Final Report for DE-SC0002298
Agency Number: DE-PS02-09ER09-01

An Advanced Network and distributed Storage Laboratory (ANDSL) for Data Intensive Science

Miron Livny
Computer Sciences Department
University of Wisconsin-Madison

The original intent of this project was to build and operate an Advanced Network and Distributed Storage Laboratory (ANDSL) for Data Intensive Science that will prepare the Open Science Grid (OSG) community for a new generation of wide area communication capabilities operating at a 100Gb rate. The plan was to engage a team of four dedicated scientists that will design, develop and operate a state of the art laboratory that will be fully integrated into the organizational and operational infrastructure of the OSG. The ANDSL was planned to include host-based hardware and software required to profile, emulate and study real-life distributed storage workloads at rates of 100 Gbps. The laboratory was designed to utilize the **Department Of Energy's Advanced Network Initiative (ANI) 100GbTestbed** as connections to the fabric become available so that OSG stakeholders can use it to profile real-life **data intensive science applications** as well as experiment with advanced storage and data management technologies**.** The goal of the original program of work was to close the adoption gap, decreasing the time between availability of production 100Gbps links and the end-to-end use of such a network to enable scientific discovery.

Given the significant cut in our proposed budget – no funding for hardware and half the number of requested staff positions – we had to significantly change the scope of the ANDSL. We focused our effort on the software aspects of the laboratory – workload generators and monitoring tools and on the offering of experimental data to the ANI project. For the experimental work, we used storage and compute capabilities deployed by other related efforts as these capabilities became available and connected to the ANI 100Gb testbed. We mainly used such storage and compute "end-point" capabilities at ANL and NERSC leveraging the compute and storage capabilities offered by Magellan.

The main contributions of our project are twofold: early end-user input and experimental data to the ANI project and software tools for conducting large scale end-to-end data placement experiments.

Figure 1 provides an overview of the ANI middleware testbed that was used for running our end-to-end tests. The architecture of the testbed and the syntactic workload driving the tests was the outcome of intensive discussion with other groups of the ANI project and our OSG colleagues to make sure that it is consistent with both the envisioned ANI and the LHC data grid architectures.

