# One Ring Cannot Rule Them All

**Gary Ogasawara**
**Cloudian Inc.**

© 2014 Cloudian, Inc. All rights reserved.



Cloudian

# Motivation

© 2014 Cloudian, Inc. All rights reserved.

# Why Object Storage?

To the application user, the logical “object” matters,

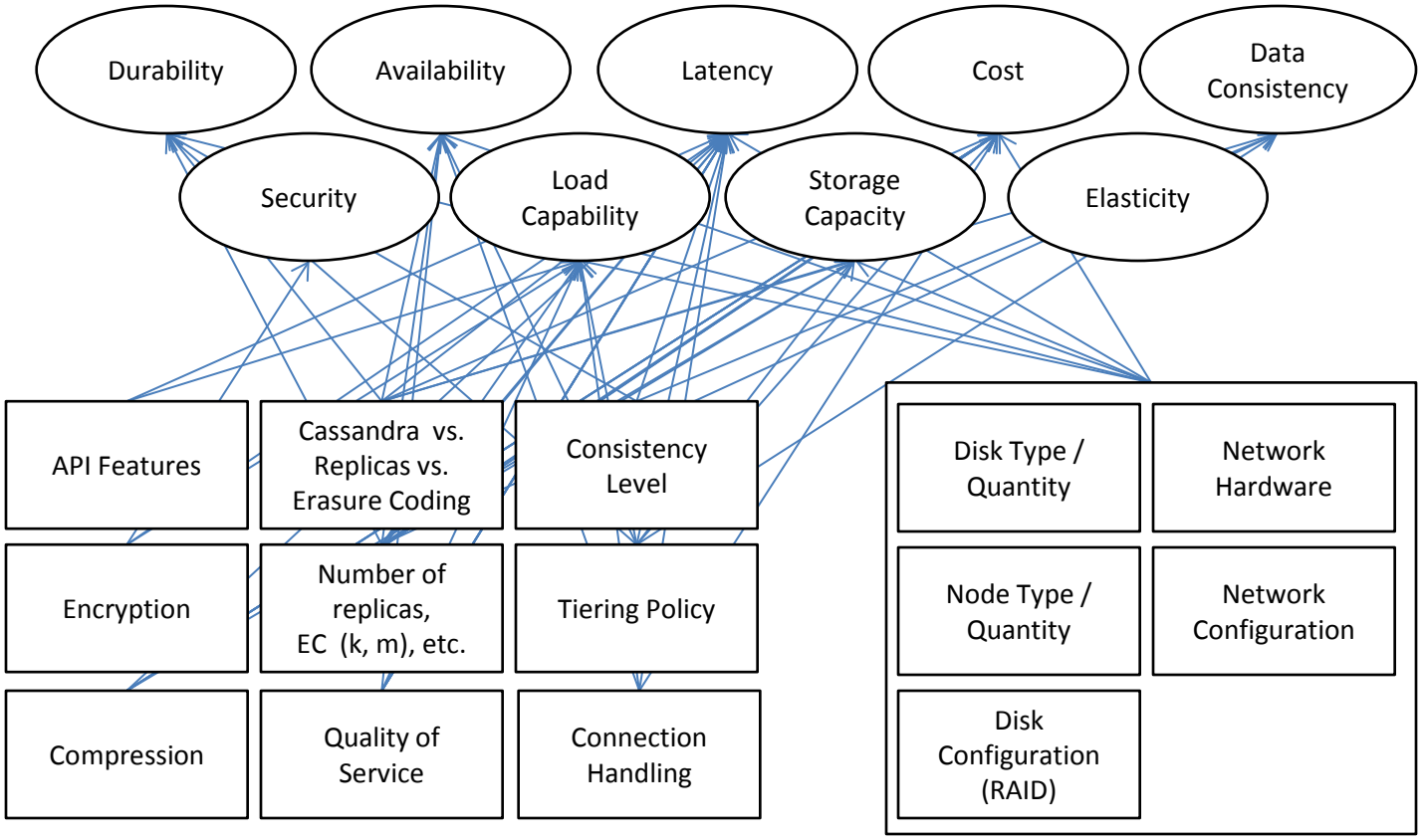
Not how it’s physically stored (e.g., pieces, versions, location).

© 2014 Cloudian, Inc. All rights reserved.



Cloudian

# How Object Storage?




*Random variables to optimize*

*Control inputs*

To optimize overall system need to be able to have fine-grained and dynamic control inputs.

© 2014 Cloudian, Inc. All rights reserved.



Cloudian

# HyperStore Policy Engine

Store and tier objects intelligently using rules and current information ⟹

*Dynamic storage policies optimized for each object*.


© 2014 Cloudian, Inc. All rights reserved.

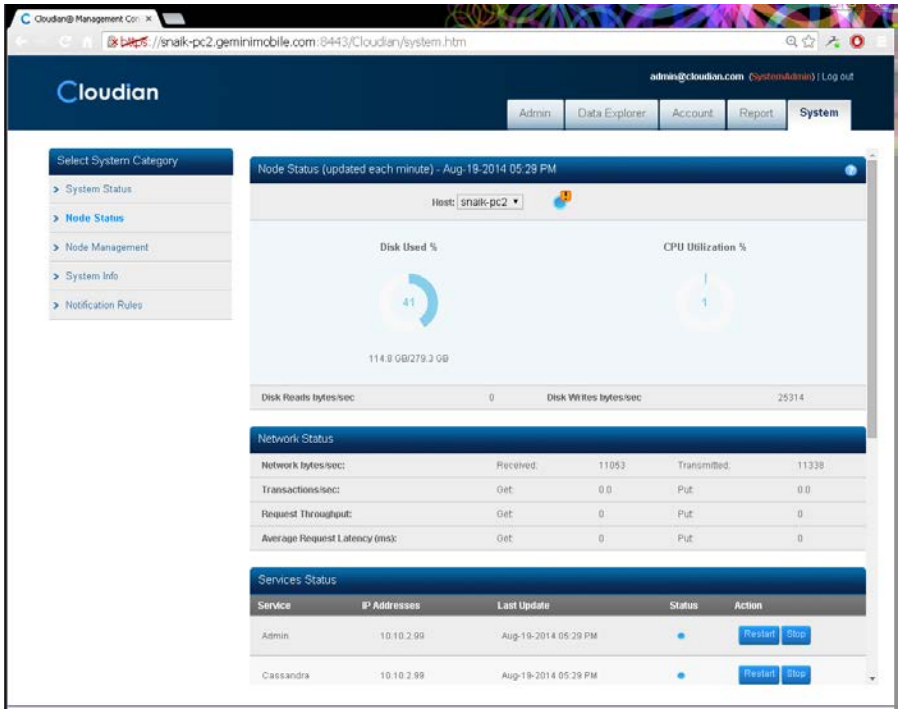
# System Overview

© 2014 Cloudian, Inc. All rights reserved.

# System Overview

1. **S3 Compatible** – Basic and advanced Amazon S3 features implemented.
2. **Optimized** for any workload
3. **Small Start/Elastic:** True P2P Architecture, 10TB to Exabytes
4. **Reliable**: True Multi-DC, Dynamic Consistency, AWS Location Constraint
5. **Turnkey:** Rich, flexible management features (QoS, Reporting, Chargeback, Admin GUI & API)

This talk

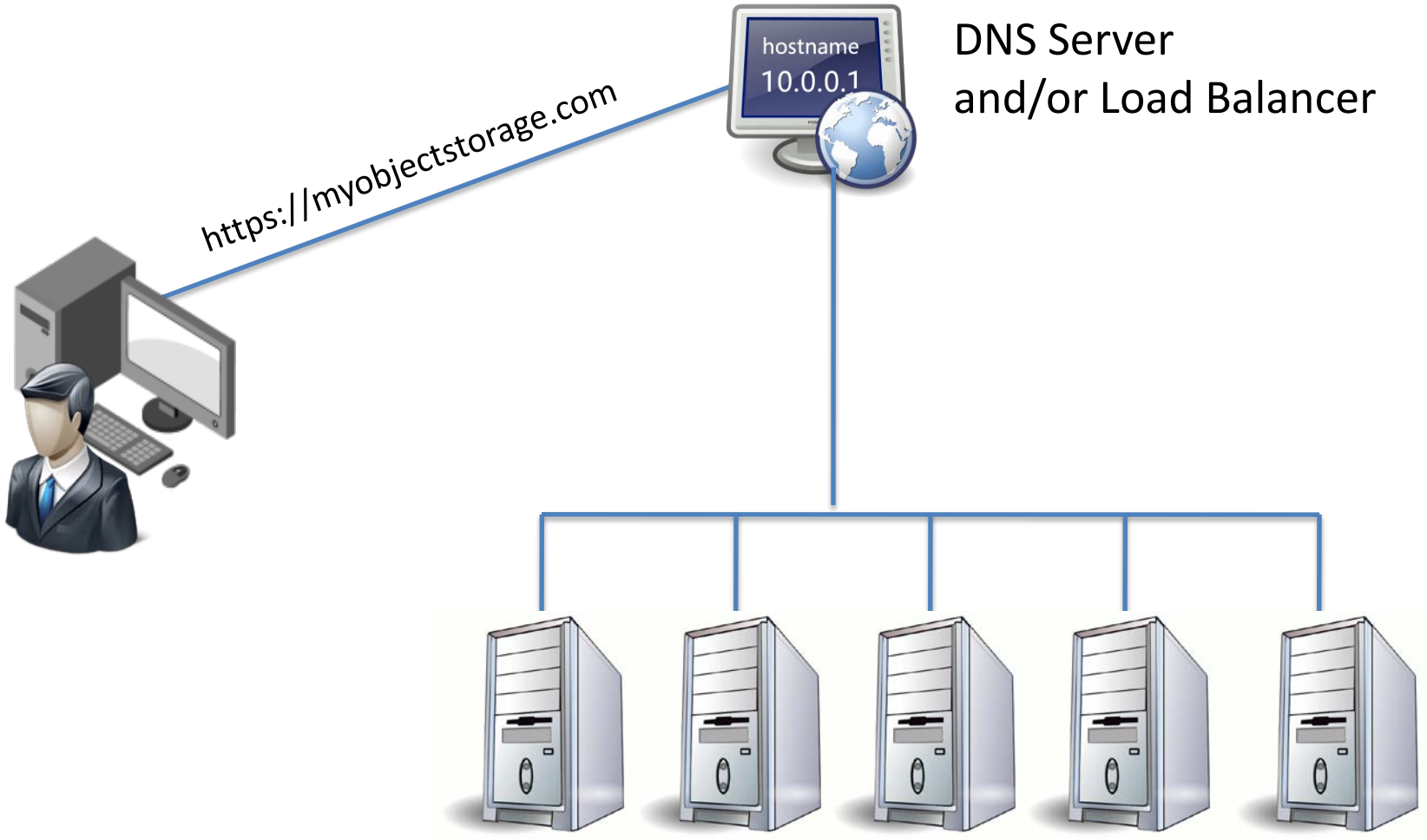




© 2014 Cloudian, Inc. All rights reserved.



