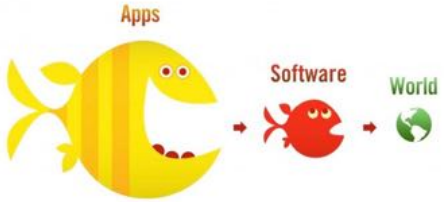
SCALITY

The Infrastructure of Petabyte-Scale Scientific Data Archiving

Bradley King Chief Architect

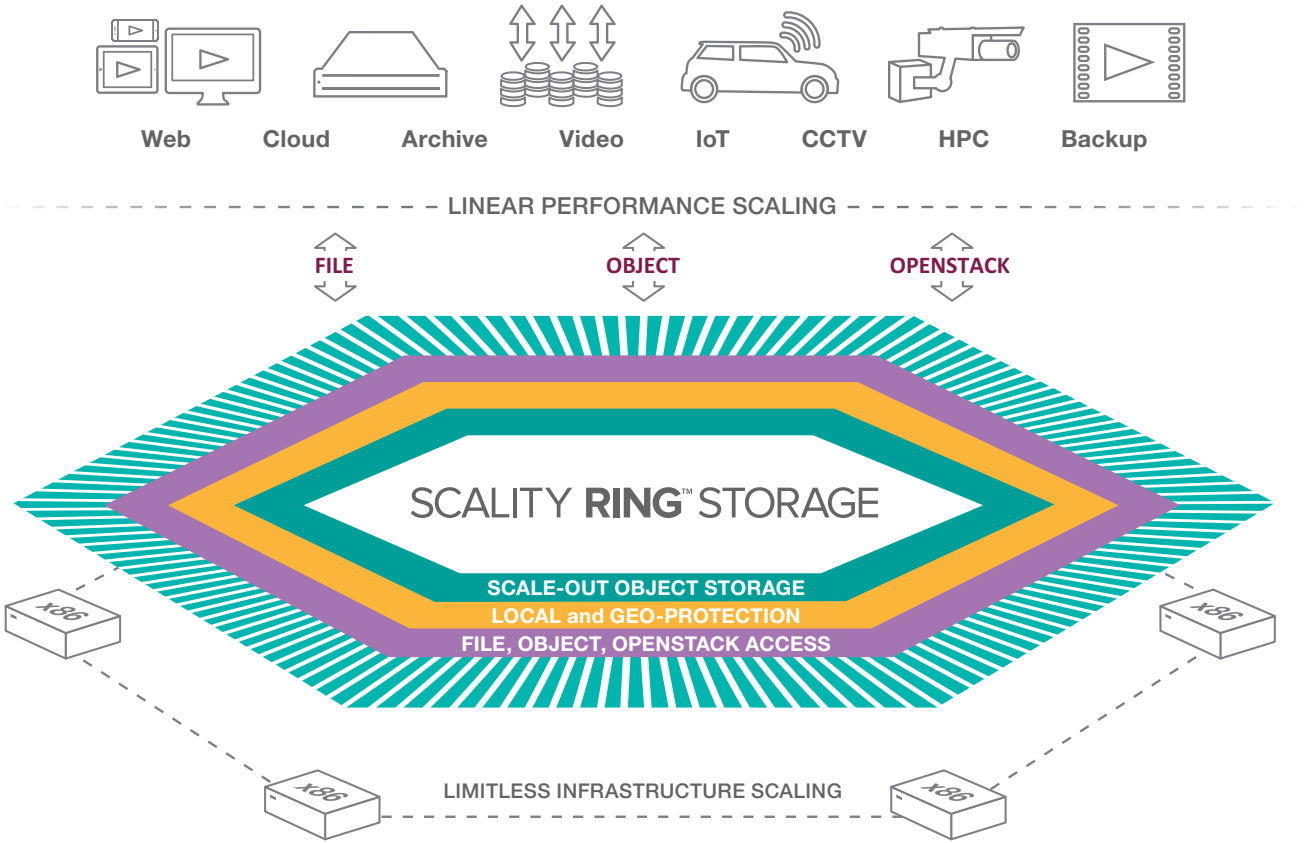
TERATEC – Juin 2016

Some Easily Visible Trends



- Digital Data has become a key source of knowledge for scientific discovery and business opportunity: The Human Genome Project, Fraud Detection, BlaBlaCar, Square Km array, Connected Automobiles, etc.
- Storage volumes are growing fast! > 50% annually: Multi-petabyte archives that need frequent access are becoming common
- As Marc Andreessen says “Software is Eating the World” – Hardware is becoming increasingly standard and thus a commodity
- As Pat Helland of Salesforce.com states “Accountants don’t use erasers” It has become too expensive to delete data!
- Ubiquitous Access to Data is increasingly important
- Hard Drives are getting larger/slower

