridFTP transfers are initiated on bnl-1 and data is sent from bnl-2 and newy-1 to bnl-1 via the following three GridFTP control methods

- Local: 3 parallel globus-url-copy commands are initiated on bnl-1 as third party transfers (server-server mode);
- Fermilab-controlled GridFTP transfer via port-forwarding: 3 parallel globus-url-copy commands are initiated on a machine at Fermilab, connecting through the VPN gateway;
- Globus Online-controlled GridFTP: 3 transfers are initiated using gsissh to cli.globus.org one after another. Timing is taken from the email reports, starting from the start of the first transfer to the end of the last one.

2.1.3. Observations

Figure 3 shows a histogram that compares data transfer throughput measurements for local and remote use of GridFTP and Globus Online for various file sizes.

- For large and medium size files, comparing the results for locally and remotely-controlled GridFTP, we observe that the latter is affected in minimal part by the control channel overhead. Globus Online, however, is affected to a greater degree because of the higher latency.
- Small files suffer from some overhead for locally initiated transfers, but have a high overhead for the high-latency control channels. This affects the resulting bandwidth.
- Globus Online auto-tuning appears to be more effective on medium files than on large ones.

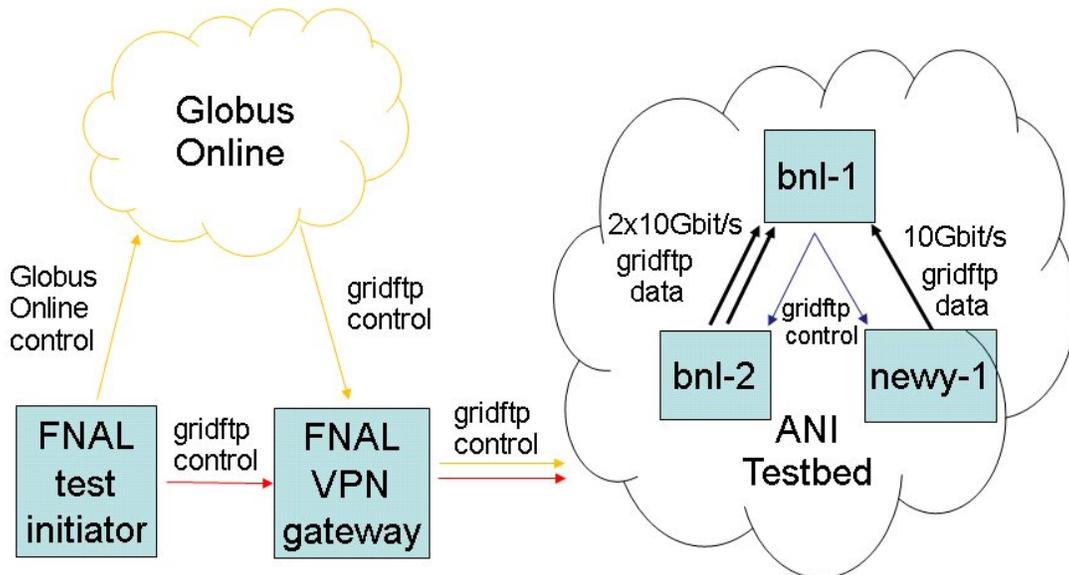


Figure 2: LIMAN test bed diagram.

