



# Lustre Deployment Using Intel Omni-Path Interconnect

Brian Johanson

J. Ray Scott

# Pittsburgh Supercomputing Center

- Cooperative effort of:
  - Carnegie Mellon University
  - University of Pittsburgh
- In operation 30 years in 2016
- Supported by:
  - National Science Foundation
  - National Institutes of Health
  - Department of Energy
  - Department of Defense
  - The Commonwealth of Pennsylvania
  - Foundations and industry



# BRIDGES

A PITTSBURGH SUPERCOMPUTING CENTER RESOURCE

*Bridges: From Communities and Data to Workflows and Insight*



The \$9.65M *Bridges* acquisition is made possible by National Science Foundation award #ACI-1445606



<http://psc.edu/bridges>



# XSEDE

Extreme Science and Engineering  
Discovery Environment

- NSF award under the eXtreme Digital solicitation
  - TeraGrid Phase III: eXtreme Digital Resources for Science and Engineering (XD), NSF 08-571
  - PSC is an XSEDE Awardee, Service Provider and active participant
- A consortium of advanced digital services
  - support a growing portfolio of resources and services
    - advanced computing, high-end visualization, data analysis, and other resources and services
    - interoperability with other infrastructures
  - a virtual organization providing dynamic distributed infrastructure

# Bridges Technology Partners

-  **Hewlett Packard Enterprise**
  - Compute Servers
  - Design and Installation
-  **intel**
  - CPU technology
  - Omni-Path Architecture (OPA) Interconnect
  - Lustre file system

-  **nVIDIA**
  - Computational GPUs
-  **SUPERMICR**
  - Storage Servers

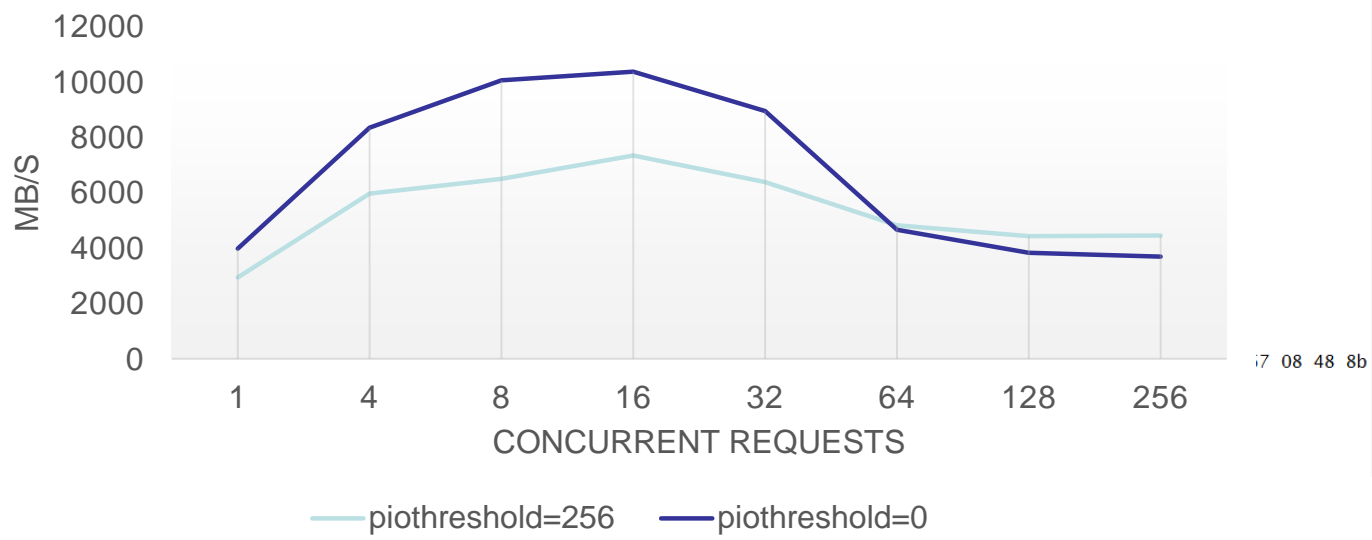
# What You Don't Want To See – Where We Started

```

[ 1029.661909] BUG: unable to handle kernel paging request at fffff9004a52c000
[ 1029.668906] IP: [] update_sge+0x2c/0xc0 [hfi1]
[ 1029.674952] PGD 103f80000: [ffff88022193d08] update_sge+0x2c/0xc0
[ 1029.680766] Ooops: 0000000000000000 (0)
[ 1029.684020] Module:
[ 1029.684020] CPU: 15 PID: 644 Comm: kworker/15:1H Tainted: P OE
[ 1029.684020] Hardware name: Intel(R) Xeon(R) CPU @ 2.00GHz/Intel(R) Xeon(R) CPU E5-2680 v4/BIOS 2.0 12/28/2015
[ 1029.684020] task: fffff9004a52c000 task.ti: fffff9004a52c000
[ 1029.684020] RIP: 0010:[ffff88022193d08] update_sge+0x2c/0xc0 [hfi1]
[ 1029.684020] RSP: 0018:ffff88022193d08 EFLAGS: 00010202
[ 1029.684020] RAX: fffff9004a52c000 RBX: 0000000000000000 RCX: 0000000000000000
[ 1029.684020] RDX: fffff9004a52c020 RSI: 0000000000000000 RDI: fffff9004a52c000
[ 1029.684020] RBP: fffff9004a52c000 R08: 0000000000000000
[ 1029.684020] R10: 0000000000000004 R11: 0000000000000000
[ 1029.684020] R13: 0000000000000000 R14: 0000000000000000
[ 1029.684020] FS: 0000000000000000(0000) GS:ffff88022193d000
[ 1029.684020] CS: 0010 DS: 0000 ES: 0000 CR0: 00000000
[ 1029.684020] CR2: fffff9004a52c000 CR3: 0000000000000000
[ 1029.684020] DR0: 0000000000000000 DR1: 0000000000000000
[ 1029.684020] DR3: 0000000000000000 DR6: 0000000000000000
[ 1029.684020] Stack:
[ 1029.684020] fffff9004a52c000 fffff9004a52c000 fffff9004a52c000
[ 1029.684020] fffff9004a52c000 fffff9004a52c000 fffff9004a52c000
[ 1029.684020] Call Trace:
[ 1029.684020] [] hfi1_verbs_s
[ 1030.006044] [] ? do_rc_ack+
[ 1030.012073] [] hfi1_verbs_s
[ 1030.018447] [] hfi1_do_send
[ 1030.024549] [] process_one_
[ 1030.030387] [] worker_threa
[ 1030.035967] [] ? rescuer_th
[ 1030.041805] [] kthread+0xc
[ 1030.046688] [] ? kthread_cr
[ 1030.053220] [] ret_from_for
[ 1030.058627] [] ? kthread_cr
[ 1030.065152] Code: 1f 44 00 00 55 8b 57 1c 89 f
[ 1030.085127] RIP: [] update_s
[ 1030.091263] RSP: fffff9004a52c000
[ 1030.094758] CR2: fffff9004a52c000
    
```



