Exploration of adaptive network transfer for 100 Gbps networks

Climate100: Scaling the Earth System Grid to 100Gbps Network

February 1, 2012

Project period of
April 1, 2011 through December 31, 2011

Principal Investigators
Alex Sim[1], Dean N. Williams[2]

Project member: Mehmet Balman[1]
Project Website: http://sdm.lbl.gov/climate100

1 Lawrence Berkeley National Laboratory
2 Lawrence Livermore National Laboratory

Table of Contents

1 Overview

Large scientific experiments and simulations running at precision levels provide better insight into scientific phenomena. However, the amount of data from these experiments and large simulations is in the multi-petabyte range, and expected to grow exponentially. This data explosion exist in nearly all fields of science, including astrophysics and cosmology, high energy physics, material science, climate modeling, fusion, and biology, to name a few. In addition, these scientific data need to be shared by increasing numbers of geographically distributed collaborators connected by high performance networks.
The Climate100 project would study massive datasets, emerging 100 Gbps networks, and data transport and management technologies to enable at-scale experimentation, in particular with climate data management and transport. The result of the project would improve the understanding and use of network technologies and transition the science community to a 100 Gbps network and beyond for production and research.

1.1 The Earth System Grid Federation and Expectations from 100-Gbps System

The ESGF is a successful international collaboration that is recognized as the leading infrastructure for the data management and access of large distributed-data volumes in climate-change research. ESGF, which originated in the Earth System Grid, has evolved to encompass tens of data centers worldwide, collectively holding tens of petabytes of data, and serving tens of thousands of users through ESG P2P Gateways and Data Nodes.
For the IPCC AR5 CMIP-5 data archive is over 30 distributed data archives totaling over 10 PB. The CMIP-5 Replica Centralized Archive (RCA), which the two-dozen major international modeling groups from Japan, U.K., Germany, China, Australia, Canada and elsewhere will replicate to LLNL to form, is estimated to exceed 1.2 PB of data set volume. Not all data will be replicated to CMIP-5 RCA at LLNL, but the majority of the 10 PB of data will be accessible to users from the ESGF P2P Gateways. Figure 1 shows the envisioned topology of the ESGF based on 100-Gbps ESnet network connections to provide a network of geographically distributed Gateways, Data Nodes, and computing facilities in a globally federated, built-to-share scientific discovery infrastructure.
It is projected that by 2020, climate data will exceed hundreds of exabytes (XB). While the projected distributed growth rate of climate data sets around the world is certain, how to move and analysis ultra-scale data efficiently is less understood. Today's average gigabit Ethernet on ESGF is capable of speeds up to 1-Gbps (moving up to 10 TB a day). Tomorrow's 100-Gbps Ethernet speeds, moving up to 1 PB a day, are needed to efficiently deliver large amounts of data to computing resources for expediting state-of-the-art climate analysis. The DOE resources at ALCF, NERSC and OLCF over 100-Gbps are of interest to ESGF for climate analysis.

This document gives a brief overview on the technical progress in Climate100 project for the project period from April 1, 2011 to December 31, 2011.

