Climate100: Scaling the Earth System Grid to 100Gbps Network

Final ARRA Report

April 30, 2011

Project period of
October 1, 2009 through March 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

The climate community has entered the petascale high performance computing era and faces a flood of data as more sophisticated Earth system models provide greater scientific insight on complex climate change scenarios. With many petascale data warehouses located globally, researchers depend heavily on high performance networks to access distributed data, information, models, analysis and visualization tools, and computational resources. In this highly collaborative decentralized problem-solving environment, a faster network—on the order of 10 to 100 times faster than what exists today—is needed to deliver data to scientists and to permit comparison and combination of large (sometimes 100s of TB) datasets generated at different locations. This extreme movement and intercomparison of data is not feasible using today's 10 Gigabit per second (Gbps) networks. Therefore the Earth System Grid Center for Enabling Technologies (ESG-CET) architecture needs to be ensured that it scales to meet the needs of the next generation network speeds of 100 Gbps.

The Climate100 project would integrate massive climate datasets, emerging 100 Gbps networks, and state-of-the-art data transport and management technologies to enable realistic at-scale experimentation with climate data management, transport, and analysis and visualization in a 100 Gbps, 100 Petabyte world. The result of the Climate100 project would improve the understanding and use of network technologies and transition the climate community to a 100 Gbps network for production and research.

This document gives a brief overview on the technical progress in Climate100 project for the project period from October 1, 2009 to March 31, 2011, and completes the final ARRA report for Climate100 project.

2. The Earth System Grid (ESG) Use Cases and Expectations from 100-Gbps System

2.1. Earth System Grid and large volume of data sets

The Earth System Grid (ESG), a consortium of seven laboratories (Argonne National Laboratory [ANL], Los Alamos National Laboratory [LANL], Lawrence Berkeley National Laboratory [LBNL], Lawrence Livermore National Laboratory [LLNL], National Center for Atmospheric Research [NCAR], Oak Ridge National Laboratory [ORNL], and Pacific Marine Environmental Laboratory [PMEL]), and one university (University of Southern California, Information Sciences Institute [USC/ISI]), is managing the distribution of massive data sets to thousands of scientists around the world through ESG science Gateways and Data Nodes. For the forthcoming CMIP-5 (IPCC AR5) archive, which will be fully populated in 2011, is expected to have over 30 distributed data archives totaling over 10 PB. The Community Climate System Model, version 4 (CCSM4) and the Community Earth System Model version 1 (CESM1) will submit roughly 300 TB of output out of the 1 PB of data generated to the CMIP-5 archive. The two-dozen (or so) other major modeling groups (e.g. from Japan, U.K., Germany, China, Australia, Canada and elsewhere) will create similar volumes of data with merely a fraction of the data migrating to LLNL to form the CMIP-5 Replica Centralized Archive (RCA), which is estimated to exceed 1.2 PB of data set volume. Not all data will be replicated at LLNL's Program for Climate Model and Intercomparison (PCMDI) CMIP-5 RCA, but the majority of the 10 PB of data will be accessible to users from the ESG federated Gateways. Figure 1 shows the envisioned topology of the ESG enterprise system based on 100-Gbps ESnet network connections to provide a network of geographically distributed Gateways, Data Nodes, and computing in a globally federated, built-to-share scientific discovery infrastructure.

Although perhaps one of the more important climate data archives, CMIP-5 is only one of many archives managed (or planning possible management) under ESG. These includes: The Atmospheric Radiation Measurement (ARM) data, the Carbon Dioxide Information Analysis Center (CDIAC) data, the AmeriFlux observational data, the Carbon-Land Model Intercomparison Project (C-LAMP) data, the North American Regional Climate Change Assessment Program (NARCCAP) data, and other data from wide-ranging climate model evaluation activities and other forms of observations.

It is projected that by 2020, climate data will exceed hundreds of exabytes (1 XB, where 1 XB is 10^{18} bytes). 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 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 Magellan computing resources at ALCF and NERSC over 100-Gbps are of interest to ESG for climate analysis.

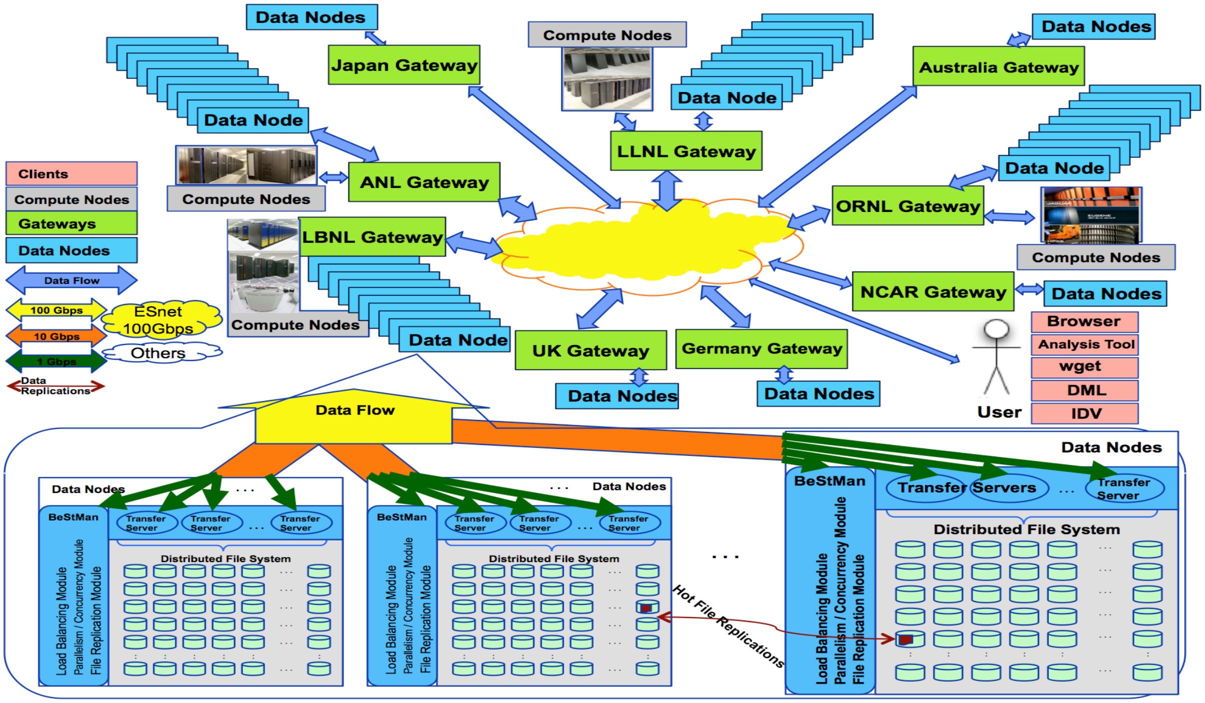


Figure 1: The envisioned topology of the ESG enterprise system based on 100-Gbps ESnet network connections.

2.2. Use Cases

Climate data sets are characterized by large volume of data and large numbers of small sized files; to handle this issue the ESG uses the Bulk Data Mover (BDM) application as a higher-level data transfer management component to manage the file transfers with optimized transfer queue and concurrency management algorithms. The BDM is designed to work in a “pull mode”, where the BDM runs as a client at the target site. This choice is made because of practical security aspects: site managers usually prefer to be in charge of pulling data, rather than having data pushed at them. However, the BDM could also be designed to operate in a “push mode”, or as an independent third-party service. The request also contains the target site and directory where the replicated files will reside. If a directory is provided at the source, then the BDM will replicate the structure of the source directory at the target site. The BDM is capable of transferring multiple files concurrently as well as using parallel TCP streams. The optimal level of concurrency or parallel streams is dependent on the bandwidth capacity of the storage systems at both ends of the transfer as well as achievable bandwidth on the wide-area-network (WAN). Setting up the level of concurrency correctly is an important issue, especially in climate data sets, because of the smaller files. We have test results showing that parallel streams do not have much effect on transfer throughput performance when concurrent transfers are well managed in the transfers of these climate data sets.

