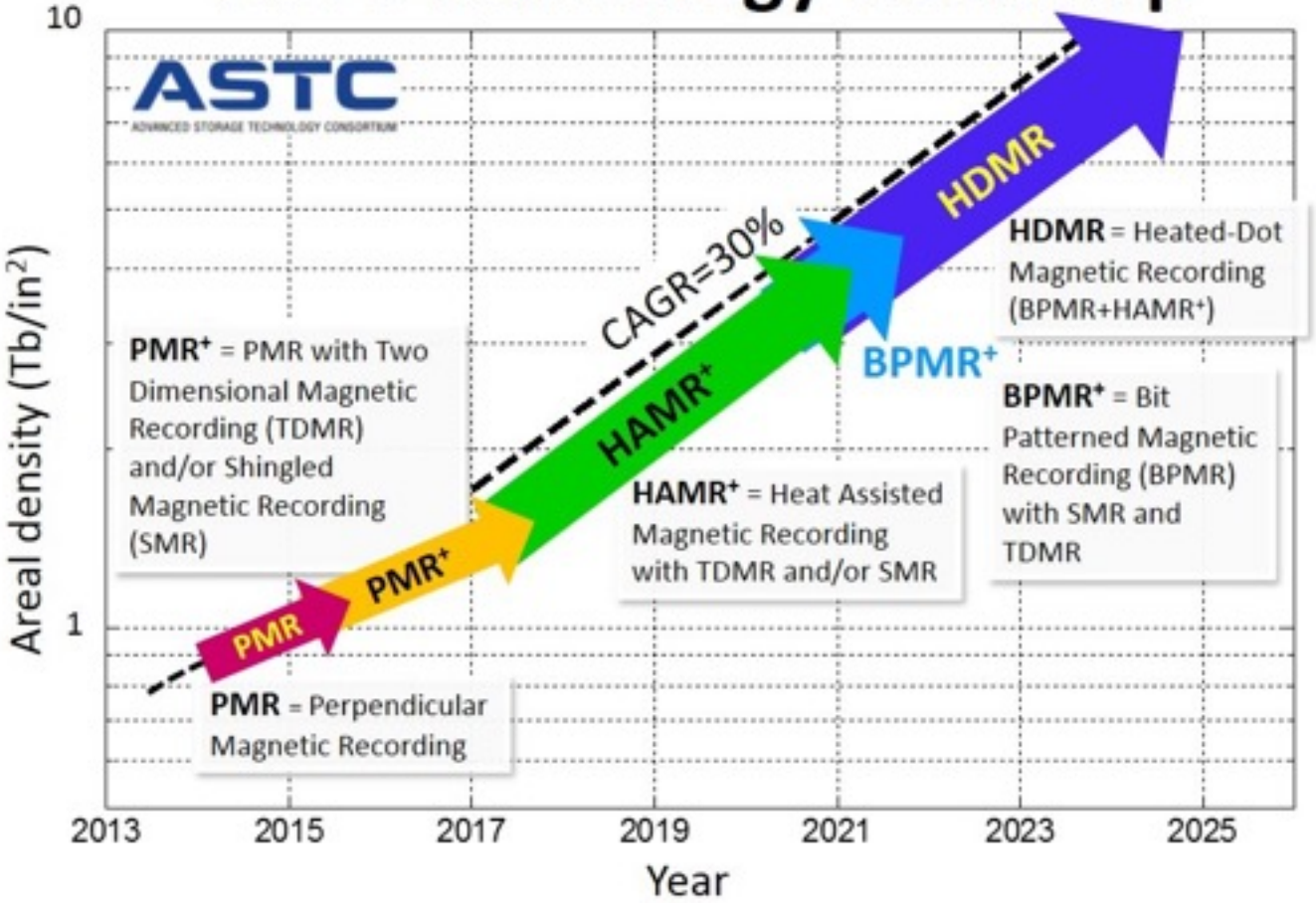
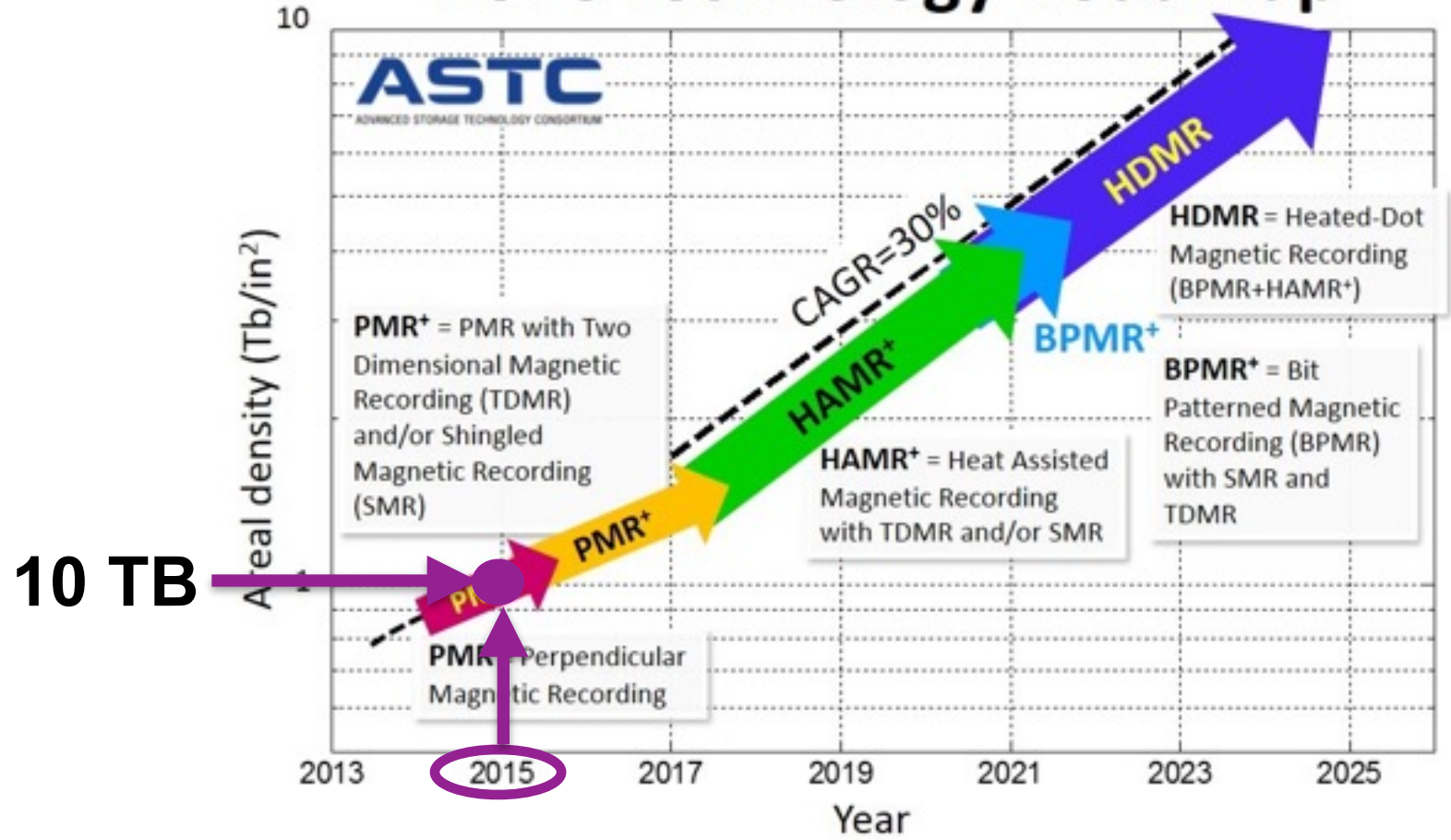
Expanding CERN's EOS Storage with Seagate Kinetic Disk Technology

