DuraCloud

Open technologies and services for managing durable data in the cloud

DURASPACE™

Michele Kimpton, CBO DuraSpace
"Repositories in the Cloud" Seminar, Feb 2010
mkimpton@duraspace.org

Open Source Portfolio
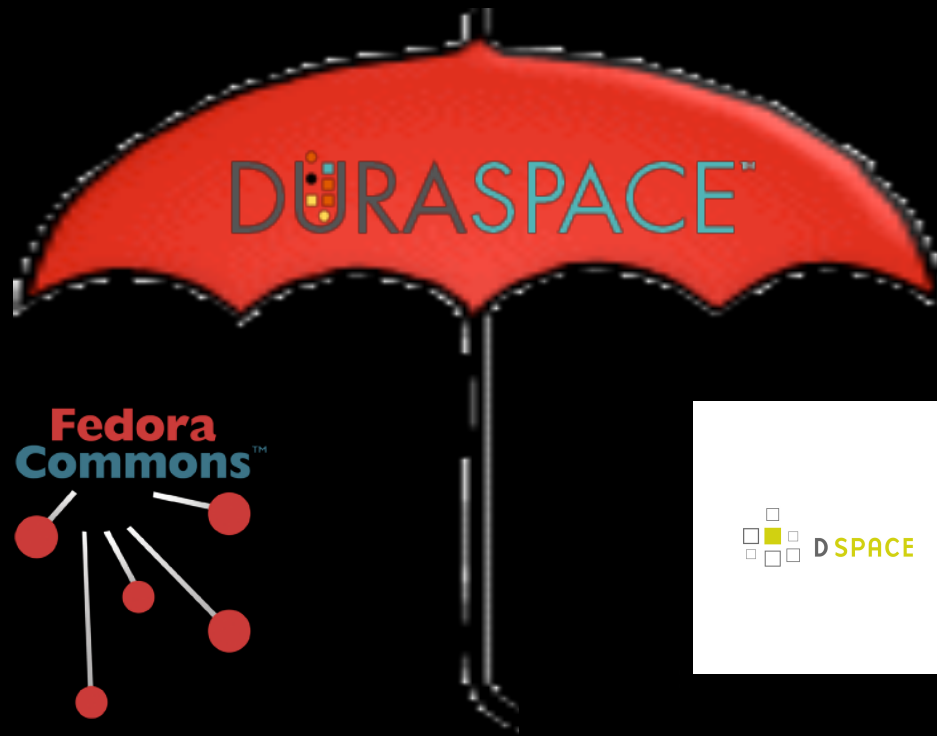
Implications for our future work

Challenges (from survey 1/22/2010)

Preservation support is hard to implement consistently

"Our preservation support is collection based where we have had grants or specific initiatives. There is no system effort."

"Where it is prioritized as mission critical, it is being done well. It is not being done well where it is not mission critical."

"We have not invested enough to make it a service of which we are proud..."

"Collection development and storage are more important than computing. "

Key Advantages
completed 1/22/2010
145 participants higher ed

Most Impactful Advantages Electronic Survey	Responses
Scalability	79
Remote, Off Campus Storage of Digital Assets	64
Ease of Implementation	54
Flexibility	53
Don't Have to Staff Locally	39
Cost	33
Elasticity	26
Pay for Use	14
Other	5

Key Challenges
completed 1/22/2010
145 participants higher ed

Key Challenges Electronic Survey	Responses
Trusting Third Party to Manage Critical Assets	64
Long-term Reliability of Solution	52
Data Security	51
Performance and Bandwidth Concerns	37
Loss of Control	34
Administrative Burden of SLAs	17
Transparency of Solution	16
Concerns about Data Lock-in	16
Less Customizable	10
Other	12

Likely to use cloud services in next 12 months

Percentage of electronic survey respondents noting it is "very likely" or "likely" they will use cloud compute or cloud storage services to manage, store or provide access to digital collections in the next twelve months.