Scality's Technology is Object Based



The **Scality RING** is a software-defined storage

20-Byte Object Key-space

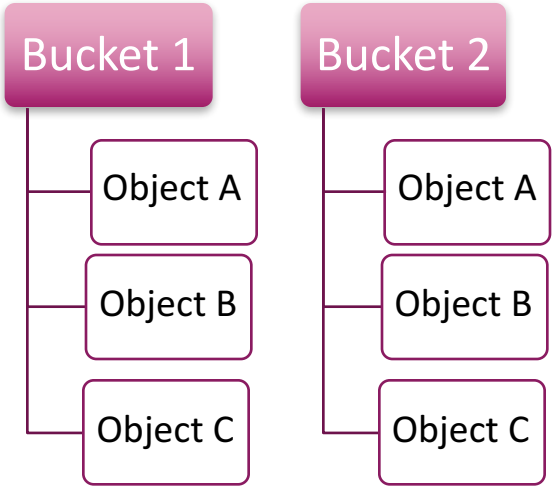
Fully Distributed System

Flexible Replication and Erasure Coded Protection Schemes

Filesystem on Object or Pure Object

SCALITY

Data Models: Object vs Posix

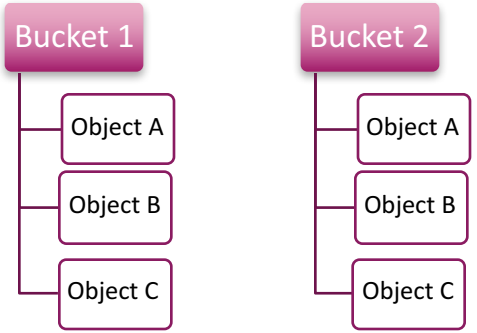


Part of the filesystem tree

- Flat Data Model with collections called "Buckets" or "Containers"
- Objects are written and overwritten not byte-wise modified

- Hierarchical Data Model
- Files are randomly writable in byte-wise fashion

Locking and Consistency Comparisons



Object Storage

- Eventual Consistency is *Acceptable*
- Last writer wins is standard behavior

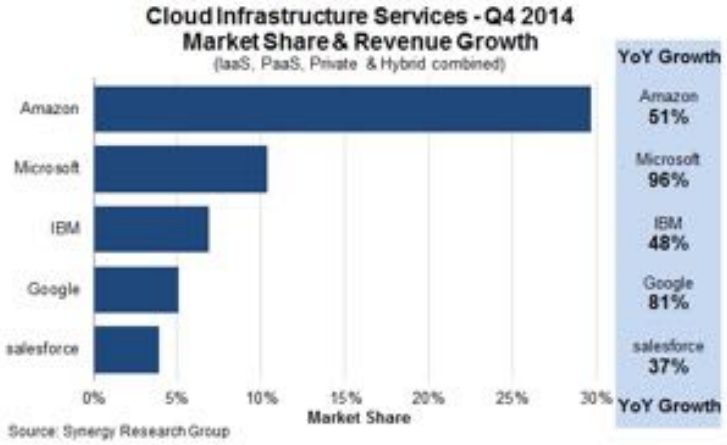
Posix

- System-wide Consistency is required
- Range locking is expected

Red Blue Consistency – Much like Amdahl's law – Coherency is Expensive
Principle: Can we ask applications to manage coherency? If so how?
DAOS and others trying to answer these questions…