LNETH Selftest - Concurrent write test

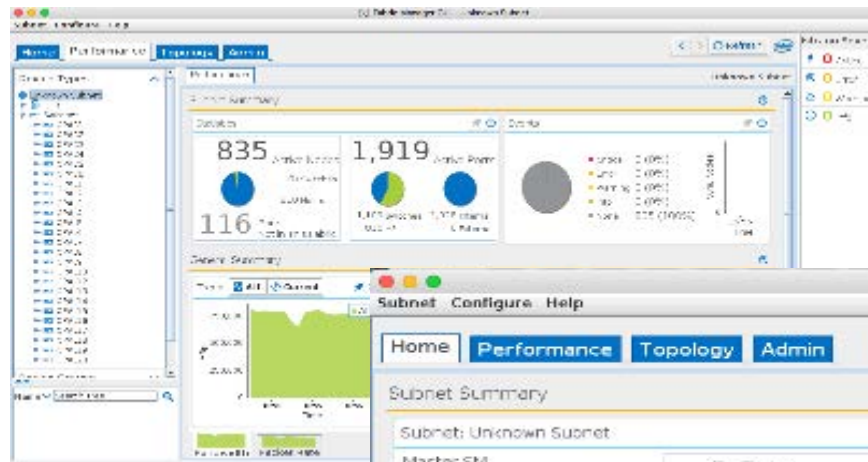


# Omni-Path Background

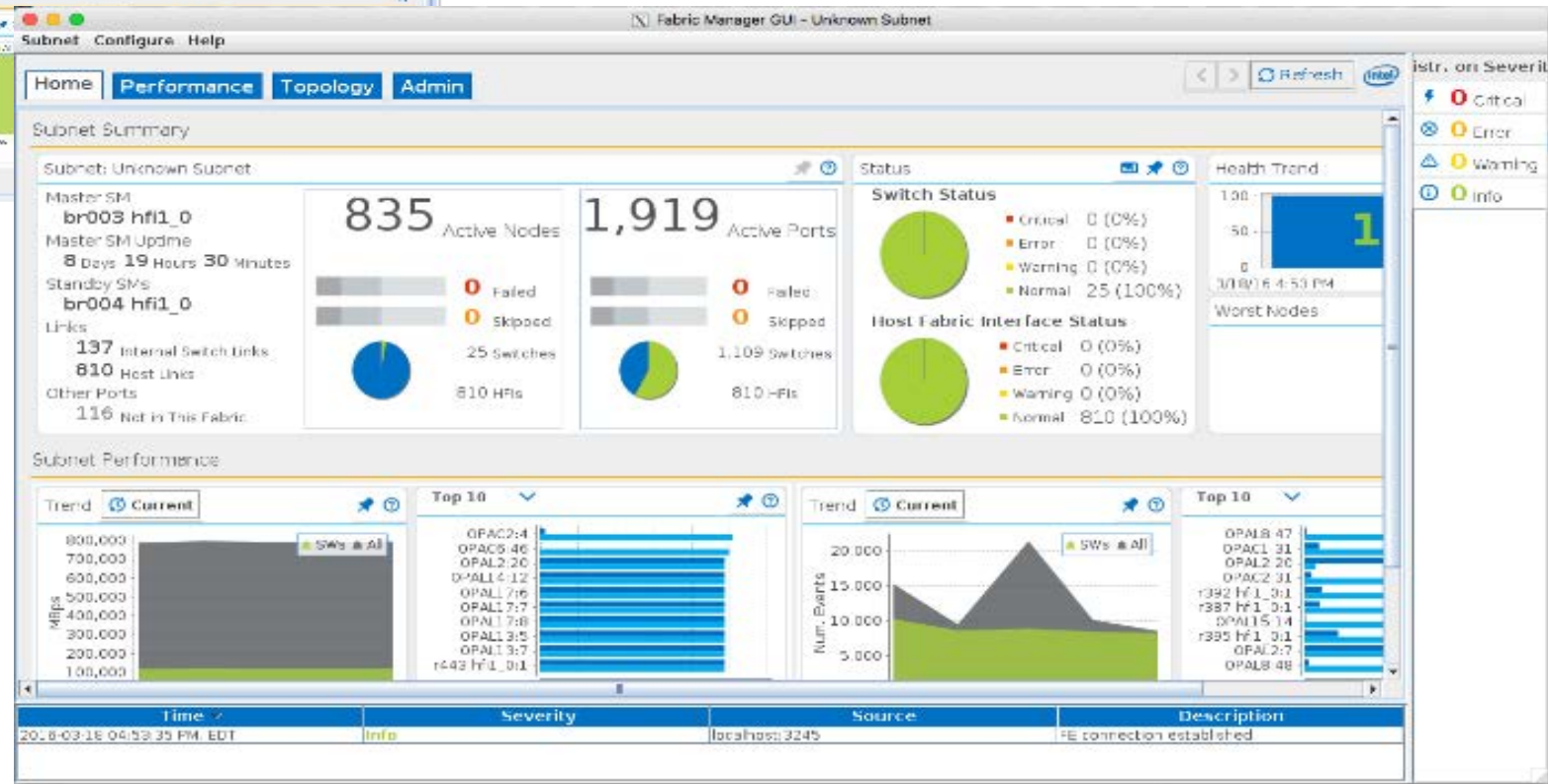
- 48 port leaf switches
- PCIe Gen 3 x16 Host Fabric Interface (HFI)
- 160M messages per second injection rate on mpi
- Quality of Service using flows
- OSU benchmarks measured:
  - 12.38 GB/s bandwidth
  - < 1  $\mu$ sec latency
- Open Fabrics support
- LNET support
- IP support