James Hughes
Paul Lensing

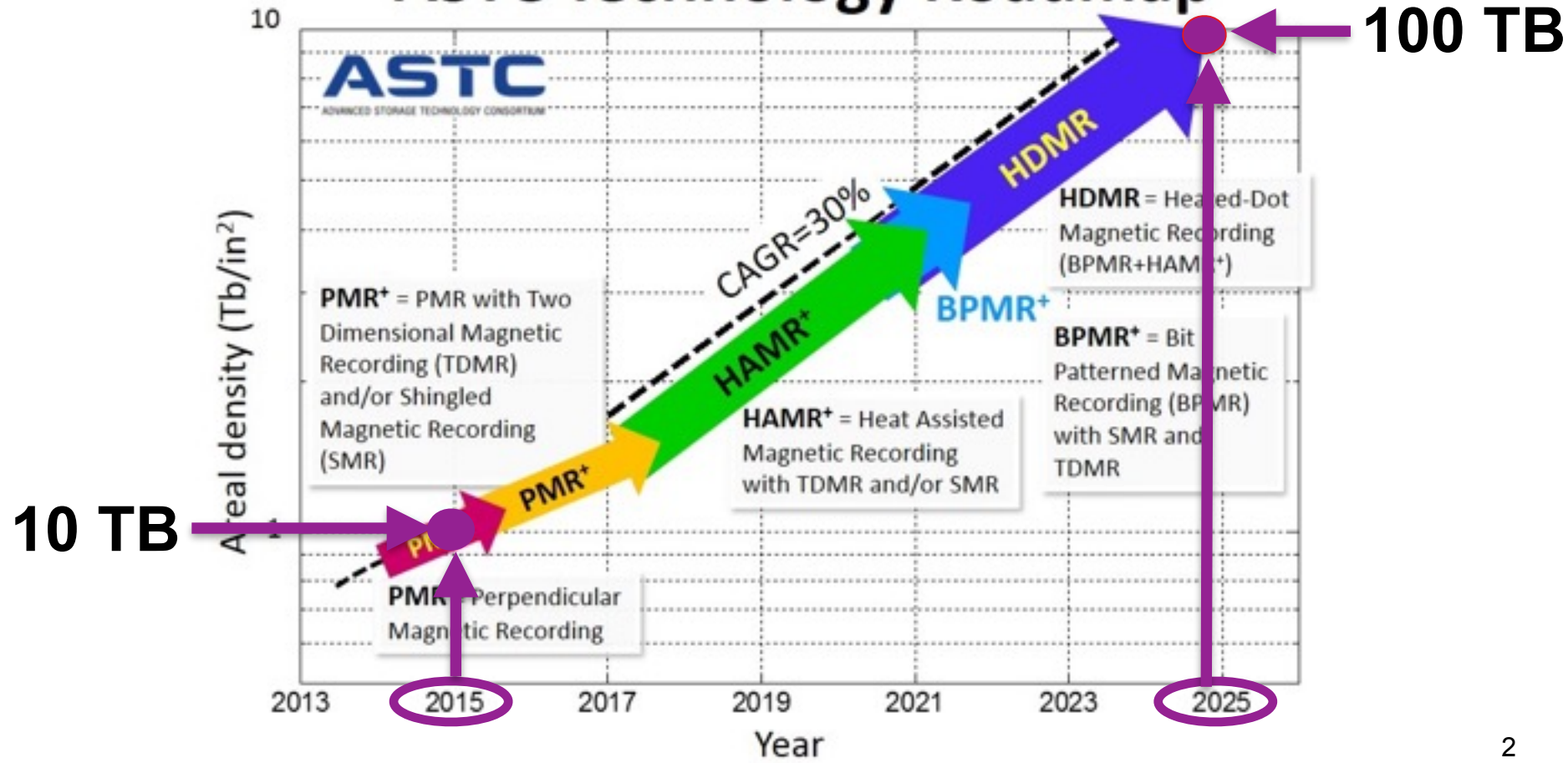
ASTC Technology Roadmap



ASTC Technology Roadmap



ASTC Technology Roadmap



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 bods:

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 a '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

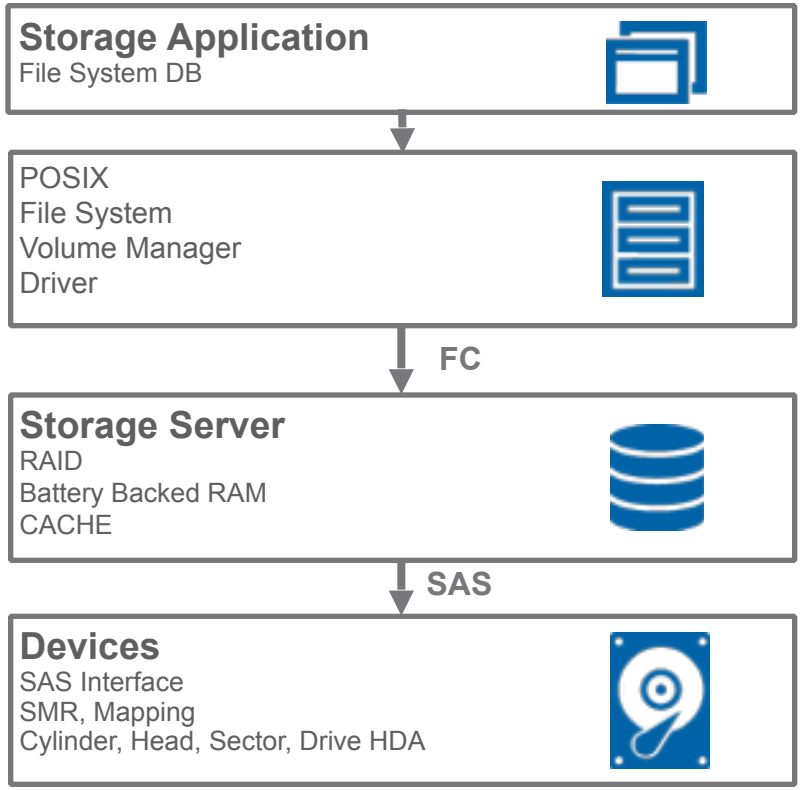
Managing Multi-Vendor Devices with System Centre 2012

