SNIA CSI | CLOUD STORAGE

IP-Based Object Drives Now Have a Management Standard

Live Webcast

April 20, 2017
10:00 am PT

Today's Presenters

Enrico Signoretti
OpenIO

David Slik
Co-Chair, SNIA Cloud TWG
NetApp

Erik Riedel
Dell EMC

Alex McDonald
Co-Chair SNIA CSI, NetApp

SNIA at a glance

SNIA CSI | CLOUD STORAGE

160
unique member companies

3,500
active contributing members

50,000
IT end users & storage pros worldwide

Learn more: snia.org/technical

@SNIA

SNIA Legal Notice

- The material contained in this presentation is copyrighted by the SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.
- NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

Agenda

- Object Drive Overview, David Slik
- Some Products & Observations, Enrico Signoretti
- Experiments & Experiences, Erik Riedel

SNIA CSI | CLOUD STORAGE

David Slik – NetApp

What are IP Based Drives?

- Interface changed from SCSI based to IP based (TCP/IP, HTTP)
- Channel (FC/SAS/SATA) interconnect moves to Ethernet network
- Want to get involved? Join the SNIA Object Drive TWG at: https://members.snia.org/apps/org/workgroup/objecttwg/

What is driving the market?

- A number of scale out storage solutions expand by adding identical storage nodes incrementally
  - Typically use an Ethernet interface and may be connected directly to the Internet
- Open source examples include:
  - Scale out file systems
    - Hadoop's HDFS
    - Lustre
  - Ceph
  - Swift (OpenStack object storage)
- Commercial examples also exist

Who would buy IP Based Drives?

- System vendors and integrators
  - Enables simplification of the software stack
- Hyperscale Data Centers
  - Using commodity hardware and open source software
- Enterprise IT
  - Following the Hyperscale folks

Traditional Block Storage Device

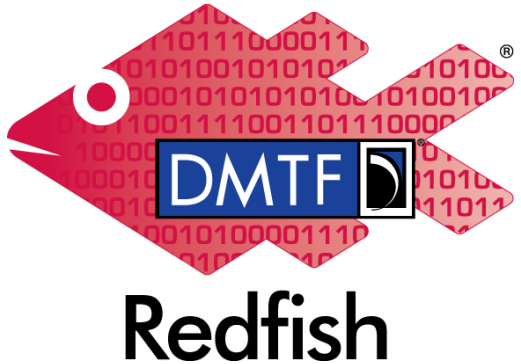
- Interconnect via SATA/SAS/PCIe
  - Locally addressable
  - connected directly or via local fabric to storage servers
- Traditional block protocols (SATA/SAS/NVMe)
  - Block addressable
  - designed for reliable transport
  - Long lived local communication
  - Coordinated concurrent accesses

IP Based Drive

- Interconnect via Ethernet
  - Globally addressable
  - provide storage services over network protocols directly to clients
- IP Based Protocol
  - Higher-level storage services
  - Error tolerant protocol (e.g., TCP/IP)
  - Transitory global communication
  - Independent concurrent accesses

IP-Based Drive Management

- The Object Drive TWG produced a specification for scalable management of IP Based Drives
- Based on the RedFish management specification from DMTF
  - https://www.dmtf.org/sites/default/files/standards/documents/DSP0266_1.0.1.pdf
- Uses Odata (OASIS) for RESTful interface
- Minimizes and simplifies the management of resources
- SNIA standard specifies common features and references the other standards



Management Activities for Object Drives

- As a device, how do I connect to a network?
  - How do devices physically negotiate to connect to a network?
  - How do devices configure themselves to talk TCP/IP over a network?
  - Address assignment, name resolution, time services, etc.
  - How do devices discover where they physically are located?

Management Activities for Object Drives

- As a manager, how do I discover devices?
  - & discover what devices are available to manage?
  - & distinguish my devices from other peoples devices?
- How do I find out
  - where these devices are located?
  - how these devices are connected?
  - if any of this changes?