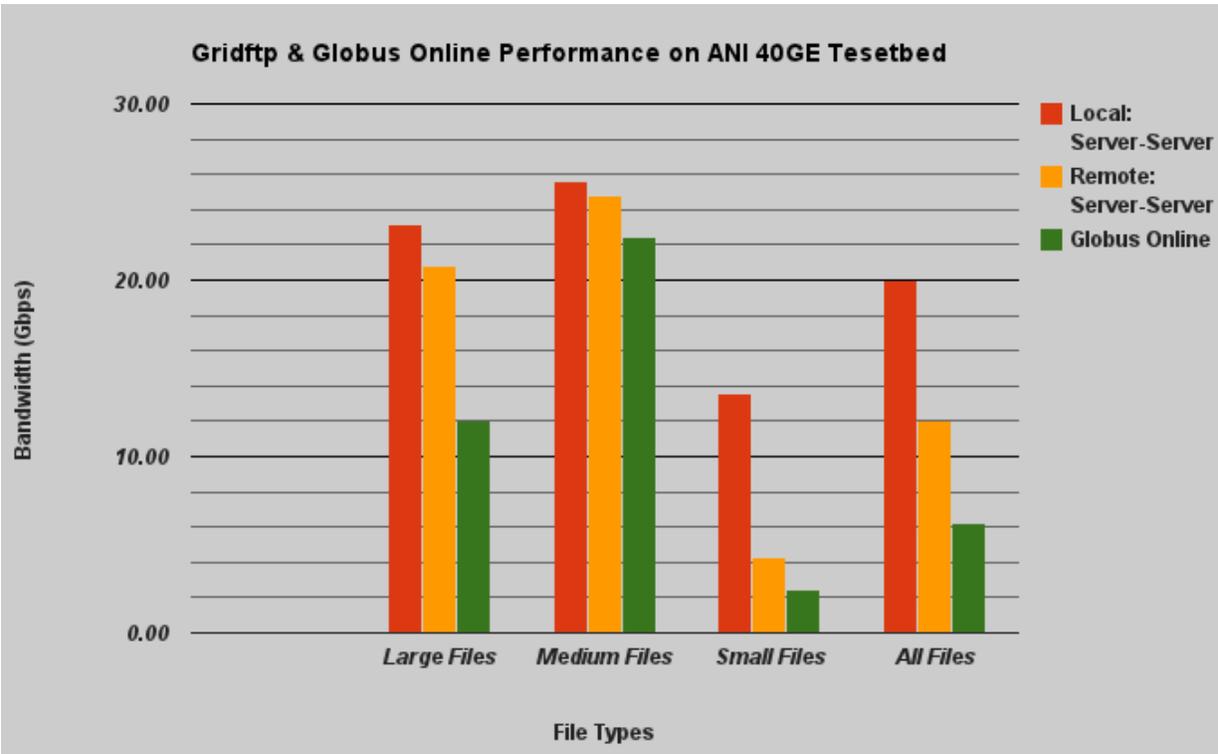


Figure 3: LIMAN test results.

2.2. SuperComputing 2011 with a shared 100 Gb/s network

ANI provided a shared 100 Gb/s network for its participant at the Super Computing conference in 2011 (SC11)[9]. The Grid & Cloud Computing Department of Fermilab, in collaboration with UCSD, demonstrated the use of 100 Gb/s networks to move 30 TB of CMS data in one hour with GridFTP.

2.2.1. SC11 demo test bed hardware and network

The demo test bed consisted of 15 machines (8 cores, Intel Xeon CPU 2.67 GHz and 48 GB RAM) at National Energy Research Science Computing Center (NERSC) and 26 machines (12 cores, same processor and 48 GB RAM) at Argonne National Laboratory (ANL). All machines had 10 Gb/s network cards for the fast network and regular 1 Gb/s for the control. NERSC and ANL were connected through a 100 Gb/s network organized in two subnets, so that some machines at NERSC and ANL were in a subnet, some in another. The ANL machines could be accessed from a UCSD virtual machine via a condor batch system. NERSC provided interactive access. ESnet and NERSC both have Alcatel-Lucent 7750 routers deployed on site to route 100 Gb/s traffic. At ANL, a Brocade router fed into 2 Juniper EX4500 switches.

2.2.2. Tests and observations

Machines at NERSC ran GridFTP servers. Both data and control channels were forced to bind to the fast network. Each server received connections from one or two machines at ANL from its own subnet. GridFTP clients were running on ANL machines. Since we found that large number of clients per core and large TCP window size result in better performance, we had each machine run 48 globus-url-copy (4 per core) with 2 parallel streams and a FTP data channel buffer of 2 MB (globus-url-copy -p2 -tcp-bs 2097152).

10 CMS files of about 2 GB were transferred repeatedly from memory to memory as the access to storage systems was identified as a bottleneck to this network-focused demonstration. A total of 30 TB of data was transmitted during a one hour demo session.

Table 1 shows the test configurations and results for test (T) and demo (D) sessions.

	GUC/ Core	GUC Streams	GUC TCP Window size	Files/ GUC	MAX BW(Gb/s)	Sustain BW(Gb/s)
T1	-	-	-	-	-	-
D1	1	2	Default	60	65	50
T2	1	2	2 MB	1	65	52
D2	1	2	2 MB	1	65	52
T3	4	2	2 MB	1	73	70
D3	4	2	2MB	1	75	70

Table 1: The throughput measurements using SC2011 demo test bed.

