Streaming Exascale Data over 100Gbps Networks

Mehmet Balman
Computational Research Division
Lawrence Berkeley National Laboratory
Berkeley, CA 94720

OCT 08, 2012
IEEE Computing Now

Data-intensive computing is one of the crosscutting themes in today's computer research. Many scientific activities depend on large-scale data analysis, scientific simulations, and verification of experimental results. Moreover, experimental facilities are distributed around the world, and scientific communities need to engage in large collaborative efforts.

Consequently, there have been recent efforts to improve the network infrastructure to support increasing demand for information sharing between geographically distributed data centers. One of these efforts is the Advance Network Initiative (ANI[1]) supported by the US Department of Energy (DoE) to deploy 100Gbps interconnects between national laboratories.

As a part of the ANI project, ESnet[2] and Internet2[3] worked together to deliver a 100Gbps prototype network available for researchers at the Supercomputing (SC11) conference. This network, as shown in Figure 1, provided a 100Gbps connection between the National Energy Research Scientific Computing Center (NERSC[4]) in Oakland, California; the Argonne Leadership Class Facility (ALCF[5]) near Chicago; and Oak Ridge Leadership Class Facility (OLCF[6]) in Oak Ridge, Tennessee.

[1] Advanced Network Initiative http://www.es.net/RandD/advanced- networking-initiative

[2] Energy Sciences Network http://www.es.net

[3] Internet2 http://www.internet2.edu

[4] National Energy Research Center http://www.nersc.gov

[5] Argonne Leadership Class Facility http://www.alcf.anl.gov

[6] Oak Ridge Leadership Class Facility http://www.olcf.ornl.gov

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California

In addition, the ANI project also includes a 100Gbps network testbed connecting high-speed hosts between NERSC and Argonne National Laboratory. The ANI testbed[7] has been available to researchers since after the SC11 demo and can be booked to test and evaluate next-generation high-bandwidth networks. At each site are three hosts with a maximum of 40Gbps bandwidth capacity connected to the 100Gbps backbone.

High-bandwidth networks are poised to provide new opportunities in tackling large data challenges in today's scientific applications. However, increasing the bandwidth is not sufficient by itself; we need careful evaluation of future high-bandwidth networks from the applications' perspective.

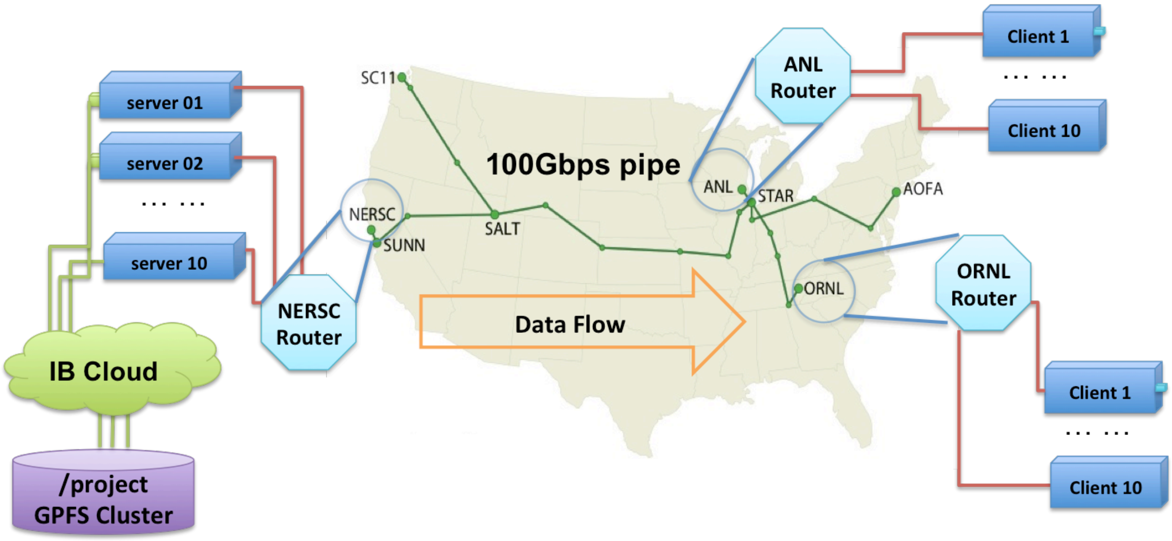


Figure 1: SC11 100Gbps Demo Configuration.

