IP-Based Object Drives Now Have a Management Standard

Live Webcast

April 20, 2017
10:00 am PT
SNIA at a glance

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
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
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](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?
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?
- How do I monitor drive hardware health?
- How do I monitor drive data health?
- How do I monitor drive application health?
- How do I tell when something is unhealthy?
- How do I tell when something is about to fail?
- How do I tell when something has failed?
- How do I 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
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
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
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
Scale

<table>
<thead>
<tr>
<th>Generation</th>
<th>Srvs</th>
<th>Encs</th>
<th>Disks</th>
<th>TB Drv</th>
<th>TB Rack</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gen1 (2008)</td>
<td>16</td>
<td>16</td>
<td>240</td>
<td>1TB</td>
<td>240TB</td>
</tr>
<tr>
<td>Gen2 (2010)</td>
<td>6</td>
<td>24</td>
<td>360</td>
<td>2TB</td>
<td>720TB</td>
</tr>
<tr>
<td>Gen3 (2012)</td>
<td>8</td>
<td>8</td>
<td>480</td>
<td>4TB</td>
<td>1.9PB</td>
</tr>
<tr>
<td>Gen4 (2014)</td>
<td>8</td>
<td>8</td>
<td>480</td>
<td>6TB</td>
<td>2.9PB</td>
</tr>
<tr>
<td>Gen5 (2015)</td>
<td>8</td>
<td>8</td>
<td>480</td>
<td>8TB</td>
<td>3.8PB</td>
</tr>
<tr>
<td>Gen5+ (2016)</td>
<td>8</td>
<td>8</td>
<td>784</td>
<td>10TB</td>
<td>7.8PB</td>
</tr>
<tr>
<td>Gen6 (2017)</td>
<td>4</td>
<td>8</td>
<td>784</td>
<td>12TB</td>
<td>9.4PB</td>
</tr>
</tbody>
</table>

Atmos

ECS

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
- sleds
- trays
- modular
### Density

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

<table>
<thead>
<tr>
<th>Year</th>
<th>Disks (raw) @ 3TB</th>
<th>Disks (protected)</th>
<th>Racks @ 480 disks</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 PB</td>
<td>1,700 disks</td>
<td>2,700 disks</td>
<td>6 racks</td>
</tr>
<tr>
<td>20 PB</td>
<td>6,700 disks</td>
<td>11,000 disks</td>
<td>23 racks</td>
</tr>
<tr>
<td>50 PB</td>
<td>17,000 disks</td>
<td>27,000 disks</td>
<td>56 racks</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>Disks (raw) @ 6TB</th>
<th>Disks (protected)</th>
<th>Racks @ 480 disks</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 PB</td>
<td>830 disks</td>
<td>1,300 disks</td>
<td>3 racks</td>
</tr>
<tr>
<td>20 PB</td>
<td>3,300 disks</td>
<td>5,300 disks</td>
<td>12 racks</td>
</tr>
<tr>
<td>50 PB</td>
<td>8,300 disks</td>
<td>13,000 disks</td>
<td>28 racks</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>Disks (raw) @ 10TB</th>
<th>Disks (protected)</th>
<th>Racks @ 780 disks</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 PB</td>
<td>500 disks</td>
<td>670 disks</td>
<td>1 rack</td>
</tr>
<tr>
<td>20 PB</td>
<td>2,000 disks</td>
<td>2,700 disks</td>
<td>4 racks</td>
</tr>
<tr>
<td>50 PB</td>
<td>5,000 disks</td>
<td>6,700 disks</td>
<td>9 racks</td>
</tr>
</tbody>
</table>
## Scale Out

<table>
<thead>
<tr>
<th>RU</th>
<th>NILE DENSE</th>
<th>GB E</th>
<th>Servers 4 Node</th>
<th>Voyager 60 Disk</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>GbE</td>
<td>10 GbE</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

23 PB petabytes
48 nodes
2,880 disks
FLEXIBILITY
Experiments in Flexibility (Goal)
Experiment – SAS Switching (2012)
Experiment – Kinetic (2014)

from Seagate material “Kinetic Open Storage - Enabling Break-through Economics in Scale-out Object Storage”
• 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**

### 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)
Management Connectivity (2017)

Node 1

Node 2

Node x

1Gb Mgmt Switch

VLAN 1 (private mgmt)

VLAN 2 (diag)

Kinetic DAE

Kinetic DAE

Kinetic DAE

deployment network
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

- 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](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
- Follow us on Twitter @SNIACloud
Thank you.