3. The ANI 100 Gb/s Test Bed

3.1. Hardware and network

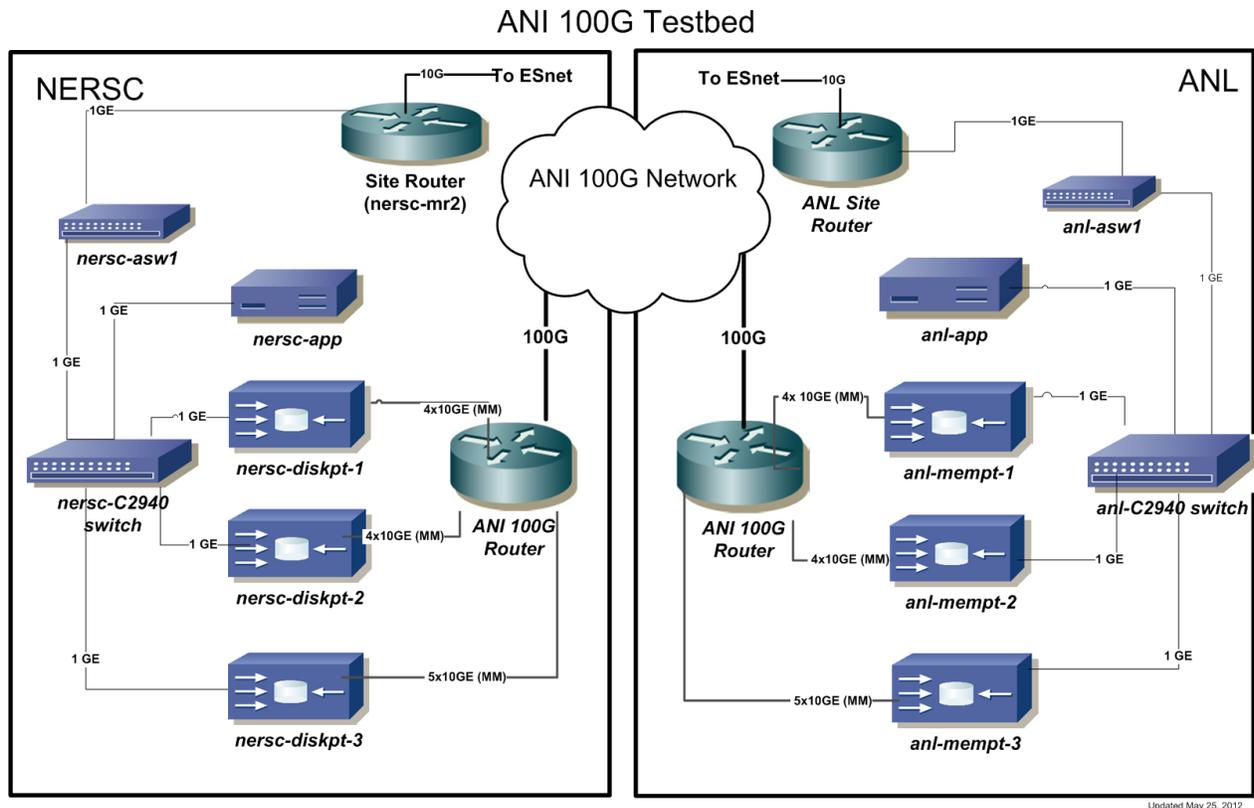
The ANI 100 Gb/s test bed is composed of three different types of hosts: physical performance I/O servers, memory performance servers, and virtual machines. Virtual machines were not involved in testing. Figure 4 shows the network configuration of the test bed.

3.1.1. Physical performance I/O servers

Three physical performance I/O servers at NERSC, nersc-diskpt-[1-3], are each powered by 2 Intel Xeon Nehalem E5650 2.67GHz processors with 48GB (12x4GB) DDR3-1066MHz ECC/REG RAM installed on a Supermicro X8DAH+ F Motherboard. I/O servers are specifically designed and tuned to handle over 10 Gb/s multi-stream sequential read and write flows to/from RAID disk. These hosts are connected to the 100 Gb/s data plane with several 10 Gb/s ports. Each host has a combination of single (10G-PCI2-8S2-2S+E) and dual port (10G-PCI2-8C2-2S+E) 10 Gb/s PCI interface cards from Myricom. RAID is configured with 16 Seagate 500GB SAS HDD 7200 rpm ST3500620SS hard drives. Each I/O performance host runs CentOS 5.7.

3.1.2. Memory performance servers

Three memory performance hosts at ANL, anl-memtp-[1-3], are designed to run applications that are CPU intensive working with dataset in memory. They are connected to the 100 Gb/s data plane with several 10 Gb/s ports. The default operating system on these hosts is CentOS 6.0. The memory hosts are each powered by 2 AMD 6140 8 core 2.6 GHz processors with 64 GB (8x8GB) DDR3-1333MHz ECC/REG RAM installed on a Supermicro H8D Gi-F motherboard. Like the I/O performance hosts, each memory performance host has a combination of single and dual port 10 Gb/s PCI interface cards from Myricom.



Updated May 25, 2012

Figure 4: ANI 100 Gb/s test bed. Courtesy of ESnet/ANI.

3.2. Connecting to the test bed

10 Gb/s interfaces on the performance hosts at NERSC and ANL are part of a 100 Gb/s private network not accessible from the general Internet, similarly to the LIMAN test bed. Each of these hosts also has one 1 Gb/s interface connected to the respective site routers through a switch. Traffic to the 1 Gb/s interfaces is fully routed within the ANI network and is used for port-forwarding. Inbound access to the hosts is only available through a VPN. A machine at Fermilab, *anidev.fnal.gov*, provides access to the 100 Gb/s test bed through the VPN (section 4.1). To overcome these limitations for our tests of GridFTP and Globus Online (section 4), we used the same technique for port forwarding discussed above (sec 2.1.1). Figure 5 shows the round trip time (RTT) between the hosts involved.

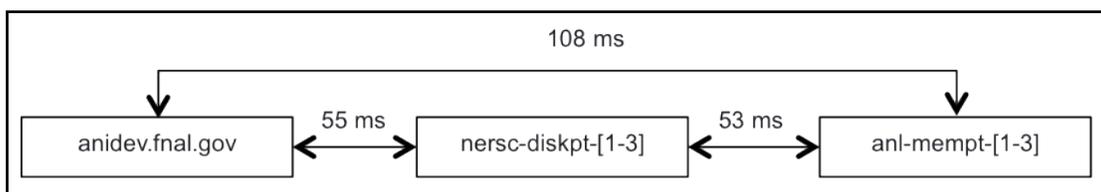


Figure 5: Test bed connections with RTT.

