

Seagate Kinetic Open Storage Platform

James Hughes

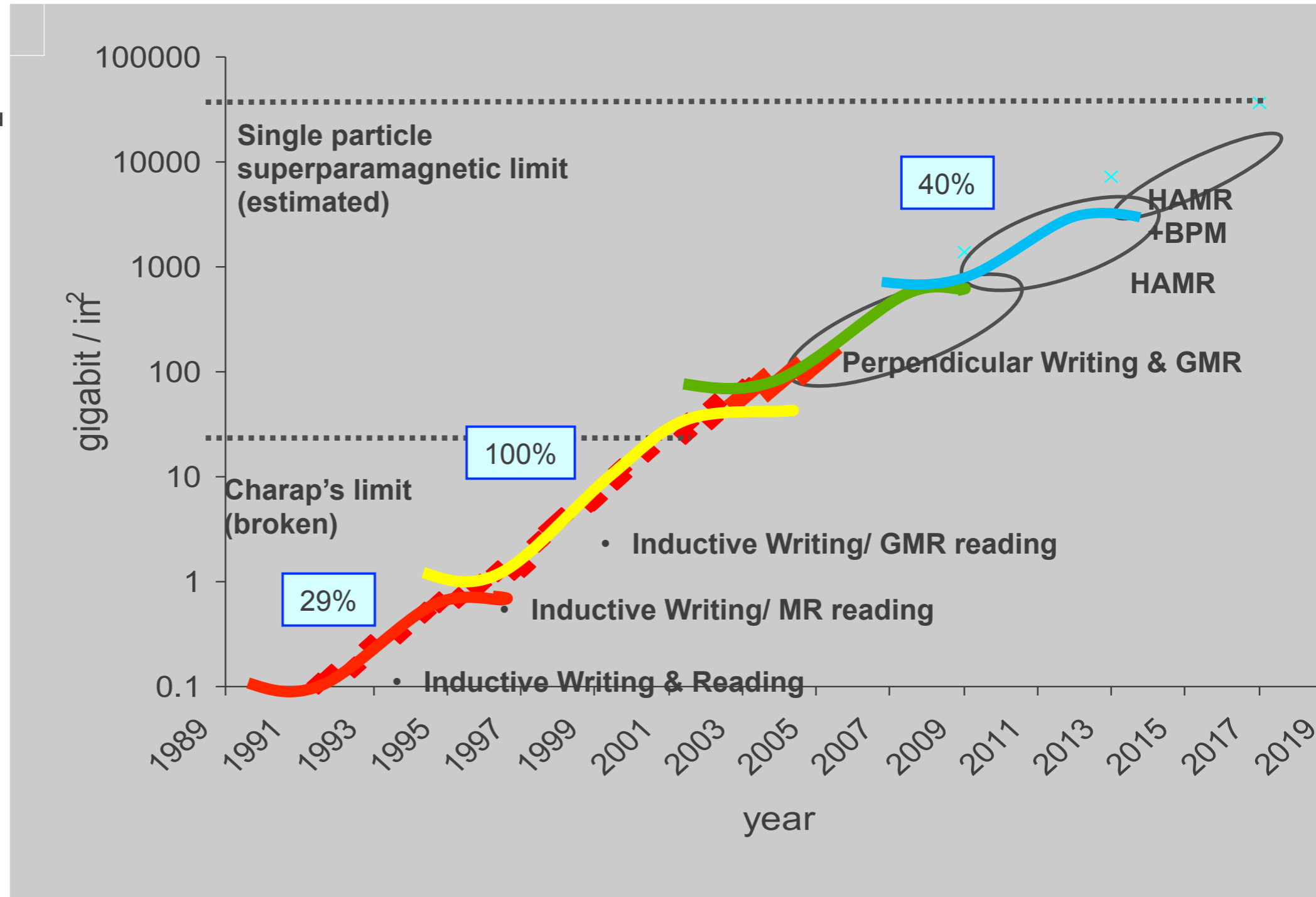
...and many others



Areal Density Growth

- Areal Density CAGR 40%
- Transfer Rate CAGR 20%

- Late 1990s – super paramagnetic limit demonstrated through modeling
- Perpendicular expected to extend to 0.5-1 Tb/in²
- Additional innovations required at that point
 - heat-assisted recording
 - bit patterned media recording



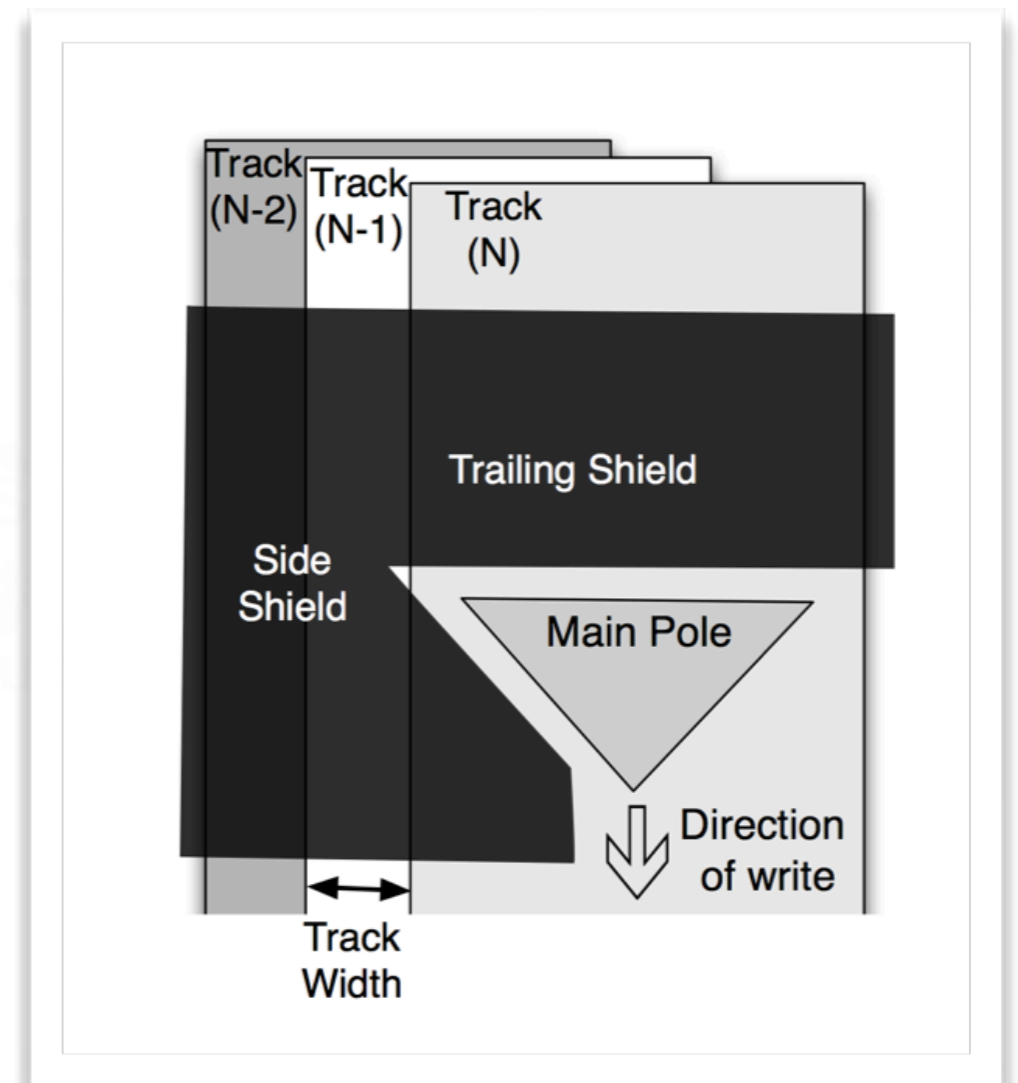
Shingled Disks

Write head larger than read head

- Turns Disk into a sequentially written media

All updates to data and metadata are written sequentially to a continuous stream, called a log

Disk API of sectors is no longer “natural”



Log Structured Storage

How much is erased on a reposition?

- Tape - the remainder of the tape
- Shingled disk - the remainder of the track group
- Flash - the entire page

All persistent Storage systems do/will implement log structure

- e.g. “NoSQL Database of sectors”

Does it make sense to layer a database on top of a database?

- Could we use the log structure of the media to provide a more natural storage systems, not mimicking an antique paradigm?



DATA CENTER > STORAGE

Seagate: Fibre Channel? RAID? SATA? File System? All RUBBISH

App to disk via ethernet, baby. The rest of you, clear out your desks

By Chris Mellor, 22nd October 2013 [Follow](#) 5,083 followers

45

RELATED STORIES

What's the first Kinetic Ethernet hard drive? Psst, it's the 4TB Terascale

Seagate to storage head:

[Free Regcast : Managing Multi-Vendor Devices with System Centre 2012](#)

Seagate is building hard disk drives with a direct Ethernet interface and object-style API access for scalable object stores, a plan which - if it works - would destroy much of the existing, typical storage stack.

Drives would become native key/value stores that manage their own space mapping with accessing applications simply dealing at the object level with gets and puts instead of using file abstractions.

Seagate says it has developed its Kinetic technology because the existing app-to-drive storage stack is clumsy, inefficient and delays data access. Put an Ethernet interface

Enterprise Backup and Recovery

Gartner Best Practices for Repairing the Broken State of Backup Report.



MOST READ

MOST COMMENTED

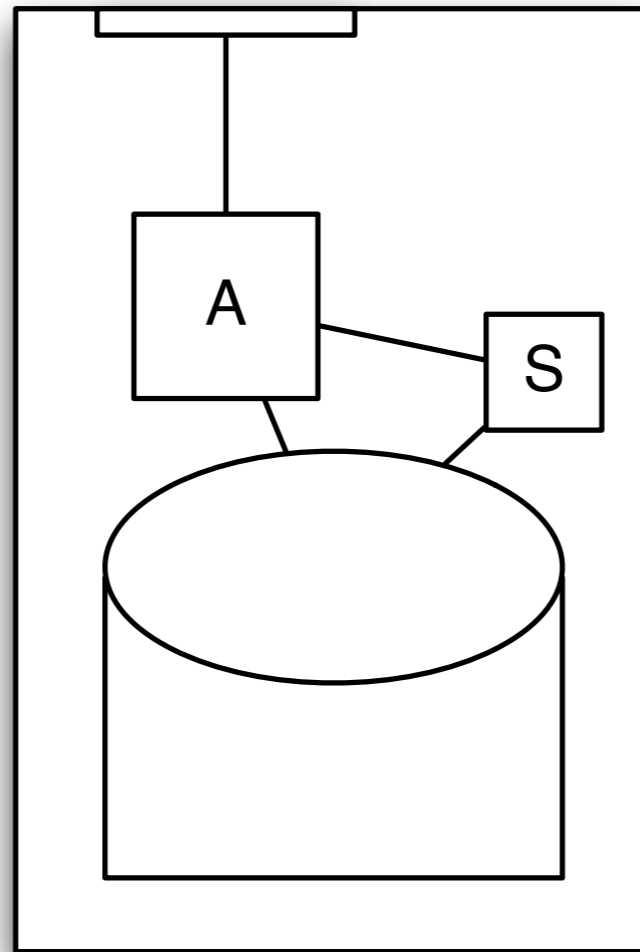
LG G Flex: A new cheeky curvy mobe with 'SELF-HEALING' bottom

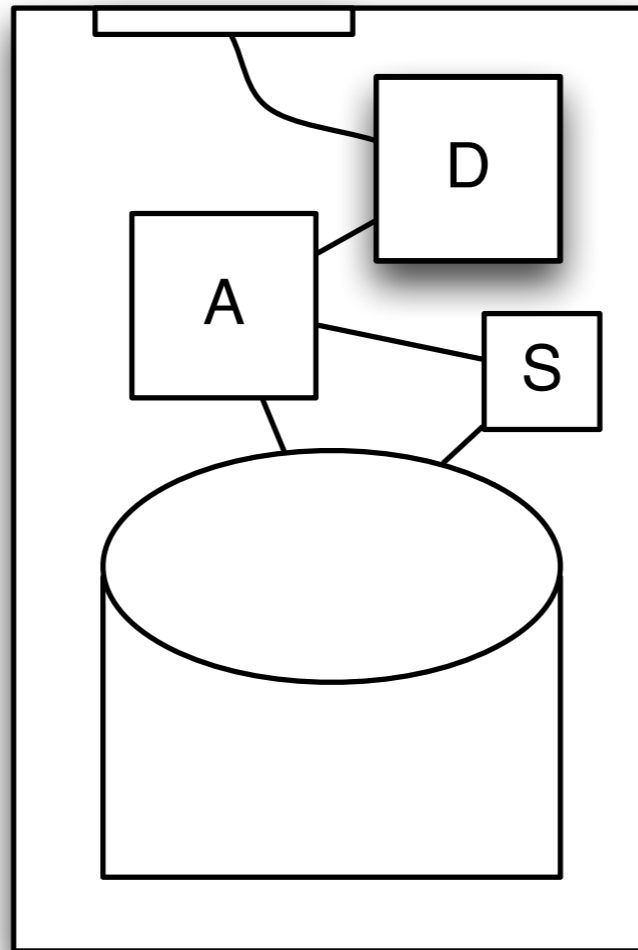
Everything's going to be all white: Google Nexus 5 mobe expected Friday

In a meeting with a woman? For pity's sake

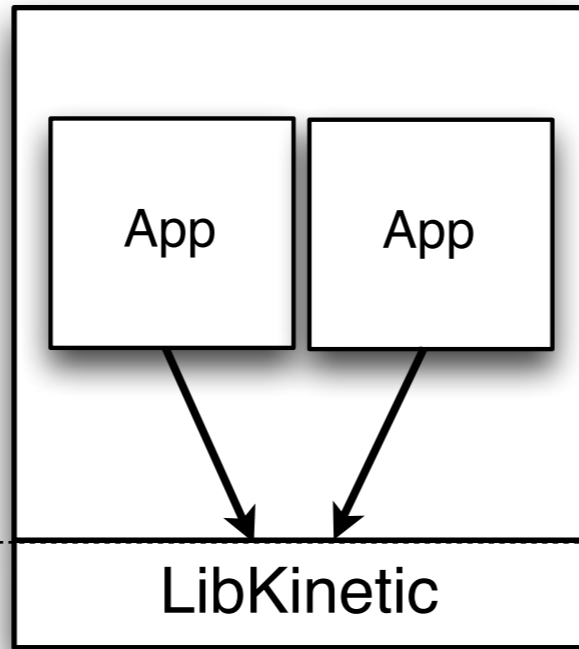
Seagate is building hard disk drives with a direct Ethernet interface and object-style API access for scalable object stores, a plan which - if it works - would destroy much of the existing, typical storage stack.

Drives would become native key/value stores that manage their own space mapping with accessing applications simply dealing at the object level with gets and puts instead of using file abstractions.