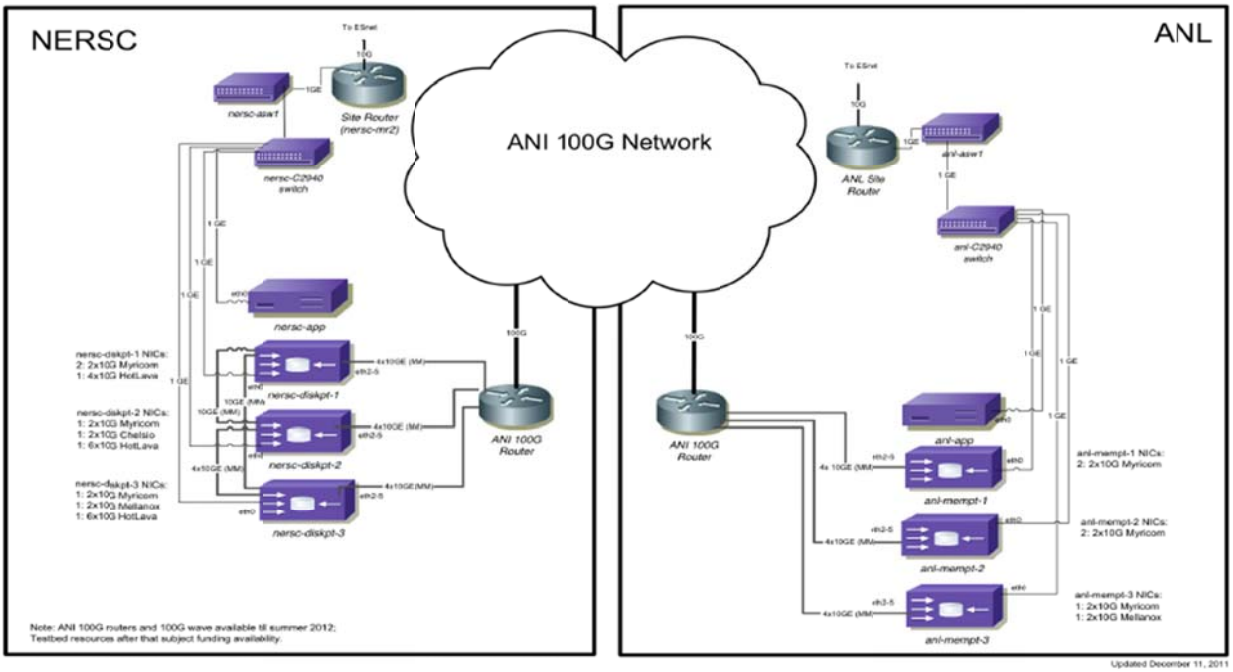

Figure 1 ANI Middleware Testbed

We deployed the OSG middleware stacks and CMS applications in the ANI testbed to demonstrate the readiness of 100Gb network for large scale data transfer and processing. By using virtual machines for managing the various application environments we were able to quickly adapt to new demands and to provide necessary isolation from operating system up to application level. It enabled us to decouple networking infrastructure, storage system and application environment and to ease the system management and configuration. This highly modular computing strategy offers designers with plenty of flexibility and leverage to fit the end-to-end I/O needs of various science projects.

We used the GlideIn Workload Management System (GlideInWMS) and the Condor High Throughput Computing system to deploy and operate the synthetic workloads that included Memory to Memory, Disk to Memory and Memory to Disk applications as well as real-life CMS applications. In order to support large scale data placement experiments that involve new networking and I/O technologies we made the following enhancements to the Condor System:

1. Provisioning of I/O and networking capacity for moving the input and output sand boxes of jobs.
2. Support for multiple data transfer protocols
3. Monitoring of network and I/O consumption resource consumption of the SchedD
4. Deployment and coordination of very large ensembles of test applications

In addition to these software improvements, we performed scalability tests on the ANI testbed, as well as elsewhere. On the ANI testbed, we showed that gridftp transfers can in aggregate operate at close to the full 100Gbps bandwidth as long as sufficiently high bandwidth endpoints are available as sources and sinks. We performed memory-to-memory tests at 93Gbps sustained using all 12 x 10Gbps NICs at both endpoints simultaneously. In this test, 4 gridftp server-client pairs per NIC were used. The corresponding disk-to-memory bandwidth achieved was 80Gbps, and memory-to-disk was substantially smaller. We concluded form this that the

endpoint storage systems were not capable of sinking the full 100Gbps in our disks. Disk-to-disk tests were thus not pursued further.

During a not cost extension phase of the project we used the experience we gained from our work with the ANI testbed to build a prototype of a data placement laboratory across two campuses – UCSD and UW-Madison. Each site of the laboratory includes a dedicated storage cluster and a storage element that is integrated into the local LHC T2 facilities. We are working on extending this laboratory to include other sites both in the US and abroad.

The work performed resulted in contributions to a number of OSG documents [1-6] as well as conference contributions [7,8]. All public OSG documents are available on the web at: http://osg-docdb.opensciencegrid.org/cgi-bin/DocumentDatabase/

[1] Specification of OSG Use Cases and Requirements for the 100 Gb/s Network Connection (OSG-doc-1008)

[2] Measurement of BeSTMan scalability (OSG-doc-1004)

[3] Using Condor glideins for distributed testing of network-facing services (OSG-doc 937)

[4] UCSD data transfers in bandwidth challenge of Supercomputing 2009 (OSG-doc 936)

[5] Hadoop Distributed Filesystem for the Grid (OSG-doc 911)

[6] Scalability of Network Facing Services in the Open Science Grid (OSG-doc 1015)

[7] Roadmap for Applying Hadoop distributed file system in Scientific Grid Computing (presented at ISGC 2010)

[8] Study of WAN data transfers with Hadoop-based SE (presented at CHEP 2010)