Each end of the 100 Gb/s connection between NERSC and ANL is connected to 3 nodes. Each node has 4 10 Gb/s NICs. Figure 6 illustrates the connection between the two sites and subnet configuration of the network.

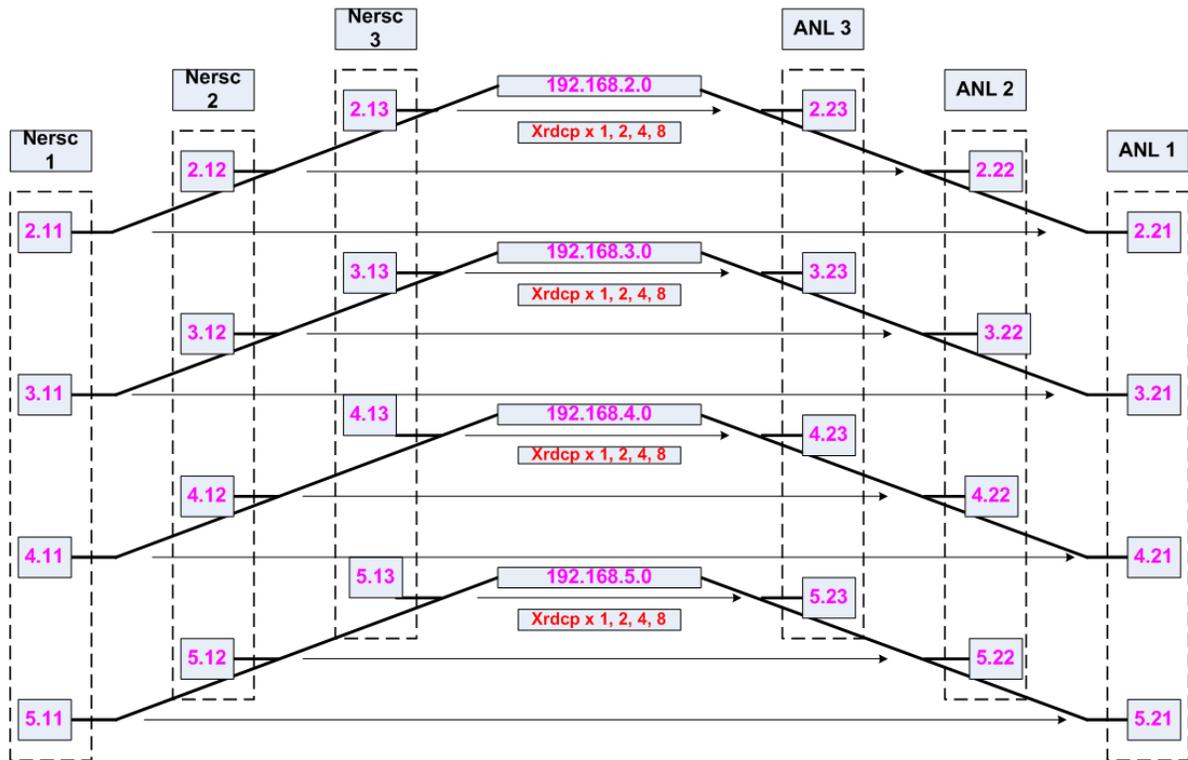


Figure 6: Illustration of connections between NERSC and ANL nodes.

3.3. Basic Network Capacity Test with nuttcp

To validate the capacity of the network, we use nuttcp, a basic network testing tool used also by the ESNet ANI team. We want to ascertain the following:

- Does each of the 12 10 Gb/s NICs provides a 10 Gb/s transfer rate?
- Do the 4 NICs in one node provide a 40 Gb/s transfer rate?
- Do the three nodes in one site provide an aggregate 100 Gb/s transfer rate?

Our results show that we can achieve the nominal maximum rate, in accordance with the results from the ESNet ANI team. In detail:

- Test of one (10 Gb/s) NIC to one NIC: 9.89 Gb/s
- Test from one node (4 x 10 Gb/s NICs) to one node: 39 Gb/s
- Aggregate throughput using 10 TCP streams: 99 Gb/s

We can use these measurements as a reference to see how much additional overhead GridFTP, Globus Online, and xrootd introduce.

4. GridFTP and Globus Online Tests on 100 Gb/s

Following the methodologies developed for the LIMAN test bed (section 2.1), we conduct tests of GridFTP and Globus Online to saturate the 100 Gb/s network with different file sizes. With respect to nuttcp (section 3.3), the GridFTP protocol introduces additional overhead due to the increased security

and reliability. We tune the parameters available in the GridFTP protocol to mitigate the effects of these overhead. These are particularly challenging for small files (section 4.3), considering the higher latencies with respect to the LIMAN test bed (section 3.2).

For Globus Online, the latencies for the control channel are higher than for GridFTP. In fact, as for the LIMAN test bed, the control channel was forwarded through a VPN gateway to the appropriate GridFTP servers inside the ANI network (section 4.1). Globus Online auto tuning and our manual tuning were not sufficient to mitigate the effects of high latency in achieving high bandwidth (section 4.3).

4.1. Establishing connectivity for the control channel

Each of the six performance hosts runs four GridFTP servers (one per NIC) under xinetd. Each server listens to the default control port (2811) on each 10 Gb/s interface.