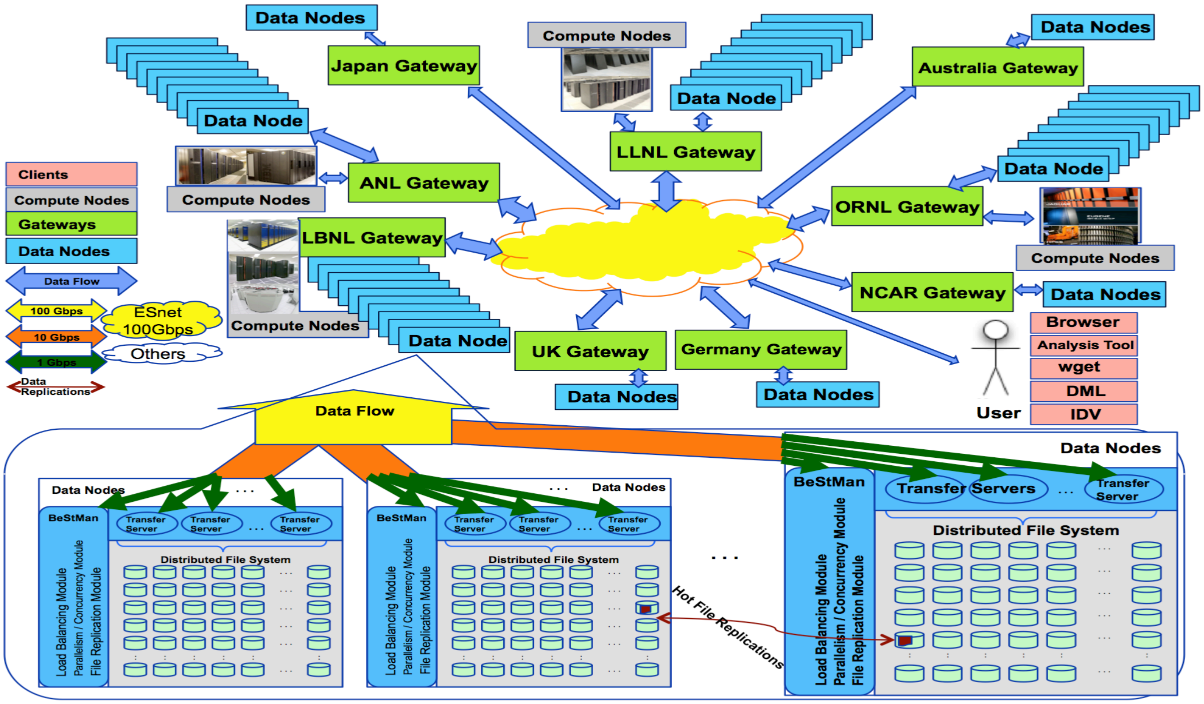

Figure 1: The envisioned topology of the ESGF based on 100-Gbps ESnet network connections.

2 Project Tasks

2.1 Network data transfer optimization

2.1.1 Integration of adaptive data transfer algorithm into data movement tools

Characteristics of the communication infrastructure determine which action should be taken when tuning data transfer operations in order to obtain high transfer rates. Local area networks and wide area networks have different characteristics, so they demonstrate diverse features in terms of congestion, failure rate, and latency. In most cases, congestion is not a concern in dedicated high bandwidth networks. However, the latency wall in data transfers over high bandwidth connections is still an issue. Enough data should be obtained from the applications and storage layers for high throughput performance. Using multiple data transfer streams is a common technique applied in the application layer to increase the network bandwidth utilization. However, the achievable end-to-end throughput and the system load in communicating parties might change during the period of a data transfer, especially when a large volume of data is transferred. Therefore, the number of multiple connections needs to be adjusted dynamically according to the capacity of the underlying environment.

In our previous work we have identified a new algorithm [1] for handling significant end-to-end performance changes (including storage and network bandwidth) in the dynamic network environments as well as controlling the desired data transfer performance. Our adaptive data transfer (ADT) algorithm is designed to calculate the dynamic transfer parameters such as the concurrency level from a simple throughput prediction model with information from current data transfer operations instead of using predictive sampling, and does not depend on external profilers for active measurements. The number of streams is set dynamically in an adaptive manner by gradually increasing the number of concurrent connections up to an optimal point. This also enables us to adapt varying environmental conditions to come up with a high-quality tuning for best system and network utilization. We have prototyped ADT algorithm as a library, which calculates dynamic transfer adjustment parameters to enhance end-to-end data transfer performance.

2.1.2 Data movement and performance optimization for ESGF

The Bulk Data Mover (BDM), the data transfer management tool in the ESGF community, has been managing the massive dataset transfers efficiently with the pre-configured transfer properties in the environment where the network bandwidth is limited. BDM is a high-level data transfer management component with optimized transfer queue and concurrency management algorithms. The integrated test runs in shared environment show its effectiveness from the adaptive transfer management in BDM in obtaining optimal transfer performance for the climate datasets that are characterized by large volume of files with extreme variance in file sizes.

We have engaged in IPCC CMIP-5 dataset replications among international ESGF data centers, and collaborated in network data transfer performance optimization.

- BADC, UK → NERSC
- NERSC → LLNL/PCMDI: network and firewall related problems are identified and fixed to improve the network transfer performance
- NERSC, USA → ANU/NCI: replication for both CMIP-5 and CMIP-3
- DKRZ, Germany → NERSC and LLNL

We also have collaborated with NERSC to support faster network data transfers by having NERSC as a data hop, which NERSC provided 245TB of disk allocation for 6 months to accommodate climate data replication efforts. Current replication size is about 50TB.

We have collaborated with ESnet for Trans-Atlantic network issues and all other network performance issues.

We have collaborated with other climate data modeling centers for the efficient replication techniques.

