PSC
PITTSBURGH SUPERCOMPUTING CENTER

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

PSC
PITTSBURGH SUPERCOMPUTING CENTER

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

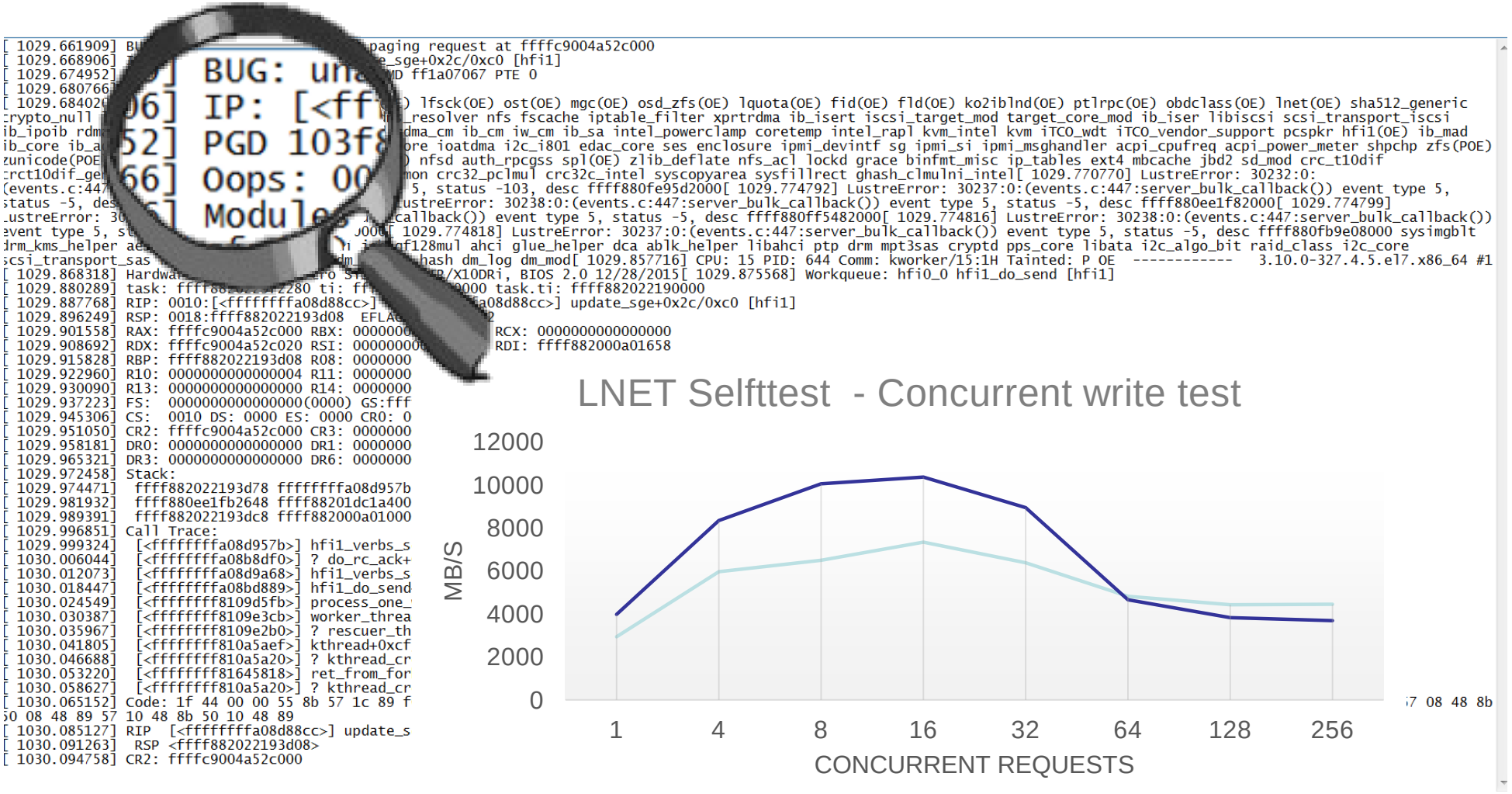
PSC PITTSBURGH SUPERCOMPUTING CENTER

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
- SUPERMICRO
  - Storage Servers