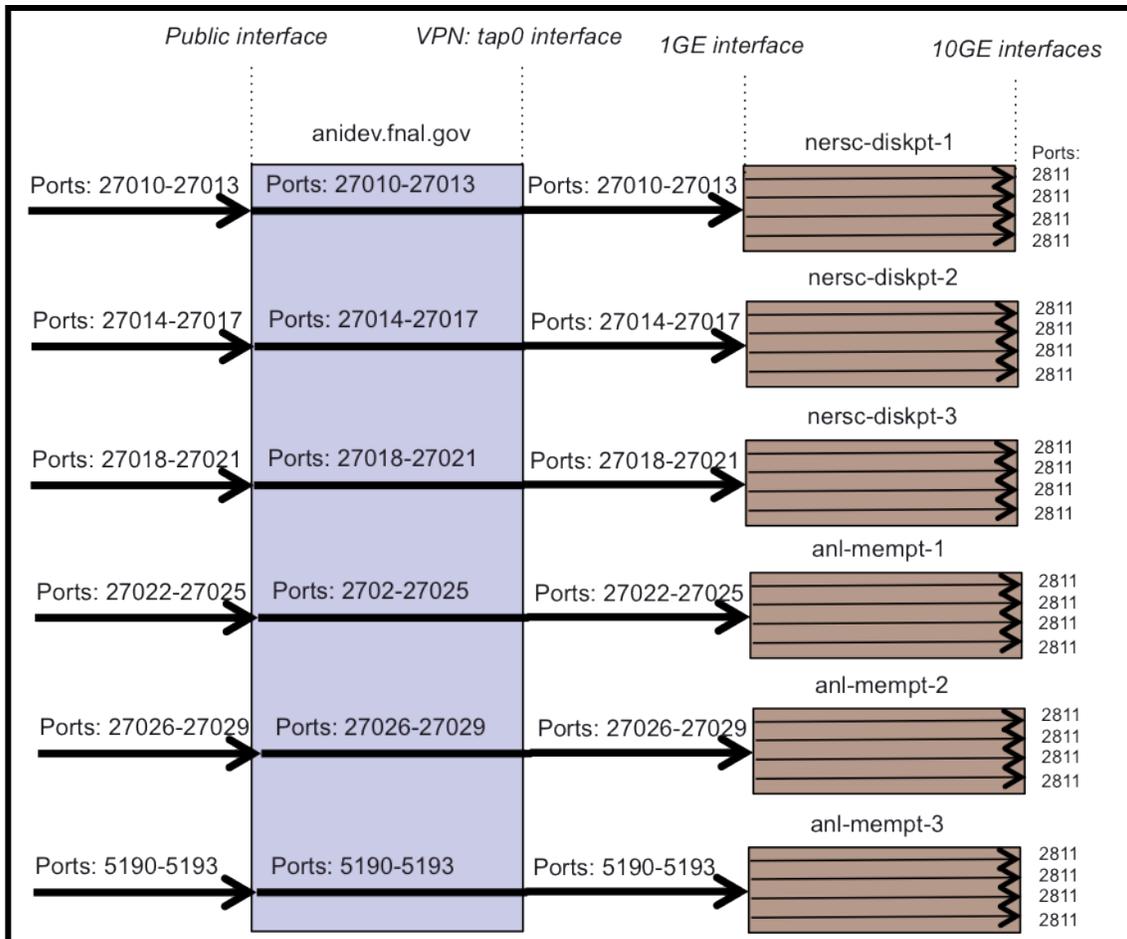


Figure 7: Port-forwarding scheme used for GridFTP and Globus Online tests.

One machine on the Fermilab network, *anidev.fnal.gov*, provides inbound network access to these GridFTP servers. This machine offers two network interfaces: one for the general Internet and one (virtual) for the VPN network. To gain access to the GridFTP servers from the general Internet, four port-forwarding steps are required (figure 7). The port forwarding tables are defined so that each of 24 inbound ports on *anidev.fnal.gov* is eventually mapped to one of the 24 GridFTP servers in the fast network. An

inbound connection to *anidev.fnal.gov* on one of these 24 ports is forwarded to the VPN interface on same machine. This is then forwarded to the “control” (1 Gb/s) interface of the performance host corresponding to that port. Rather than going directly to the 10 Gb/s NICS, this step is necessary because the “control” network is the only one fully routed (section 3.2). In turn, this is further forwarded to the 10 Gb/s interface on that performance host, where the GridFTP server is listening to port 2811.

4.2. Test dataset

The same dataset as in the LIMAN test bed are used (section 2.1.2) but they were sent a different number of times; that is, the total amount of data transferred varied. All the tests performed transferred files from disks on NERSC hosts to the memory on ANL hosts. Note that the files transferred were small enough to be cached in the operating system’s file system RAM buffer cache; therefore, access to disk was not a bottleneck for the measurement.

4.3. Tests details

We have run three types of GridFTP tests, which we have labeled local client-server, local server-server, and remote server-server. With “local”, we mean that the GridFTP control is local to the 100 Gb/s network; with “remote”, the control is outside the VPN and it is port-forwarded. For every test, we transferred each dataset multiple times.

