EarthCube Cloud Services and Data Management

Chaitan Baru, SDSC
Tim Ahern, IRIS
Fran Boler, Charles Meertens, UNAVCO
Ian Foster, Ravi Madduri, University of Chicago

SUMMARY

Geoscience data archives have established a sound reputation for providing reliable and efficient access to the increasing amounts of data that is being collected. The pressure will only increase in terms of having to manage increasing amounts of data at the same levels of service to the end user. We recognize an opportunity for exploiting recent trends in cloud computing to assist geoscience data archives with both the infrastructure and business process aspects of data management. In particular, the San Diego Supercomputer Center (SDSC) has recently announced its SDSC Cloud for storage of academic, research datasets. There is also an effort underway to provide a solution for meeting NSF's Data Management plan using this cloud infrastructure. Also, the research data management as a service (Globus Online) solution developed at the University of Chicago's Computation Institute (CI) is gaining a lot of traction in the research community

Preliminary discussions among IRIS, UNAVCO, SDSC, and CI lead us to believe that there is, indeed, an opportunity for organizations like IRIS and UNAVCO—and perhaps other geoscience data facilities—to exploit the capabilities of the SDSC Cloud for modest-cost, high-quality storage and computing, and Globus Online's research data management as a service for automation of the operational processes required for data management. Initially, this can be simply for data backup, and also for failover. This white paper provides some of the details of what could be achieved with this approach.

1. Cloud Computing

Cloud computing concepts have become extremely popular across the information technology industry. A key aspect of cloud computing is virtualization of services so that the client (end user) is not aware of where the service is provided. This virtualization can be at the application level itself, e.g., email on the Web, or at the infrastructure level, e.g., servers and storage. The recently announced SDSC Cloud initially provides storage capability, and will soon also provide computational capability.

1.1 SDSC Cloud Storage Services

Access to storage objects is via URLs, which provides a level of abstraction. Data may be replicated within the system, including replicated to offsite locations. The SDSC Cloud implementation is based on the open-source OpenStack software and consists of 5.5PB of disk. The data is replicated within the cloud, and users can also request for offsite replication of data.

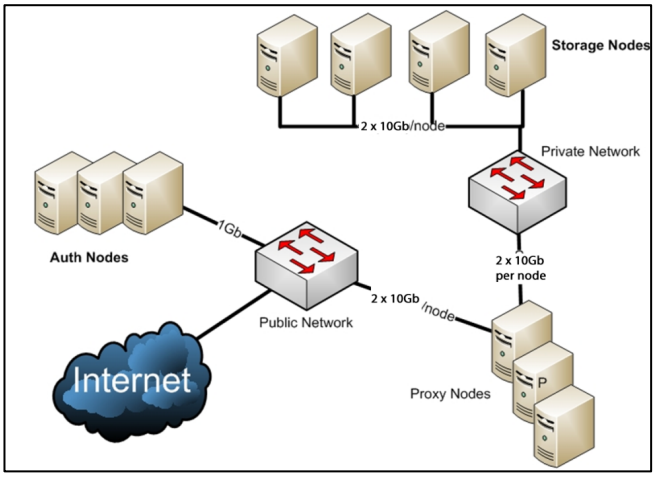

Figure 1. OpenStack-based SDSC Cloud

SDSC's Cloud Storage provides academic and research partners a convenient and affordable way to store, share, and archive data, including extremely large data

sets. The object based storage system and multiple interface methods make the system easy to use for the average user, but also provide a flexible, configurable, and expandable solution to meet the needs of more demanding applications.

Utilizing the OpenStack Swift Object Storage software, files (also known as objects) are written to multiple physical storage arrays simultaneously, ensuring at least two verified copies exist on different servers at all times. The system provides continuous automatic data verification for durability. If needed, a 3rd copy can be replicated to an off-site storage partner facility, further reducing the potential for data loss.

Files of any size can be stored in the cloud, from small personal document collections to multi-terabyte backup sets routed directly to the cloud by Rackspace or S3 API compliant applications. Cloud Backup package solutions are also available, using SDSC's CommVault Backup service. More information about the SDSC Cloud is available at http://cloud.sdsc.edu.

Once objects are in the cloud they are immediately available over the web to other users, the public, or to only a restricted set of users, using an access control mechanism. By assigning a URL to each file or container created, SDSC Cloud storage provides a simple way for researchers and other users to effortlessly share any amount of data.

2. Software as a Service (SaaS) and Globus Online

Software as a Service (SaaS) is a technique in which software functionality is made available to users as a service without users installing anything on their machines. SaaS has gained a lot of traction in recent years with popularity of services like Gmail, Google docs, and Salesforce increasing.

