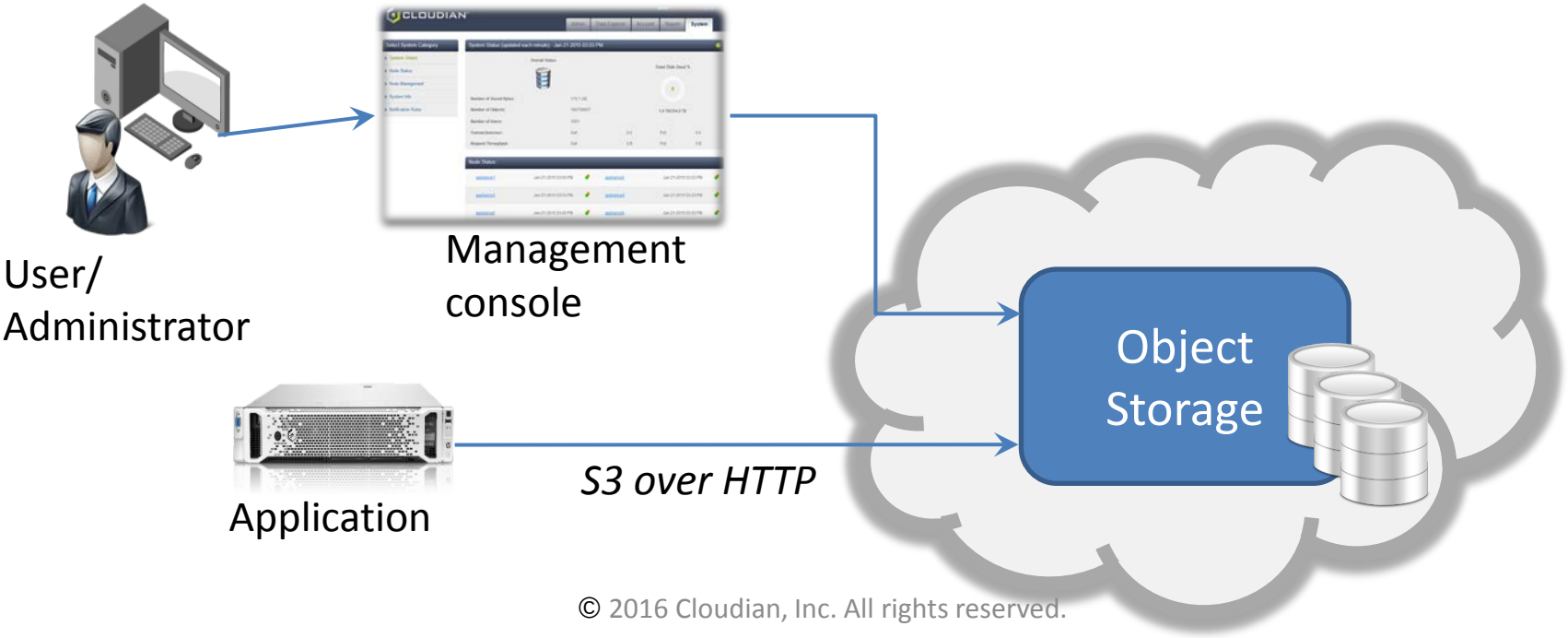
SDC
STORAGE DEVELOPER CONFERENCE
SNIA ■ SANTA CLARA, 2016

Dynamic Object Routing

Balaji Ganesan
Bharat Boddu
Cloudian

HyperStore System Overview

1. **Full Amazon S3 API Compatibility**, including error codes.
2. **Multi-datacenter**, peer-to-peer architecture. **No single point of failure**.
3. **Multi-tenant**: QoS controls, billing, reporting by user and group.
4. **Elastic Capacity**: Small start and scale-out as needed.
5. **Management/Monitoring Console or REST API**
6. **Easy to Deploy :** Packaged software or Appliance



Why Object Storage

To the application user, the logical “object” matters, Not how it’s physically stored (e.g., pieces, versions, location).