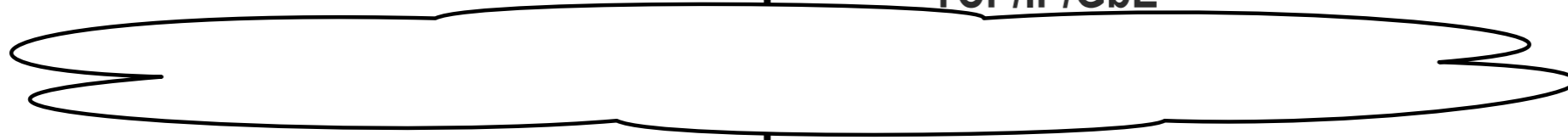


- Application
- Clustering
- Management

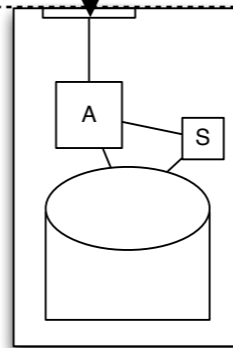


- Interconnect

- ProtoBuf
- TCP/IP/GbE



- Storage

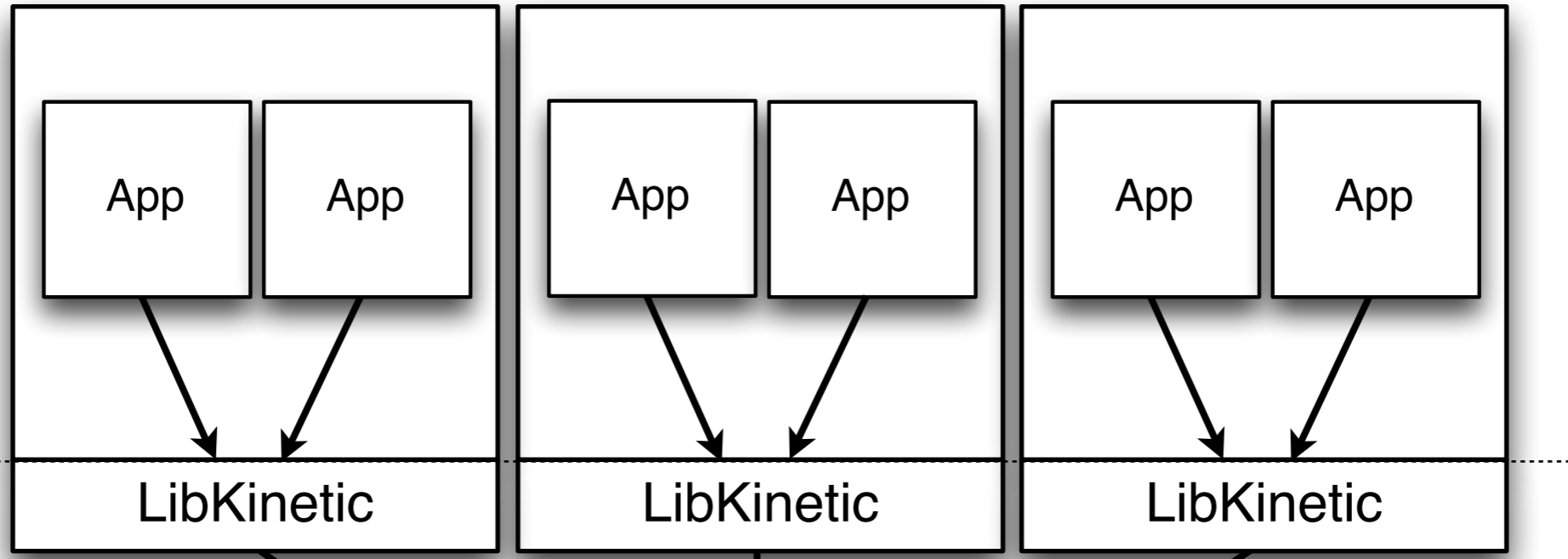


- Proprietary
- to System Vendor

- GPL
- Standard

- Proprietary
- to Seagate

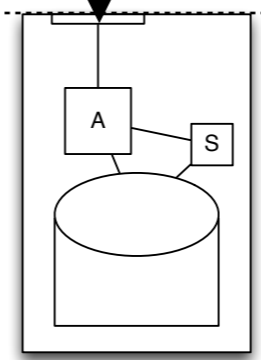
- Application
- Clustering
- Management



- Interconnect

- ProtoBuf
- TCP/IP/GbE

- Storage

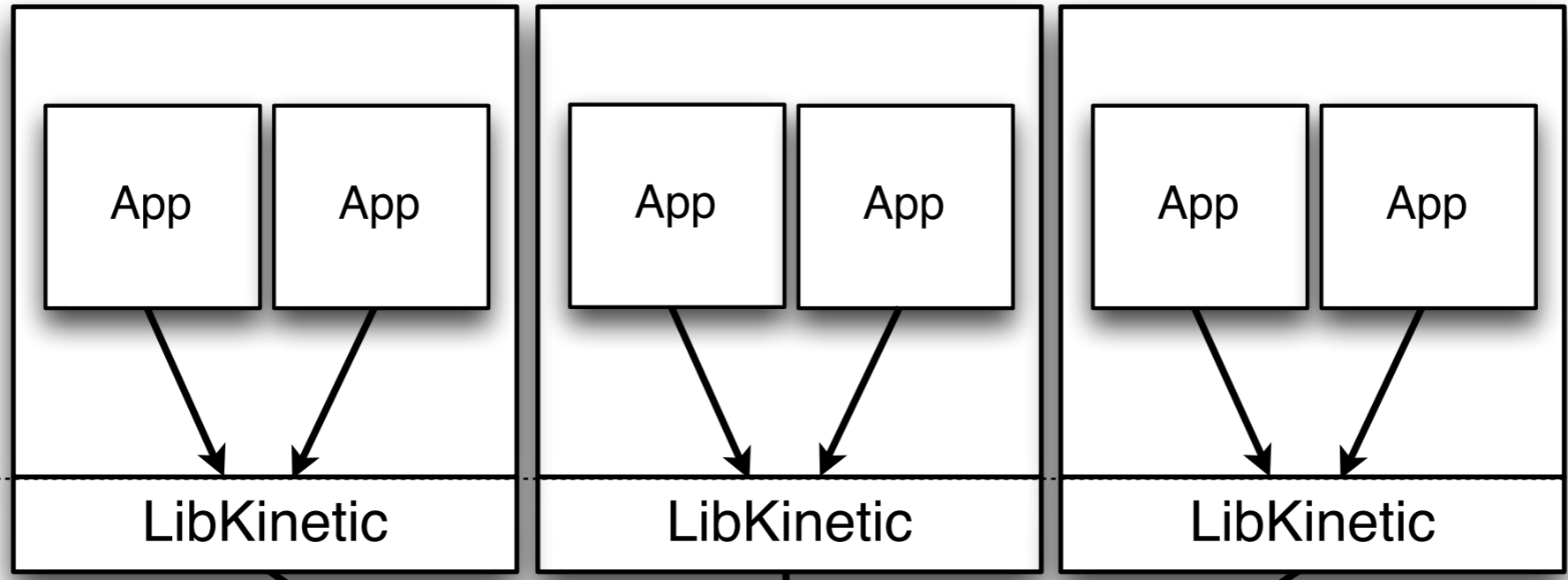


- Proprietary
- to System Vendor

- GPL
- Standard

- Proprietary
- to Seagate

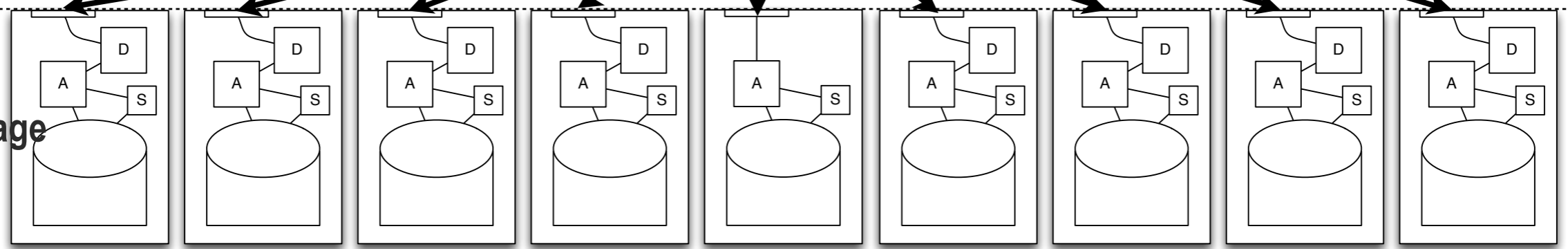
- Application
- Clustering
- Management



- Interconnect

- ProtoBuf
- TCP/IP/GbE

- Storage



- Proprietary
- to System Vendor

- GPL
- Standard

- Proprietary
- to Seagate

SAS versus



Kinetic Open Storage



- Standard form factor
- 2 SAS ports
- SCSI command set
 - data = read (LBA, count)
 - write (LBA, count, data)
 - LBA :: [0, max]
 - data :: count * 512 bytes
 - CRC on cmd and PI on block

- Standard form factor
- 2 Ethernet ports (same connector)
- Kinetic key/value API
 - value = get (key)
 - put (key, value)
 - delete (key)
 - key :: 1 byte to 4 KiB
 - value :: 0 bytes to 1 MiB
 - HMAC on cmd and SHA on value

