

# DuraCloud

Open technologies and services for managing durable data in the cloud

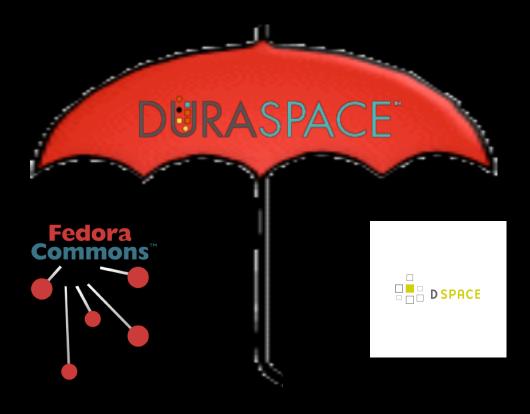


Michele Kimpton, CBO DuraSpace





## Open Source Portfolio











#### 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:

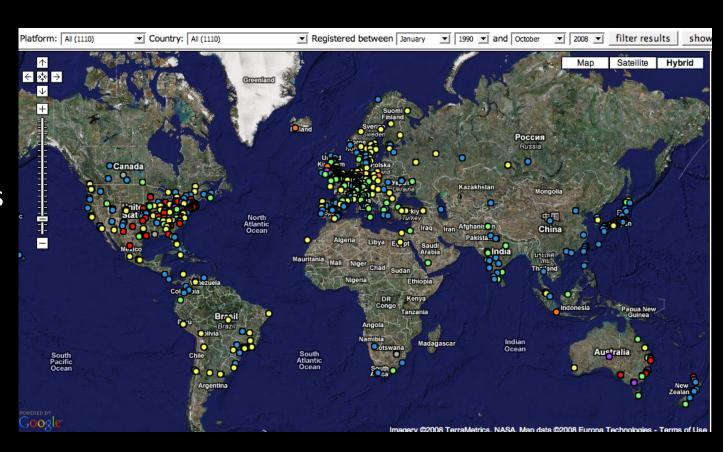
Develop business model to sustain the nonprofit and open technologies we support





#### **DSpace and Fedora Installations**

Universities
Research Centers
Libraries
Archives
Cultural Heritage
Government
More...





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

collaborative

more







more distributed

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).



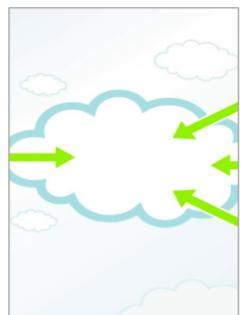


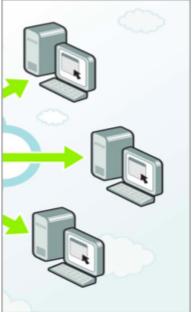


## Cloud services

- Software as a Service
- Infrastructure as a Service
- Platform as a Service











## **Public Cloud Services**

Elastic web-based infrastructure for storage and compute

















## **Economies of Scale and Cost**

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

| Technology* | Cost Medium Datacenter | Cost Large<br>Datacenter |
|-------------|------------------------|--------------------------|
| Network     | \$95 per Mbit/sec/mo   | \$13 per Mbit/sec/mo     |
| Storage     | \$2.20 per Gbyte/mo    | \$.40 per Gbyte/mo       |
| Admin       | 140 servers/admin      | >1000 servers/admin      |

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



#### **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?\*

Reduced hardware requirements
Reduced costs – pay-as-you go model
Reduced staff requirements – allows for focus on more critical tasks
Flexibility
Potential for group collaboration
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