PSC PITTSBURGH SUPERCOMPUTING CENTER

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



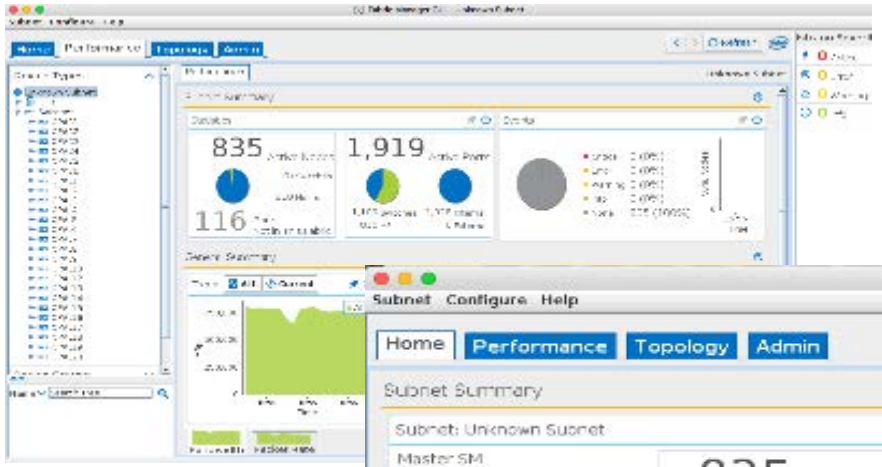
PSC
PITTSBURGH SUPERCOMPUTING CENTER

Omni-Path Background

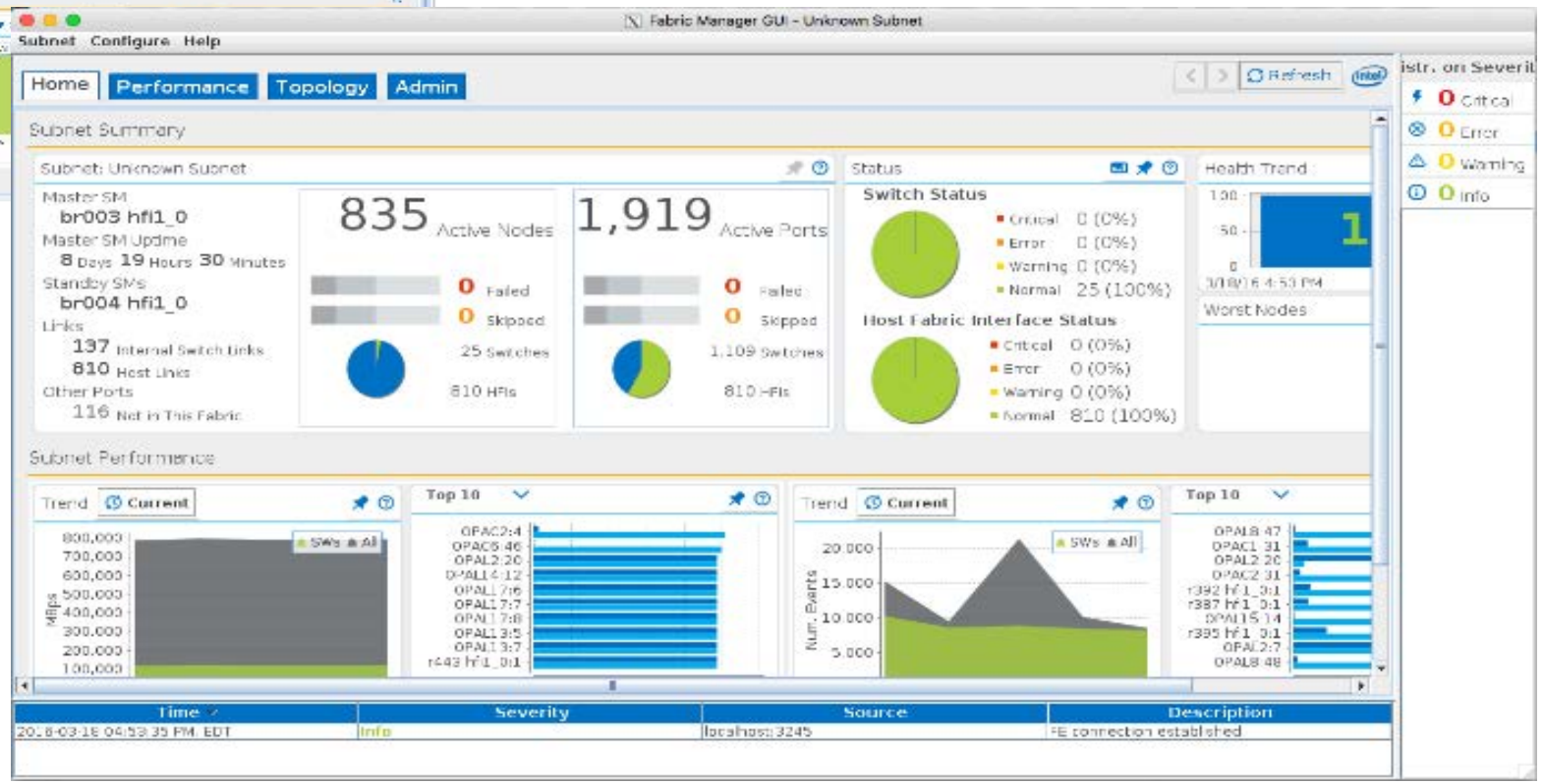
- 48 port leaf switches
- PCIe Gen 3 x16 Host Fabric Interface (HFI)
- 160M messages per second injection rate on mpi
- Quality of Service using flows
- OSU benchmarks measured:
  - 12.38 GB/s bandwidth
  - < 1 µsec latency
- Open Fabrics support
- LNET support
- IP support

