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

Alex Sim, Lawrence Berkeley National Laboratory
Dean N. Williams, Lawrence Livermore National Laboratory
Mehmet Balman, Lawrence Berkeley National Laboratory

This document summarizes the design of the high-level architecture of the Earth System Grid (ESG) climate application for the Energy Sciences Network (ESnet) 100-gigabit-per-second (Gbps) Advance Network Initiative (ANI) project. It also describes the Department of Energy's (DOE's) climate community network requirements defined under ESG. In addition, climate network use case scenarios describing the combination of ESG software and ESnet network—aggregated under the Climate100 project—will assist in understanding the effectiveness of ANI's 100-Gbps backbone DOE laboratory connections for the climate community.

Climate100 is a joint collaboration between ESG and ESnet to provide climate researchers worldwide with a software and network infrastructure to access: data, information, models, analysis tools, and computational resources. ESG is a key data management and transport activity for climate science, including three major ESnet bandwidth activities: receiving data from major modeling centers (including periodic updates); replicating data to other key data holdings or sites; and responding to requests from users for portions of the data. Each of these three major activities use network bandwidth on the order of the size of the repository. For the upcoming multi-petabyte Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report (AR5) Coupled Model Intercomparison Project, phase 5 (CMIP-5) archive, the repository will increase by 3 orders of magnitude, but at best, the network bandwidth is only going up by 1 order of magnitude. The replicated data sites, spreading out the demand, will resolve part of the gap. All of the connected sites will need the fastest available network to deliver results to customers.

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 import 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.

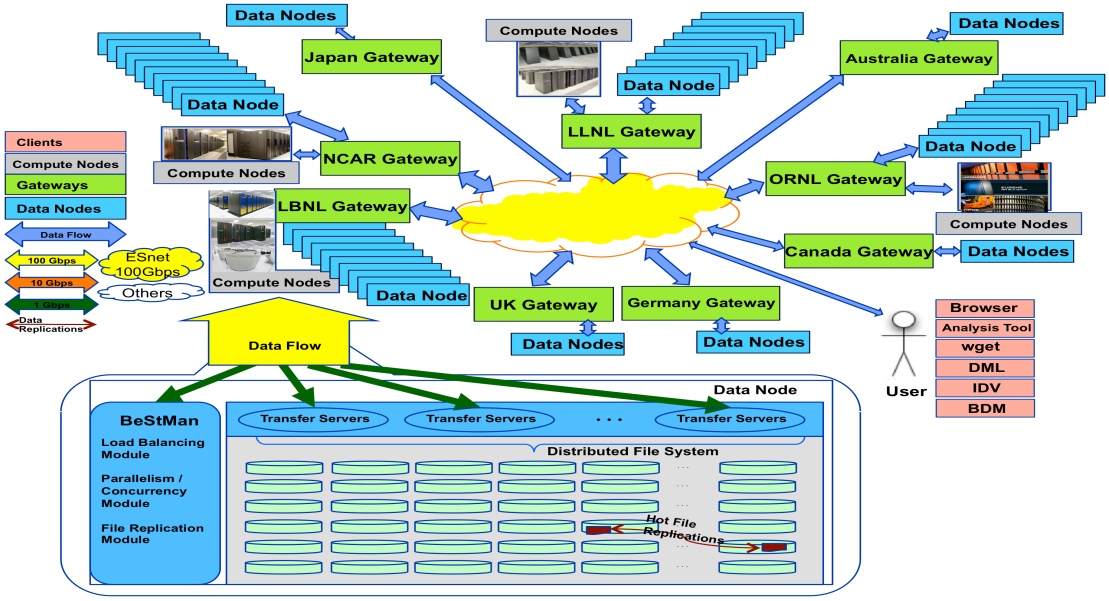

Figure 1: The envisioned topology of the ESG enterprise system based on 100-Gbps ESnet network connections.

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.

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.