http://www.meritalk.com/2009-cloud-consensus.php



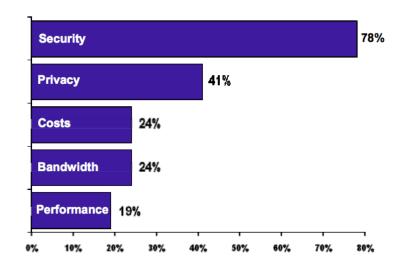
#### Barriers



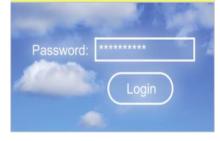
#### **Cloud Barriers**

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?\*



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%





http://www.meritalk.com/2009-cloud-consensus.php



## Here to stay



#### **Cloud Impact**

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

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

#### **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





http://www.meritalk.com/2009-cloud-consensus.php



## **DuraCloud Proposition**

#### Trust and durability in the cloud

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

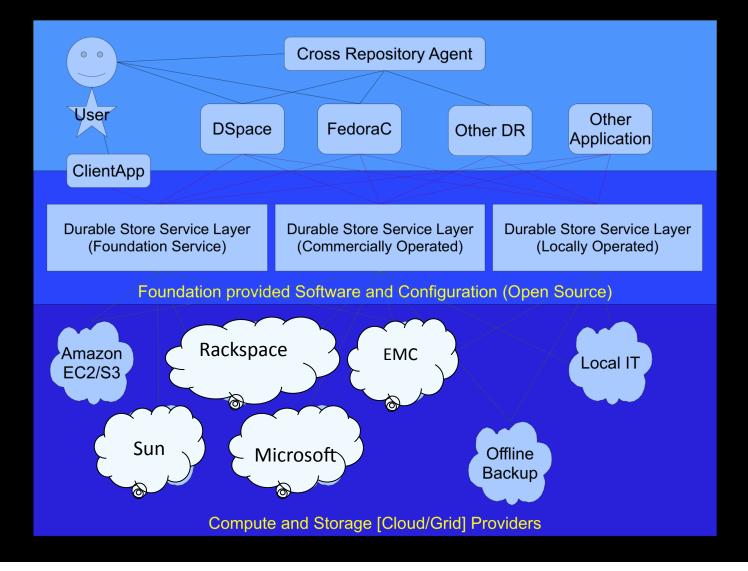




#### DuraCloud

#### A web based service enabling management of Data in the cloud

DuraCloud mediating web Service







## Vision: Preservation Support









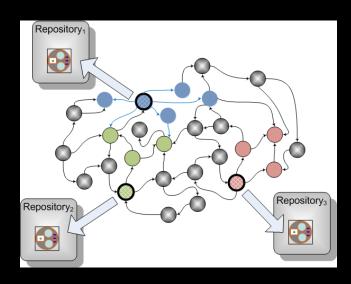


DuraCloud: content replication, auditing, and repair





#### Vision: Shared infrastructure

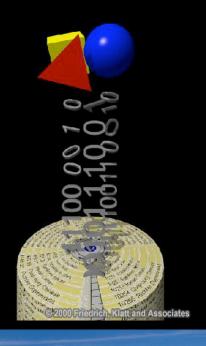


DuraCloud: collaboration and data linking of stored objects



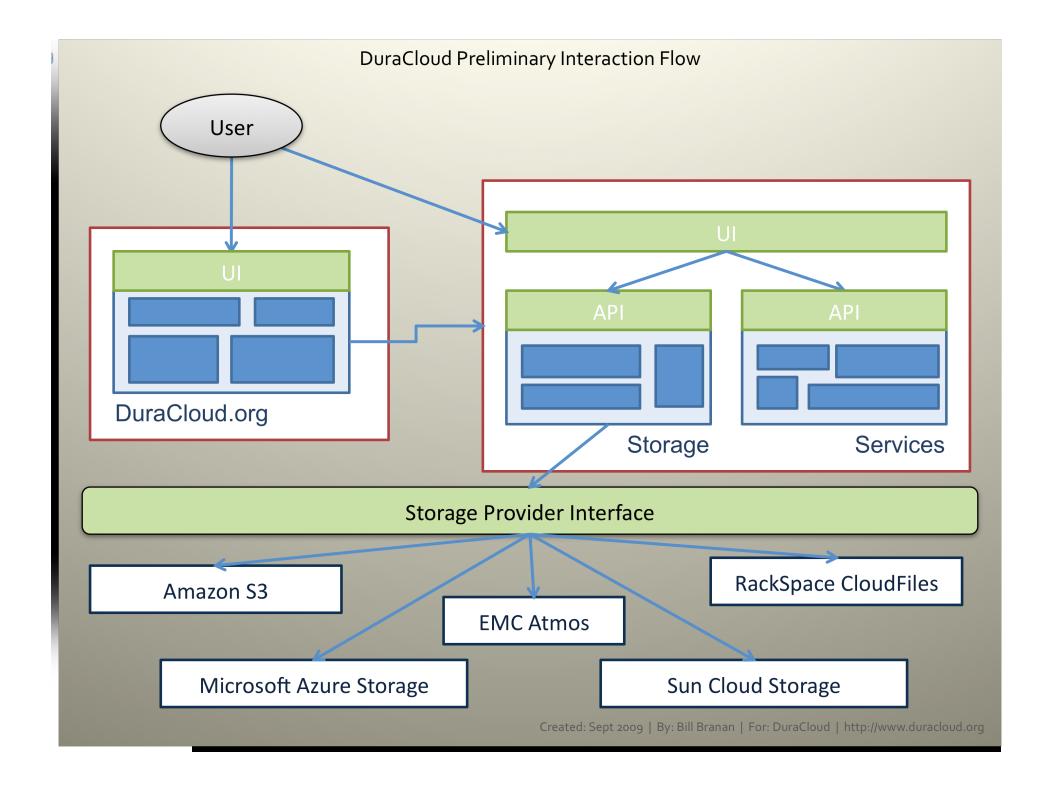


## Vision: Data Analysis and Mining

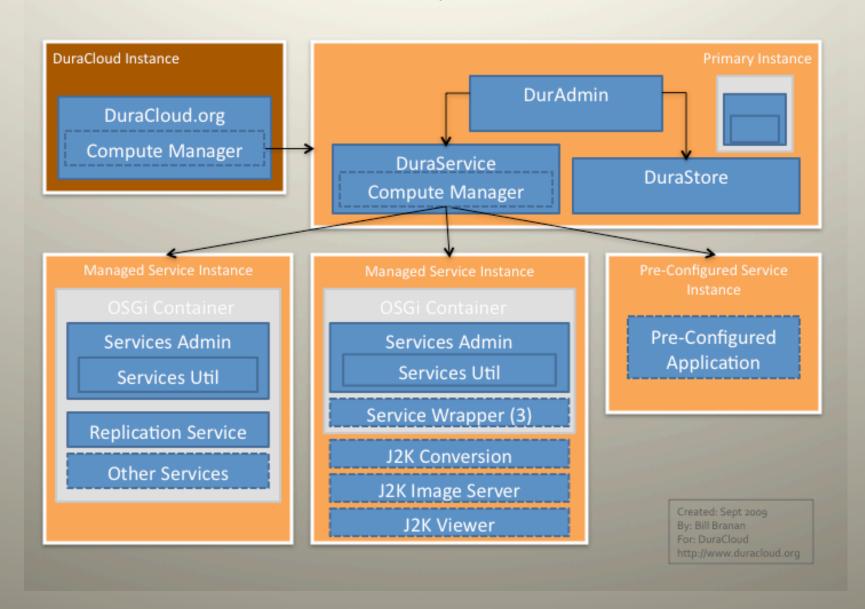


**DuraCloud: running large compute jobs on stored content** 





#### DuraCloud Preliminary Server Architecture





# 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









Selected 2 initial pilot partners



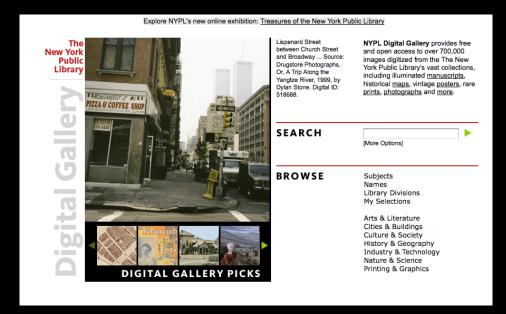






## NYPL pilot

#### 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



News and Updates from the Biodiversity Heritage Library

available online at http://www.biodiversitylibrary.org

