DuraCloud: Data Integrity Monitoring in the Cloud

Digital Preservation Partners Meeting
July 21, 2010

Andrew Woods
awoods@duraspace.org
Overview

• What is DuraCloud?
• Fixity service use case
• Basic flow
• Cost and performance
• Next steps
What is it?

• Cloud-based service offered by the not for profit organization, DuraSpace

• An open source, cloud storage/compute application
  – Focused on preservation support and
  – Data access for reuse and sharing

• Cloud storage across multiple commercial & non-commercial providers

• An open canvas for cloud-based services
Fixity use case

• DuraCloud user has replicated content across one or more cloud stores

• Need for periodic verification of bit integrity

• Seeking balance between cost & trust
0: Content Topology
1: Data load
1a: Replicate
1b: MD5 export
2: Determine MD5s*

...running fixity service
3: Compare & Report
0: Trust vs. Cost

Trust in...
- Underlying storage providers
- DuraCloud and opensource software
- Requester of service (administrator)
1: Trust vs. Cost

Three approaches:
- Request stored value
  [inexpensive & fast]

- Stream out content & re-calculate
  [compute intensive & slow]

- Stream out content & re-calculate with salt
  [user intensive, compute intensive & slow]
2: Determine MD5s*

Options for providing expected MD5
• With initial listing
• After MD5 calculation
2a: MD5 at non-primary

Additional cost of processing content not local to compute
Next steps

• Scalability
  – MD5 calculation across Hadoop cluster

• Multi-administration efficiency
  – On-demand compute at secondary provider

• Event logging
Thank you

Requesting comments & review

https://wiki.duraspace.org/display/duracloud/Fixity+Service

http://duracloud.org
https://wiki.duraspace.org/display/duracloud/DuraCloud
https://svn.duraspace.org/duracloud/trunk/