Towards Energy Efficient Data Management in HPC: The *Open Ethernet Drive* Approach

Anthony Kougkas, Anthony Fleck, Xian-He Sun
Outline

- Introduction
- Background
- Evaluation results
- Conclusions
- Future directions
Introduction

- What is an Open Ethernet Drive (OED)?
- Who makes them?
- Why do we need one?
Open Ethernet Drive

- An “intelligent” storage device in a 3.5” form factor
- ARM-based CPU
- Fixed-size RAM
- Ethernet card
- ...and a disk drive.
Open Ethernet Drive ecosystem

- Kinetic Open Storage Project (8/2015) created by
  - Seagate
  - Western Digital (HGST)
  - Toshiba

- Joined by

<table>
<thead>
<tr>
<th>Cisco</th>
<th>Cleversafe (IBM)</th>
<th>DELL</th>
</tr>
</thead>
<tbody>
<tr>
<td>DigitalSense</td>
<td>NetApp</td>
<td>Open vStorage</td>
</tr>
<tr>
<td>RedHat</td>
<td>Scality</td>
<td></td>
</tr>
</tbody>
</table>
Why an Open Ethernet Drive in HPC?

• Two main reasons:
  – Optimize global I/O performance
  – Reduce energy consumption
I/O optimization using OED

- **Processor-per-disk** database machines (1983), perform simple queries on disk exploiting locality.
- **Active Storage** (1998), proposed to offload some computations to storage servers.
- **Decoupled Execution Paradigm** (2013), specialized data nodes perform computations to minimize the data movement.
- **Active Burst Buffer** (2016) perform in-situ visualization and/or analysis.
- **OED** encapsulates a lot of the necessary tech in a small, affordable device that will enable extra functionality.
Energy and cost savings

- Designed with low-powered mobile components.
- OED small factor requires less space.
- And thus, more efficient cooling.
- Less and easy maintenance.
Outline

- Introduction
- Background
- Evaluation results
- Conclusions
- Future directions
OED architecture

- Designed to bring computation closer to the data.
- Presented in enclosures of multiple such drives.
- Enclosures have an embedded switched fabric (60Gbit/s).
- Runs Linux OS (Debian 8.0).
- Internal components are subject to each implementation.
OED use cases

- Mirantis, collaborated with HGST to deploy Openstack’s Swift object store, Ceph’s OSDs and GlusterFS bricks.
- Cloudian, deployed its own Hyperstore service on an enclosure of 60 OED drives.
- Skylable, deployed their object store service SkylableSX.
- All of the above concluded that OED is the perfect building block for an energy efficient and horizontally scalable storage cluster.

Can we bring it to HPC and harness its strengths?
Outline

- Introduction
- Background
- Evaluation results
- Conclusions
- Future directions
Test environment

- Three categories:
  - Hardware components with benchmarks
  - Overall device with real applications
  - Energy consumption (Watts)

- Software used:
  - Stress-ng
  - SysBench
  - Iperf
  - Out-of-core sorting
  - Vector addition
  - Descriptive statistics

<table>
<thead>
<tr>
<th>Feature</th>
<th>OED</th>
<th>Personal Computer</th>
<th>Server Node</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU</td>
<td>ARM 32bit 1-core (1GHz)</td>
<td>AMD Athlon X4 4-cores (3.7GHz)</td>
<td>2xAMD Opteron 8-cores (2.3GHz)</td>
</tr>
<tr>
<td>RAM</td>
<td>2GB DDR3 1600Mhz</td>
<td>16GB DDR3 2400Mhz</td>
<td>8GB DDR2 667Mhz</td>
</tr>
<tr>
<td>Disk</td>
<td>Megascale DC4000.B 4TB 7200rpm</td>
<td>Seagate Barracuda 1TB 7200rpm</td>
<td>WD 250GB 7200rpm</td>
</tr>
<tr>
<td>Network</td>
<td>1 Gbit/s</td>
<td>1 Gbit/s</td>
<td>1 Gbit/s</td>
</tr>
<tr>
<td>OS</td>
<td>Debian 8.0</td>
<td>Ubuntu 14.04</td>
<td>Ubuntu server 9.04</td>
</tr>
<tr>
<td>Kernel</td>
<td>3.14.3</td>
<td>4.4.0-34</td>
<td>2.6.28</td>
</tr>
<tr>
<td>Year</td>
<td>2014</td>
<td>2015</td>
<td>2009</td>
</tr>
</tbody>
</table>
CPU performance