Over 250 years of biodiversity literature to support the work of scientists, researchers, students and you.















#### BHL RECENT ADDITIONS

L'Illustration horticole : v. 20 1873 (added: 09/11/2009) - 09/11/2 009

L'Illustration horticole : v. 19 1872 (added: 09/11/2009) - 09/11/2

The Floricultural Cabinet and Florist's Magazine. VOL 1 (added: 09/10/2009) - 09/10/2

Voyage de La Pérouse autour du monde / Atlas (added: 09/10/2009) - 09/10/2 TUESDAY, AUGUST 25, 2009

#### Book of the Week: BHL, EOL, and Marine Life



freely available to people around the world is an ambitious goal embraced by many in the scientific community, and it represents the mission behind both the Biodiversity Heritage Library and the Encyclopedia of Life (www.eol.com), of which BHL is a cornerstone institution. While the Biodiversity Heritage Library strives to digitize the published librarbure

biodiversity information

#### ABOUT BHL

Twelve major natural history museum libraries, botanical libraries, and research institutions have joined to form the Blodiversity Heritage Library. The participating libraries have over two million volumes of biodiversity literature collected over 200 years to support the work of scientists, researchers, and students in their home institutions and throughout the world.

#### LABELS

News (10

-back up copy entire corpus (40 TB data)

-have multiple copies including Europe

-Do compute intensive data mining over corpus





#### Pilot use cases

#### <u>NYPL</u>

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: <a href="http://duraspace.org">http://duraspace.org</a>
Wiki: <a href="http://www.fedora-commons.org/">http://duraspace.org/</a>
Confluence/display/duracloudpilot/
Mkimpton@duraspace.org