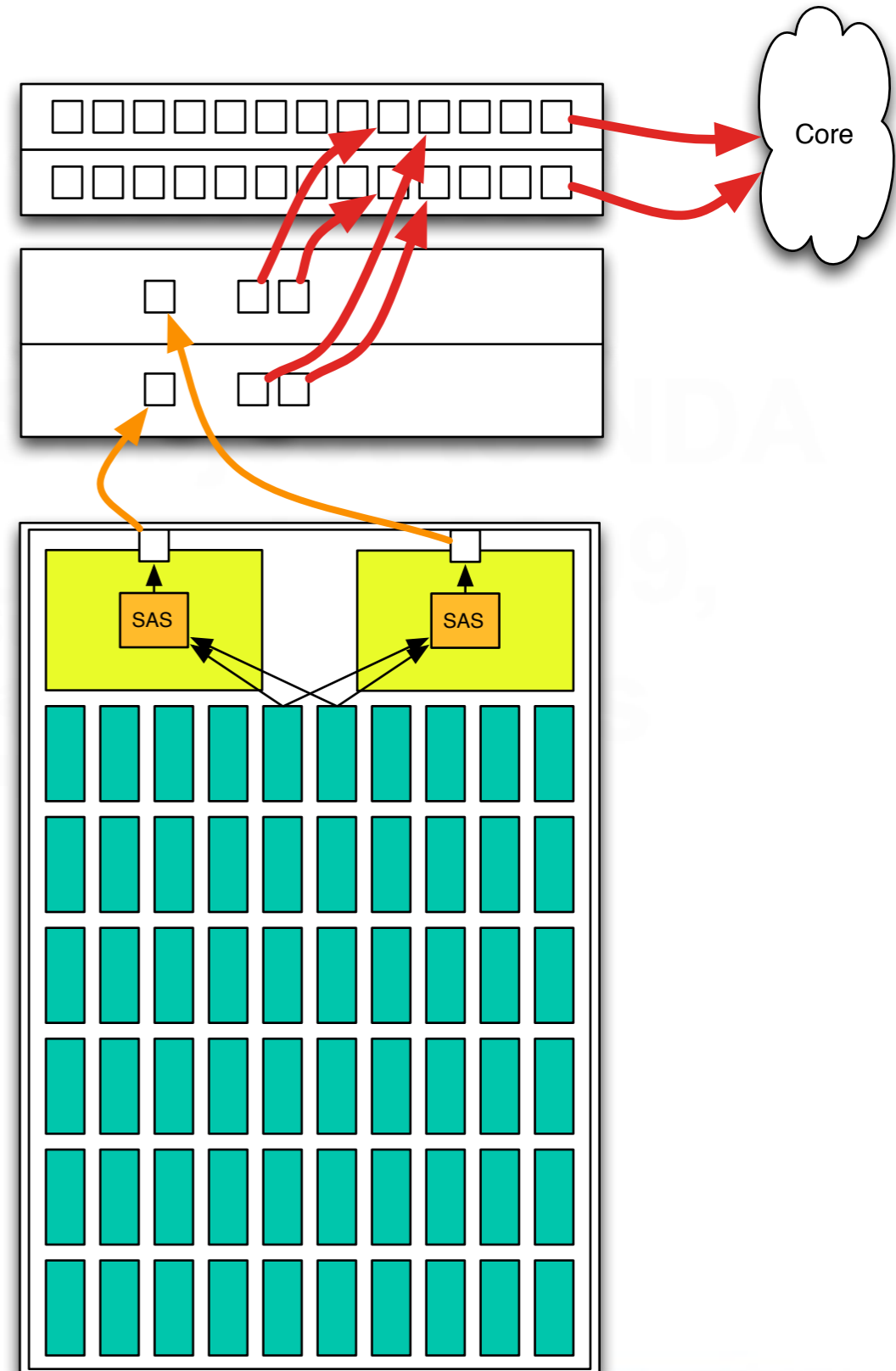
Typical HA High Density

Intel server

- Double Socket
- 48GB Ram
- 1000w

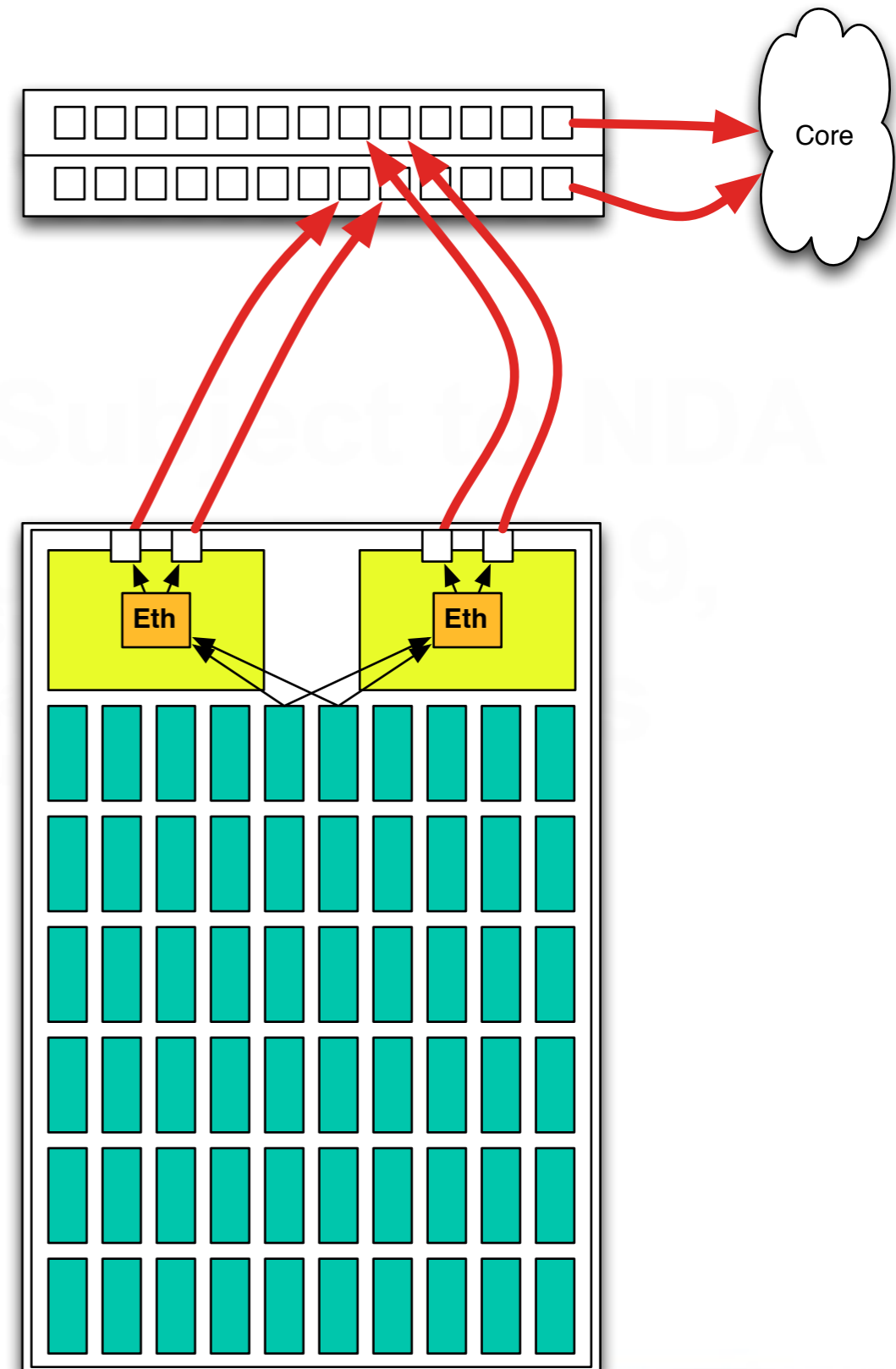
SAS tray

- Connected to the server



Low cost HA Configuration

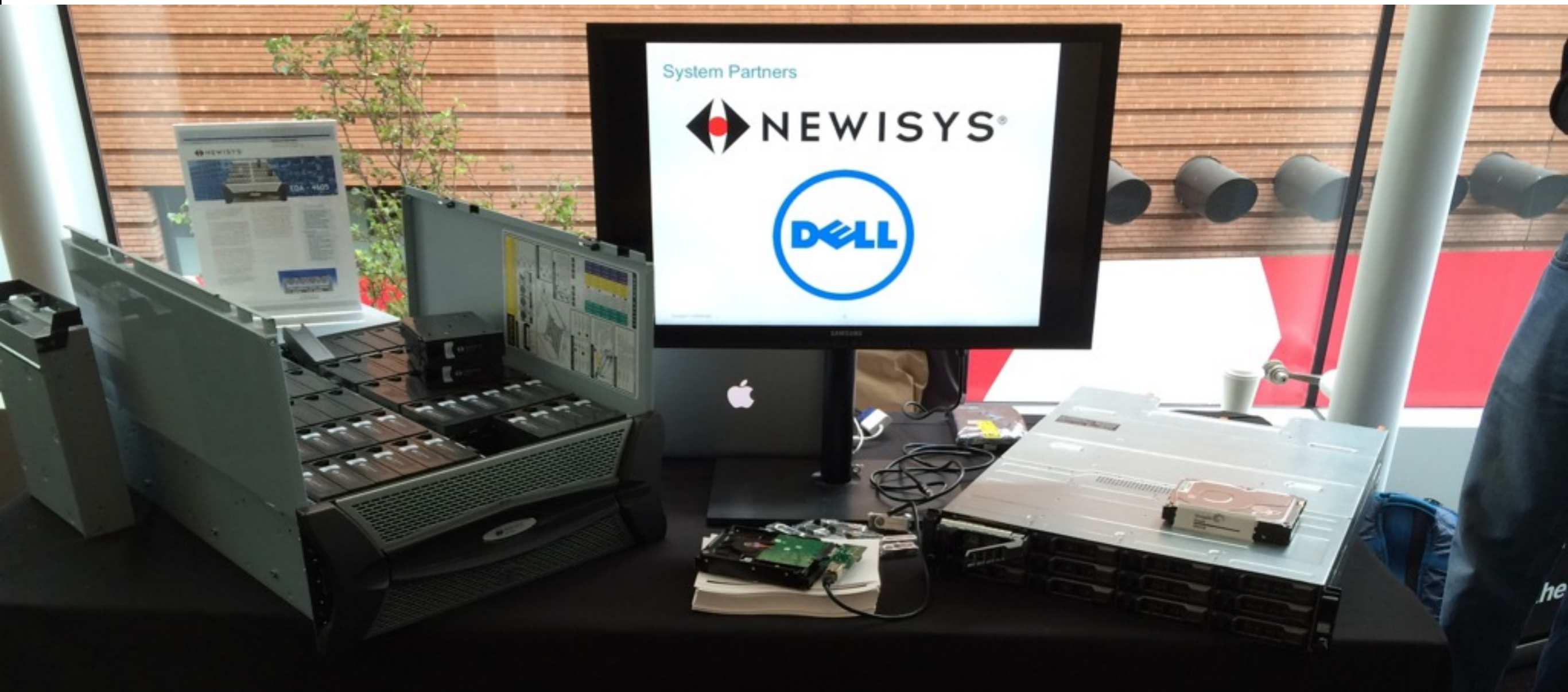
- Each drive talks to both switches
- Each switch has 2 by 10Gb/s Ethernet
- Kinetic Tray talks directly to ToR
- No servers



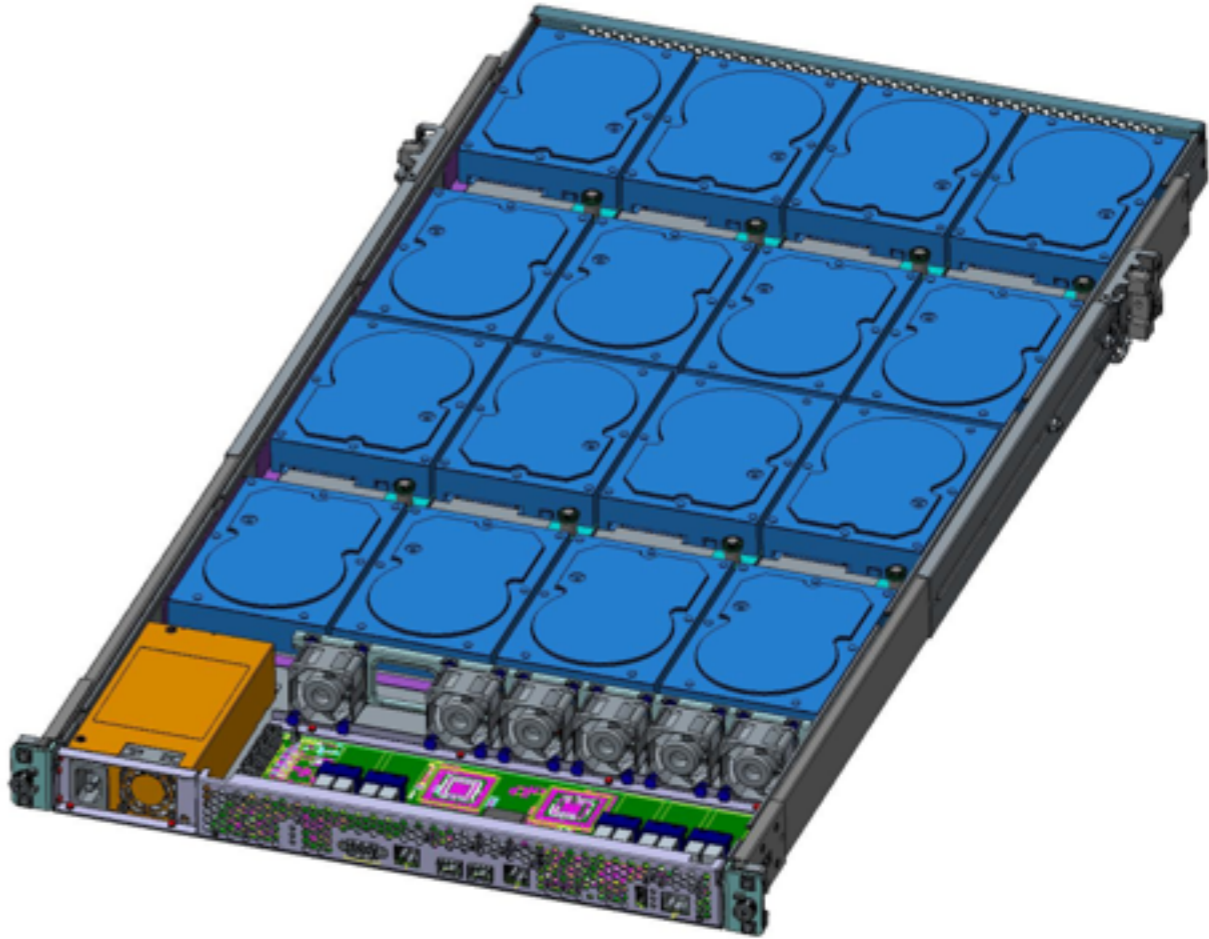
System Hardware

Typical JBOD architecture

- Does not require a server, just JBODs to the ToR Switch
- 10 JBODS × 60 drives × 4TB = 2.4PB/Rack



System Hardware



Kinetic *Device*

RPC to Key/Value database

P2P (Device to Device) copy of key ranges

Communicate using common TCP/IP features

Multiple masters - Data sharing between machines

Configurable caching per command

- WriteThrough, WriteBack, Flush

Local space management

Kinetic *Device*

High Density HDD

- \ll \$/GB
- \ll \$/GB/s

Flash based persistent storage

- \ll \$/Transaction/s

Other storage system

- Ram, Memristor, etc.
- Some persistent, some not

All communicating using P2P to move between tiers

Kinetic *System*

Clustering (performance, reliability, management)

Compatibility with large scale applications (S3, etc.)

Centralized Management

- Reliability, availability, durability

Kinetic *Systems*

Distributed Hash Table

- Quorum or Paxos based

Object Storage System

- Swift, Ceph, HDFS, Scality

Posix

- Scality, Ceph, H-Flat

RYO

Goals of API

Data movement

- Get/put/delete/getnext/getprevious
- Versioned (== for success), options

Range operations

Multiple masters

- Authentication/Integrity/Authorization

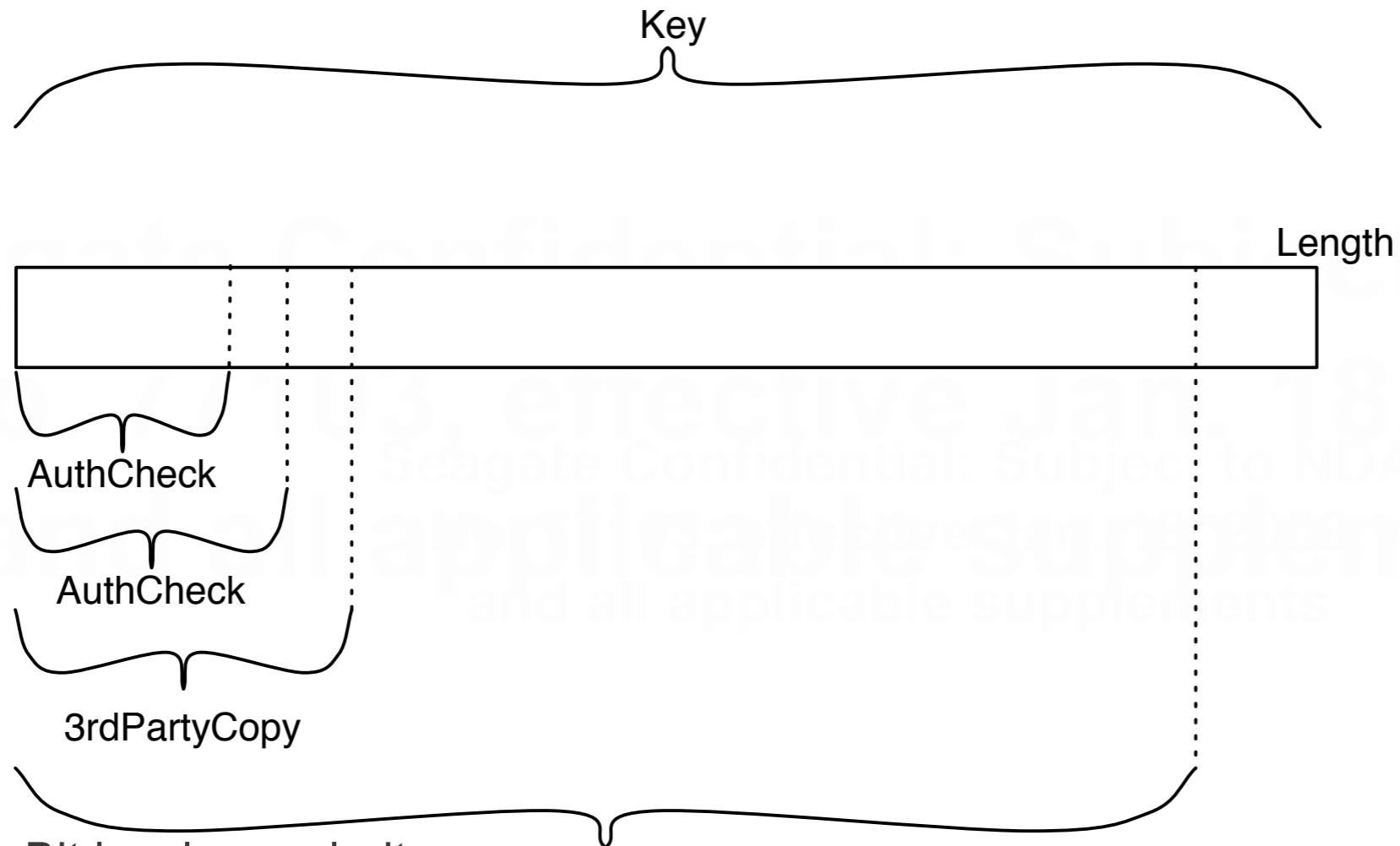
Cluster-able

- Simple cluster configuration version enforcement

P2P copy

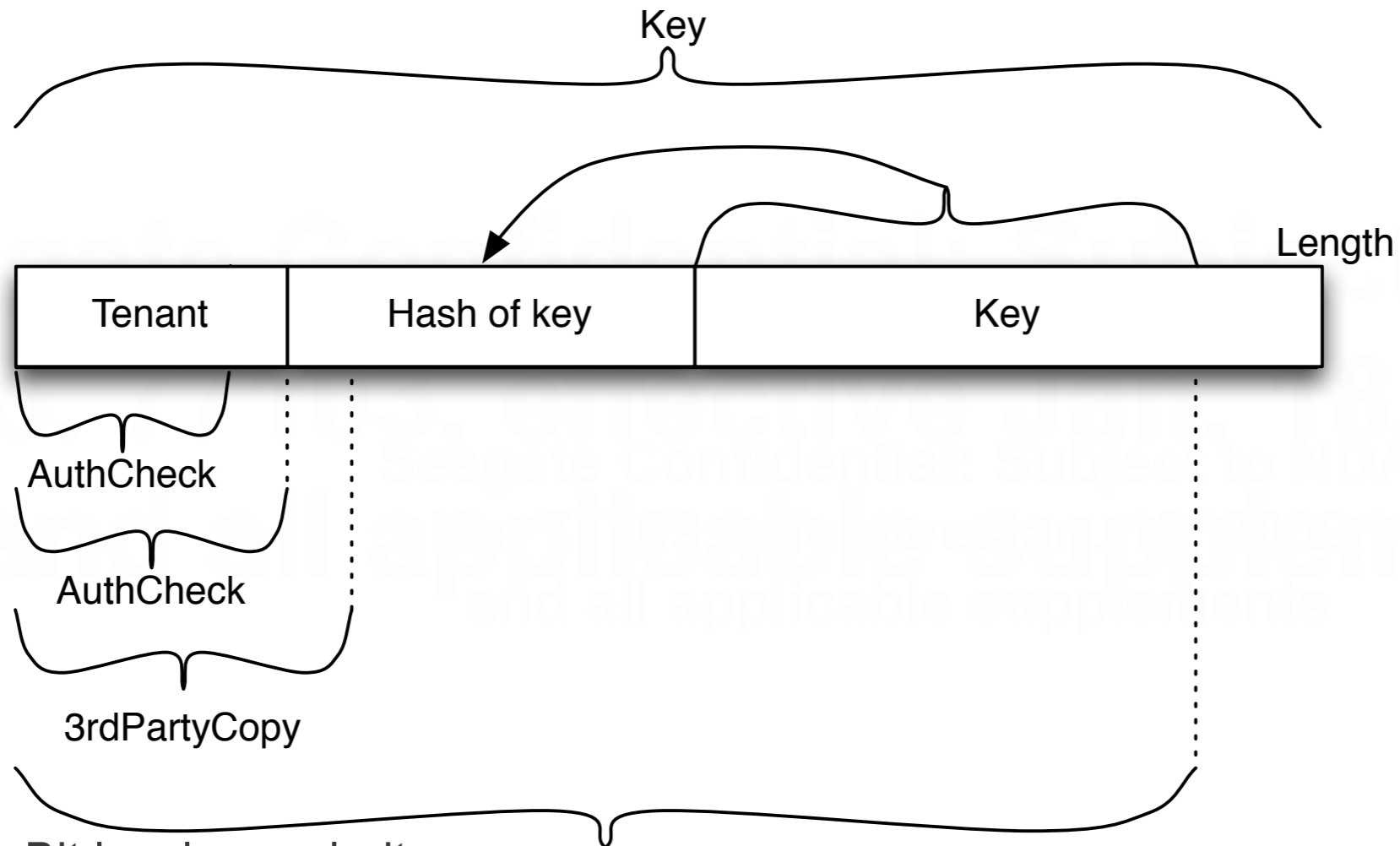
Management

Sample key format (1 of 2)