Figure 2 shows the overview of the data replications with BDM over 100-Gbps networks. When source directory and target directory are determined for replications, BDM is launched with a pre-configured concurrency, and starts transferring files concurrently. For transfer failures for any reasons except those invalid source paths, file transfers will be retried few times more.

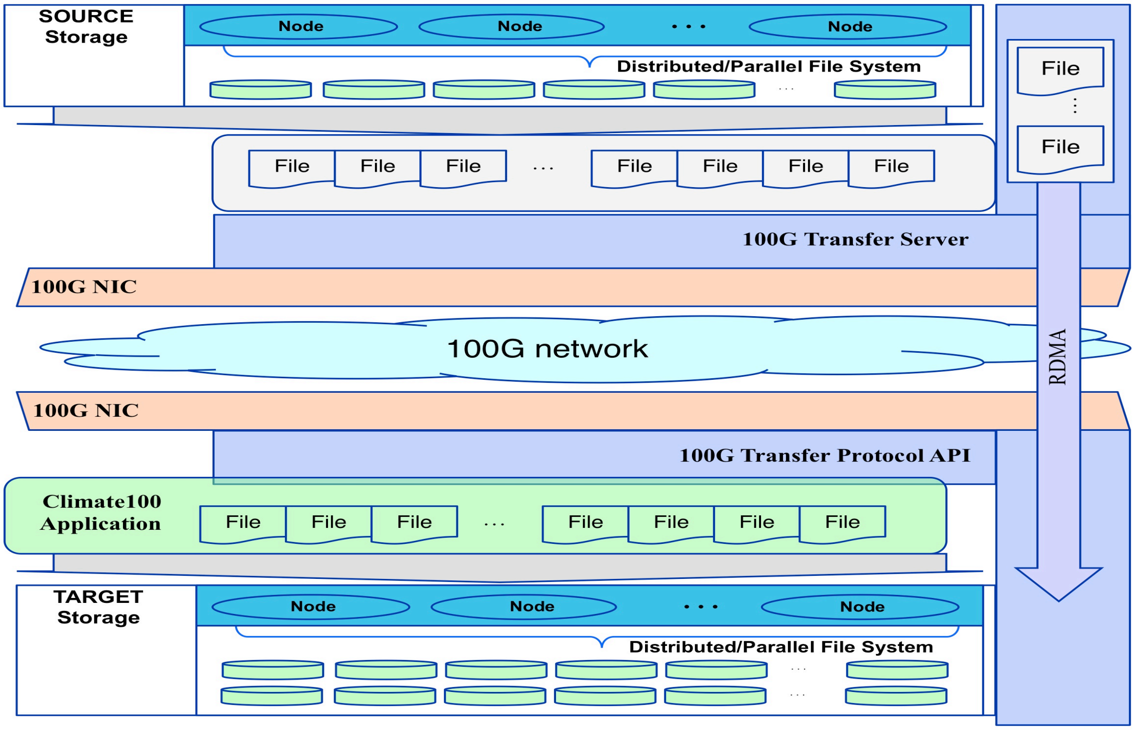


Figure 2: Overview of the data replication use case in Climate100.

The following use cases are targeted for testbed including Magellan facilities, and simulate the production use cases as in the example scenarios.

1) Use Case 1: Data replication from one source host to one target host

This use case is a common climate data replication scenario. This use case can be tested with one host on the source end hosting a transfer server and another host on the target end pulling data from the source server, all over a 100-Gbps network.

Scenario 1: Core dataset needs to be mirrored from source node to target node. For example, BADC node at UK needs to mirror all or part of core dataset from LLNL node. A request token is returned and transfer start asynchronously at a target node at BADC by pulling data from the source node at LLNL. The user can check the status of transfer request using request token. Approximate volume of data is 1.2 PB at the maximum depending on what target node requests to mirror. This process could take about 1 day over a 100-Gbps network connection.

2) Use Case 2: Data replication from many source hosts to one target host

This use case is another common climate data replication scenario. This use case can be tested with Magellan resources on the source end hosting many transfer servers and a host on the target end pulling data from the source servers, all over a 100-Gbps network.

Scenario 2: Data generation sites, with computers specialized for running models are LLNL, NCAR, ORNL, Japan, UK, Germany, Australia, and Canada. Authorized user logs onto LLNL's Gateway and issues a request to collect large-scale data from multiple source nodes (i.e., LLNL, NCAR, ORNL, Japan, UK, Germany, Australia, and Canada) in order to generate a temperature ensemble of the global models. The target node initiates the data transfer by pulling data from the multiple source nodes. A request token is returned to the user, and transfer starts asynchronously. The user can check the status of transfer requests with the request token. Approximate volume of data is 1 PB, depending on the data set.

3) Use Case 3: Data access from many source hosts to many target hosts

This use case is a main climate data analysis scenario. This use case can be tested with Magellan resources on the source end hosting many transfer servers and another Magellan resources on the target end pulling data from the source Magellan servers.

Scenario 3: Thousands of users log onto the LLNL Gateway to search and browse data. They simultaneously request data subsets consisting of hundreds of thousands of files located on the LLNL, NCAR, and ORNL data nodes. Transferable URLs are returned to each user and users start transfers concurrently. The given size that any one user can access and download is approximately 10 TB. The system manages the I/O requests in parallel and balances loads on transfer servers.

2.3. Assessment on requirements

As all the data set resides at LLNL, we first need to move the data to either Argonne Leadership Computing Facility (ALCF) or NERSC, with National Energy Research Scientific Computing Center (NERSC) being the first choice due to the distance from LLNL, and make data transfer tests from NERSC to ALCF, unless LLNL is on 100-Gbps networks. When LLNL is on 100-Gbps networks, we can have test runs on ALCF and NERSC and pull the data from LLNL directly.

1) Common requirements to all use cases

- 100 Gbps backbone network environment
- Java 1.6 or later
- Posix compliant file system

2) Additional minimal requirements for use case 1 (one-to-one data movement)

- 100 Gbps Transfer API at the destination host
- 100 Gbps Transfer Server at the source host
- 100 Gbps network performance from the source host to the destination host
- 100+ Gbps capable storage backend performance at the source host
- 100+ Gbps capable storage backend performance at the destination hosts
- Minimum 22.5 TB storage space required per 30 minute test at source and destination

3) Additional minimal requirements for use case 2 (many-to-one data movement)

- 100 Gbps Transfer API at the destination host
- (Any speed) Transfer Server at the source hosts

    - 1-100+ number of source hosts are needed, depending on each host capacity
- 1/10/100 Gbps network performance from the source hosts
    - 1-100+ number of source hosts are needed, depending on each host capacity
    - 100 Gbps network performance between the two Magellan facilities would be ok.
- 100 Gbps network performance to the destination host
- 10+ Gbps capable storage backend performance at the source hosts
    - 10-100+ number of source hosts are needed, depending on each storage host capacity
- 100+ Gbps capable storage backend performance at the destination hosts
- Minimum 22.5 TB storage space required per 30 minute test runs at destination
    - Depending on each host capacity, minimum (22.5 TB / N source hosts) storage space at the sources is required for 30 minute test

4) Additional minimal requirements for use case 3 (many-to-many data movement)