Category		Subcategory	Percentage
Non-US			47.7%
US Institutions	Institution Size	Large, very large	47.2%
		Medium	68.8%
		Small, very small	42.9%
	Enrollment Profile	RU/VH	52.1%
		RU/H, DRU	50.0%
		Master's S, M and L	46.2%
		Bac and Assoc	57.1%
	Public/ Private	Public	46.9%
		Private	59.3%

Institutional needs: managing digital collections

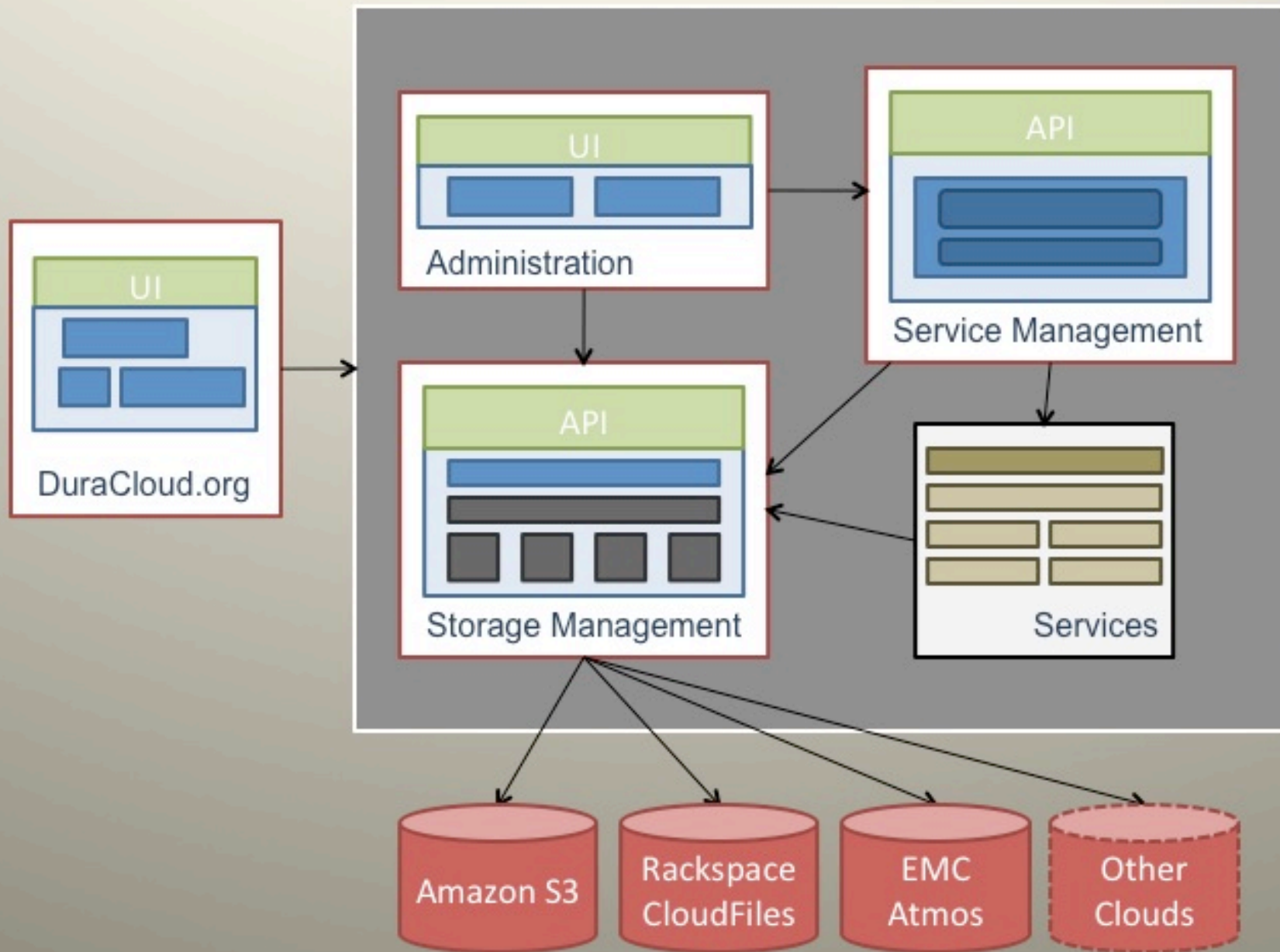| Service Area	Importance	Extent Need is Met	Difference	*Likelihood to Use Cloud Services*
Remote secondary storage of digital collections	3.54	2.60	0.94	*3.09*
Preservation support	3.35	2.17	1.18	*2.88*
Intra-institution shared collections	3.11	2.47	0.64	*2.69*
Inter-institution shared collections	2.72	2.07	0.65	*2.67*
Compute services	2.80	2.25	0.55	*2.54*
Online primary storage	*3.51*	*2.97*	*0.53*	*2.29*

