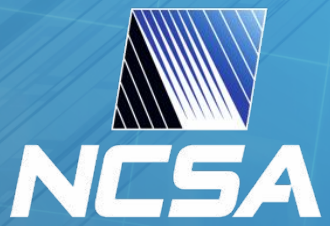
Globus and a Multi-Site, Multi-Petabyte Workflow

JD Maloney

Storage Engineer

National Center for Supercomputing Applications (NCSA)



Globus World 2016

National Center for Supercomputing Applications
University of Illinois at Urbana–Champaign

Terraref Project

- Goal to accelerate breeding and commercial release of high-yield bio energy sorghum
- Terabytes of imaging data will be captured by UAVs, tractor based systems, and indoor imaging systems
- Main instrument is a LemnaTec Field Scanalyzer located at the USDA Arid Land Research Station in Maricopa, AZ that will generate 4-5 TB/day many days of the year over 4 years

Challenges

- Data flow: Multiple sites across the country including Maricopa Ag Center, Danforth Plant Center (WashU), Kansas State, sequencing lab, others
- Bandwidth limits: Maricopa has 1Gb uplink to Phoenix Data Center, will need to saturate for many hours per day
  - Intermittent outages due to equipment failure, planned maintenance periods, etc. will require additional catch up
- Data availability: Scientists across the country will need access to data after it is run through the pipeline

Data Flow Solution with Globus

- Multiple sites sending data to NCSA for processing
- Pipeline launch needs to be done automatically as files arrive so a unified method of transfer was necessary so we could key off an event (completed Globus transfer)
- Creates a template that can be used if/when new data sources are added
- All raw data is also backed up to NCSA Nearline storage (HPSS), which will be done using Globus transfer mechanisms as well

Bandwidth Solution with Globus

- Limited network bandwidth at sites requires high utilization of the network bandwidth that is in place
- Transfers need to be threaded to push the LAN connection
  - This is especially important during times when we will be catching up on data ingest following an outage period
- Connection stability may not always be solid, Globus automatic tracking and retrying of transfers frees up our time, not having to baby sit data movement

Data Availability Solution with Globus

- Huge deliverable of project is to make data available to the scientific community, we'll be using Globus with a Shared Endpoint
- When science teams need to download bulk amounts of data, Globus makes it easy for them to do so without having to go through account creation on our systems
- Transfers to others can also be done quickly with the ability of Globus to parallelize the transfer automatically

Data Flow Diagram

Gather

Share

globus online

Compute

Backup

NCSA

Questions