- (Any speed) Transfer Server at the source hosts
    - 1-100+ number of source hosts are needed, depending on each host capacity
- 1/10/100 Gbps network performance from the source hosts
    - 1-100+ number of source hosts are needed, depending on each host capacity
    - 100 Gbps network performance between the two Magellan facilities would be ok.
- 1/10/100 Gbps network performance to the destination host
    - 1-100+ number of destination hosts are needed, depending on each host capacity
- 1-10+ Gbps capable storage backend performance at the source hosts
    - 10-100+ number of source hosts are needed, depending on each storage host capacity
- 1-10+ Gbps capable storage backend performance at the destination hosts
    - 10-100+ number of source hosts are needed, depending on each destination host capacity
- Minimum (22.5 TB / N destination hosts) storage space at the destination is required for 30 minute test, depending on each destination host capacity
- Minimum (22.5 TB / N source hosts) storage space at the sources is required for 30 minute test, depending on each source host capacity

2.4. Climate100 Phased Goals

Climate100 brings together participants from three areas: applications, infrastructure, and middleware/network research. In the application area, the Climate100 project includes the active participation of ESG-CET to ensure that testbed and technology development activities focus on the specific needs of the climate community. ESG-CET participates in the following areas:

- Provide real data and use cases scenarios for project testing and experimentation;
- Get climate data into the testbed via specified use case scenarios using scaled ESG-CET Gateway and Data Nodes technologies;
- Connect testbed to important ESnet sites (i.e., ANL, LBNL, LLNL, NCAR, ORNL) - these DoE sites have already been established as ESG-CET Gateway and/or Data Node sites where data and resources must be distributed;

- Connect testbed to important global international data centers (i.e., UK, Germany, Japan, and Australia)—these sites have also been established as ESG-CET Gateways and/or Data Nodes sites where data and resources must be distributed; and
- Ready and coordinate crosscutting efforts to transfer from the current 10 Gbps network to the future 100 Gbps production network.

In the infrastructure area, Climate100 also includes the active participation of ESnet, who provides the 100 Gbps network used for experimentation, and collaborate with Climate100 researchers to ensure that instrumentation required for effective end-to-end transfers is available. Finally, Climate100 includes middleware and network researchers. Once end-to-end 100 Gbps connectivity has been provided, the effective use of that network connectivity for climate science is a middleware and network problem.

- Phase 1
  - The primary goal is to move beyond the current machine hardware capability with multiple 10-Gbps connections and to prepare for extension to higher data transfer performance with the coming ESnet 100-Gbps network and multiple distributed storage systems.
  - Test environment involving LLNL and NERSC.
- Phase 2
  - Extend the phase 1 operating environment to ALCF and ORNL for ESG data archives on ANI testbed.
  - The primary goal is to make use of the available 100-Gbps network capability with the designed data transfer framework at ALCF, LBNL/NERSC, and ORNL for ESG data archives and to continue work on data transfers including LLNL.
- Phase 3
  - High-performance data transfers technique over the 100-Gbps network environment will be extended to the broader ESG community, if resource permits, and these research activities will be prepared for ESG production activities.

2.5. Research Tasks and Integration Challenges on the testbed and 100-Gbps network environment

a) Study of the data movement protocols over 100 Gbps network and integration with client applications.

b) Study of the data transfer test cases over 100-Gbps networks and contribute to the system enhancements.

c) Simulation of Round-Trip Time (RTT)/network delays on 100-Gbps testbed. It is understood that large files in size will be transferred for testing on the testbed to avoid complexity of dataset characteristics in large variance of file sizes and file system and storage I/O.

d) Backend storage input/output (I/O) performance for climate data sets. Climate data sets consists of a mix of large and small sized files, and generally much smaller than High Energy Physics (HEP) data files. Parallel/Distributed file system performance on small files is generally very poor. The typical file size distribution in climate dataset in Intergovernmental Panel on Climate Change (IPCC) Coupled Model Intercomparison Project, phase 3 (CMIP-3) indicates that most of the data files have less than 200MB of file size (~60-70% of all files), and among those smaller files, file sizes less than 20MB have the biggest portion (~30% of all files).

3. Project Tasks

3.1. Efficient Data Transfer Mechanism

The Bulk Data Mover (BDM), the data transfer management tool in the Earth System Grid (ESG) 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. We have studied dynamic transfer adjustment to enhance end-to-end data transfer performance. In addition to that, a mathematical model for optimal concurrency estimation and dynamic transfer parameter adjustment for higher throughput in a highly available bandwidth has been developed. The test runs in real shared environment show that the dynamic transfer management in BDM with the dynamic transfer estimation for approximating the initial behavior of the transfer performance is effective in obtaining optimal transfer performance as well as controlling the data transfers at the desired performance for the climate datasets that are characterized by large volume of files with extreme variance in file sizes.

1) Climate Data with File Size Distribution

The climate datasets consist of a mixture of large and small files. The typical file size distribution in climate dataset in Intergovernmental Panel on Climate Change (IPCC) Coupled Model Intercomparison Project, Phase 3 (CMIP-3) indicates that most of the data files have less than 200MB of file size, and among those smaller files, file sizes less than 20MB have the biggest portion. Using parallel streams, in general, improves the performance of datasets with large files. However, when the file size is small, using parallel streams may have adverse affects on the performance. Therefore, file size distribution plays an important role in transfer management. The ***Figure 3*** shows typical file size distribution of a climate dataset in IPCC CMIP-3, showing majority of file sizes are less than 200MB (shown left), and majority of smaller files are around 10-20MB range (shown right).

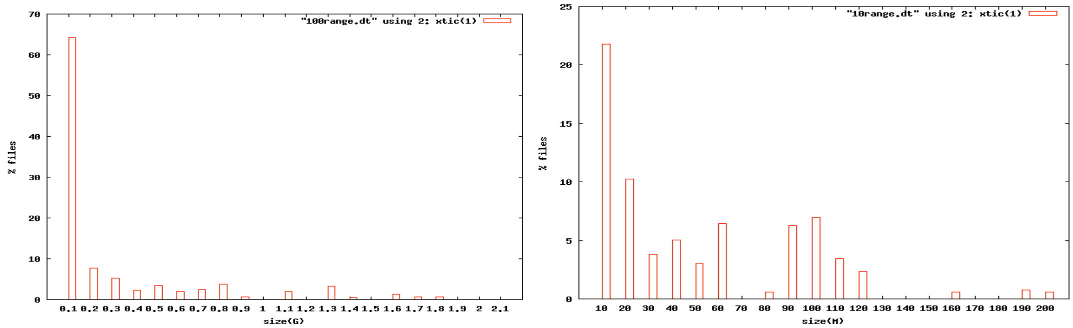


Figure 3: Typical file size distribution of a climate dataset in IPCC CMIP-3, showing majority of dataset file sizes less than 200MB (shown left), and majority of smaller files are around 10-20MB range (shown right).

2) BDM Logger (for debugging and visualizing transfer management)

We have developed a fast and efficient tool to analyze BDM logs. BDM logger is an alternative log analysis and visualization tool. It is specifically designed to capture and see detailed views in queue management. BDM logger examines all log entries for each file transfer operation. It divides entire transfer period into varying time steps and calculates concurrency and throughput metrics in a very efficient way for each time step. We designed BDM logger, as an additional tool to be used along with Netlogger, in order to see effects of concurrency and throughput in queue management. This tool is also used for debugging purposes. It enabled us to see the effect of transfer management algorithms at first hand. Using this tool, we have enhanced our methodology in queue management and made several updates.

