CAPACITY-ORIENTED LUSTRE FILESYSTEM: PUSHING THE LIMITS WITH OAK STORAGE
OAK STORAGE?

Persistent storage for research data
Capacity-oriented
Low-cost
Multi-cluster: mounted on Sherlock, SCG and XStream
SFTP, Globus, NFS and SMB Gateways available
Optional per-group backup-to-the-cloud using restic
Oak storage growth

400 TB/month

Oak storage average allocations since the beginning (April 2017)
OAK ULTRA-HIGH DENSITY LUSTRE I/O CELL
OAK ULTRA-HIGH DENSITY I/O CELL (2020)

IB FDR “Storage Fabric”

- IB FDR
- SAS 4-lane 12Gb/s links
- Lustre OSS (HA)
- SAS Storage Array
- Infiniband
- JBOD SAS I/O modules
- 102 x 14 TB HDDs

- 8 x
- 2 x 56 = 112 Gb/s
- 4 x 48 = 192 Gb/s
- 2 x 48 = 96 Gb/s (single JBOD)

SAS 12Gb/s SW

SAS 4-lane 12Gb/s links

Lustre MDS server

SAS 4-lane 12Gb/s links

JBOD SAS I/O modules

Infiniband
SAS Storage Array
Lustre MDS server
SAS

8 x
102 x 14 TB HDDs
SECRET SAUCE: SWITCHED SAS BACKEND

- Astek A54812-SW-04 12-Port (48-Lanes) Gen3 SAS Switch
- 12Gbps per lane, 48Gbps/port
JBOD: HGST DATA102
SAS BACKEND TOPOLOGY
FAT-TREE 2:1

- Excellent performance and resiliency vs. SAS ring or daisy-chaining
- JBOD hot-plug
- 2 SAS controllers (HBA) per server needed to manage the number the SAS devices: ~1,650 (!)
9 PB usable
11.5 PB raw

816 drives
8 x 102

40 OSTs/OSS
Up to 80 OSTs/OSS in failover mode

< $0.00009/GB/month
Cost of all equipments of a fully populated rack, over 4 years
SEE ALSO
OAK WEB SITE

https://oak-storage.stanford.edu/
sasutils
Serial Attached SCSI (SAS) Linux utilities and Python library
sasutils

- Display SAS fabric tree and provide aggregated view of devices
- `sas_discover, sas_devices, sas_counters, ses_report`
- Based on sysfs (and also `sg3_utils` and `smp_utils`)
- Support SES Enclosure Nickname

- Published in EPEL 7 and soon in EPEL 8
- Or use “pip install sasutils”
sasutils :: sas_discover (topology)

[root@oak-io1-s1 ~]# sas_discover -v

-oak-io1-s1
|--host1 SAS9300-8e
    `--8x--expander-1:0 ASTEK
        |--3x--end_device-1:0:0
            `--enclosure io1-sassw1 ASTEK
        |--4x--expander-1:1 QCT
            |-- 60 x end_device -- disk
            `--1 x end_device -- enclosure io1-jbod1-0 QCT
        |--4x--expander-1:2 QCT
            |-- 60 x end_device -- disk
            `--1 x end_device -- enclosure io1-jbod2-0 QCT
        |--4x--expander-1:3 QCT
            |-- 60 x end_device -- disk
            `--1 x end_device -- enclosure io1-jbod3-0 QCT
        |--4x--expander-1:4 QCT
            |-- 60 x end_device -- disk
            `--1 x end_device -- enclosure io1-jbod4-0 QCT
        |--4x--expander-1:5 QCT
            |-- 60 x end_device -- disk
            `--1 x end_device -- enclosure io1-jbod5-0 QCT
        |--4x--expander-1:6 QCT
            |-- 60 x end_device -- disk
            `--1 x end_device -- enclosure io1-jbod6-0 QCT