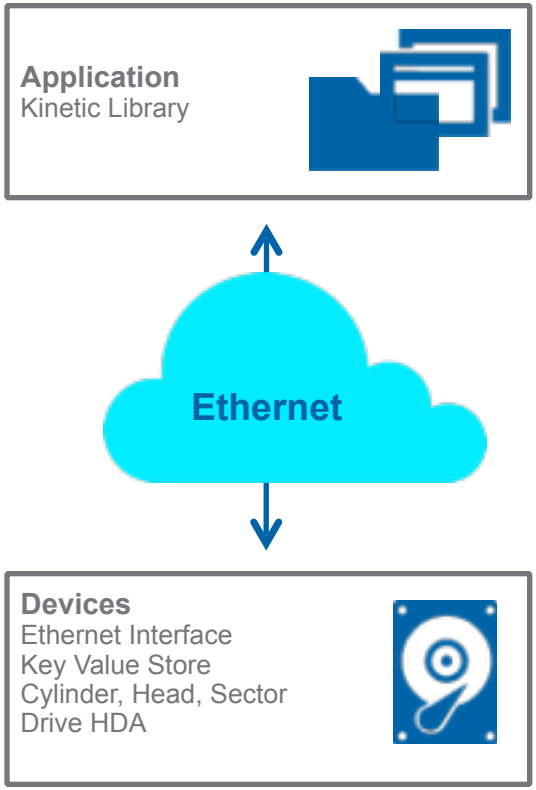
Seagate is building hard disk drives with a direct Ethernet interface and object-style API access for scalable object stores, a plan which - if it works - would destroy much of the existing, typical storage stack.

Drives would become native key/value stores that manage their own space, map

Data Center Stack

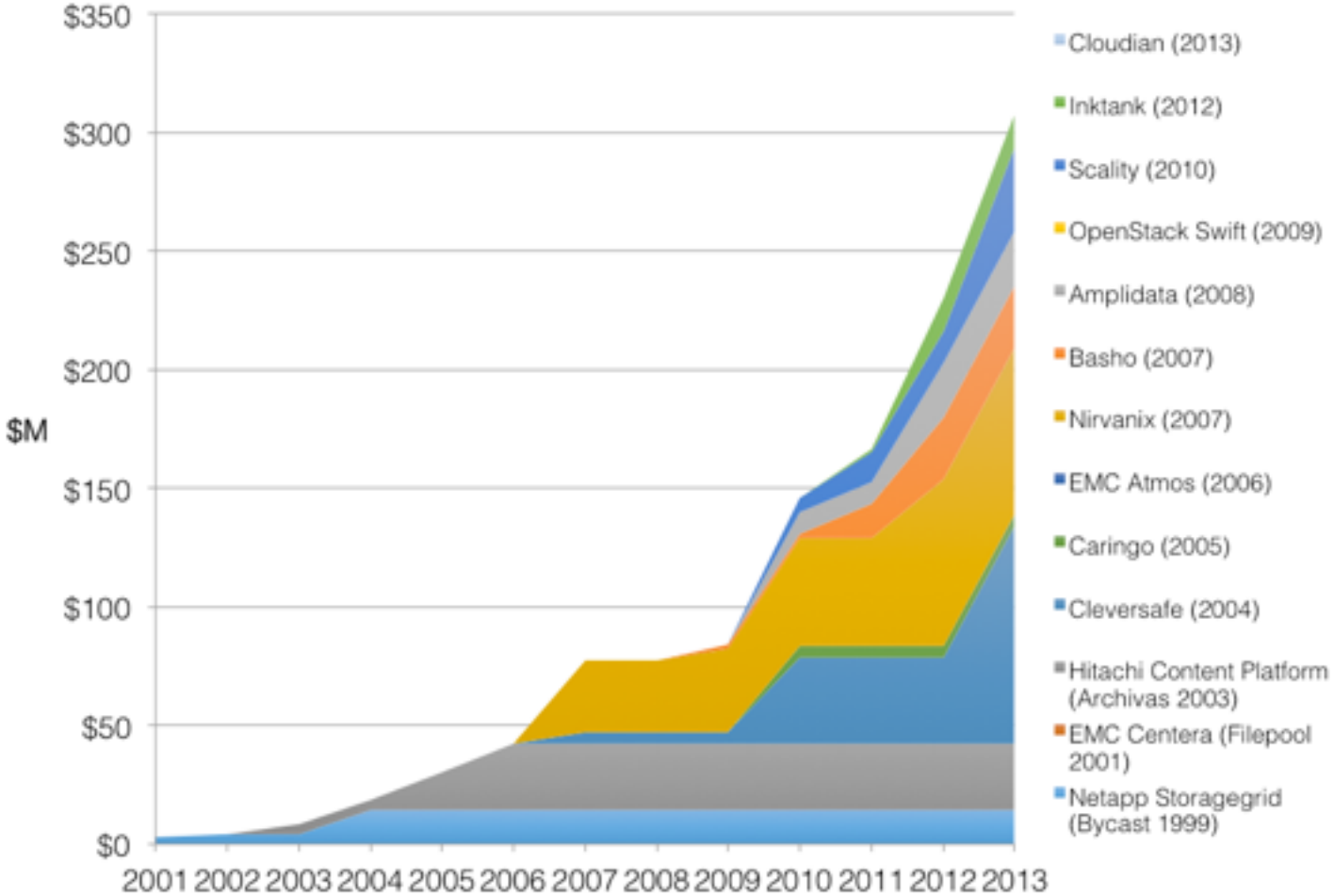
Conventional vs Kinetic





13 Years of Object Storage Funding

https://oxygencloudblog.files.wordpress.com/2013/09/13-years-of-object-storage.png



What is Object Storage

Information that is written, read, deleted, replaced but not “modified in place”

- Amazon S3
- Tape
- EMC Centera
- Powerpoint

Object Storage System

- Name (Key) / Object (Value)
- A storage system for unlimited sized objects

Swift, Scality, etc.