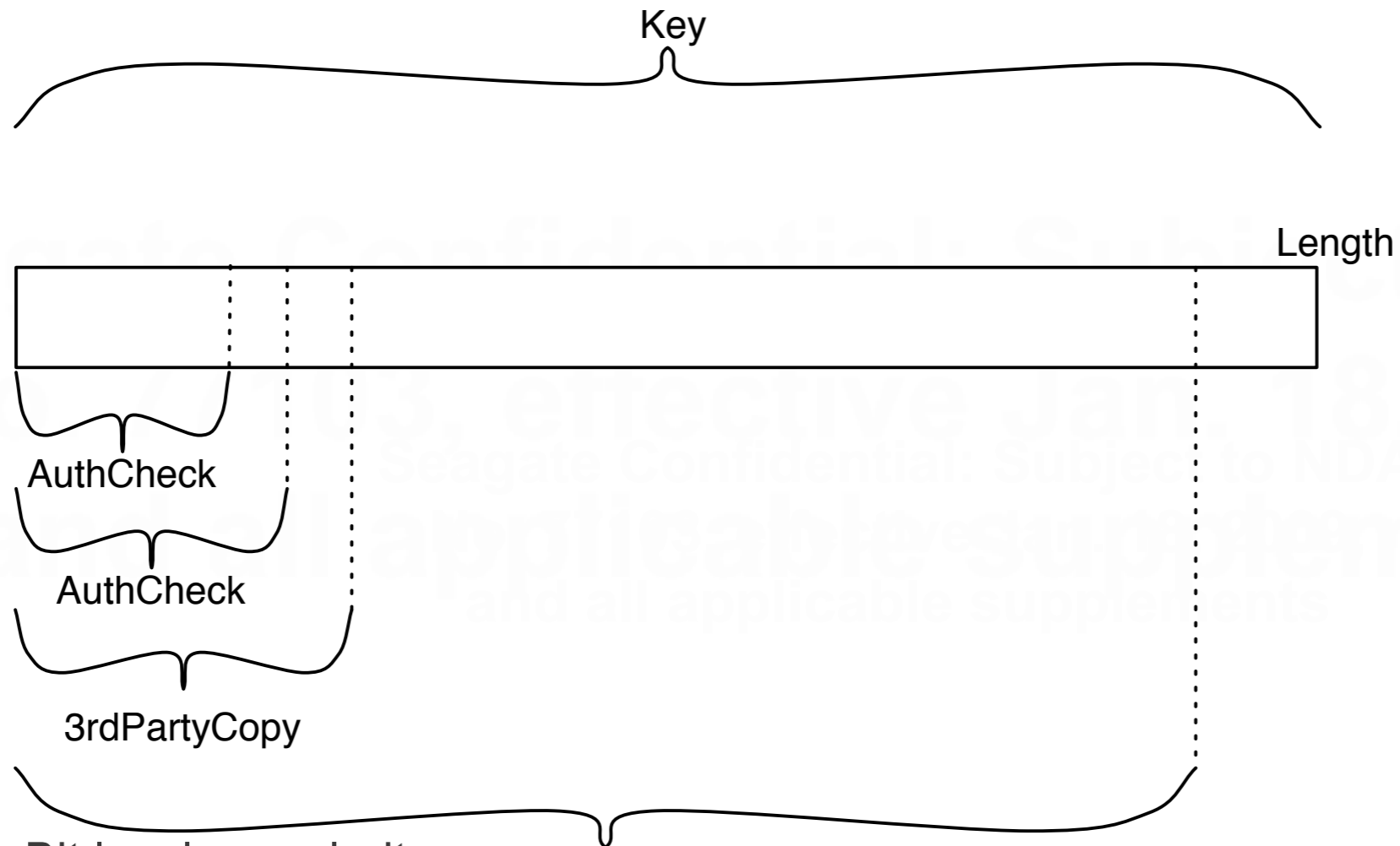
- Bit level granularity
 - » 3rd party starts at top bit of first byte with bit length
 - » Auth check starts at any bit and has bit length
- Individual system vendors can do as they wish

Sample key format (1 of 2)



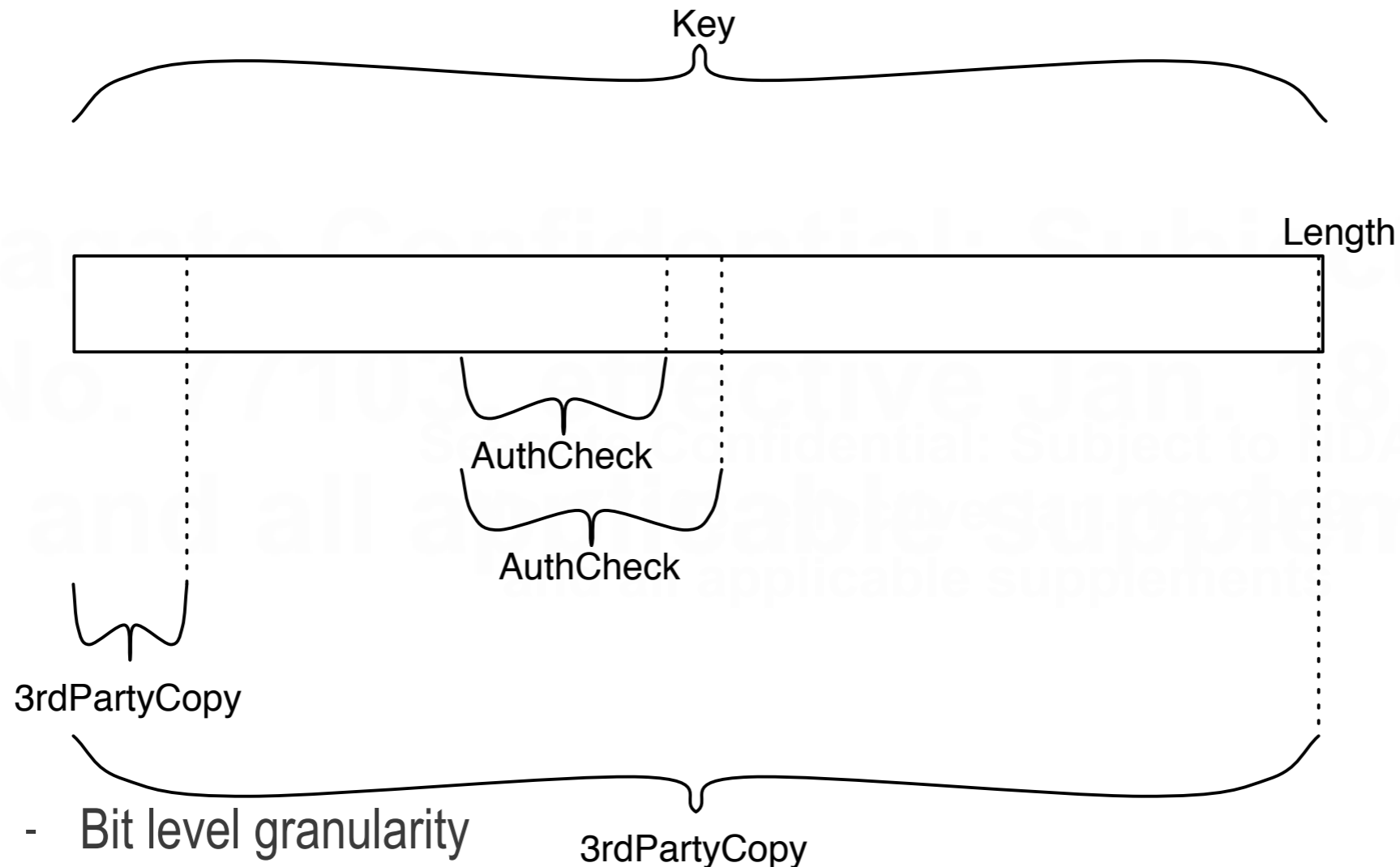
- Bit level granularity
 - » 3rd party starts at top bit of first byte with bit length
 - » Auth check starts at any bit and has bit length
- Individual system vendors can do as they wish

Sample key format (1 of 2)



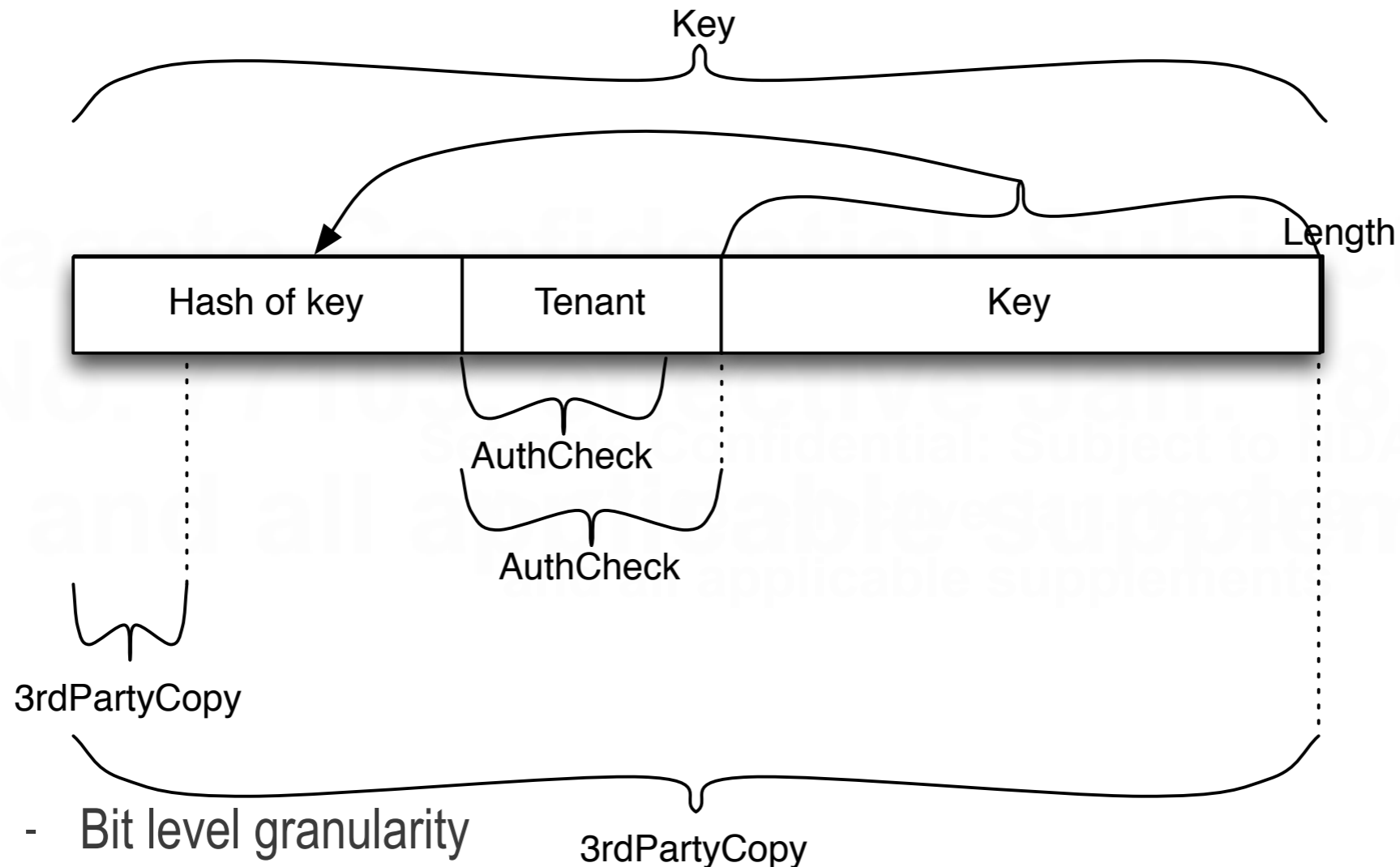
- Bit level granularity
 - » 3rd party starts at top bit of first byte with bit length
 - » Auth check starts at any bit and has bit length
- Individual system vendors can do as they wish

Sample key format (2of2)



