Executive Summary

The DSpace Foundation and Fedora Commons joined together in July 2009, as a unified non-profit organization named DuraSpace. The first new offering to emerge from the new organization is a new service and cloud based technology named DuraCloud. DuraCloud is targeted at academic libraries, national and public libraries, universities, and other cultural heritage organizations that are responsible for digital preservation and ensuring perpetual access to digital content.

DuraCloud is a hosted service and open technology, which allows organizations and end users to effectively utilize public cloud services. The DuraCloud platform builds upon existing cloud infrastructure for the purpose of increasing durability and re-use of digital content. A key design feature of DuraCloud is to leave the basics of pure storage to those who do it best (storage providers) and to overlay storage solutions with additional functionality that is essential to ensuring long-term access and ease of use. The service provides baseline functionality that begins with the ability to replicate and distribute content across multiple cloud providers. It adds value over and above storage by enabling the deployment of services to support access, preservation, re-use, and sharing of content stored in the cloud.

DuraSpace has been funded by the NDIPP program sponsored by Library of Congress, to run a pilot program testing and deploying the DuraCloud technology and service around several use cases with pilot partners. The pilot program commenced in July 2009 and will finish in Spring 2010. There are three partners participating in the pilot program: New York Public Library, Biodiversity Heritage Library and WGBH. Each of the pilot partners has identified a particular use case for DuraCloud of interest to their institution, and content for testing in the pilot. Data includes images, video, and text in a variety of formats. Upon completion of the pilot program the DuraSpace organization intends to release the open technology and launch the service.

Description of the DuraCloud Service

DuraCloud is designed to act as a mediator between institutional or end-user applications and a variety of 3rd party cloud services. The purpose of the service is to provide a trusted intermediary that offers different levels of service aimed at making digital content (1) durable - meaning it is accessible for long periods of time, providing permanence and (2) usable - meaning that it can be re-exposed or dynamically transformed to fit within in a variety of application contexts. From the technology

standpoint, the DuraCloud team is currently developing the service to the specifications described in a set of architectural blueprints. Key features of the service are:

a. **Transparently push content to multiple 3rd party storage providers so all users can take advantage of cost effective internet based storage.** The envisioned service will provide the ability to send content to one or more underlying storage providers. The idea is that DuraCloud will not directly host content, meaning that the service, itself, is NOT a big data center. Instead the service will store only what's necessary to mediate storage and retrieval of content with 3rd party storage providers.

b. **Storage configuration:** the envisioned service can be configurable so that users/customers can select to have their content sent to ***one or many*** storage providers via our mediating service. Note, this is motivated by the well recognized principle in the archiving and preservation communities that best practice for longevity of content is to ensure that content is stored multiple times, and ideally in different systems.

c. **Value added services**: the envisioned service will add value to what the underlying storage providers offer. We will be particularly focused on providing services that enable longevity of content and facilitate flexible use/re-use. These services will be provided as a sort of Chinese menu where users can choose to subscribe to them or not. Current services planned to be deployed at launch include:

    1. **Preservation support services**: Replication, file format transformation, and bit integrity checking
    2. **Access and reuse services**: Image viewing and editing, video streaming and editing, faceted browse and search

d. The service will be built on open source technologies. The envisioned service will be built as open source software, keeping with the open source principles promoted by both Fedora Commons and the DSpace Foundation.

e. **Hosted or Run-Your-Own:** The DuraSpace organization will run the DuraCloud technology as a web based hosted service. Since the service will be built on open source technologies, it can be deployed so that others can pick it up and run their own local instance of the service to create their own hybrid cloud or cloud consortium network.