3) Concurrency vs. Parallelism

Figure 4 shows that a typical climate dataset transfer over a shared network. It shows transfer throughput performance from two data sources at LLNL to one destination at NERSC over time in seconds on different concurrency and number of parallel streams. Data transfers with less parallel streams show more consistency in file transfer rates throughout the request. According to our experiments, parallel streams do not have much effect in the transfer performance for this type of datasets. Based on this observation, we conclude that concurrent transfers are essential for our case in which we have many small files. Thus, we focus on modeling concurrency level. We improved BDM to maintain the number of concurrency throughout entire transfer without gaps between file transfers

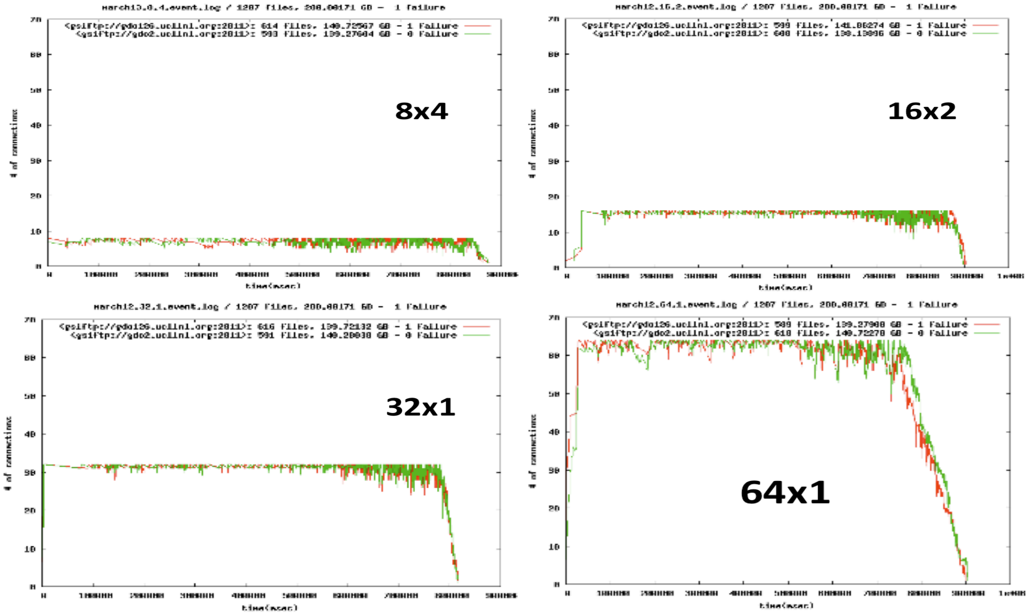


Figure 4: Climate data replication from LLNL to NERSC over shared network showing concurrent GridFTP transfers and load balancing over multiple data sources.

4) Dynamic adaptation algorithm for efficient data movement

We have designed a new approach in which we set the concurrency level dynamically. Instead of making measurements with external profilers to set the level of concurrency, transfer parameters are calculated using information from current data transfer operations. Thus, the network would not have extra packets, and extra load is not put onto the system due to extraneous calculations for exact parameter settings. The number of multiple streams is set by observing the achieved application throughput for each transfer operation, and parameters are gradually adjusted according to the current performance merit. The transfer time of each operation is measured and the total throughput is calculated. The best throughput for the current concurrency level is recorded. The actual throughput value of the data transfers is calculated, and the number of multiple streams is increased if the throughput value is larger than the best throughput seen so far. In this dynamic approach, we try to reach to a near optimum value gradually, instead of finding the best parameter achieving the highest throughput at once. We underline the fact that the focus is on application-level tuning such that we do not deal with low-level network and server optimization. We adjust the number of multiple streams according to the dynamic environmental conditions, and also taking into the consideration of the fact that there might be other data transfer operation using the same network resources. Our dynamic tuning model has been explained in detail in a recent paper [5.1.1].

Based on these studies, we have developed an advanced data movement tool with adaptive tuning capability. Our dynamic approach has been implemented using GridFTP API. We have studied several approaches of this dynamic adaption and enhanced the algorithm after intensive testing and analysis. ***Figure 5*** shows another climate data transfers from LLNL to NERSC for 4.8 TB of a climate dataset from two source servers to one destination. Transfer throughput was consistent most of the time throughout the request, as expected. In the middle of the dataset transfers, low performance was detected, as shown in the middle of the next plot, but the number of concurrency was still at 64 all together. This caused each concurrent connection performance to be much lower, and may have caused packet loss too. The dynamic transfer adjustment can help this case in minimizing overhead of slow data transfers during the low performance period, and the client tool can reduce the number of concurrent transfers to maximize the per-connection throughput which could maximize the resource usability during those time. ***Figure 6*** shows the throughput performance over time from a similar data replication of 1228 files in ~305 GB from NERSC to ANU, one of ESG Gateways in Australia over a shared 10Gbps network connection, and it shows the consistent throughput performance over the well-managed transfer connections throughout the request.

Adaptive Data Transfer (ADT) is designed to automatically adjust and maximize transfer throughput dynamically, for the movement of large climate data sets over WAN. Our current prototype design focuses on transferring large-scale climate datasets around the world. We specifically implement efficient methodologies according to the characteristics of climate data sets. The client application includes a very efficient thread management component. Also, it provides an asynchronous buffer management library that makes file I/O operations efficient. This library enables an abstraction layer for I/O operation. We have performed some local experiments, and observed vast performance improvement with the asynchronous buffer management, since it hid the latency in I/O operations. We also plan to benefit from the client tool for our RDMA and iWARP experiments.