object storage device

- Key/Value store
- Limited value length (1M, defacto)
- Building block

Kinetic

noSQL Market

Flash based Kinetic will clearly focus on the noSQL market

- Latency and KVOP/sec matter

Voldemort

- Linkedin, 100% Flash

Cassandra

- Apple's, >75,000 nodes storing >10 PB of data.
- Netflix, 2,500 nodes, 420TB, over 1 trillion requests per day)
- Chinese search engine Easou, 270 nodes, 300 TB, over 800 million requests per day,
- eBay, >100 nodes, 250 TB.
- Constant Contact, CERN, Comcast, eBay, GitHub, GoDaddy, Hulu, Instagram, Intuit, Netflix, Reddit, The Weather Channel, and over 1500 more companies

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

Seagate Copyright © 2013

Kinetic Architecture

Remote Procedure Call to Key/Value database

- Ordered database (next, previous)

Database is currently LevelDB

- Writes are at the end of the log.
- Deletes are tombstones placed at the end of the log
- Log structured writing xMB Sorted String Tables (SST)
- Garbage collection reads SSTs and writes SSTs
- Same function of FTL

Background operations

- Media Scan
- Garbage collection

Processing in the drive (future)

Kinetic API

Remote Procedure call to Key/Value Database

- Request out / Response back
  - Not REST

Get/Put/Delete

- 4k Key/1M Value

Atomic Operations

- Single record
- Batch

Peer 2 Peer (drive to drive copy)

- No server involved

Background Operations

- Media Scan

Developer Kit

4 drives

- 10 port switch

More resources

- developers.Seagate.com

github.com/Seagate

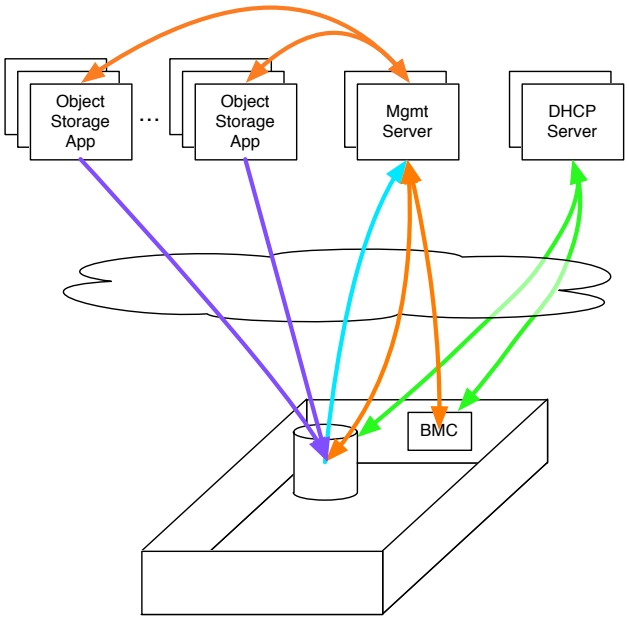
- Libraries
- Simulator
- Protocol
- All open source