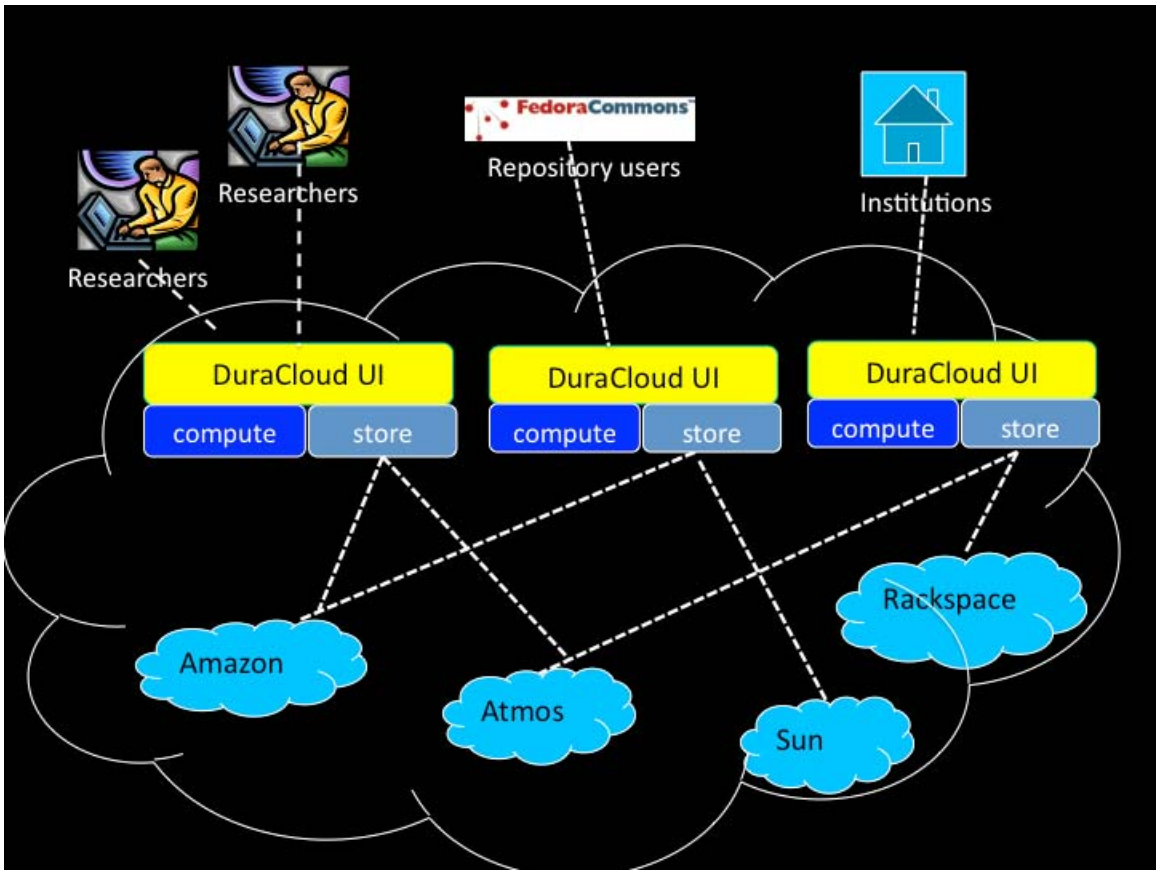

Figure 1: Notional View of DuraCloud

This illustration shows multiple DuraCloud applications running in a cloud environment. There are three main components to the DuraCloud architecture, the user interface, the compute layer where services are deployed, and the storage layer that manages where content is stored within the cloud provider network A DuraCloud instance can support a single institution, where they are the sole users, or can be used shared among multiple institutions are end users to support collaborative projects.

Pilot program

The goal of the pilot program is to fully test and further develop the DuraCloud technology by working with partners that have real data and use cases. For the first phase we have selected three pilot partners from a candidate pool of twelve organizations. The outcome of a successful pilot program will be the public launch of the DuraCloud service, which will include a set of open source applications and tools that will include the following capabilities:

- Ability to replicate content to multiple cloud providers through a single web interface (multiple copies, multiple geographic areas, multiple administrations)

- Data integrity checking and monitoring across multiple cloud providers
- Ability to put, get, and move content between a local repository and DuraCloud through extensions to the DSpace and Fedora software.
- Ability to get, put, move content across multiple cloud providers through the DuraCloud web interface

* Ability to run a set of compute services on top of content in the cloud through the DuraCloud web interface (such as data transformations, data mining, indexing, media services, and hosting).

Goals of the Pilot Program

Our intent is to develop the DuraCloud service, in concert with pilot partners, to achieve a state where there is demonstrable evidence that the DuraCloud technology works and provides cost/benefit advantages compared to other IT solutions. At a minimum we must demonstrate (1) a running service with a management console supporting provisioning, monitoring and billing capabilities, (2) initial release of open source software to allow for the development of intra/consortium cloud networks (3) ingest and replication of content to at least two cloud providers, (4) plug-ins for DSpace and Fedora repositories, (5) multiple preservation and access services able to run contiguously in DuraCloud (6) easy access to stored content via a simple API, and (7) reliability and cost effectiveness. We believe that the demonstration of this set of capabilities is a necessary condition to obtaining commitment from pilot partners to become early adopters of the DuraCloud service.

DuraSpace has met with several major commercial cloud providers to discuss our plans. All providers were intrigued with the DuraCloud value proposition and they understood how we could add value to what they offer. Each was interested in working with us in developing our first prototype and we signed Non-Disclosure Agreements with each organization to enable this work to commence. Our goal is to develop relationships with cloud providers that will enable us to evaluate their technologies, advocate for features that serve our customers, and negotiate lowest costs. Since the DuraSpace organization will host the DuraCloud service, we are positioning as a trusted intermediary to third party cloud services. We thus believe it is essential for our non-profit organization to have access to knowledge of cloud provider offerings. This will enable use to assess fitness-of-purpose of cloud provider services so that we can ensure that our overall DuraCloud goals are achieved and customer requirements are met.