Climate100: Scaling the Earth System Grid to 100Gbps Network

Progress Report for the Period
July 1, 2010 through September 30, 2010

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

Project member: Mehmet Balman[1]

Project Website: http://sdm.lbl.gov/climate100

Table of Contents

[1] Lawrence Berkeley National Laboratory

[2] Lawrence Livermore National Laboratory

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 will 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 will 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 period from July 1, 2010 to September 30, 2010.

2. ESG Data Node and Gateway Installation

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

We have used the recent ESG distribution, which includes latest software releases, and first deployed ESG Data Node and then ESG Gateway at NERSC systems. Dedicated machines at NERSC have been allocated for ESG Gateway and ESG Data Node services. We have made a new deployment for ESG services, and the services are actively being tested. One of the main challenges we have faced during our deployment at NERSC is on the customization of the services. We have been documenting our entire deployment experience for ESG Data node and Gateway services, and shared this information at our project page (https://sdm.lbl.gov/wiki/Projects/EarthSystemGrid/WebHome)

2.1. Summary

Recent Progress includes

- Active participation in ESG community for deployment of ESG services,
- Installation and testing of ESG software stack and publishing mechanism,
- ESG Data Node and ESG Gateway deployment at NERSC systems.

Future Activities include

- Complete deployment of ESG Data Node and Gateway with CMIP-3 dataset publications.

3. IPCC AR4 CMIP-3 Climate Data Replication

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 Intergovernmental Panel on Climate Change (IPCC) the Fourth Assessment Report (AR4) phase 3 of the Coupled Model Intercomparison Project (CMIP-3) datasets (~35TB) from LLNL to NERSC with Bulk Data Mover (BDM), and make it accessible to climate community through a “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. Figure 1 and Figure 2 show the throughput performance and connection management of the dataset transfers from LLNL to NERSC.

3.1. Summary

Recent Progress includes

- Complete IPCC AR4 CMIP-3 dataset replication from LLNL to NERSC,

Future Activities include

- Engage in the replications within ESG federation with CMIP-3 dataset at NERSC.