# Omni-Path Fabric Manager (FM) GUI



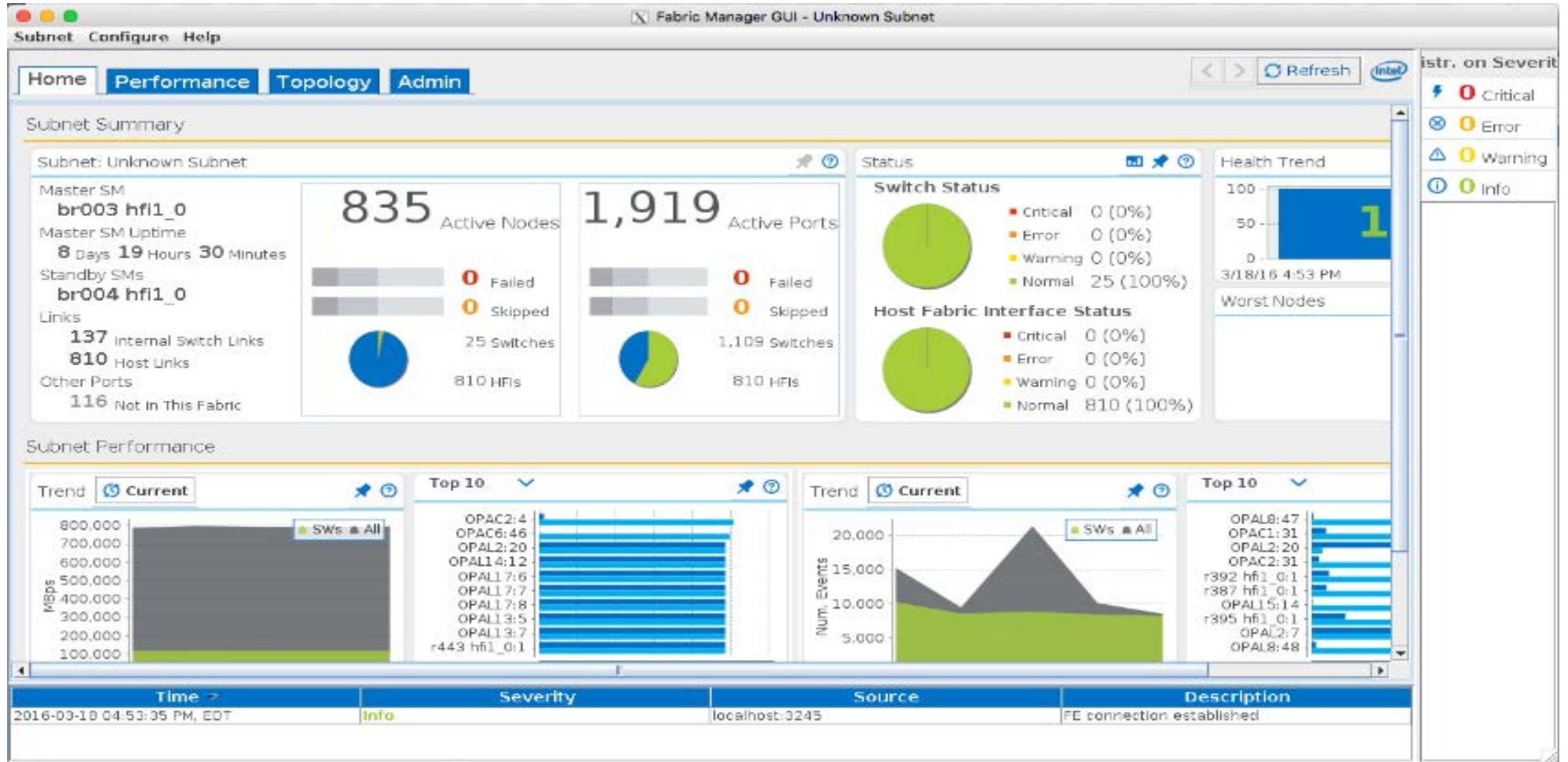
- Presentation of OPA Fabric Manager and Performance Manager data
- Concise summaries
- Click to dive deeper



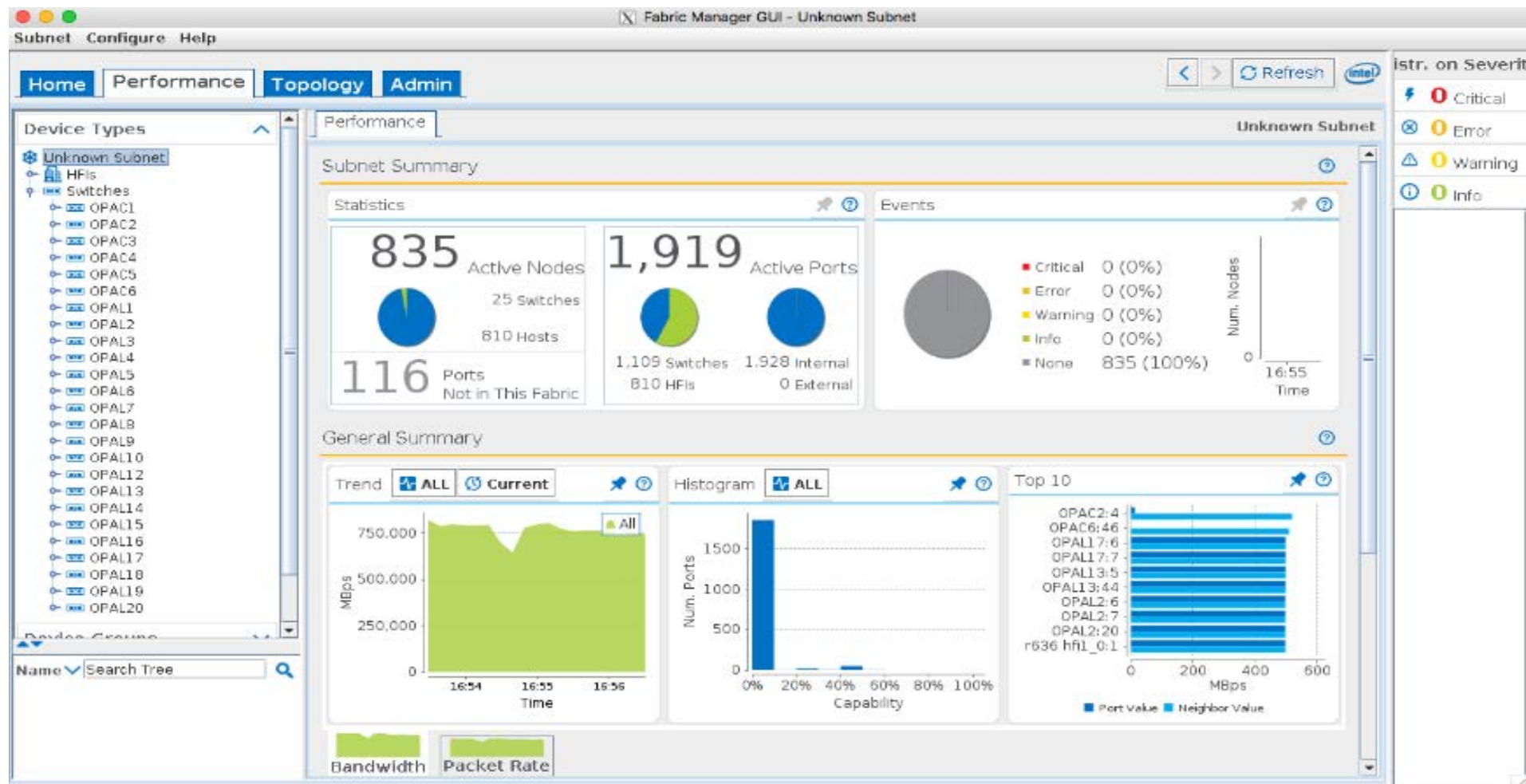
*Rich, full suite of textual and TUI tools too!*



