

## IP-Based Object Drives Now Have a Management Standard

**Live Webcast** 

April 20, 2017 10:00 am PT





Enrico Signoretti OpenIO



David Slik Co-Chair, SNIA Cloud TWG NetApp

Erik Riedel Dell EMC



Alex McDonald Co-Chair SNIA CSI, NetApp





160

unique member companies



3,500 active contributing members



50,000 IT end users & storage pros worldwide

### Learn more: snia.org/technical







- 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.





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



## David Slik – NetApp



- 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/



- 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



- 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**



#### SATA/SAS/PCIe

#### ATA/SCSI /NVMe

Disk and/or NVM

#### **Traditional Drive**

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



- The Object Drive TWG produced a specification for scalable management of IP Based Drives
- Based on the RedFish management specification from DMTF
  - <u>https://www.dmtf.org/sites/default/files/standards/documents/</u> <u>DSP0266\_1.0.1.pdf</u>
- Uses Odata (OASIS) for RESTful interface
- Minimizes and simplifies the management of resources
- SNIA standard specifies common features and references the other standards





- 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?



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?



As a manager, how do I configure devices?

- CPU Firmware?
- drive firmware?
- the network?
- ... & select the drive application?



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?



- 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?



- 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





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



## **Enrico Signoretti**



### 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







# 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







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



SDS



#### 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





### **Erik Riedel**



## SCALE



#### **Mechanicals**







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

#### Density

2012	Disks (raw) @ 3TB	Disks (protected)	Racks @ 480 disks
5 <b>PB</b>	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**

RU

NILE DENSE

GbE 10 GbE

10 GbE

Voyager

60 Disk





23 PB petabytes 48 nodes 2,880 disks



## FLEXIBILITY

#### **Experiments in Flexibility (Goal)**



RU	NILE DENSE		RU	NILE DENSE
40	GbE		40	GbE
39	10 GbE	IL	39	10 GbE
38	10 GbE	IL	38	10 GbE
37	Server 4 Node	I L	37	Server 4 No
36	berrer intode	1 1	36	Server 4 ne
35	Server 4 Node		35	Server 4 No
34			34	
33	Vovager		33	Vovage
32			32	
31	10 Disk		31	30 Dis
30			30	
29	Vovager	I F	29	Vovage
20	10.01	-	20	
26	10 DISK	I F	26	30 DIS
25		I F	25	
24	Voyager	I F	24	Voyage
23			23	
22	TODISK		22	30 DIS
21	1/2012 202		21	1/
20	voyager		20	voyage
19	10 Dick		19	20 Dic
18	TODISK		18	
17	Vouagor	1 [	17	Vovag
16	voyagei	I [	16	
15	10 Disk	I L	15	30 Dis
14	10 0151	I L	14	30 813
13	Vovager		13	Vovage
12			12	
11	10 Disk		11	30 Dis
10			10	
9	Vovager	-	9	Vovage
0	10.01	-	0	
6	10 DISK	I F	6	30 DIS
5	11	1 1	5	X/
4	voyager		4	voyage
3	10 Dick		3	20 Dic
2	TODISK		2	
1	Not Used	I [	1	Not Used
		_		
c	640 I B		1	.7 F D
0	0 nodec		Q	nodos
C	o nodes			nodes
0				
Ö	80 disks			U CIISK

NILE DENSE	RU	NILE DENSE	
GbE	40	GbE	
10 GbE	39	10 GbE	
10 GbE	38	10 GbE	
Server 4 Node	37	Server 4 Node	
	36		
Server 4 Node	33	Server 4 Node	
	22		
Voyager	32	Voyager	
20 Dick	31	60 Dick	
SUDISK	30	OU DISK	
Moungar	29	Moupgor	
voyagei	28	voyagei	
30 Disk	27	60 Disk	
50 BISK	26	00 0131	
Vovager	25	Vovager	
	24		
30 Disk	23	60 Disk	
	22		
Voyager	21	Voyager	
20 0:-1-	19	CO Diala	
30 DISK	18	OU DISK	
Mouagor	17	Moupgor	
voyager	16	voyager	
30 Disk	15	60 Disk	
30 0131	14	00 DISK	
Vovager	13	Vovager	
	12		
30 Disk	10	60 Disk	
	10		
Voyager	8	Voyager	
20 Dick	7	60 Dick	
SUDISK	6	OU DISK	
Vovager	5	Vovagor	
VUyagei	4	VUyagei	
30 Disk	3	60 Disk	
	2		
Not Used	1	Not Used	
.9 PB	3.8PB		
nadaa	9 nodos		
noaes	o nodes		
	100 10 1		
0 disks	480 disks		

	40	10 GbE	
Ē	39	10 GbE	
Γ	38	GbE	
	37	Server 4 Node	
	36		
	35	Server 4 Node	
L	34		
L	33	Server 4 Node	
L	32	Server 4 Node	
L	31	Server 4 Node	
L	30		
L	29	Server 4 Node	
L	28		
Ļ	27	Server 4 Node	
L	26		
H	25		
H	24		
ŀ	23		
ŀ	22		
ŀ	21		
H	20		
H	19	Empty	
ŀ	18		
ŀ	1/	Empty	
ŀ	10		
ŀ	10	Titon /Vinotic	
ŀ	13	manykinetic	
H	12	84 Disk	
ŀ	11		
ŀ	10		
ŀ	9	Titan/Kinetic	
ŀ	8	manykinetic	
Ŀ	7	84 Disk	
ŀ	6		
ŀ	5		
ŀ	4	Titan/Kinetic	
ŀ	3	many Killetic	
Ŀ	2	84 Disk	
ŀ	1		