For local client-server tests, on each NERSC host we invoked one GridFTP client command (*globus-url-copy – GUC*) to transfer files from every 10Gb/s interface on the ANL hosts. GUC also controlled the interface selected for control and data channel on the NERSC hosts.

For local server-server tests, on each NERSC host we invoked 48 GUC, controlling third party transfers between every 10 Gb/s interface on the NERSC hosts to every 10 Gb/s interface on the ANL hosts.

For remote server-server tests, we ran the same number of GUC as in the local case; however, the clients controlled the third-party transfers from Fermilab, outside of the VPN, through port-forwarding (section 4.1).

Globus Online tests involved transferring files from the NERSC endpoint to the ANL endpoint. The NERSC and ANL endpoint were created with the 12 GridFTP servers on the NERSC and ANL hosts respectively.

Tables 2 and 3 show different parameters used for different data sets in GridFTP and Globus Online.

Test Type	Dataset	GUC -p	GUC -cc	GUC -pp	GUCS per host	Times dataset transferred per GUC	Total Data Transferred (MB)	Transfer Time (s)
Local: Client-Server	Large	4	4	No	12	3	1,548,288	140.89
	Medium	-	4	Yes	12	5	367,200	38.22
	Small	-	-	Yes	16	250	95,904	262.44
Local: Server-Server	Large	4	4	No	48	3	6,193,152	534.22
	Medium	-	4	Yes	48	5	1,468,800	129.22
	Small	-	-	Yes	16	250	95,904	305.67
Remote: Server-Server	Large	4	4	No	48	3	6,193,152	543.33
	Medium	-	4	Yes	48	5	1,468,800	143.67
	Small	-	-	Yes	48	250	287,712	1,116.33

Table 2: GridFTP test parameters

Test Type	Dataset	--perf-p	--perf-cc	--perf-pp	Number of transfer requests	Times dataset transferred per GO transfer	Total Data Transferred (MB)	Transfer Time (s)
Globus	Large	4	4	32	48	3	2,064,384	262.61
Online	Medium	-	4	32	48	5	489,600	137.58
	Small	-	-	32	48	250	95,904	333.83

Table 3: Globus Online test parameters.

Dataset	Local: Client-Server (Gb/s)	Local: Server-Server (Gb/s)	Remote: Server-Server (Gb/s)	Globus Online (Gb/s)
Large	87.92	92.74	91.19	62.90
Medium	76.90	90.94	81.79	28.49
Small	2.99	2.57	2.11	2.30

Table 4: GridFTP & Globus Online performance measurements.

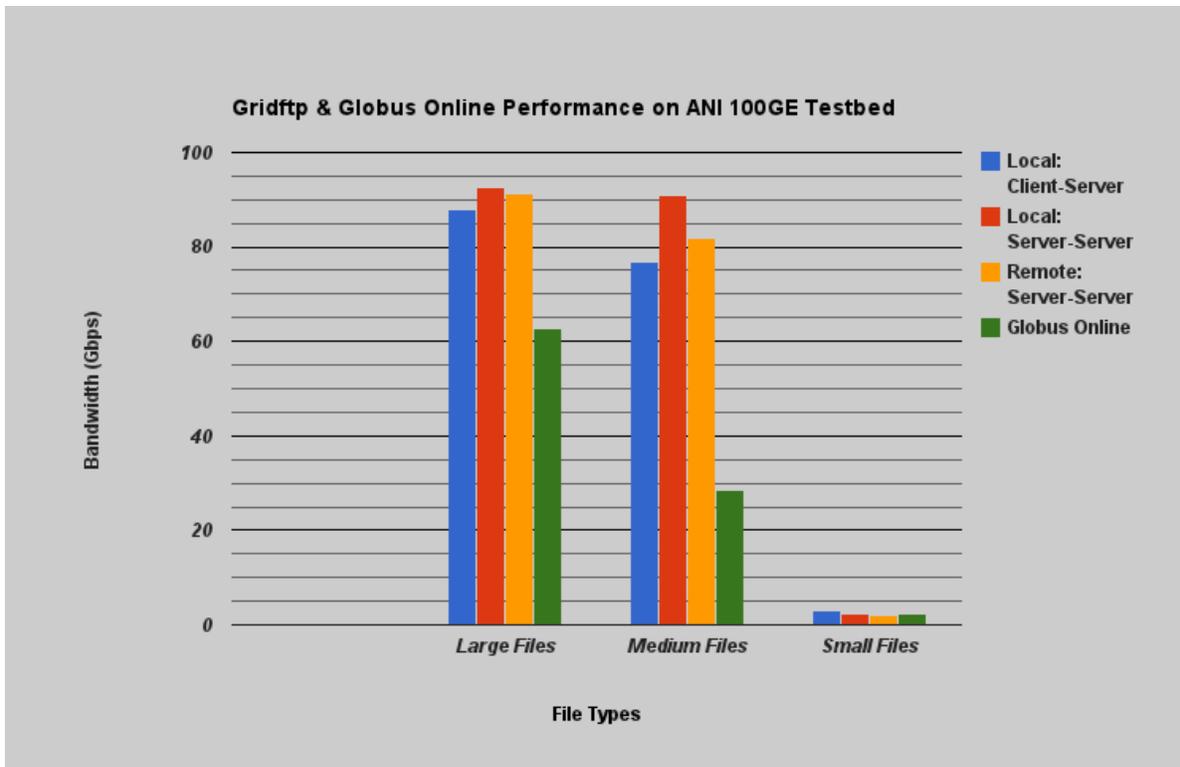


Figure 8: GridFTP & Globus Online performance comparison.