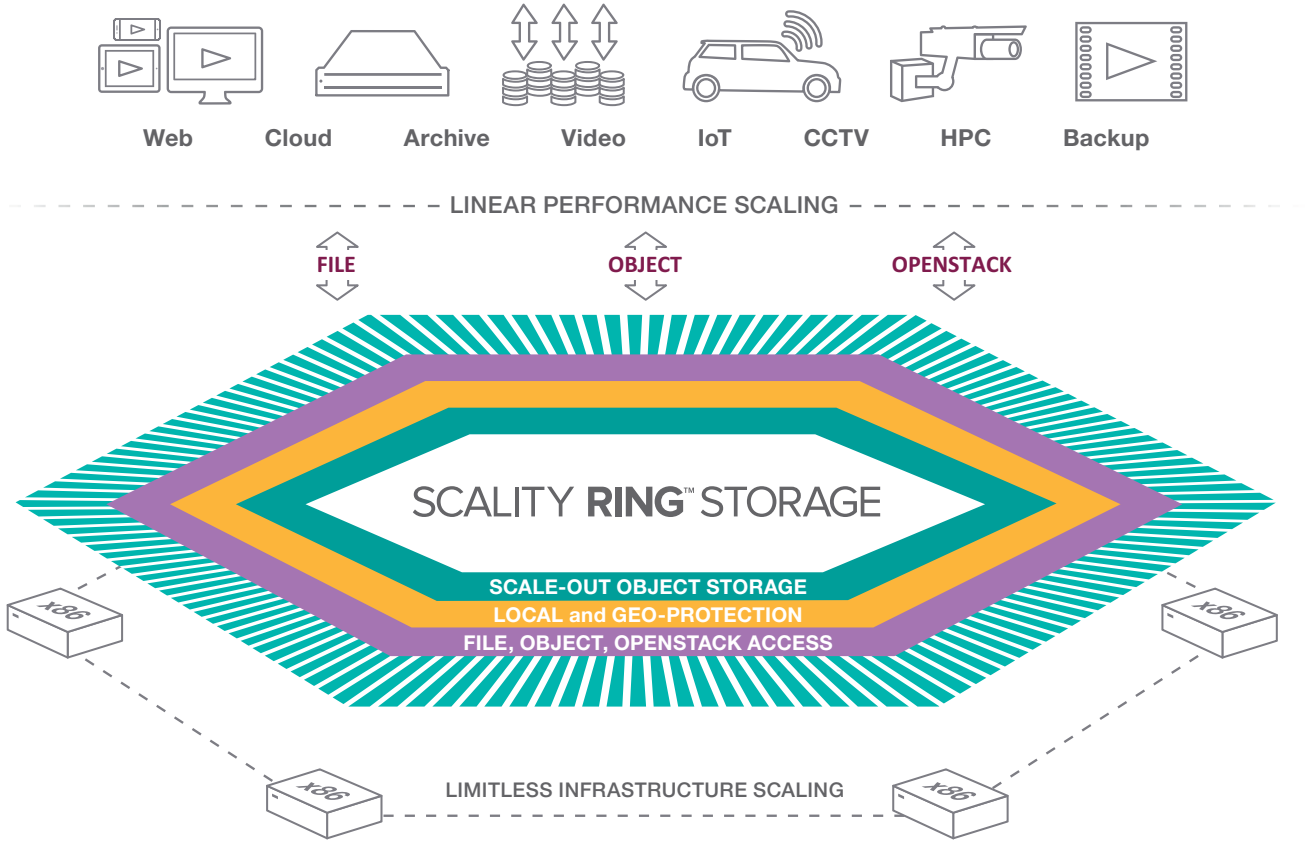
Which Object Interface?

- Amazon S3 is becoming the de-facto standard for object storage at the expense of Swift and CDMI
- AWS is *the* dominant player in IaaS – S3 is *the* storage of the Cloud!
- Developers develop for the leader(s)



SCALITY

Where is Scality Headed?



