DURA SPACE

DuraCloud Pilot Program: utilizing cloud infrastructure as part of your preservation strategy

Michele Kimpton
Project Director, DuraCloud
NDIPP Partner meeting
July 21, 2010

DURACLOUD™

Overview

- What is DuraCloud
- Basic architecture
- Pilot program experiences
- Future direction

DuraSpace not for profit

Implications for our future work

more distributed

More collaborative

more web-oriented

more open

more interoperable

Cloud Infrastructure

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).

DuraCloud Platform

Open technology and hosted service for utilizing cloud infrastructure for preservation support and access services

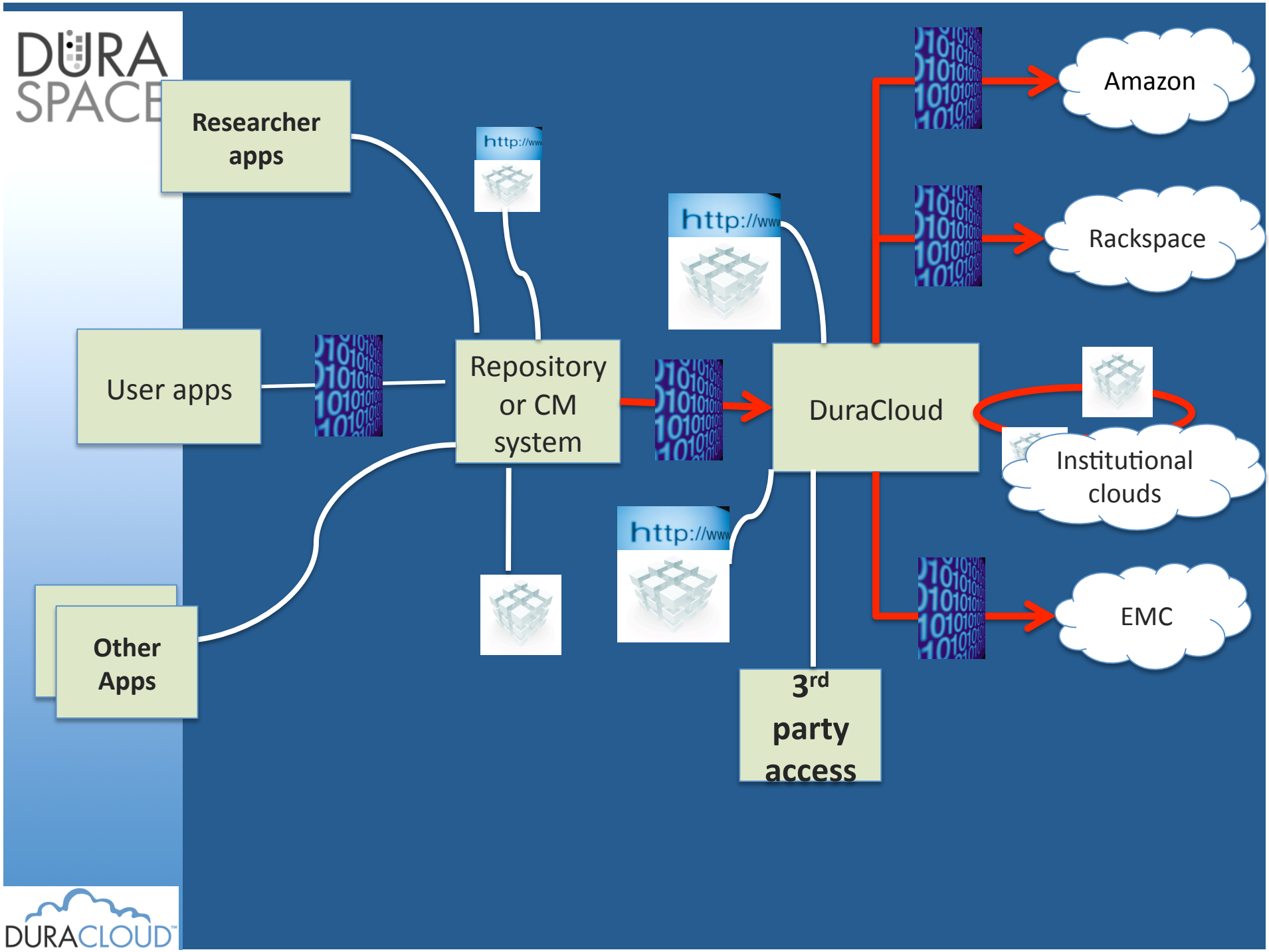
Architectural Features:

Interoperable across multiple cloud providers

Web enabled

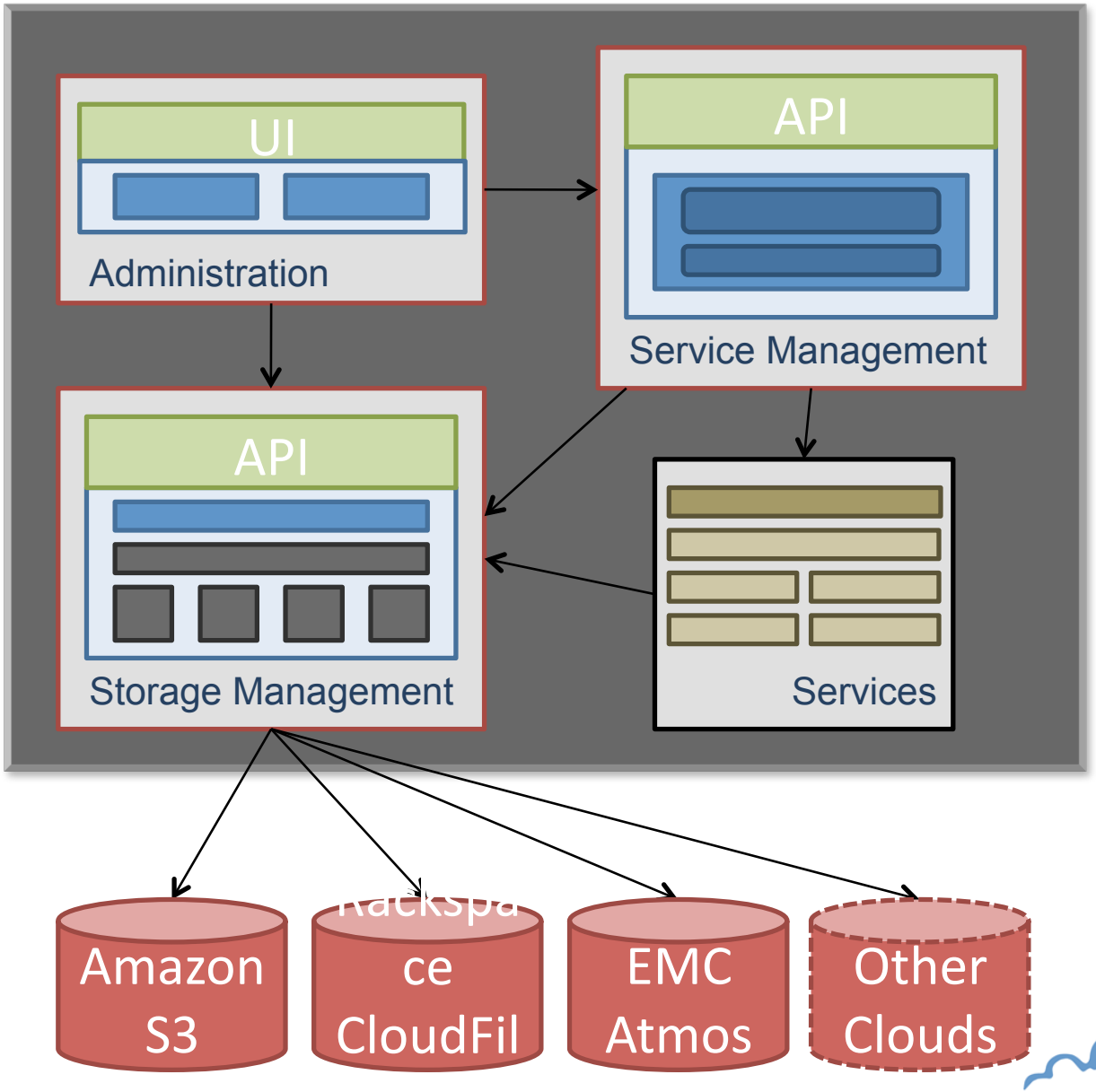
Built on highly scalable, flexible shared infrastructure

Open API's for easy integration


DURA SPACE
Researcher apps
User apps
Other Apps
Repository or CM system
DuraCloud
3rd party access
Amazon
Rackspace
Institutional clouds
EMC
DURACLOUD