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

Bootstrapping devices



Green is good

- 1200w
- $7,800

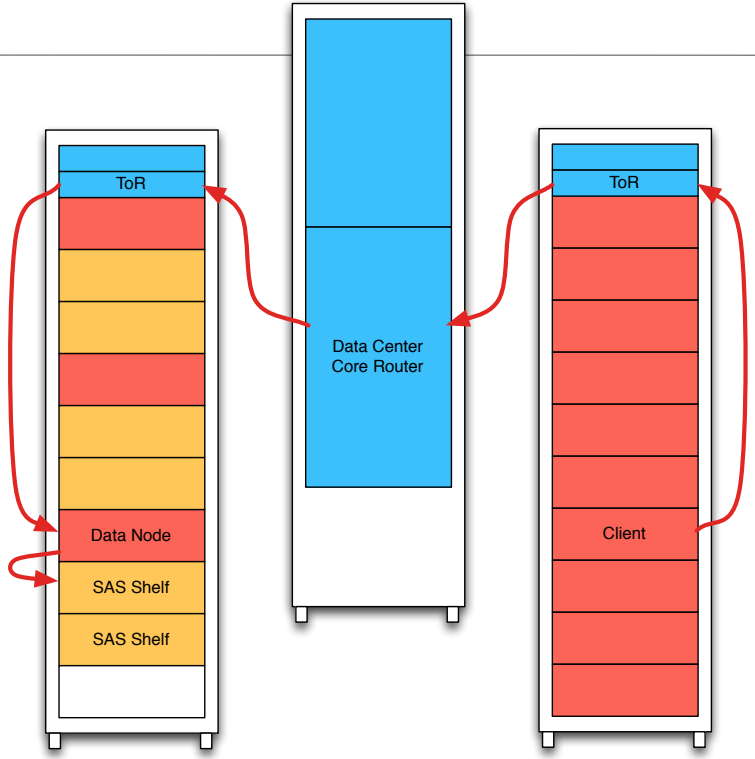
Conventional

- 400w
- $4,200

Kinetic

Seagate Titan Enclosure

Existing Traffic Flow



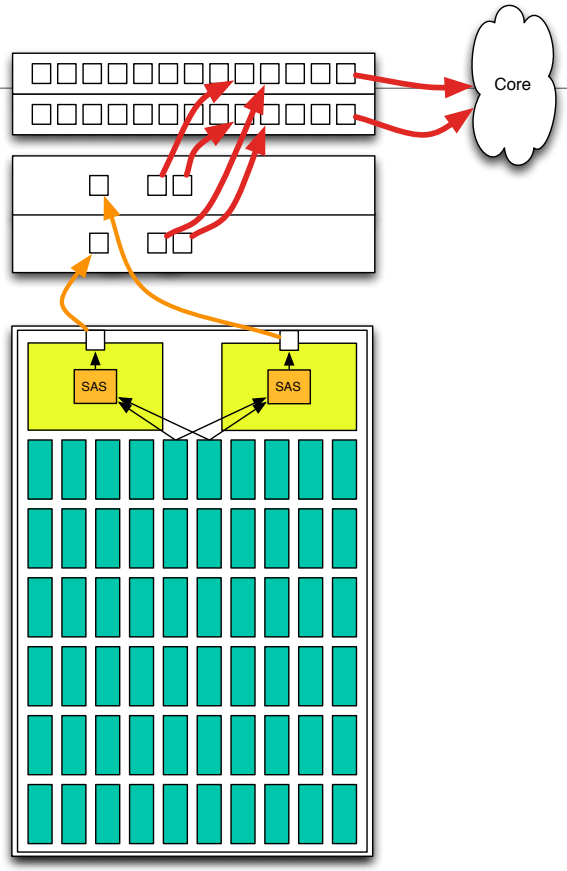
Typical HA High Density

Intel server

- Double Socket
- 48GB Ram
- 1000w

SAS tray

- Connected to the server



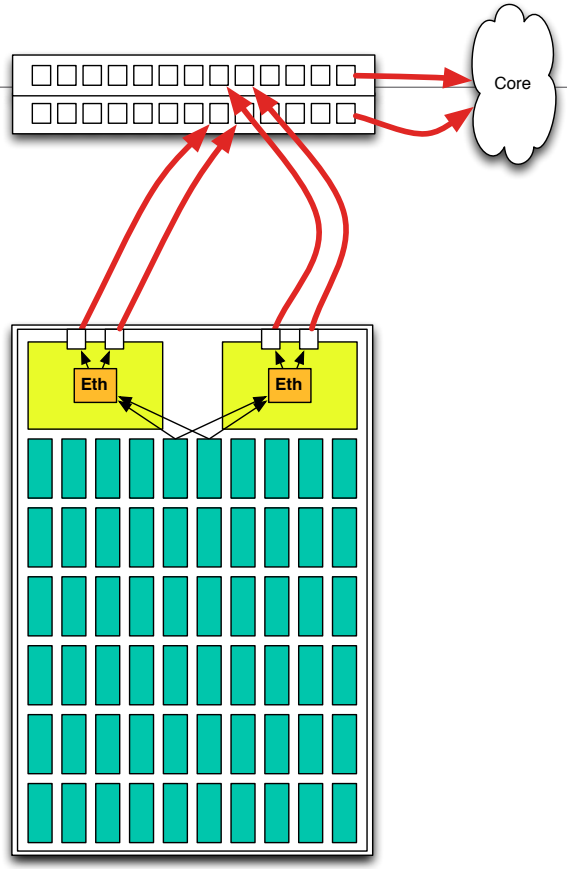
Low cost HA Configuration

Each drive talks to both switches

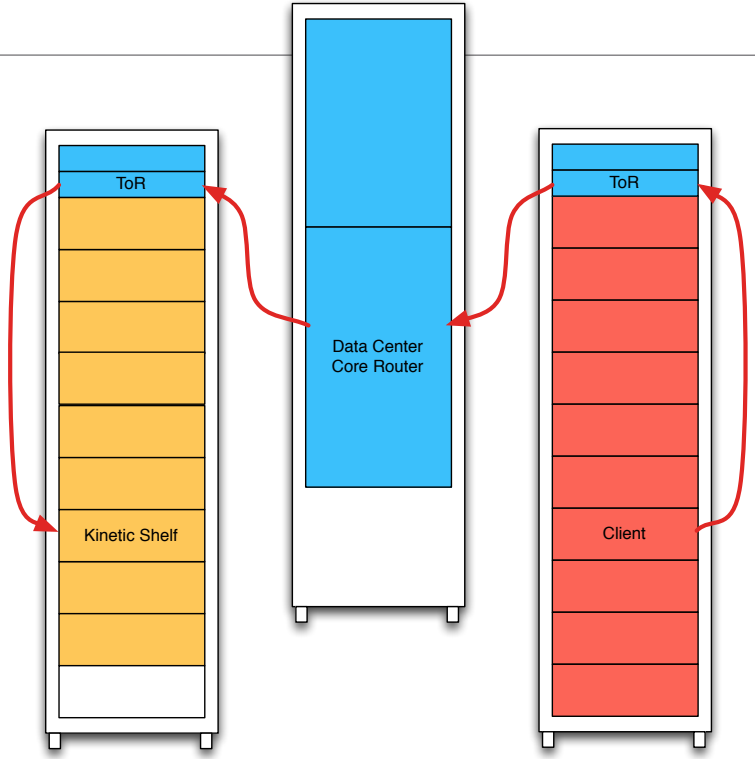
Each switch has 2 by 10Gb/s Ethernet

Kinetic Tray talks directly to ToR

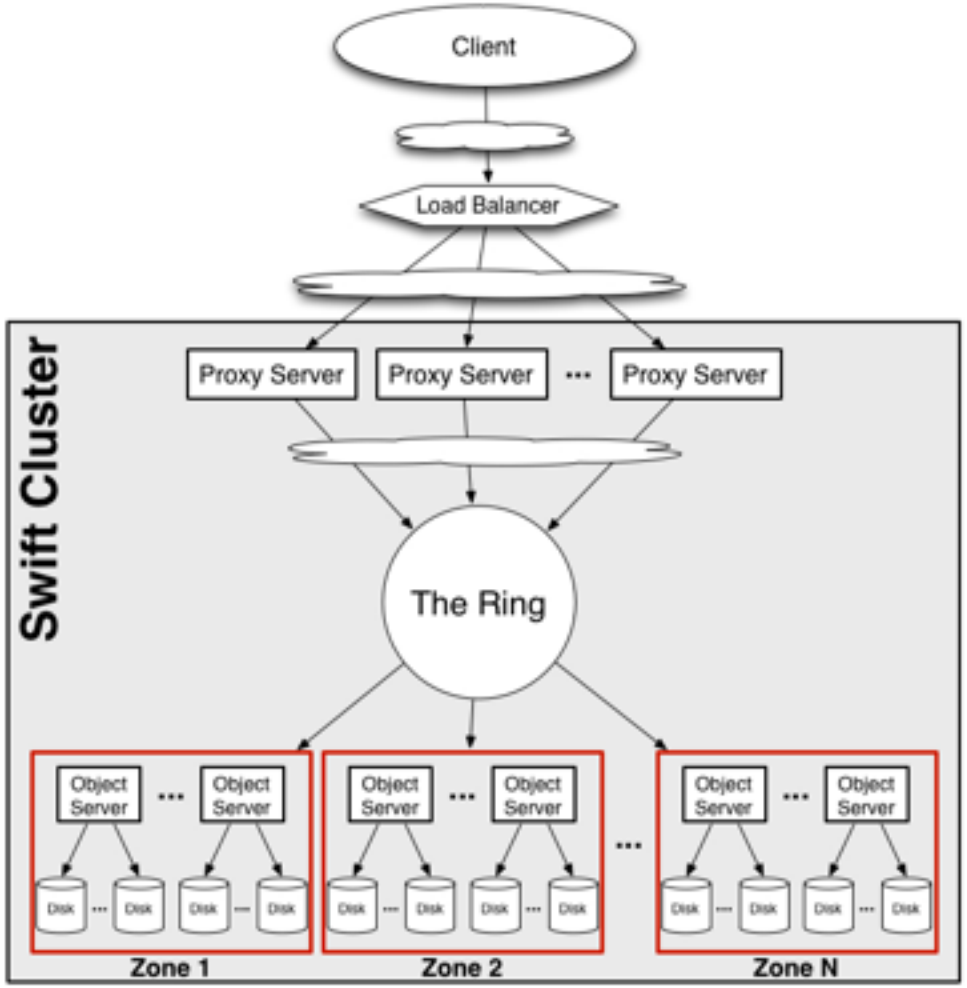
No servers



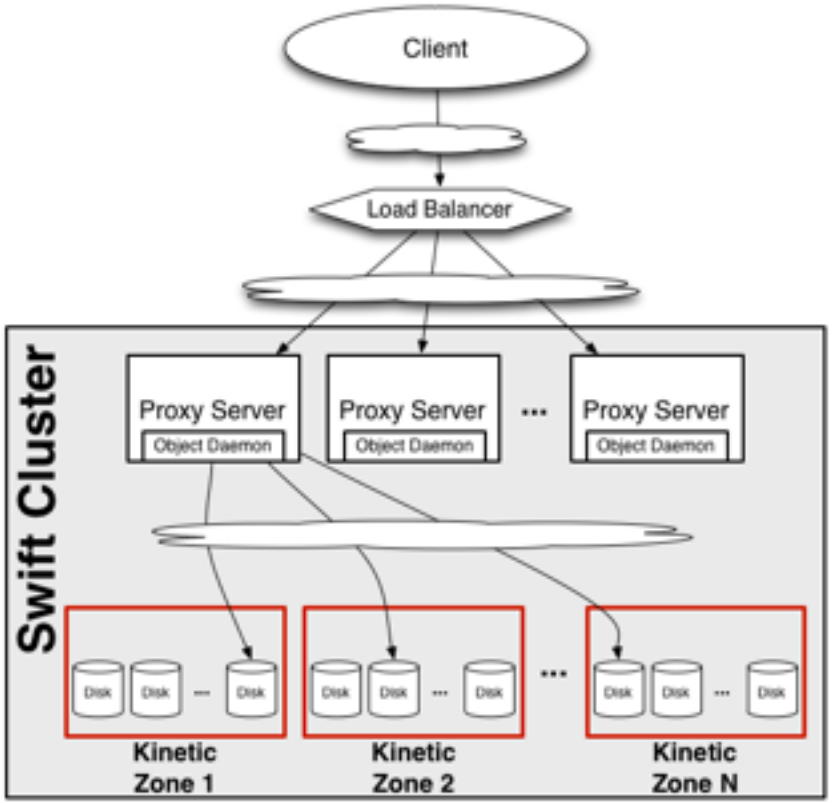
