DuraCloud

A service provided by DuraSpace

Sandy Payette and Michele Kimpton
The commercialization of data storage and services contributes an important component of the data CI environment needed to harness the potential of our information-rich world. However, private-sector storage and services are not the solution to all digital data needs. For some digital data considered to be “in the public interest” (such as census data, official records, critical scientific data collections, and a variety of irreplaceable data), a greater level of trust, monitoring, replication, and accountability is required to minimize the likelihood of loss or damage and ensure the data will be there for a very long time. For such community data sets, stewardship by a trusted entity (such as libraries, archives, museums, universities, and institutional repositories), whose mission is the public good rather than profit, is generally required.
Vision: Federated Repositories and Cyberinfrastructure

Heaven

DuraCloud
What About the Cloud?

A style of computing where massively scalable IT-related capabilities are provided “as a service” using Internet technologies to multiple external customers. (Gartner, 6/08).
“The services themselves have long been referred to as Software as a Service (SaaS). The datacenter hardware and software is what we will call a Cloud. When a Cloud is made available in a pay-as-you-go manner to the general public, we call it a Public Cloud; the service being sold is Utility Computing.

We use the term Private Cloud to refer to internal datacenters of a business or other organization, not made available to the general public.”

Examples of Cloud Services

• Software as a Service (SAAS)
  – e.g., Google Apps

• Cloud Computing
  – e.g., Amazon Elastic Compute Cloud (EC2)

• Cloud Storage
  – e.g., Amazon Simple Storage Service (S3)
Cloud Services

Elastic web-based infrastructure for storage and compute
DuraCloud Proposition

Trust and durability in the cloud
What have we learned from our users?
Challenges
(From our communities)

Digital preservation and archiving is hard to achieve, even just basic replication

Easy and elastic provisioning of shared infrastructure (also across institutions!)

Robust compute environments for large indexing jobs, data mining and analysis of large datasets

Making digital content more accessible and usable to researchers
DuraCloud - basics

- Replicate to multiple storage providers
- Replicate to multiple geographic areas
- Monitor and audit digital assets
- Compute services in cloud next to content

Hosted by DuraSpace not-for-profit org
Partnerships with cloud providers
“Pay for use” for services and storage
DuraCloud

*Trusted management of and access to durable digital assets in the cloud*

DuraSpace

Mediating Service

Durable Store Service Layer
- (Foundation Service)
- (Commercially Operated)
- (Locally Operated)

Foundation provided Software and Configuration (Open Source)

Compute and Storage [Cloud/Grid] Providers

- Amazon
- Google
- EMC
- Local IT
- Sun
- Microsoft
- Offline Backup
DuraCloud
Basic Architecture

DuraCloud – Notional View

- DuraCloud.org
  - Website and Account Administration

- DuraCloud Application
  - Dashboard
  - Access
  - Deposit
  - Replicate
  - Admin

- Customer Primary Data Store
- Customer Secondary Data Store
- Primary Compute Cloud Provider (Amazon EC2 and others)
- Primary Storage Cloud Provider (Amazon S3, others)
- Secondary Storage Cloud Provider

Replicates to
Stores in
DuraCloud

Making the Cloud Durable

Use Cases

Partnerships and Pilots
Digital preservation is essential but difficult to implement

- Tools and processes unproven
- Limited IT support
- Capital expenditures limited
- Task can be overwhelming (replication, migration, emulation, etc.)
Challenge

Barriers to making digital content more accessible and useful to researchers

• Systems not interoperable
• Heterogeneous applications/platforms
• Lack of commons standards
• Non-elastic compute capability
Advantages – Cloud Services

- Flexibility
- Scalability
- Pay for use
- Easy to implement
- Cost
Economies of Scale and Cost

Public cloud providers drive cost down through scale, location and virtualization technology

<table>
<thead>
<tr>
<th>Technology*</th>
<th>Cost Medium Datacenter</th>
<th>Cost Large Datacenter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network</td>
<td>$95 per Mbit/sec/mo</td>
<td>$13 per Mbit/sec/mo</td>
</tr>
<tr>
<td>Storage</td>
<td>$2.20 per Gbyte/mo</td>
<td>$.40 per Gbyte/mo</td>
</tr>
<tr>
<td>Admin</td>
<td>140 servers/admin</td>
<td>&gt;1000 servers/admin</td>
</tr>
</tbody>
</table>

Large Datacenters (tens of thousands of computers)
Medium Datacenters (thousands)

Source: Hamilton, Internet-Scale Service Efficiency, LADIS Workshop (Sept 08)
Issues

- Security
- Transparency
- Data lock in
- SLA's
- Trust
Initial capabilities

- Replication, up to three providers (including local store)
- Web based “Dashboard”
- Data integrity checking and monitoring
- Can push content from DSpace and Fedora repository platforms via plug-ins
- Pay per use
- Initial compute services on content
Additional services

• Other DuraSpace-provided services on top of content stored in the cloud
  – Search
  – Aggregation
  – Streaming
  – Migration
  – Hosting repositories
Enable others to build and deploy services and apps in DuraSpace environment
Use Cases:
DuraCloud with Cloud Storage

- Online backup for text, images, datasets, video, audio
- Enable preservation via multiple copies, geographies, administrations
- Elastic provisioning of temporary or permanent storage for projects or jobs
Use Cases:
DuraCloud with Cloud Compute

• Streaming service for video
• JPEG2000 image engine
• Indexing and other processing heavy jobs
• Repositories in cloud
• Data and text mining over open data
• Aggregation and web 2.0 tools on open content and collections
DuraCloud
Underlying software

• Open core
  ✓ Core components available for others to build on and run
  ✓ Open source - apache license

• Architecture to create cloud networks
  ✓ Public clouds
  ✓ Private clouds
  ✓ University consortia

• Also useful in research partnerships
Critical success factors

- Ease of use - simplicity
- Trusted partner for end user
- Cost effective
- Elastic, scalable, flexible
- Establish key partnerships with cloud preferred cloud service providers
- Build community of developers and users
Partners and Pilots

- Selected initial cloud providers
  - Amazon
  - Sun
  - Microsoft
  - EMC
- Selected initial 3 pilot partners
  - New York Public Library
  - Biodiversity Heritage Library
  - TBD (in selection process now)
Timeline

- Alpha DuraCloud service – June 2009
- Begin pilots – September 2009
- Pilot data loading and testing – Fall 2009
- Plug-ins for repository platforms – Q4 2009
- Beta for repository community - Q1 2010
- Pilot testing with compute services Q1 2010
- Report pilot results – Q1 2010
- Launch production service Q2 2010
For more information:

DuraSpace Organization: http://duraspace.org

DuraCloud Service: http://duracloud.org (soon)