One of the scientific applications that took advantage of the ANI 100Gbps prototype network was the distribution of climate data in the Earth Systems Grid (ESG), between geographically separated supercomputing facilities for scientific data analysis. Climate data is one of the fastest-growing scientific datasets, and it is shared/distributed among many research institutions around the world.

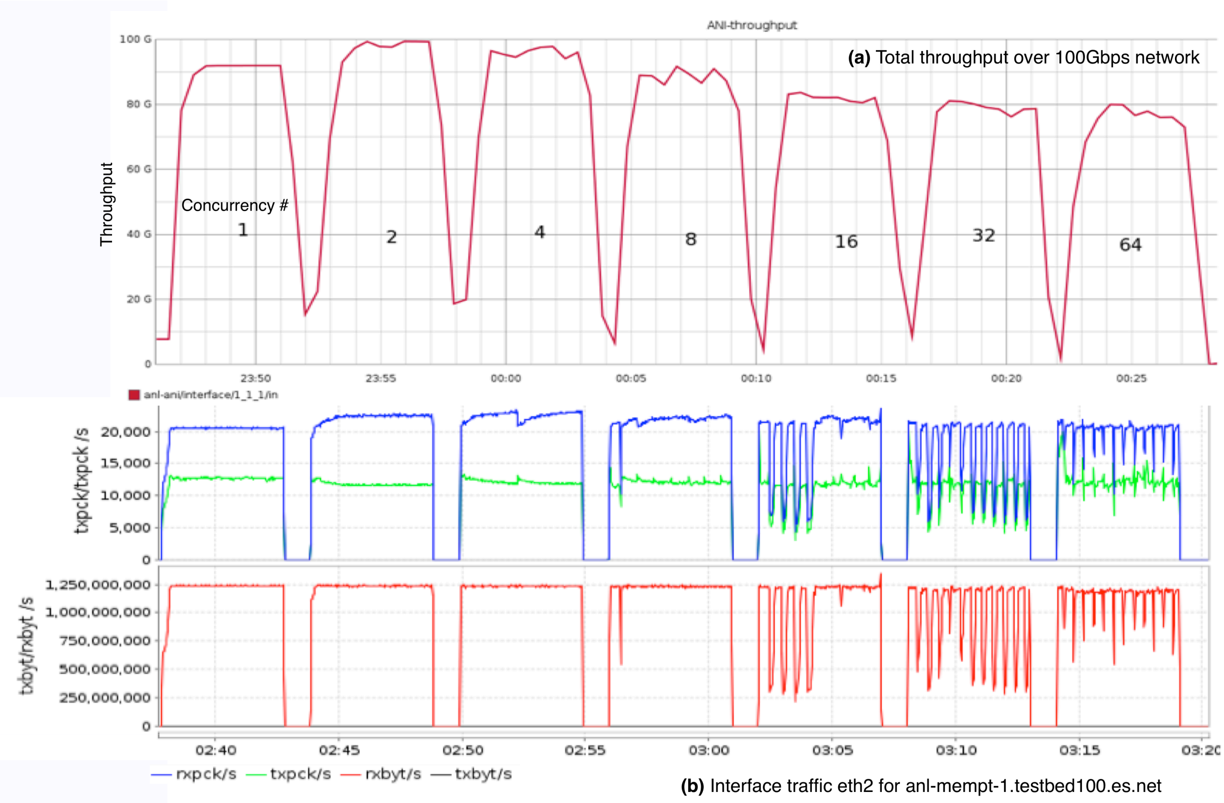
The Earth System Grid Federation (ESGF[8]) provides necessary middleware and software to support end-user data access and data replication between partner institutions. High-performance data movement between ESG data nodes is an important challenge, especially between geographically separated data centers, as many institutions collaborate on the generation and analysis of simulation data.