2.0 PB

24 nodes

252 disks

ECS FLEX

RU

			7	
5	10 GbE		6	
4	1 GbE		5	
3	Server 4 Nede		4	1
2	Server 4 Noue		3	
1	Kinetic 12d		2	
			1	
96 TB 4 nodes 12 disks				

10 GbE	15	
10 GbE	14	
1 GbE	13	
Server 4 Node	12	
Server 4 Noue	11	
Kinetic 12d	10	
Kinetic 12d	9	
	8	
97 TR	7	
12 10	6	
nodes	5	
	4	
1 disks	3	T
	2	

4 24

33

15	10 GDE	
14	10 GbE	
13	1 GbE	
12	Sonyor 4 Nodo	
11	Server 4 Noue	
10	Kinetic 12d	
9	Kinetic 12d	
8	Kinetic 12d	
7	Kinetic 12d	
6	Kinetic 12d	
5	Kinetic 12d	
4	Kinetic 12d	
3	Kinetic 12d	
2	Kinetic 12d	
1	Kinetic 12d	

960 TB 4 nodes 120 disks

### **Experiment – SAS Switching (2012)**



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







LSI<sup>™</sup> SAS6160 Switch

right

1:120 Disk Node with HA SAS and 2 Service/Management Nodes





### **Experiment – Kinetic (2014)**

![](_page_35_Picture_1.jpeg)

![](_page_35_Picture_2.jpeg)

- Newisys ٠ EDA-4605 Enclosure
  - 60-disk
  - dual 10 GbE controllers
  - 4x 10 GbE uplinks

![](_page_35_Picture_7.jpeg)

Ultra-Efficient Ethernet-Direct HDD Storage System delivering Object storage to Cloud and Big Data Deployments

PRODUCT SHEET

The Newisys EDA-4605 paves the way for the latest developments in Storage technology targeting Ethernet-direct Hard Disk Drive technology. Optimized for Object storage, the EDA-4605 implements an ultra-efficient storage platform for Cloud and Big Data deployments. It is ideal for scale-out and distributed storage solutions.

Unlike traditional storage boxes, the EDA-4605 provides redundant Ethernet fabrics that connect directly to the disk drives, eliminating many layers of overhead and enabling new levels of storage scalability. Whether deployed in Cloud installations, for Big Data or the traditional data center, the EDA-4605 delivers object storage at unprecedented efficiency.

The first disk product to leverage the EDA-4605 is the new Seagate Kinetic Open Storage drive.

With up to 60 x 3.5" Seagate Kinetic Open Storage drives per 4U enclosure, the industry-unique Newisys EDA-4605 is an ultra-dense, space and power saving, complexity-reducing storage solution. The Newisys EDA-4605 fits nicely into a standard 19" wide, 1m deep, rack that easily creates a 15 drives/U object storage building block. With 4TB drives, this can deliver 2.4 Petabyte per standard 42U rack, and can easily scale out beyond that.

![](_page_35_Picture_13.jpeg)

 Optimized for Ethernet-Direct HDDs, such as the novel Seagate Kinetic Open Storage Drives • Ideal building block for Object Storage deployments Reduces complexity and improves efficiency for Cloud and Big Data Object Storage installations Full-featured, highly available, high performance Object Storage

 Product Features
 Four 10GbE connections to the datacenter • Redundant, 1GbE connections to each of the 60 HDDs • Dual, redundant, hot-pluggable Ethernet Switch and Management (ESM) modules Dual, redundant, hot-pluggable, high efficiency power supply and fan units (PSU) Redundant hot-pluggable, system blowers implemented in the PSUs Modular design increases product configuration flexibility Standard chassis customiza and branding available

EDA-4605 Dual Ethernet

Switch and Management

- 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)**

RU 

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

![](_page_36_Picture_2.jpeg)

CS FLEX	RU	ECS FLEX	211
10 GbE	40	10 GbE	40
10 GbE	39	10 GbE	39
GbE	38	GbE	38
Rinjin	37	Rinjin	37
4 Blade	36	4 Blade	36
Rinjin	35	Rinjin	35
4 Blade	34	4 Blade	34
Rinjin	33		33
4 Blade	32	4 Blade	32
	31		31
	30		30
Titan/Kinetic	29		29
	28		28
84 DISK	27		27
	26		26
	25		25
Titan/Kinetic	24		24
94 Dick	23		23
84 DISK	22		22
	21		21
	20		20
Titan/Kinetic	19		19
84 Dick	18		18
04 DISK	17		17
	16		16
	15		15
Titan/Kinetic	14		14
84 Disk	13		13
04 DISK	12	14 DISK	12
	11		11
_	10		10
Titan/Kinetic	9		9
84 Disk	8		8
	7		7
	6		6
-	5		5
Titan/Kinetic	4		4
84 Disk	3		3
-			
	1		1

ECS FLEX 10 GbF GbE Rinjin

#### Parts List

9x Rinjin servers (36 nodes)

SNIA. | CLOUD

CSI | STORAGE

- 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)

![](_page_37_Figure_0.jpeg)

![](_page_38_Figure_0.jpeg)

![](_page_39_Picture_0.jpeg)

## SUMMARY

![](_page_40_Picture_1.jpeg)

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

![](_page_41_Picture_0.jpeg)

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

![](_page_42_Picture_1.jpeg)

- 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

![](_page_43_Picture_0.jpeg)

# Thank you.