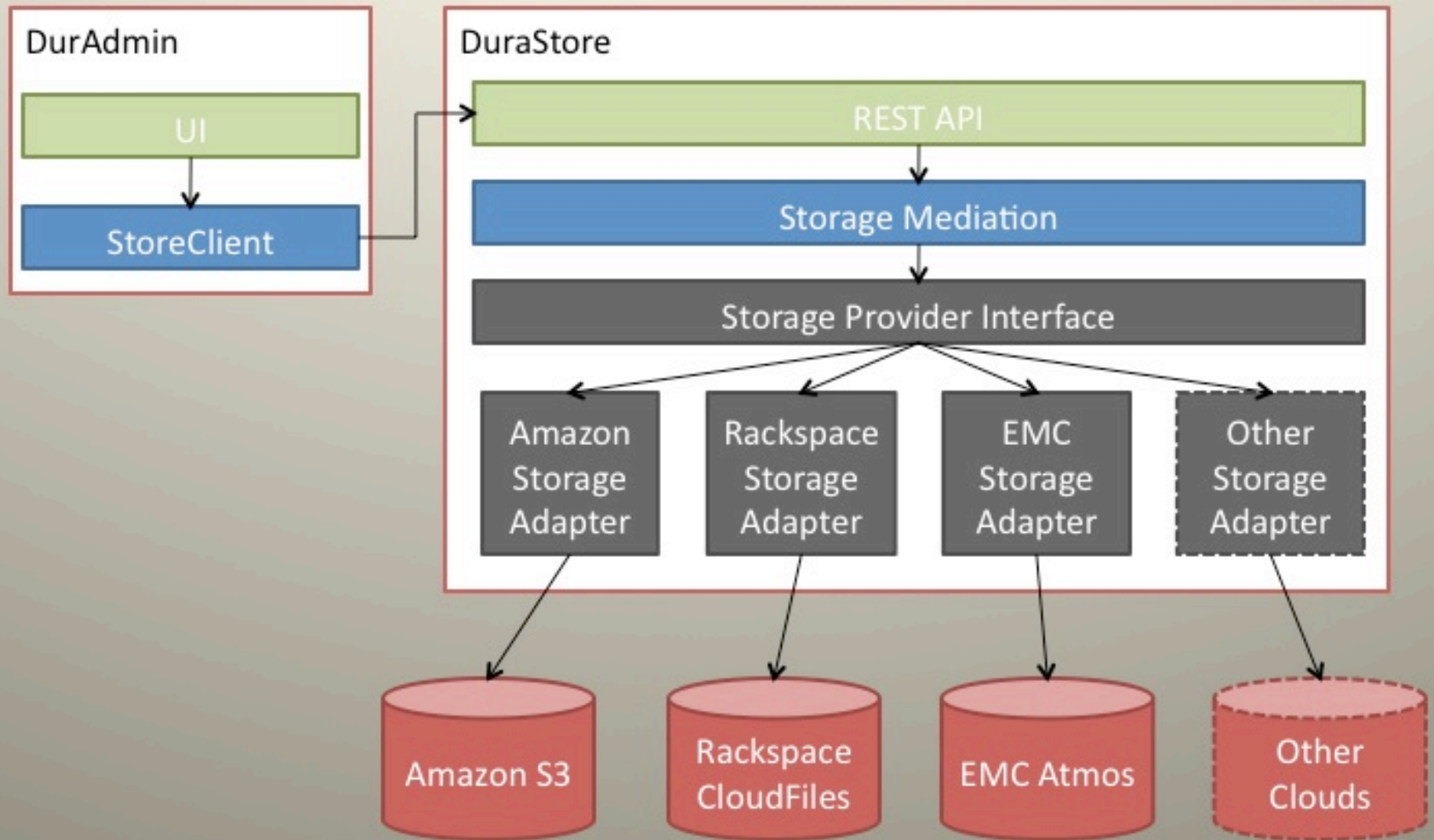# Services in the cloud for durable digital content
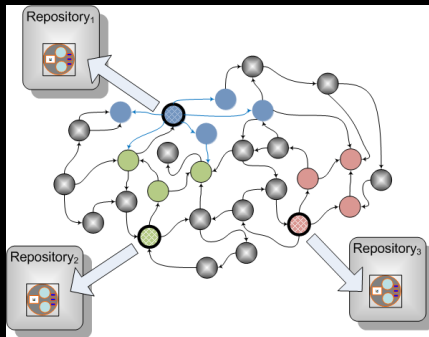
DuraCloud Platform: Allow organizations to utilize