DURA SPACE

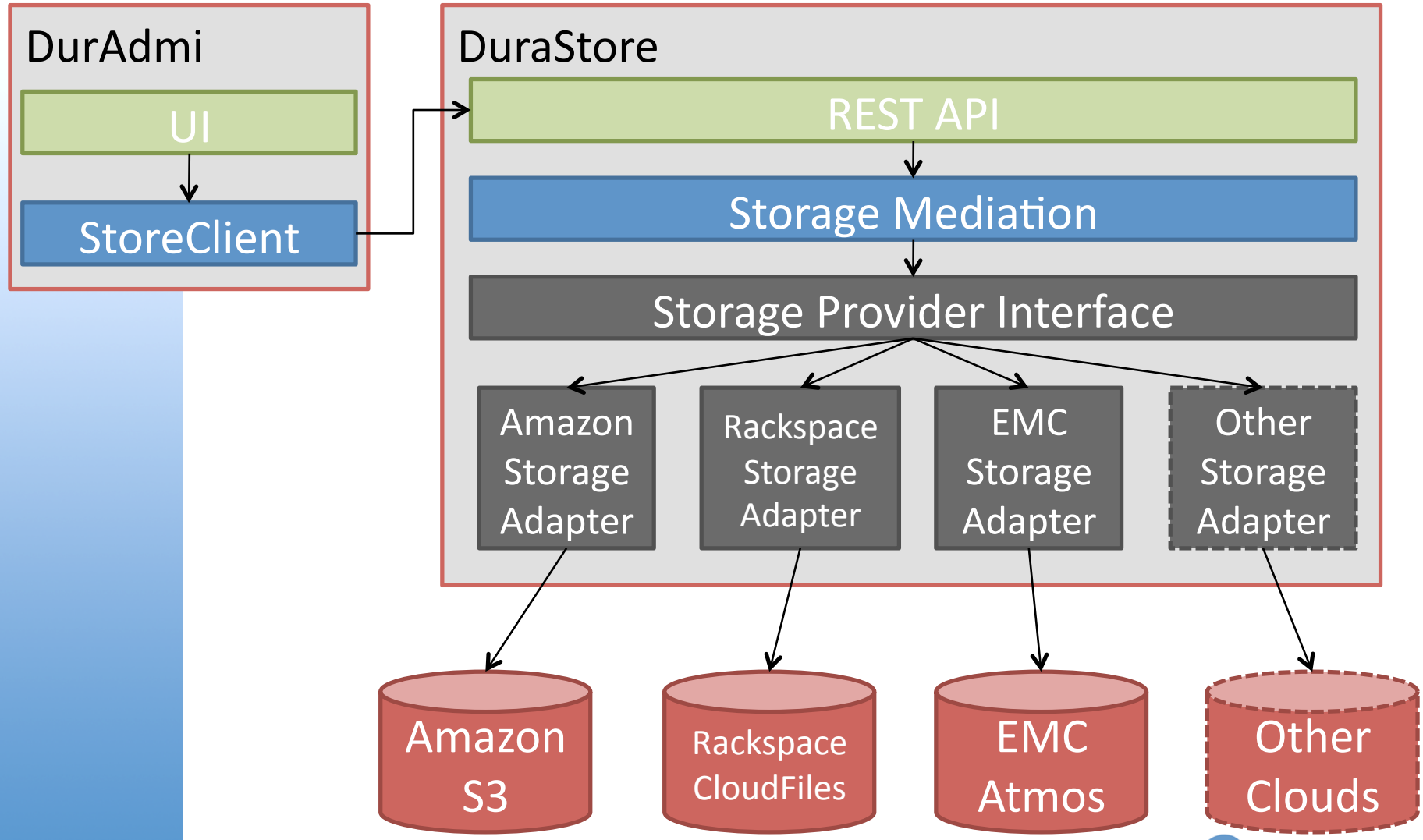
DuraCloud High Level Interaction



DURACLOUD™

Updated: July 2010 | http://www.duracloud.org

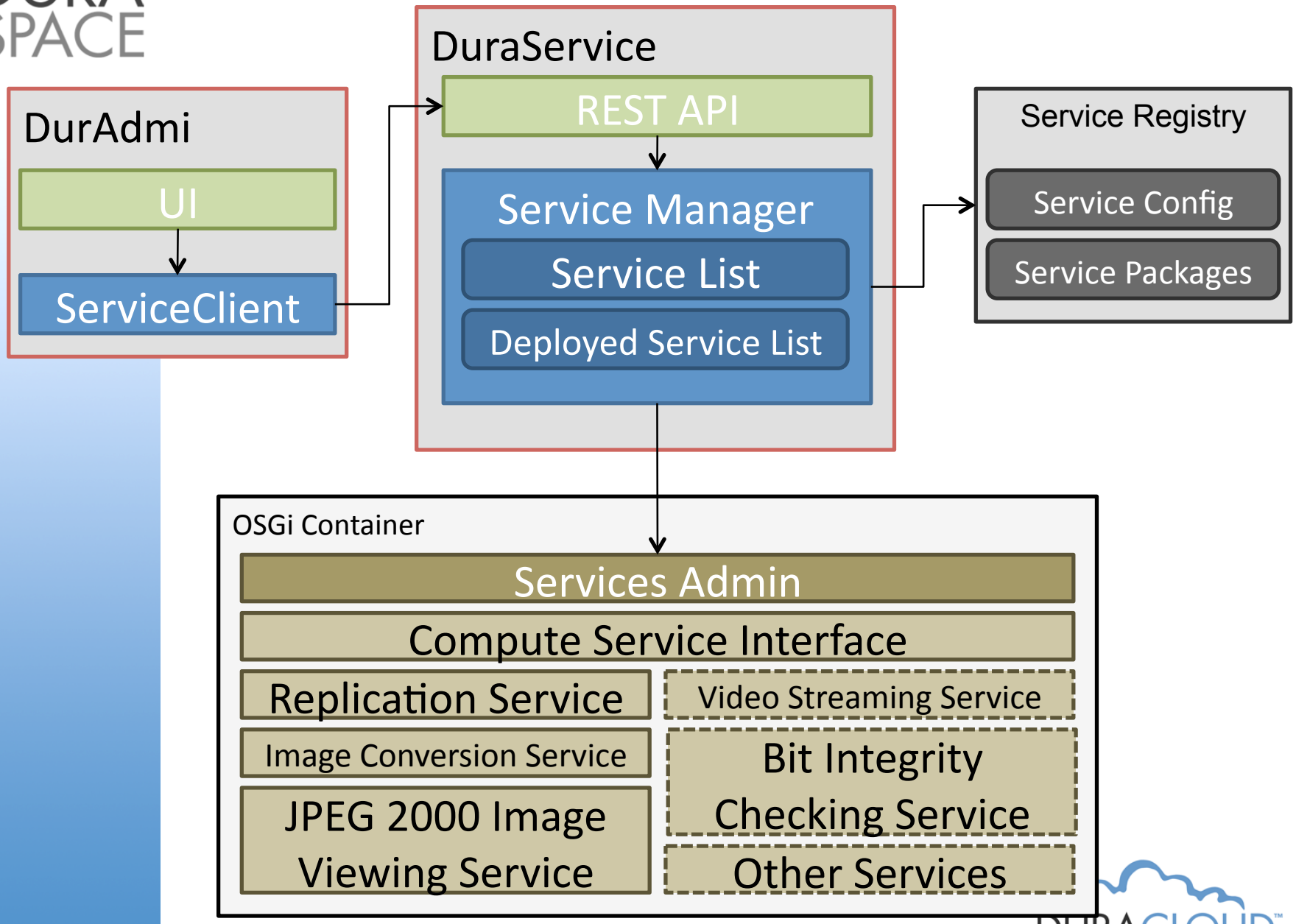
DuraCloud Storage Architecture



Updated: July 2010 | http://www.duracloud.org

DURA SPACE

DuraCloud Service Architecture



DURACLOUD™

Updated: July 2010 | http://www.duracloud.org

Services and Capabilities

Replication

Image Viewing

Image Transformation

Media Streaming

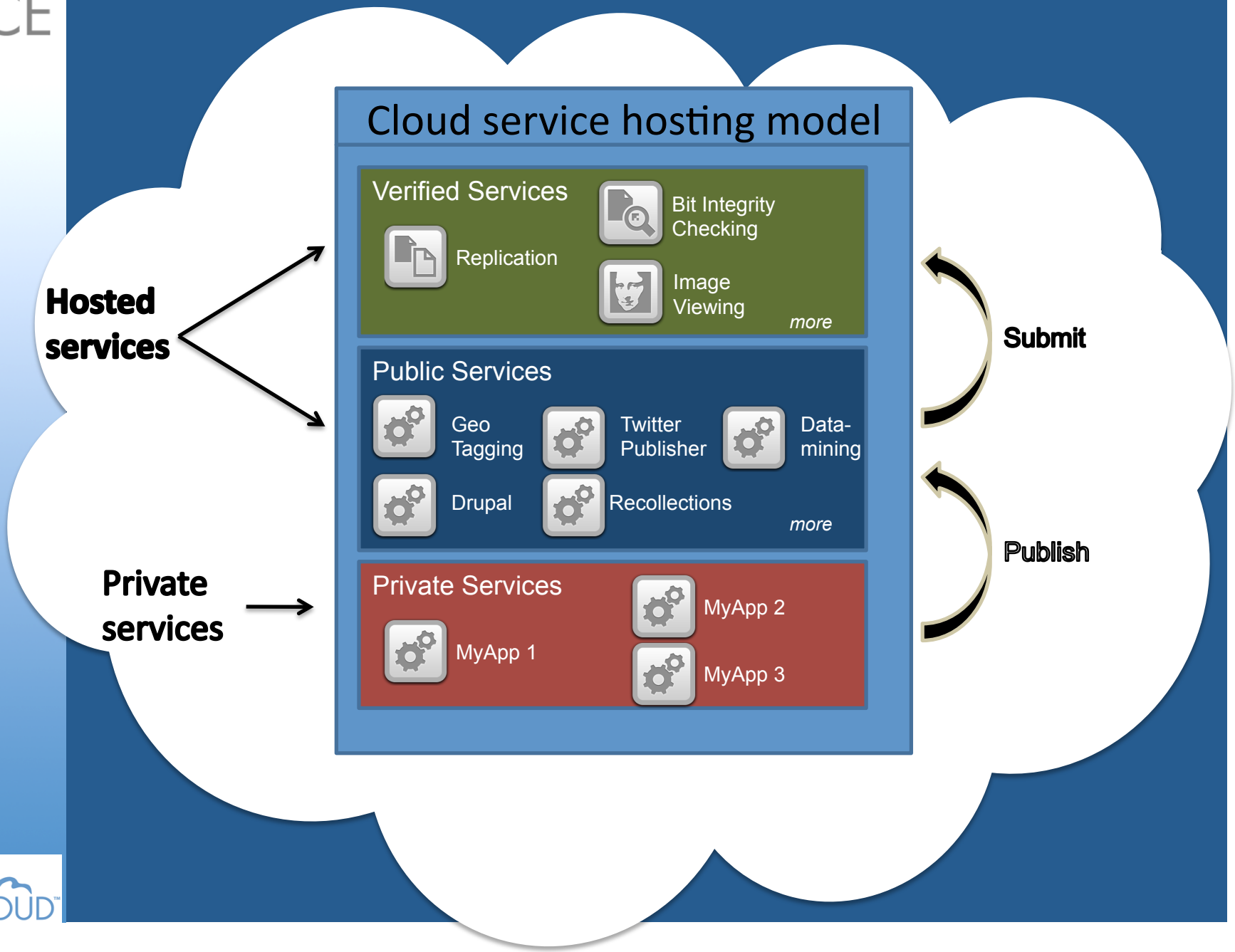
Bit Integrity Checking

…more on roadmap

File format validation

Parallel processing

DuraCloud Shared Services Vision

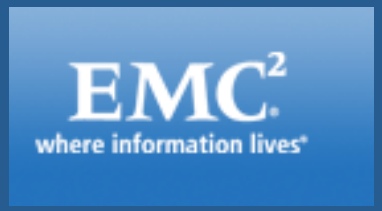


Purpose of Pilot Program

- Engage with users whom have concrete use cases and want to test the software
- Real data at scale
- Uncover the obstacles
- Engage community response and assessment
- Learn and assess benefits and limitations of the cloud

Partners and Pilots

- Selected initial cloud providers



- Selected 3 initial pilot partners

NYPL DuraCloud Pilot Goals

- Preservation
  - Migration of service files from an unsupported format to a supported format – Mr. Sid to JP2000
  - Data integrity checking of new format
  - Ingest of new data streams associated with existing objects in Fedora Repository

