DuraCloud

A service provided by

Sandy Payette
Chief Executive Officer
DuraSpace
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).
Vision: Preservation Support

Heaven

DuraCloud: content replication, auditing, and repair
Vision: Federated Repositories and Cyberinfrastructure

DuraCloud: repositories and data linking of stored objects
Vision: Shared collections

DuraCloud: access via JPEG2000 engine on stored images
Vision: Data Analysis and Mining

Heaven

DuraCloud: running large compute jobs on stored content
Vision: Your Ideas...

DuraCloud: run your application as service on content

Heaven
DuraCloud Proposition

Trust and durability in the cloud
“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
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 useable 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
DuraCloud
Basic Architecture

DuraCloud – Notional View

- DuraCloud.org
  - Website
  - Account Administration

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

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

Replicates to
Stores in
DuraCloud Flow

- Customer WebDav Client
- Customer Browser
- Customer Webapp
- DuraDav Webapp
- Service Application
- Main Webapp
- Storage Provider Interface
  - Amazon S3
  - RackSpace CloudFiles
  - Microsoft Azure Storage
  - EMC Atmos Storage Utility
  - Sun Cloud Storage
Making the Cloud Durable

Use Cases

Partnerships and Pilots
Advantages – Cloud Services

- Flexibility
- Scalability
- Elasticity
- 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)
Concerns - Cloud

- Security
- Transparency
- Data lock in
- SLA’s
- Trust
How DuraCloud can help...

- Hosted service by non-for-profit org
- Network of preferred cloud providers
- Trusted broker
- Ease of use
- Flexibility and scalability
- Risk mitigation by diversification
- Cost effective
DuraCloud Core Service
(motivated by preservation)

- Replicate to multiple storage providers
- Replicate to multiple geographic areas
- Manage and monitor content and services through web based “Dashboard”
- Integrity checking, alerting, repair
- “Pay for use” for services and storage
- Integrated billing
DuraCloud Menu of Services
(motivated by access, sharing, linking)

• Use cloud compute on stored cloud content
• Optional services
  ✓ Search
  ✓ Aggregation
  ✓ Streaming
  ✓ Migration
  ✓ Hosting repositories
• Enable others to deploy their services and apps in DuraCloud 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
- Staging area for repository ingest
- 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
  ✓ Open API
  ✓ Core components as open source
  ✓ Intent: Apache-style license

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

• Also useful in research partnerships
Providers and Pilot Partners

• Selected initial cloud providers
  ✓ Amazon
  ✓ Sun
  ✓ Microsoft
  ✓ EMC
  ✓ Rackspace

• Selected initial 3 pilot partners
  ✓ New York Public Library
  ✓ Biodiversity Heritage Library
  ✓ TBD (in selection process now)
Timeline

- DuraCloud APIs published – July 2009
- Begin pilots (Alpha service) – Sept 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 initial service Q2 2010
For more information:

DuraSpace Organization: [http://duraspace.org](http://duraspace.org)

DuraCloud Service: [http://duracloud.org](http://duracloud.org) (soon)
Thank You