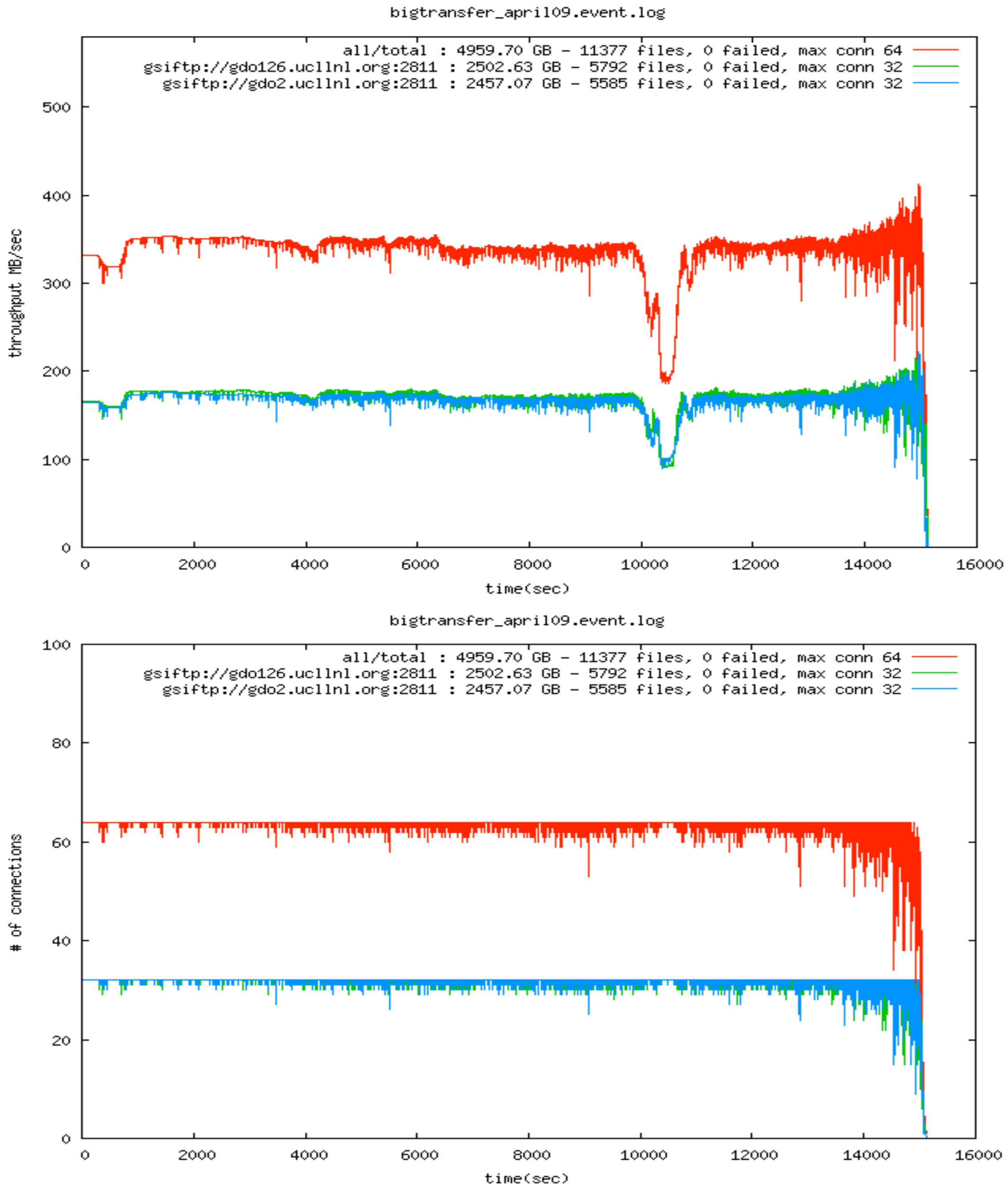


Figure 5: Climate data replication from LLNL to NERSC over shared network. Transfers from 11208 files in 4.8TB of climate dataset from two sources at LLNL to one destination at NERSC with 32 concurrency and 1 parallel stream for each data source show throughput history over time in seconds on the top and the number of concurrency over time in seconds on the bottom.

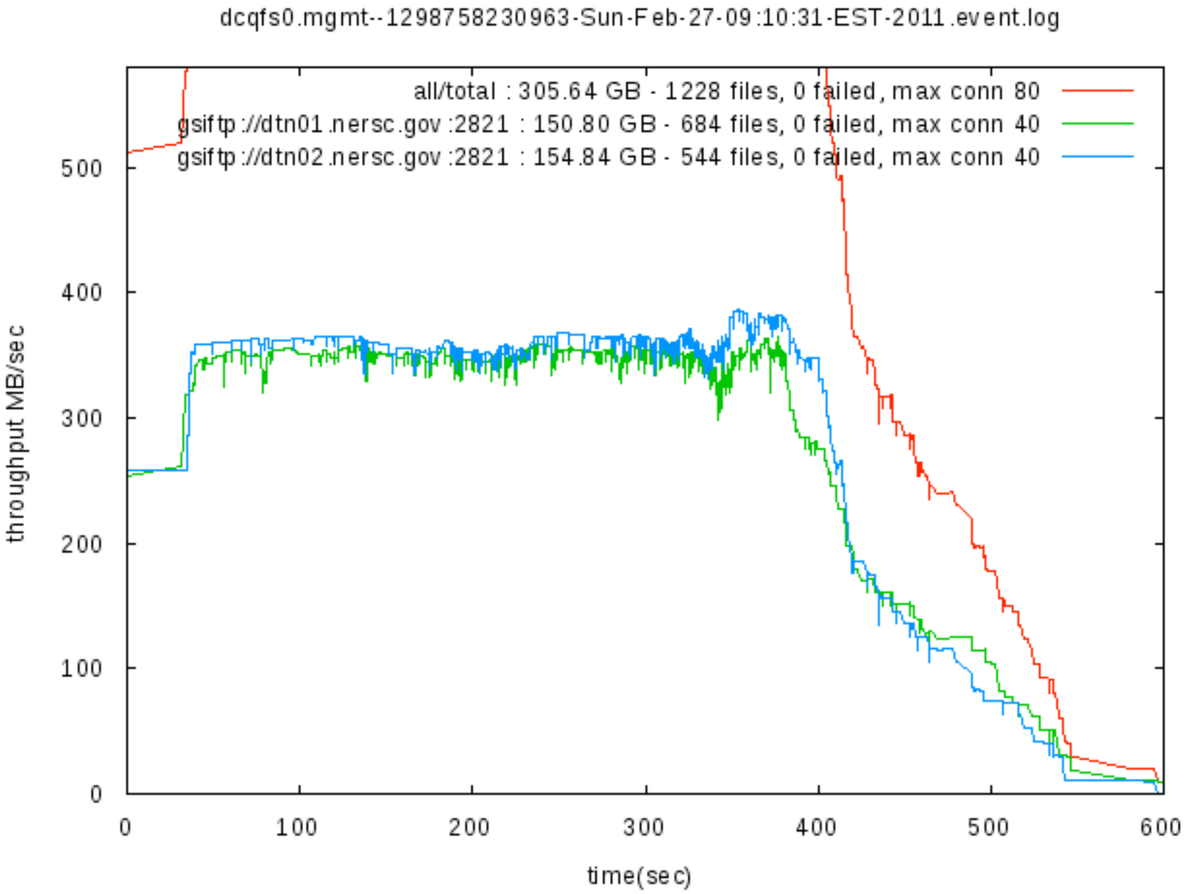
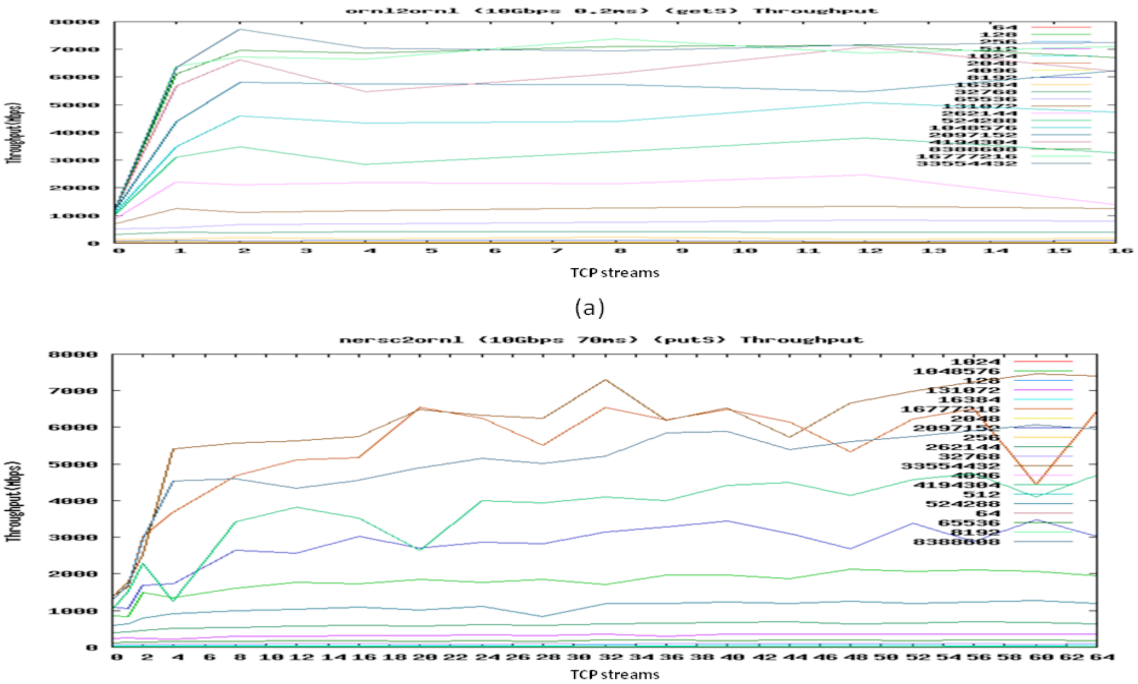


Figure 6: Climate data replication from NERSC to ANU in Australia over shared network. Transfers from 1228 files in ~305GB of climate dataset from two sources at NERSC to one destination at ANU with 40 concurrency for each data source on a shared network of 10Gbps connection show throughput history over time in seconds.

