Real-time data transfer over 100 Gbps WAN networks using UDT based application suite

Renuka Arya
Center for Data Intensive Science
University of Chicago
Chicago, Illinois 60637
Email: rarya@uchicago.edu

Joshua S. Miller
Center for Data Intensive Science
University of Chicago
Chicago, Illinois 60637
Email: jsmiller@uchicago.edu

Mark W. Murphy
Department of Human Genetics
University of Chicago
Chicago, Illinois 60637
Email: murphymarkw@uchicago.edu

Joe Sislow
Center of Data Intensive Science
University of Chicago
Chicago, Illinois 60637
Email: madopal@uchicago.edu

Allison P. Heath
Center for Data Intensive Science
University of Chicago
Chicago, Illinois 60637
Email: aheath@uchicago.edu

Robert L. Grossman
Center for Data Intensive Science
University of Chicago
Chicago, Illinois 60637
Email: robert.grossman@uchicago.edu

***Abstract*—We plan to demonstrate how large scale providers of data ("data commons") can peer with each other so that a researcher using cloud based computing services at one of the sites can access data transparently over wide area 100 Gbps connections at one of the other sites. For this demonstration, we plan to connect the Open Science Data Cloud at the University of Chicago and 100 Gbps ESnet testbed at Oakland, California with a wide area 100 Gbps network. We also plan to access genomic data at the National Institute of Health/National Center for Biotechnology Information, European Bioinformatics Institute, University of Amsterdam and University of Edinburgh over 100 Gbps networks. We plan to use a variety of applications, including applications that leverage the UDT based protocol.**

I. Overview

In our SC14 workshop demonstration, we plan to analyze large genomic data sets across 100 Gbps wide area networks between data centers in Chicago, (Illinois), Oakland (California) and Hinxton (South Cambridgeshire). We plan to use genomic data from the 1000 Genomes Project and from the ENCODE Project for this demonstration. The goal is to show how researchers at one site can perform local genomic computations with data that is located at one of the other sites.

II. Problem Statement

With the increasing size of scientific datasets, the traditional model of downloading and analyzing datasets locally is no longer feasible in many cases. One approach is to build what the biomedical informatics communities are calling *data commons*, or more simply *commons*, in which large amounts of data, compute, and bioinformatics tools are co-located and integrated. We discuss some of the challenges building data commons and how geographically distributed data commons can interoperate over 100 Gbps networks.

III. Demonstration Overview

The demonstration will take place between the Open Science Data Cloud (OSDC) at the University of Chicago, the ESnet testbed at Oakland, California, the European Bioinformatics Institute (EBI) in Hinxton, UK, the National Institute of Health/National Center for Biotechnology Information (NIH/NCBI) in Bethesda MD, the University of Amsterdam in The Netherlands, and the University of Edinburgh in the UK.

Figure 1 contains the network diagram of the expected network connections between UChicago and its partners that will be used during the demonstrations at SC2014.

UDT (UDP-based Data Transfer Protocol) is a reliable UDP based application level data transfer protocol designed for moving large data sets over wide area high performance networks with large bandwidth delay products (udt.sf.net). UDT uses its own reliability control and congestion control mechanisms that provide a high speed alternative to TCP over wide area networks with high bandwidth delay products. It is highly configurable and also capable of accommodating various congestion control algorithms.

Our group has built an application suite around UDT to make it easier to use for data intensive applications. In this demonstration, we will be showing several applications built around UDT, including UDR, Udpipe and Parcel. UDR is a wrapper around rsync that enables rsync to use UDT. Udpipe is a bidirectional network application that uses pipes. Parcel is an alternative to SCP that can be used for secure and recursive directory transfers. All of these applications are open-source and are available on our GitHub page (https://github.com/LabAdvComp). These applications can be installed across various platforms, including Linux, BSD, and OS X.