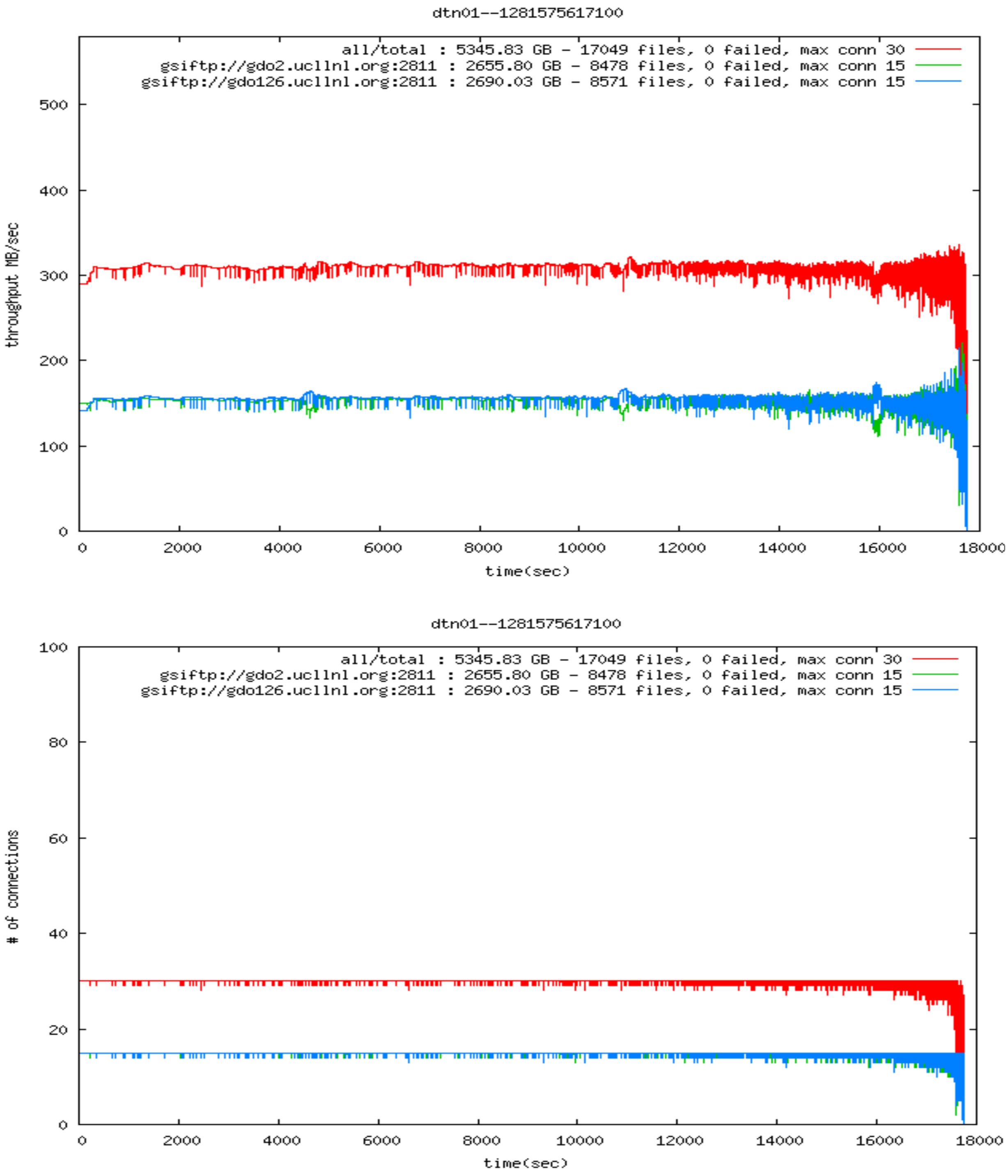

Figure 1: Climate data replication from LLNL to NERSC over shared network. Transfers from 17049 files in ~5.3TB of climate dataset from two sources at LLNL to one destination at NERSC with 15 concurrency and 1 parallel stream for each data source show throughput history over time in seconds on the top and the number of concurrency over time in seconds on the bottom. It shows the consistent throughput performance well-managed transfer connections throughout the request.

