PDSW-DISCS'16
Monday, November 14th
Salt Lake City, USA

ILLINOIS INSTITUTE OF TECHNOLOGY

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

Anthony Kougkas, Anthony Fleck, Xian-He Sun

Outline

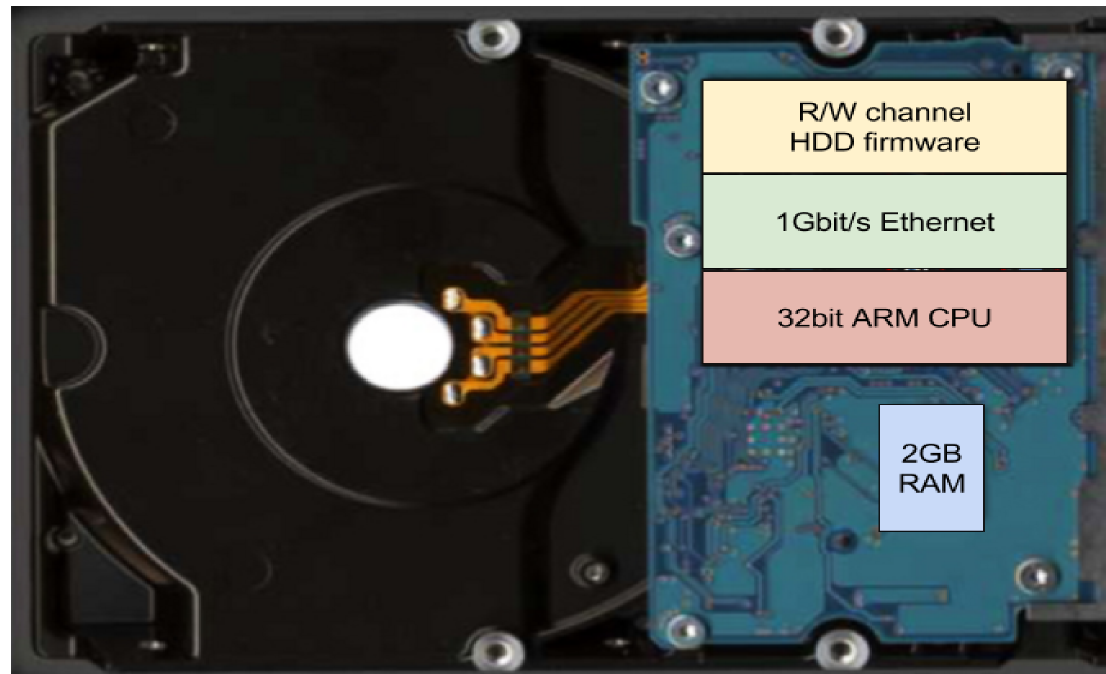
- Introduction
- Background
- Evaluation results
- Conclusions
- Future directions