HyperStore Use Cases

Private Cloud

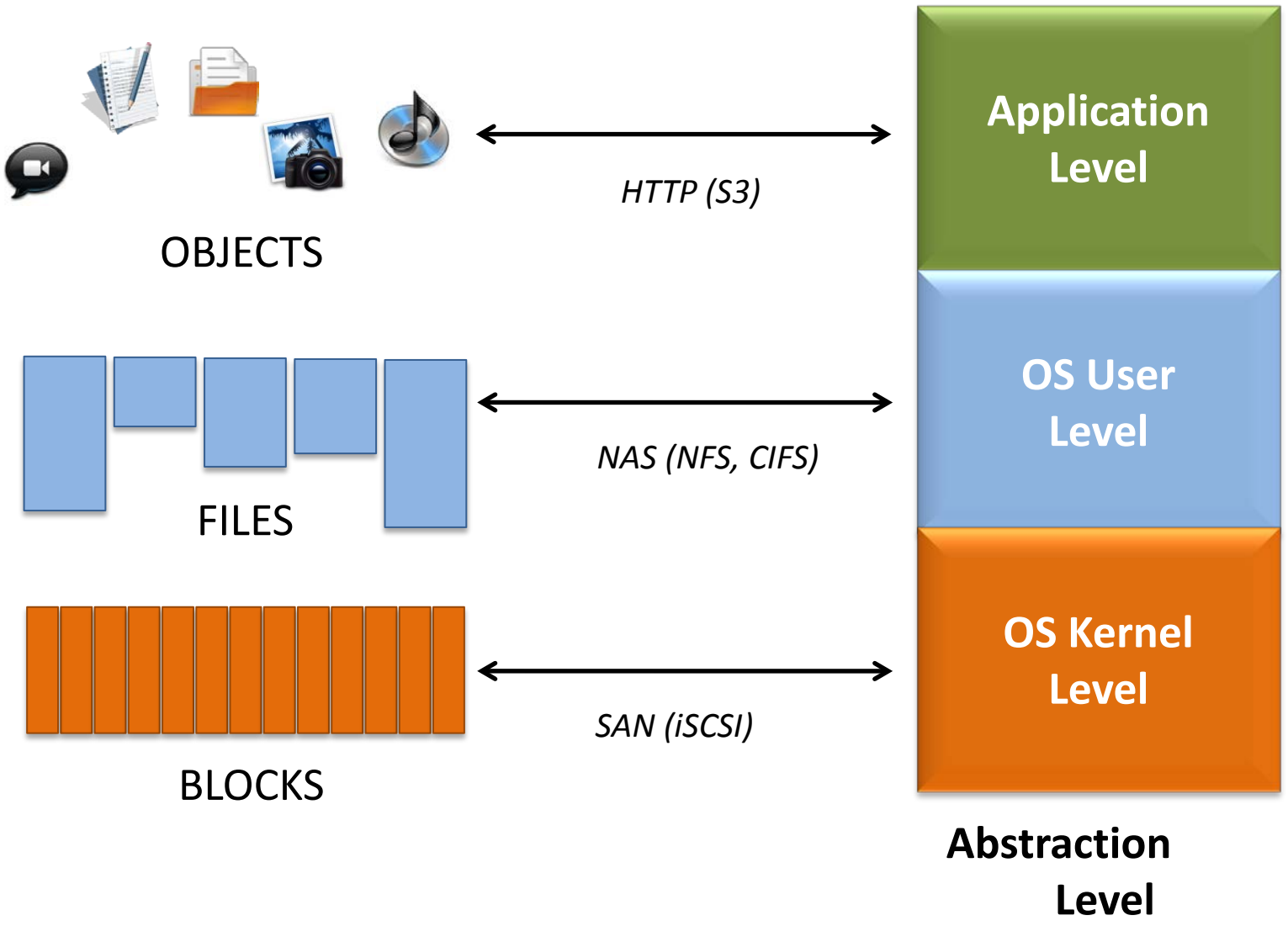
Backup Archive

File Distribution & Sharing

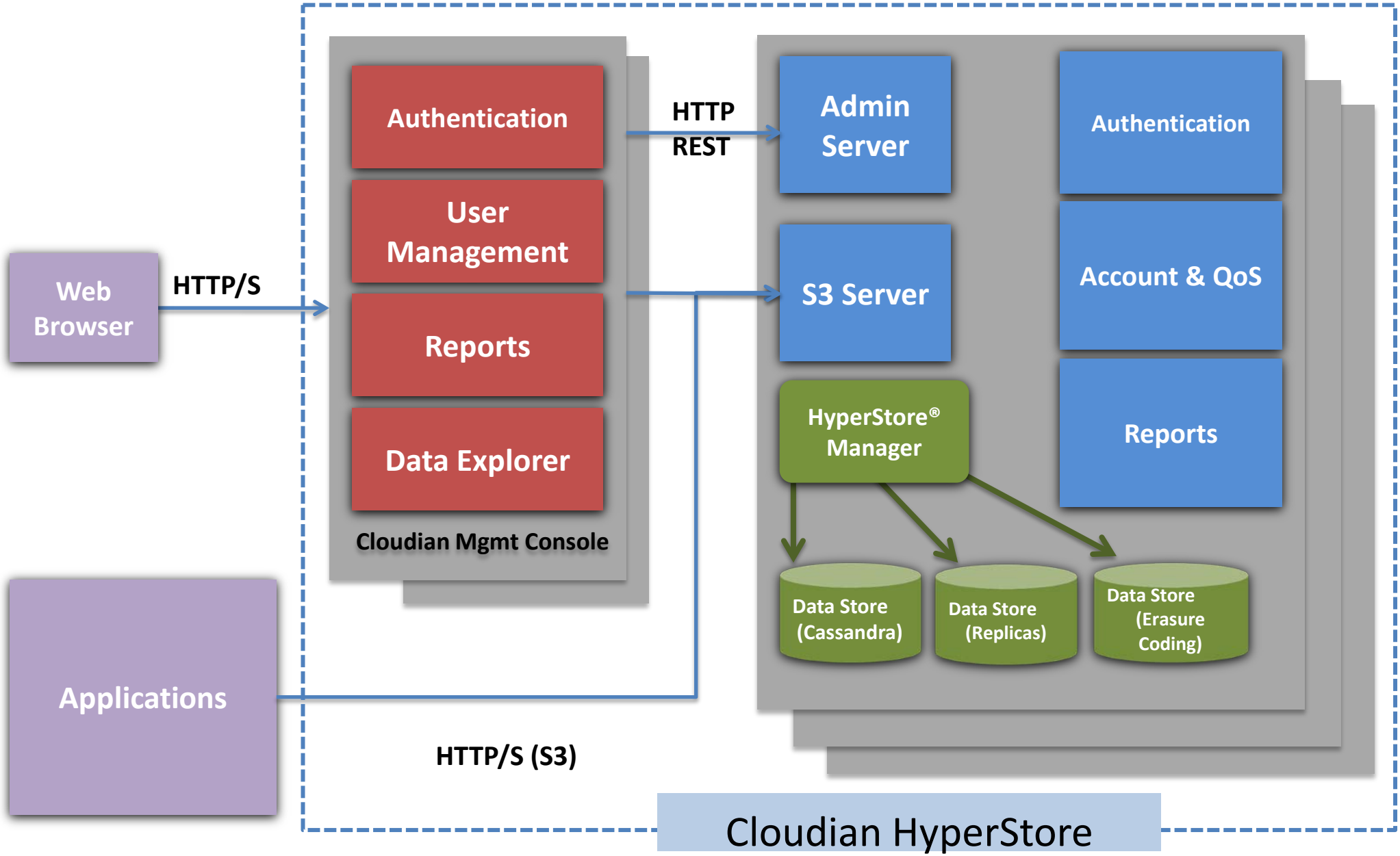