The **Scality RING** is software-defined storage

We're working at being 100% reliable, and unlimited in scale

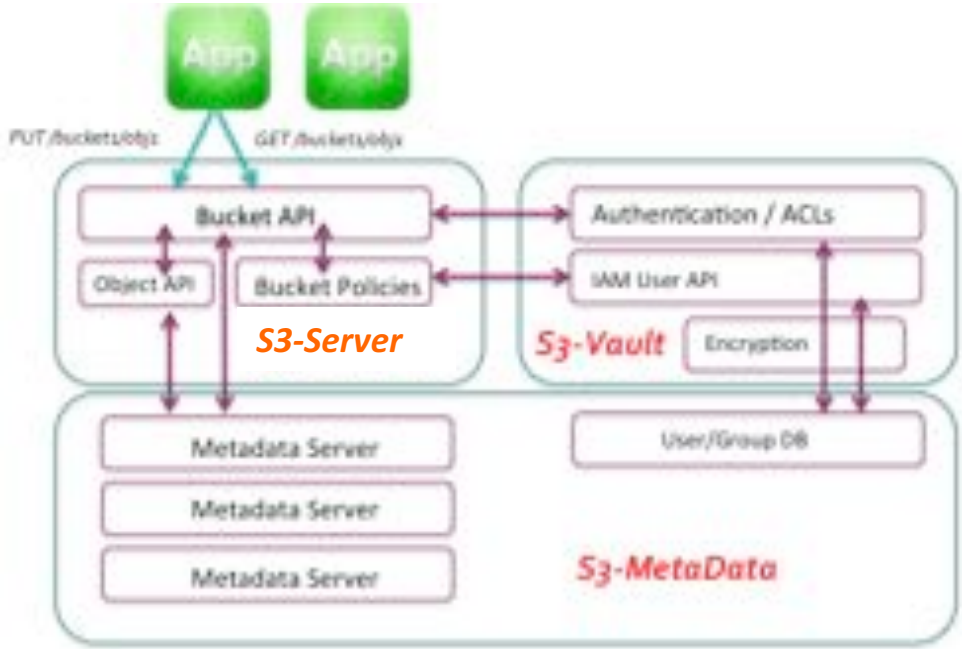
We're embracing Object Storage Models

We've become expert at managing distributed data storage

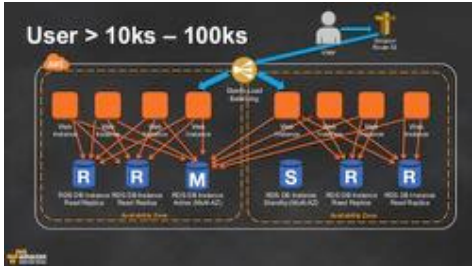
SCALITY

S3 Connector: Three Key Components

- **S3-Server**
  - S3 compatible API servers
  - Responds to standard S3 http requests
  - Standard S3 headers & response codes
  - Multi-connector scale-out
- **S3-MetaData**
  - A distributed metadata database service
  - Supports fast Bucket & object listing
- **S3-Vault**
  - Security, Identity & Authentication Service
  - Provides Accounts/Keys
  - Supports S3 IAM Users, roles
  - Interoperable with AD for directory services (via SAML)

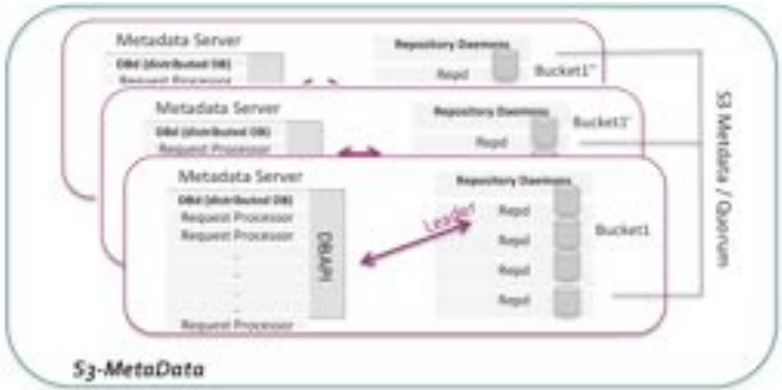


S3 Connector: Scale-Out



- Bucket Scale Out – at Cloud Scale
  - Buckets may be users, groups, departments, companies or customers
  - May be O(100's) or O(1,000,000) depending on deployment model and use case
  - O(billion) objects per Bucket
  - O(1,000) HTTP clients per Bucket with scalable sharing / coherency across connectors