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

Cisco	Cleversafe (IBM)	DELL
DigitalSense	NetApp	Open vStorage
RedHat	Scality	

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

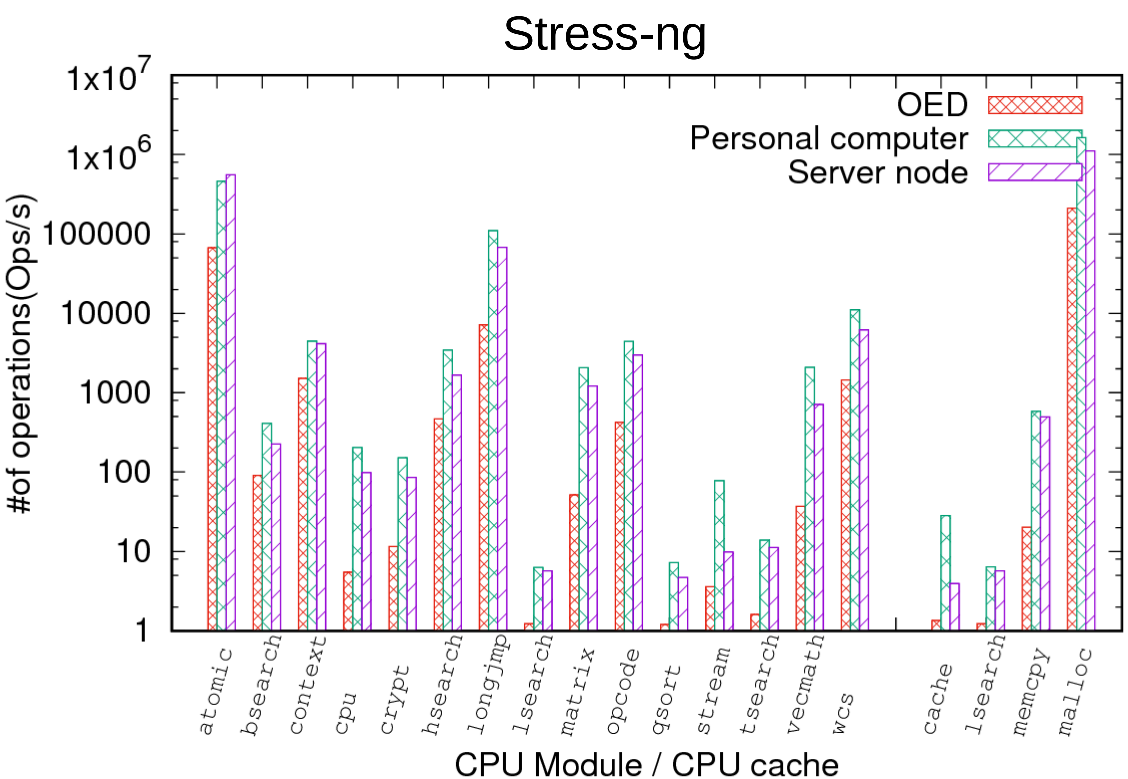
- Introduction
- Background
- Evaluation results
- Conclusions
- Future directions

Test environment

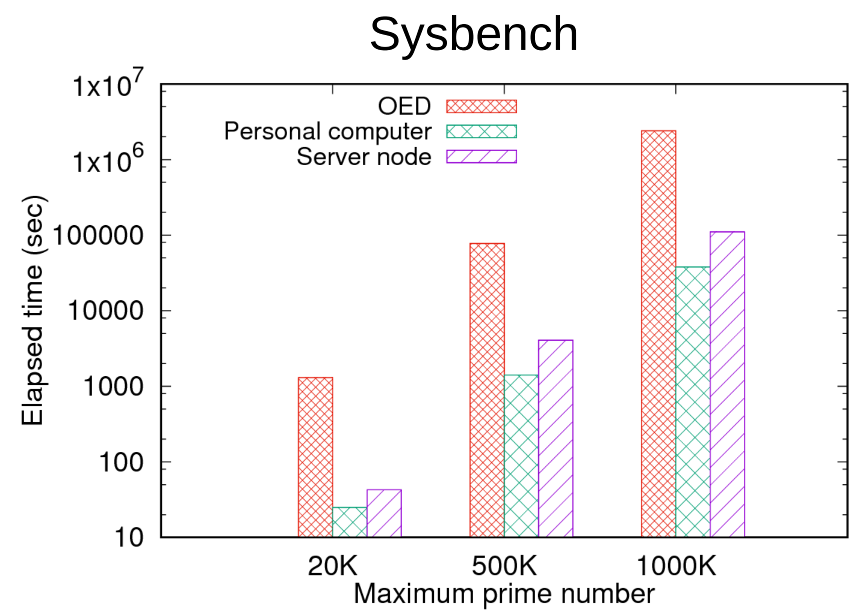
- Three categories:
  - Hardware components with benchmarks
  - Overall device with real applications
  - Energy consumption (Watts)
- Software used:
  - Stress-ng
  - SysBench
  - Iperf
  - Out-of-core sorting
  - Vector addition
  - Descriptive statistics