Management Activities for Object Drives

- As a manager, how do I configure devices?
  - CPU Firmware?
  - drive firmware?
  - the network?
  - … & select the drive application?

Management Activities for Object Drives

- As a manager, how do I keep devices secure?
  - find out when security updates are available?
  - push security updates to devices?
  - maintain operations during updates?
- How do I keep my devices up to date?
  - tell when firmware updates are available?
  - push firmware updates to my devices?
  - maintain operations during updates?

Management Activities for Object Drives

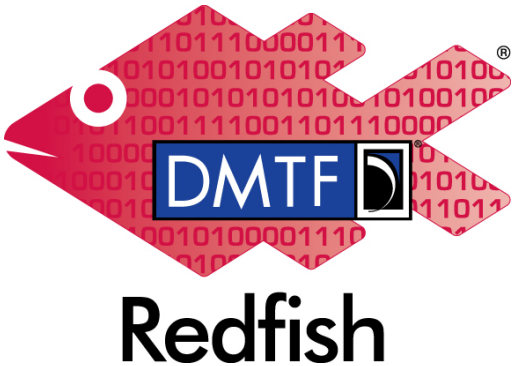
- How do I know when things are failing?
- How do I
  - monitor environmental health?
  - monitor drive hardware health?
  - monitor drive data health?
  - monitor drive application health?
  - tell when something is unhealthy?
  - tell when something is about to fail?
  - tell when something has failed?
  - identify what needs to be done to replace a failed device?

IP Based Drive Management

- Specification is now a SNIA Technical Position (Standard)
  - IP-Based Drive Management Specification v1.0
- **IP-Based Drive Characteristics and Requirements**
  - Describes the physical form factors, electrical and link layer requirements
  - Has a Taxonomy of various possible drive types with protocol and other information
- **IP-Based Drive Management**
  - Describes the device discovery and management
  - Assignment of IP address
  - Discovery of Basic Services
  - Redfish based management

Using Redfish to manage drives

- The following services are used (but not limited to these):
  - Account Service
  - Session Service
  - Chassis Collection
  - Manager Collection
  - Computer System Collection
  - Update Service (recommended)
- "ChassisType" property is "IPBasedDrive"
- The Redfish implementation should support the Redfish standard Drive entity.



Three mockups available

- What are the mockups?
  - Examples of management interfaces to an IP Based drive system
  - Redfish schemas
  - Located at http://www.snia.org/object-drives
- What mockups are available
  - Simple IP Based drive mockup
    - Single drive
    - Dual network connections
  - IP Based drive array mockup
    - Single manager
    - Multiple drives arranged hierarchically

SNIA CSI | CLOUD STORAGE

Enrico Signoretti

Nano-node

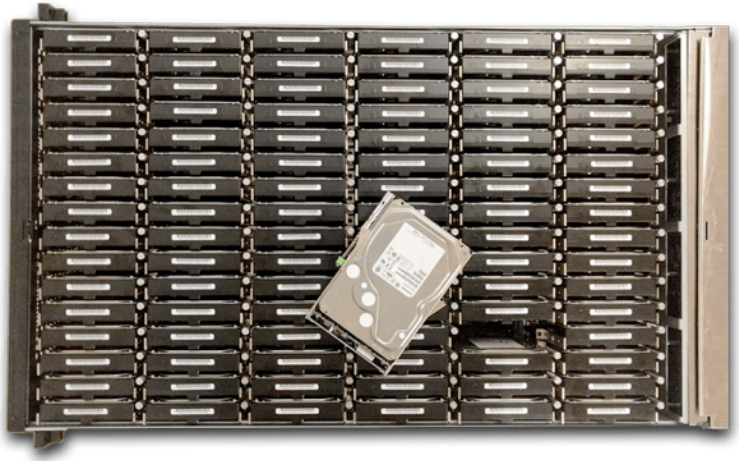
Hyper Scalable Storage

- Dual-core ARM-v8 CPU
- RAM, flash memory, 2 * 2.5gb/s Ethernet links
- 3W power consumption and HDDs Power management
- Supports 8,10,12 TB HDDs