Cloudian

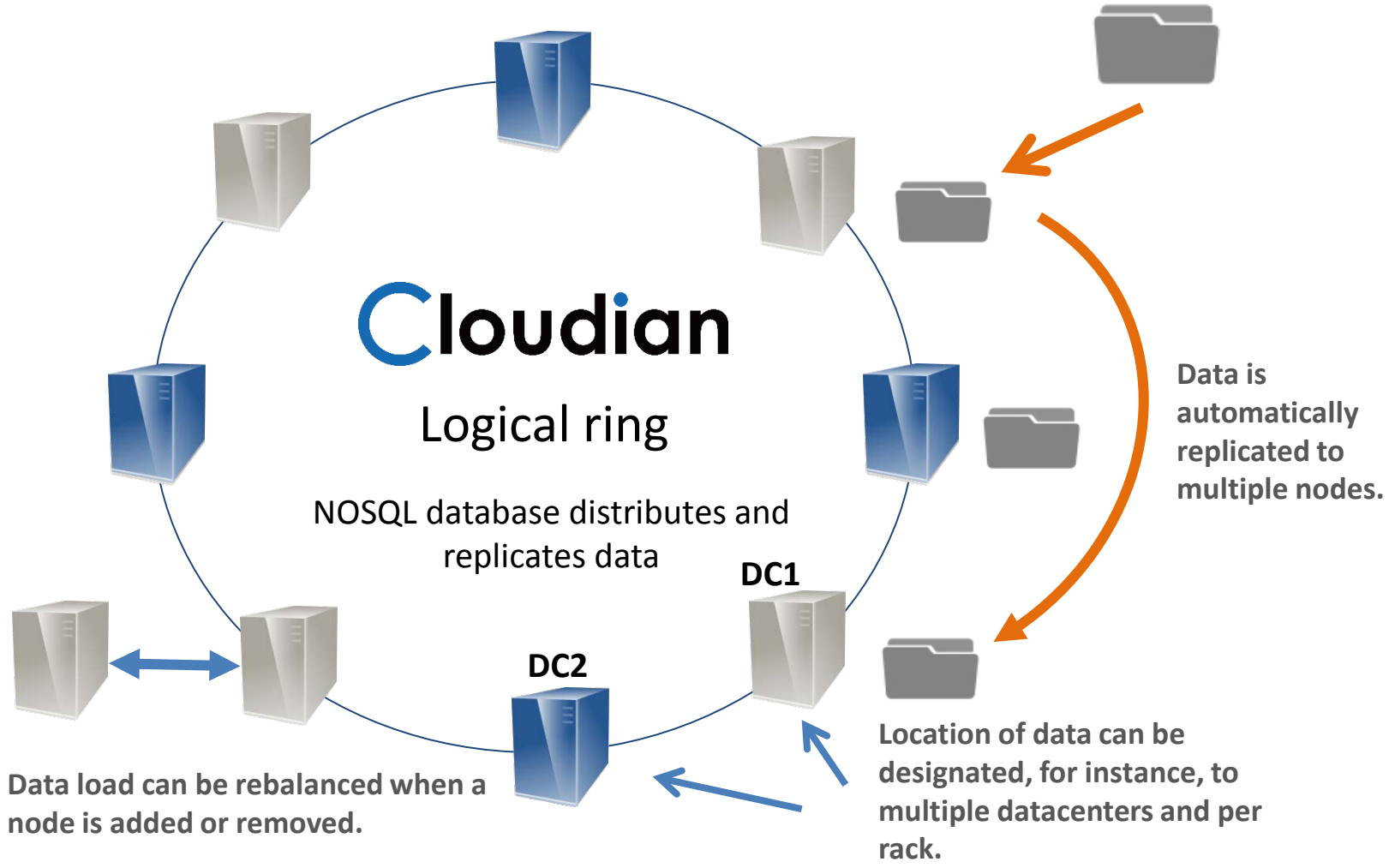
# High-level system view




© 2014 Cloudian, Inc. All rights reserved.

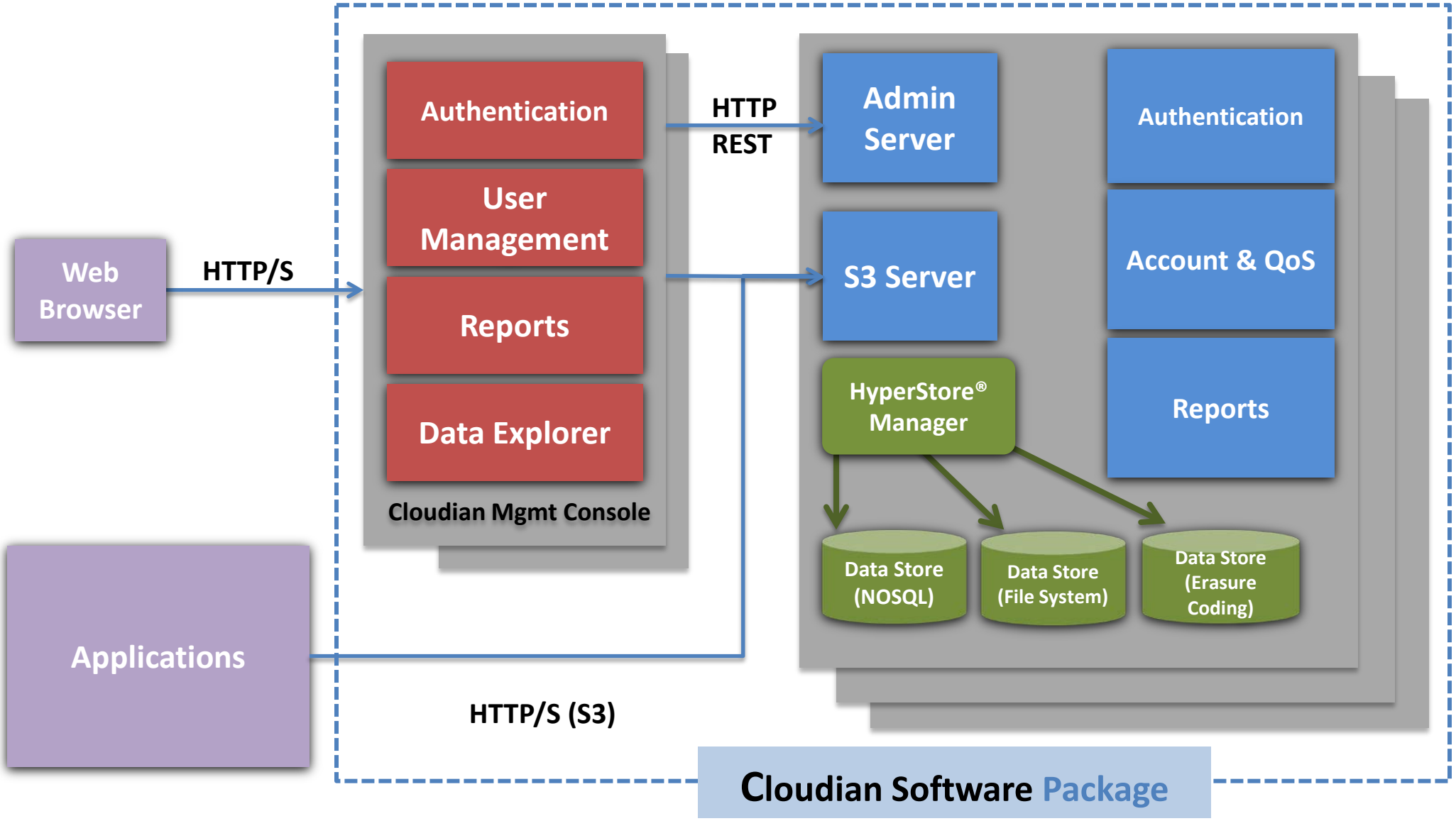
Cloudian

# Elastic, distributed and reliable




© 2014 Cloudian, Inc. All rights reserved.

