DataONE Member Node Pilot Integration With TeraGrid[*] [†]

Nicholas C. Dexter
University of Tennessee, Knoxville
ndexter@utk.edu

John W. Cobb
Oak Ridge National Laboratory
cobbjw@ornl.gov

Dave Vieglais
University of Kansas
vieglais@ku.edu

Matthew B. Jones
National Center for Ecological Analysis and Synthesis, University of California, Santa Barbara
jones@nceas.ucsb.edu

Mike Lowe
Indiana University-Purdue University Indianapolis
jomlowe@iupui.edu

ABSTRACT

The NSF DataONE [1] DataNet project and the NSF TeraGrid [2] project have initiated a pilot collaboration to deploy and operate the DataONE Member Node software stack on TeraGrid infrastructure. The appealing feature of this collaboration is that it opens up the possibility to add large scale computing as an adjunct to DataONE data, metadata, and workflow manipulation and analysis tools. Additionally, DataONE data archive and curation services are exposed as an option for large scale computing and storage efforts such as TeraGrid/XSEDE. With this joint effort, DataONE also brings an open, persistent, robust, and secure method for accessing Earth sciences data collected by science communities such as The National Evolutionary Synthesis Center's Dryad [3], The Ecological Society of America's Ecological Archive [4], NASA's Distributed Active Archive Center at the Oak Ridge National Laboratory [5], the USGS's National Biological Information Infrastructure [6], the Fire Research & Management Exchange System [7], the Long Term Ecological Research Network [8], and the Knowledge Network for Biocomplexity [9].

Beginning with an April 1st, 2011, allocation, the DataONE Core Cyberinfrastructure Team has been working with the IU Quarry [10] virtual hosting service, and more generally with the TeraGrid data area, on this pilot implementation. The implementation includes multiple virtual servers in order to test different reference implementations of the common DataONE Member Node RESTful web-service functions [11]. These implementations include implementation as a Metacat server [12], as well as a Python Generic Member Node developed by DataONE [13]. The implementations will also mount TeraGrid-wide global storage services (DC-WAN [14] and Albedo [15]) and thus allow integration of input and output of large scale computational runs with wide area archival data and metadata services.

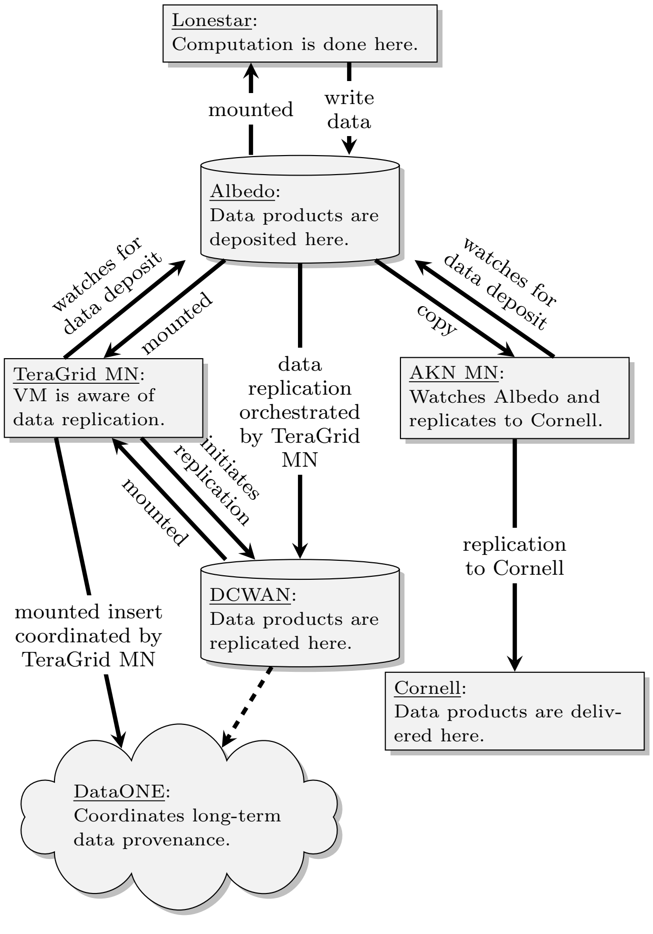
Lonestar:
Computation is done here.
mounted
write data
Albedo:
Data products are deposited here.
watches for data deposit
mounted
watches for data deposit
copy
TeraGrid MN:
VM is aware of data replication.
data replication orchestrated by TeraGrid MN
AKN MN:
Watches Albedo and replicates to Cornell.
initiates replication
mounted
replication to Cornell
DCWAN:
Data products are replicated here.
mounted insert coordinated by TeraGrid MN
Cornell:
Data products are delivered here.
DataONE:
Coordinates long-term data provenance.

Figure 1: Workflow diagram of data production, deposition, replication, and insertion into DataONE using the TeraGrid Member Node.

[*] Oak Ridge TeraGrid efforts sponsored by the US National Science Foundation under Cooperative Agreement OCI-0504074. This work was performed at Oak Ridge National Laboratory, managed for the US D.O.E. by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725.

[†] DataONE efforts sponsored by the US National Science Foundation under Cooperative Agreement OCI-0830944.