SLS 4U96

No Single Point of Failure

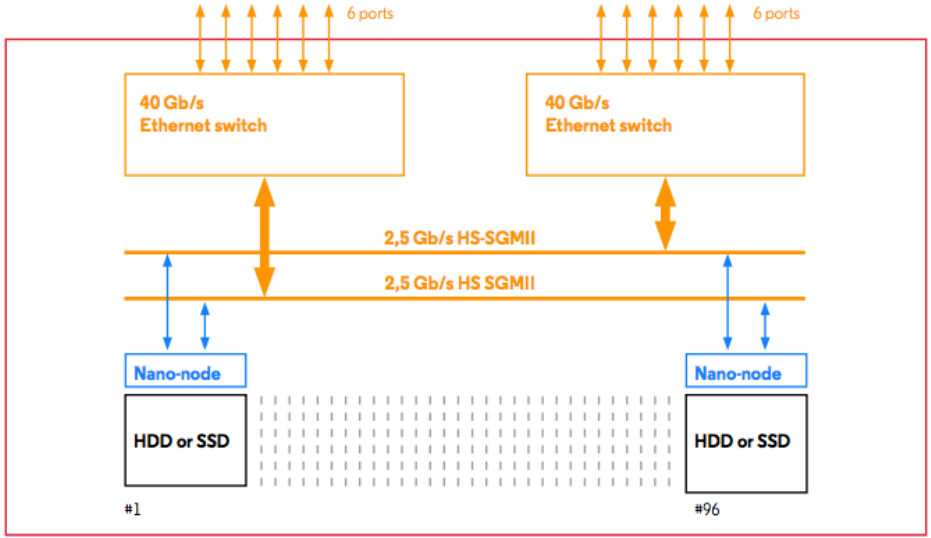
- N+1 power supplies and cooling units
- Chassis Management
- 2x 6-port 40gb/s Ethernet switches for front-end and back-to-back expansion
- Up to 96 hot-swap nano-nodes



SLS 4U96

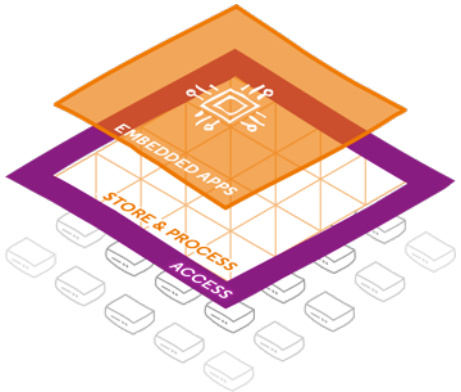
No Single Point of Failure

- N+1 power supplies and cooling units
- Chassis Management
- 2x 6-port 40gb/s Ethernet switches
  for front-end and back-to-back expansion
- Up to 96 hot-swap nano-nodes



Same software, same capabilities

- Standard Object APIs to leverage natively the platform: OpenIO REST/HTTP, Amazon S3 and OpenStack Swift
- Industry File-Sharing Protocols: NFS, SMB, AFP and FTP
- Several data protection schemes and cluster topologies
- Ease of Use. GUI, APIs, CLI
- Lightweight backend design
- Grid for Apps: event-driven framework for Serverless computing



SNIA CSI | CLOUD STORAGE

Erik Riedel

SCALE

Scale

		srvs	encs	disks	TB drv	TB rack
Atmos	gen1	16	16	240	1TB	240TB
	gen2	6	24	360	2TB	720TB
	gen3	8	8	480	4TB	1.9PB
ECS	gen4	8	8	480	6TB	2.9PB
	gen5	8	8	480	8TB	3.8PB
	gen5+	8	8	784	10TB	7.8PB
	gen6	4	8	784	12TB	9.4PB

Gen1 (2008)
1TB

Gen2 (2010)
2,3TB

Gen3 (2012)
3,4,6TB

Gen4 (2014)
6TB

Gen5 (2015)
8TB

Gen5+ (2016)
10TB

Gen6 (2017)
12TB

Mechanicals

back-to-back

drawers

drawers

sleds

modular

trays

modular

Density

Updated from "Long-Term Storage", presented at Library of Congress Workshop in September 2012