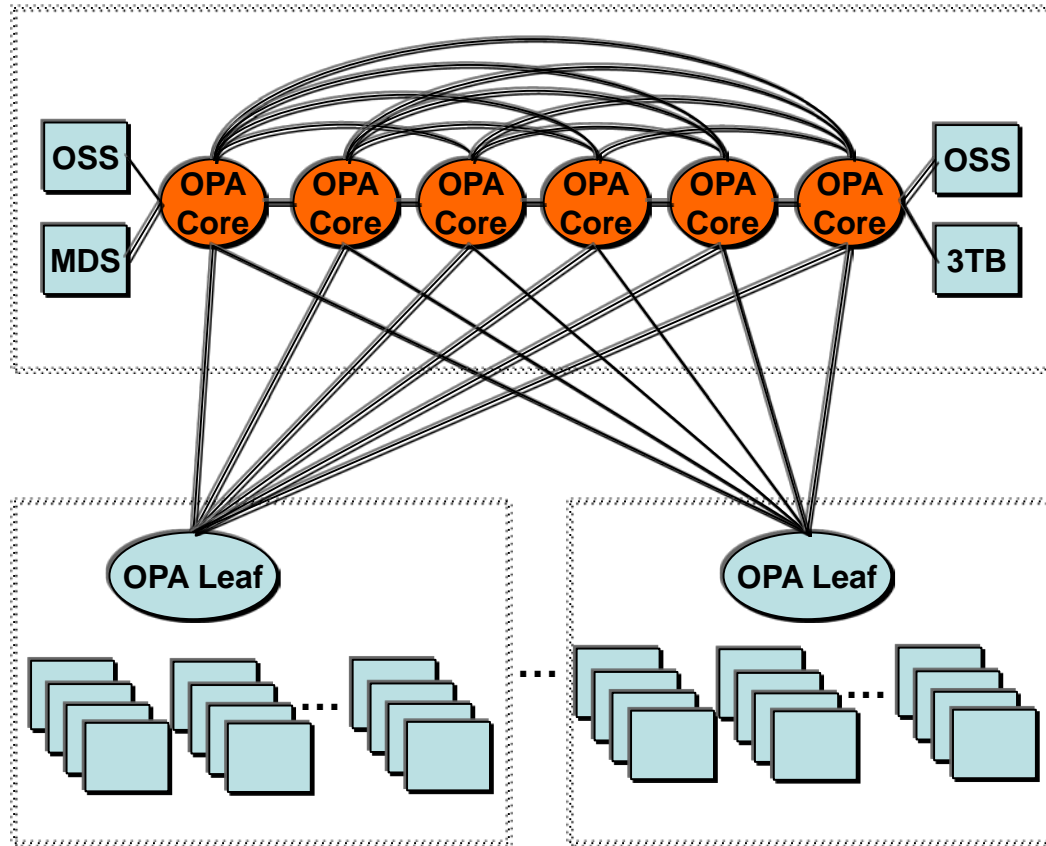
# OPA FM Home—Fabric status at a glance



# OPA FM Performance-Counter exploration tool

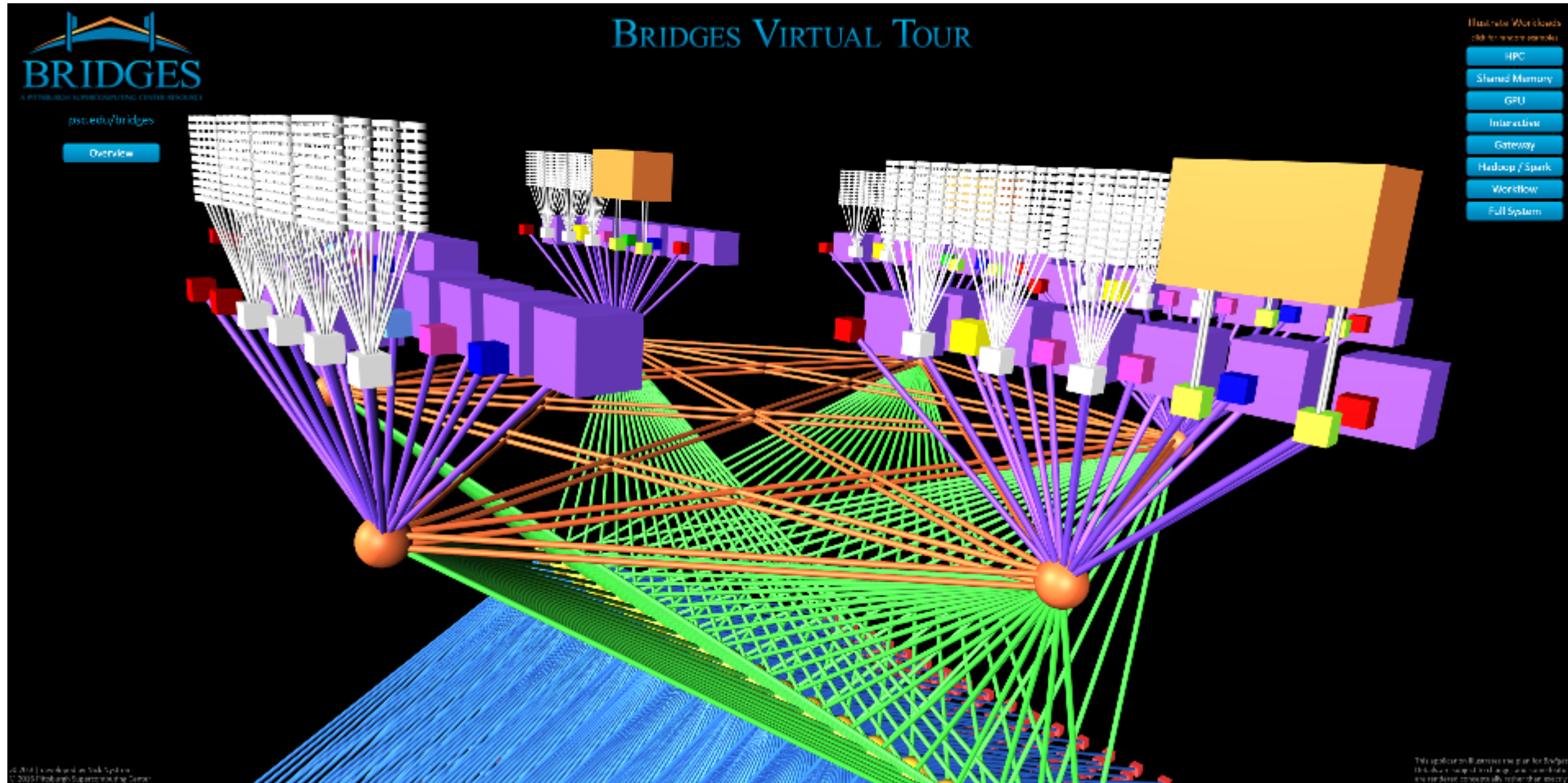


# Bridges Omni-Path Topology



- One 100Gbps Omni-Path *HFI* adapter per host
- 48-port Omni-Path Switches throughout
  - 6 “core” switches at top level
    - Full “PSC original” mesh between core switches (200Gbps)
    - Select hosts connected to core
  - 20 “leaf” switches
    - uplinks to every core switch
    - up to 42 hosts