PSC
PITTSBURGH SUPERCOMPUTING CENTER

Omni-Path Fabric Manager (FM) GUI



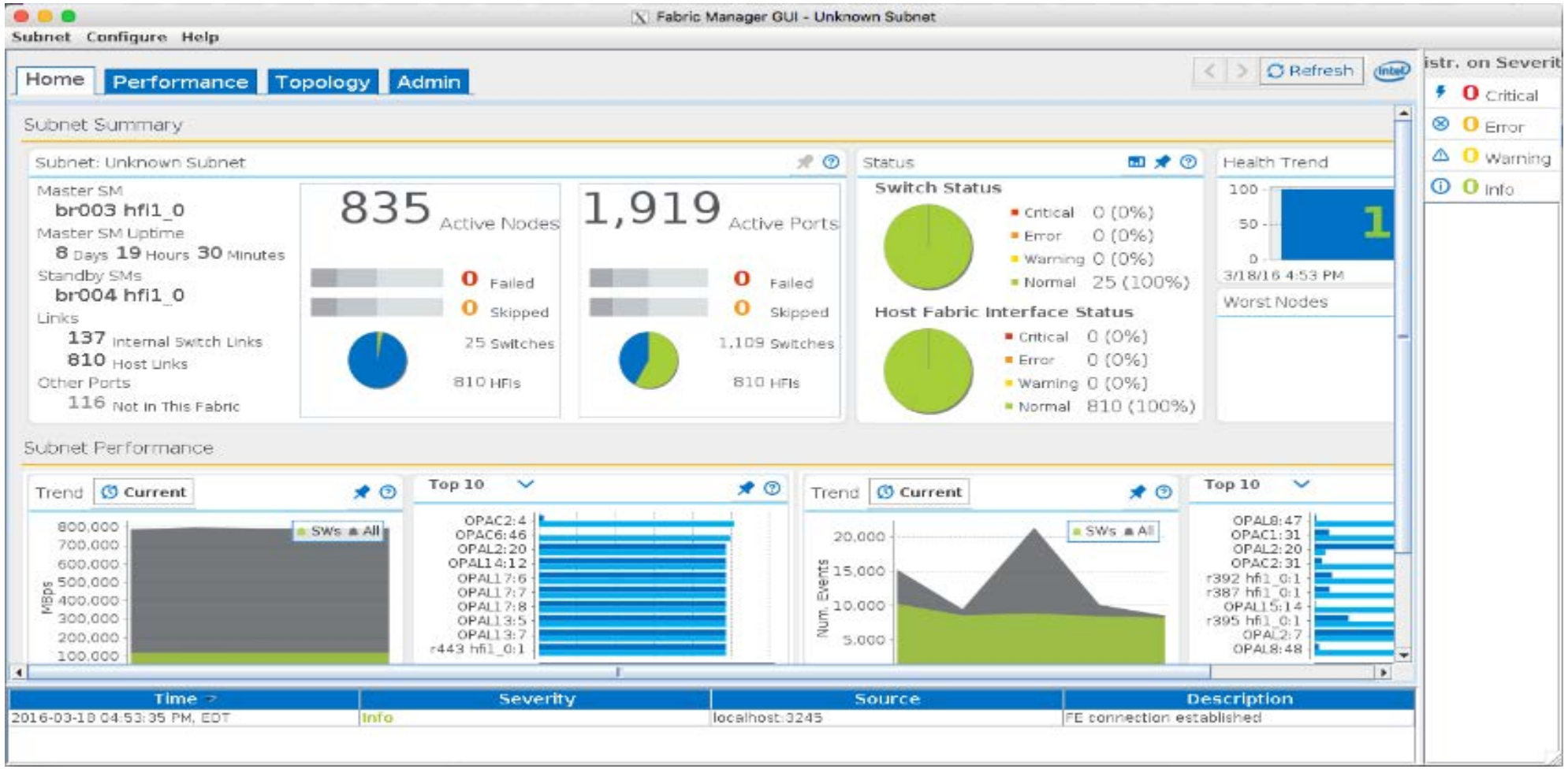
- Presentation of OPA Fabric Manager and Performance Manager data
- Concise summaries
- Click to dive deeper



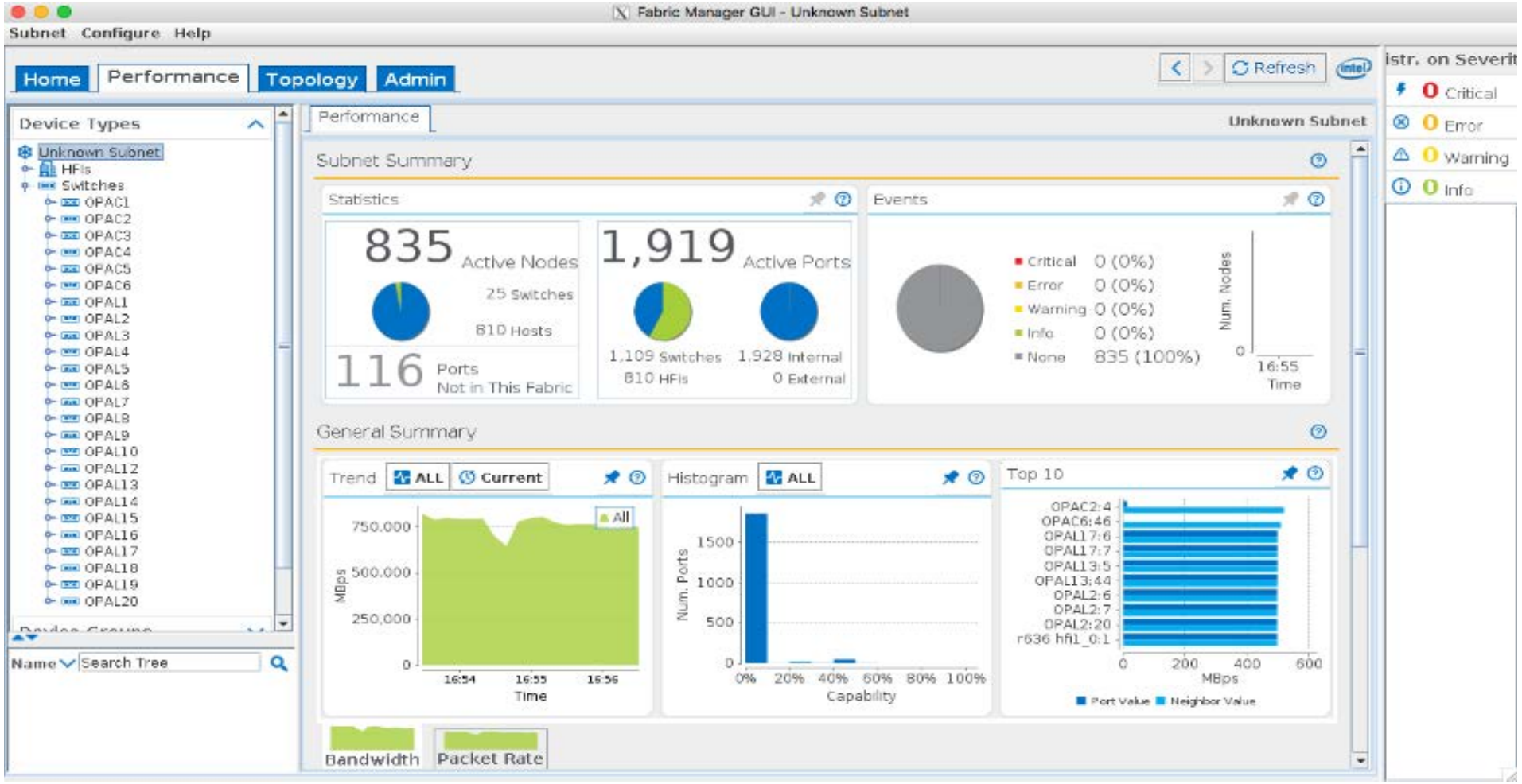
Rich, full suite of textual and TUI tools too!