Media Content Store

Data Analytics

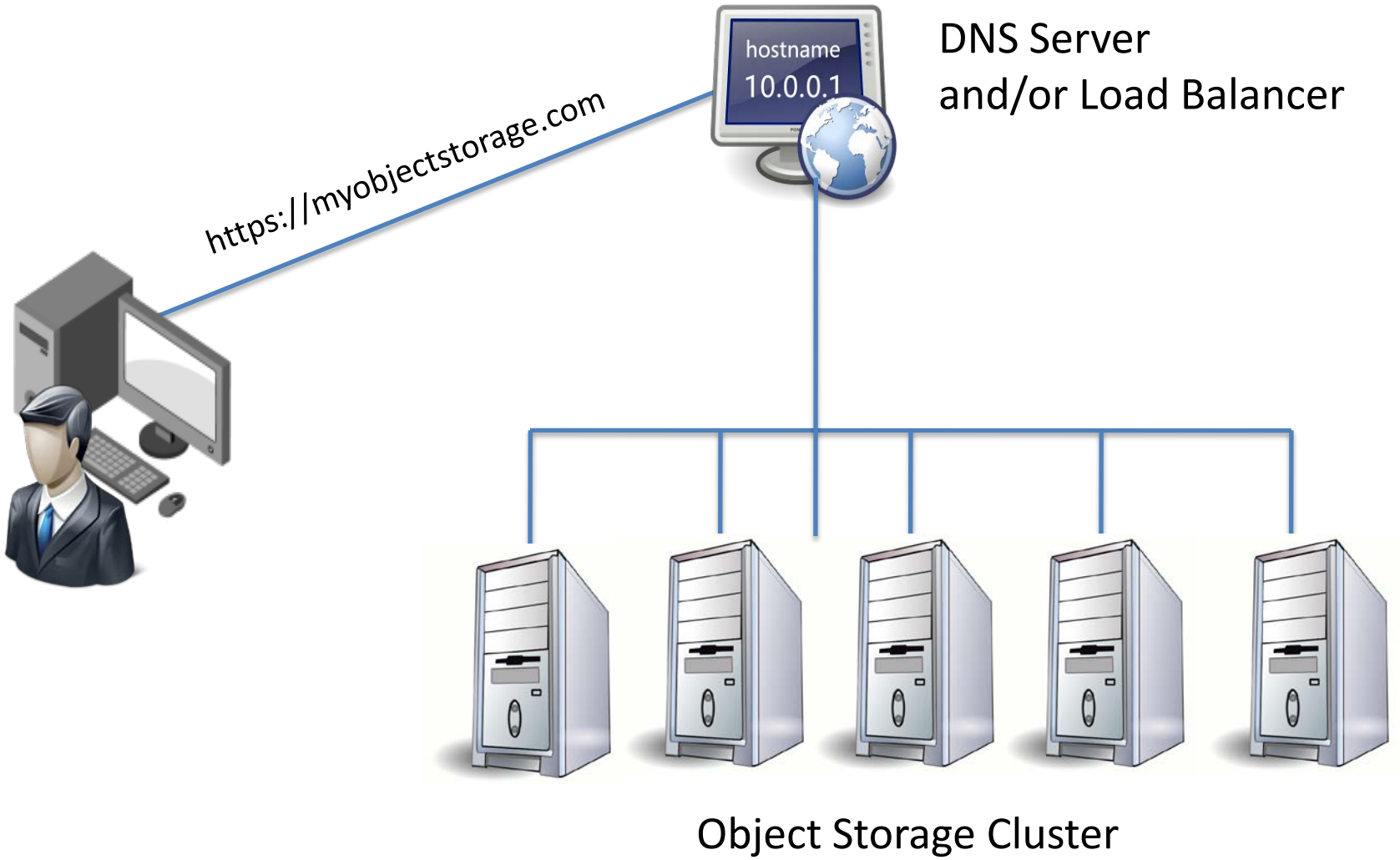
Object vs. File vs. Block Storage



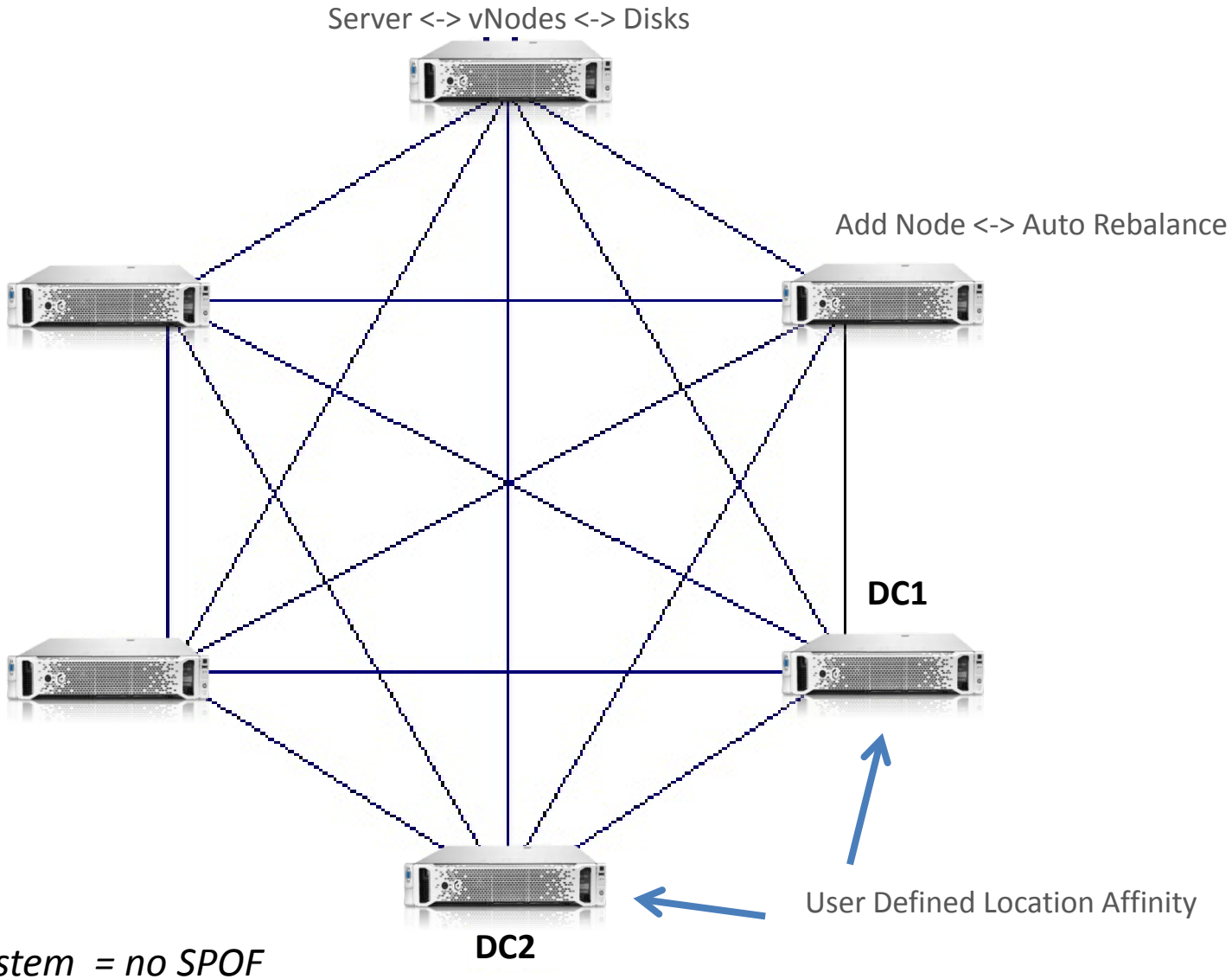
Logical Architecture