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

CPU performance



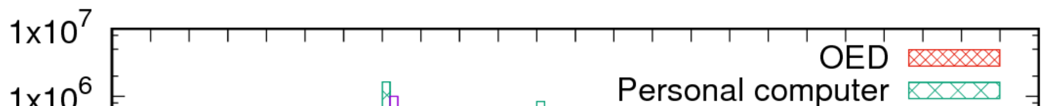
16x slower than personal computer
9x slower than server node



50x slower than personal computer
30x slower than server node

RAM performance

Stress-ng



Sysbench

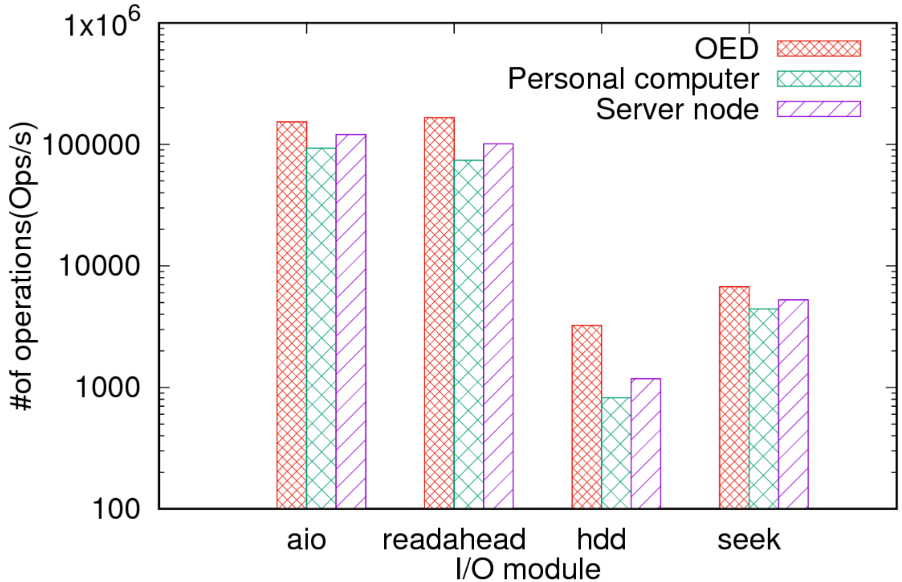
Memory Results	OED	Personal Computer	Server Node
max bandwidth	8 GiB/s	60 GiB/s	35 GiB/s
average bandwidth	4.2 GiB/s	24 GiB/s	8.9 GiB/s
min latency	0.5 ns	0.2 ns	0.3 ns
average latency	3.5 ns	2.1 ns	2.5 ns