PSC
PITTSBURGH SUPERCOMPUTING CENTER

OPA FM Home—Fabric status at a glance

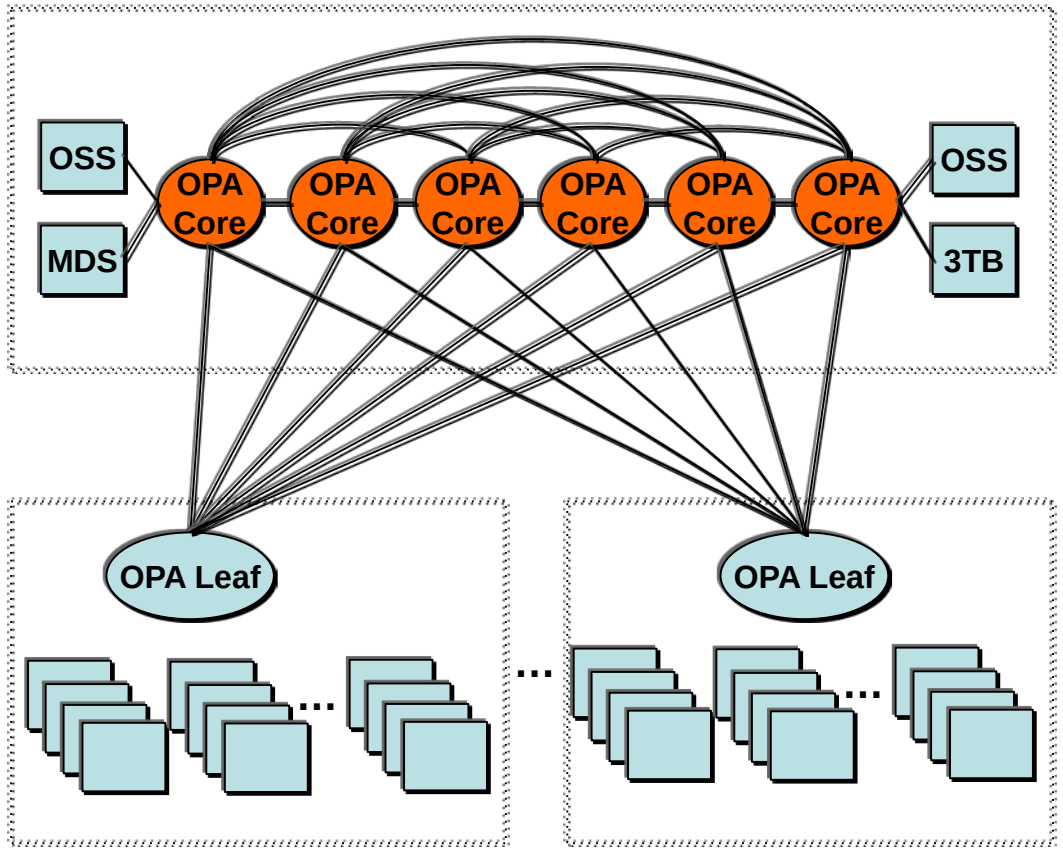


OPA FM Performance-Counter exploration tool



PSC PITTSBURGH SUPERCOMPUTING CENTER

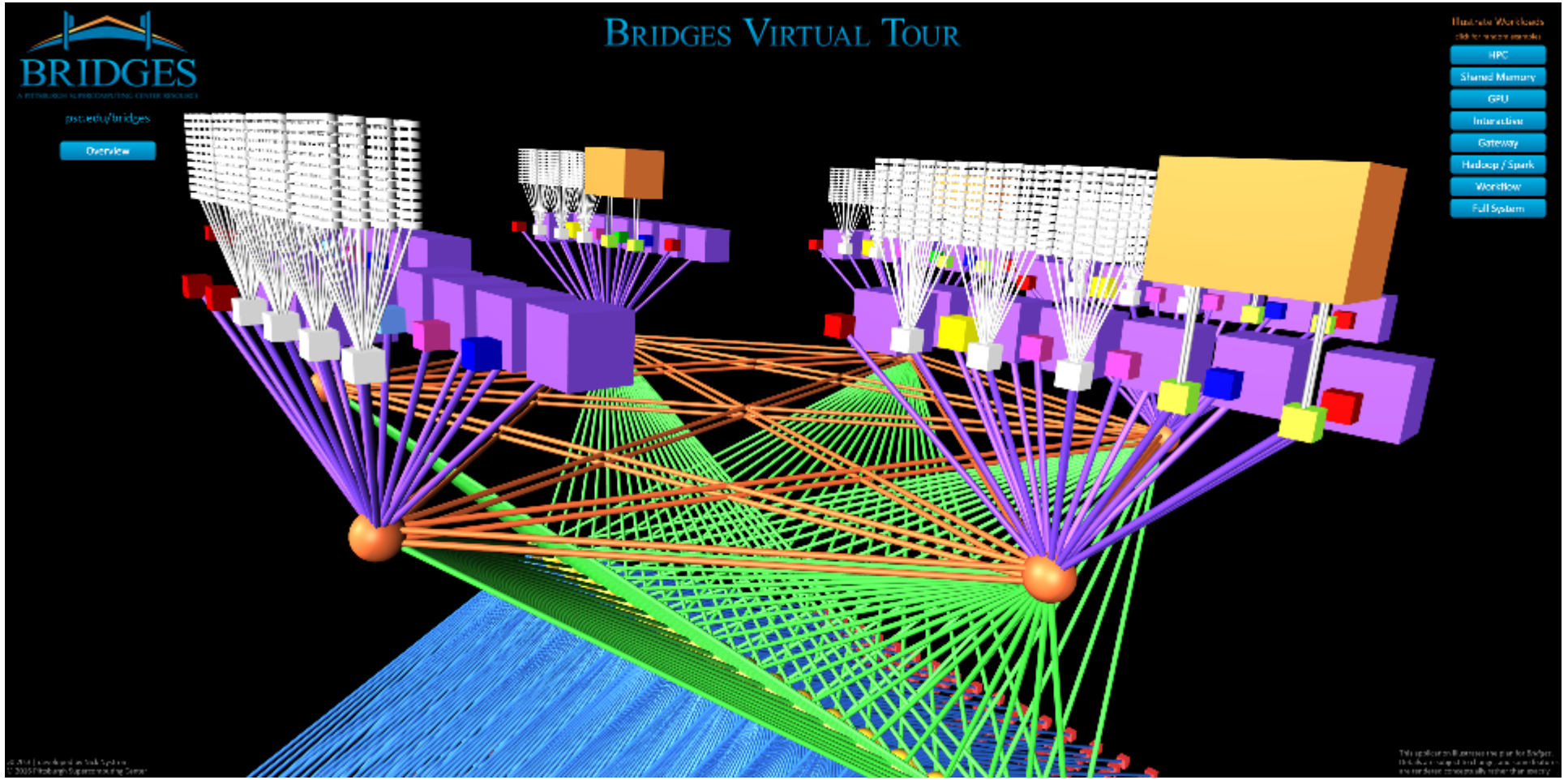
Bridges Omni-Path Topology



- 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