High-level System View

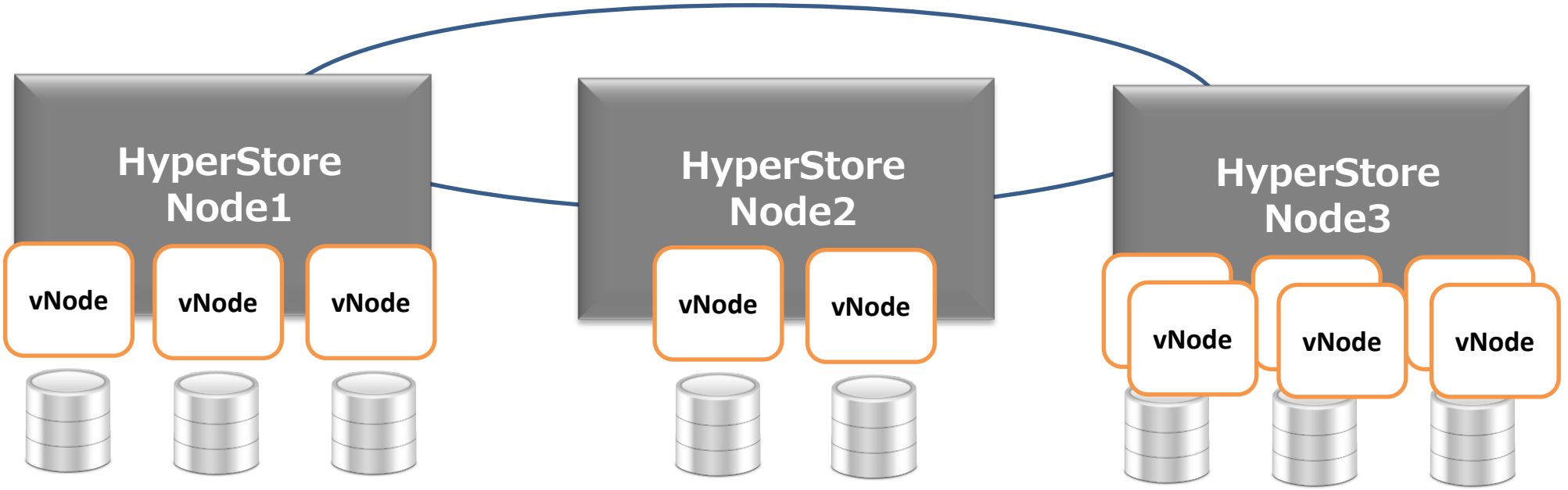


Distributed & Elastic Geo Cluster



- ✓ *Peer-to-peer system = no SPOF*
- ✓ *Distributed Everything = Data , Metadata, Configuration*

vnodes



- Vnodes are mapped to physical disks. Then one disk failure only affects those vnodes.
- Max 256 vnodes per physical node. No token management. Tokens randomly assigned.
- Increased repair speed in case of disk or node failure
- Allows heterogeneous machines in a cluster