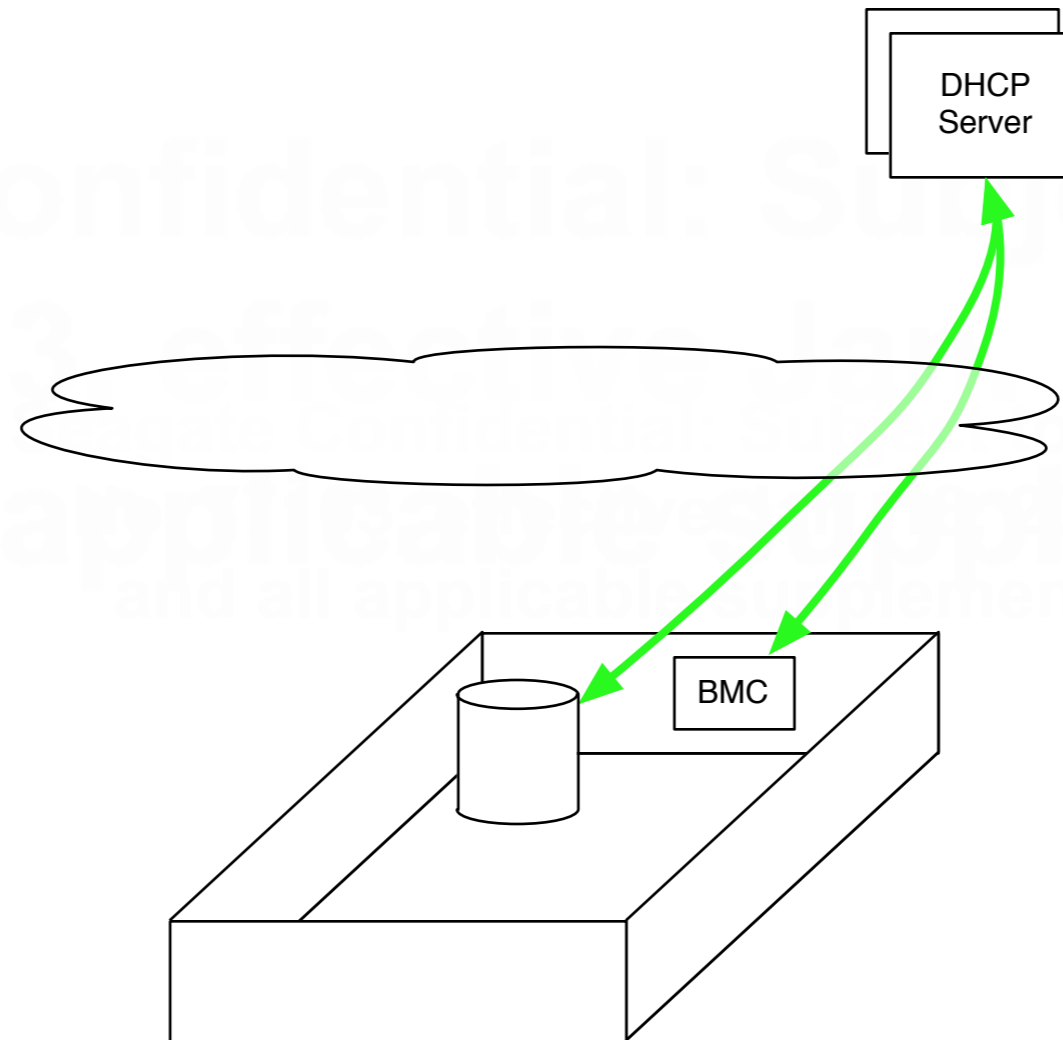
- Bit level granularity
 - » 3rd party starts at top bit of first byte with bit length
 - » Auth check starts at any bit and has bit length
- Individual system vendors can do as they wish

Sample key format (2of2)

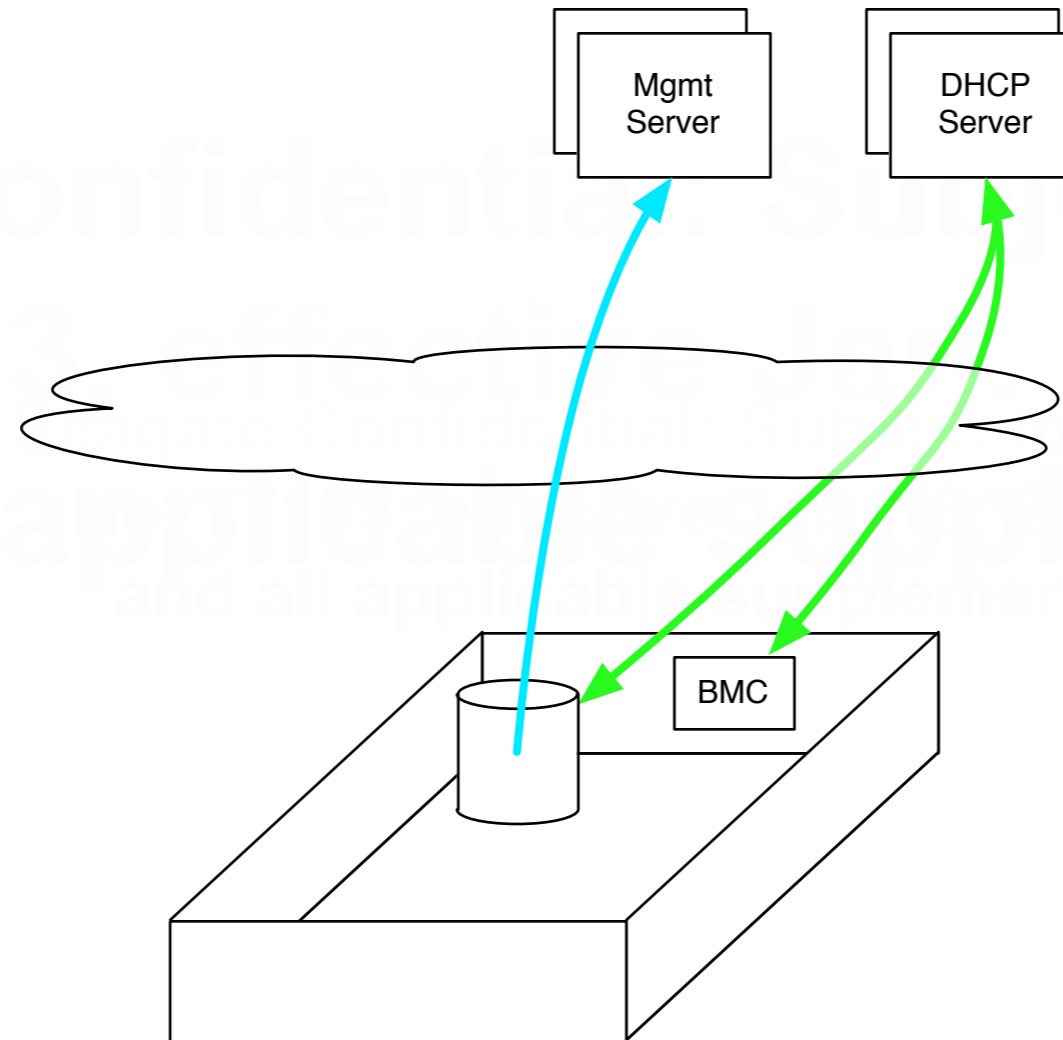


- Bit level granularity
 - » 3rd party starts at top bit of first byte with bit length
 - » Auth check starts at any bit and has bit length
- Individual system vendors can do as they wish