SNIA CSI | CLOUD STORAGE

2012	Disks (raw) @ 3TB	Disks (protected)	Racks @ 480 disks
5 PB	1,700 disks	2,700 disks	6 racks
20 PB	6,700 disks	11,000 disks	23 racks
50 PB	17,000 disks	27,000 disks	56 racks

2014	Disks (raw) @ 6TB	Disks (protected)	Racks @ 480 disks
5 PB	830 disks	1,300 disks	3 racks
20 PB	3,300 disks	5,300 disks	12 racks
50 PB	8,300 disks	13,000 disks	28 racks

2016	Disks (raw) @ 10TB	Disks (protected)	Racks @ 780 disks
5 PB	500 disks	670 disks	1 rack
20 PB	2,000 disks	2,700 disks	4 racks
50 PB	5,000 disks	6,700 disks	9 racks

Scale Out

3.8PB/8n480d

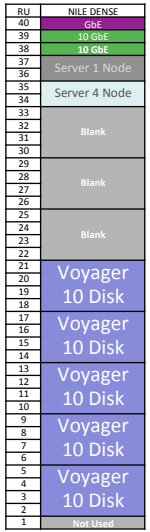
RU	NILE DENSE
40	GbE
39	10 GbE
38	10 GbE
37	Servers 4 Node
36	
35	Servers 4 Node
34	
33	Voyager 60 Disk
32	
31	
30	
29	Voyager 60 Disk
28	
27	
26	
25	Voyager 60 Disk
24	
23	
22	
21	Voyager 60 Disk
20	
19	
18	
17	Voyager 60 Disk
16	
15	
14	
13	Voyager 60 Disk
12	
11	
10	
9	Voyager 60 Disk
8	
7	
6	
5	Voyager 60 Disk
4	
3	
2	
1	Not Used

23 PB petabytes
48 nodes
2,880 disks

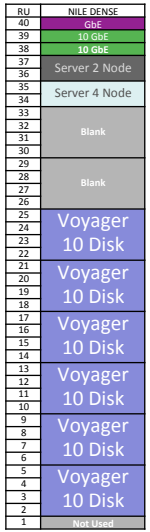
FLEXIBILITY

Experiments in Flexibility (Goal)

