Climate100: Scaling the Earth System Grid to 100Gbps Network

Progress Report for the Period
October 1, 2010 through December 31, 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 October 1, 2010 to December 31, 2010.

2. ESG Gateway and Data Node Deployment

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 successfully configured and deployed ESG Gateway (esg.nersc.gov) at NERSC. Further, we have updated the ESG software stack at esg-datanode.nersc.gov. We also have administered ESG services on the NERSC machines, and we developed and implemented necessary scripts for a production-level services.

In order to join the ESG Gateway consolidation, we have configured the required security components and generated certificates for the ESG federation. This step includes active participation in ESG community. We have tested dataset publication and created a project for IPCC AR4 CMIP3 datasets. ESG-NERSC is one of the Gateway nodes listed in ESG

federation. We have updated and released Gateway certificates to be updated in other Gateway systems for ESG federation.

2.1. Summary

Recent Progress includes

- Active participation in ESG community for deployment of ESG services,
- Upgrade of ESG Gateway software stack to version 1.2,
- Reconfiguration NERSC Gateway for ESG-NERSC specifics,
- Preparation for ESG federation testing and OpenID authentication.

Future Activities include

- Complete ESG federation of ESG-NERSC Data Node and Gateway with CMIP-3 dataset publications,
- Delegation of ESG systems support to NERSC Science Gateway team.

3. 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 and ANL testbed. This gave us the base for studying further data transfers over and RDMA over Ethernet (RDMAoE) on ANI testbed.

We have configured NERSC Magellan testbed machines for GridFTP over RDMAoE in collaboration with ANL, which lead 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 test machines. We have studied SoftiWARP from IBM, and started testing them on ANI test machines. We have continued our collaboration with ANI FTP100 group as well.

3.1. Summary

Recent Progress includes

- Studied and tested open fabric and data transfers over RMDA over IB with Mellanox ConnectX 2 cards,
- Studied RDMAoE, iWARP and SoftiWARP,
- Initiated a new collaboration with ANL GridFTP team (GridFTP over RoCE),
- Helped configuration of Chelsio driver at NERSC test machines (cxgb3toe),
- Prepared a RDMAoE data movement demonstration between NERSC and ANL, during Supercomputing Conference 2010,
- Configured OFED drivers on ANI test machines for Myricom cards (myri10ge),
- Obtained SoftiWARP from IBM, and experimented on ANI test machines,
- Continued collaboration with ANI FTP100 group.

Future Activities include

- Experiment SoftiWARP over WAN,
- Enhancement of simple client and server application, based on test results from SoftiWARP over WAN,
- Integration with FTP100 data transfer server with RDMAoE in the client application tool,
- Continue collaboration with ANI FTP100 group and ANL GridFTP team.

4. Adaptive tuning for efficient data movement

We have developed an advanced data movement tool with adaptive tuning capability. 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 tool provides a dynamic approach to set concurrency level to optimize data transfer operations. We have implemented our adaptive algorithm, and prepared all necessary components for a production-level data transfer tool.

Our dynamic tuning model has been explained in a recent paper, and we have presented our paper in the 22nd International Conference on Parallel and Distributed Computing and Systems (PDCS2010) in Marina Del Rey, CA. Our client tool implements the adaptive tuning algorithms for dynamically setting the concurrency level, given in the recent paper.

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. The client tool supports a modular architecture. Currently, we have developed drivers for GridFTP. We also plan to benefit from the client tool for our RDMA and iWARP experiments.

4.1. Summary

Recent Progress includes

- 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) tool,

Future Activities include

- Prepare test scenarios and tune ADT, test performance of the GridFTP driver in ADT,
- Perform experiments over WAN with ADT,
- Utilize implemented libraries for RDMA experiments, and explore possibilities for an efficient driver using iWARP/SoftiWARP
- Integration with FTP100 data transfer server with RDMAoE in the client application tool.

5. 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.