4.4. Observations

Table 4 and figure 8 summarize the bandwidth utilization achieved while transferring datasets using GridFTP and Globus Online. Based on our observation, GridFTP performed well for large and medium files after we tuned the options for number of concurrent servers, number of parallel streams and pipelining (tables 2 and 3); however, as shown in Figure 8, the performance of Globus Online was lower than that of GridFTP. Probable reasons are that there wasn't as much control over the tuning options as for GridFTP and that Globus Online is the system in our tests with the highest control latencies. We are investigating this issue further in collaboration with the Globus Online Team.

Both GridFTP and Globus Online performed poorly for small files. Tuning transfer parameters resulted only in minor changes in performance. We believe that the high latency on the control channels for remote GridFTP tests and the Globus Online tests affected the performance. Also the Local tests on small files suffered from high latency between the two endpoints at NERSC and ANL, which is greater than what it was between the endpoints in the LIMAN test bed Local tests.

5. Xrootd data transfer and comparison with GridFTP

Xrootd is a popular data location and transfer tool in the high-energy physics community. Xrootd provides the capability to manage a cluster of hosts as storage nodes as well as a basic data transfer protocol. The xrootd package consists of the storage server xrootd and the client tool xrscp. In our tests, we focus on the data transfer capability of xrootd and compare it with GridFTP.

As for GridFTP, our goal is to saturate the individual and aggregate network interfaces with xrootd traffic. In general we observe that the protocol requires overhead to initiate, maintain and close the transfer. In order to minimize these, we have investigated running tests with multiple concurrent file transfers and with multiple parallel data streams for each file. We also tested the effects of overhead on file sizes.

5.1. Tests details

We transfer files with xrootd from memory to memory to overcome the bottleneck of file systems. Similarly to GridFTP, the file system kernel cache holds the input after the first read. Unlike GridFTP, though, xrootd does not allow writing to /dev/null to dump the data to memory at the destination. To overcome this limitation, we have set up a RAM disk on the receiving end of the data transfer. This setup, however, limits the amount of data for writing to the total size of the RAM disk. Because of this, we cannot explore the complete parameter space for our tests with this setup.

For GridFTP, we ran multiple parallel transfers concurrently to reduce the effect of the protocol overhead. Unlike GridFTP, the parallel transfers must be managed with independent clients, rather than specifying an option to a single client. In addition, like for GridFTP, we attempted to configure xrootd to transfer files over multiple streams. The xrootd package provides the option to enable multiple streams for the transfer of one file but it did not work properly in our system.

In summary, our tests consist of up to 8 instances of xrootd clients (xrscp) writing files to the xrootd server on a RAM disk.

5.2. Observations

For files larger than 2 GB our tests are limited by the size of the RAMdisk on the server side; therefore, we estimate the aggregate throughput by scaling the results for one NIC. For the other file sizes, we performed a direct bandwidth measurement (see table 5).

Dataset	1 Client (Gb/s)	2 Clients (Gb/s)	4 Clients (Gb/s)	8 Clients (Gb/s)
Large 1-NIC (8 GB)	3	5	7.9	N / A
Large, 1-NIC (2 / 4 GB)	2.3 – 2.7	3.5 – 4.4	5.6 – 6.9	7.7 – 8.7
Medium (64 MB / 256 MB)	2.9 – 8.8	5.7 – 14.7	11.2 – 23.9	22 – 39
Small (256 KB / 4 MB)	0.03 – 0.19	0.07 – 0.38	0.11 – 0.76	0.1 – 1.4

Table 5: Throughput measurements for xrootd clients and servers.

We make the following observations, based on these measurements.

- The xrootd server scaled well with increasing number of clients. Single network interfaces were utilized at about 80% with 4 clients.
- In the case of medium and large files, aggregate throughputs could not be measured directly due to the size of the RAMdisk at the servers. As shown in table 6, for the small file sizes that fit the RAMDisk (510 MB and 1 GB), we measured scale factors for 1 NIC with respect to the aggregate (12 NIC) throughputs. These take into account that 3 machines used concurrently may not be as efficient in driving 4 NIC as 1 machine in driving 1 NIC. We assume that applying the scale factor for 1 GB to 4 and 8 GB files gives us a rough estimate of the aggregate throughputs for these medium and large file sizes. For example, the aggregate throughput estimate for 4 GB is calculated as $8.7 \text{ Gb/s} \times 0.83 \times 12 = 86.7 \text{ Gb/s}$.
- For small files, as with GridFTP, xrootd shows poor performance. As with GridFTP, we will explore the tuning of operating system parameters to overcome this effect.