We will be demonstrating how these applications can be used to provide transparent access to genomic data for researchers doing local compute at one of the sites *using data that is located at one of the other sites*. This is an example of what we call *data peering*, in which a data commons can provide transparent access to data to researchers located at

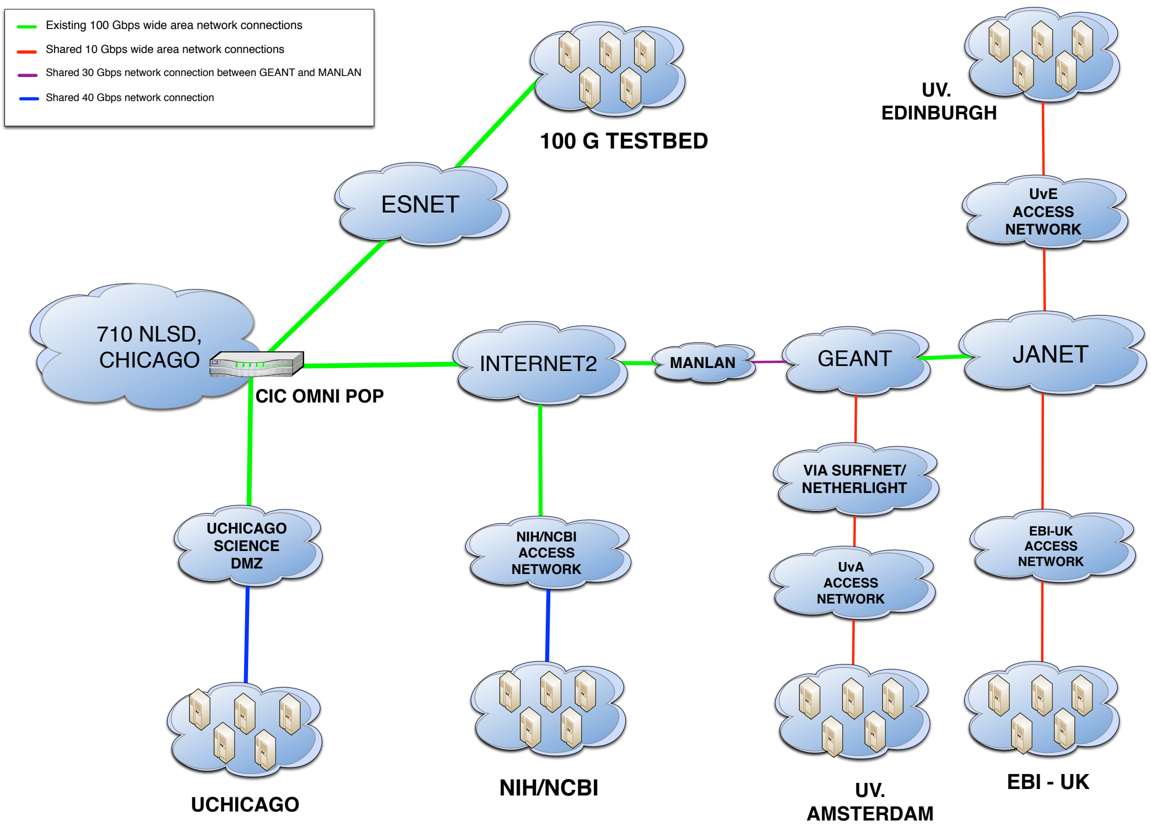

Fig. 1: Network diagram for the the proposed demonstration.

another site offering cloud based computing services.

IV. Network resources at SCinet and Other Networks

A demonstration will use high speed network connections between a data commons located at the University of Chicago and the other sites participating in the demonstration, but will not stream data to the SC 14 exhibit floor.

V. Conclusion

One emerging approach to providing researchers with the ability to analyze large genomic datasets is to co-locate genomic data, compute, and bioinformatics tools in commons and genomic clouds. In practice, there will be a number of such facilities and it is important that mechanisms be developed so that they can interoperate. In this demonstration, we will show one way that this can be done over 100 Gbps using a suite of protocols that use the UDT high performance data transport protocol. In this demonstration, we will use an infrastructure that supports jumbo frames and no intermediate firewalls.