# Logical architecture

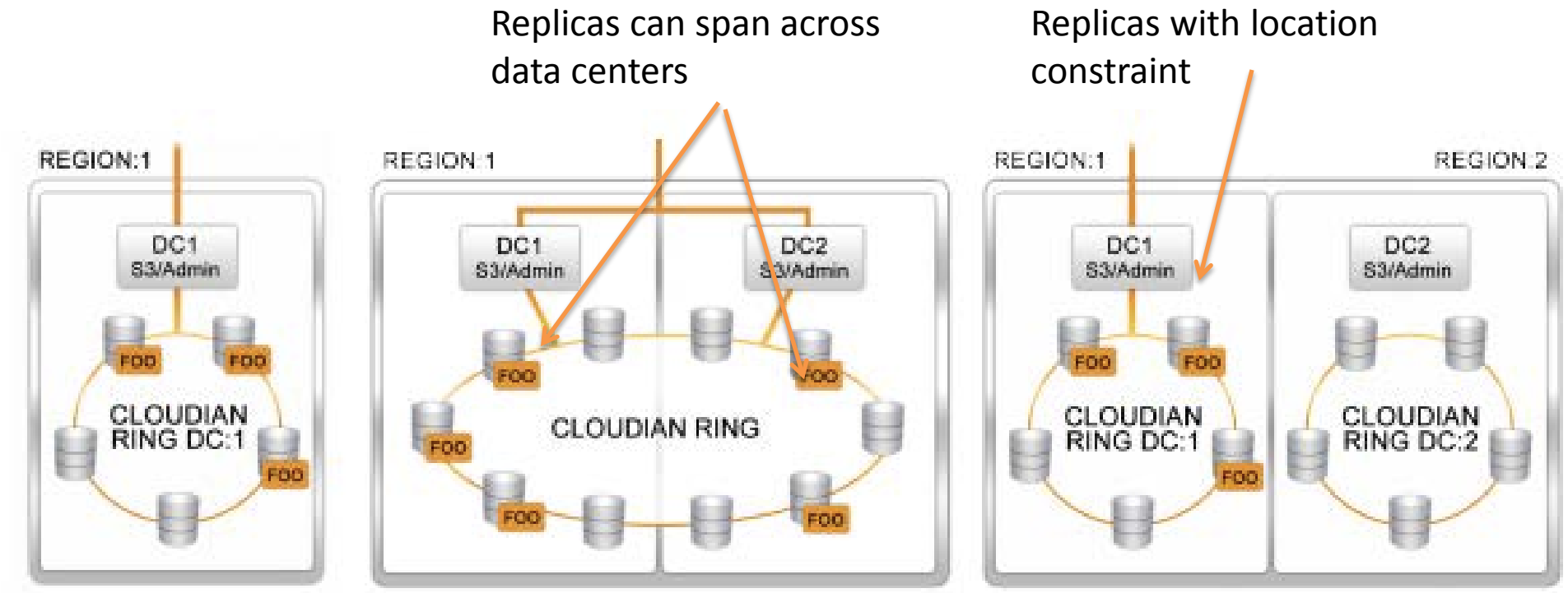



© 2014 Cloudian, Inc. All rights reserved.



Cloudian

# Multi data center




**Single-Data Center Single Region**

**Multi-Data Center Single Region**

**Multi-Data Center Multi Region**

© 2014 Cloudian, Inc. All rights reserved.



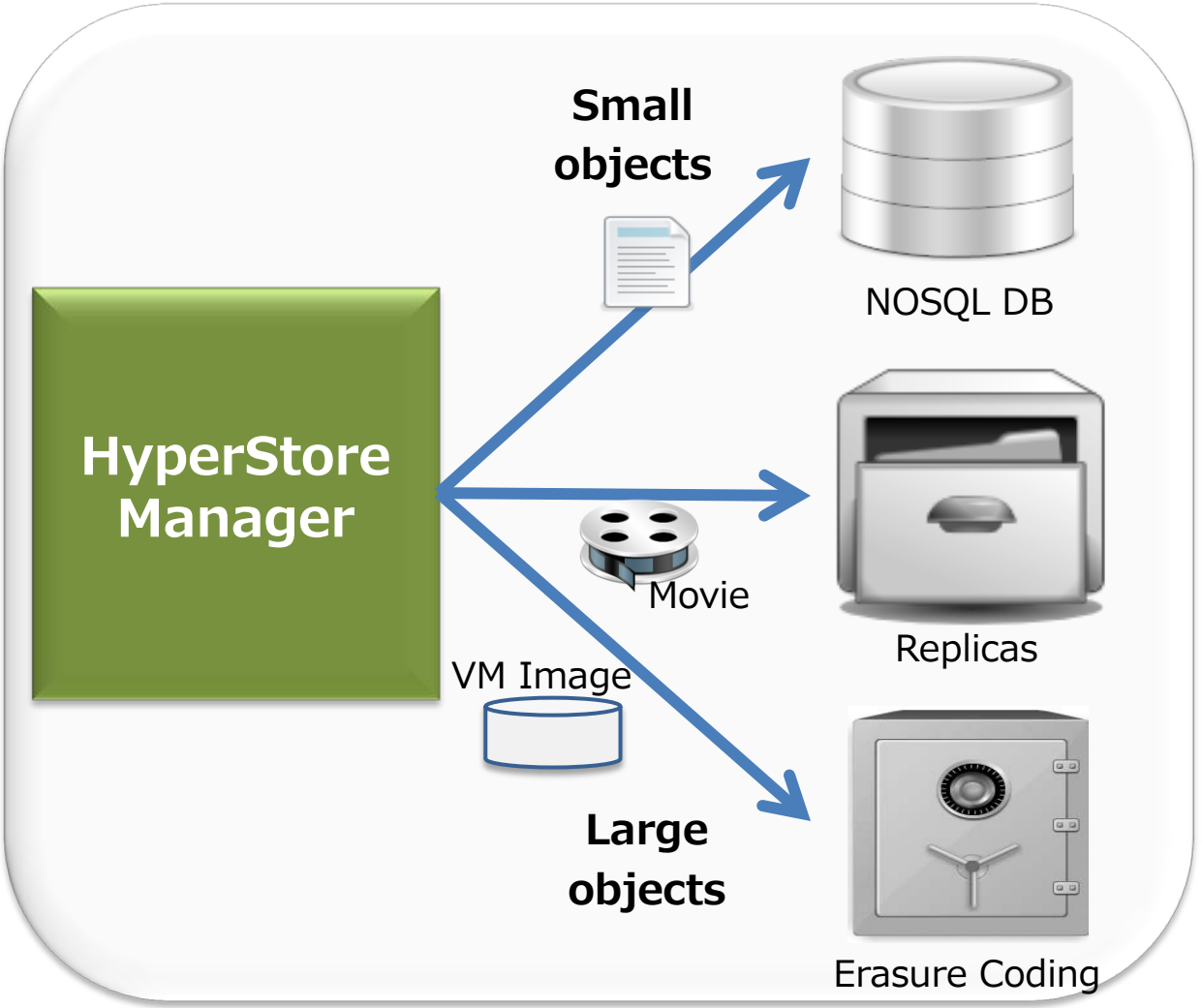
Cloudian

# HyperStore: Storage Type

© 2014 Cloudian, Inc. All rights reserved.

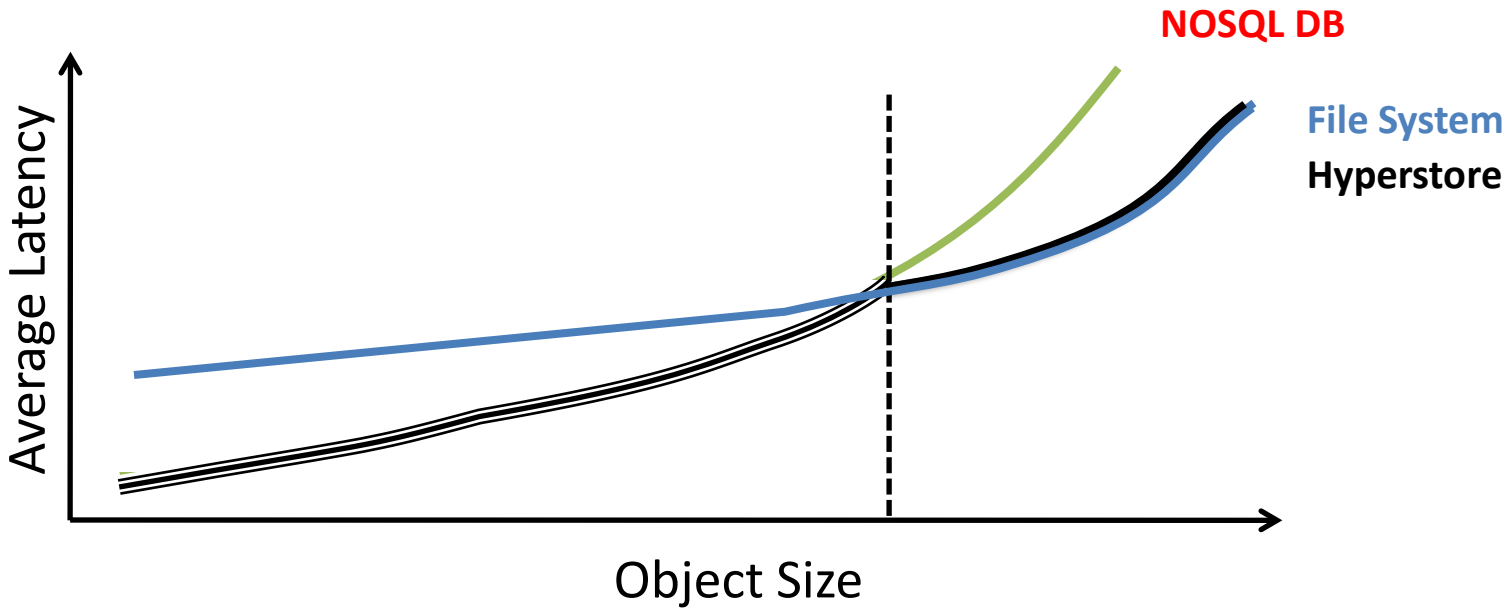
# Policy: Storage Type

- Policies tailored for different object types
- Large object support
- 30% faster reads
- 400% faster writes
- 4x more TPS
- 1.6x better disk utilization




© 2014 Cloudian, Inc. All rights reserved.