5) GridFTP Experiments over a 10G network

We have performed several experiments with various file sizes by changing the number of concurrent TCP streams. We have measured the overall throughput performance over the number of concurrent TCP streams under different round trip time (RTT) values when different sizes of files are transferred. We have shared our observations with NERSC network engineers. We conclude that latency directly affects the behavior of the throughput performance curve. Figure 7.a shows throughput performance on a 10-Gbps network with round-trip time 0.5ms. As seen in Figure 7.b, more TCP streams are needed to fill the network bandwidth when latency is higher.



(b)

***Figure** 7*: Total throughput over the number of streams; (a) rtt=0.5 ms, (b) rtt=70ms

In addition to GridFTP experiments, we have also implemented a simple transfer server/client application to test the effect of system performance on achievable network throughput. We used this application to profile system calls and measure the practical effect of memory usage in a transfer over network. Our simple transfer server/client application was able to saturate the network, and use the entire bandwidth available, which is hard to achieve using GridFTP (due to its protocol overhead). We plan to improve this internal application and make it more stable to use in the future experiments.

6) Analyzing Concurrency Level: A Power-Law Model

We also observed that we could use power-law to come up with a simple prediction schema. We see that the relationship between the number of multiple streams and the throughput gain can be approximated by a simple power-law model. We presented a power-law approximation to predict the number of multiple streams. The achievable throughput varies as a power of the number of streams where the scaling exponent is related to the round-trip time. It represents the tradeoff between the gain and the cost of adding TCP streams in a data transfer operation over a network. This power law model is used along with the dynamic adaptation algorithm. The goal is to set the initial number of multiple streams that would be calculated in the first phase. This will be used as the base point in the second phase of the algorithm, where we gradually adjust for optimum tuning. Note that we try to obtain a good starting point that will be used later for fine-tuning with dynamic adaptation.

The power-law approximation is modeled as $T = (n / c)^{(RTT / k)}$, where T is achievable throughput in percentage, n is the number of multiple streams ($n > 0$), RTT is the round trip time, and c and k are constant factors. Unlike other models in the literature that try to find an approximation model for the multiple streams and throughput relationship, this model only focuses on the initial behavior of the transfer performance. As in ***Figure 8***, test runs show achievable throughput over the number of concurrent transfers in different *RTT* values. When *RTT* is low, the achievable throughput starts high with the low number of streams and quickly approaches to the optimal throughput. When *RTT* is high, more number of streams is needed for higher achievable throughput. A simple throughput prediction model for approximating the initial behavior of the transfer performance would be effective in quickly obtaining high transfer performance in BDM, and dynamic transfer adjustment contributes to the management in BDM for the optimized as well as controlled transfer performance. We are in the process of implementing this model inside BDM. We also plan to enhance our mathematical approach by conducting several other experiments in various environments.

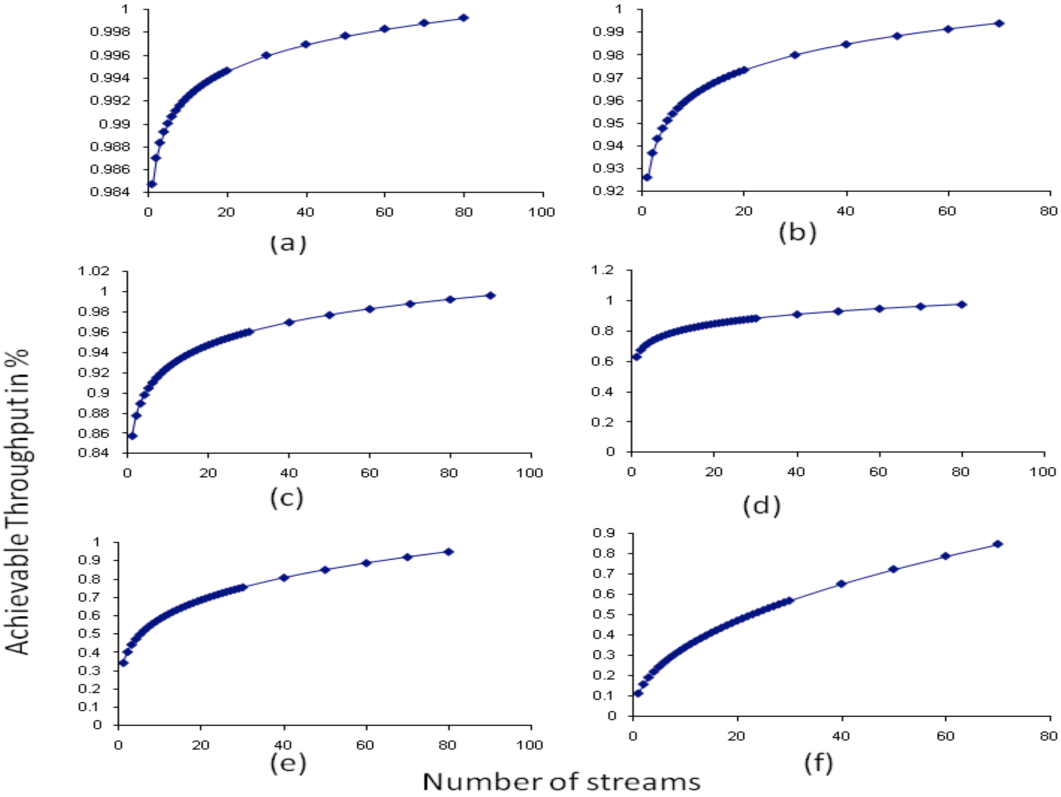


Figure 8: Achievable throughput in percentage over the number of streams with low/medium/high *RTT*;
(a) *RTT*=1ms, (b) *RTT*=5ms, (c) *RTT*=10ms, (d) *RTT*=30ms, (e) *RTT*=70ms, (f) *RTT*=140ms

7) Summary

Completed tasks include

- Studied characteristics of application data by intensively analyzing file size distribution, which directly affects data transfer performance,
- Developed a tool for analyzing BDM logs, for debugging and testing queue management algorithms,
- Conducted many tests over 10Gbps network with GridFTP using different parameters,
- Developed a simple client-server application for experimenting system overhead in a network transfer,
- Designed a new algorithm for setting optimal concurrency level for dynamic adaptation,

- Developed a mathematical model for throughout prediction,
- Designed a hybrid algorithm using dynamic adaptation and initial prediction based on our mathematical model,
- Performed literature review on throughput optimization and data transfer management,
- Implemented an advanced data transfer client which supports automatic performance tuning and adapts to dynamic environments to maximize data transfer performance,
- Developed major components for Adaptive Data Transfer (ADT),
- Studied dynamic adaptation approach and power-law model, and performed experiments on different test environments,
- Studied effects of large dataset replications among international collaborating sites over wide area shared network as well as the reserved SDN,
- Prepared test scenarios and experimented performance of the GridFTP driver in ADT,
- Utilized library implementations for RDMA experiments, and explored possibilities for an efficient driver using iWARP and SoftiWARP.

3.2. ESG Gateway and Data Node Deployment at NERSC

We have actively participated in the ESG-CET community to learn specific needs and support data management requirements of Climate Research over 100-Gbps networks. Our participation in climate community enabled us to provide real-life data and real use cases for testing and also experimenting underlying network infrastructure for Climate100.

We have analyzed climate datasets and obtained particular characterization of application data, which is transferred and replicated between collaborating partners. Our main goal is to enhance data transport technology based on application requirements to ensure that climate community is ready for the next generation high bandwidth networks. In order to support our development effort in Climate100, we replicated IPCC AR4 CMIP-3 datasets (~35TB) from LLNL to NERSC and make it accessible to climate community by deploying an “ESG Data Node” and publishing climate data over an “ESG Gateway” at NERSC. This data will also be available to be used in our Climate100 testbed.