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

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

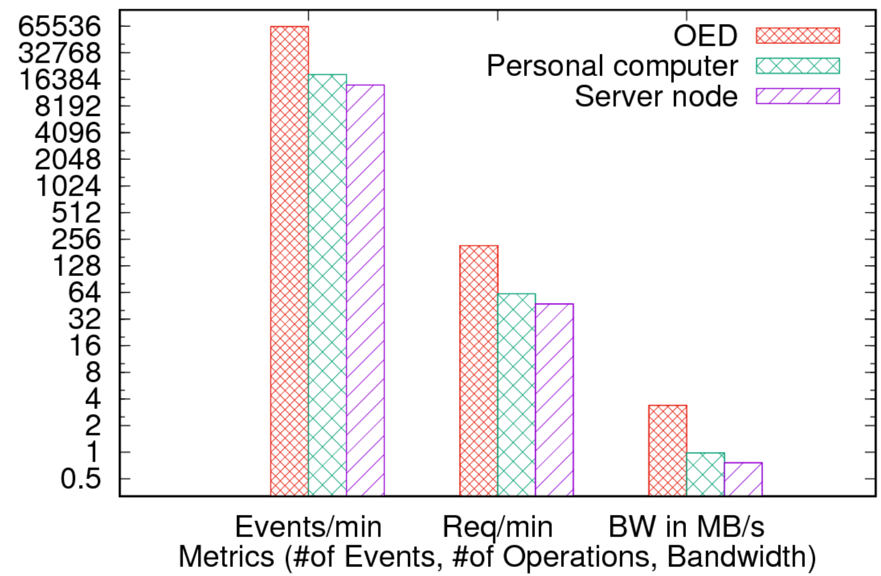
Disk performance

Stress-ng



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

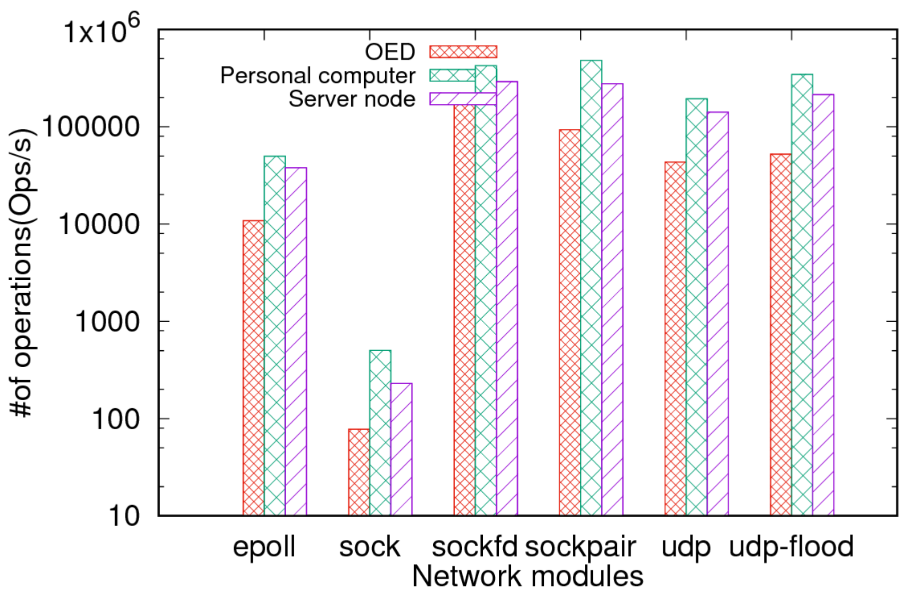
Sysbench



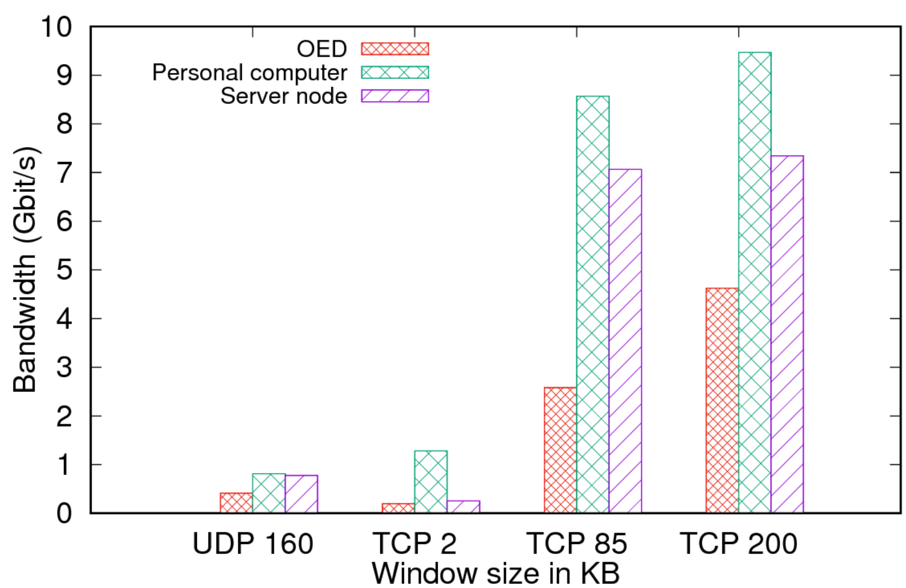
4.5x faster than personal computer
3.5x faster than server node