# Average latency vs. object size

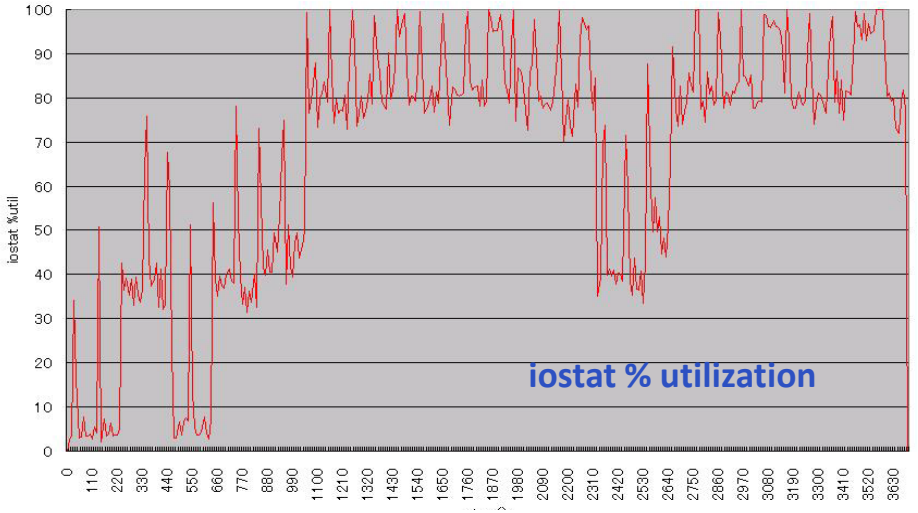


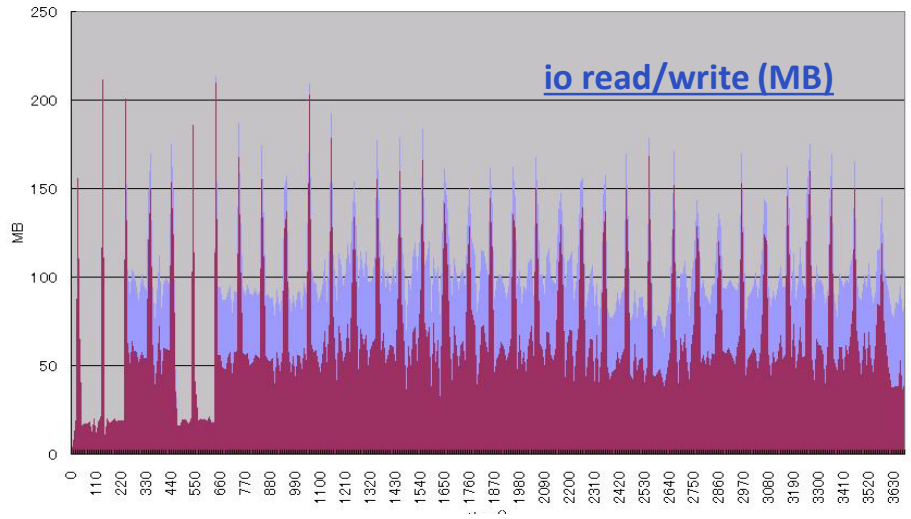

- Optimal storage type selection depends on object size.

© 2014 Cloudian, Inc. All rights reserved.

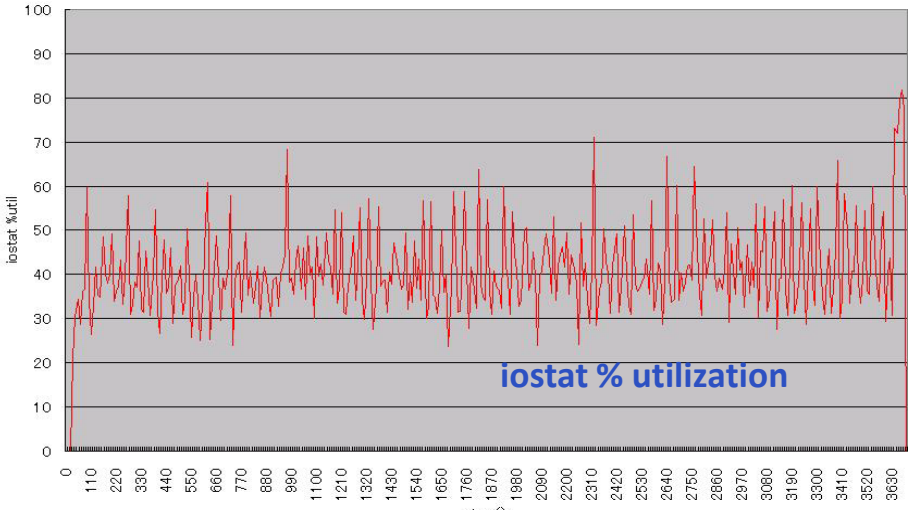
# Better performance with less variance

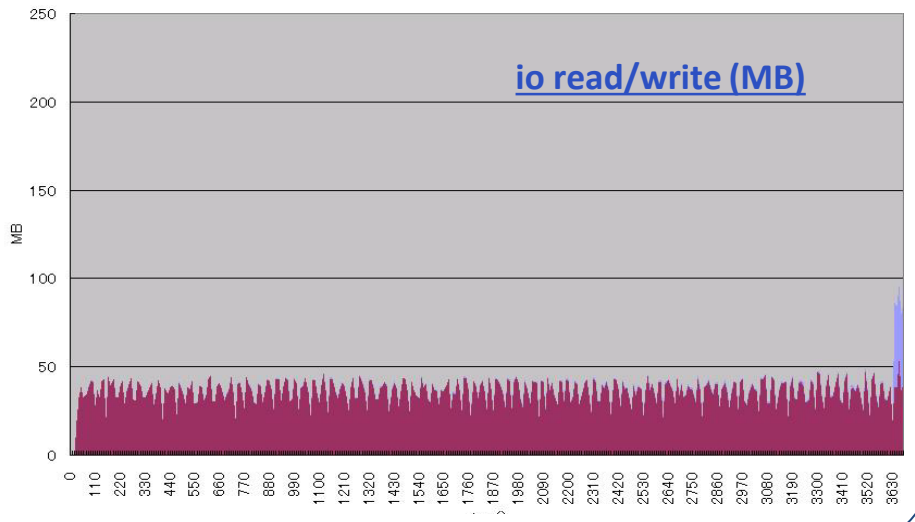
## Single Storage Type




## Multiple Storage Types



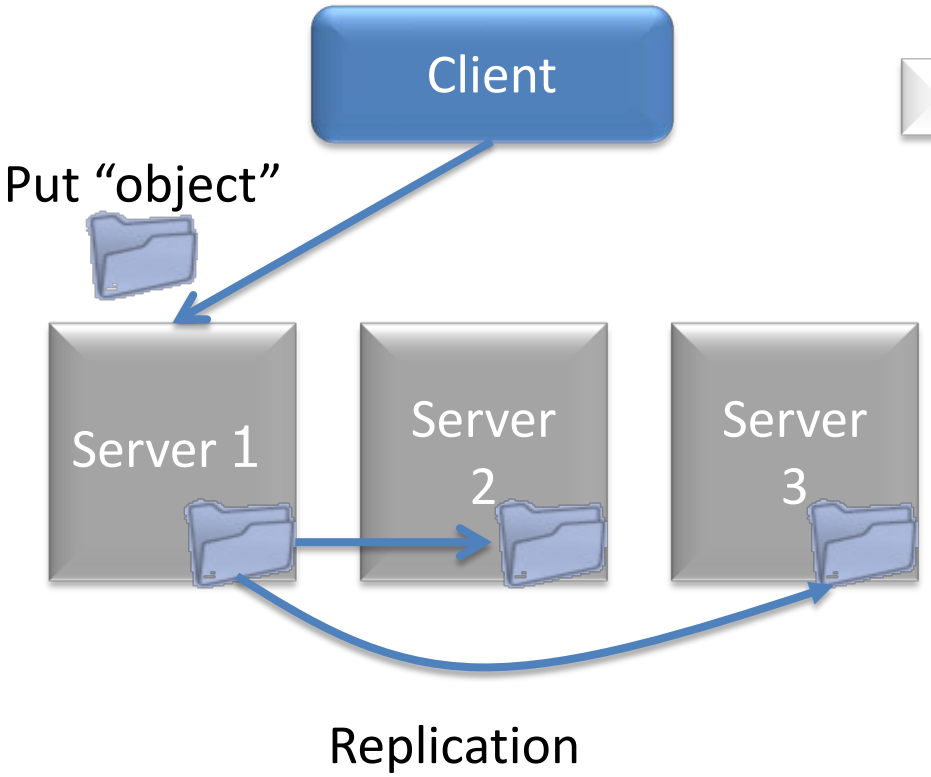

© 2014 Cloudian, Inc. All rights reserved.



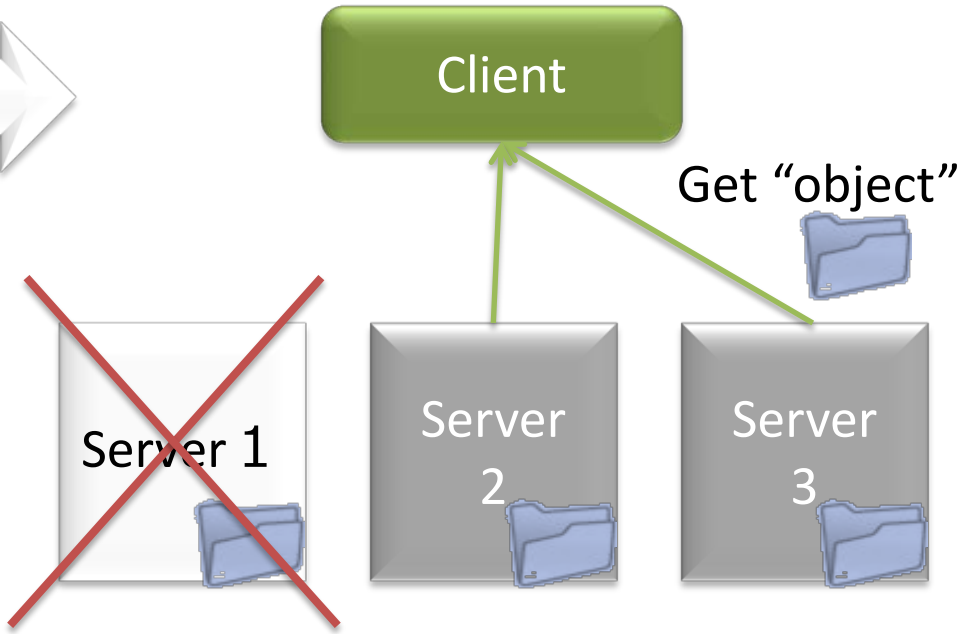
Cloudian

# Replication

One object is replicated to multiple servers




Even in the case of a server failure, client can get the object



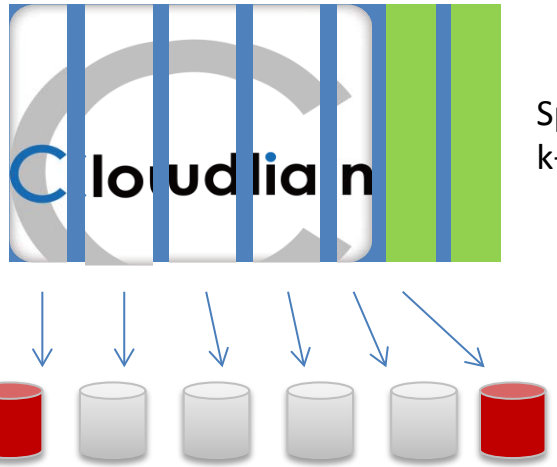

© 2014 Cloudian, Inc. All rights reserved.