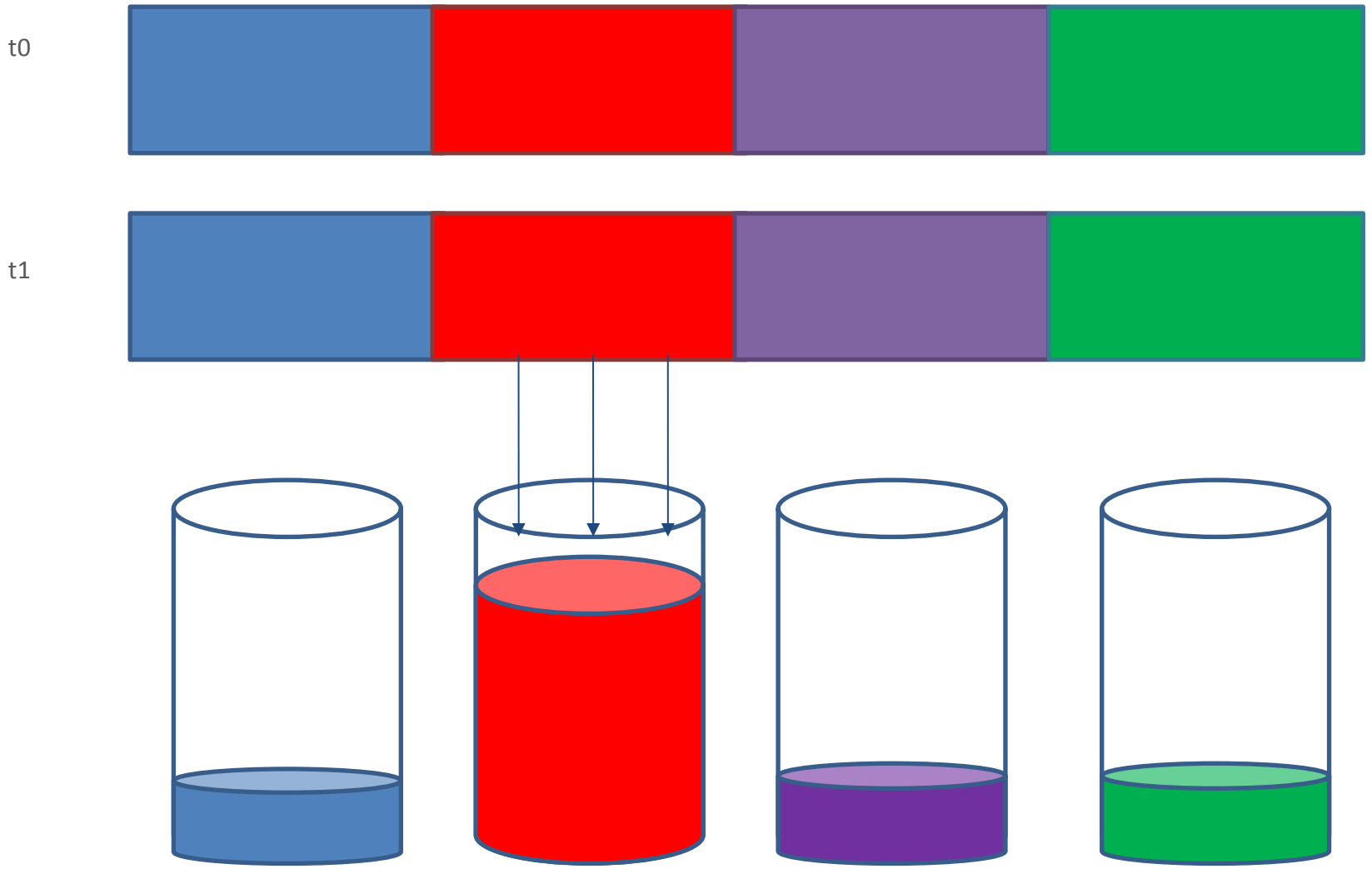
HyperStore Data distribution

- Each object is assigned with a unique identified called Object ID.
- Object ID consists of two parts
- MD5 hash of object name
- Object last modified time
- Objects are immutable.
- When an existing object is overwritten, a new Object ID is created with same MD5 hash of object name but with a different timestamp

Static Mapping Table

Disk	vNode
Disk1	A1,a2,a3,a4…..aN
Disk2	B1,b2,b3,b4...bN
Disk3	C1,c2,c3….cN
Disk4	D1,d2,d3...dN