Kinetic Traffic Flow



Swift


Client
Load Balancer
Swift Cluster
Proxy Server
Proxy Server
Proxy Server
The Ring
Object Server
Object Server
Object Server
Object Server
Object Server
Object Server
Disk
Zone 1
Zone 2
Zone N

Swift


Client
Load Balancer
Swift Cluster
Proxy Server
Object Daemon
Proxy Server
Object Daemon
...
Proxy Server
Object Daemon
Disk
Disk
...
Disk
Disk
Disk
...
Disk
...
Disk
Disk
...
Disk
Kinetic
Zone 1
Kinetic
Zone 2
Kinetic
Zone N

Conventional EOS

Gateways

Clients

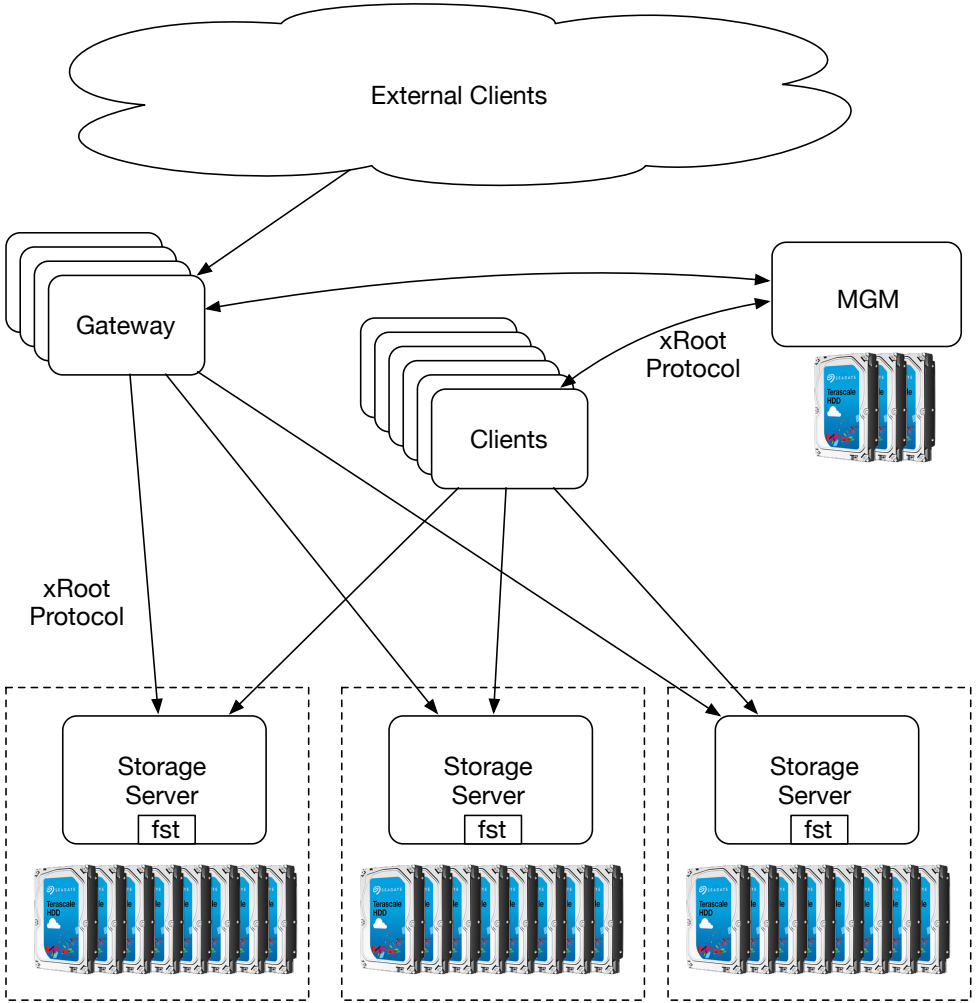
- External Clients Through Gateway
- Internal Clients Connecting Directly

Metadata Server

- MGM (EOS Management Module)
- Namespace, Pool Configuration, Placement & Access Scheduler