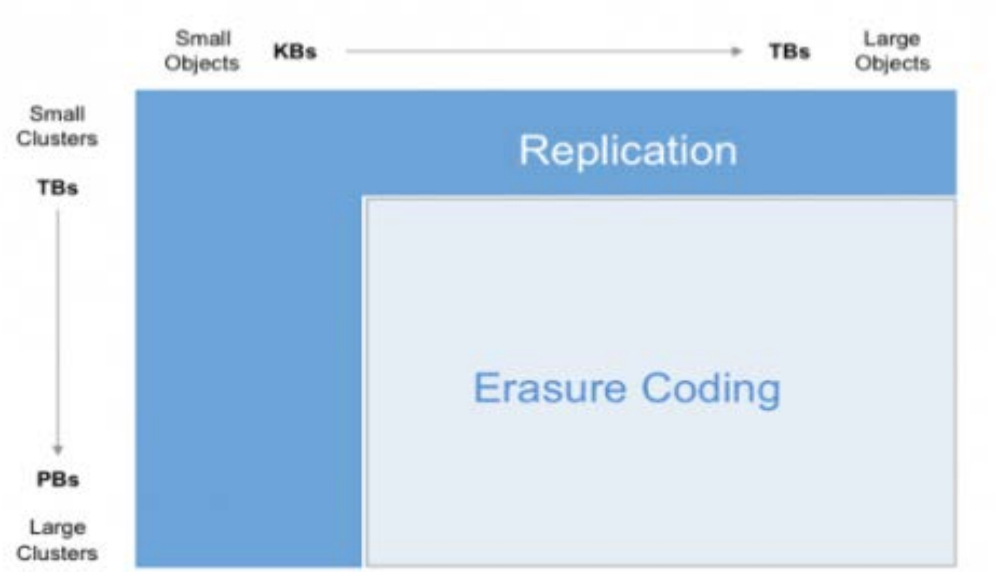
Cloudian

# Erasure coding


Divide one object into
- k data fragments
- m coding fragments



Spread object across k+m servers




http://soacloud.ulitzer.com/node/2702747

**Configurable**

- Storage type per-bucket and per-object: {EC, Replicas, Cassandra}
- Number of data & coding fragments: k and m.
- Consistency level: {QUORUM(k), ALL}
- Fragment placement on unique nodes: {T, F}

© 2014 Cloudian, Inc. All rights reserved.

# Large object support

- **Chunking**
  - Break single objects into smaller chunks when storing
  - Distribute chunks across the cluster for better performance
- **Multi-part**
  - Parts uploaded independently and in any order
  - Single parts can be re-transmitted
  - After all parts are uploaded, then presented as a single object
  - All parts are stored in file system, regardless of threshold configuration

© 2014 Cloudian, Inc. All rights reserved.

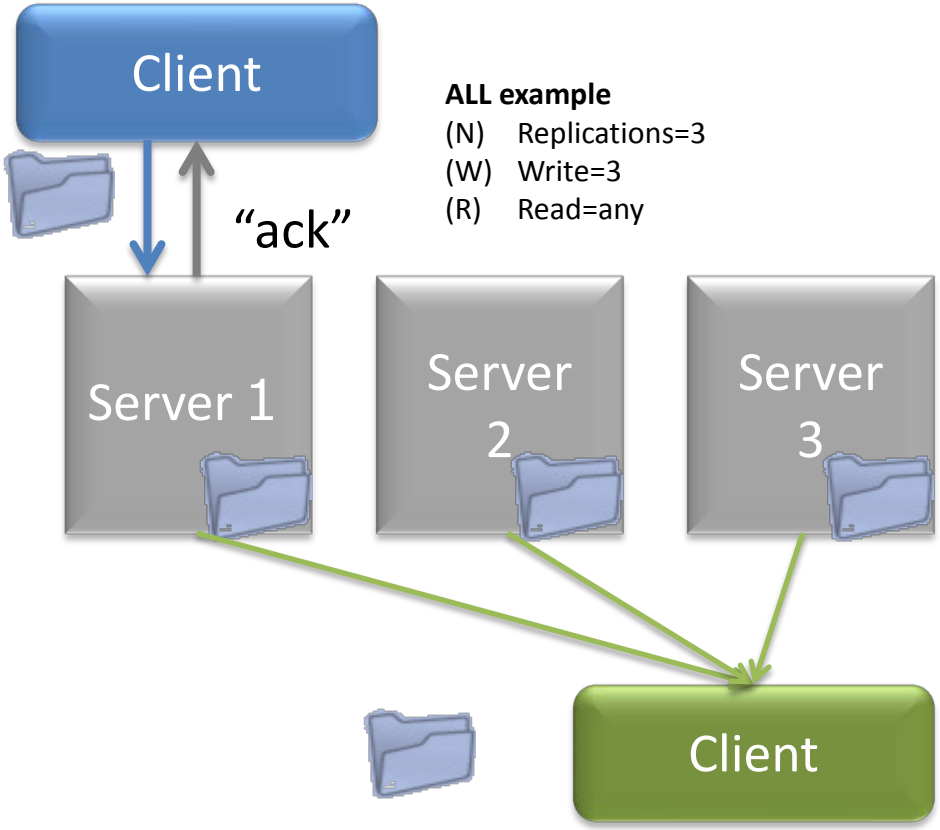
# HyperStore: Consistency Level

© 2014 Cloudian, Inc. All rights reserved.

# Consistency Level Example

CL=ALL

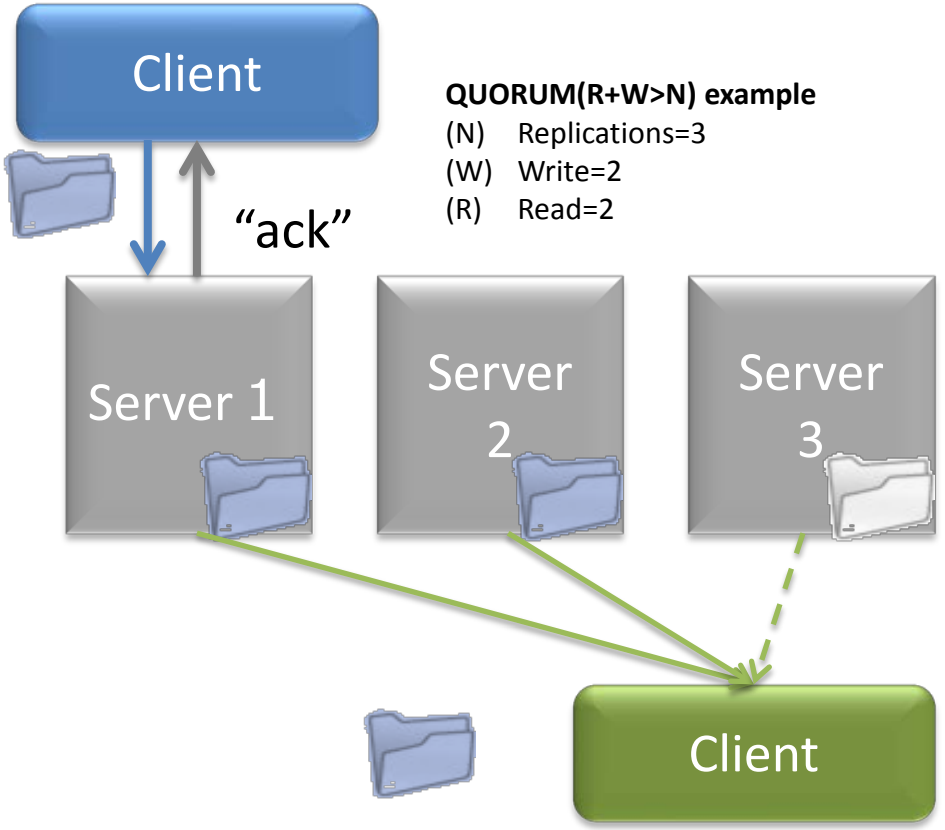
After “all” objects are replicated, Client receives acknowledgement




- **Data is always consistent**

CL=QUORUM

After “some” objects are replicated, Client receives acknowledgement




- **Data is consistent by QUORUM**

© 2014 Cloudian, Inc. All rights reserved.



Cloudian

# Consistency levels

WRITE

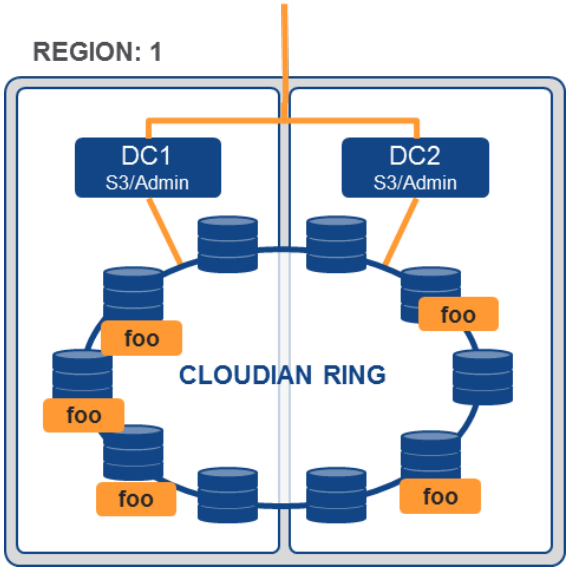
| Level | Description |
|---|---|
| ONE | A write has been written to at least 1 replica's commit log and memory table before responding to the client. |
| QUORUM | Ensure that the write has been written to N / 2 + 1 replicas before responding to the client. |
| LOCAL_QUORUM | Similar to QUORUM but replicas are in same data center as coordinator node. |
| EACH_QUORUM | Ensure that the write has been written to a quorum of replicas in *each datacenter in the cluster* before responding to the client |
| ALL | All replicas must have received the write, otherwise the operation will fail. |