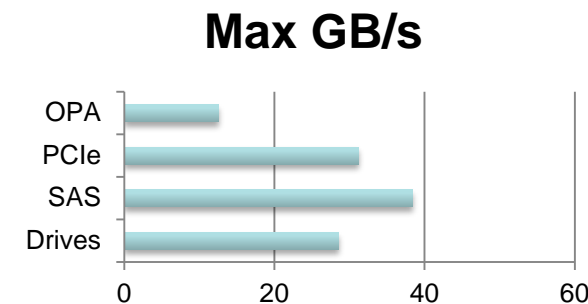
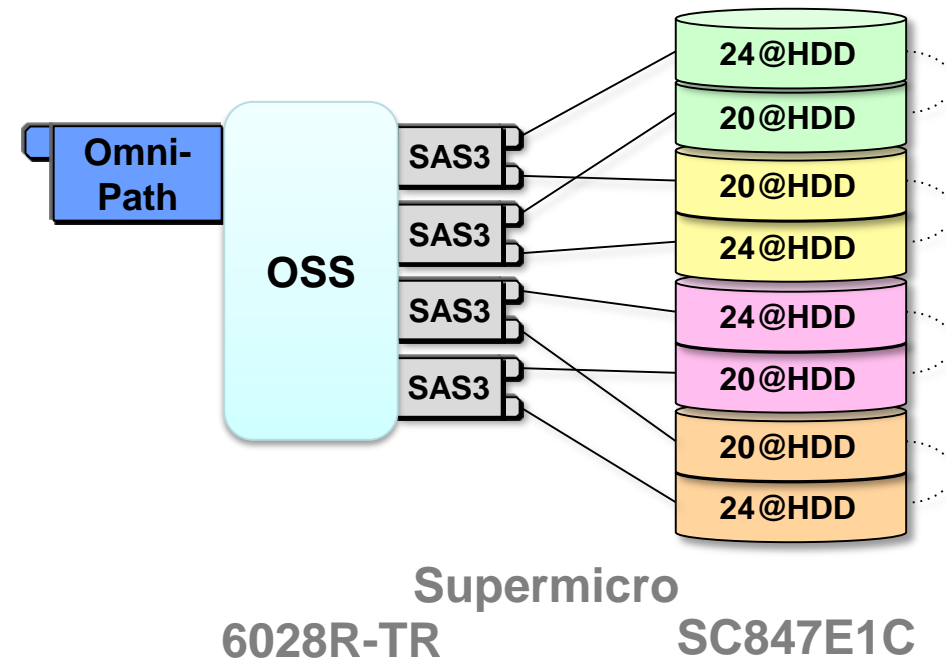
# Bridges Virtual Tour (<http://staff.psc.edu/nystrom/bvt>)



# Bridges OSS Layout

- Two E5-2650v3 (12@2.5GHz)
- 128GB RAM
- Internal disks
  - One 4TB SATA HDD (OS)
  - One 4TB SATA HDD (spare)
  - One 960GB SSD
- One Omni-Path HFI (100Gbps)
  - Link to “core” OPA switch
- Four 8 × 12Gbps SAS3 HBAs
- Four 20-bay 12Gb SAS backplanes
  - 8 × 12Gbps uplink
  - 20@4TB SATA drives
- Four 24-bay 12Gb SAS backplanes
  - 8 × 12Gbps uplink
  - 24@4TB SATA drives