2.1.3 Summary

Recent progress includes

- Developed a prototype library for ADT algorithm for adaptive adjustment on transfer parameters,
- Integrated ADT library into BDM data movement management tool for ESGF to support adaptive transfers according to dynamic environments,
- Studied effects of large dataset replications over wide area shared network among ESGF sites, and
- Help CMIP-5 data replication and network transfer optimization, in collaboration with ESnet and NERSC,
  - from BADC to LLNL,
  - from BADC to NERSC,
  - from NERSC to LLNL,
  - from NERSC to ANU/NCI,
  - from DKRZ to NERSC, and
  - from DKRZ to LLNL.
- Temporary NERSC disk space allocations (6 months) for 245TB to support CMIP-5 data replication efforts.
- Collaboration with ESnet and international data modeling centers for CMIP-5 dataset replications

and network performance.

Future activities include

- Estimate initial number of multiple streams: The latency directly affects throughput, and more transfer streams are needed to fill the network bandwidth when latency is higher. The challenge in adaptive data transfer is to estimate the initial number of streams based on round-trip time (RTT) between the source and destination hosts. We will experiment with different models to estimate the initial number of transfer streams, which will be used for gradual adjustment for optimum tuning.
- Determine the interval between the transfer parameter adjustment points: The instant throughput performance is measured, but it may not be appropriate to make adjustment on the number of transfer streams after every measurement point. Considering the possibility of minor fluctuations in the network throughput, a threshold value is needed based on some transfer property such as the transferred data size before determining changes in the achievable throughput and adjusting the number of concurrent streams. We plan to experiment with different models to determine the threshold value.
- Identify network transfer optimization issues in international climate data replication, in collaboration with ESnet and NERSC,

2.2 Exploring the remote data access over the network in a climate analysis on cloud computing

CMIP-5 climate community plans to support the analysis on the Gateway server side with CMIP-5 dataset. Climate analysis has unique dependency and requires a large dataset from several Climate Data Nodes. The analysis processing requires large amount of computing power and network performance. Climate Gateway portal does not have such computing power, and upon the user analysis request, customized analysis virtual machine would be created dynamically. The analysis VM is then submitted to the cloud computing facilities in the climate research community. The analysis VMs retrieve the input data from the remote data source on Climate Data Nodes over the network, and the analysis results are transferred to Climate Gateway portal for the user to review and retrieve it.

The study was done by running a tropical storm tracking analysis on Magellan Cloud Computing. The main purpose of this study is to explore possible opportunities and options for future climate analysis over the network with the coming large bandwidth such as 100Gbps network, and to get familiar with Magellan cloud computing resources from the following aspects:

- Network usability in the climate analysis on the cloud computing environment
  - Input data access over WAN from the remote climate data nodes,
  - Coordination of analysis process and the input dataset,
  - Analysis results are stored remotely.
- Effects of remote data access in climate analysis on the cloud computing environment
  - Large amount of input data,
  - The number of analysis VM instances and analysis time,
  - The shared computing resources and network bandwidth.

Transferring large amount of files in many concurrent connections with a limited network bandwidth is problematic in general. Due to the nature of the analysis application in our test case, data retrieval would be expected to be the main bottleneck, and the overall performance will get affected adversely when the number of tasks accessing the data increases.

2.2.1 Test case

Current tropical storm tracking analysis takes ~2 hours with 60 VM instances (This is a simplified numbers for easier calculations), and it is a CPU intensive analysis. The sample input dataset is about 0.5 TB for 15 years with lower resolution. In the previous tests, we have observed that the total analysis time in our analysis case gets reduced linearly to the number of processes [1]. If 600 analysis VM instances can be run, the analysis can be completed in less than 20 minutes if the network can provide ~560MBps (4.5Gbps) to the entire analysis and all 600 analysis VM instances are active simultaneously.

The analysis program for tracking tropical storms in climate simulation dataset works on a single file independently and produces relatively small output for further processing such as trajectory detection. We have prepared custom VM instances with necessary software components installed and configured. In a VM image, an initialization script is triggered automatically when the instance boots up. NERSC Magellan uses Eucalyptus Cloud management software. We have started VM instances using private IP addresses, and defined basic security groups, without any requirements for external access into VM instances.

In order to process the entire dataset using VM instances, we have designed a simple data coordination in which we define an input queue consisting of input files, and distribute files one at a time to VM instances. Each VM instance retrieves a file, analyzes it, and then sends the output back to the data server if any cyclone is detected. Once a VM instance is started, it connects to the coordinator service over HTTP, and requests a file. The coordinator service selects a file from the input FIFO queue, updates its database, and returns the file path back to the instance. After that, the file is transferred using GridFTP and stored into a temporary disk space. This temporary disk space is automatically mounted by Eucalyptus, and it is faster than S3 or Walrus. After the input file is downloaded, the analysis code is run, and the result output is transferred using GridFTP.

