DuraCloud
Managing durable data in the cloud

Michele Kimpton, Director DuraSpace
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
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

Elastic web-based infrastructure for storage and compute
Challenge

Digital preservation is essential but difficult to implement

- Tools and processes unproven
- Limited IT support
- Resources unavailable
- Task can be overwhelming (replication, migration, emulation, etc.)
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

• Stability
• Transparency
• Data lock in
• SLA’s
• Trust
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
Available to run internally- open source
Additional services

- Other DuraSpace-provided services on top of content stored in the cloud
  - Search
  - Aggregation
  - Streaming
  - Migration
  - Hosting repositories
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
- Hosting 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 within community
- 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
  - Sun Microsystems
  - Amazon Web Services
  - Rackspace
  - EMC

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

Thank You, New Yorkers!
Pilot use cases

• Ingest large quantity of material

• Replicate to multiple cloud platforms

• Manage replication and monitoring

• Run services
For more information:

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