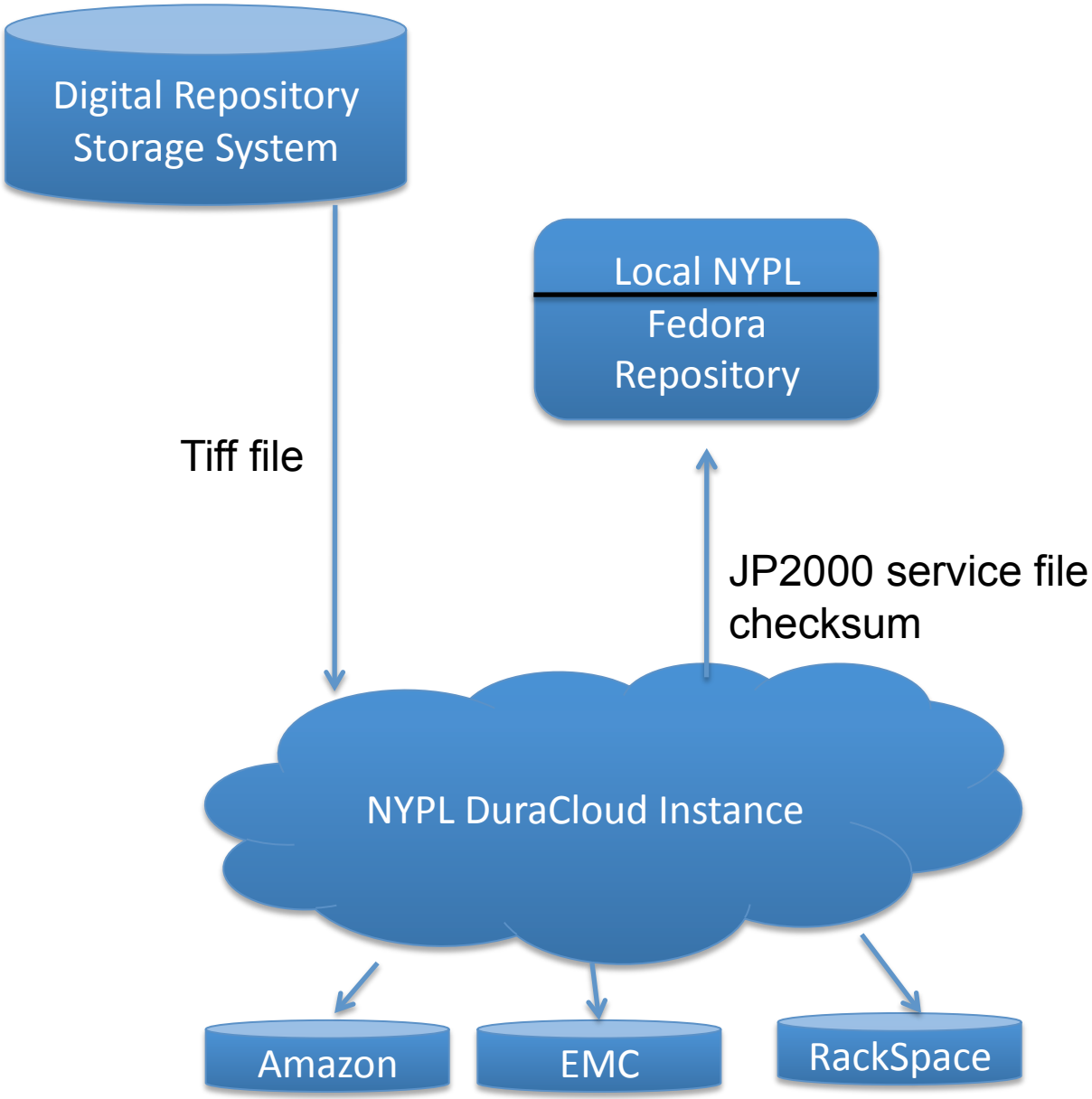
Access

- Reduce number of service files
- Reduce number of services used in delivering service files
- Provide reliable and dependable service and access to those service files

Key Advantages Cloud provides NYPL

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

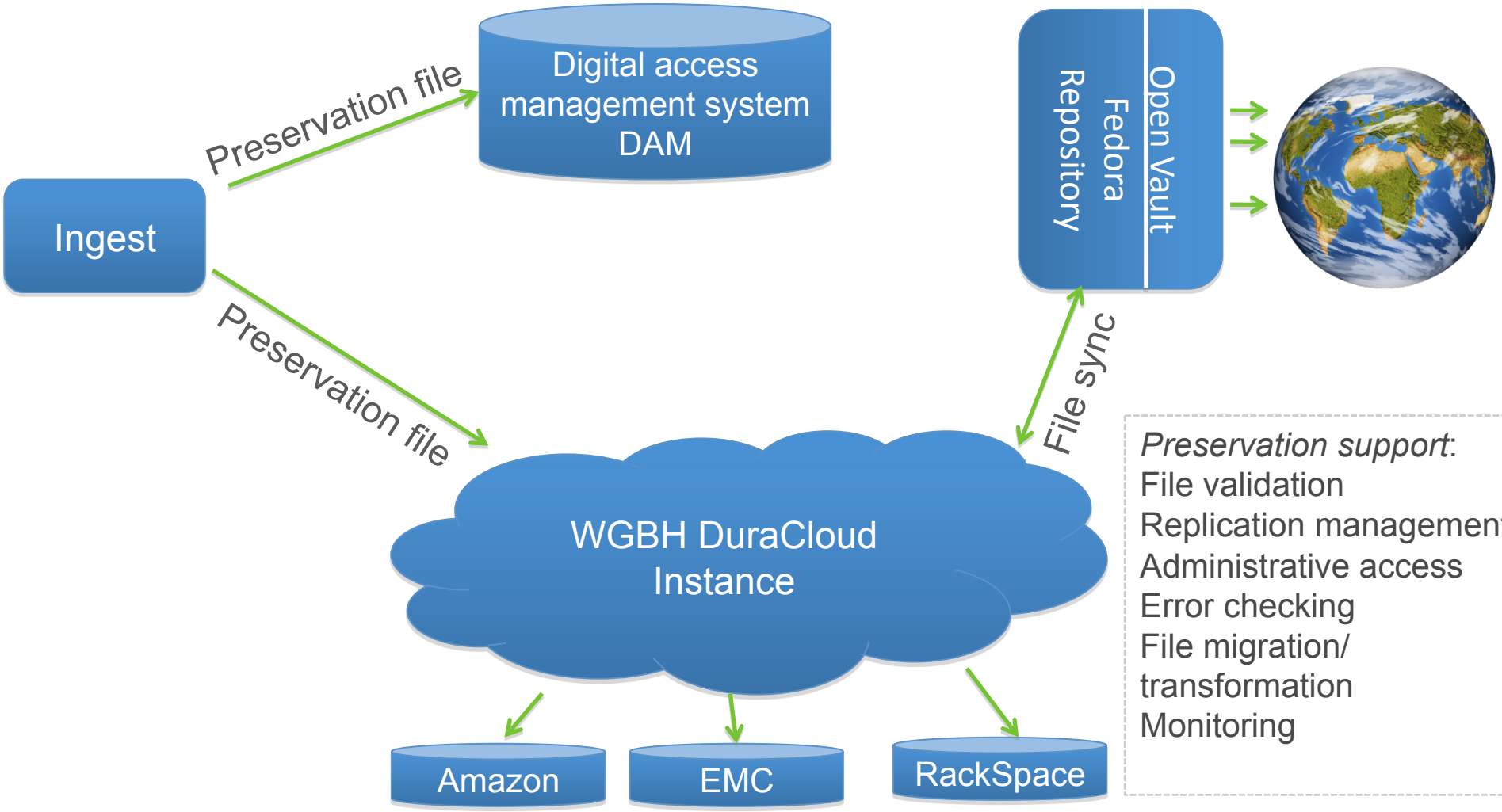
NYPL Preservation services utilizing DuraCloud