320 TB
4 node
40 disks

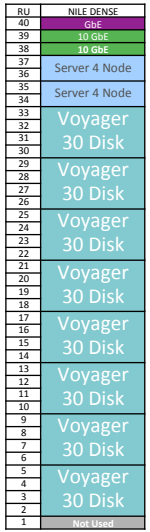


400 TB
5 nodes
50 disks



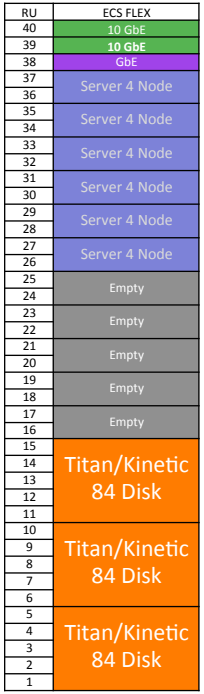
480 TB
6 nodes
60 disks

640 TB
8 nodes
80 disks

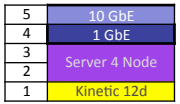


1.9 PB
8 nodes
240 disks

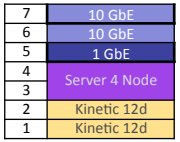
3.8PB
8 nodes
480 disks



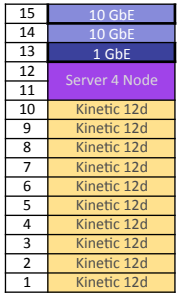
2.0 PB
24 nodes
252 disks



96 TB
4 nodes
12 disks



192 TB
4 nodes
24 disks



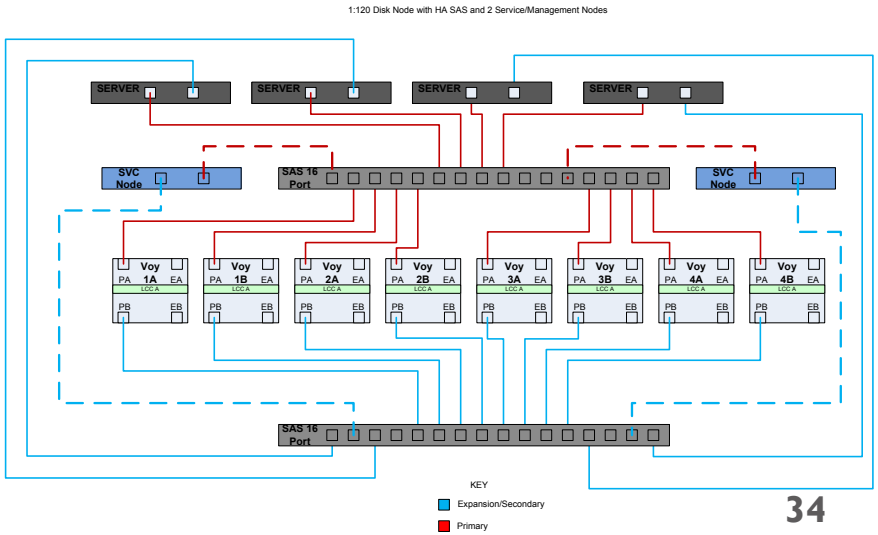
960 TB
4 nodes
120 disks

Experiment – SAS Switching (2012)

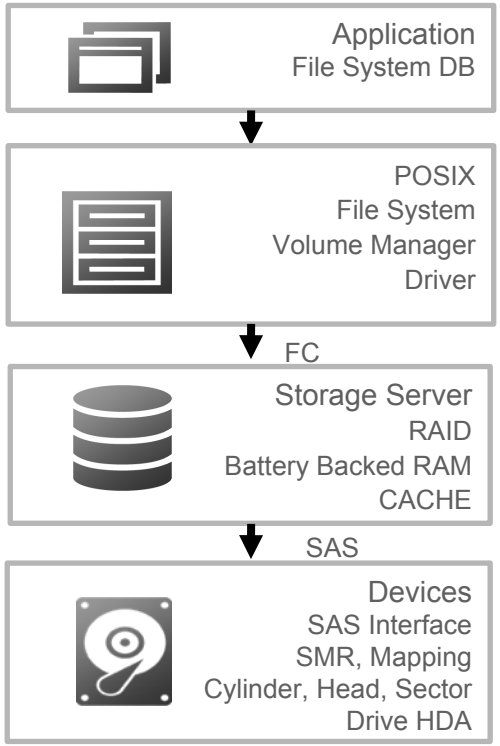
40	turtle	
39	rabbit	
38	left	right
37	lehi	murray
36	layton	logan
35	orem	ogden
34	provo	sandy
33	eight	
31		
29	seven	
27		
25	six	
23		
21	five	
19		
17	four	
15		
13	three	
11		
09	two	
07		
05	one	
03		
01	EMPTY	