2.2.2 Test results on network performance

Since each VM instance needs to download an input file separately, there would be multiple network connections to the source GridFTP server. We expected that the computation for the analysis and data transfers would overlap, considering the fact that VM instances are started in different time intervals and they run independently. Data transfer times were usually dominated by the overall processing time in the CPU intensive analysis. Therefore, there were as many network connections as the number of instances running.

We experienced dramatic performance degradations in the network, especially when we started many VM instances. For example, the RTT value to the data server is usually less than 1ms, but the RTT value increases and becomes more than 100ms when we have started several instances and they all start downloading files from the source GridFTP servers. We have observed the same behavior even with small number of VM instances such as 16 or 20 instances only.

In the test case on Magellan, the main bottleneck is identified in the disk performance in the VM instance for a single transfer operation. We may use tmpfs or ramdisk to enhance the transfer performance since we do not need to store the input data files. One of our goals is to run many VM instances to minimize the overall analysis time. However, when we have many analysis VM instances, the network management in the cloud computing environment does not scale and the network becomes the main bottleneck. Therefore, improving file I/O performance by using tmpfs or ramdisk would not help much if there were many connections towards data source over the network. Note that Magellan Eucalyptus Cloud operates in Managed-noVLAN mode, and each VM instance connects to the Eucalyptus gateway (a single IP, the management host) to route outside the local virtual network.

In the test case, the average throughput was around 120Mbps while downloading a data file to the virtual temporary disk space, inside the VM instance. When data file is stored in memory, no disk I/O is issued, and the average throughput rises to 850Mbps (NERSC Magellan is connected with a 1Gbps link). On the other hand, when we initiated multiple VM instances, and each VM instance requests input files from the source data server independently so that there were multiple transfer requests at the same time, the network transfer throughput was limited by only 50Mbps on average in both cases whether it is written to disk or kept only in memory.

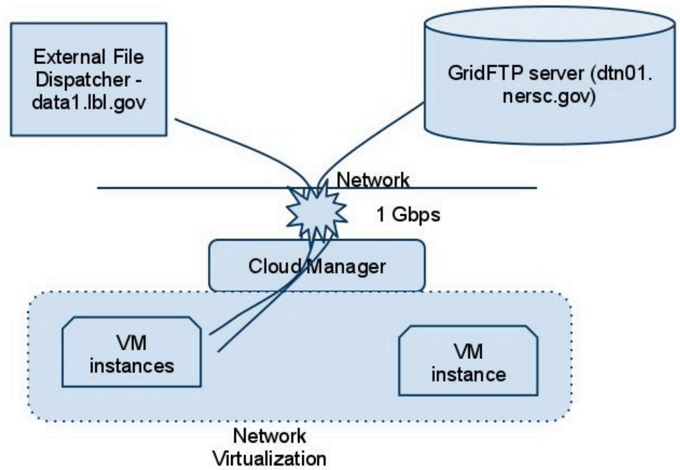

Figure 2: Network performance bottleneck.

2.2.3 Test results on number of VM instances

One of the goals in the study is how many VM instances can be started for an analysis. As the total analysis time decreased by the number of VM instances, it is expected to see many VM instances running for a large analysis. In our tests, we have observed that we were not able to initiate multiple VM instances at the same time or in a reasonable time window. It takes a considerable amount of time to load an image and start an instance. We have used a sample dataset with more than 1.5K files (0.5TB in size), and we have seen that, at most of the time, all files are already consumed before all of the VM instances are initiated. Moreover, we had been advised by the NERSC Magellan team to start 20 VM instances at a time. We were able to run 60 VM instances to analyze climate data files in 3 submissions, but had troubles with more than 60 VM instances. Entire Magellan Eucalyptus system becomes unstable, and a clean startup was needed several times after submitting with many VM instances. Finally, we coordinated a scalable test with the NERSC Magellan team and tried to claim many more resources. In the test, we have requested for 600 VM instances in 12 submissions (50 VM instances in each submission), and 450 VM instances were accepted. The Eucalyptus on Magellan was able to start 147 VM instances within a half an hour. However, the entire Eucalyptus system crashed after an hour, and a clean restart was required. During the one-hour test, we observed that 279 VM instances retrieved at least one input file from the data source, and processed those files successfully. According to our logs, we saw that 378 VM instances tried to retrieve input files. From the sample input dataset (of about 1500 files), only a half of the files were processed. The analysis time spent on a single input file was 3-4 times more than the one from the test with 60 VM instances.