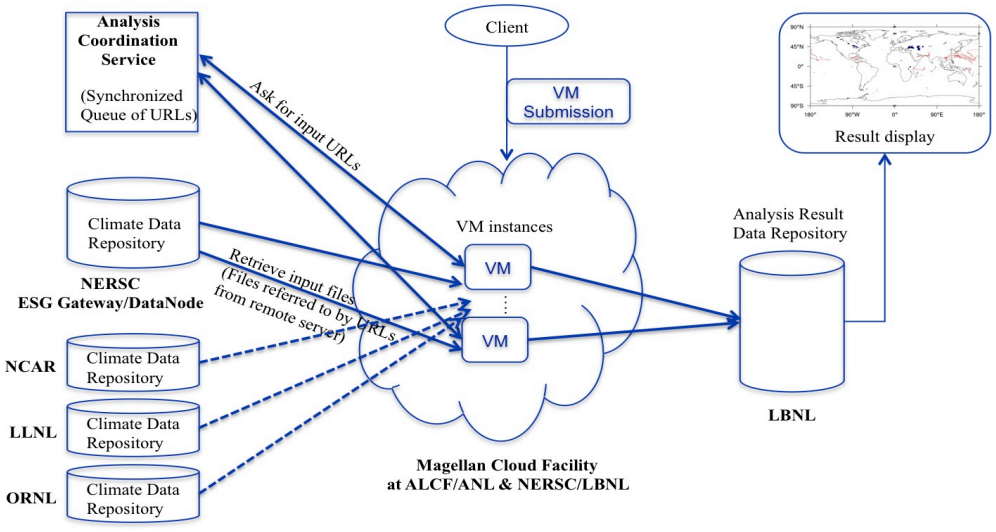

Figure 1: Analysis diagram using Magellan Cloud.

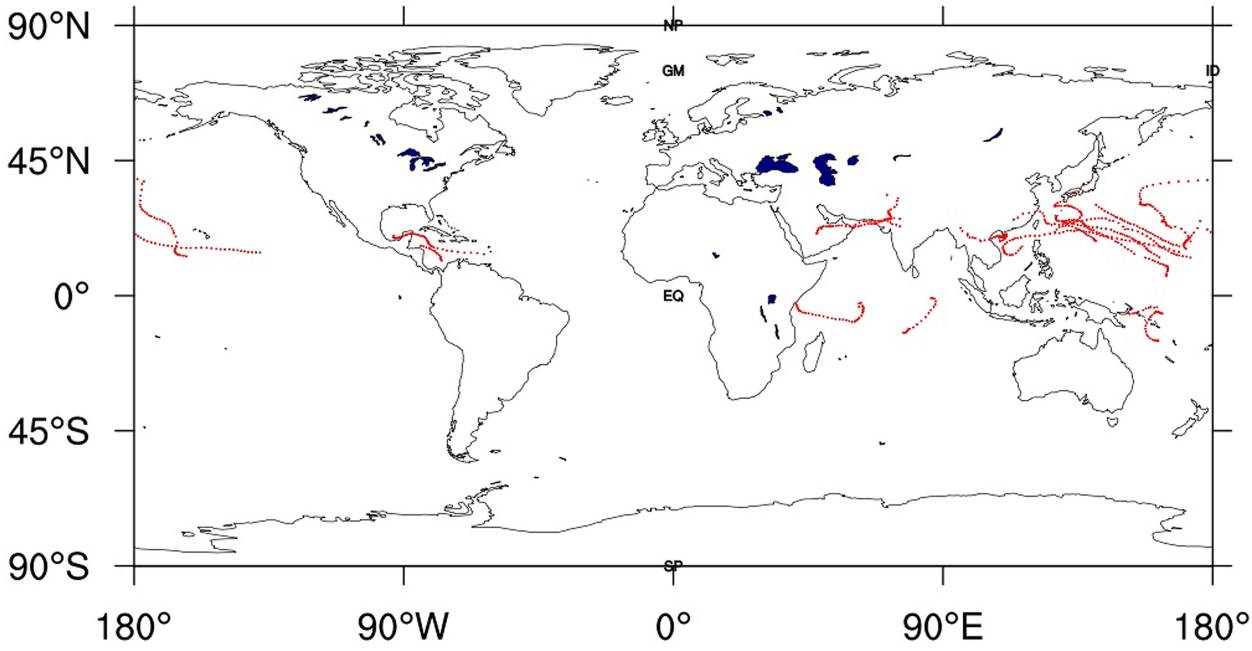

Figure 2: Results from the analysis of the simulated tropical storms, Sep. 1993, from fvCAM2.2 simulation encompassing 1979-1993.

6. Publications, Presentations and Awards

- "***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, the 2nd IEEE International Conference on Cloud Computing Technology and Science, 2010
- "***A Flexible Reservation Algorithm for Advance Network Provisioning***", **M. Balman**, E. Chaniotakis, A. Shoshani, **A. Sim**, ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis (SC'10), 2010.
- "***Finding Tropical Cyclones on Clouds***", D. Hasenkamp, **A. Sim**, M. Wehner, K. Wu, ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis (SC'10), 2010. Daren won the **Third place in ACM Student Research Poster Competition**.
- "***Adaptive Transfer Adjustment in Efficient Bulk Data Transfer Management for Climate Dataset***", **A. Sim, M. Balman, D. Williams**, A. Shoshani, V. Natarajan, The 22nd IASTED International Conference on Parallel and Distributed Computing and Systems (PDPS2010), 2010.
- "***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.