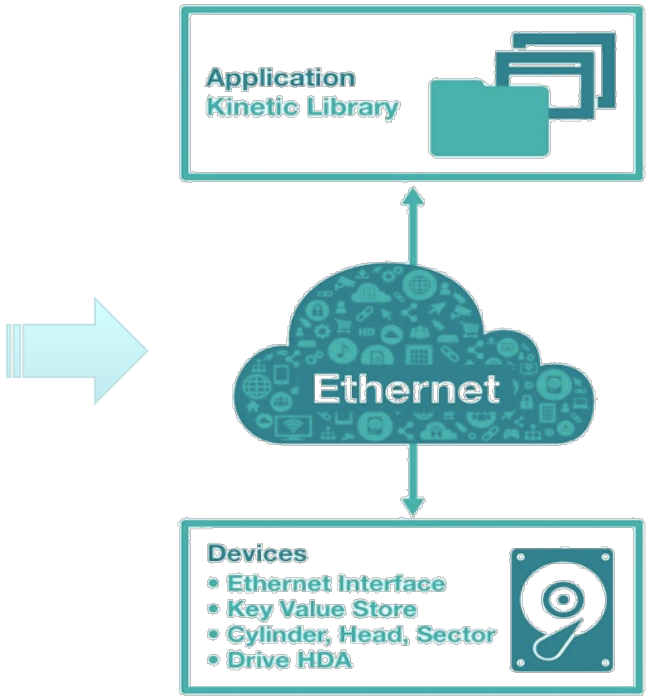
left

right



Experiment – Kinetic (2014)





from Seagate material "Kinetic Open Storage - Enabling Break-through Economics in Scale-out Object Storage"

Experiment – Kinetic (2014)

- Newisys EDA-4605 Enclosure
  - 60-disk
  - dual 10 GbE controllers
  - 4x 10 GbE uplinks

- Seagate Kinetic Ethernet drive
  - 4TB in October 2014 (2x 1 GbE network)
  - 8TB in September 2015 (2x 2.5 GbE network)

Experiment – Kinetic 2nd Generation (2017)

12 nodes 504 disks
240 cores
4,032 TB raw
160 Gbps

RU	ECS FLEX
40	10 GbE
39	10 GbE
38	GbE
37	Rinjin 4 Blade
36	
35	Rinjin 4 Blade
34	
33	Rinjin 4 Blade
32	
31	
30	Titan/Kinetic 84 Disk
29	
28	
27	
26	
25	Titan/Kinetic 84 Disk
24	
23	
22	
21	
20	Titan/Kinetic 84 Disk
19	
18	
17	
16	
15	Titan/Kinetic 84 Disk
14	
13	
12	
11	
10	Titan/Kinetic 84 Disk
9	
8	
7	
6	
5	Titan/Kinetic 84 Disk
4	
3	
2	
1	