Ethernet performance

Stress-ng



Iperf

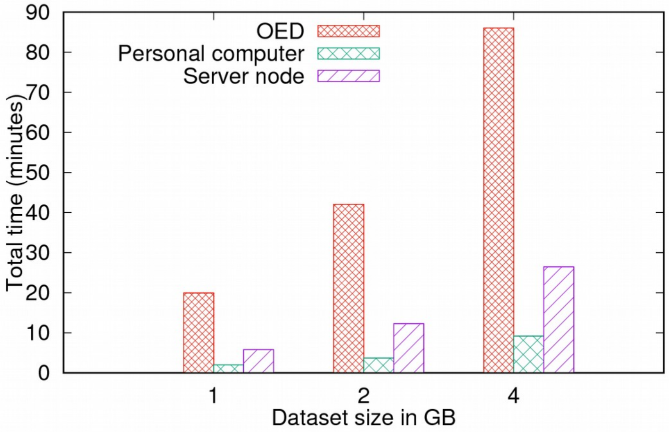


2-6x slower than personal computer
1-4x slower than server node

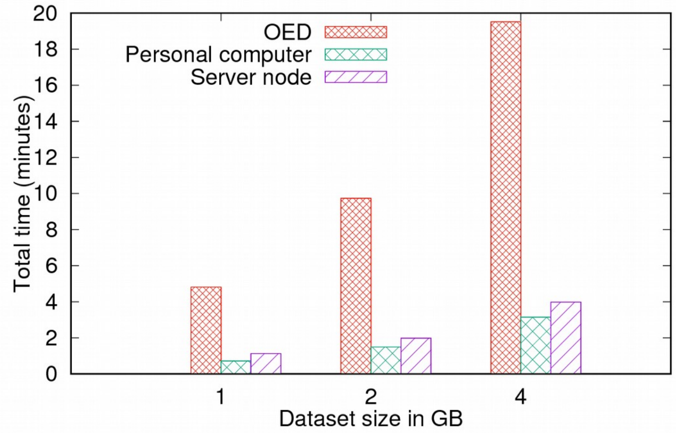
3x slower than personal computer
2x slower than server node