READ

| Level | Description |
|---|---|
| ONE | Returns the response from the first replica causing a consistency check in a background thread. |
| QUORUM | Returns the record with the most recent timestamp once (N/2 + 1) replicas has responded. |
| LOCAL_QUORUM | Returns the record with the most recent timestamp once (N/2 + 1) replicas in the same data center as the coordinator node has reported. |
| ALL | Returns the record with the most recent timestamp once all replicas have responded. The read operation will fail if a replica does not respond. |

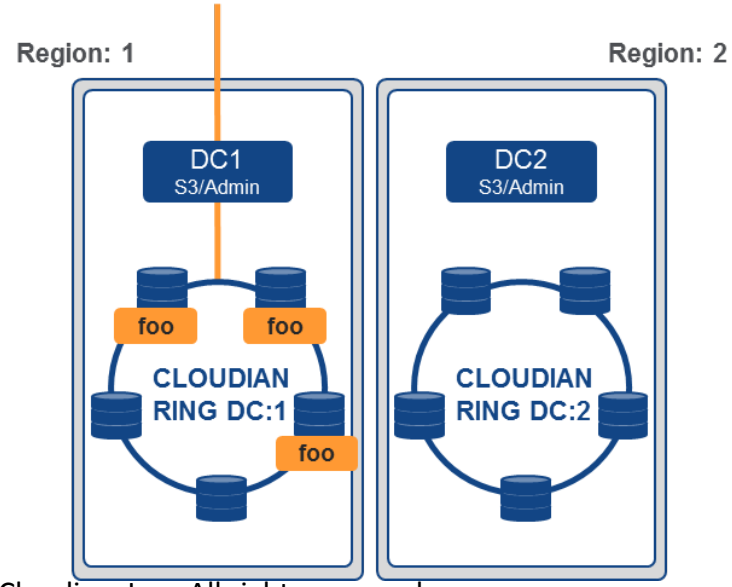
© 2014 Cloudian, Inc. All rights reserved.

# Configurable consistency across DCs




## Multi-DataCenter, Single Region

- 2 DCs, 1 Region, 5 replicas
- # of replicas per DC
- Consistency level: {QUORUM, LOCAL_QUORUM, EACH_QUORUM, ONE, ALL}




## Multi-DataCenter, Multi Region

- 2 DCs, 2 Regions, 3 Replicas
- Location constraint at bucket level.
- Objects not shared across regions.
- Span buckets across regions (virtual buckets)
- QoS & Billing based on region
- In each region, standard multi-DC configuration.

© 2014 Cloudian, Inc. All rights reserved.

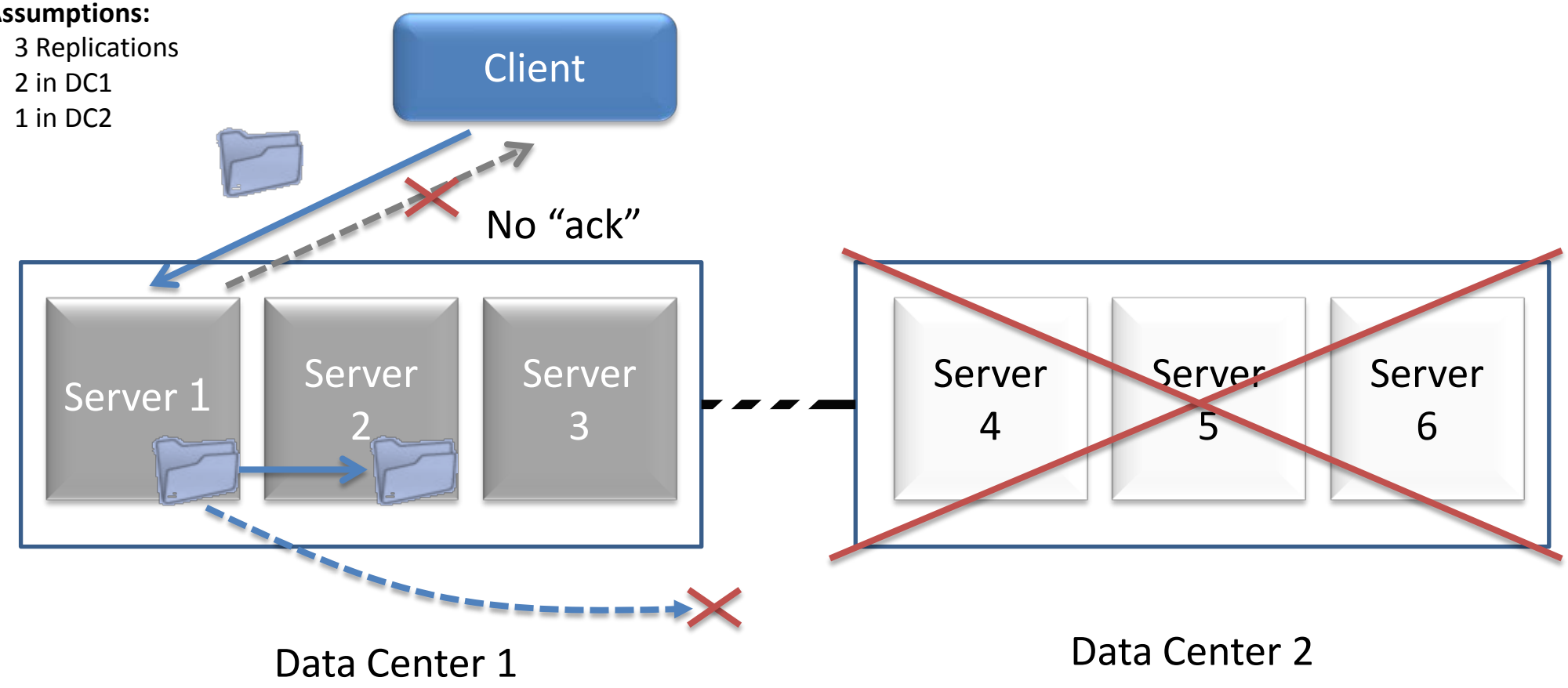


Cloudian

# Problem: CL=ALL for Multi-DC

- "All" or "QUORUM" cannot be completed in the case of failure
- "All" or "QUORUM" is slower for clients/applications

**Assumptions:**
- 3 Replications
- 2 in DC1
- 1 in DC2

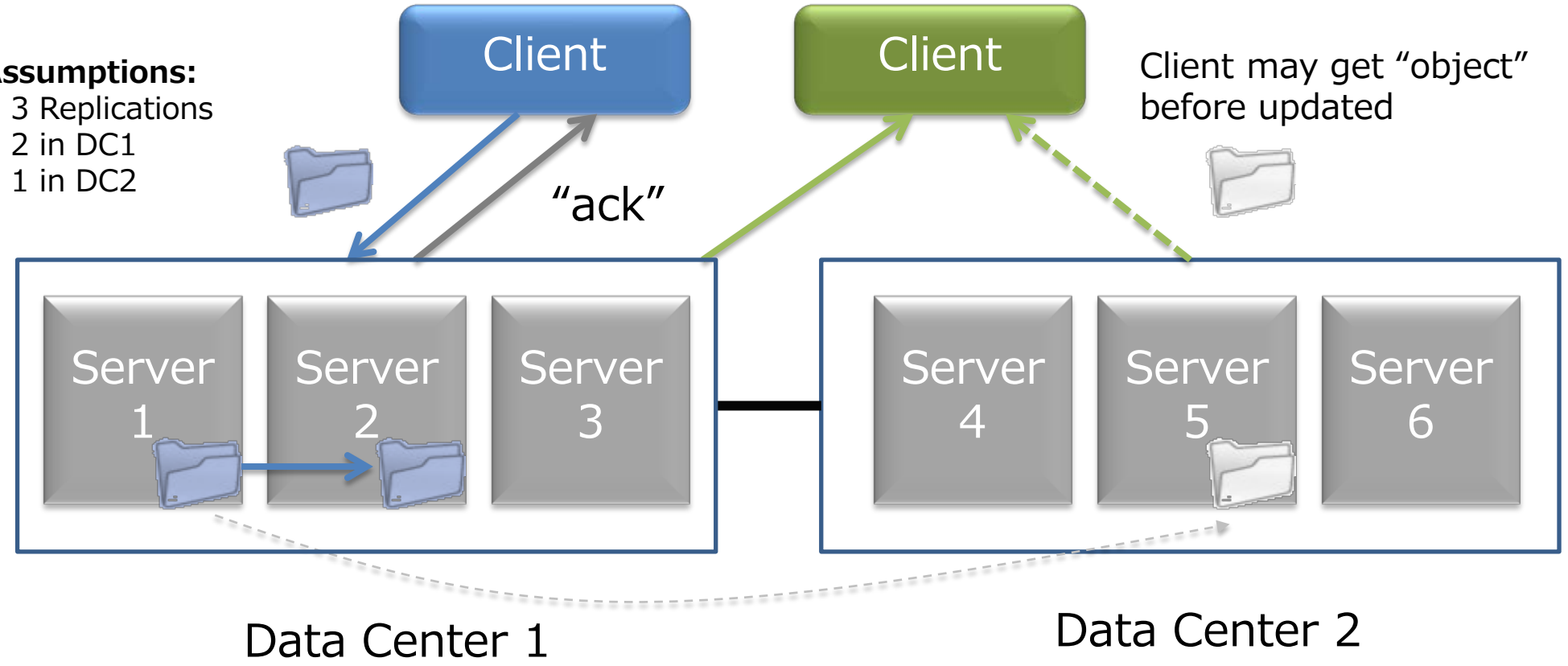



© 2014 Cloudian, Inc. All rights reserved.

# Problem: CL=LOCAL_QUORUM for Multi-DC

"Local _Quorum" cannot provide "data consistency" in the whole system

**Assumptions:**
- 3 Replications
- 2 in DC1
- 1 in DC2




**Data is consistent only when object is read from DC1**

© 2014 Cloudian, Inc. All rights reserved.