Total 176 (+3 internal) drives

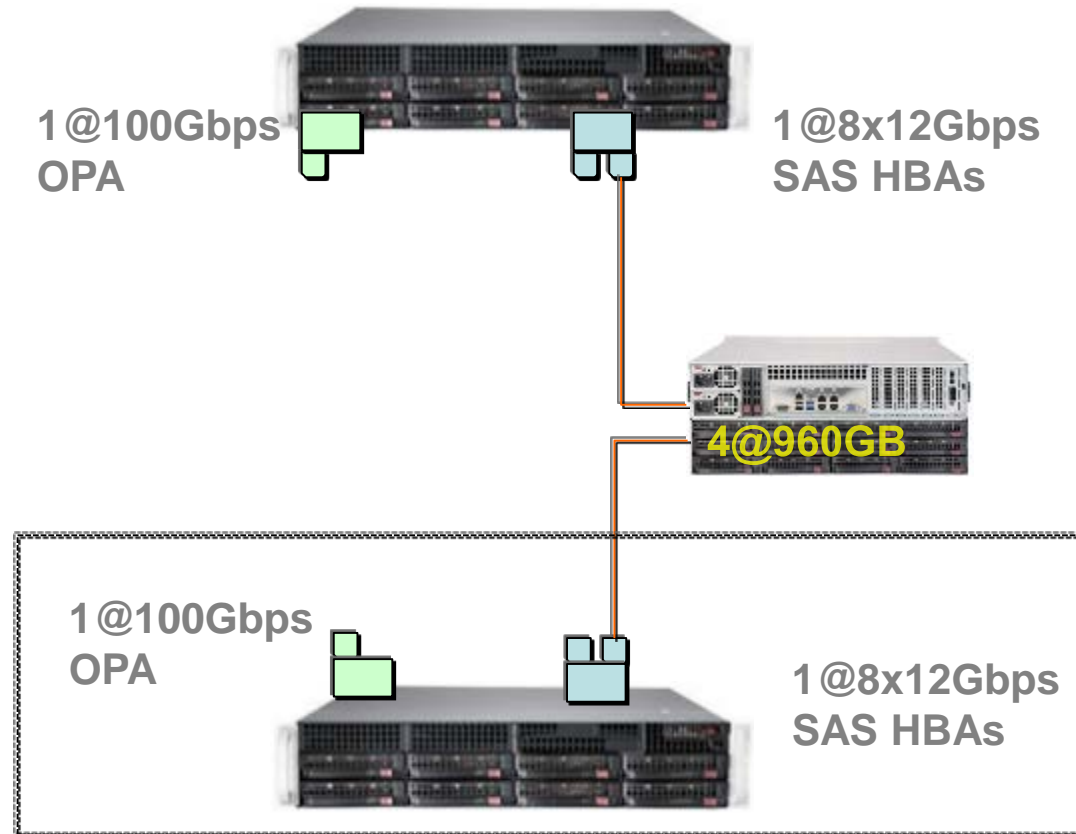


# Bridges MDS Layout

Each MDS consists of:

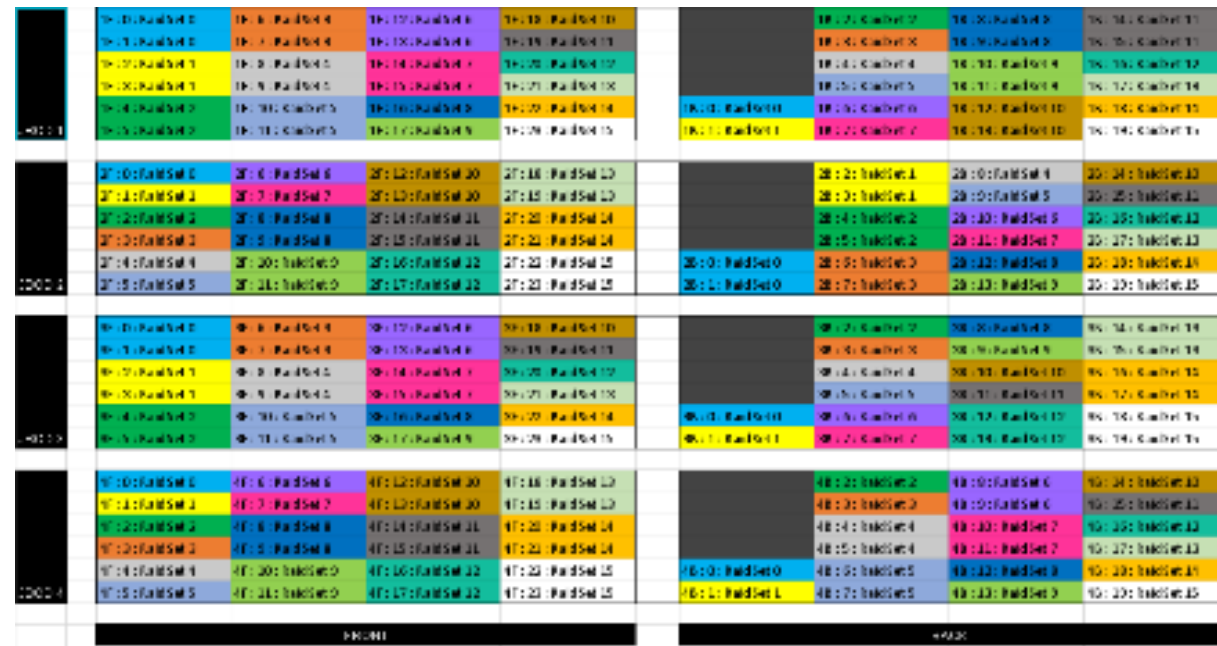
- Two E5-2695v3 (14@2.4GHz)
- 384GB RAM
- Internal disk
  - One 3TB SATA HDD (OS)
- One Omni-Path HFI (100Gbps)
  - Link to “core” OPA switch
- One 8x 12Gbps SAS3 HBAs
- One 20-bay 12Gb SAS backplanes
  - 8x 12Gbps uplink
  - 4@ 960GB SATA SSD

*Note: HA partner attached to same SAS3 backplane with access to same SATA drives.*