Storage Servers

- FST (EOS File Storage Module)
- Expose Local Disks to EOS



EOS with Kinetic

Kinetic Drives

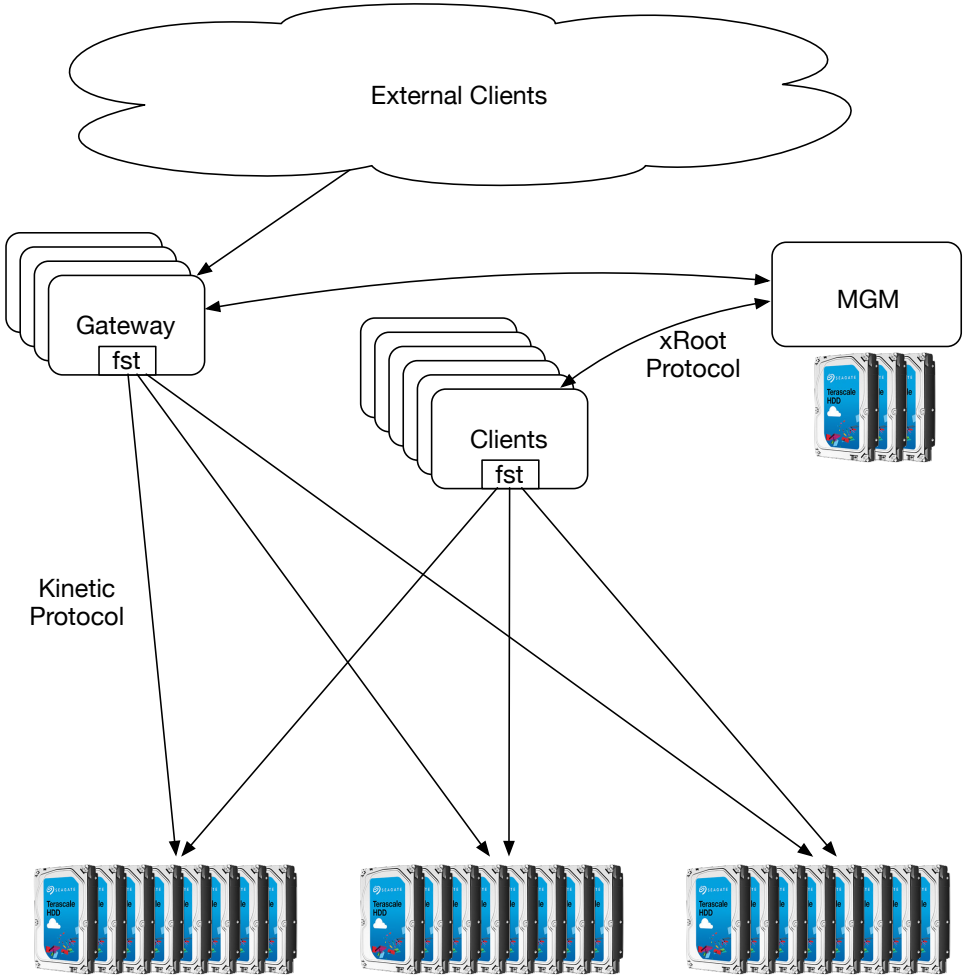
- Ethernet Storage
- Not Locked to Single Server
- Accessible from Multiple FST