Cloudian

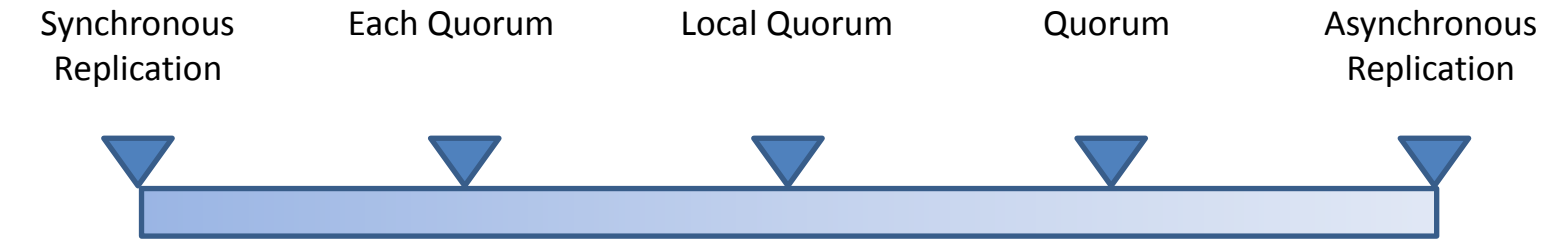
# Automatic dynamic consistency level

**Basic CLs**

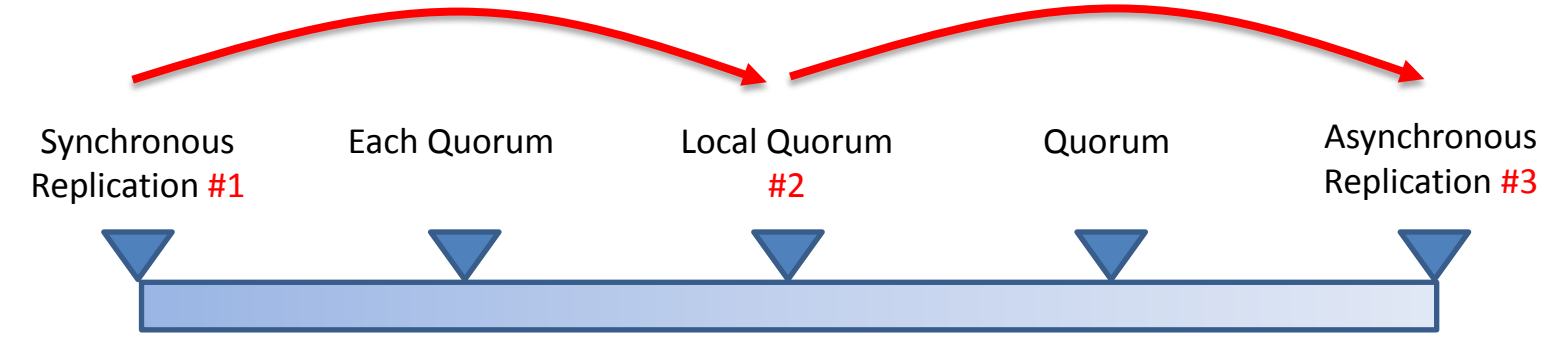

*+ No Data Loss*
*- Failures if DR site not reachable*

*+ Not dependent on availability of DR site*
*- Data Loss in case of Disaster*

**Advanced CLs**




**Advanced CLs with Automatic Change**




© 2014 Cloudian, Inc. All rights reserved.



Cloudian

# HyperStore:
# QoS

© 2014 Cloudian, Inc. All rights reserved.

# Quality of Service (QoS)

Operator:

- Prevent a user or group from impacting other users/groups.
- ~~Guarantee a specific level of performance to a user/group.~~

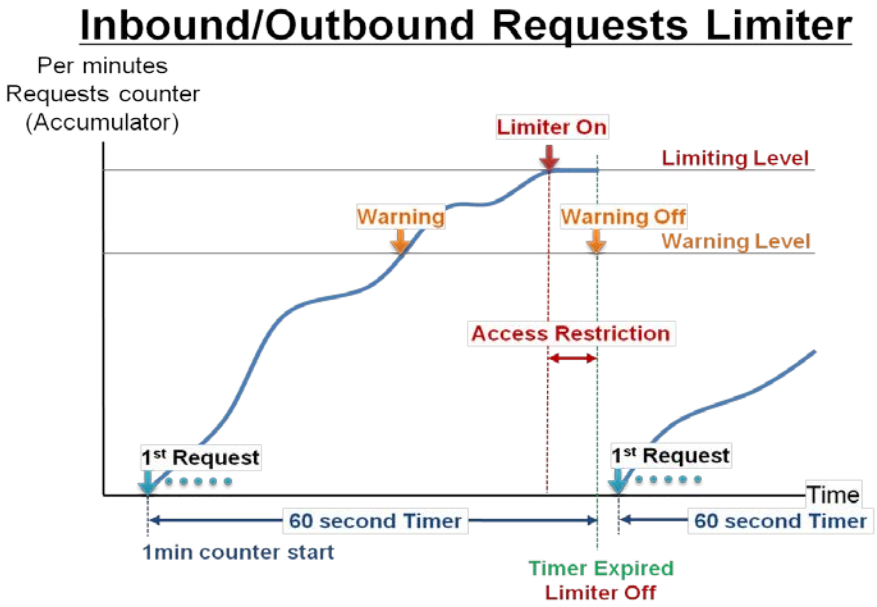
User/Group:

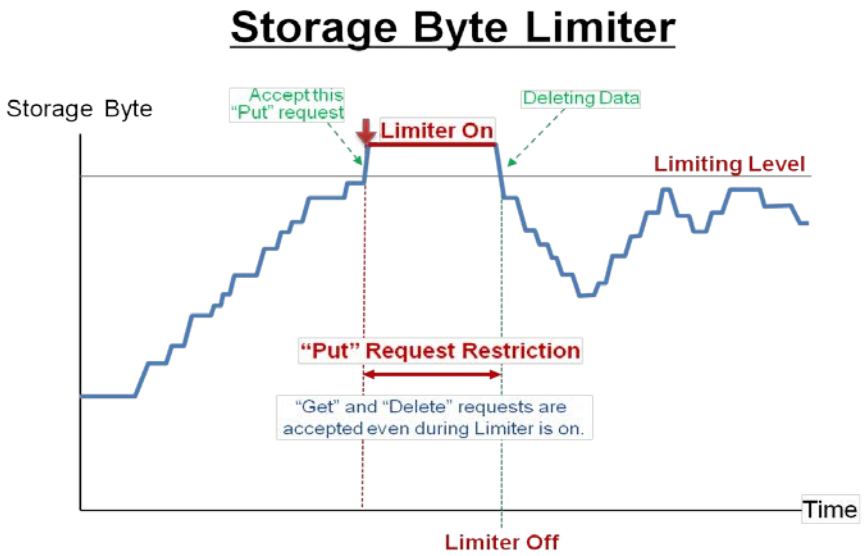
- Don’t exceed a preset limit.

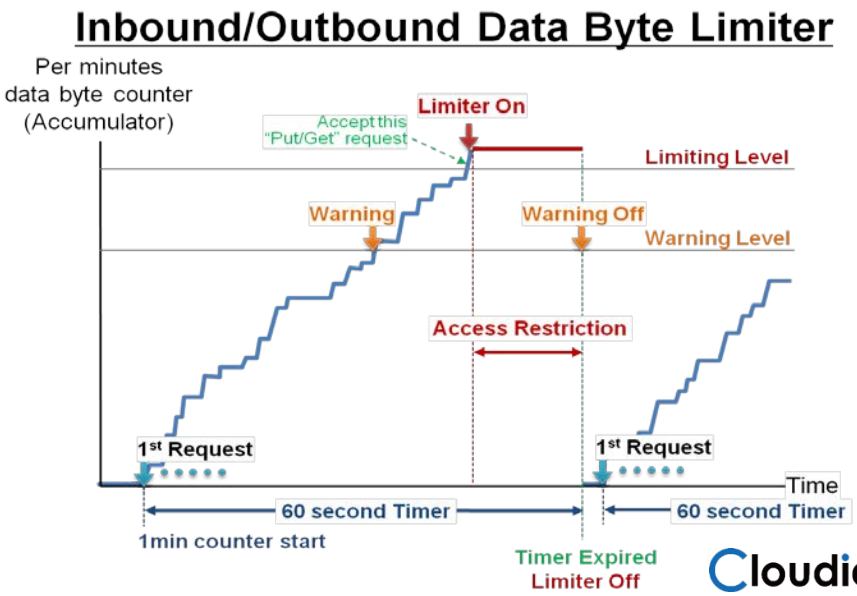
© 2014 Cloudian, Inc. All rights reserved.

# Quality of Service(QoS) Management

- Configurable maximum limits per-region at per-user, per-group, system level.
  - Requests/minute
  - Storage bytes
  - Storage objects
  - Data Bytes Inbound
  - Data Bytes Outbound
- While limit is reached, requests are rejected.




© 2014 Cloudian, Inc. All rights reserved.