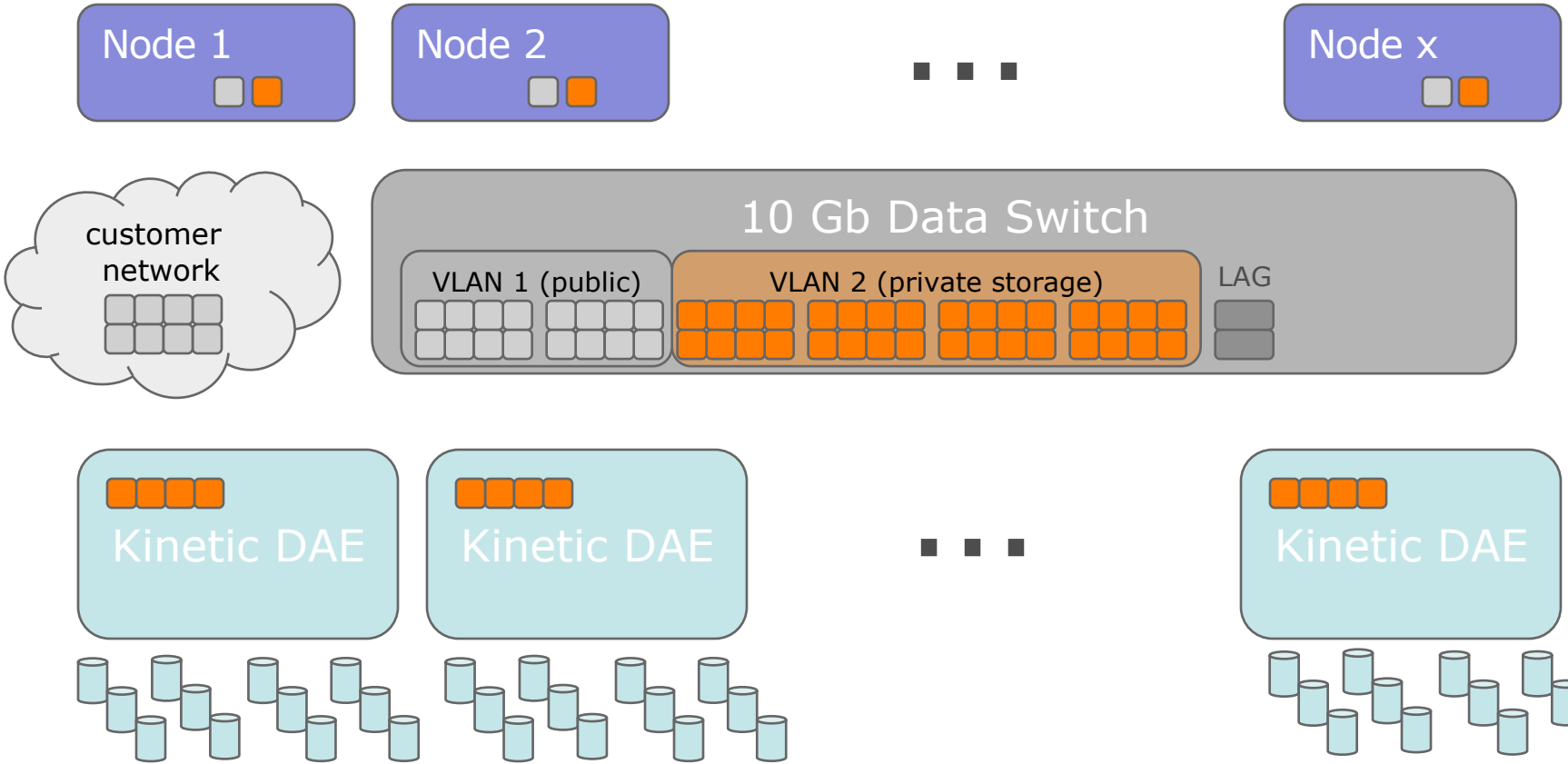
RU	ECS FLEX
40	10 GbE
39	10 GbE
38	GbE
37	Rinjin 4 Blade
36	
35	Rinjin 4 Blade
34	
33	Rinjin 4 Blade
32	
31	
30	Titan/Kinetic 14 Disk
29	
28	
27	
26	
25	Titan/Kinetic 14 Disk
24	
23	
22	
21	
20	Titan/Kinetic 14 Disk
19	
18	
17	
16	
15	Titan/Kinetic 14 Disk
14	
13	
12	
11	
10	Titan/Kinetic 14 Disk
9	
8	
7	
6	
5	Titan/Kinetic 14 Disk
4	
3	
2	
1	

RU	ECS FLEX
40	10 GbE
39	10 GbE
38	GbE
37	Rinjin 4 Blade
36	
35	Rinjin 4 Blade
34	
33	Rinjin 4 Blade
32	
31	
30	Titan/Kinetic 14 Disk
29	
28	
27	
26	
25	Titan/Kinetic 14 Disk
24	
23	
22	
21	
20	Titan/Kinetic 14 Disk
19	
18	
17	
16	
15	Titan/Kinetic 14 Disk
14	
13	
12	
11	
10	Titan/Kinetic 14 Disk
9	
8	
7	
6	
5	Titan/Kinetic 14 Disk
4	
3	
2	
1	