PSC PITTSBURGH SUPERCOMPUTING CENTER

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

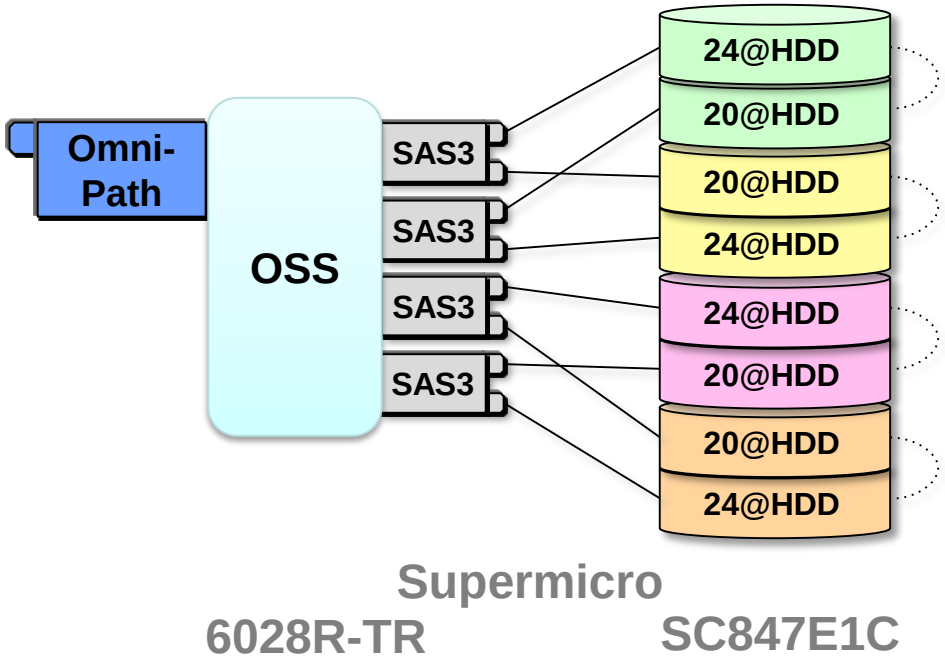


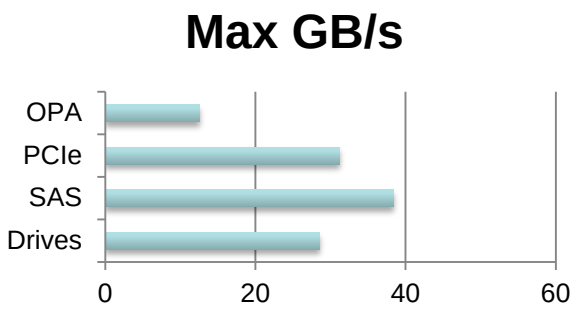
PSC
PITTSBURGH SUPERCOMPUTING CENTER

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



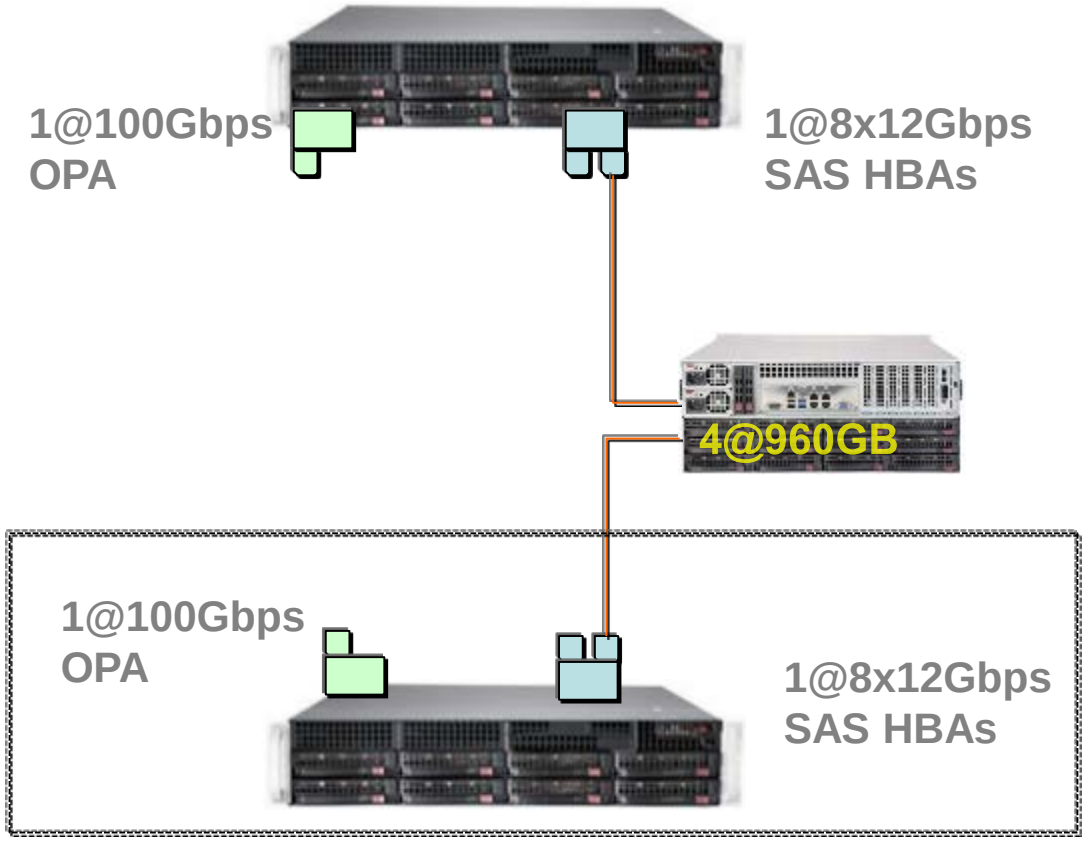