The Globus Online (GO) hosted data movement service leverages software-as-a-service (SaaS) methods to provide fire-and-forget file transfer, automatic fault recovery, and high performance, and to simplify the use of multiple security domains while requiring no client software installation to submit and monitor requests. Globus Connect provides for easy installation of a GridFTP server on a user's machine, so that it can participate in GO transfers. The result is a system that provides automatic fault recovery, high performance, and easy-to-use security, to virtually all researchers and facilities.

Since its launch in November 2010, the number of Globus Online users has grown to 3000. Globus Online has been adopted as the preferred data movement solution by centers such as NERSC, and has been integrated into earth science data systems such as Earth System Grid. More information about Globus Online is available at www.globusonline.org.

3. Using the SDSC Cloud and Globus Online

Preliminary discussions have occurred among IRIS, UNAVCO, and SDSC about using the SDSC Cloud service initially for offsite data backup and replication. SDSC and Globus Online have also discussed the use of GO software as a service approach. File objects and other backup files could relatively easily be backed up to the SDSC Cloud from IRIS or UNAVCO using Globus Online. The initial sense is that this might lead to cost savings at data archive sites, not only in terms of storage hardware but also in terms of the time of technical and management personnel.

Initial experiences with Globus Online suggest similar opportunities for cost savings via outsourcing of otherwise labor-intensive tasks. For example, it is common for groups to spend person-months developing software to manage data movement from experimental facilities to archives, and then spend further time managing these services. Globus Online can reduce both startup and ongoing costs to close to zero.

Performing simple backup of data, while useful and important, does not realize the full potential of what could be achieved. There is interest in treating the SDSC Cloud as a failover environment, which means that it should be possible to serve the data requests served by the IRIS/UNAVCO sites using the SDSC Cloud. IRIS and UNAVCO utilize a set of Web services to

access their data. Thus, the primary requirement for achieving this is to host the same set of Web services in the SDSC Cloud, in conjunction with the data that is stored there. SDSC is already planning to deploy a Eucalyptus-based compute cloud that would sit alongside the storage cloud. This would be the ideal setup for then deploying the necessary services as part of the SDSC Cloud.

3.1 Linking to Commercial (public) Clouds

Cloud environments such as the Amazon Web Services are referred to as "public" clouds (even though they are run by private organizations), since their only barrier to entry is the ability to pay for the service. Anyone willing and able to pay can use the service.

At SDSC, we have considered a couple of options that would allow intermingling of such public clouds with the SDSC Cloud. Since the OpenStack software used in the SDSC Cloud is fully compatible with the AWS S3 interfaces, it would be quite simple to move references from one cloud to the other. Thus, users who may already be using the AWS cloud can make the transition to the SDSC Cloud relatively painlessly. The reverse is also true—users who decide to move from the SDSC Cloud to AWS or other similar environments would be able to do that easily. More important, there may be various reasons why an application may wish to use both clouds, i.e., distribute their data across the SDSC Cloud and some other cloud. This would also be possible as long as the interfaces being used are S3-compatible. The reason for using more than one cloud environment could be for cost and/or performance reasons, and for achieving greater robustness for the data.

Globus Online has prototyped support for Amazon S3 access, and can easily be extended to allow for flexible distribution of data among multiple cloud services, and dynamic selection of data source.

4. NSF Data Management

The above setup indirectly serves the NSF Data Management plan since, as NSF data archives, IRIS and UNAVCO themselves adhere to the NSF Data Management plan, and the system described above supports those operations. However, the SDSC Cloud can also be utilized directly to serve the needs of the NSF Data Management Plan. As part of its campus Research Cyberinfrastructure Initiative, UC San Diego has established a Data Curation activity as a joint activity between the UCSD Library and SDSC.

As described at the UCSD Data Curation website (http://rci.ucsd.edu/services/data-curation.html), data curation involves managing data to ensure they are fit for contemporary use and available for discovery and reuse. Archiving and preservation are subsets of the larger curation process, which is a much broader, planned, and interactive process. The curation process is well understood by the UC San Diego Libraries and has long been applied to non-digital scholarly materials. Archiving refers to ensuring that data are properly selected, appraised, stored, and made accessible. The logical and physical integrity—including security and authenticity—of the data are maintained. While preservation refers to ensuring that items or collections remain accessible and viable in subsequent technology environments. The overall set of services provided as part of the UCSD research cyberinfrastructure are described at http://rci.ucsd.edu.

The capabilities provide by the SDSC Cloud and the services offered via the UCSD data curation activity, together embody a complete set of data management services that can be utilized by any EarthCube-based activity for their data management. Globus Online provides a natural framework for delivering those services to a large number if people.

5. Collaborations

This activity was initiated by a discussion among IRIS, UNAVCO, SDSC, and CI. The SDSC Cloud and Globus Online are both available for use by any research/academic project, thus the potential for collaborations is limitless. This activity can serve as the cloud platform and data management plan for EarthCube in general.