Token Range Static Mapping



Problems With Static Mapping

- Uneven Disk Usage
- Complex Failure Handling

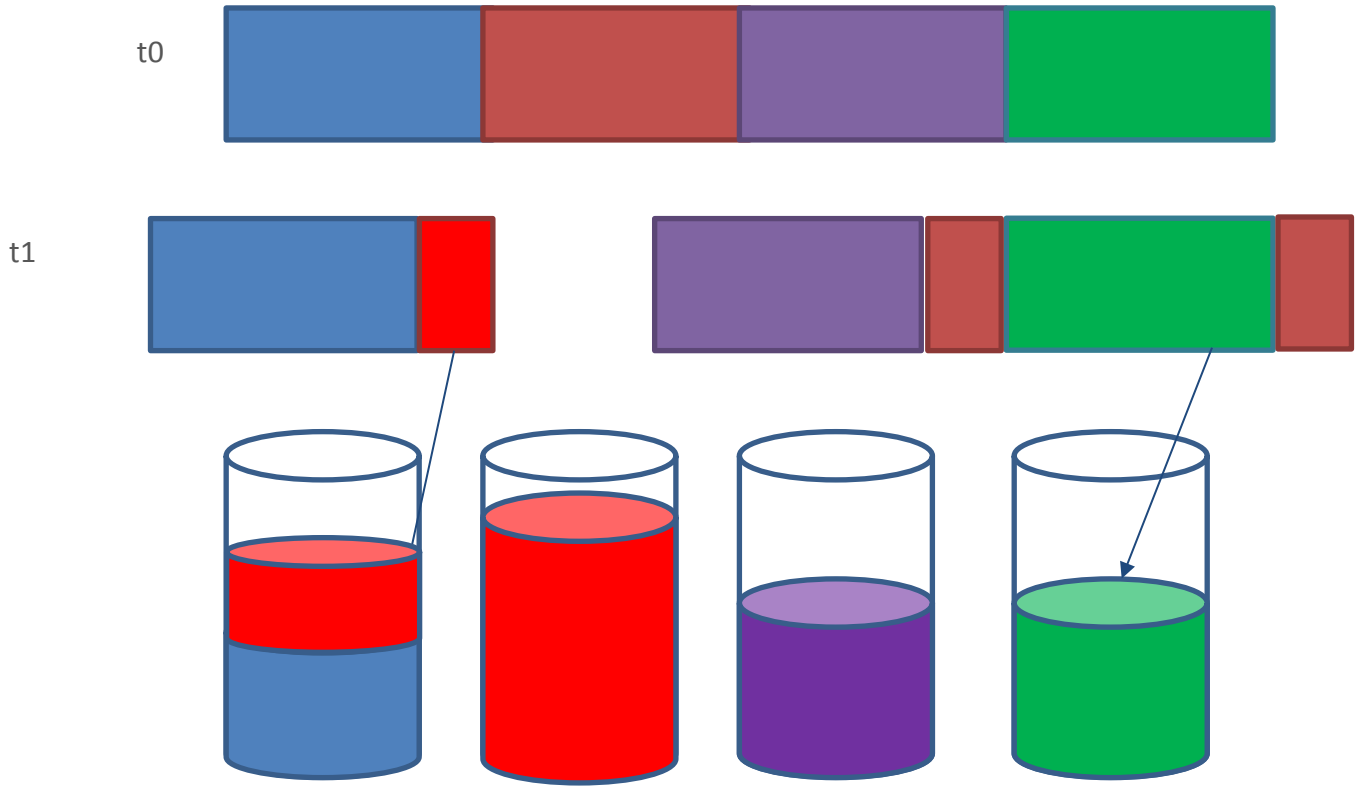
Initial Solution

- Use a tool to move data from heavily used disk to less used disk
- Tool needs to be run manually
- Lot of data movement.
- Complex to recover from errors if data movement fails.

Dynamic Object Routing

- A routing table is used to determine object's storage location.
- Object hash value as well as its insertion timestamp is used to determine the object's storage location.
- Each hash bucket is assigned initially to one of the available disks and a routing table entry is created for that hash bucket with timestamp 0.
- When a hash bucket's storage disk utilization is greater than overall average disk utilization, another less used disk is assigned to that hash bucket with a new timestamp.
- All new objects to that hash bucket will be stored in new disk. Existing objects will be accessed from old disk using the routing table.
- This method will avoid moving data.

Smart Disk Balancing



Routing Table

vNode	Routing
A1	[{time:T, disk:disk1}, {time:T1, disk:disk2}]
A2	[{time:T, disk:disk1}]
B1	[{time:T, disk:disk2}, {time:T2, disk:disk3}]
C1	[{time:T, disk:disk4}]

DOR Implementation

- Periodically run checks on disks usage
- If we notice an imbalance it will change the tokens pointing from “highly used disk to low used disk”

Other Advantages of DOR

- Disk failure handling without affecting service
- Disk maintenance handling
- Enables adding multiple nodes to Cluster simultaneously

THANK YOU

More Info, free trial, demo, PoC:

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