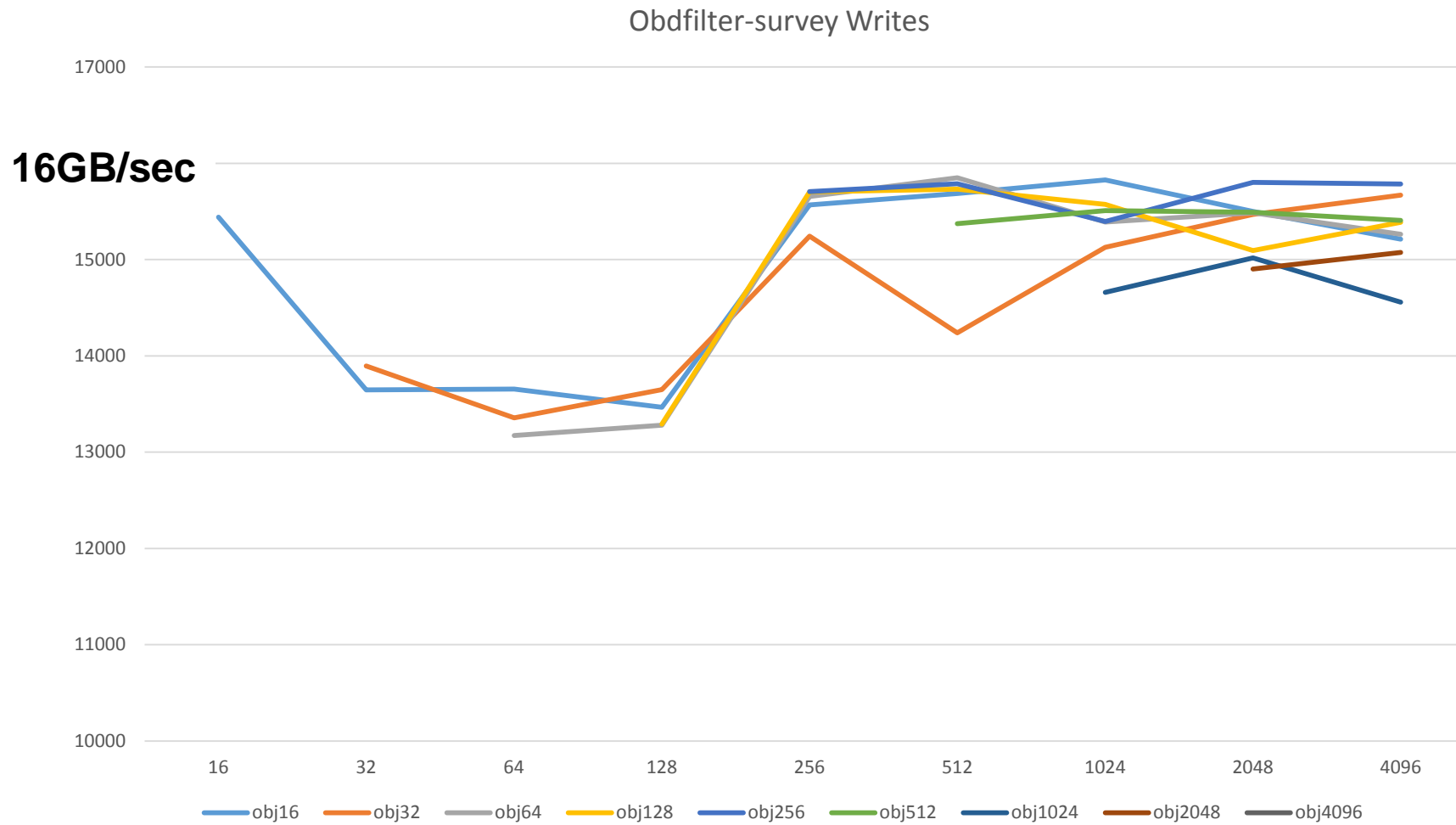


# Software

- CentOS 7.2 Kernel 3.10.0-327.3.1
- Lustre 2.8.0 RC5 -> 2.8.0 GA
- ZFS 0.6.5.5
- OSS
  - Sixteen 11-disk raidz2 pools per OSS
- MDS
  - 2 Mirrors striped