Outcomes

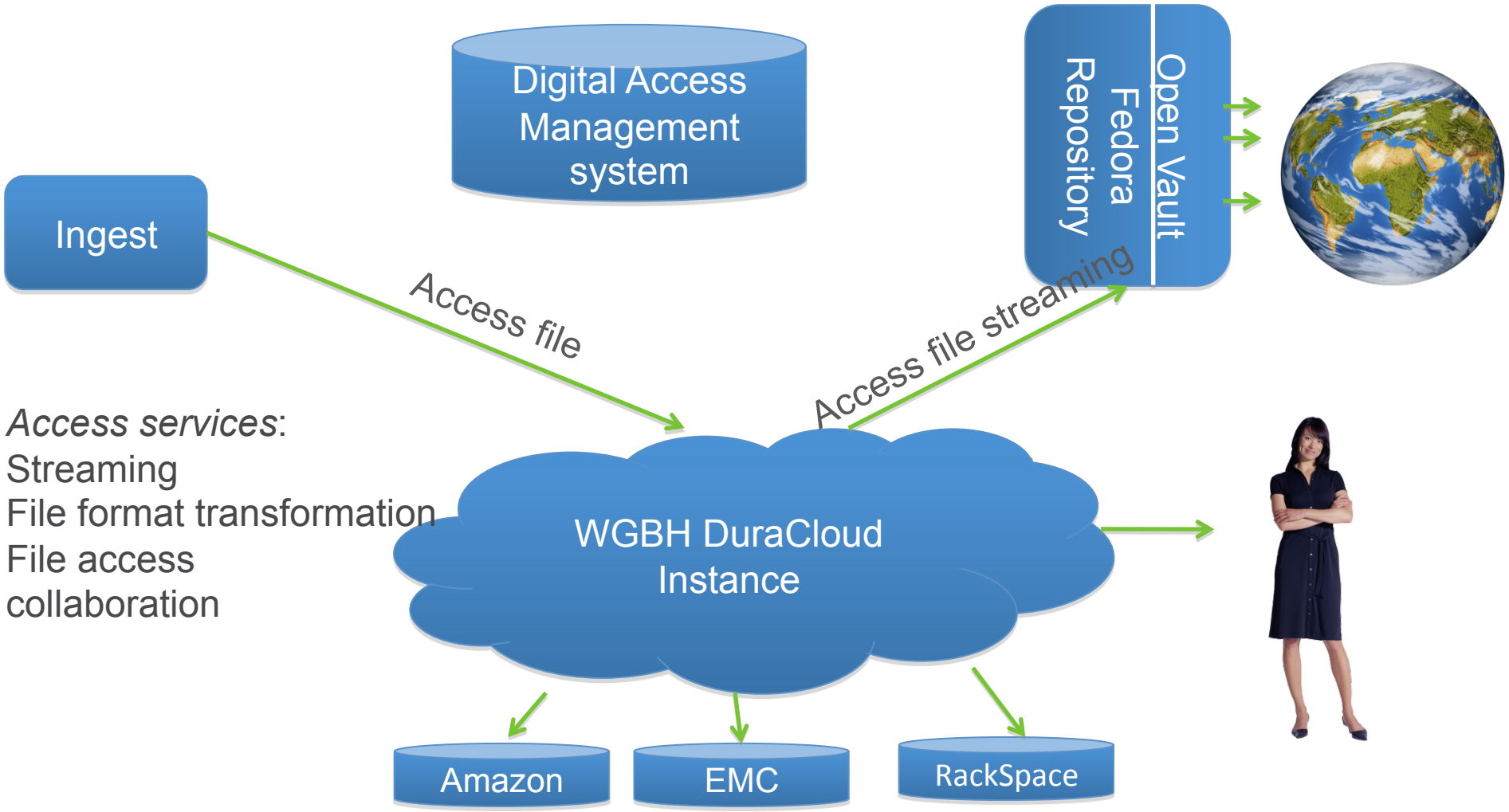
- Loaded 10 TBs of TIFFS to cloud
- Were able to view, convert, and serve a subset of those files via Adore Djatoka image service
- We were able to download and verify end products of conversion process with the DuraCloud API and JHOVE
- We were able to demo the process of chunking and storing multi-gigabyte media files with the DuraCloud sync tool
- Delays in hardware did not allow for provisioning additional services from DuraCloud

WGBH Preservation using DuraCloud



Preservation support:
File validation
Replication management
Administrative access
Error checking
File migration/ transformation
Monitoring