The new architecture of the ESG-CET can be analyzed in three steps. We have global metadata service in the first level, which provides a common database for federation among multiple sites. Next, we have gateways in which user can search and browse metadata. Gateways provide interface to user communities, and they direct users to a data node for access when data is requested. Data nodes hold actual data as well as metadata and update the associated gateway by publishing datasets. A data node basically includes the following components:

- ESG Data Node service
- Thredds server
- CDAT (Climate Data Analysis Tools)
- GridFTP server (with ESG authorization mechanism enabled and configured)
- A database and web server with servlet capabilities (Tomcat, PostgreSQL, Python, etc.)

We have used the recent ESG distribution, which includes latest software releases, and successfully configured and deployed ESG Gateway (esg.nersc.gov) and Data Node (esg-datanode.nersc.gov) at NERSC. We have developed and implemented necessary scripts for a production-level services, and prepared a deployment guide for ESG Data node components based on our experience. In this guide, we elaborate on creating a dataset catalog, listing files in Thredds server and publishing on Gateway, setting up security filters, re-initializing catalog

information, and creating the initial ESG configuration. We gave details about authentication and communication with the ESG Gateway on our guide, explained all required components and listed how to configure ESG software stack. In addition to that, we noted several possible problems and their solutions. We have been documenting our entire deployment experience, and shared this information at our project page (https://sdm.lbl.gov/wiki/Projects/EarthSystemGrid/). In order to join the ESG Gateway federation, we have configured the required security components and generated certificates for the ESG federation. We have updated and released Gateway certificates to be updated in other Gateway systems for ESG federation.
We have tested dataset publication and created a project for IPCC AR4 CMIP3 datasets on ESG Gateway and Data Node at LBNL/NERSC at the production level. ESG-NERSC is one of the Gateway nodes listed in ESG federation. When IPCC AR5 CMIP5 datasets are available, those datasets will be available on NERSC ESG Gateway and Data Node.

1) Summary

Completed tasks include

- Active participation in ESG-CET community for deployment of ESG Data Node,
- Successful deployment of ESG Gateway and Data Node at NERSC,
- Completed IPCC AR4 CMIP-3 dataset replication from LLNL to NERSC,
- Installation and testing of ESG software stack and publishing mechanism,
- Completed ESG federation of ESG-NERSC Gateway and Data Node with CMIP-3 dataset publications,
- Active participation in ESG federation testing with OpenID authentication.

3.3. Remote Direct Memory Access (RDMA) based data movements

The Remote Direct Memory Access (RDMA) is the protocol that data movements on 100Gbps network may benefit from. We have studied open fabric and data transfers over RDMA over InfiniBand (IB) on NERSC Magellan and ANL testbed. This gave us the base for studying further data transfers over RDMA over Ethernet (RDMAoE) on ANI testbed.
We have configured NERSC Magellan testbed machines for GridFTP over RDMAoE in collaboration with ANL, which led a demonstration at Supercomputing Conference 2010 (SC'10) in New Orleans, LA. We also have studied Internet Wide Area RDMA Protocol (iWARP) as an alternative, and performed some experiment on ANI testbed machines. We have studied SoftiWARP on ANI test machines. This SoftiWARP approach may help us use the RDMA protocol in data movements on client machines without RDMA and iWARP enabled cards in them. We have prepared several data movement test cases for wide-area network based on these technologies.
We initiated and continued collaboration with ANL GridFTP team and Ohio State University RDMA team (GridFTP over RoCE). We also initiated and continued our collaboration with ANI FTP100 group, and the bi-weekly conference call was started in September, 2010.

1) Summary

Completed tasks include

- Studied open fabric and data transfers over RMDA over IB,

- Experimented open fabric and data transfers over RMDA over IB with Mellanox ConnectX 2 cards,
- Studied RDMAoE, iWARP and SoftiWARP,
- Helped configuration of Chelsio driver at NERSC Magellan testbed machines (cxgb3toe),
- Prepared a RDMAoE data movement demonstration between NERSC and ANL, during Supercomputing Conference 2010,
- Configured OFED drivers on ANI testbed machines for Myricom cards (myri10ge),
- Experimented with SoftiWARP on ANI testbed machines,
- Experimented SoftiWARP over WAN,
- Maintained collaboration with ANL GridFTP team and Ohio State University RDMA team (GridFTP over RoCE),
- Maintained collaboration with ANI FTP100 group, with the bi-weekly conference calls.

3.4. Large-scale climate simulation analysis on Cloud computing

We have experimented large-scale climate simulation analysis on ANI Magellan Science Cloud. We explored the possibility of using the emerging cloud computing platform in a set of virtual machines (VMs) to parallelize sequential data analysis tasks in analyzing trends of tropical cyclones. This approach allows the users to keep their familiar data analysis environment in the VMs, while we provide the coordination and data transfer services to ensure that the necessary input and output are directed to the desired locations. This work extensively exercises the networking capability of the cloud computing systems, and has revealed a number of weaknesses in the current cloud system. In our tests, we were able to scale the parallel data analysis job to a modest number of VMs, and achieve a performance that is comparable to running the same analysis task using Message Passing Interface (MPI). However, compared to MPI based parallelization, the cloud-based approach has a number of advantages. The cloud-based approach is more flexible because the VMs can capture arbitrary software dependencies without requiring the users to re-write their programs. The cloud-based approach is also more resilient to failure; as long as a single VM is running, it can make progress while the whole analysis job fails as soon as one MPI node fails. In short, this study demonstrates that a cloud computing system is a viable platform for distributed scientific data analyses traditionally conducted on dedicated supercomputing systems.

We also explored different techniques of job coordination algorithms and effects of faster network environment such as 100Gbps networking environment in cloud computing, and identified a few test cases for climate analysis. I/O intensive climate analysis may be in need of different techniques of coordination under faster network environment than CPU intensive climate analysis.

In collaboration with NERSC Magellan team, we plan to explore different types of job coordination techniques, such as Hadoop MapReduce with different backend file systems on flash disk drives as well as spinning disks, batch systems in combination of Hadoop or MPI, and an external job coordination service, with local or remote dataset access. This would allow us to explore further study in the intelligent analysis framework support for climate datasets depending on the analysis characteristics and data access environment including network performance.

1) Summary

Completed tasks include

- Studied different job coordination algorithms,
- Identification of climate analysis test cases,
- Collaboration with NERSC Magellan team.

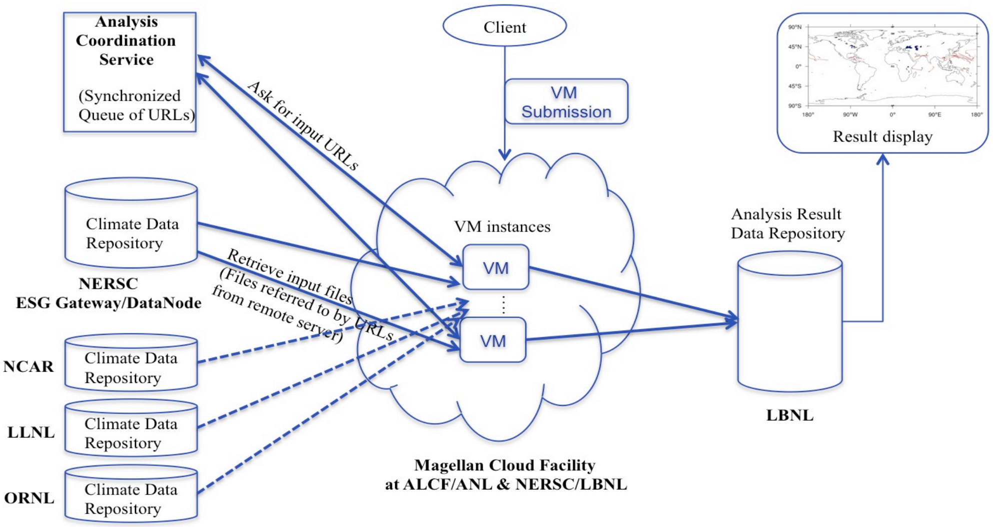


Figure 9: Analysis diagram using Magellan Cloud.