--host12 SAS9300-8e
    `--8x--expander-12:0 ASTEK
        |--3x--end_device-12:0:0
            `--enclosure io1-sassw2 ASTEK
        |--4x--expander-12:1 QCT
            |-- 60 x end_device -- disk
            `--1 x end_device -- enclosure io1-jbod1-1 QCT
        |--4x--expander-12:2 QCT
            |-- 60 x end_device -- disk
            `--1 x end_device -- enclosure io1-jbod2-1 QCT
        |--4x--expander-12:3 QCT
            |-- 60 x end_device -- disk
            `--1 x end_device -- enclosure io1-jbod3-1 QCT
        |--4x--expander-12:4 QCT
            |-- 60 x end_device -- disk
            `--1 x end_device -- enclosure io1-jbod4-1 QCT
        |--4x--expander-12:5 QCT
            |-- 60 x end_device -- disk
            `--1 x end_device -- enclosure io1-jbod5-1 QCT
        |--4x--expander-12:6 QCT
            |-- 60 x end_device -- disk
            `--1 x end_device -- enclosure io1-jbod6-1 QCT
**Zero configuration tool**

**Display aggregate view of SAS end devices per enclosure**

**Easily spot discrepancy**

**Use -v to get a verbose view of all SAS devices**

```
[root@oak-io4-s1 ~]# sas_devices
Found 2 SAS hosts
Found 38 SAS expanders
Found 6 enclosure groups
Enclosure group: [io4-jbod3][io4-jbod3]

<table>
<thead>
<tr>
<th>NUM</th>
<th>VENDOR</th>
<th>MODEL</th>
<th>REV</th>
<th>PATHS</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>HGST</td>
<td>HUH721212AL5200</td>
<td>A640</td>
<td>2</td>
</tr>
</tbody>
</table>

Total: 60 block devices in enclosure group

Enclosure group: [io4-jbod4][io4-jbod4]

<table>
<thead>
<tr>
<th>NUM</th>
<th>VENDOR</th>
<th>MODEL</th>
<th>REV</th>
<th>PATHS</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>HGST</td>
<td>HUH721212AL5200</td>
<td>A640</td>
<td>2</td>
</tr>
</tbody>
</table>

Total: 60 block devices in enclosure group

Enclosure group: [io4-jbod5][io4-jbod5]

<table>
<thead>
<tr>
<th>NUM</th>
<th>VENDOR</th>
<th>MODEL</th>
<th>REV</th>
<th>PATHS</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>HGST</td>
<td>HUH721212AL5200</td>
<td>A640</td>
<td>2</td>
</tr>
</tbody>
</table>

Total: 60 block devices in enclosure group

Enclosure group: [io4-jbod6][io4-jbod6]

<table>
<thead>
<tr>
<th>NUM</th>
<th>VENDOR</th>
<th>MODEL</th>
<th>REV</th>
<th>PATHS</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>HGST</td>
<td>HUH721212AL5200</td>
<td>A640</td>
<td>2</td>
</tr>
</tbody>
</table>

Total: 60 block devices in enclosure group

Enclosure group: [io4-jbod2][io4-jbod2]

<table>
<thead>
<tr>
<th>NUM</th>
<th>VENDOR</th>
<th>MODEL</th>
<th>REV</th>
<th>PATHS</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>HGST</td>
<td>HUH721212AL5200</td>
<td>A640</td>
<td>2</td>
</tr>
</tbody>
</table>

Total: 60 block devices in enclosure group

Enclosure group: [io4-jbod1][io4-jbod1]

<table>
<thead>
<tr>
<th>NUM</th>
<th>VENDOR</th>
<th>MODEL</th>
<th>REV</th>
<th>PATHS</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>HGST</td>
<td>HUH721212AL5200</td>
<td>A640</td>
<td>2</td>
</tr>
</tbody>
</table>

Total: 60 block devices in enclosure group
```
SES stands for **SCSI Enclosure Services**

The initiator can communicate with the enclosure using a specialized set of SCSI commands to access power, cooling, and other non-data characteristics.

**Examples of use**

Status only monitoring

```
$ ses_report --status --json
```

SES Element descriptors metrics (suitable for Graphite)

```
$ ses_report --carbon
```
https://github.com/stanford-rc/sasutils