Bridges MDS Layout

Each MDS consists of:

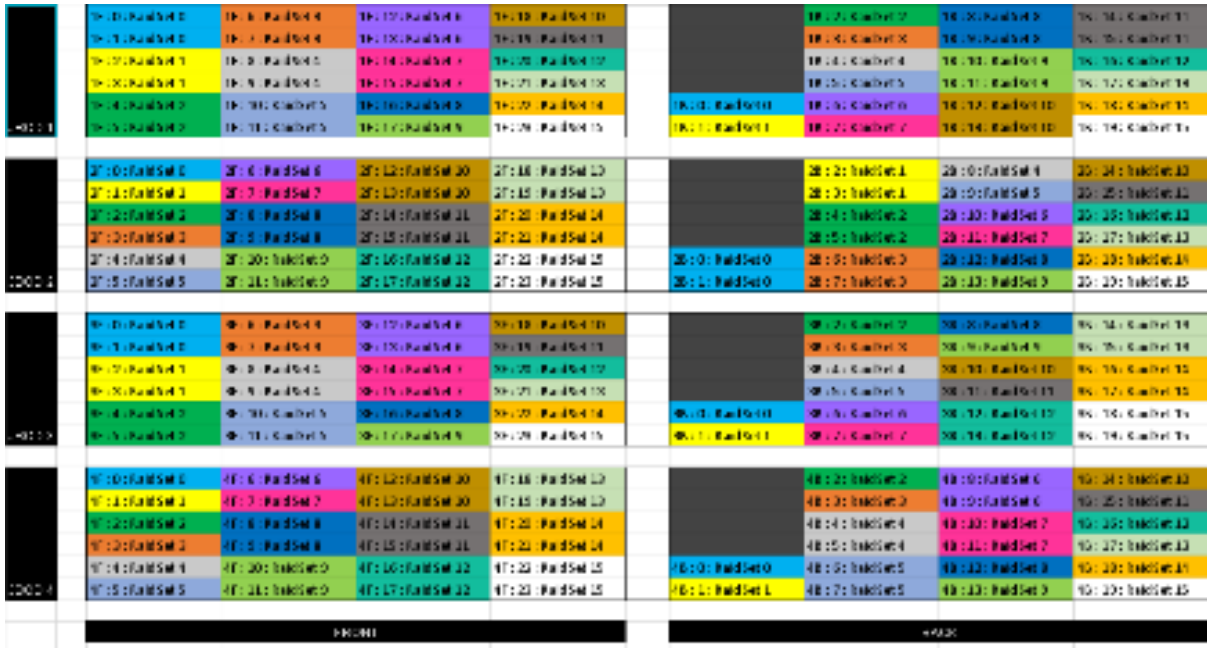
- Two E5-2695v3 (14@2.4GHz)
- 384GB RAM
- Internal disk
  - One 3TB SATA HDD (OS)
- One Omni-Path HFI (100Gbps)
  - Link to “core” OPA switch
- One 8x 12Gbps SAS3 HBAs
- One 20-bay 12Gb SAS backplanes
  - 8x 12Gbps uplink
  - 4@ 960GB SATA SSD