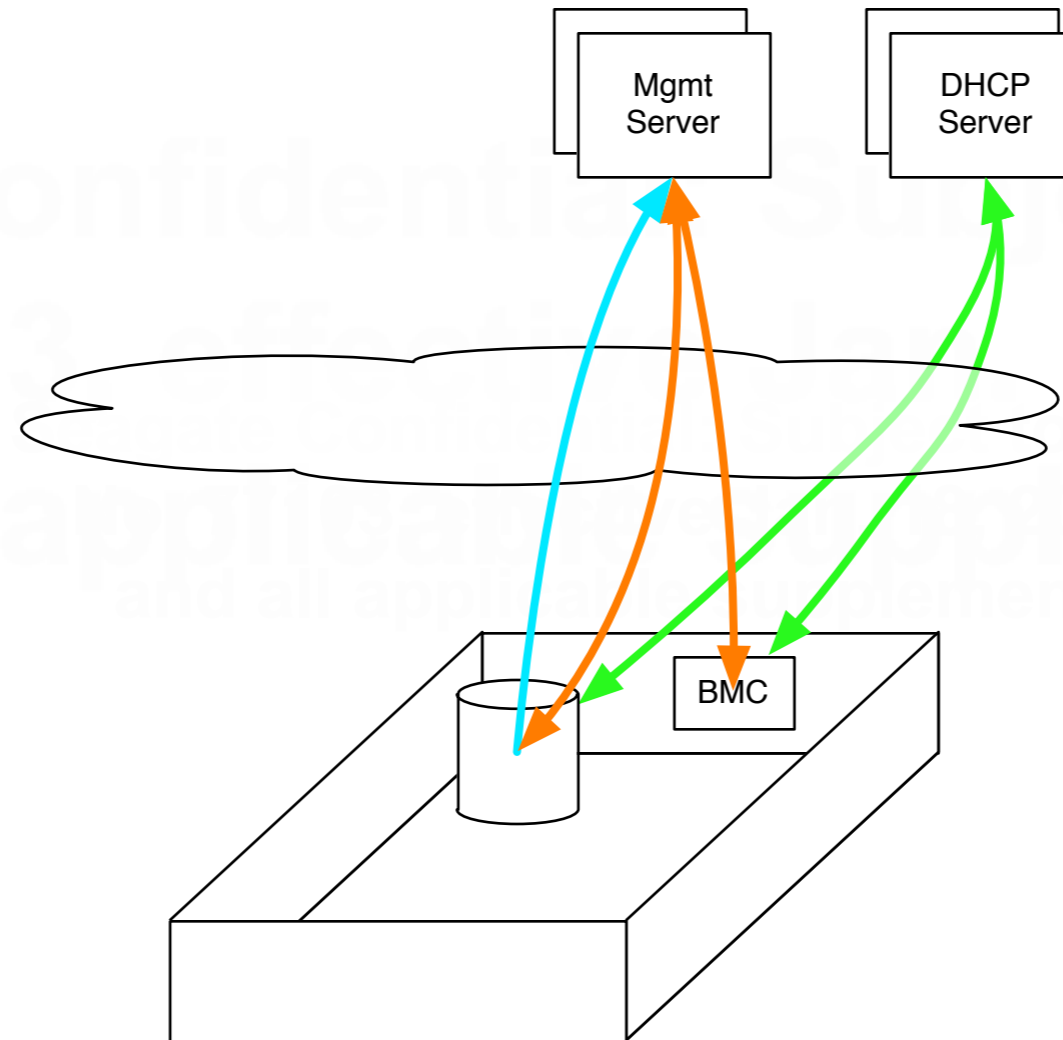
Bootstrapping devices



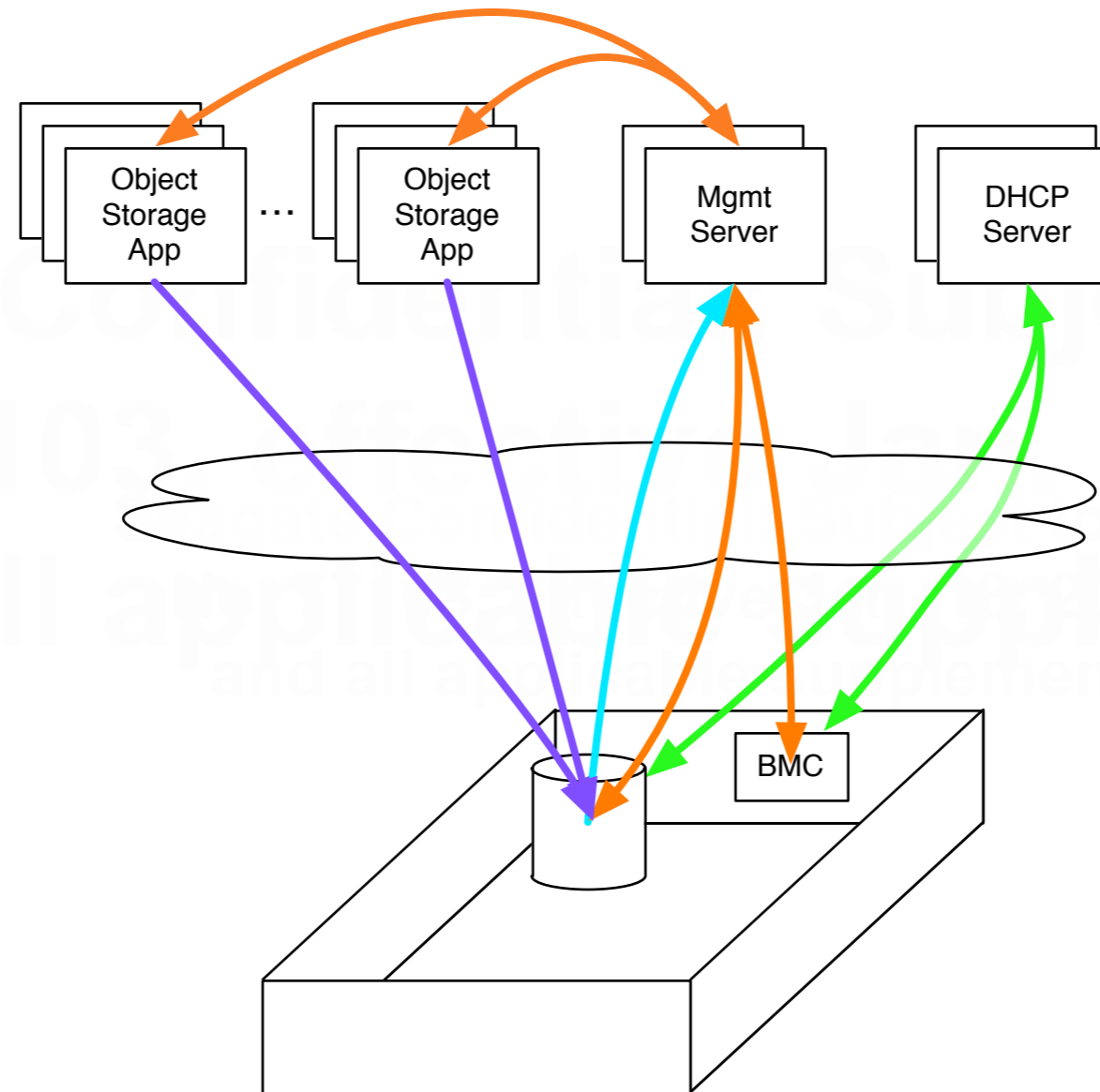
Bootstrapping devices



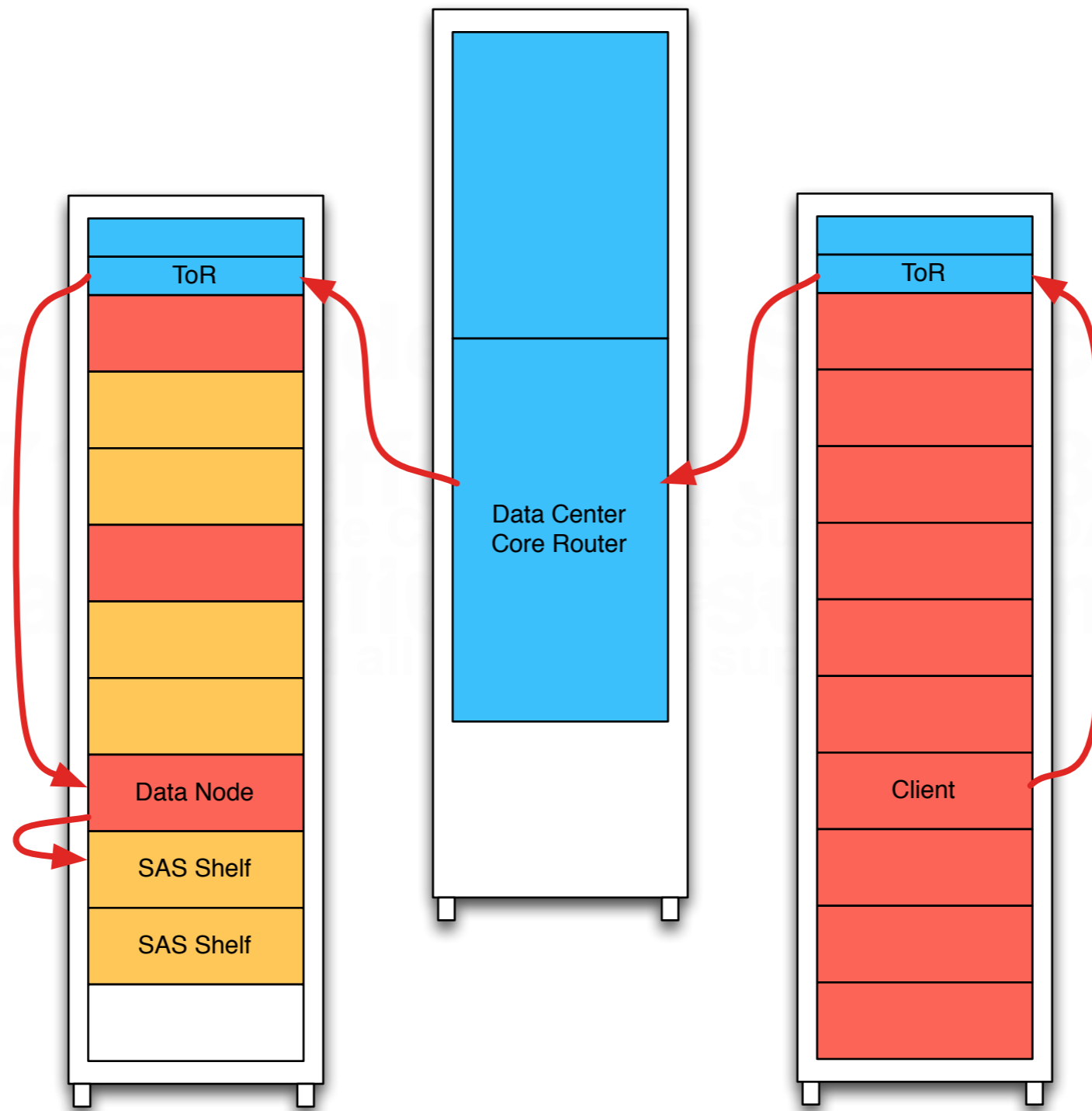
Bootstrapping devices



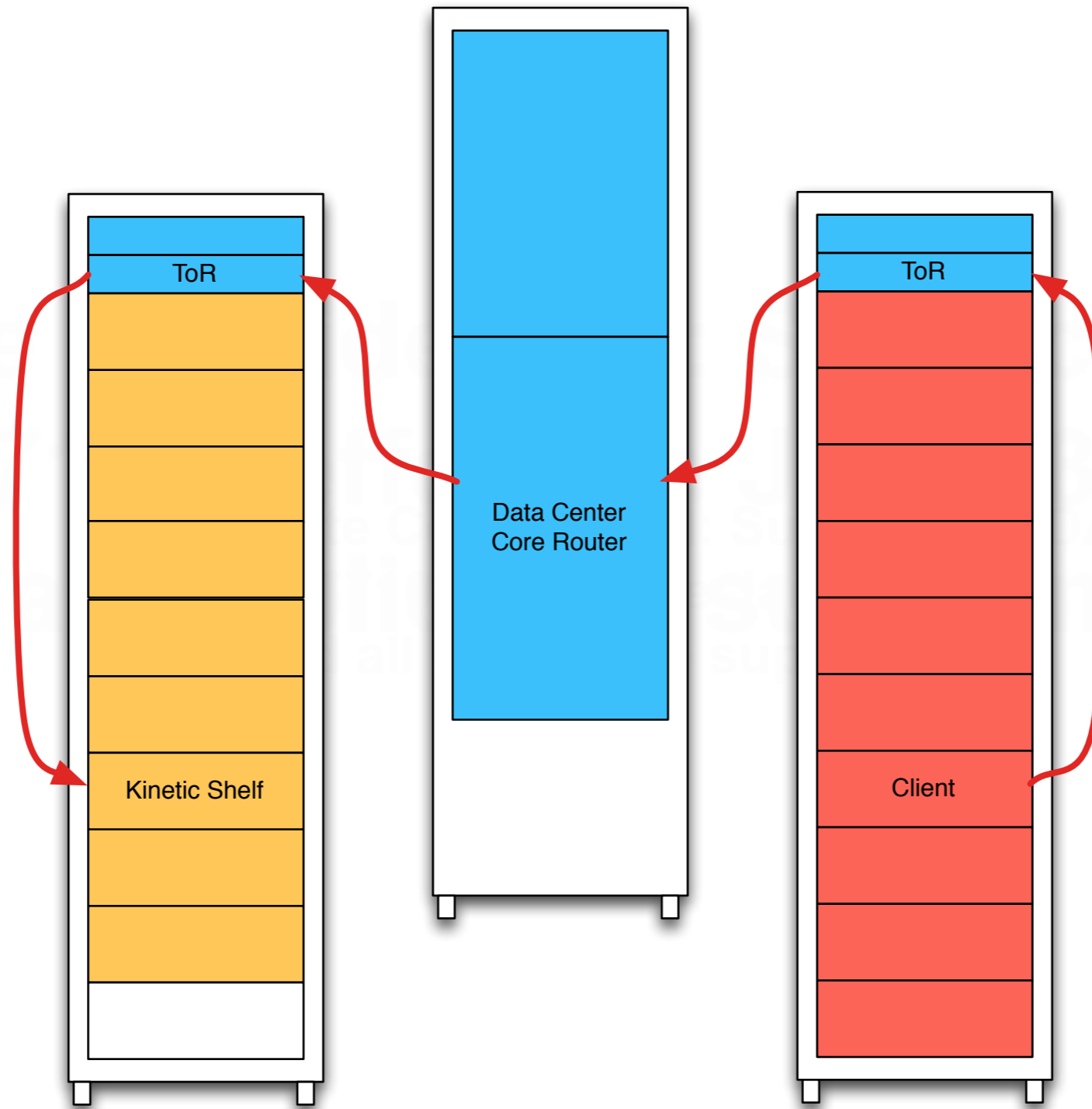
Bootstrapping devices



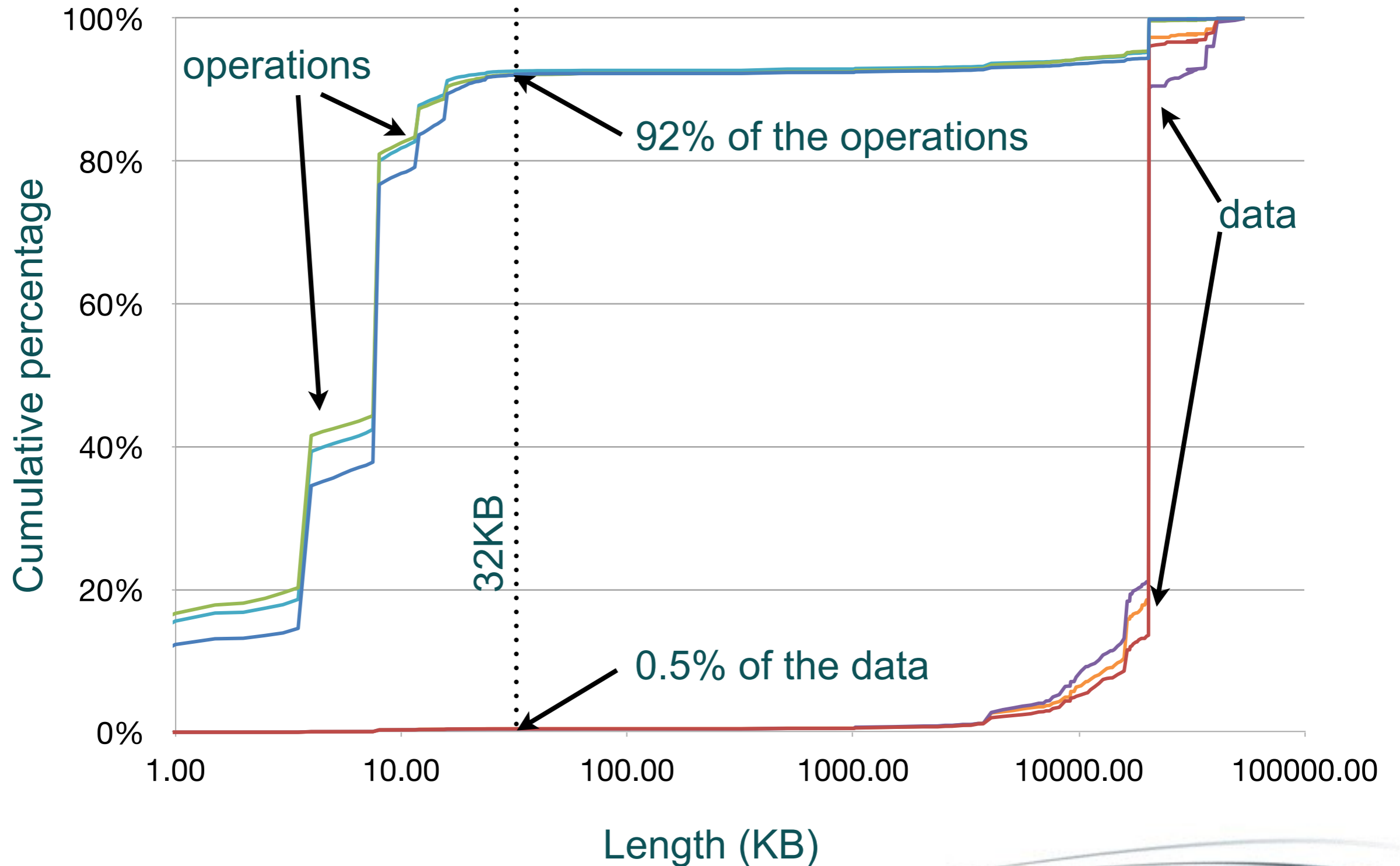
Existing Traffic Flow



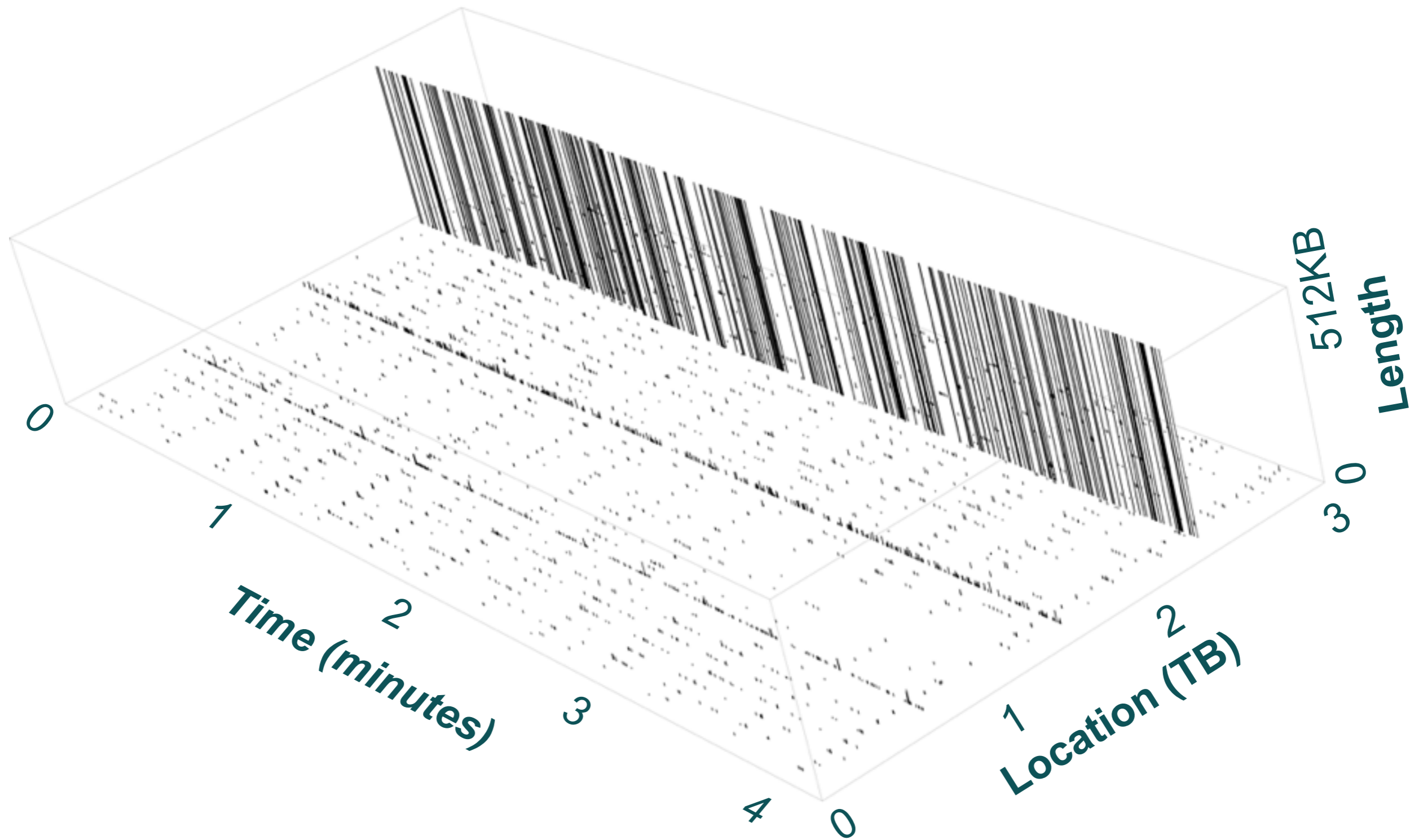
Kinetic Traffic Flow



Cumulative operations ordered by length



Map of Operations

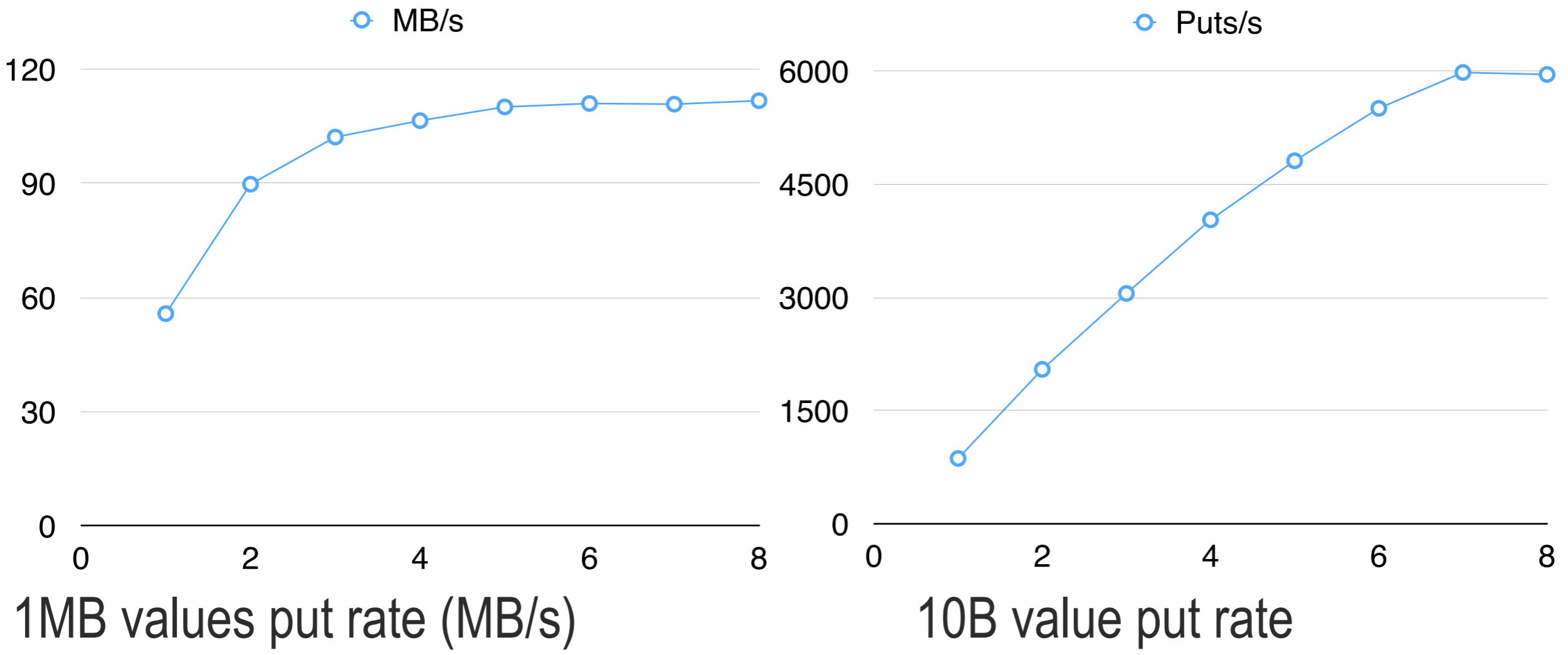


Performance Metrics

Same normal performance expectations

- Sequential Write: 50MB/s
- Random Write: 50MB/s
- Sequential Read: 50MB/s
- Random Read: 20% slower than traditional drives