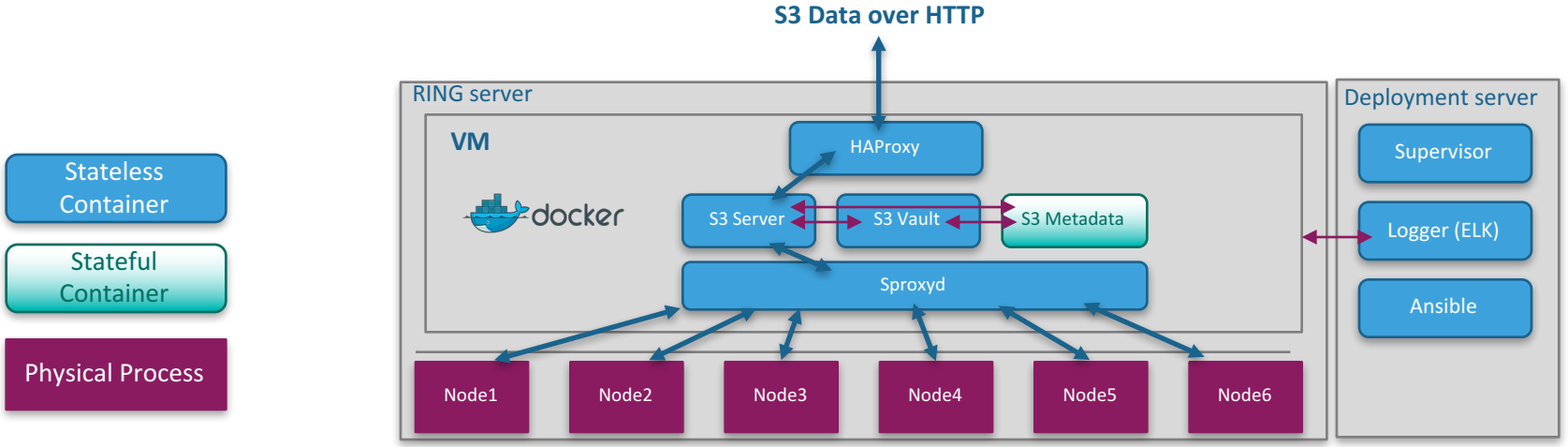
- S3 Distributed Bucket Metadata Engine (the really hard part)
  - Scalable and high-performance metadata engine
  - Enables distributed updates to Buckets from multiple "S3 Connectors"
  - Highly-available clustered design to provides consistency & availability after failures