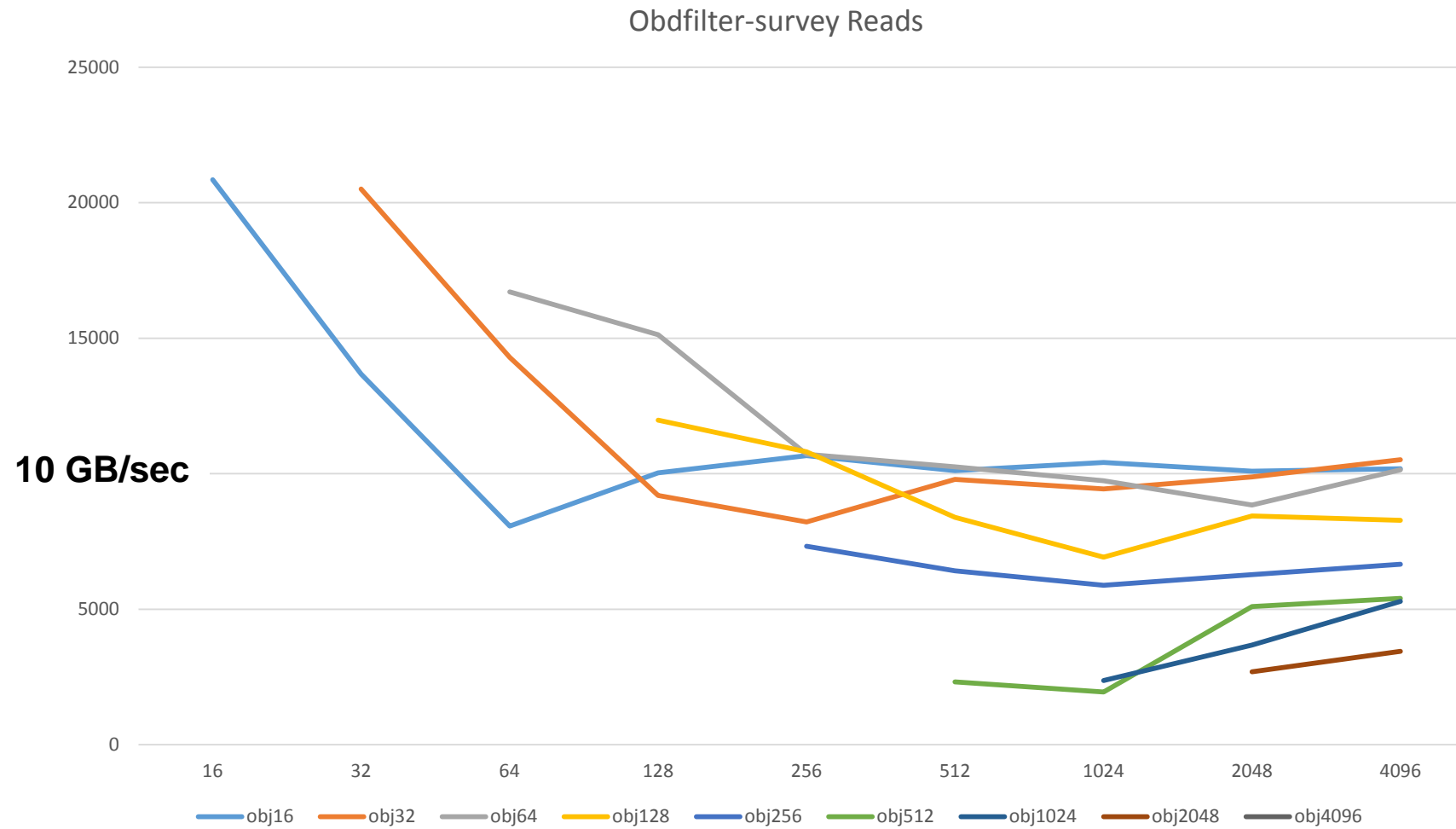


# OST Zpool Write Performance

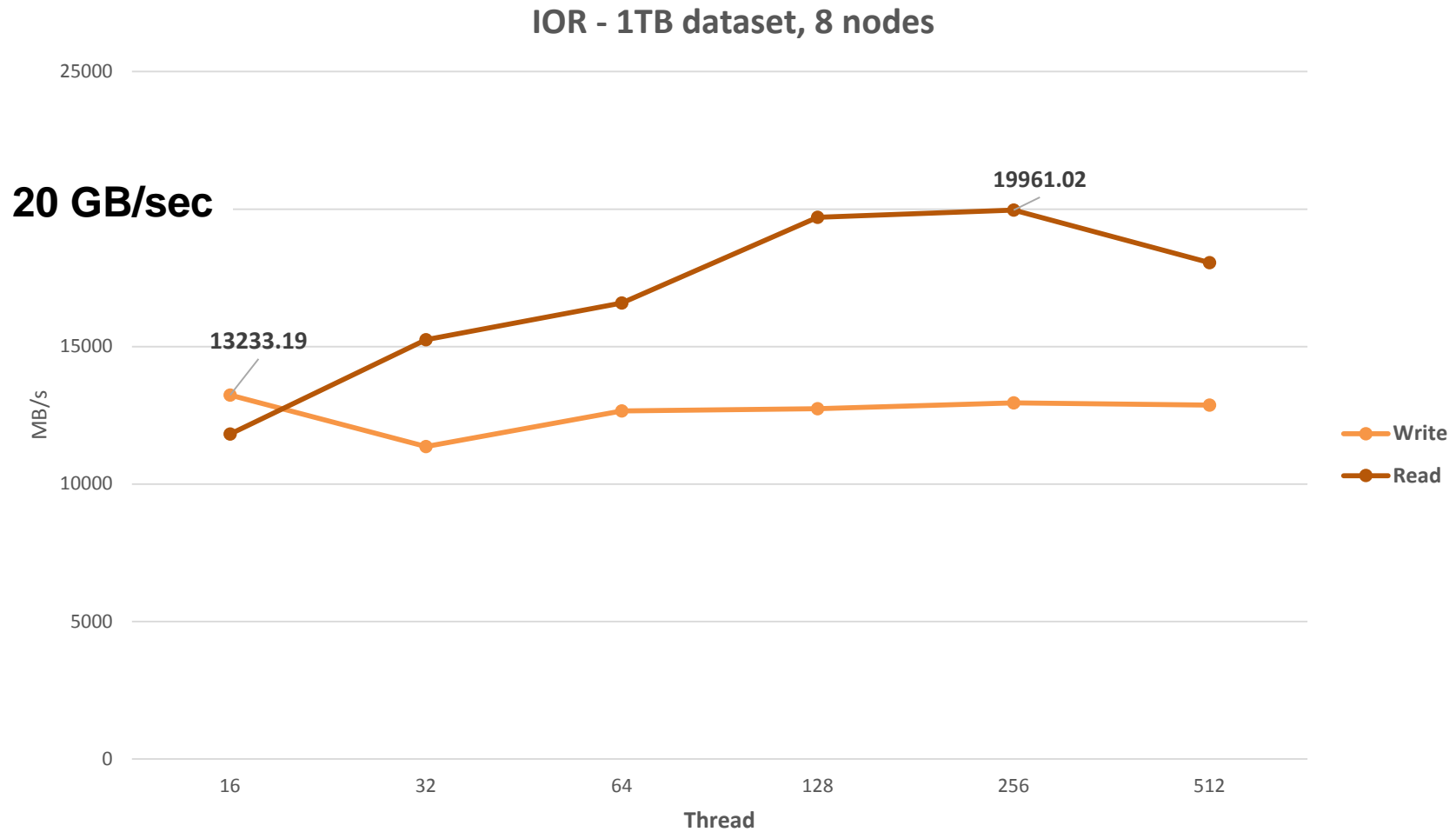




# OST Zpool Read Performance



# Client Performance



# Credits and Contacts

- Intel
- Brian Johanson : [bjohanso@psc.edu](mailto:bjohanso@psc.edu)  
Bridges System Administrator
- Nick Nystrom : [nystrom@psc.edu](mailto:nystrom@psc.edu)  
Bridges Principal Investigator (PI)
- J. Ray Scott : [scott@psc.edu](mailto:scott@psc.edu)  
Bridges Co-PI
- Jason Sommerfield : [jasons@psc.edu](mailto:jasons@psc.edu)  
PSC Research Technology Infrastructure Developer

Thank You!