Real Applications

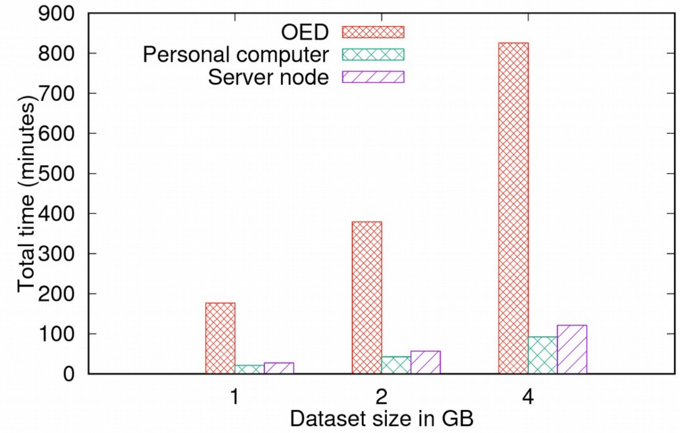
Sorting



Vector Addition

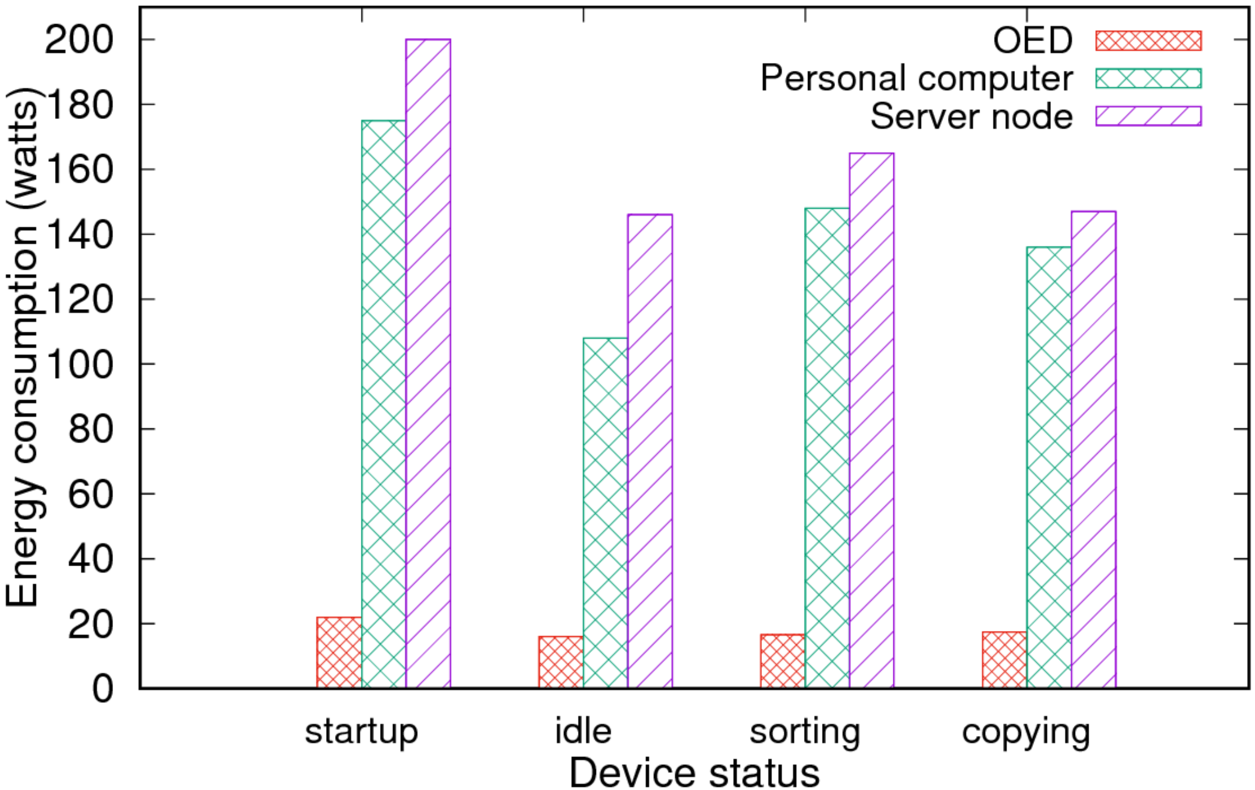


Desc. Statistics



Let’s just say OEDs are currently slower :(

Energy consumption



- Higher Performance comes with a cost.
- OED needs 1/10th of the power compared to an average node.
- Sorting integers took 3x more time on the OED but consumed 1/14th of watts needed per sorting unit.
- Sorting 4GB of integers:
  - OED → 1380w
  - Server → 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.

