Cloud Task Replica

Repository Preservation Tools
Open Repositories 2009 - Atlanta
Richard Rodgers
MIT Libraries
cloud computing

- dynamic capacity: elastic
- high availability
- > storage: compute, database, more
- new programming model
- WOA - service bus in the sky
- lightweight protocols
problem space: replication

- replication != backup
- time decay of trust - needs maintenance
- coordination costs $$$$ 
- who’s watching the detectives ?
- impermeable system boundaries
- sizing forecast uncertainty
reliable messaging

- enables asynchronous handling
- queue = list of messages
- coordination of work, non-persistent
- access controlled, encryptable
- cheap: $0.01 per 10k messages
- Amazon SQS + S3
roles

- decompose work into distinct replaceable agents
- archive = content home
- replicator = manages copies
- auditor = implements and enforces policy
- role != institution
process model

- a message queue for each role
- message post triggers activity asynchronously
- bucket brigade - message is a handoff or acknowledgment
- storage is abstracted (cloud in prototype)
workflow: replication
workflow: removal

- archiver
- replicator
- auditor
- S3
workflow: audit
message semantics

- web-standard URI addressing
- entities: packages, ORE maps
- content model agnostic
- entity checksums for integrity
- standard identifiers for actors
self-managed deployment

mit
peered deployment

mit
gatech
service provider deployment

mit

mit

mit
todo

- plumbing only - replication requires more
- all policy definition and agreements OOB
- address business model
- content packaging/description
- expand skeletal prototype
- stress at scale
opportunities

- federated & large scale problems
- distributed registries
- metadata harvesting
- subject overlays
- preservation workflows, micro-services
thanks
extra credit