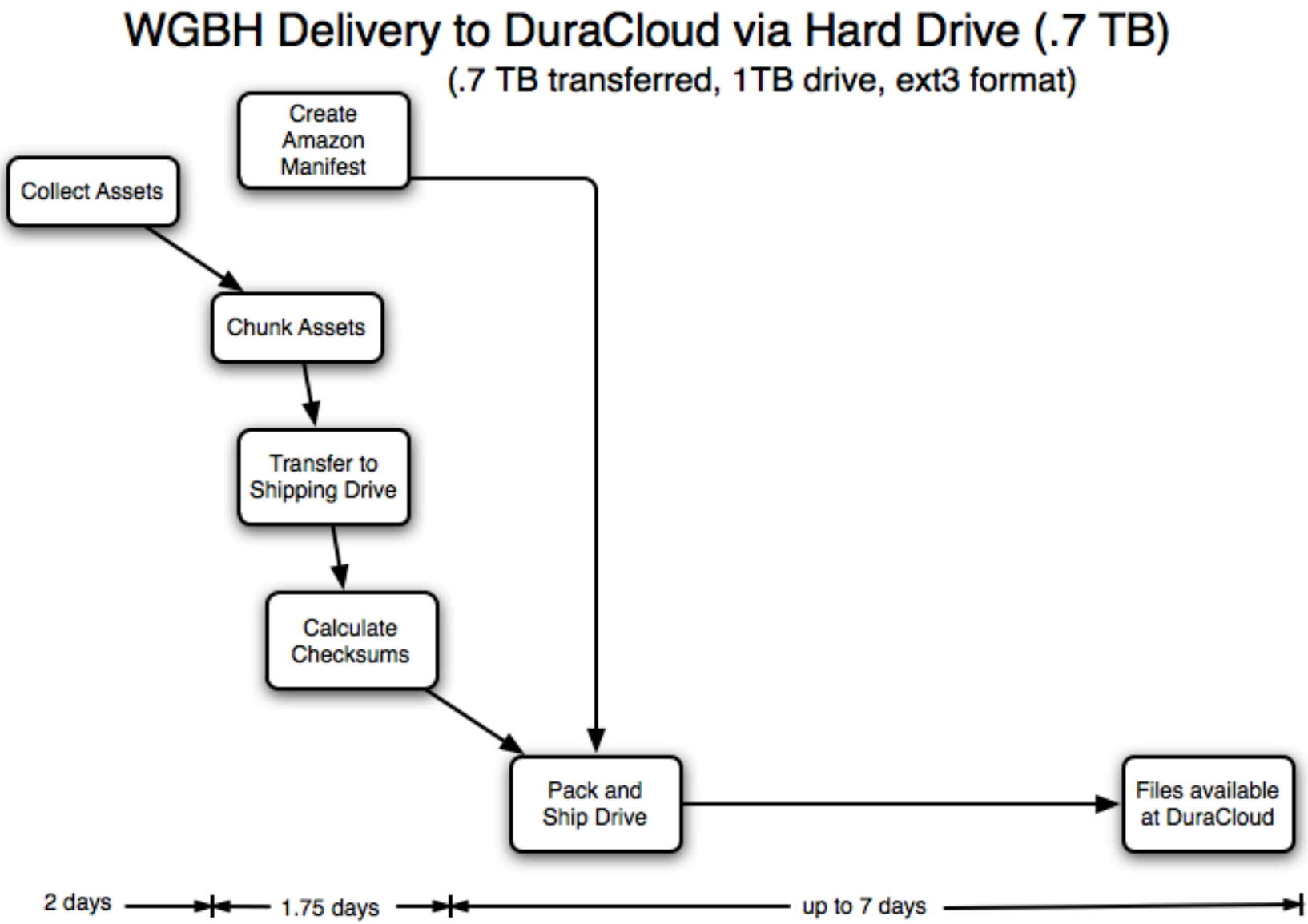
WGBH Access using DuraCloud



Access services:
Streaming
File format transformation
File access
collaboration

© 2010 WGBH

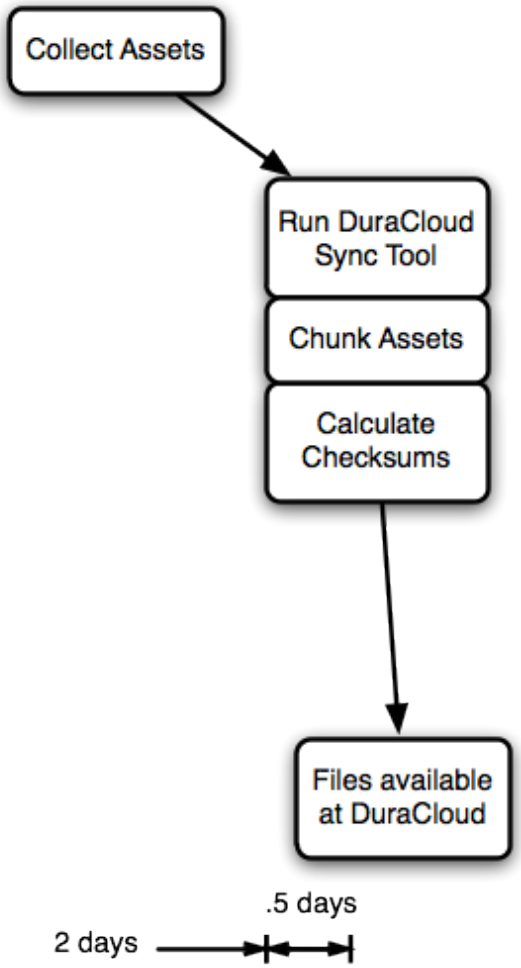
Sending disks to the cloud



Using tubes and wires

WGBH Delivery to DuraCloud via Internet

(.7 TB transferred, 300Mbps pipe)



Outcomes

- 5.5 TB of audio & video uploaded
  - Preservation and access files
- Still working with sync tool
- Streaming service works
  - But still need to integrate with Open Vault web site (for access)

© 2010 WGBH

Achievements during Initial Pilot

- DuraCloud integrated with 3 cloud storage providers
- Pilot partners loaded 30 TB into Duracloud
- Integrated and deployed multiple independent services
- Developed tools to overcome limitation of 5 GB file size and ease data loading