cloud infrastructure easily offering data storage, data replication, preservation support and access services
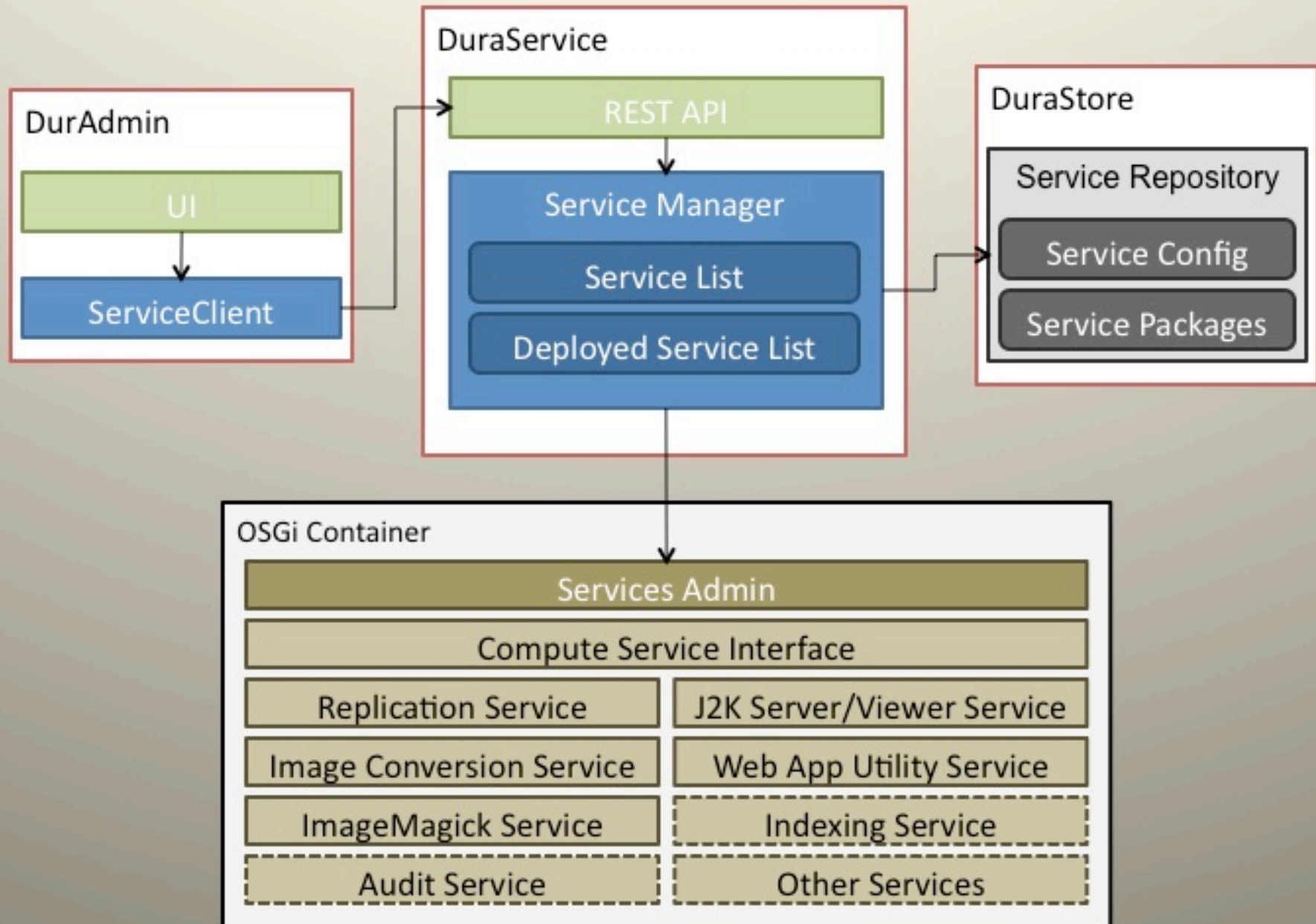
DuraCloud High Level Interaction