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



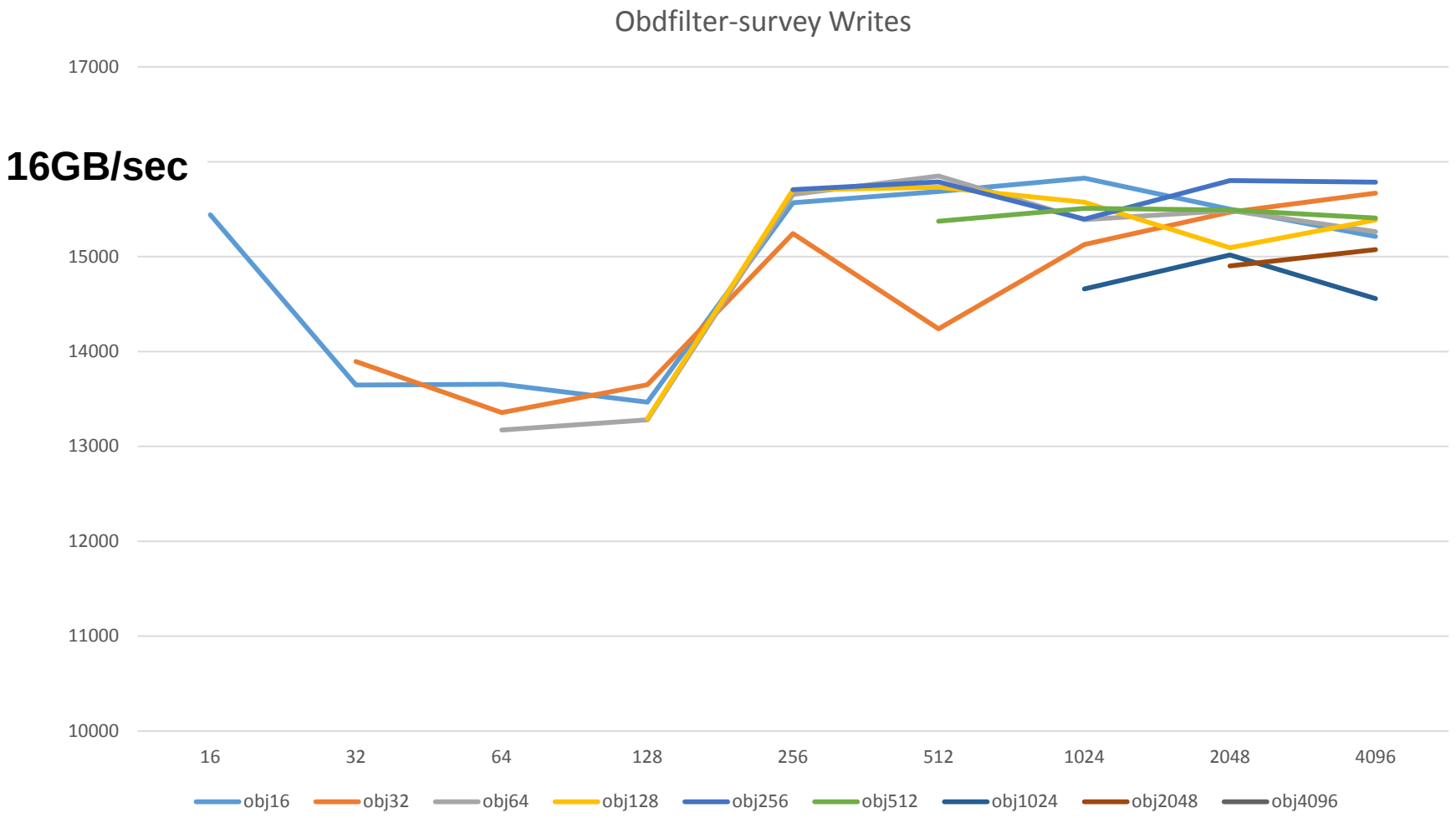
Software

- 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



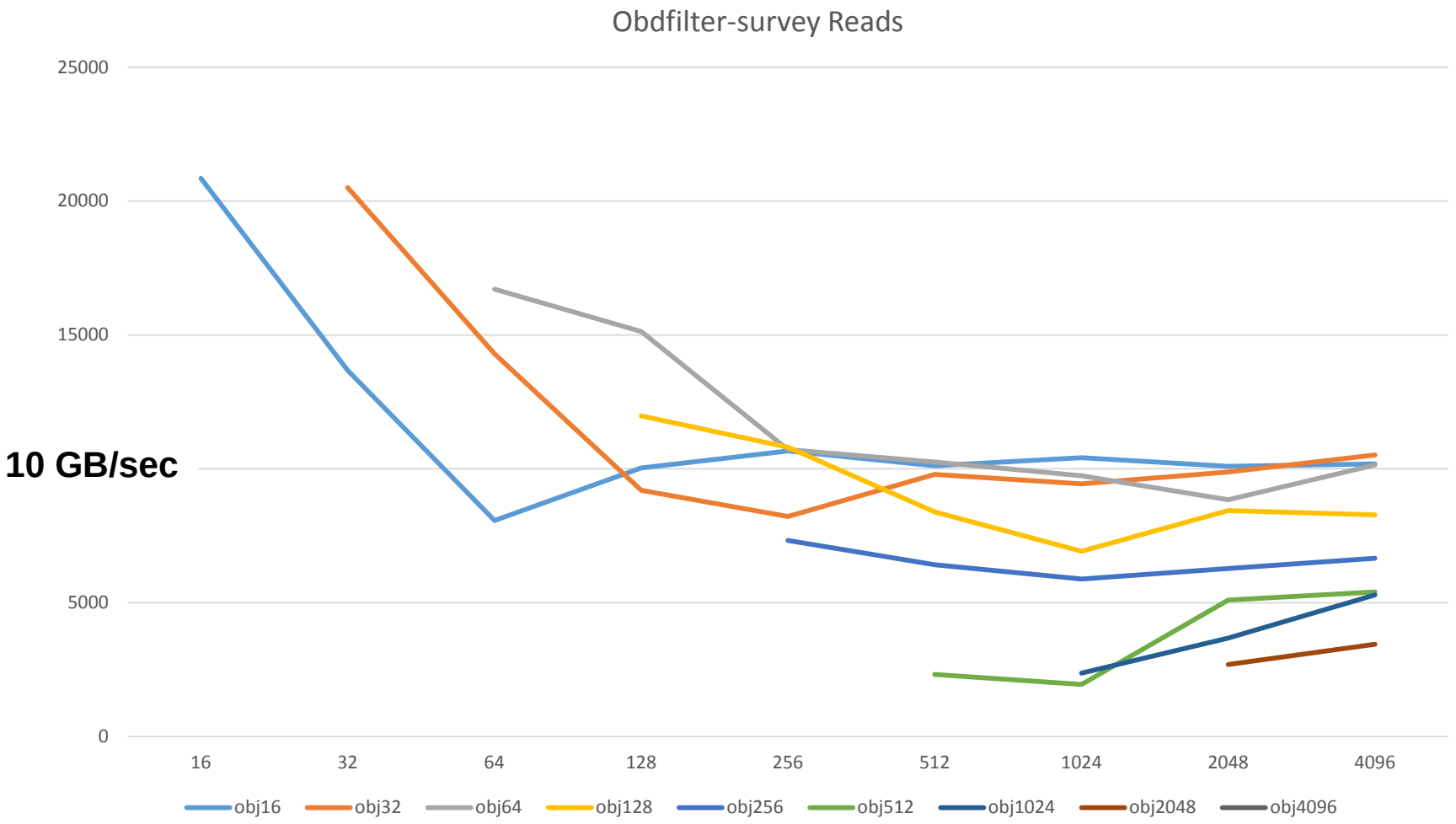