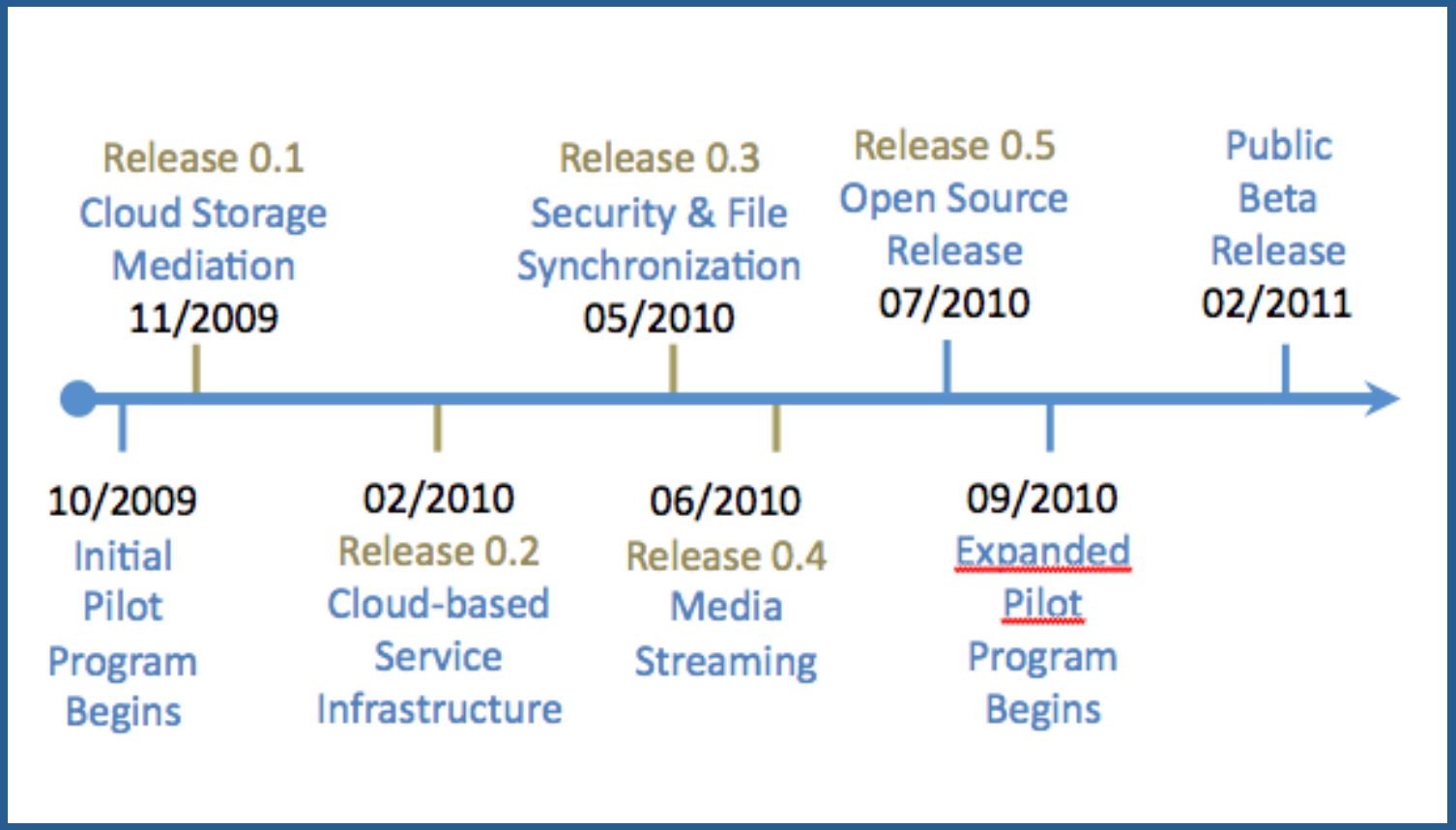
Lessons Learned

- Content transfer requires time and effort
- Internet Latency can be high
  - Minimize transactions across the wire
  - Data should be close to compute
  - Minimize data transfer with local host
- Storage more mature than compute
- Must allow for "eventual consistency"
- Market still developing

Extended Pilot Partners

University	Use Case	Repository
Rice U	Preservation	DSpace, meta archive
Hamilton College	Access/international collaboration	Fedora
Northwestern U	Preservation books, audio, image	Fedora
U of PEI	Image viewing/hosting	Fedora/Islandora
Cornell U	Data stream access and preservation	Fedora
ICPSR	Access and Preservation	Fedora
SUNY Buffalo	Preservation	DSpace
IUPUI	Preservation	DSpace
Rhodes College	Image Access	DSpace
North Carolina State U	Preservation	DSpace
CARL	Preservation and Services	Fedora
Orbis Cascade Alliance	Preservation and Services	DSpace
MIT	Preservation, OAIS compliance	Dspace

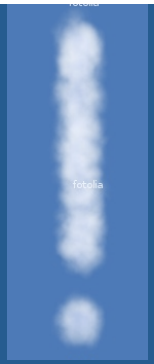
Key milestones



DuraCloud now available open source

- Open core
  - ✓ Open API
  - ✓ Open Source
  - ✓ Apache license
- Architecture to create cloud networks
  - ✓ Public clouds
  - ✓ Private clouds
  - ✓ University consortia
- Partner implementations/Integrations

Thank You



DuraSpace organization: www.duraspace.org
Wiki: wiki.duraspace.org/display/duracloud/
DuraCloud project page: duracloud.org
DuraCloud demonstration: link available at duracloud site
DuraCloud open source: wiki.duraspace.org/display/duracloud/DuraCloud