Dataset (GB)	1 NIC measurements (Gb/s)	Aggregate Measurements (12 NIC) (Gb/s)	Scale Factor per NIC	Aggregate estimate (12 NIC) (Gb/s)
0.512	4.5	46.9	0.87	–
1	6.2	62.4	0.83	–
4	8.7 (8 clients)	–	0.83	86.7
8	7.9 (4 clients)	–	0.83	78.7

Table 6: Aggregate throughput estimation using a scale factor from medium size files.

6. Future Work

Our program of work is driven by the needs of the Fermilab stakeholders. The technologies investigated so far, GridFTP and xrootd, are among the most popular among them. Several other technologies are relevant and will be prioritized based on demand. These technologies include Squid[10] and CVMFS[11], the Storage Resource Broker (SRM), iRODS[12], dCache [13]and NFS v4.1, and Lustre[14].

The process of prioritization is necessary because the current funding for ANI will end in August 2012, although a proposal for an extension has been submitted. In July, we expect Fermilab to join the 100 Gb/s ESNet network, making available 50 Gb/s for R&D efforts in the short term. We plan to use the network at Fermilab to continue our program of work in conjunction with ANI.

7. Conclusion

The Fermilab Computing Sector has developed the High Throughput Data Program (HTDP) to test the functionality and performance of major end-to-end analysis systems of our stakeholders on 100 Gb/s networks. The middleware components of these systems that we tested are GridFTP, Globus Online, and xrootd. Recent work included running tests of data transfers on the ESN Net Advanced Network Initiative between NERSC and ANL for 3 different dataset sizes.

We have tuned GridFTP transfer parameters to maximize transfer bandwidth. In particular, we have tuned the number of concurrent file transfers, the number of parallel transfer streams, and the number of commands “pipelined” over the control channel. We have tuned similar parameters for the GridFTP transfers controlled by Globus Online. Because xrootd does not provide options to directly control these parameters, we have partially emulated the effect of parallel transfers by instantiating multiple client instances.

Both GridFTP and xrootd showed good performance over large and medium size files; small files, however, showed poor data transfer throughput, irrespectively of the transfer parameters. In response, we are planning on tuning operating system parameters instead. These will affect file system kernel caching and process scheduling.

Globus Online was tested forwarding the control channel port to the GridFTP servers inside the network through a VPN gateway. We believe that the resulting high latency is partly responsible for the comparatively lower bandwidth achieved. Another reason might be that options to tune Globus Online are not as extensive as for the direct use of GridFTP. We are working with the Globus Online team to improve overall bandwidth.

We plan to continue the evaluation of middleware relevant to our stakeholders while funds for the ANI last.

8. Acknowledgements

Fermilab is operated by Fermi Research Alliance, LLC under Contract number DE-AC02-07CH11359 with the United States Department of Energy. We thank the DoE ESnet ANI team for their help and assistance.

References

- [1] Energy Sciences Network. Accessed on Jun 1, 2012. <http://www.es.net>
- [2] The U.S. Department of Energy. Accessed on Jun 1, 2012. <http://energy.gov>
- [3] The Advanced Networking Initiative. Accessed on Jun 1, 2012. <http://www.es.net/RandD/advanced-networking-initiative/>.
- [4] F. Bachmann et al, “XSEDE Architecture – Level 1 and 2 Decomposition” Feb 21, 2012. White paper. <https://www.xsede.org/>
- [5] R.Pordes et al. “The Open Science Grid”. In Proceedings of the CHEP’04 Conference, Interlaken, Switzerland, September 27th - October 1st, 2004 2004. Published on InDiCo.
- [6] W. Allcock, J. Bresnahan, R. Kettimuthu, M. Link, C. Dumitrescu, I. Raicu, I. Foster, The Globus Striped GridFTP Framework and Server, in Proc. of the 2005 ACM/IEEE conference on Supercomputing, pp.54-64, Seattle, Washington USA, November 2005.
- [7] Foster, I. Globus Online: Accelerating and democratizing science through cloud-based services. *IEEE Internet Computing*(May/June):70-73, 2011.
- [8] A. Dorigo, P. Elmer, F. Furano, and A. Hanushevsky. XROOTD-A Highly scalable architecture for data access. *WSEAS Transactions on Computers*, 1(4.3), 2005.
- [9] SuperComputing 2011. Accessed on Jun 1, 2012. <http://sc11.supercomputing.org/>
- [10] D. Wessels, *Squid: The Definitive Guide*, O'Reilly & Associates, Inc. Sebastopol, CA, USA ©2004,

ISBN:0596001622

- [11] P Buncic et al, CernVM – a virtual software appliance for LHC applications, 2010 J. Phys.: Conf. Ser. 219 042003
- [12] Rajasekar, A., Moore, R., Hou, C.-Y., Lee, C.A., Marciano, R., de Torcy, A., Wan, M., Schroeder, W., Chen, S.-Y., Gilbert, L., Tooby, P. and Zhu, B. iRODS Primer: Integrated Rule-Oriented Data System. Morgan and Claypool Publishers, 2010.
- [13] P. Fuhrmann. dCache: the commodity cache. In Twelfth NASA Goddard and Twenty First IEEE Conference on Mass Storage Systems and Technologies, Washington DC, Spring 2004.
- [14] P. J. Braam. The Lustre storage architecture. <http://www.lustre.org/documentation.html>, Cluster File Systems, Inc., Aug. 2004.