Object Storage
Some Input to Brainstorming discussion

Dirk Duellmann, IT-DSS

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Object Stores

- Basic Idea
  - store user defined variable length byte sequences (objects) rather than fixed length blocks
    - abstracts lower level of media handling - like a file
    - but with constrained data modification semantics
      - eg create, read, delete
      - no update
  - Usually implementing media redundancy
    - mostly now using distributed object replicas
      - avoiding eg RAID recovery problems
    - several plan to add erasure-encoding (Reed Solomon or more advanced)
      - to be more space efficient
  - identified by object ID with simpler semantics than eg posix file name semantics
    - eg no (scalable) iteration over content
      - application side keeps track of object cataloging
- Goal:
  - locally clustered store which scales better than posix (eg NAS)
  - in access performance, price and operational effort

For example

- CEPH
  - redundant object storage with client side calculated (more scalable) placement decision (CRUSH)
  - RADOS - native access
    - S3 / Swift via gateway -> scalability impact?
  - additional consolidation possibilities for sites
    - Block storage (eg for VM local space) used in IT AI project
    - CEPH File System
      - not yet supported - but “almost awesome”
- Interest from several projects to evaluate
  - CASTOR: match high-speed tape drives to “slow” disk cache for migration/recall

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Object -> Cloud Storage

- Semantically similar
  - but typically accessed via http extensions like S3 or swift
    - may tie-in easier with existing http caching components like SQUID
  - trivial namespace scaling via bucket separation
    - user chooses placement via object name (url)
  - commercial storage-as-service offerings and quasi-standard via Amazon docs exist
    - advantage: if “standard” service offered by a larger set of sites is needed
    - likely more suitable for volume scalability than single client performance
      - this depends more on the backend implementation than the access protocol

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Server-less Object Storage

- Eg Seagate Kinetics Drive
- Single disk talks object storage protocol to client over a direct TCP/IP port
  - and organise replication/failover with other disks in a (LAN) networked disk cluster
  - open access library for app development
- Other disk vendors are probably (re-)evaluating this approach
  - Why now?
    - shingled disk technology comes with natural match in semantic constraints: eg no data/metadata updates
  - Early stage with several open questions
    - port price for disk network / price gain via reduced server CPU?
    - standardisation of protocol/semantics to allow app development at low risk of vendor binding?

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it