An important problem in dealing with climate data movement is the "lots-of-small-files" problem. Each climate dataset includes many relatively small files. The state-of-the-art data movement tools require managing each file movement separately. Dealing with many files imposes extra bookkeeping overhead, especially over high-latency networks. In addition to network optimization, data transfers require appropriate middleware for managing and transferring a large number of files efficiently.

[7] ANI Testbed http://sites.google.com/a/lbl.gov/ani-testbed/

[8] Earth System Grid Federation http://esgf.org/

A common approach is to use application level parallelism to provide fast disk access inside the end systems, and to use concurrent transfers to overcome the latency cost in the network. In general, multiple sockets are used simultaneously to fill the high-capacity network pipe.



(a) Total throughput over 100Gbps network

(b) Interface traffic eth2 for anl-mempt-1.testbed100.es.net

Figure 2: (a) Total throughput vs. the number of concurrent memory-to-memory transfers, (b) interface traffic, packages per second (blue) and bytes per second, over a single NIC with different number of concurrent transfers. Three hosts, each with four available NICs, and a total of 10 10Gbps NIC pairs were used to saturate the 100Gbps pipe in the ANI Testbed. Ten data movement jobs, each corresponding to a NIC pair, at source and destination started simultaneously. Each peak represents a different test; 1, 2, 4, 8, 16, 32, 64 concurrent streams per job were initiated for 5min intervals (e.g. when concurrency level is 4, there are 40 streams in total).

We have observed that the use of too many TCP sockets oversubscribes the network and causes performance degradations. Figure 2 shows the effects of concurrent transfers in the ANI testbed. Each peak in these graphs represents a different test with a different number of concurrent operations. Another important drawback in using many transfer streams simultaneously is that the end system resources are over provisioned. Increasing the level of parallelism results in unnecessary CPU load, contention in memory access, increase in context switches, and as a result starvation in system resources.

On the other hand, data movement requests, both for bulk data replication and data streaming for large-scale data analysis, deal with a set of many files. Instead of moving data from a single file at a time, the data movement middleware should handle the entire data collection. Therefore, we've developed a new approach, called Memory-mapped Zero-copy Network (MemzNet) channels, that provides dynamic data channel management and block-based data movement.

Figure 3 shows the underlying system architecture. The idea is to combine files into a single stream on-the-fly. In our tool, data files are aggregated and divided into simple data blocks. Blocks are tagged and streamed over the network. Each data block's tag includes information about the content inside.

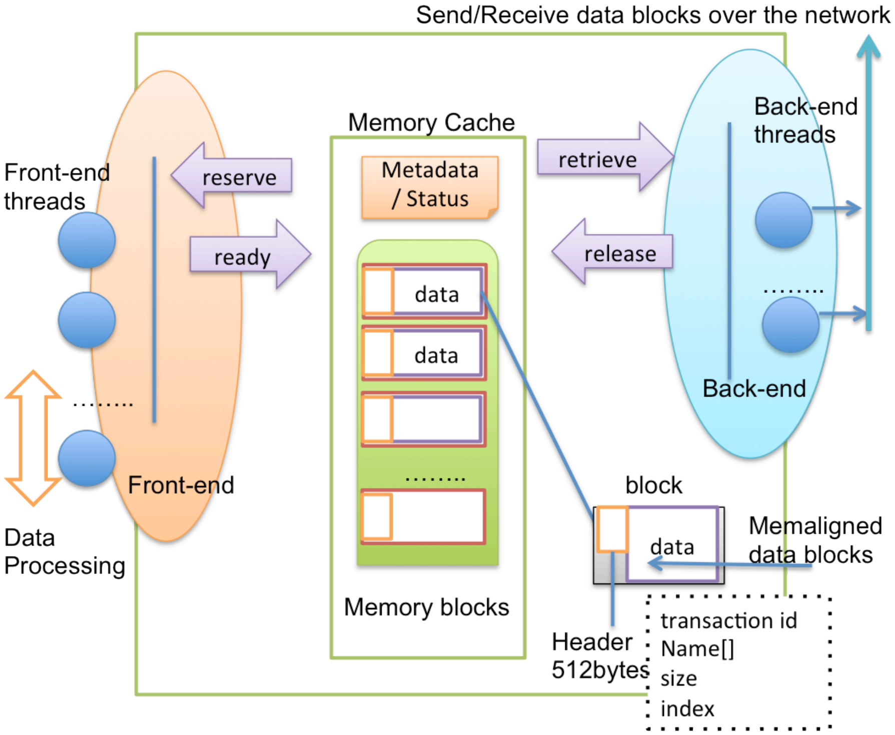
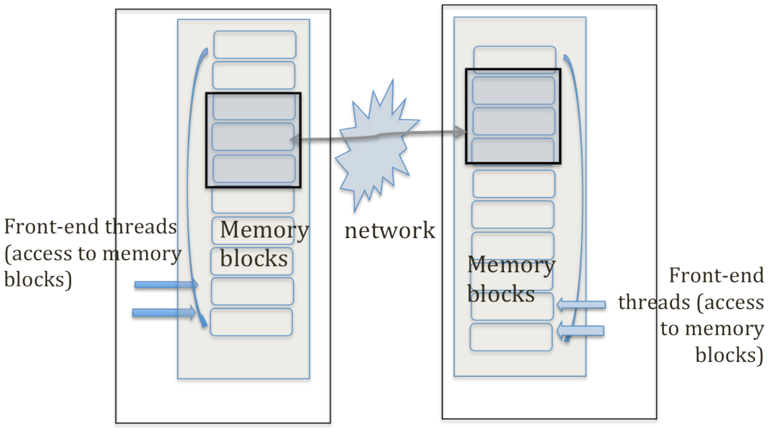