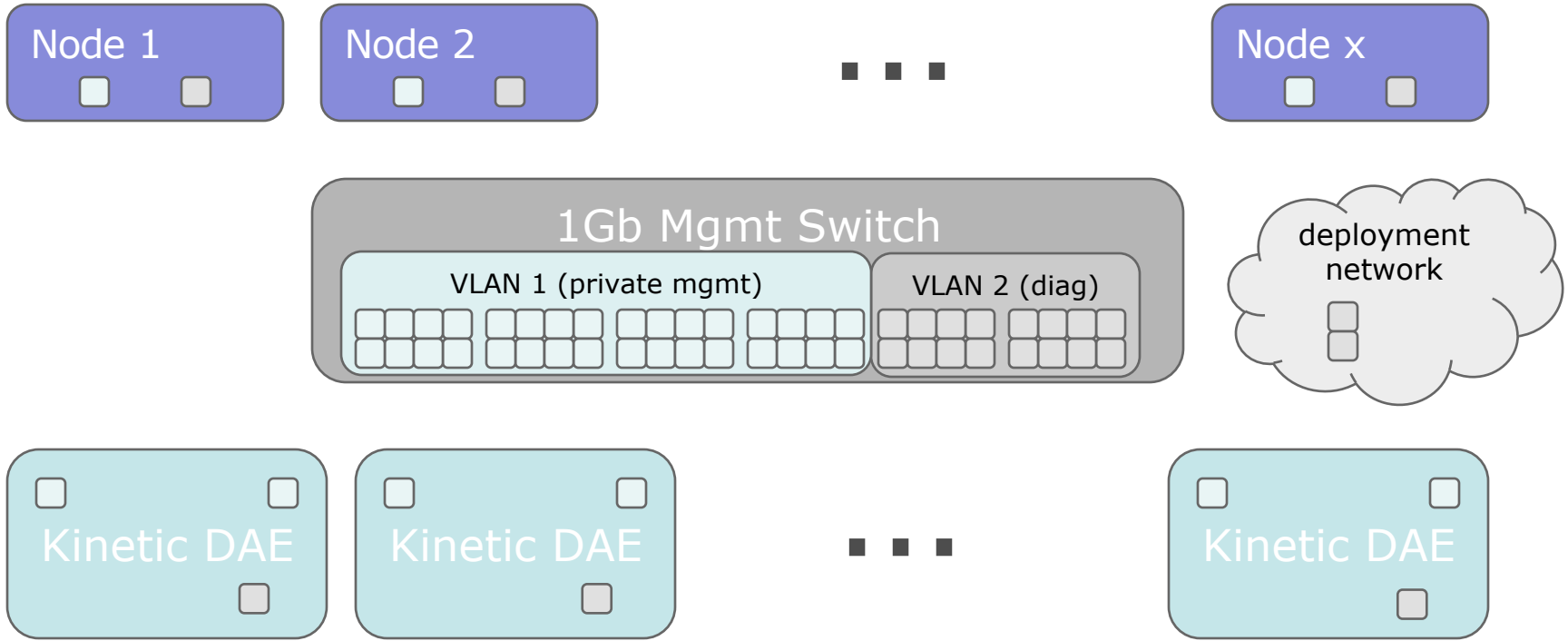
Parts List

- 9x Rinjin servers (36 nodes)
- 2x 10 GbE per node SFP+
- 6x 10 GbE data switches (Arista 64-port SFP+)
- 3x 1 GbE mgmt switches (Arista 48-port Cat6)
- 18x Titan enclosures (dual controller, 4x 10 GbE uplinks)
- 84 * 6 + 14 * 15 = 714
- 714x Kinetic/8TB drives
- SFP+ twinax cables (data)
- Cat6 cables (mgmt)

I/O Connectivity – Data Path (2017)



Management Connectivity (2017)



SUMMARY

Software-Defined Storage

- Scale-out storage is all about density (PB/rack) and cost ($/TB)
  - achieved by simplicity
  - less components, less cables
  - less code, less layers
- Many deployments need flexibility
  - start small, grow large
  - adjustable compute/storage ratios
  - purchase-time choice is good; dynamic choice is even better
- Ethernet drives offer this flexibility & scalability

Today's Speakers

- Object Drive Overview, David Slik
- Some Products & Observations, Enrico Signoretti
- Experiments & Experiences with Object Drives, Erik Riedel

After This Webcast

SNIA CSI | CLOUD STORAGE

- Please rate this webcast. We value your feedback
- This webcast and a copy of the slides will be on the SNIA Cloud Storage website and available on-demand
  - http://www.snia.org/forum/csi/knowledge/webcasts
- A Q&A from this webcast, including answers to questions we couldn't get to today, will be on the SNIACloud blog
  - http://www.sniacloud.com/
- Follow us on Twitter @SNIACloud

SNIA CSI | CLOUD STORAGE

Thank you.