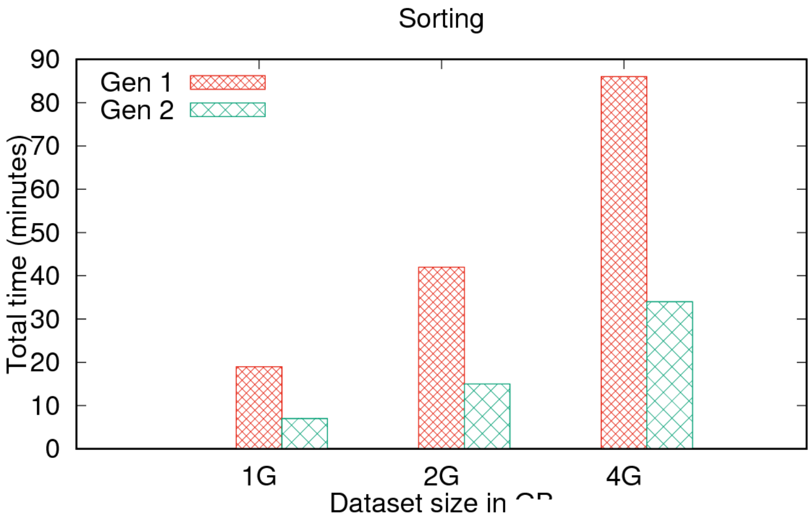
Outline

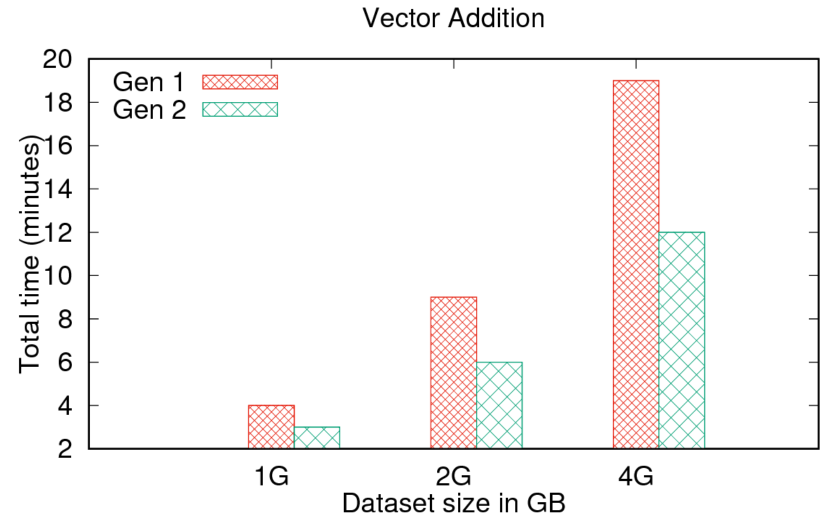
- Introduction
- Background
- Evaluation results
- Conclusions
- Future directions

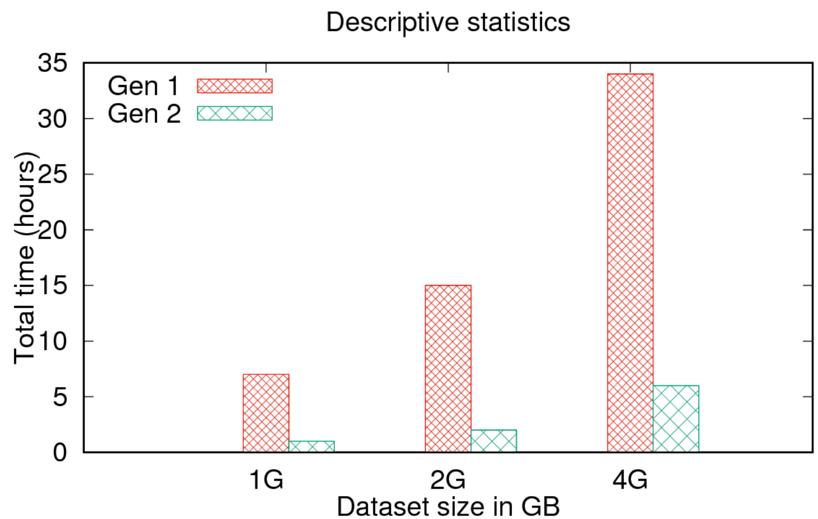
Future work

- 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…







ILLINOIS INSTITUTE OF TECHNOLOGY

Q & A

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.