Figure 3: Server/client Architecture for the New Data-Streaming Tool.

In the server, the front end is responsible for the preparation of data, and the back end is responsible for sending data over the network. On the client side, the back-end component receives data blocks and feeds the memory cache, so the corresponding front end can get and process data blocks. Those layers are tied to each other with a block-based pre-allocated memory cache, implemented as a set of shared memory blocks. These memory caches are logically mapped between client and server. Figure 4 gives the overall idea behind MemzNet.

MemzNet introduces dynamic data channel management and asynchronous movement of data blocks. Data is sent in larger chunks when it is ready. A single stream is sufficient to fill up the network pipe. This is analogous in concept to on-the-fly "tar" approach bundling and sending many files together. Moreover, the data blocks can be received and sent out of order and asynchronously. Since we don't use a control channel for bookkeeping, all communication is mainly over a single data channel, over a fixed port.

Bookkeeping information is embedded inside each block. This is beneficial for ease of firewall traversal over a wide area. Also, we can increase/decrease the number of multiple streams (if necessary) without the need to set up a connection channel, since each block includes its bookkeeping information.


Front-end threads (access to memory blocks)
Memory blocks
network
Memory blocks
Front-end threads (access to memory blocks)

memory caches are logically mapped between client and server

Figure 4: Framework for the Memory-Mapped Network Channels.

In the SC11 100Gbps demo, the data from IPCC Fourth Assessment Report (AR4) phase 3, real data from the Coupled Model Intercomparison Project (CMIP), with total size of 35TB, was used in our test and demonstration. In addition to experiments with file-centric state-of-the-art transfer tools (FTP variants), we have also tested our new tool that aggregates files into blocks.

We have observed better performance and efficiency with our tool in transferring large datasets especially with many files. CMIP-3 data was staged successfully into the memory of computing nodes across the country at ANL and ORNL from NERSC data storage over the 100Gbps network.

Next-generation high-bandwidth networks have their peculiarities, and hence need to be studied in detail for optimum performance. Using 100Gpbs efficiently is a feasible goal for the future, but careful evaluation is necessary to avoid inadequate data transfer protocols, poorly tuned network parameters, and host bottlenecks in end systems. The effects of using multiple NICs in multicore environments will play an important role in the overall network performance. Another important aspect is the effect of the application design.

MemzNet is an example specifically designed to take advantage of high-bandwidth networks for high throughput data movements. Future high-bandwidth networks will lead to novel middleware tools and system management software to support efficient resource utilization, and automated tuning and optimization for high throughout network operations.

Acknowledgements: This work is supported by the Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy, under contract no. DE-AC02-05CH11231. This research used resources of the ESnet Testbed, which is supported by the Office of Science of the U.S. Department of Energy under contract DE-AC02-05CH11231