Write Performance Results

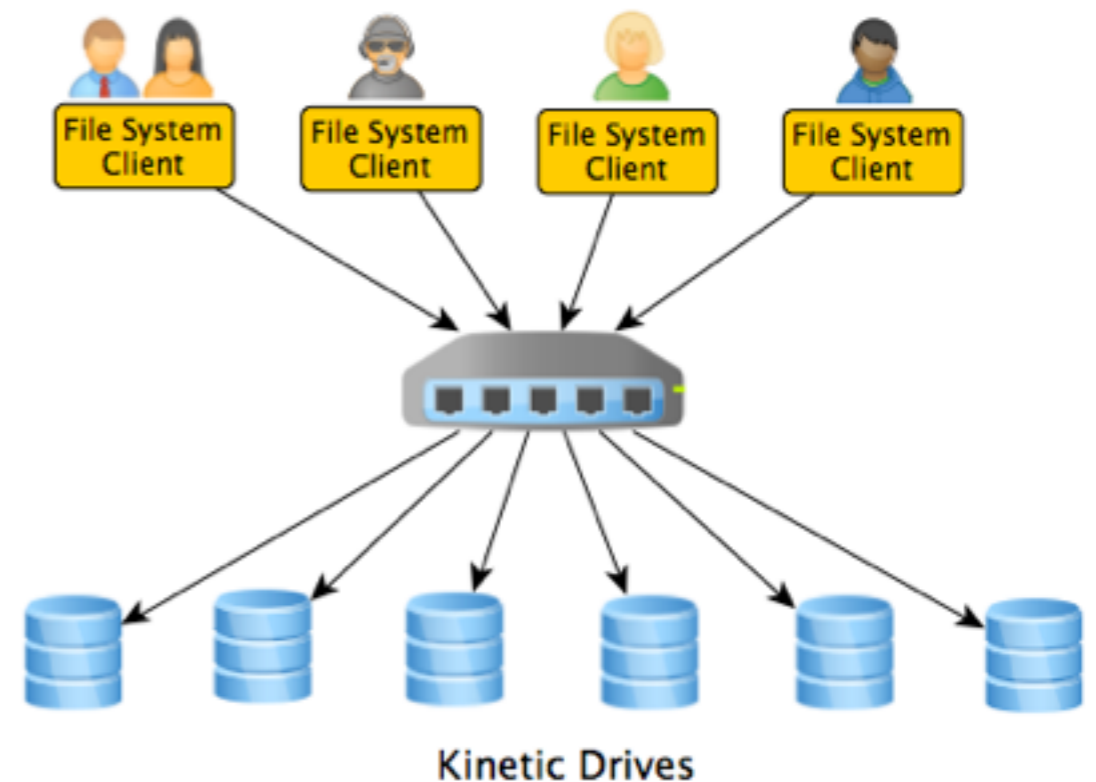
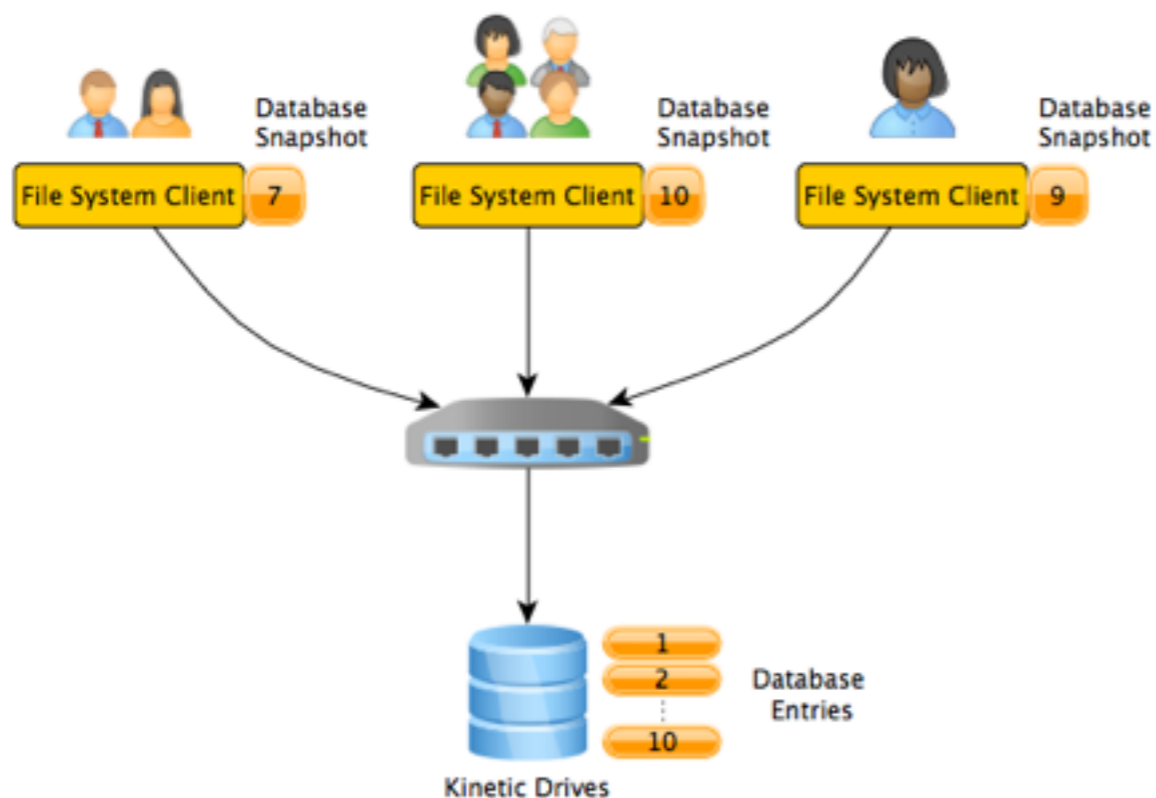


H-Flat: Posix over Kinetic

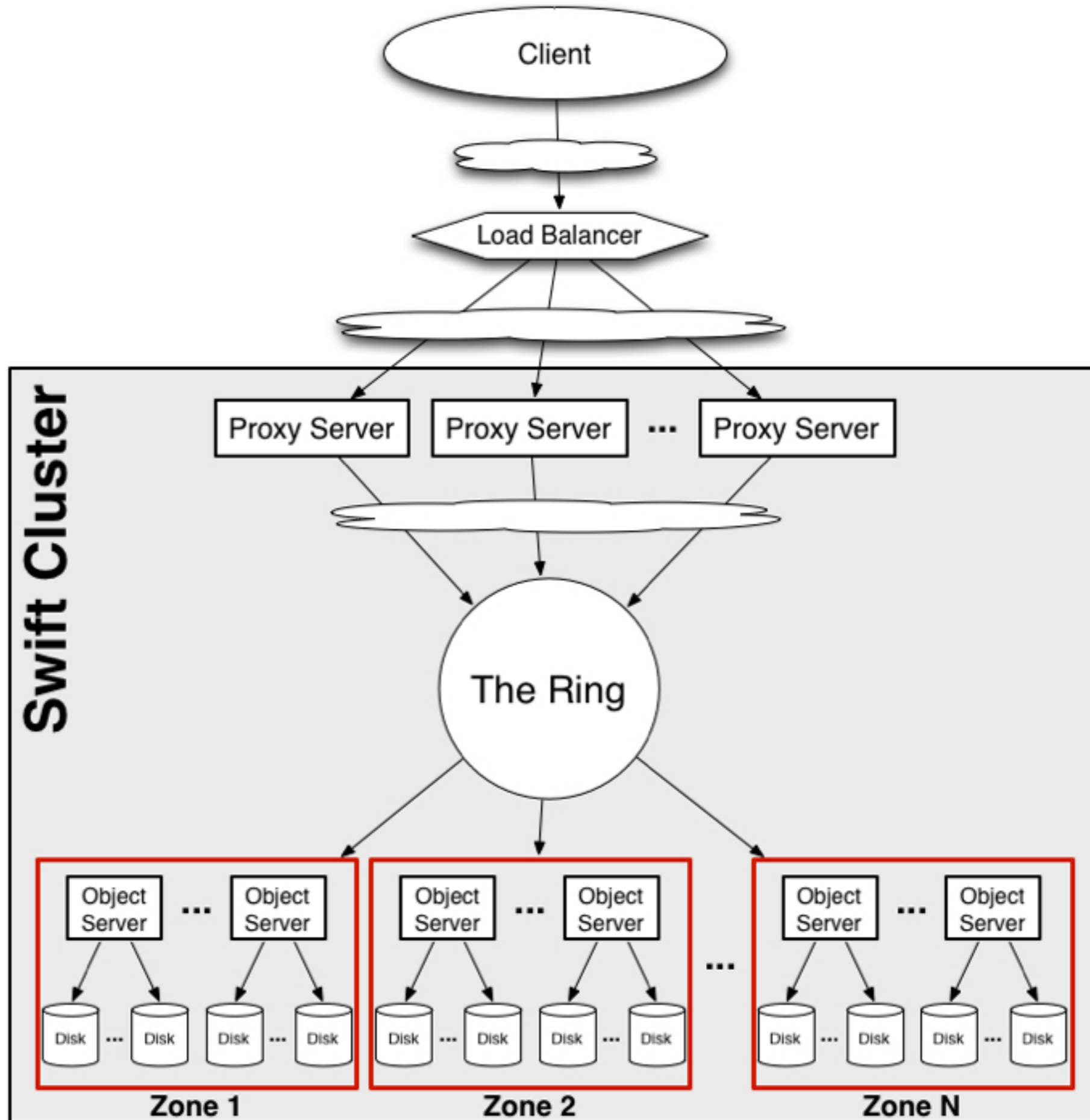
Hierarchical Functionality, Flat Namespace

- standard hierarchical file system semantics (POSIX) while using a flat namespace internally
- combine the performance and scalability of a key-value storage system with a file system interface.