UI

DuraCloud.org

UI
Administration

API
Service Management

API
Storage Management

Services

Amazon S3

Rackspace CloudFiles

EMC Atmos

Other Clouds

DuraCloud Storage Architecture

DurAdmin
- UI
- StoreClient

DuraStore
- REST API
- Storage Mediation
- Storage Provider Interface
  - Amazon Storage Adapter
  - Rackspace Storage Adapter
  - EMC Storage Adapter
  - Other Storage Adapter

- Amazon S3
- Rackspace CloudFiles
- EMC Atmos
- Other Clouds

DuraCloud Service Architecture

DuraService

REST API

Service Manager

Service List

Deployed Service List

DurAdmin

UI

ServiceClient

DuraStore

Service Repository

Service Config

Service Packages

OSGi Container

Services Admin

Compute Service Interface

Replication Service	J2K Server/Viewer Service
Image Conversion Service	Web App Utility Service
ImageMagick Service	Indexing Service
Audit Service	Other Services

Preservation Services

-ability to replicate content to multiple providers and locations

-ability to synchronize backup with primary store or repository system

-access to content through web based interface

-ability to do bit integrity checking

-ability to do file format transformations

Partners and Pilots

- Selected initial cloud providers

- Selected 3 initial pilot partners

NYPL pilot

Digital Gallery Collection

Use case: back up online preservation copy to Fedora, file format transformation
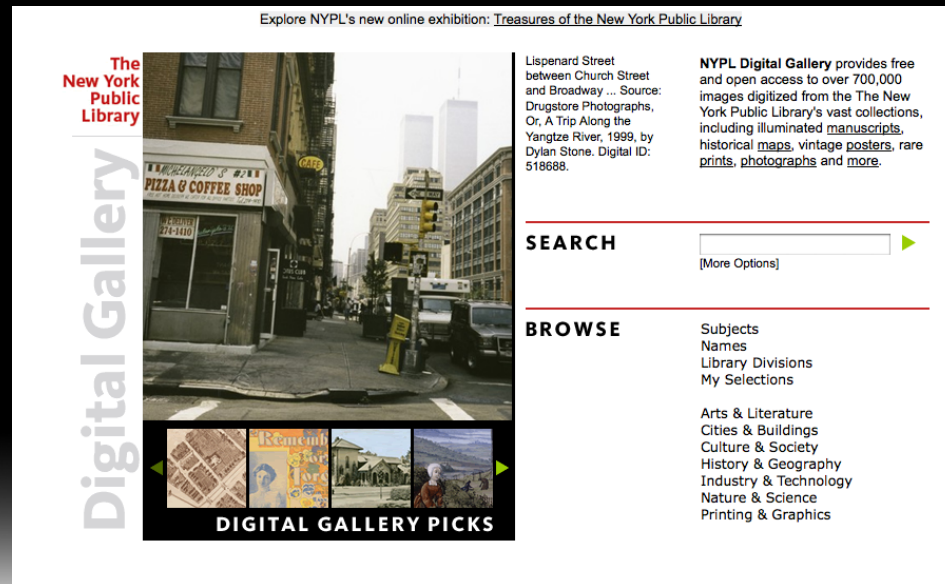
-back up copy all TIFF images (10 TB data)

-transformation from Tiff to JPEG 2000 using Imagemagick

-run J2k image server in cloud

-Push JPEG 2000 back into Fedora Repository

BHL pilot

BioDiversity Heritage Library

Use case: Find the best cost competitive solution for keeping multiple copies in multiple geographies, easily accessible.

-back up copy entire corpus (40 TB data- JPEG, Tiff

-have multiple copies including Europe

-Run J2K image server in cloud

WGBH Media Library and Archives

Use case: Provide backup preservation for video files from repository and other sources, and create derivative files for access and streaming.

- Archive large video files
- Provide public access to streaming versions
- Transcode files in cloud
- Edit files where appropriate to sell clips
- Give third party access to cloud store for processing and access

Challenges

- Provisioning bandwidth at local institution to transfer data

- Transferring large files over the wire (over 5 GB is rejected, found issues in transfer over 1 GB)

- Consistency of operation of 2nd tier providers (EMC, RackSpace)

- Enabling others to easily build on platform

- Best process for integration of 3rd party applications into hosting service

- Cost effective bit integrity checking

- Balancing ease of use and more sophisticated functionality

Advantages of hosted platform

Strategic partnerships with cloud providers

- Better pricing
- Transparency
- Early notification

Ease of implementation for end user

Multiple copies in multiple geographies/ administrations through one interface

✓ Access to broad number of services relevant to the repository community

Timeline

- Begin pilots– September 2009

- DuraCloud Alpha Pilot release- Oct 2009

- Pilot data loading and testing – Fall 2009

- Beta for repository community – Q2 2010

- Pilot testing with software services Q2 2010

- Cloud partner evaluations complete-Q3 2010

- Hosting service pricing and SLA's complete-Q3 2010

- Report pilot results – Q3 2010

- Code available open source-Q3 2010

- Launch production service Q4 2010

Next Steps(Feb-April)

- V.2 release complete

    - Replication, web access and viewing, file format conversion, J2K image server, bit integrity checking

- Launch Fedora and DSpace plug ins

- V.3 release primary features

    - Synchronization with local repository(Fedora and DSpace)

- Expand pilot in April to include 15 new users, to connect with current repositories

- Continue to test robustness and performance of commercial cloud partners

Thank You

For more information:

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