2.2.4 Summary

In summary, the network performance gets very poor on cloud computing if there are multiple connections downloading files from the remote sites. Transferring many files in many concurrent connections with a limited network bandwidth is problematic in general. Current cloud environments add additional limitations for data intensive applications in multiple layers, such as network gateway and disk access. Designing a data gateway that coordinates data distribution might improve the overall performance, but each connection from VM instances needs to go through a network gateway service on the cloud computing environment. (VM instances cannot access directly to the underlying network devices.) NERSC Magellan Eucalyptus cloud network was connected with only 1Gbps. Another issue in our analysis case is in the multiple VM instances unable to start at the same time or in a reasonable time window. Although over 350 VM instances were started in an hour in our tests, it takes a considerable amount of time to load and start VM instances in general.

For data access, retrieving input data files from the remote data source with GridFTP is still a good option in our test case, compared to S3 in which we first need to store all input data into S3 buckets and then retrieve them back inside VM instances. Data access through S3 is not better than via GridFTP as they both are limited by the performance of the cloud head node and by the capability of the network management in Magellan Eucalyptus cloud.

Overall, data intensive applications such as climate analysis might not fully benefit from the cloud computing resources in the current technical state of cloud computing environments in Eucalyptus system. Significant design consideration in Eucalyptus system seems to be needed for scientific cloud for network-aware data intensive applications such as our climate analysis case. Or, our climate analysis case needs to find different ways to use the scientific cloud, with minimal network communication and I/O requirements.

2.3 Study on Remote Direct Memory Access protocol over WAN for data movements

The Remote Direct Memory Access (RDMA) is the protocol that data movement on 100 Gbps network may benefit from. We have studied open fabric and data transfers over RDMA over ANI testbed. We explored the following items:

- Study and experiment open fabric and data transfers over RDMA over Converged Ethernet (RoCE), Software RoCE, Internet Wide Area RDMA Protocol (iWARP) and Software iWARP (Soft-iWARP),
- Implement a client application tool based on RDMA,
- Experiment RDMA-based data transfers for climate dataset over ANI testbed.

2.3.1 Summary

- RoCE-based data transfer requires layer-2 based network reservation, and can maximize the throughput performance when end-systems are configured properly and coordinated well.
- Soft-RoCE based data transfer also requires layer-2 based network reservation, and the throughput performance is good when end-systems are configured properly and coordinated well.
- iWARP and Soft-iWARP based data transfers perform very poorly.

Overall, data intensive applications such as climate research might not fully benefit from the RDMA based data movement in the current state of RDMA technology and costs. Significant configuration and coordination efforts in the end-system configurations as well as lower costs for the RDMA-enabled device seem to be needed for scientific network-aware data intensive applications such as our climate research case.

2.4 Supercomputing 2011 Climate100 Demonstration

The SC'11 demonstration, titled Scaling the Earth System Grid to 100Gbps Networks, showed the ability to use underlying infrastructure for the movement of climate data over 100Gbps network. Climate change research is one of the critical data intensive sciences, and the amount of data is continuously growing. Climate simulation data is geographically distributed over the world, and it needs to be accessed from many sources for fast and efficient analysis and inter-comparison of simulations. We used a 100Gbps link connecting National Energy Research Scientific Computing Center (NERSC) at Lawrence Berkeley

National Laboratory (LBNL), Argonne National Laboratory (ANL) and Oak Ridge National Laboratory (ORNL). In thiedemo, the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4) phase 3 of the Coupled Model Intercomparison Project (CMIP-3) dataset was staged into the memory of computing nodes at ANL and ORNL from NERSC over the 100Gbps network for analysis and visualization. We demonstrated a sample climate application for the analysis of data transferred over the network. In general, climate simulation data consists of relatively small and large files with irregular file size distribution in each dataset. In this demo, we addressed challenges on data management in terms of high bandwidth networks, usability of existing protocols and middleware tools, and how applications can adapt and benefit from next generation networks.

2.4.1 Demo preparation

Profiling Data Transfers Applications for High-bandwidth Networks

Large-bandwidth provided by today's networks cannot be fully utilized due to inadequate protocols (original TCP is not designed for high-speed networks), poorly tuned protocol parameters, and most importantly underutilized capacity in end-systems. There have been several studies to tune up underlying communication protocols. As high-speed optical networks are reaching at the 100Gbps capacity, we are still limited by the end system performance and system resources (i.e. CPU and memory) becoming the main bottleneck for end-to-end data transfers, although the I/O sub-systems are able to provide high-throughput data access. One common approach is to use parallel streams, so multiple threads work simultaneously to overcome the latency cost generated by disk and memory copy operations inside the end systems. Another approach is to use concurrent transfers in which multiple transfer nodes cooperate together to generate high throughput data in order to fill the network pipe. An important drawback in using application level tuning (parallel streams and concurrent transfers) is that system resources are not used efficiently. For example, more than 10 end-nodes, each with more than 10Gbps capacity, might be necessary to fully utilize a 100Gbps link. In the demonstration, one of the focused items was on analyzing various factors that affect end-to-end data transfers. Our goal was to identify major system bottlenecks that could help optimize higher level methods for efficiency and performance and also design new data transfer protocols.