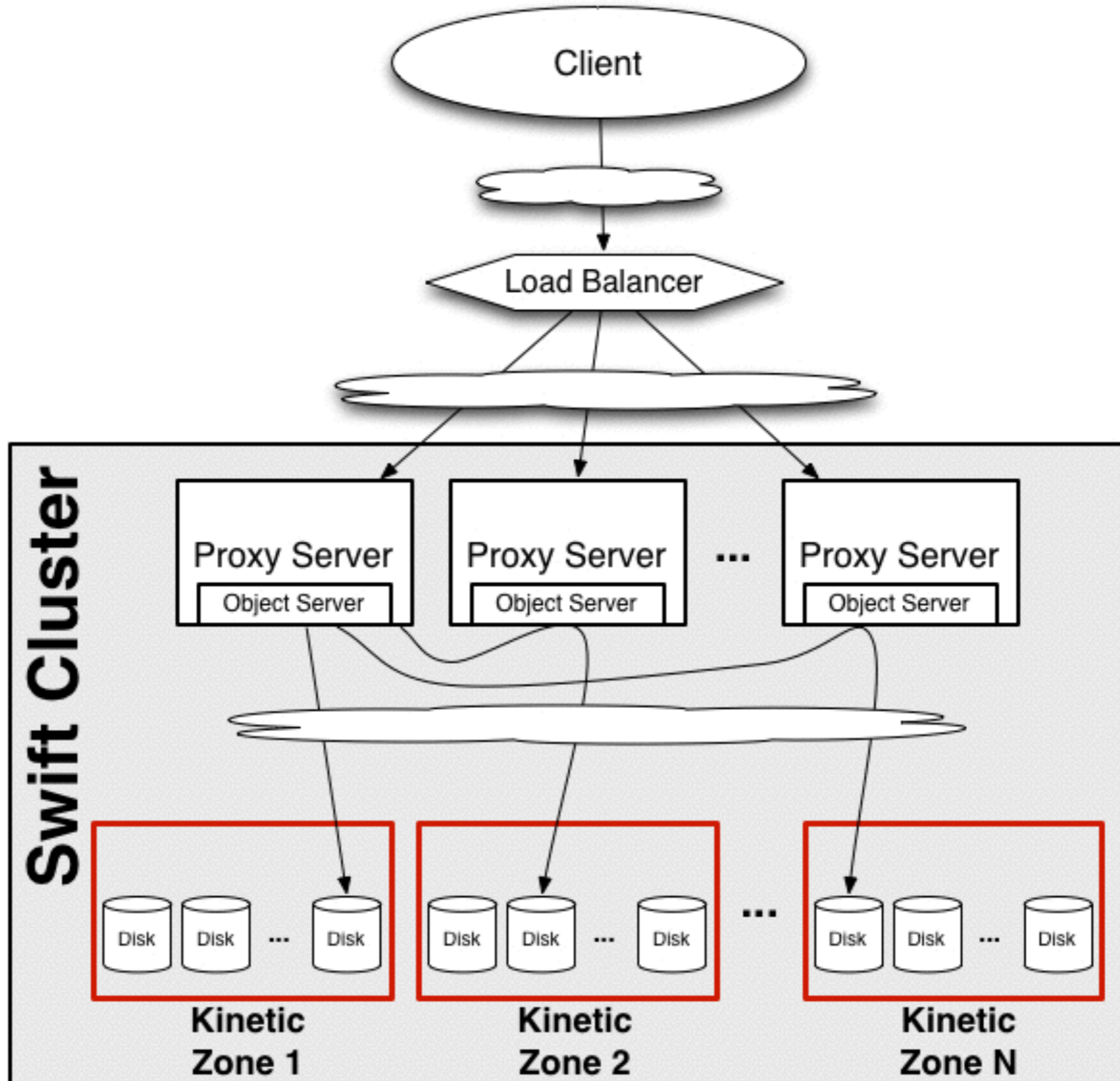
Seagate Confidential: Subject to NDA



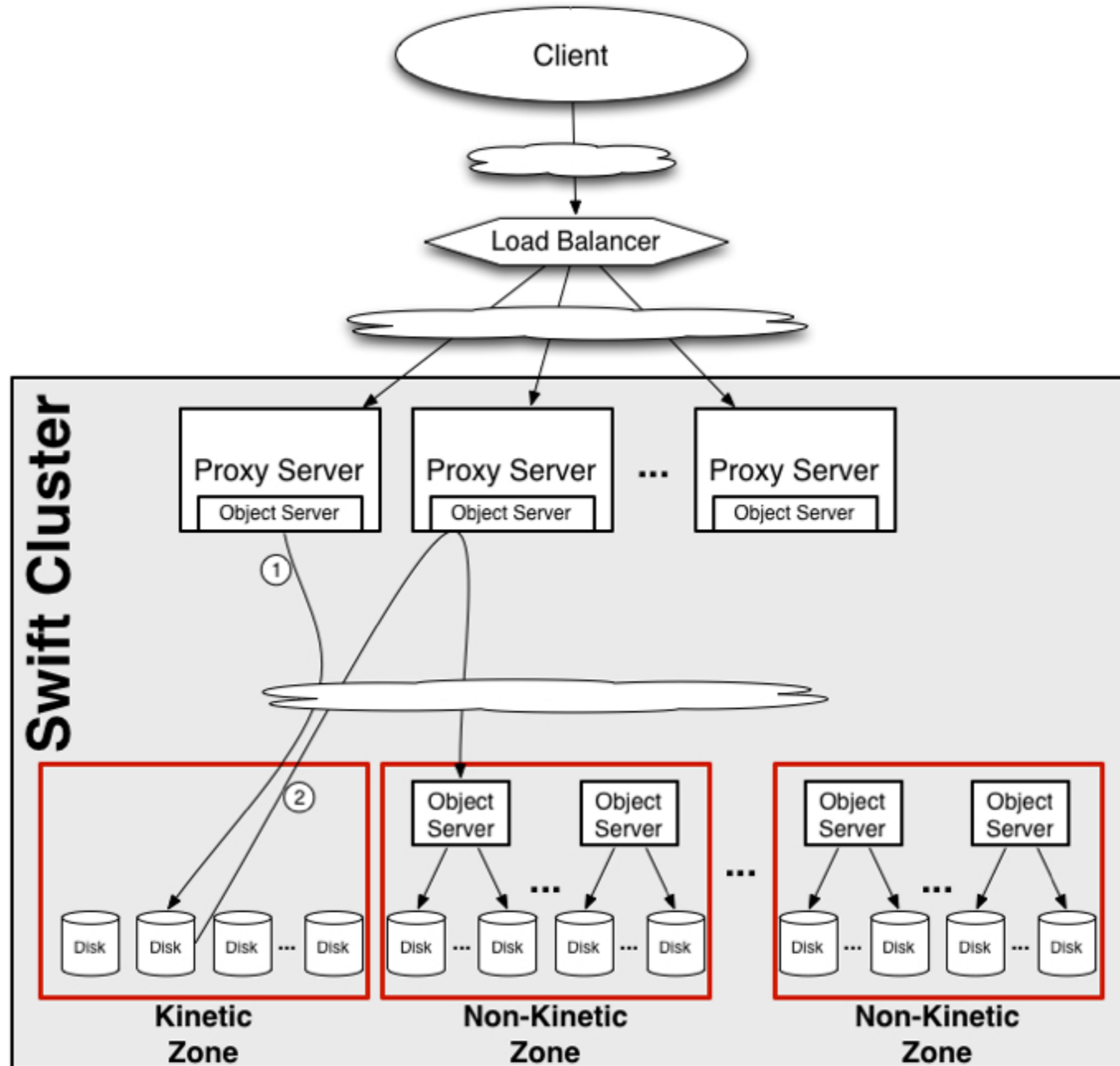
Swift



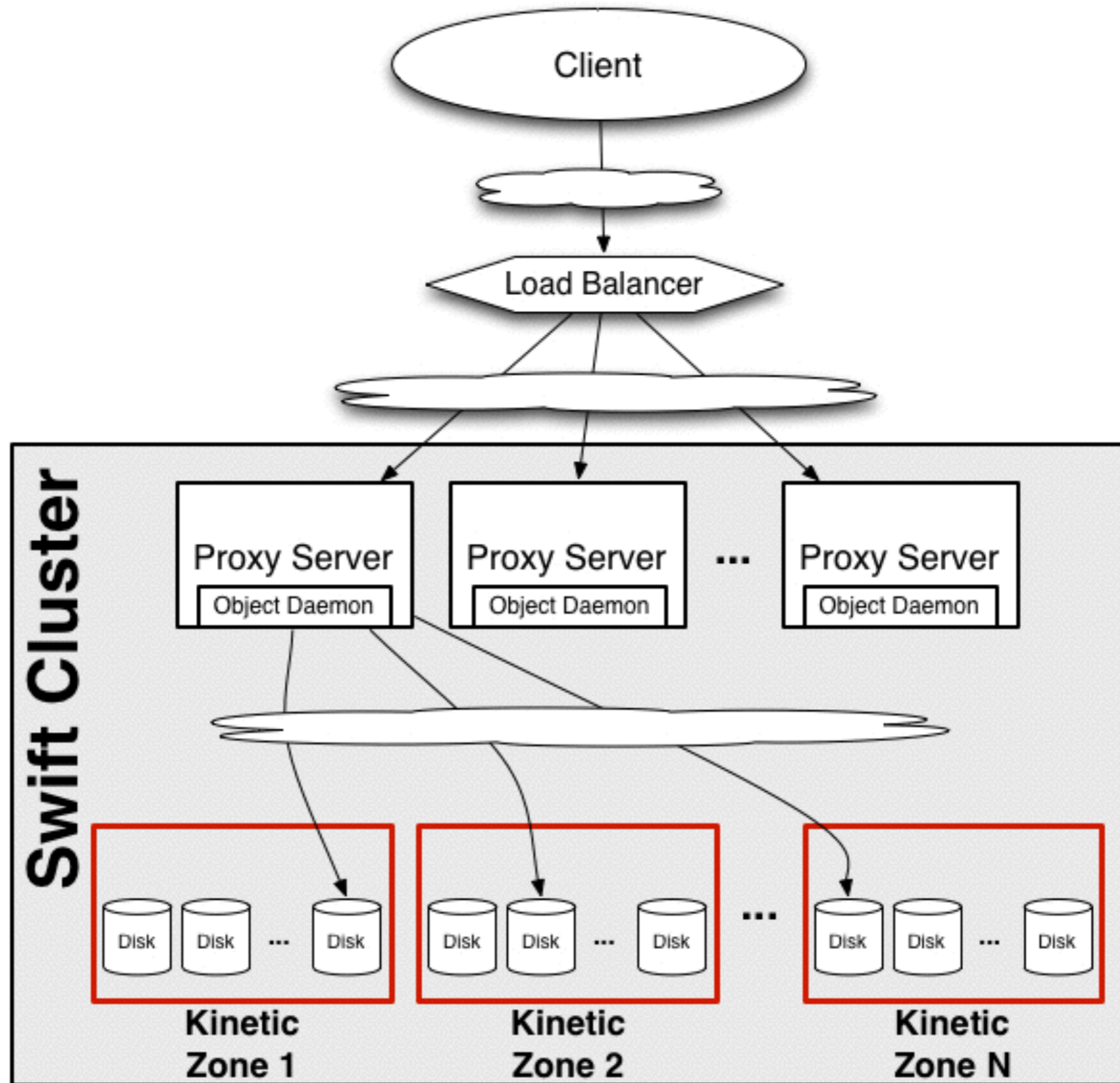
Swift



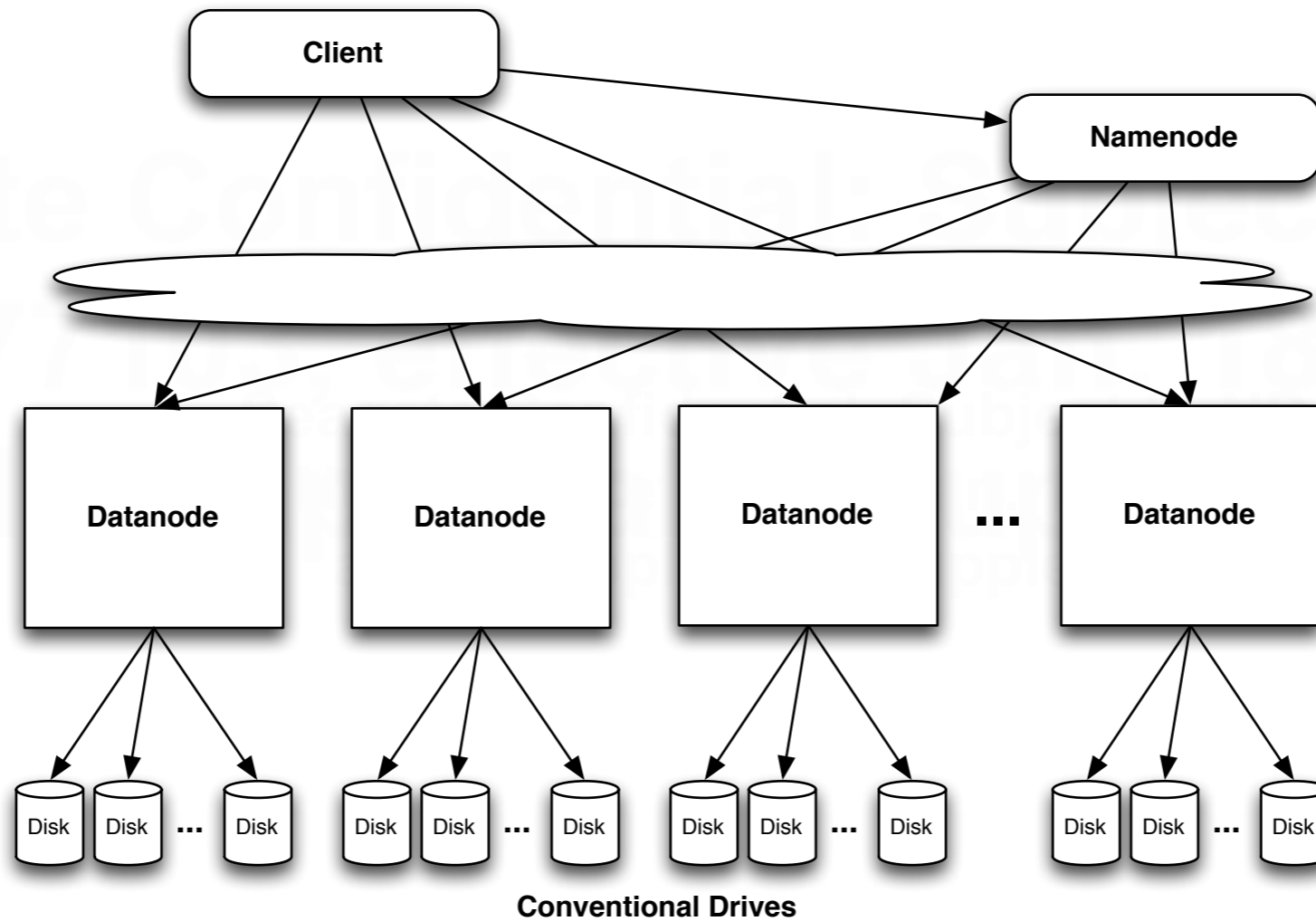
Swift



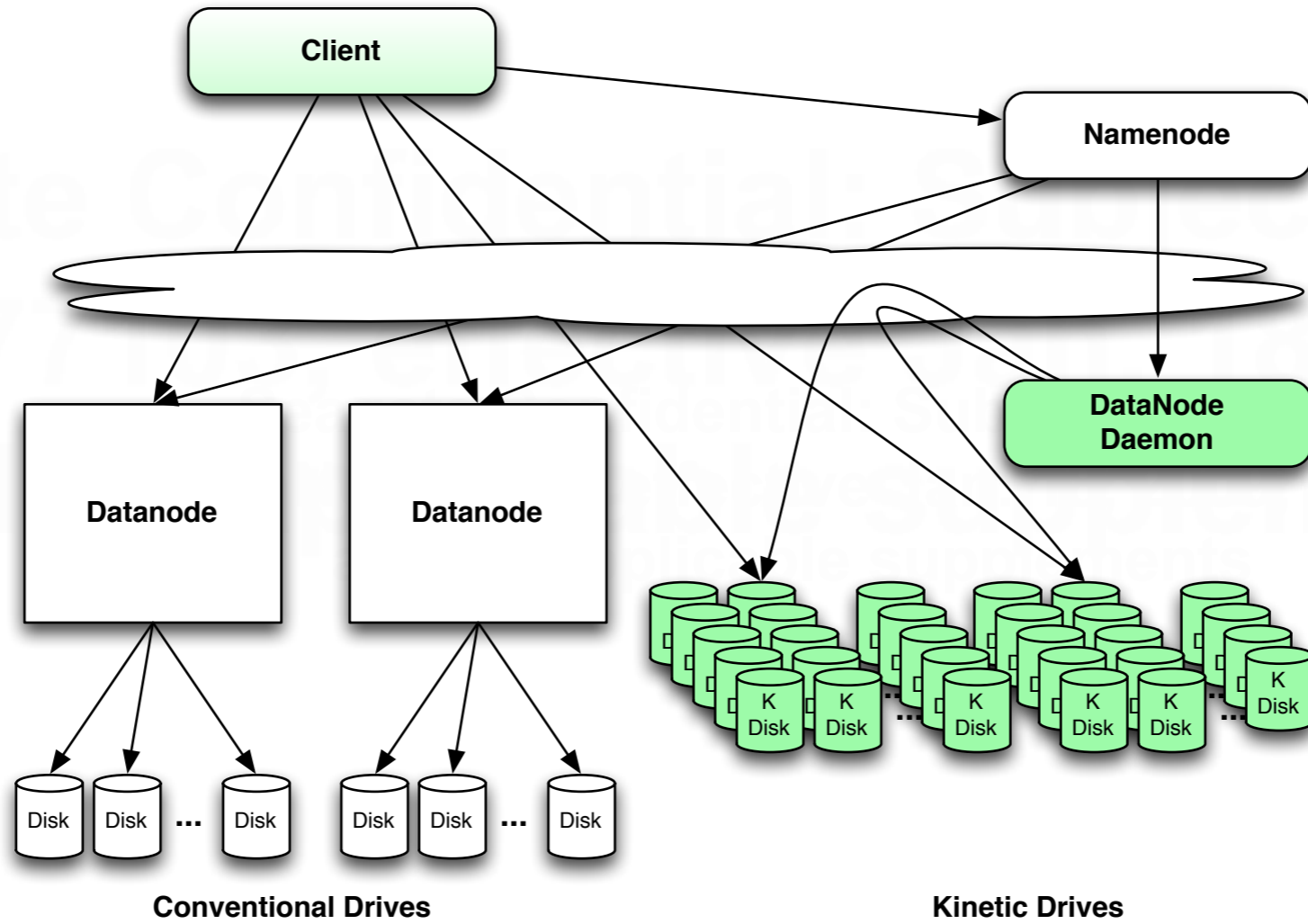
Swift



Conventional HDFS System



HDFS on Kinetic



Conclusion

Next Generation Storage Devices

- Dis-intermediates cloud applications to drive
- Enable innovation in hardware and software ecosystem
- Lower TCO

Integration with:

- Swift
- HDFS
- Basho Riak
- Ceph
- Scality

More information

- <https://developers.seagate.com>
- <http://github.com/Seagate>

Seagate Confidential: Subject to NDA
No. 77103, effective Jan. 18, 2009,
and all applicable supplements