SCALITY

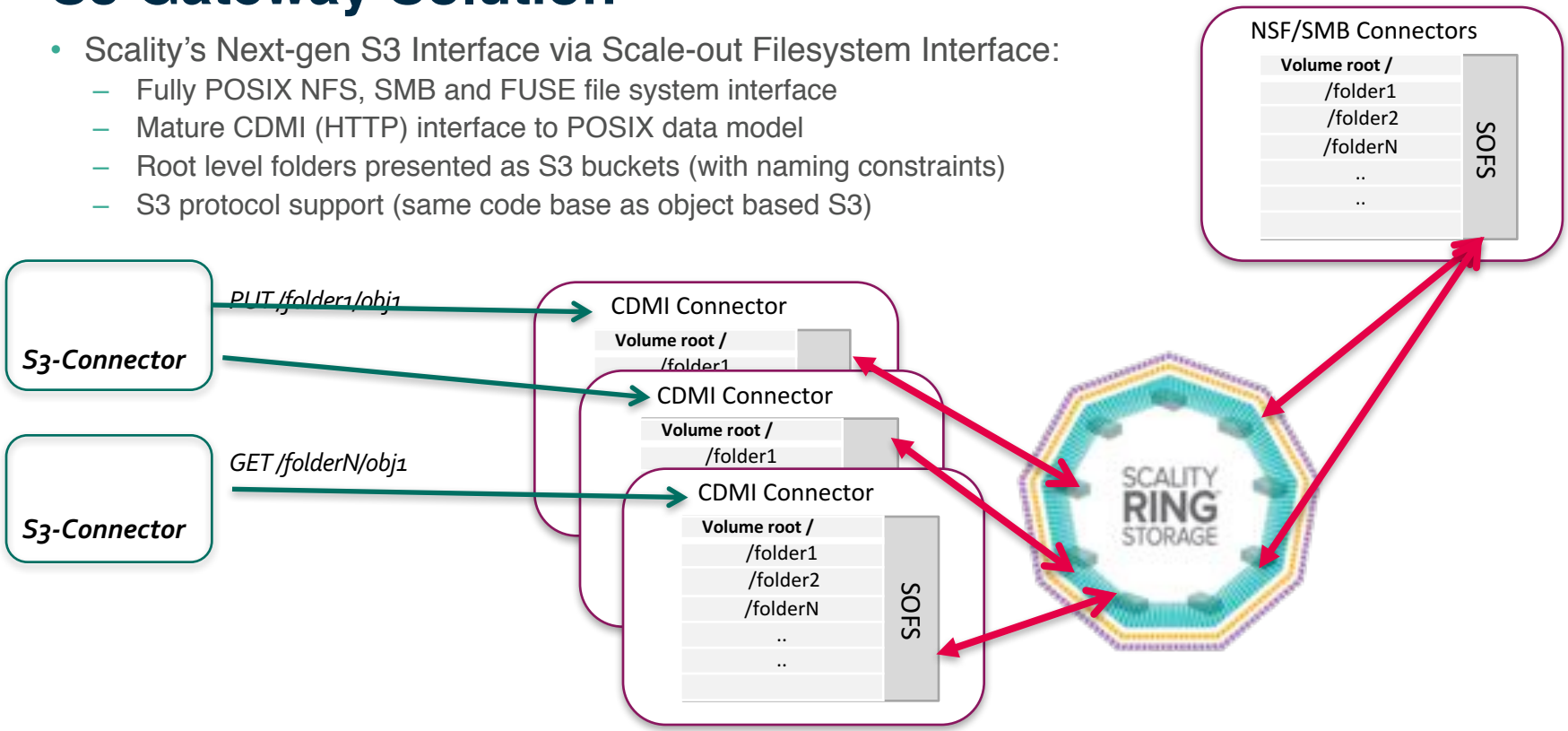
S3 Connector: Container Architecture

- S3 installs as a set of distributed services in Docker Containers
  - 8 cores, 32GB RAM per machine recommended
  - SSD for Metadata database (sized on #keys & avg. object length)
  - All network traffic may be encrypted over HTTPS/SSL
- Deployment server
  - hosts Supervisor, Logger and Ansible framework for federated deployment