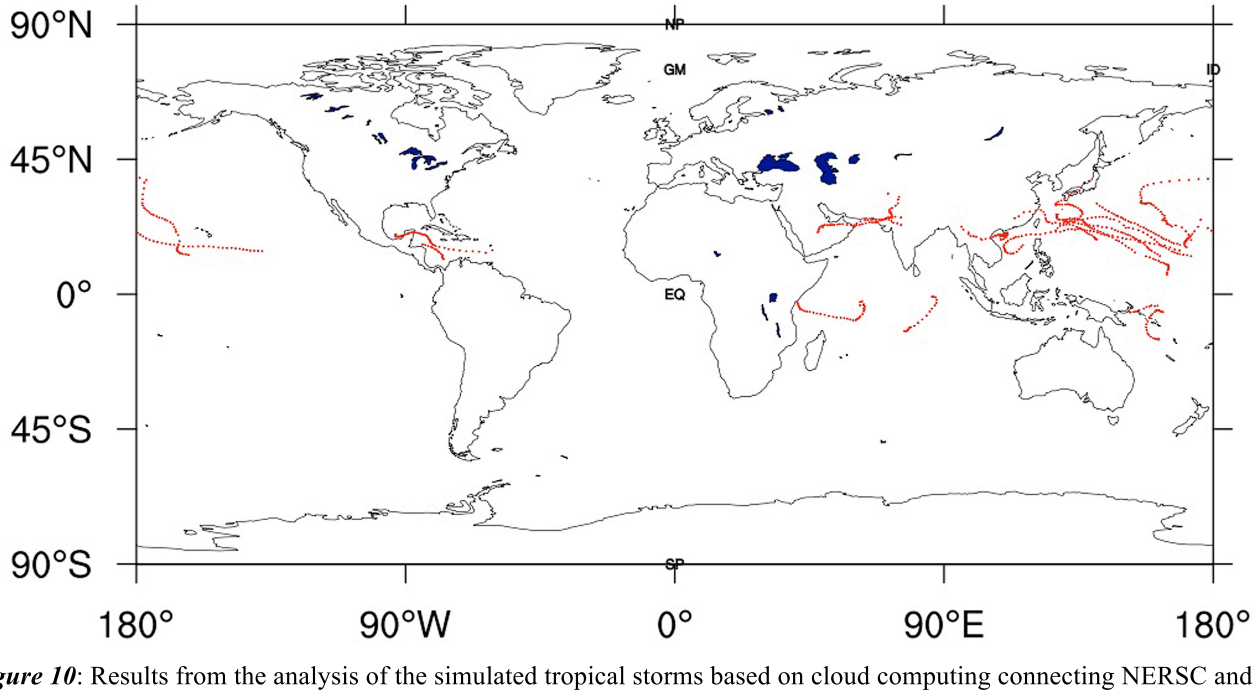


Figure 10: Results from the analysis of the simulated tropical storms based on cloud computing connecting NERSC and ALCF Magellan systems, Sep. 1993, from fvCAM2.2 simulation encompassing 1979-1993.

4. Publications, awards and presentations

Papers and talks presented during this time period:

4.1. Papers

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) "*A Flexible Reservation Algorithm for Advance Network Provisioning*", M. Balman, E. Chaniotakis, A. Shoshani, A. Sim, Proceedings of ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis (SC'10), 2010.
3) "*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.
4) "*Bulk Data Movement for Climate Dataset: Efficient Data Transfer Management with Dynamic Transfer Adjustment*", A. Sim, M. Balman, D. Williams, A. Shoshani, V. Natarajan, "Bulk Data Movement for Climate Dataset: Efficient Data Transfer Management with Dynamic Transfer Adjustment", LBNL Technical Report, LBNL-3985E, 2010.
5) "*Efficient Bulk Data Replication for the Earth System Grid*", A. Sim, D. Gunter, V. Natarajan, A. Shoshani, D. Williams, J. Long, J. Hick, J. Lee, E. Dart, "Efficient Bulk Data Replication for the Earth System Grid", Proceedings of International Symposium on Grid Computing, Data Driven e-Science: Use Cases and Successful Applications of Distributed Computing Infrastructures (ISGC 2010), 2010.
6) "*Lessons learned from moving Earth System Grid data sets over a 20 Gbps wide-area network*", Raj Kettimuthu, Alex Sim, Dan Gunter, Bill Allcock, Peer-Timo Bremer, John Bresnahan, Andrew Cherry, Lisa Childers, Eli Dart, Ian Foster, Kevin Harms, Jason Hick, Jason Lee, Michael Link, Jeff Long, Keith Miller, Vijaya Natarajan, Valerio Pascucci, Ken Raffenetti, David, Ressman, Dean Williams, Loren Wilson, and Linda Winkler, "Lessons learned from moving Earth System Grid data sets over a 20 Gbps wide-area network", 19th ACM International Symposium on High Performance Distributed Computing (HPDC), 2010. http://esg-pcmdi.llnl.gov/publications_and_documents/HPDC10_BWC_Final.pdf
7) "*Error Detection and Error Classification: Failure Awareness in Data Transfer Scheduling*", M. Balman and T. Kosar, International Journal of Autonomic Computing 2010 - Vol. 1, No.4 pp. 425 - 446, DOI: 10.1504/IJAC.2010.037516.
8) "*A New Approach in Advance Network Reservation and Provisioning for High-Performance Scientific Data Transfers*", M. Balman, E. Chaniotakis, A. Shoshani, A. Sim, LBNL Technical Report, LBNL-4091E, 2010.
9) "*Advance Network Reservation and Provisioning for Science*", M. Balman, E. Chaniotakis, A. Shoshani, A. Sim, LBNL Technical Report, LBNL-3731E, 2009

4.2. Awards

1) "*Finding Tropical Cyclones on Clouds*", D. Hasenkamp under the supervision of A. Sim, M.

Wehner, K. Wu, ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis (SC'10), 2010. Won the Third place in ACM Student Research Poster Competition.

4.3. News articles

1) "*Aussies' Data Flows Literally over the Top*", Network matters: Updates from ESnet, 4/1/2011. http://esnetupdates.wordpress.com/2011/04/01/aussies-data-flows-literally-over-the-top/.

4.4. Presentations

1) "*Searching Tropical Storms on Cloud: A Large-Scale Climate Data Analysis*", D. Hasenkamp, A. Sim, M. Wehner, K. Wu, International Symposium on Grid and Clouds (ISGC2011), 2011.
2) "*ESG and Cloud Computing with an Experience from Exploring Cloud for Parallelizing Tropical Storm Tracking*", A. Sim, Expedition Workshop, Seeing Through the Clouds: Exploring Early Communities and Markets Streamlined by Open Government Principles, Oct. 19, 2010.

5. Summary

The goal of the Climate100 project is to research, develop, and test end-to-end capabilities from the new ANI technologies in collaboration with the Earth System Grid (ESG) community. This project contributed to ensure that the software, services, and applications associated with massive climate datasets could be scaled to the anticipated 100Gbps infrastructure, providing high-performance data movement for distributed networking. ESG will deploy similar systems on multiple 10- or 100-Gbps connections to move massive data sets between its major Data Node sites in the U.S. With demonstrated success at the U.S. sites, ESG will then extend high-performance data transfers to smaller Data Nodes in the U.S. as well as to the foreign partner sites. The scaled testing of the data movements and communication connections from Climate100 project is important because the production-scale ESG systems over 100Gbps network will empower scientists to try new and exciting data exchanges that could lead to breakthrough discoveries. The result of the Climate100 project has improved the understanding and use of network technologies and helped the climate community transition to a 100 Gbps network for production and research.