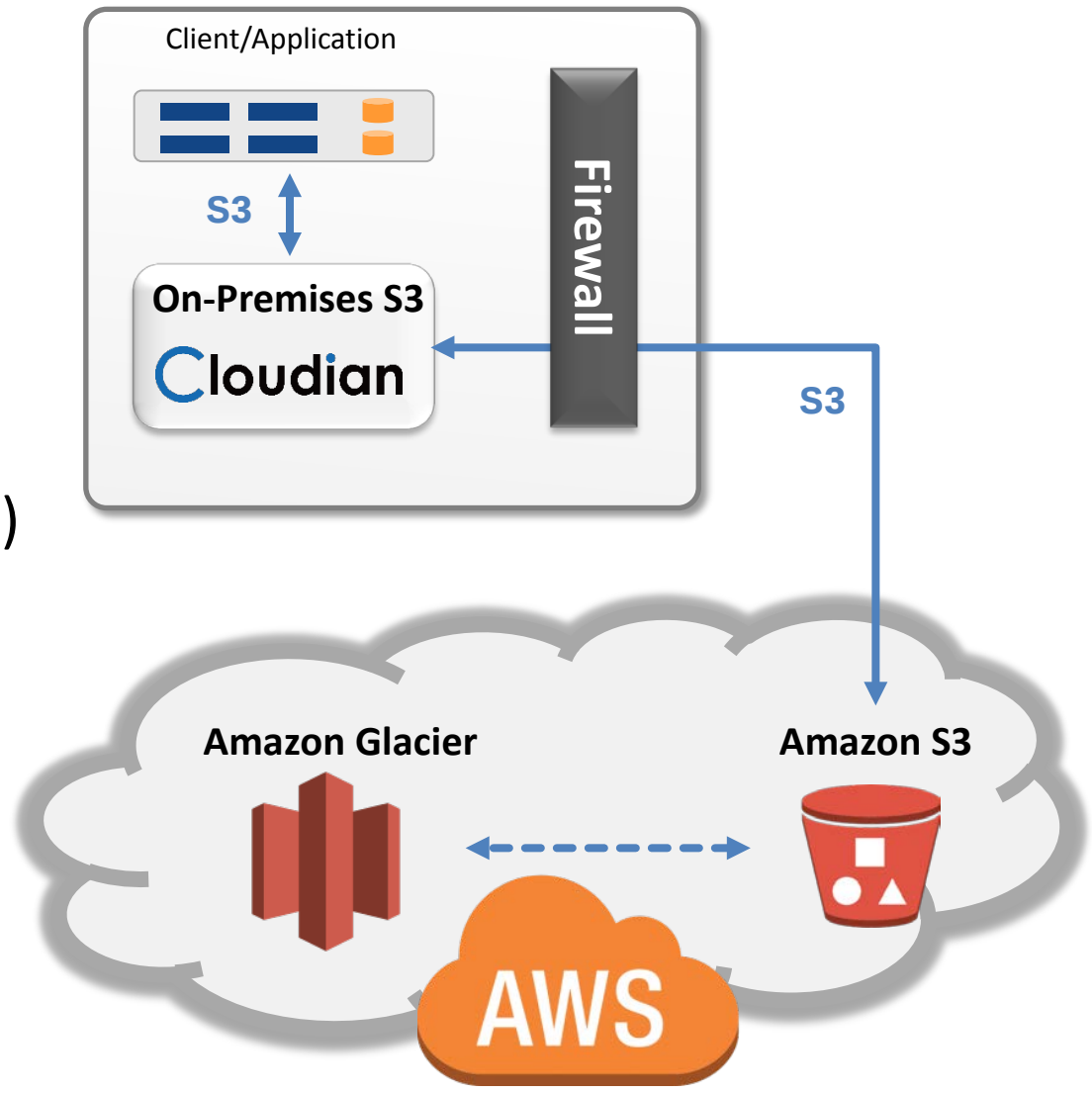
Cloudian

# HyperStore: Tiering

© 2014 Cloudian, Inc. All rights reserved.

# Tiering: Hybrid private/public

- Some data on-premise, some off-premise
- S3 bucket lifecycle policies (e.g., age) to migrate data to Amazon S3 and Glacier (or any S3 system)
- Read options:
  - Streaming
  - HTTP redirect
  - Restore
- Consolidated reports and bill




© 2014 Cloudian, Inc. All rights reserved.

# Cloudian products

## HyperStore Software

- The software only version
- Runs on commodity hardware
- Runs on commodity software: Linux, POSIX filesystem


## HyperStore Appliance

- Sold as an appliance by Cloudian or by a Cloudian Partner
- End user gets complete hardware/software solution
- No software installation needed
- 3 Models
  - HSA1024: 1U, 32GB RAM, 4xGigE NIC, 24 TB,
  - HSA1048: 1U, 32GB RAM, 4xGigE NIC, 48 TB
  - HSA2048: 2U, 64GB RAM, 2x1 GigE + 2x10GigE NIC, 48 TB, Flash Optimized

© 2014 Cloudian, Inc. All rights reserved.

# Closing

*Optimize overall system*
*by providing fine-grained controls,*
*both manually and automatically changed.*

**More Info, free trial, demo, PoC:**

- www.cloudian.com
- @CloudianStorage, @go10
- www.facebook.com/cloudian.cloudstorage

© 2014 Cloudian, Inc. All rights reserved.

# Backup

© 2014 Cloudian, Inc. All rights reserved.

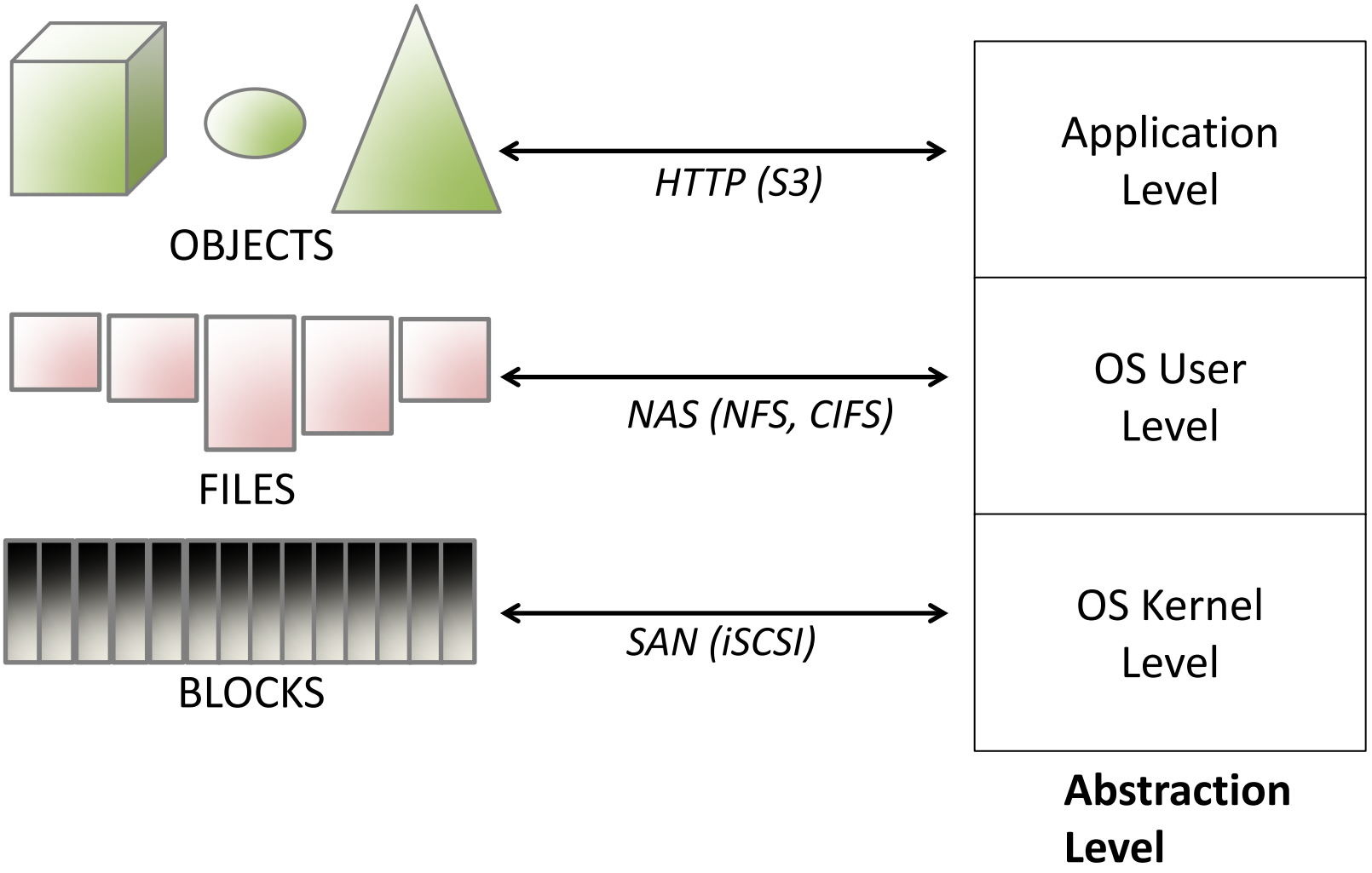
# About Cloudian

- Object storage startup in Silicon Valley
- Production hardened product
- Target market: mid- to large-enterprises & regional service providers

## CLOUDIAN PARTNERS


© 2014 Cloudian, Inc. All rights reserved.

# Objects as a higher layer of abstraction




© 2014 Cloudian, Inc. All rights reserved.

# Block & File vs. Object

## Emergence of a two-tier enterprise storage architecture

Block & File

### Faster

- For ‘hot’ data
- Flash-optimized
- IOPS-centric
- VM/VDI optimized
- Variety of approaches

Object

### Bigger

- For cool/cold data
- Object-based
- Scale-out (multi-PB)
- Software-centric
- Cloud-compatible

© 2014 Cloudian, Inc. All rights reserved.

# Object Storage Use Cases

- Enterprise backup
- Long term archiving
- Sync and share
- Remote office file storage
- Cloud storage (PaaS)












On-premise S3



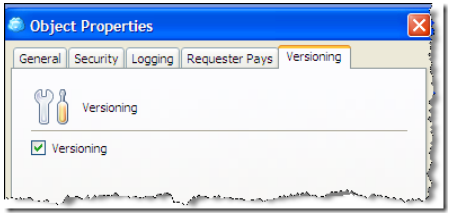
© 2014 Cloudian, Inc. All rights reserved.

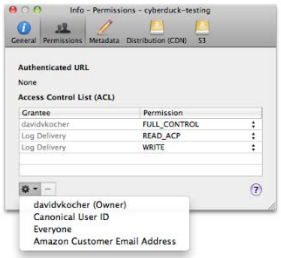


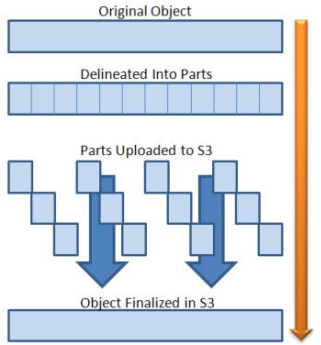
Cloudian

# S3 advanced features:

- Multi-part uploads:  allows uploading large objects in multiple parts
- Versioning:  multiple versions of same object
- Bucket Lifecycle: auto-expiration using rules
- Server side encryption: enhance confidentiality
- Location constraint:  Assign data to specific region (eg for HIPAA Compliance)
- Bucket Website: Create buckets as websites to host web content
- Access control lists (ACLs) define access rights to bucket and object
- And more...




© 2014 Cloudian, Inc.  All rights reserved.



Cloudian