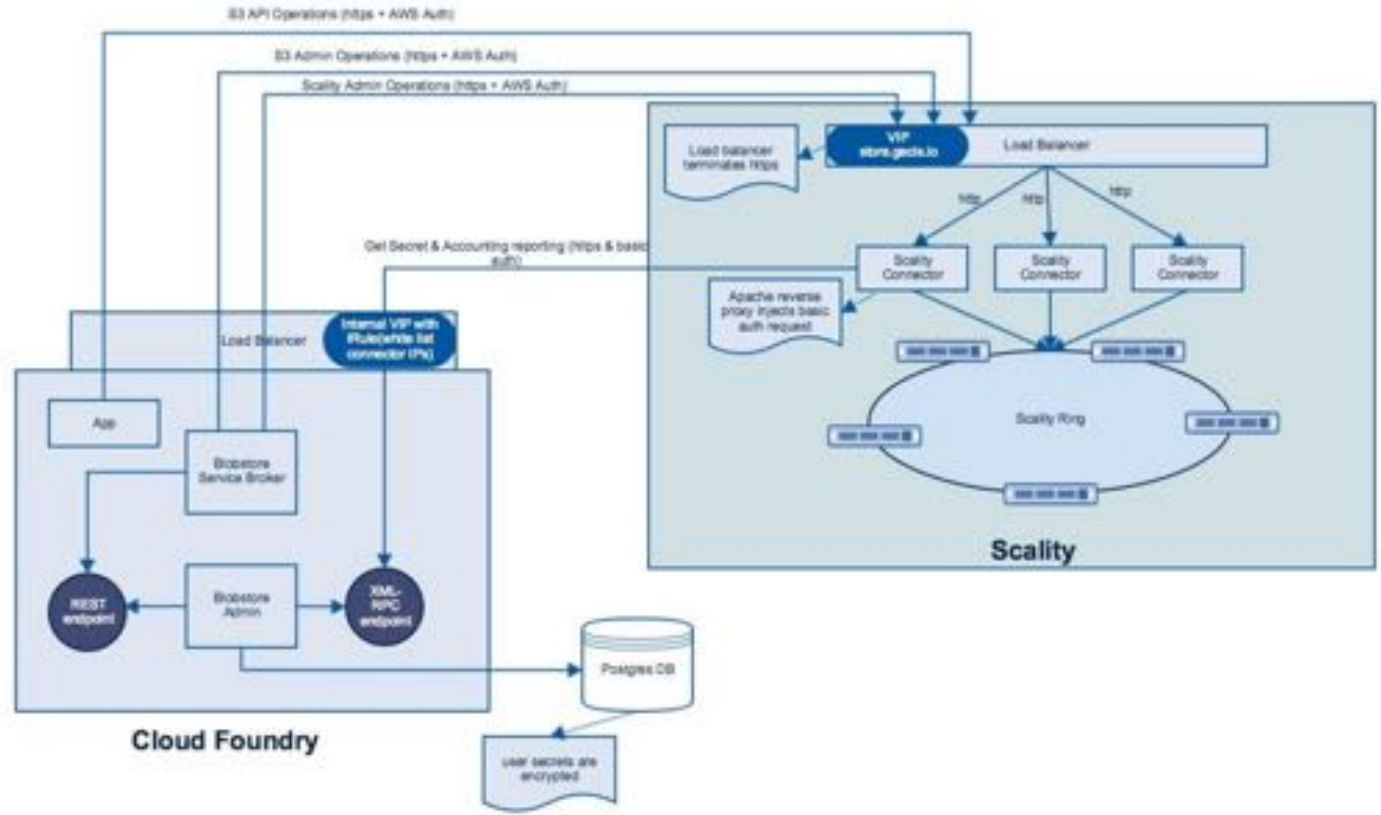
S3 Gateway Solution

- Scality's Next-gen S3 Interface via Scale-out Filesystem Interface:
  - Fully POSIX NFS, SMB and FUSE file system interface
  - Mature CDMI (HTTP) interface to POSIX data model
  - Root level folders presented as S3 buckets (with naming constraints)
  - S3 protocol support (same code base as object based S3)



Some Examples

Example IOT Architecture with Scality RING



Automobile Manufacturer with Scality RING:

Strong parallel performance for R&D application

Stores millions of kilometers of video and sensor data at lower cost

Scaled single system to **8 petabytes with mixed size servers**

SCALITY

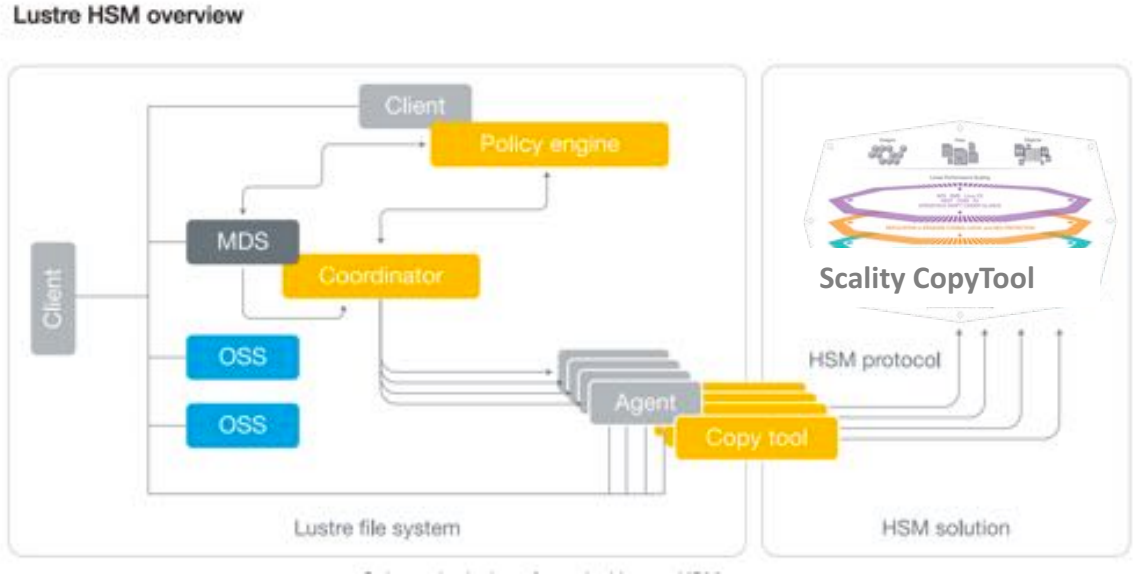
Lustre HSM Support over based CopyTool

Integrated Lustre HSM enables automated storage tiering support

- Policy manager based on “Robinhood” (developed at CEA)
- Tier 1: Scratch (100’s TB)
- Tier 2/3: Home & Archive (PB -> 10’s PB)

Scality has created an Open Source CopyTool

- Uses the published Scality Droplet library (with Sproxyd, CDMI & S3 profiles)