We have examined the overall data movement in two steps. First, data blocks are read into a memory buffers (i.e. disk I/O). Later, memory buffers are transmitted over the network (i.e. network I/O). Each step requires CPU and memory resources. Furthermore, multiple memory copy operations might be necessary unless protocols are not specifically designed for maximum system performance (e.g. O_DIRECT, RDMA, asysnc-IO, etc). In order to model the system with higher accuracy, we have measured performance of each step independently. In addition to several I/O and network benchmark tools, we also had a benefit from the custom tools implemented for simple network operations. We have performed experiments in order to understand affect of each system component separately that was involved in the end-to-end throughput optimization problem. We have compared the results from the experiments with the actual data transfer performance (using GridFTP and rFTP from FTP100 project). Our measurements include the detailed profiling of the data transfer applications during the SC'11 demo. We have collected the details about memory usage, number of context switches, time spent waiting I/O completion, user and system time, call graph of system calls, and time spent in each user operation. We expect that inefficient use of end system resources would be the major bottleneck in high-bandwidth networks. A paper is being prepared based on this study, providing an analysis of parameters that affect end-to-end data movement.

In order to support our effort to analyze performance of end-to-end data movement, we have enhanced our custom test suit and developed a new measurement tool for testing and measuring end-to-end performance. The tool comes with a network library, and it consists of two layers, front-end and back-end. The front-end of the tool is responsible of retrieving data from the disks, and the back-end is responsible of sending data over the network. Each layer works independently, so we can measure performance and tune each layer separately. These layers are tied to each other with a block-based virtual memory file. Data objects in the virtual file include information about the file ID, offset and size. This is similar to having a streaming "tar" approach while sending many files. This enables data blocks to be received and sent out out-of-order asynchronously. Figure 3 shows the data block communication in this approach. We have been testing the new 100Gbps network link between NERSC and ANL/ORNL using this customized tool with various parameters such as memory size, thread count for reading from NERSC GPFS file system, and multiple streams to increase the utilization of the available bandwidth. Since virtual memory file is mapped between client and server, the performance from our measurement tool is comparable with parallel streams and concurrent transfers features that are common in file transfer tools. We have observed that it performs slightly better than an FTP based approach. Main advantage in our tool is that the data transfer is not limited by the data characteristics such as the file size distribution in the dataset, which is common in climate datasets. Another benefit is that it is easy to watch how front-end and back-end layers are performing and to adjust the testing parameters for better performance. All custom tools are available on https://codeforge.lbl.gov/projects/clim100/ under open source license.