Internal EOS Clients

- Run FST Module for Direct Access

External EOS Clients

- Continue Using Gateway Servers



1PB test system

Twenty One Chassis

- 1u12 from Supermicro
  - 2x10Gbe

252 Kinetic drives

- 504 IP addresses
- 126GBytes of RAM cache

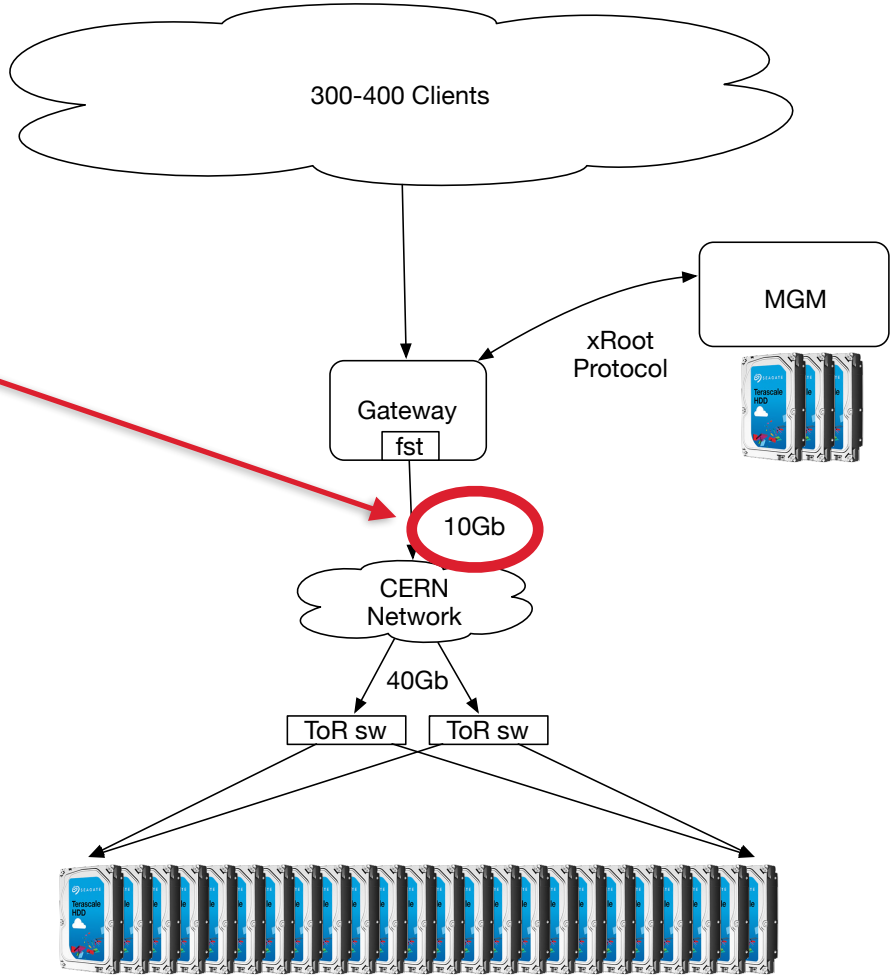
Two ToR Switches

- 40Gbe to CERN

Initial Test Setup

Single Gateway Test

- 10Gb/s Over a 10Gb/s Network



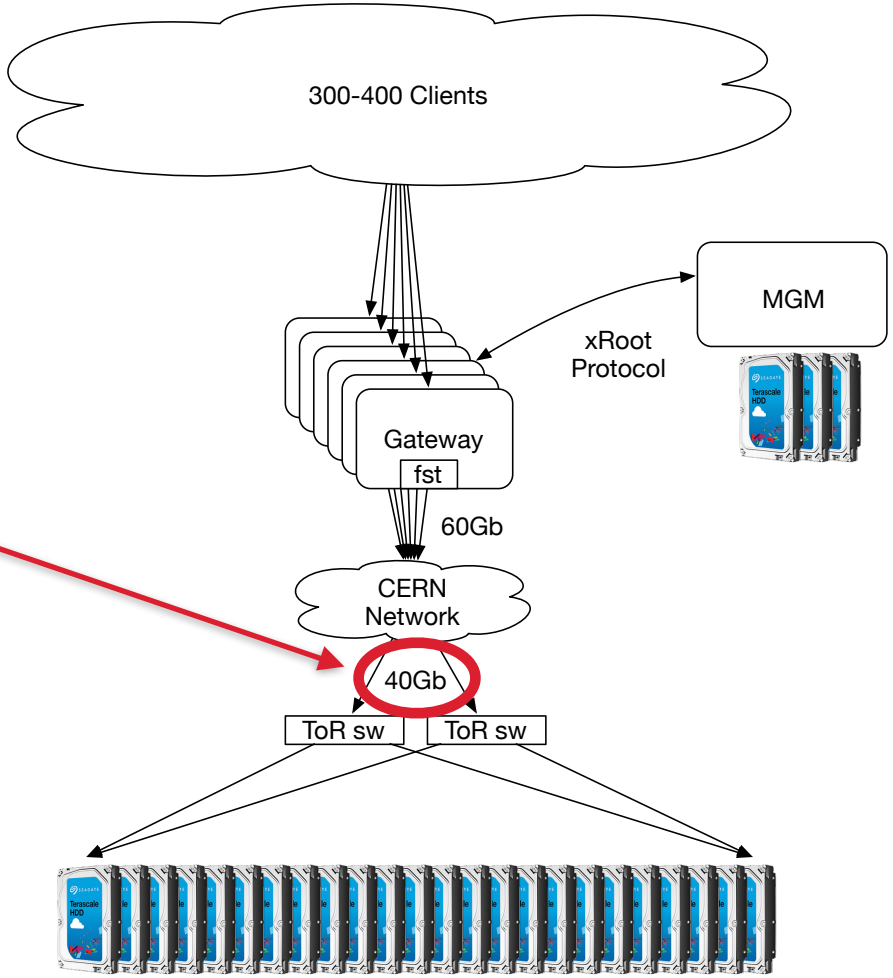
Initial Test Setup

Single Gateway Test

- 10Gb/s Over a 10Gb/s Network

Six Gateway Test

- 32Gb/s Over a 40Gb/s Network



Initial Test Setup

Single Gateway Test

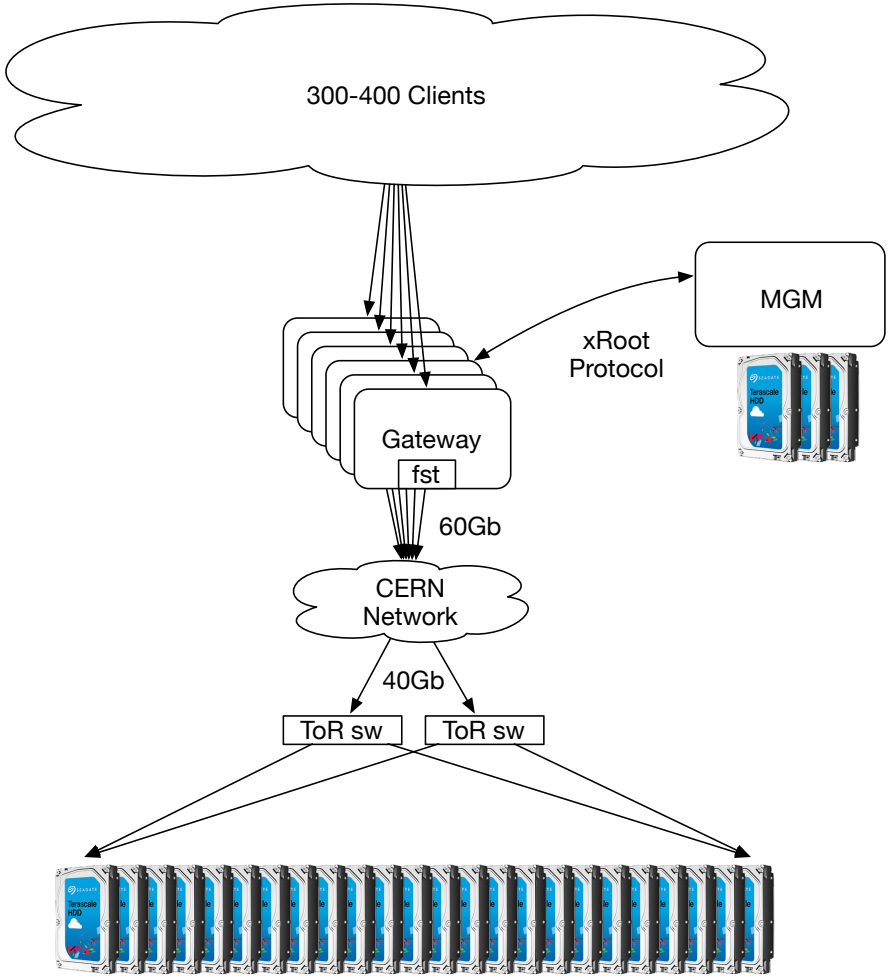
- 10Gb/s Over a 10Gb/s Network

Six Gateway Test

- 32Gb/s Over a 40Gb/s Network

More work to do

- Deploy Kinetic/FST on Datacenter Clients
- Add Cluster Logic and Erasure Coding
- Begin Testing With Real Data
- Add Second Cluster in 2016
  - Next Generation Drives



Conclusion

On The Path to Prove TCO Gains of Kinetic at CERN

- No change to Existing EoS User Applications
- Looking Forward to Storing LHC Data

Kinetic Security

Unencrypted TCP port

- Authentication
- Integrity
- Roles
- Read, write, etc.

TLS port

- Same security as unencrypted port but protected by Seagate

More information

More Information about Kinetic

- https://developers.seagate.com

Kinetic Libraries and Simulator

- http://github.com/Seagate

Kinetic Support

- http://github.com/SwiftStack/kinetic-swift

Optimizations reducing overhead

- github.com/Seagate/swift

For more information contact

- Ignacio Corderi <ignacio.corderi@seagate.com>
- James Hughes <james.hughes@seagate.com>

SEAGATE

Thank You

HTTP://Developers.Seagate.Com