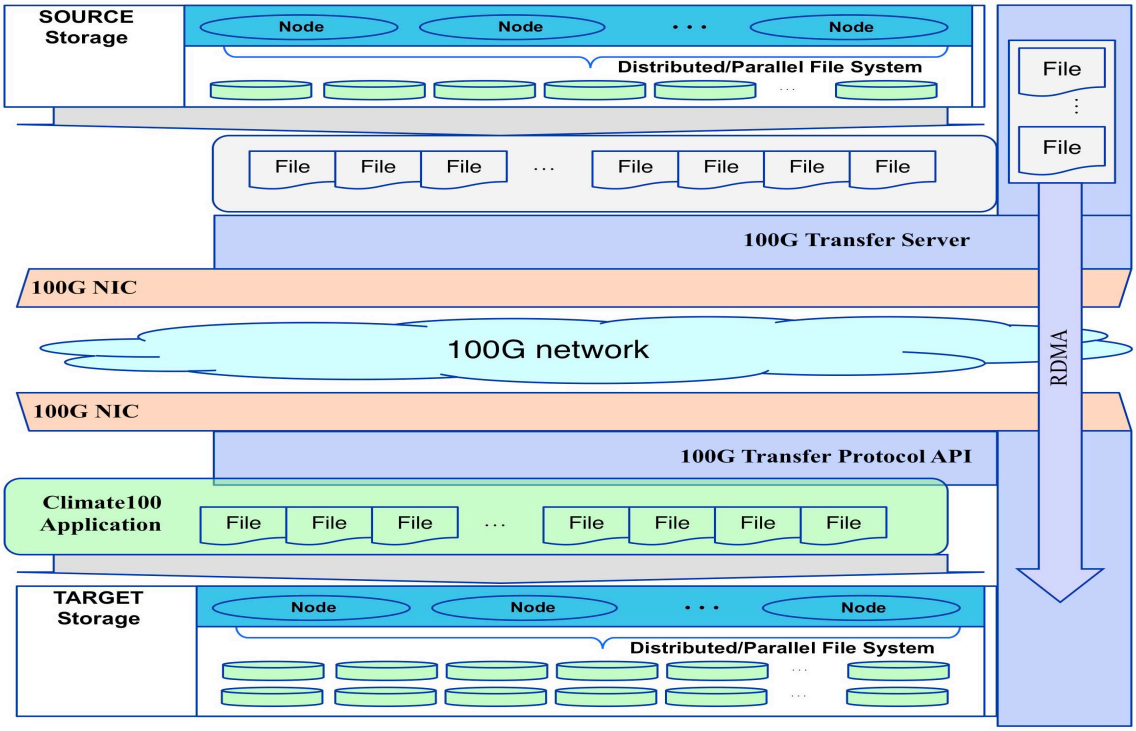

Figure 2: Overview of the data replication use case in Climate100.

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

2.1 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.

2.2 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.

2.3 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.

3 Minimal 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.

3.1 Common requirements to all use cases

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

3.2 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

3.3 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

3.4 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

4 Testbed requirements

- Posix compliant file system access
- Java 1.6 or later
- GNU Compiler Collection (GCC)
- GridFTP server
- Interactive login access
- "Good" size of storage space and "good" storage access performance
  - For 40 Gbps network testbed
    - 40+ Gbps storage performance is needed at both source and destination ends.
    - Minimum 9 TB of storage space is required for 10 minute test runs.
  - For 100 Gbps network testbed
    - 100+ Gbps storage performance is needed at both source and destination ends.
    - Minimum 22.5 TB of storage space is required for 30 minute test runs.

5 Climate100 Phased Goals

- 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.

6 Research Tasks and Integration Issues

6.1 Research Tasks

1. A review of FTP100 protocol functionalities and client Application Programming Interface (API).
2. Study and enhancements on BDM transfers over multiple 10-Gbps shared network environments.
3. Study and enhancements on BDM transfers for FTP100 client API over 100-Gbps network environments.
4. Study of the scalability of file system to be deployed for 100-Gbps network environments.
5. Study the data transfers over 100-Gbps networks and contribute to the system enhancements.

6.2 Integration challenges on the testbed and 100-Gbps network environment

1. Simulating 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.
2. 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).