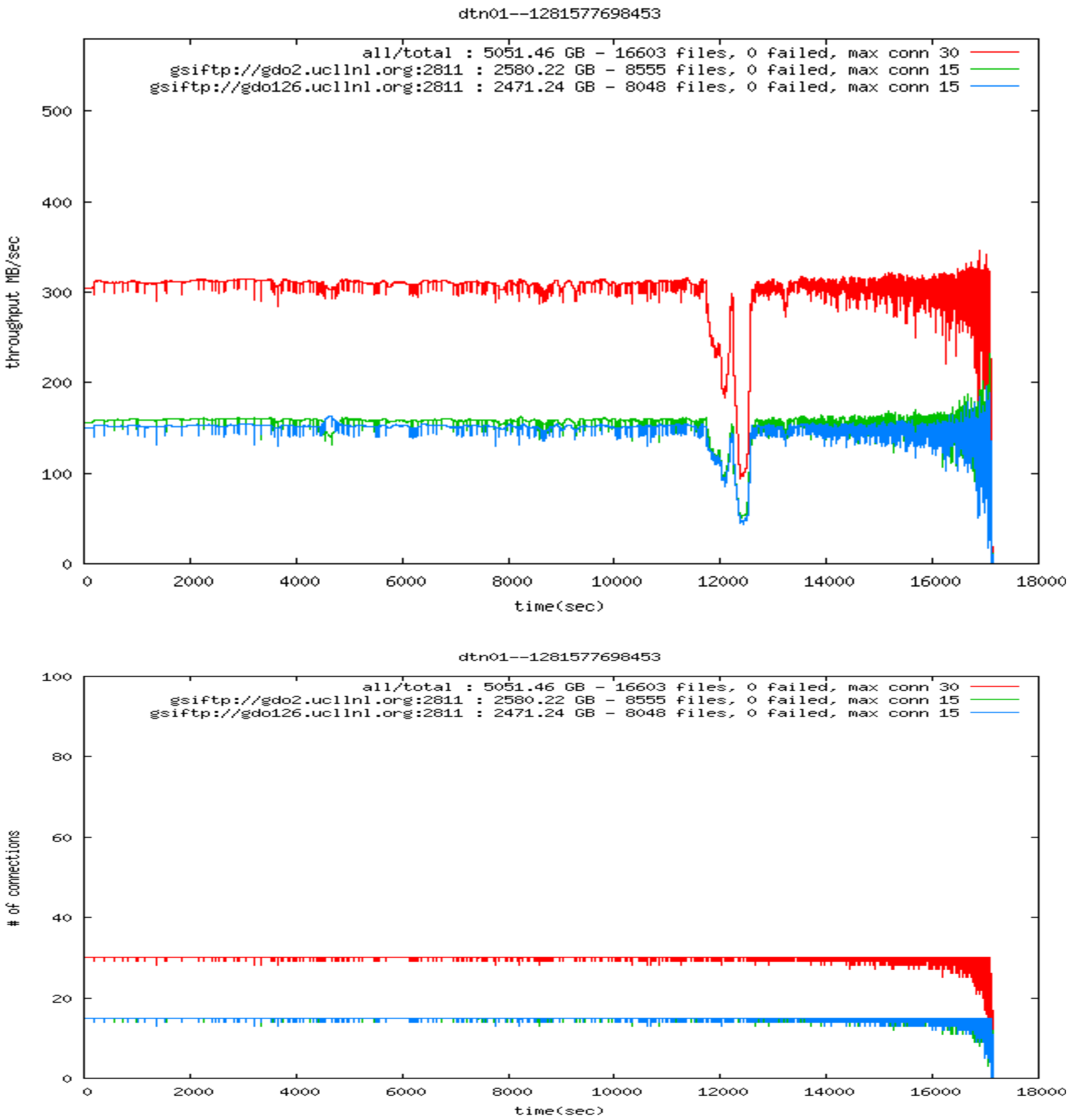

Figure 2: Climate data replication from LLNL to NERSC over shared network. Transfers from 16603 files in ~5.05TB of climate dataset from two sources at LLNL to one destination at NERSC with 15 concurrency and 1 parallel stream for each data source show throughput history over time in seconds on the top and the number of concurrency over time in seconds on the bottom. 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 plot, but the transfer concurrency was still at 30 all together. This caused each concurrent connection performance to be much lower, and may have caused packet loss too. The adaptive transfer management can help this case in minimizing overhead of slow data transfers during the low performance period, and the future release of the BDM can reduce the number of concurrent transfers to maximize the per-connection throughput which could maximize the resource usability during those time.

4. Progress towards Remote Direct Memory Access (RDMA) based data movements

The Remote Direct Memory Access (RDMA) is the protocol that data movements on 100Gbps network will benefit from. We have studied open fabric and data transfers over RDMA over InfiniBand (IB) on NERSC Magellan testbed. This gave us the base for studying further data transfers over and RDMA over Ethernet on ANI testbed in the future.

We also initiated our collaboration with ANI FTP100 group, and the weekly conference call was started in September, 2010.

4.1. Summary

Recent Progress includes

- Study open fabric and data transfers over RMDA over IB,
- Initiated collaboration with ANI FTP100 group, and set up the weekly conference calls every other week.

Future Activities include

- Study open fabric and data transfers over RDMA over Ethernet (RDMAoE),
- Implement client application tool based on RDMAoE,
- Integration with FTP100 data transfer server with RDMAoE in the client application tool.