LANL Trinity Storage Architecture

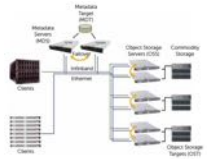
Data movement

Data movement

Data movement

MarFS

Data movement



Server Memory Data Source
- Technology: RAM
- Capacity: 2PB
- Residence: Hours
- Overwritten: Continuous

Burst Buffer
- Technology: SSD Arrays
- Capacity: 3PB
- Residence: Hours
- Overwritten: Hours
- Movement In: Automated
- Data Move: Application
- Performance: 4-6TB/s

Parallel File System
- Technology: Lustre FS
- Capacity: 78PB
- Residence: Days/Weeks
- Flushed: Weeks
- Movement In: Automated
- Data Move: Application
- Performance: 1-2TB/s

Campaign Storage
- ***Technology: Scality RING***
- Capacity: 30PB
- Residence: Months/Year
- Flushed: Months/Year
- Movement In: Manual
- Data Move: User
- Performance: 30GB/s

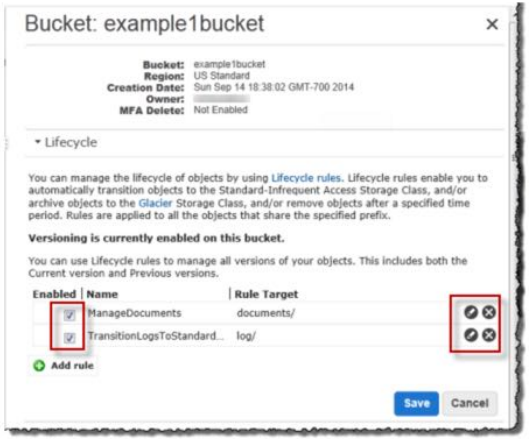
Archive Storage
- Technology: Tape
- Capacity: 50PB
- Residence: Forever
- Flushed: Never
- Movement In: Manual
- Data Move: User
- Performance: 1-10GB/s

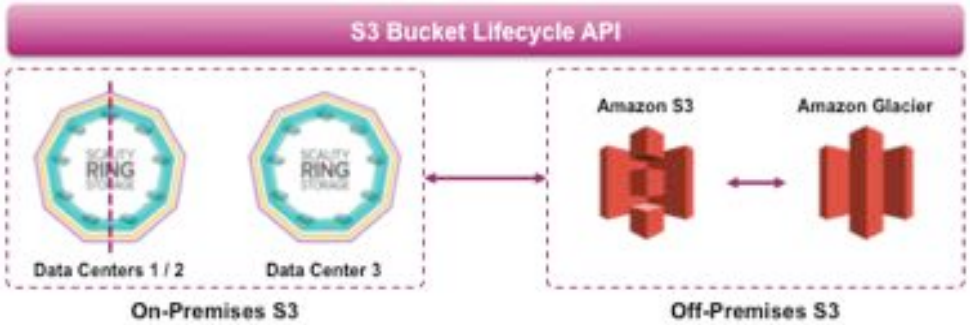
SCALITY

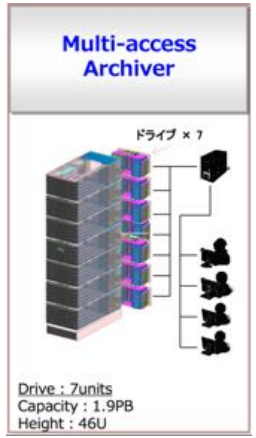
S3 Connector: Information Lifecycle Management

- Application Data has different value over time
  - Frequently used: Tier 2 RING or S3
  - Older: capacity-optimized Tier2 with reduced redundancy (S3-IA)
  - Oldest/compliance: cold-storage (e.g., Glacier)

- The S3-Connector will enable the Bucket Lifecycle API
  - Enables data lifecycle management: expiration or transition
  - Rules per Bucket - transition objects to any S3 compatible Bucket
  - Lifecycle supports “Tiering” across multiple S3 compatible Buckets (Bucket-to-Bucket OR on-premises-to-Cloud)







SCALITY

Scality S3 Server Opensource – **s**3.scality.com

1) Gain better understanding of S3
2) Give developers easy access and avoid shadow IT
3) Develop to an identical scalable interface
4) Open source access to code
5) Further increase adoption of S3