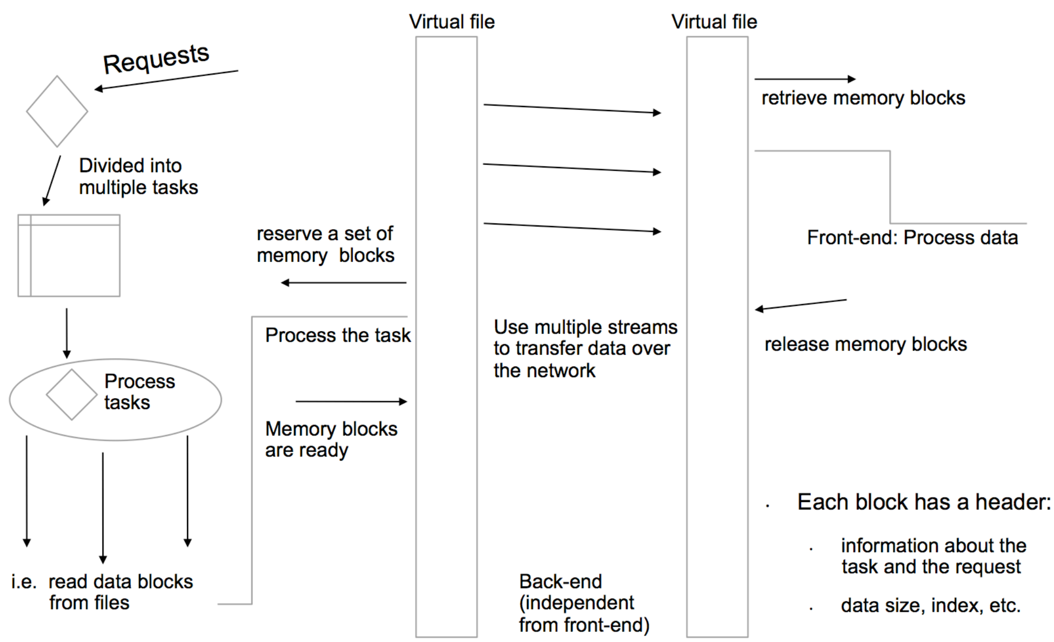

Figure 3: Data block communication in SC'11 Climate100 demonstration

2.4.2 Demo results

SC'11 Climate100 demonstration, showing climate data transfers over 100Gbps network from NERSC to ANL and from NERSC to ORNL, resulted in 83Gbps throughput on average over TCP connections. Five demo schedules were allocated at LBNL booth, and Figure 4 and Figure 5 show the throughput plots during the demonstration time. A paper is being prepared based on SC'11 experiences and study, providing an analysis of parameters that affect end-to-end data movement. Further details of demo and experiences are being prepared on https://sdm.lbl.gov/wiki /Projects/Climate100/Clim100SC11.

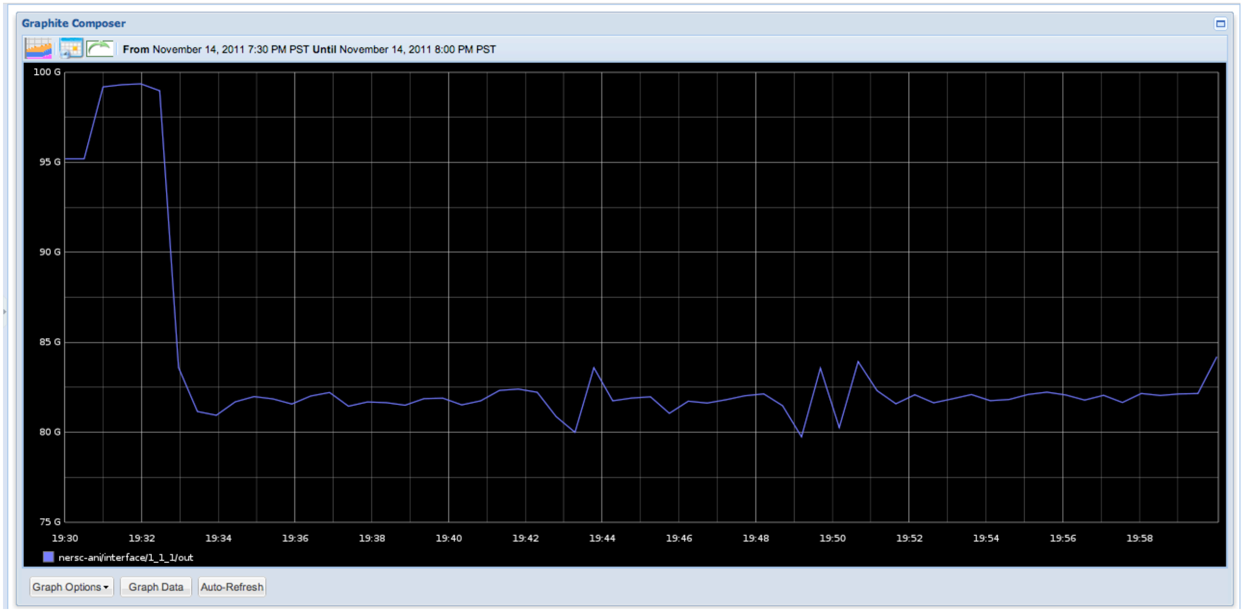

Figure 4: SC'11 Climate100 demonstration result on 11/14/2011 7:30pm-8:00pm PST, showing the data transfer throughput. (Credit: https://stats.es.net/graphite/)