Stress-ng

16x slower than personal computer
9x slower than server node

Sysbench

50x slower than personal computer
30x slower than server node
# RAM performance

## Stress-ng

<table>
<thead>
<tr>
<th>Memory Results</th>
<th>OED</th>
<th>Personal Computer</th>
<th>Server Node</th>
</tr>
</thead>
<tbody>
<tr>
<td>max bandwidth</td>
<td>8 GiB/s</td>
<td>60 GiB/s</td>
<td>35 GiB/s</td>
</tr>
<tr>
<td>average bandwidth</td>
<td>4.2 GiB/s</td>
<td>24 GiB/s</td>
<td>8.9 GiB/s</td>
</tr>
<tr>
<td>min latency</td>
<td>0.5 ns</td>
<td>0.2 ns</td>
<td>0.3 ns</td>
</tr>
<tr>
<td>average latency</td>
<td>3.5 ns</td>
<td>2.1 ns</td>
<td>2.5 ns</td>
</tr>
</tbody>
</table>

12x slower than personal computer
5x slower than server node

11x slower than personal computer
7x slower than server node
Disk performance

Stress-ng

- 2.3x faster than personal computer
- 1.7x faster than server node

Sysbench

- 4.5x faster than personal computer
- 3.5x faster than server node
Ethernet performance

**Stress-ng**

- 2-6x slower than personal computer
- 1-4x slower than server node

**Iperf**

- 3x slower than personal computer
- 2x slower than server node
Real Applications

Let's just say OEDs are currently slower :(  

11/14/16
Higher Performance comes with a cost.

- OED needs $1/10^{th}$ of the power compared to an average node.
- Sorting integers took 3x more time on the OED but consumed $1/14^{th}$ of watts needed per sorting unit.
- Sorting 4GB of integers:
  - OED $\rightarrow$ 1380w
  - Server $\rightarrow$ 3800w
Outline

• Introduction
• Background
• Evaluation results
• Conclusions
• Future directions
Conclusions

- This 1st generation of OED technology is not yet on par with the average server node in terms of performance.
- Energy savings seem promising.
- OEDs could be used to run parallel file system servers for an archival and energy efficient storage solution.
- As OED technology progresses, data-intensive operations can be accelerated by offloading computation on OEDs.
Outline

• Introduction
• Background
• Evaluation results
• Conclusions
• Future directions
Future work

- Installed MPICH and OrangeFS storage system on an enclosure of 60 OED drives.
- Initial IOR benchmarks were successful.
- The 2nd generation of OED looks very promising.
- Planning to explore the use of OED as specialized data nodes that can run operations on local data
  - Compression / decompression
  - Deduplication
  - Statistics
In the meantime...

- **Sorting**
  - Gen 1
  - Gen 2
  - Dataset size in GB: 1G, 2G, 4G
  - Total time (minutes): 0 to 90

- **Vector Addition**
  - Gen 1
  - Gen 2
  - Dataset size in GB: 1G, 2G, 4G
  - Total time (minutes): 0 to 20

- **Descriptive statistics**
  - Gen 1
  - Gen 2
  - Total time (hours): 0 to 35
  - Dataset size in GB: 1G, 2G, 4G
Towards Energy Efficient Data Management in HPC: The Open Ethernet Drive Approach

Anthony Kougkas

akougkas@hawk.iit.edu

The authors would like to acknowledge Los Alamos Lab for providing us with the prototype devices.