The pilot Member Node provides access to the DataONE Service Programming Interface (SPI), which allows Investigator Toolkit (ITK) Software to retrieve datasets relevant to TeraGrid/XSEDE analyses. Simulation and analysis data products which have been inserted into DataONE can later be accessed through Excel, MATLAB, and R interfaces [16]. Additionally, files tagged through Morpho [17] or uploaded through the web interface can be found using a FUSE [18] driver for tagged filesystem access. Furthermore, tools can be developed using the DataONE SPI and ITK client libraries for command-line, Java, and Python, which contain DataONE service call implementations. These tools would enable dynamic content retrieval and data mashups, resulting in novel analysis that use TeraGrid/XSEDE compute and storage resources. These new results, and the workflows that create them, could then be saved and archived in DataONE in a common workflow interchange format, for example Kepler [19] and VisTrails [20].

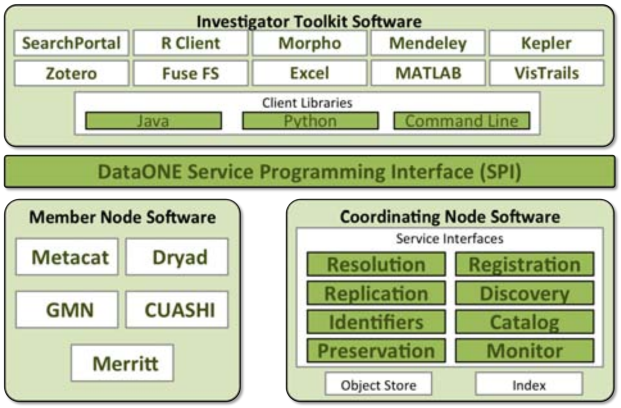

Figure 2: **Investigator Toolkit Software items and libraries.**

Initial results show a smooth and easy Metacat configuration implementation due mainly to a concise and clean interface design both on the part of DataONE and the Quarry VM infrastructure. We are also testing mounted TeraGrid-wide storage services such as DC-WAN and Albedo, and will initiate data replication from TeraGrid external DataONE nodes to Quarry member nodes. Additionally, we will present preliminary results from scale and performance of replication and science metadata consistency traffic.

We are also exploring using the Quarry DataONE Member Node implementation for production use to simplify and accelerate analysis product transport to the Cornell lab of Ornithology [21] for the 2012 eBird/State of the Bird [22] statistical analyses that will be conducted on TeraGrid resources. As data products are generated by the analysis runs on Lonestar and deposited to a mounted Albedo share, the Quarry virtual machine running the DataONE software stack will monitor output directories and initiate replication to the Data Capacitor. This staging area will hold the data products until they are able to be inserted into DataONE by the TeraGrid Member Node.

As the DataONE Member Node and Coordinating Node software implementations are still young in their development cycle, we have planned for the workflow described in Fig. 1 to carry redundancy in transporting the data to Cornell. Once the DataONE service is fully operational, starting in 2012, we will be able to directly insert science data and metadata generated on the TeraGrid/XSEDE into DataONE for long-term curation and management. In continued collaboration, we hope to provide access to DataONE services for future data-intensive research projects on the TeraGrid/XSEDE.

1. REFERENCES

[1] DataONE, http://www.dataone.org/about

[2] TeraGrid, https://www.teragrid.org/web/about/

[3] Dryad - The National Evolutionary Synthesis Center, http://www.datadryad.org/

[4] The Ecological Society of America's Ecological Archive, http://www.esapubs.org/archive/

[5] The National Aeronautics and Space Administration's Distributed Active Archive Center, http://daac.gsfc.nasa.gov/

[6] The United States Geological Survey's National Biological Information Infrastructure, http://www.nbii.gov/

[7] The United States Geological Survey's Fire Research and Management Exchange System, http://frames.nbii.gov/

[8] The United States Long Term Ecological Research Network, http://www.lternet.edu/

[9] The Knowledge Network for Biocomplexity, http://knb.ecoinformatics.org/index.jsp

[10] Quarry | Cyberinfrastructure | Indiana University Pervasive Technology Institute, http://rc.uits.indiana.edu/ci/systems/quarry

[11] Fielding, R. T. and Taylor, R. N. 2002. Principled design of the modern Web architecture. ACM Trans. Internet Technol. 2, 2 (May. 2002), 115-150. DOI=http://doi.acm.org/10.1145/514183.514185.

[12] Metacat: Metadata and Data Management Server, http://knb.ecoinformatics.org/software/metacat/

[13] Components of the DataONE Infrastructure, http://mule1.dataone.org/ArchitectureDocs-current/implementation/components.html

[14] Data Capacitor | Indiana University Pervasive Technology Institute, http://pti.iu.edu/dc

[15] Wide area data resources on the TeraGrid - Albedo, http://kb.iu.edu/data/bauy.html#albedo

[16] DataONE Investigator Toolkit Overview http://mule1.dataone.org/ArchitectureDocs-current/design/itk-overview.html

[17] KNB Data :: Morpho Data Management Software http://knb.ecoinformatics.org/morphoportal.jsp

[18] Filesystem in Userspace (FUSE) http://fuse.sourceforge.net/

[19] Kepler https://kepler-project.org/

[20] VisTrails http://www.vistrails.org/index.php/Main_Page

[21] Cornell Lab of Ornithology, http://www.birds.cornell.edu/

[22] The State of the Birds Report, http://www.stateofthebirds.org/