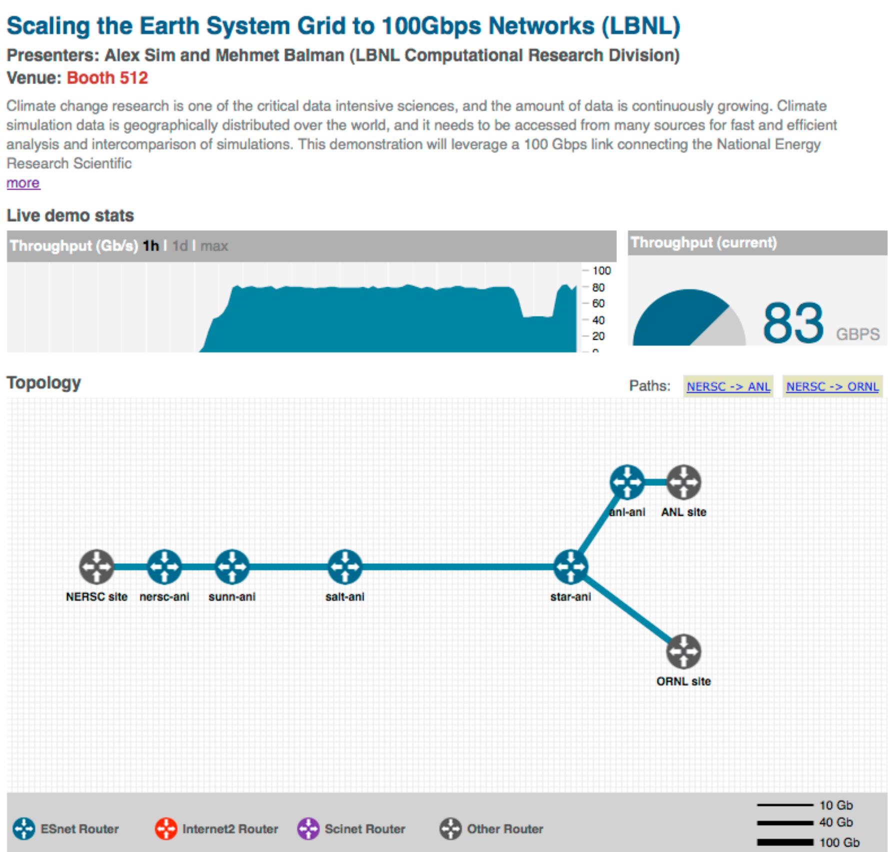

Figure 5: SC'11 Climate100 demonstration result on 11/16/2011 12:00pm-1:00pm PST, showing the data transfer throughput. (Credit: https://my.es.net/topology/sc11/demos/1).

2.4.3 Summary

Recent progress includes

- Demonstrated the outcomes from the Climate100 project scaling Earth System Grid to 100Gbps network,
- Developed a customized measurement tool for measuring end-to-end performance in separate layers of communications,
- Maintained collaboration with ANI testbed groups, with the weekly conference calls.

3 Publications, presentations and other activities

Papers and talks presented during this time period:

3.1 Presentations

1) "*Data Movement over 100Gbps Network*", M. Balman, A. Sim, ESGF P2P Meeting, LLNL, 2011.
2) "DOE's Climate and Earth System Modeling Town Hall: Climate Model Intercomparison and Visualization Efforts for Next Generation Needs – Climate100: Scaling Climate Applications to 100Gbps Network", Dean N. Williams, American Geophysical Union (AGU) Town Hall meeting, San Francisco, CA, 2011.

3.2 Demonstration

1) "*Scaling the Earth System Grid to 100Gbps Networks*", M. Balman, A. Sim, IEEE/ACM International Conference for High Performance Computing, Networking, Storage and Analysis (SC'11), Seattle, WA, 2011. https://sdm.lbl.gov/climate100/docs/SC11-demo.pdf.

3.3 Open source

1) "*Climate100 Toolkit*", A. Sim, M. Balman, LBNL CR-3098, open source under a BSD license with a grant back provision. Available on https://codeforge.lbl.gov/projects/clim100/.

3.4 Workshop

1) International Workshop on Network-Aware Data Management Workshop (NDM2011), in conjunction with the IEEE/ACM International Conference for High Performance Computing, Networking, Storage and Analysis (SC'11). Mehmet Balman as General Chair of the workshop.

4 References

[1] "*Adaptive Transfer Adjustment in Efficient Bulk Data Transfer Management for Climate Dataset*", A. Sim, M. Balman, D. Williams, A. Shoshani, V. Natarajan, Proceedings of the 22nd IASTED International Conference on Parallel and Distributed Computing and Systems (PDPS2010), 2010.

[2] "*Finding Tropical Cyclones on a Cloud Computing Cluster: Using Parallel Virtualization for Large-Scale Climate Simulation Analysis*", D. Hasenkamp, A. Sim, M. Wehner, K. Wu, Proceedings of the 2nd IEEE International Conference on Cloud Computing Technology and Science, 2010.