DuraCloud, Chronopolis and SDSC Cloud Integration

Andrew Woods
DuraSpace
165 Washington Street, Suite #201
Winchester, MA 01890
011-1-781-369-5880
awoods@duraspace.org

Bill Branan
DuraSpace
165 Washington Street, Suite #201
Winchester, MA 01890
011-1-781-369-5880
bbranan@duraspace.org

David Minor
UC San Diego
San Diego Supercomputer Center
La Jolla, CA 92093
011-1-858-534-5104
minor@sdsc.edu

Don Sutton
UC San Diego
San Diego Supercomputer Center
La Jolla, CA 92093
011-1-858-534-5085
suttond@sdsc.edu

Michael Burek
National Center for Atmospheric Research
P.O. Box 3000
Boulder, CO 80307
011-1-303-497-1202
mburek@ucar.edu

ABSTRACT
In this paper we describe the interaction of three different systems: DuraCloud, a cloud-based service provider, Chronopolis, a digital preservation network, and the San Diego Computer Center's cloud service. This interaction is targeted to developing a new storage and preservation service available to a wide range of users.

General Terms
Algorithms, Management, Design, Experimentation

Keywords
Digital preservation, cloud storage, integration

1. INTRODUCTION
Since late 2011, Chronopolis[1], the San Diego Supercomputer Center (SDSC) Cloud Storage[2], and DuraCloud[3] have been collaborating in an effort to add another layer of reliability to the field of distributed digital preservation. Each of our services have been designed to address a set of needs within the preservation community. Together, we have developed a single service that combines the archiving sensibilities of Chronopolis, the cost-effective, academic cloud storage of SDSC, and the provider-neutral access and preservation capabilities of DuraCloud. This paper will describe the details of the integration as well as follow-on activities. We will start, however, with a brief introduction to each of the constituent pieces.

2. INTEGRATION OVERVIEW
The DuraCloud/SDSC/Chronopolis integration was conceived of as a way to bridge the cost-effective dark archive, Chronopolis, with the online, dynamically accessible, provider-independent, preservation platform of DuraCloud. Prior to this effort, DuraCloud provided a mediation layer over three underlying commercial cloud storage providers: Amazon S3, Rackspace CloudFiles, and Microsoft Azure. The goals of the integration were to (1) add an academic cloud store (SDSC Cloud Service) to this list of providers supported by DuraCloud as well as to (2) enable DuraCloud users to replicate content to a geographically distributed, TRAC certified, preservation network (Chronopolis). Among other benefits, this integration supports the preservation strategy of distributing content across multiple geographic, platform, and administrative domains.

The first step in the integration process was to ensure that DuraCloud had the ability to store and retrieve content from the

[1] http://chronopolis.sdsc.edu

[2] http://cloud.sdsc.edu

[3] http://duracloud.org

SDSC Cloud Service. This initial integration required very little effort due to the fact that the SDSC Cloud exposes what is emerging as the standard cloud storage interface, OpenStack's Object Storage API. Since this is the same API offered by an existing storage provider supported by DuraCloud, Rackspace Cloudfiles, the connector code was already in place. As a result, adding SDSC as an additional underlying storage provider to DuraCloud was as simple as providing a new connection URL.

While the integration between DuraCloud and SDSC Cloud was simple, the connection to Chronopolis required more care. The model of programmatic interaction with Chronopolis is very different from that of the common cloud storage providers, and as such a means of communication between the two systems needed to be defined. The final approach defines an asynchronous RESTful integration. Over the course of several months, a joint team with representation from all three organizations (SDSC, Chronopolis, and DuraSpace) created the set of calls required in the REST API. This work defined a series of steps which would be used to move content from the SDSC Cloud to Chronopolis and back as needed, all managed by calls made from DuraCloud to the REST API.

To move content from DuraCloud to Chronopolis, DuraCloud stores content in one or more SDSC cloud storage containers then sends a request to Chronopolis to read content from those container(s). Part of this request is a manifest file detailing each content item to be transferred. Chronopolis then pulls the requested content into its data centers and compares the file checksums with the provided manifest to ensure that all content was pulled successfully. Once the transfer is validated the objects are arranged in BagIt format[4] and ingested into the Chronopolis system. The SDSC cloud service also allows custom meta name-value parameters to be assigned to objects. Using the manifest file, Chronopolis queries the SDSC cloud for any custom meta parameters and stores them with the ability to restore them if a retrieval is requested.

To retrieve content from Chronopolis, DuraCloud requests the restoration of all (or a subset) of the content back to an SDSC container, and Chronopolis performs the work of transferring the content from its data centers back to the SDSC Cloud. The inter-system communication is achieved via a REST server hosted by Chronopolis that receives requests from DuraCloud. (It should be noted that the Chronopolis REST server does not need to know that the client is a DuraCloud process. In this way, it is expected that other external systems could integrate with Chronopolis using the same methods.) The Chronopolis process behind the REST server is granted read/write access to one or more SDSC Cloud storage containers that are owned by DuraCloud.

The following three scenarios are covered by this integration: (1) A DuraCloud user wishes to have a snapshot of their content replicated to the Chronopolis preservation network. (2) A DuraCloud user wishes to restore a previously snapshotted collection from Chronopolis back to DuraCloud. (3) A DuraCloud user wishes to restore a single item of a previously snapshotted collection from Chronopolis back to DuraCloud.

3. NEXT STEPS

Due to the initial success of the DuraCloud/SDSC/Chronopolis integration a series of follow-on tasks are in process. Several end-to-end tests have proven the communication and data flow patterns. The objectives of the second round activities are to tease out any performance or technical issues as well as to discover and add any usability features that will ultimately ready the integrated system for production use.

On the technical side the next tasks will address security, inter-process communication, and performance improvements. The team will be layering security over the REST server in the form of SSL coupled with Basic-Auth. Beyond security, the API will also be extended to support a more robust back-flow communication mechanism. For example, after a content restore from Chronopolis to DuraCloud, if DuraCloud detects an error in the resultant file(s) an API method should be available to communicate that error back to the Chronopolis system. From a performance perspective we will be stressing the system to ensure that response times do not suffer at scale. We are in the process of staging a series of tests to back up and restore half a million data files up to one gigabyte in size.

As a step towards validating the capability and usability design of the integration, a set of interested institutions using DuraCloud will be invited to participate in beta testing. From the beta testing phase we expect to uncover any use cases that were not revealed in the earlier testing. Additionally, we hope to gain feedback on the general process flow and user interface. Assuming a successful period of beta testing, the expectation is that the SDSC and Chronopolis services nested under DuraCloud will be made publicly available as a production offering in the Fall of 2012.

In summary, the recognition that cloud-based, distributed, digital preservation is an increasingly emerging need, the three related technologies of Chronopolis, SDSC Cloud Service, and DuraCloud have undertaken the joint effort to provide the preservation and archiving communities with an additional layer of assurance. Not only will users be able to now select an academic cloud store in addition to existing commercial stores, they will also have the option to mirror their holdings through a dark archive network spanning across North America.

[4] https://wiki.ucop.edu/display/Curation/BagIt