DuraCloud

Open technologies and services for managing durable data in the cloud

Michele Kimpton, CBO DuraSpace
Open Source Portfolio

Fedora Commons

DuraCloud
Goals of DuraSpace

Stewardship:
Support and align open source development communities for DSpace and Fedora

Innovation:
Think beyond existing platforms
New strategies for enabling access and preservation of digital content

Sustainability:
Develop business model to sustain the non-profit and open technologies we support
DSpace and Fedora Installations

Largest share of open repositories worldwide...
... over 700 institutions tracked in our registries
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 data mining and analysis of large datasets

Making digital content more accessible and useable to researchers
Implications for our future work

- more distributed
- more collaborative
- more web-oriented
- more open
- more interoperable
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).
Cloud services

- Software as a Service
- Infrastructure as a Service
- Platform as a Service
Public Cloud Services

Elastic web-based infrastructure for storage and compute
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)
Study of 605 government IT

MeriTalk
The Government IT Network

Cloud Benefits

With Federal budgets shrinking, the cloud may offer an opportunity to save without sacrifice

What are the most significant benefits of cloud computing for your agency?*

- **57%** Reduced hardware requirements
- **45%** Reduced costs – pay-as-you go model
- **35%** Reduced staff requirements – allows for focus on more critical tasks
- **33%** Flexibility
- **24%** Potential for group collaboration
- **22%** Continuity of Operations Planning (COOP)

*While Federal and industry IT managers agree on the top three benefits of cloud computing, Federal managers are twice as likely to view COOP as a top benefit – 22% to 11%*

Yet, only 13% utilizing cloud compute today

Barriers

However, Federal IT managers see security concerns as a major hurdle to cloud advancement

What are the biggest barriers to cloud computing within your agency?*

<table>
<thead>
<tr>
<th>Barriers</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Security</td>
<td>78%</td>
</tr>
<tr>
<td>Privacy</td>
<td>41%</td>
</tr>
<tr>
<td>Costs</td>
<td>24%</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>24%</td>
</tr>
<tr>
<td>Performance</td>
<td>19%</td>
</tr>
</tbody>
</table>

Though Federal and industry IT managers both see it as the top barrier, Federal agencies are more likely than industry organizations to experience implementation challenges with security – 73% to 48%

Password: *********

Here to stay

Despite uncertainty, Federal IT managers see the cloud as more than the latest tech trend

**Confident Outlook:**

- **76%** Of Federal IT managers believe the cloud is “here to stay”
- **61%** Believe that in five years, the majority of large enterprises will rely on cloud computing for core applications

**Aggressive Timeline:**

- **2.6** Average number of years that Federal IT managers believe it will take for their agency to realize the benefits of cloud computing

> Federal IT managers are more likely than industry managers to hold this optimistic outlook – 61% to 51%

DuraCloud is a platform aimed at supporting libraries, universities, and other cultural heritage organizations that wish to provide perpetual access to their digital content. The service replicates and distributes content across multiple cloud providers and enables the deployment of services to support:

* access
* preservation
* re-use
DuraCloud

A web based service enabling management of Data in the cloud
Vision: Preservation Support

DuraCloud: content replication, auditing, and repair
Vision: Shared infrastructure

DuraCloud: collaboration and data linking of stored objects
Vision: Data Analysis and Mining

DuraCloud: running large compute jobs on stored content
DuraCloud Preliminary Interaction Flow

User

DuraCloud.org

UI

Storage

API

Services

Storage Provider Interface

Amazon S3

EMC Atmos

RackSpace CloudFiles

Microsoft Azure Storage

Sun Cloud Storage
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
Preservation Services

-ability to replicate content to multiple providers and locations

-ability to synchronize backup with primary store or repository system

-management, monitoring, audit and repair through web based interface

Hosted by DuraSpace not-for-profit org
Partnerships with cloud providers
software services

- Other DuraSpace-provided services on top of content stored in the cloud
  - Data mining
  - Video Streaming
  - Format transformation
  - Repository hosting
  - discovery
Enable others to build and deploy services and apps in DuraCloud environment

DuraCloud: run your application as a service on content
Partners and Pilots

- Selected initial cloud providers
  - Sun Microsystems
  - The Rackspace Cloud
  - Amazon Web Services
  - EMC²

- Selected 2 initial pilot partners
  - Biodiversity Heritage Library
  - The New York Public Library

Thank You, New Yorkers!
NYPL pilot

Digital Gallery Collection

- back up copy 700k images (50 TB data)
- transformation from Tiff to JPEG 2000
- run image server in cloud
- Push JPEG 2000 back into Fedora Repository
BHL pilot

BioDiversity Heritage Library

- back up copy entire corpus (40 TB data)
- have multiple copies including Europe
- Do compute intensive data mining over corpus
Pilot use cases

**NYPL**
Replication and preservation support
Format conversion
Instant provisioning of image server
Synchronization with repository

**BHL**
- Replication and preservation support
- International collaborative infrastructure
Researcher platform for data mining
Timeline

• Begin pilots (MOU’s in place) – September 2009
• DuraCloud Alpha Pilot release – Oct 2009
• Pilot data loading and testing – Fall 2009
• Beta for repository community - Q1 2010
• Pilot testing with software services Q1 2010
• Cloud partner evaluations complete - Q2 2010
• Strategic cloud partnerships in place - Q2 2010
• Pricing Model determined - Q2 2010
• Report pilot results – Q2 2010
• Launch production service Q3 2010
Critical success factors

- Ease of use - simplicity
- Trusted partner for end user
- Cost effective
- Scalable/Flexible
- Can establish key partnerships with service providers
- Can build community of developers and users
Thank You

For more information:

DuraSpace Organization: http://duraspace.org
Wiki: http://www.fedora-commons.org/confluence/display/duracloudpilot/
Mkimpton@duraspace.org