PSC PITTSBURGH SUPERCOMPUTING CENTER

OST Zpool Write Performance



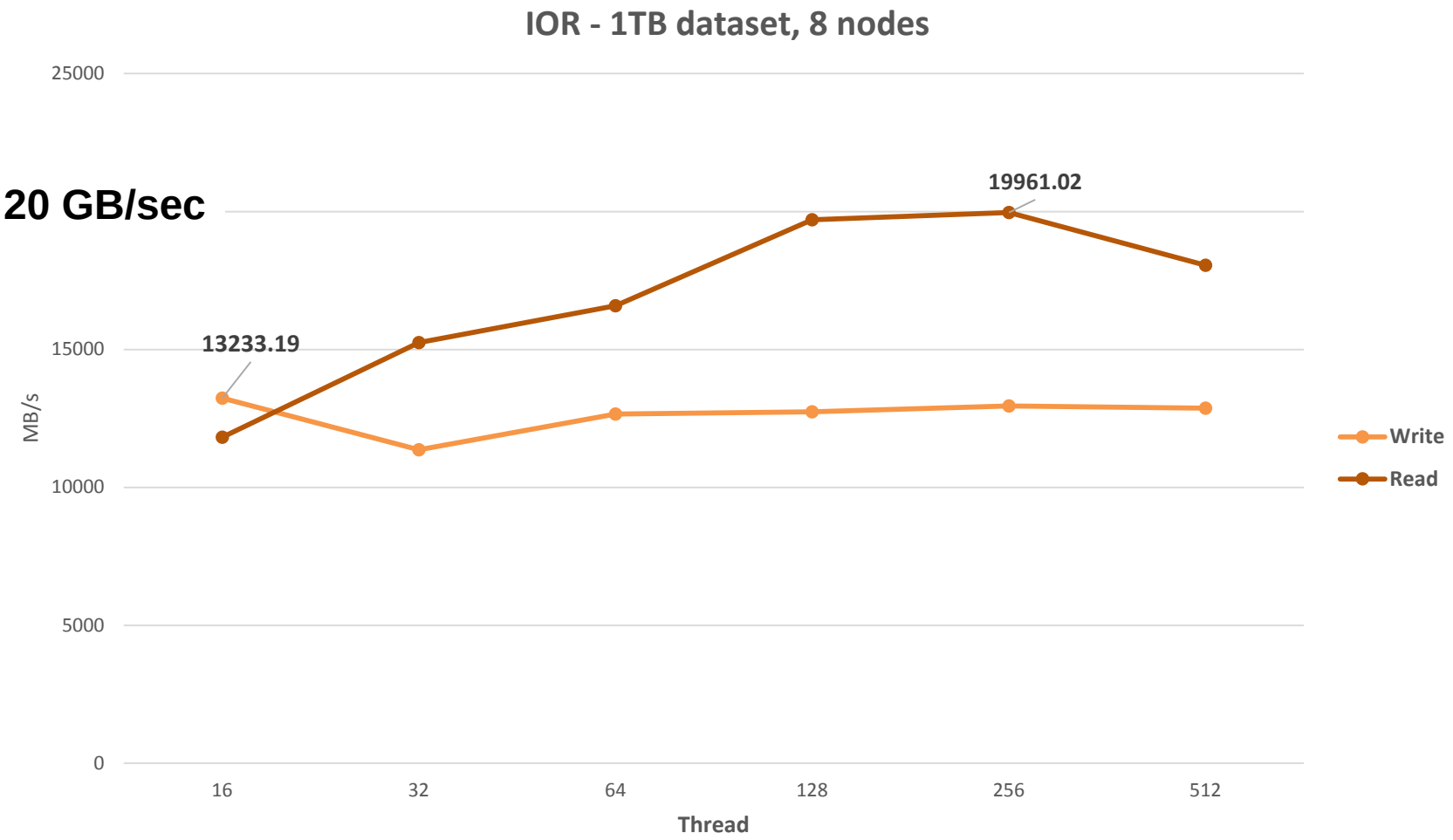
PSC
PITTSBURGH SUPERCOMPUTING CENTER

OST Zpool Read Performance



PSC
PITTSBURGH SUPERCOMPUTING CENTER

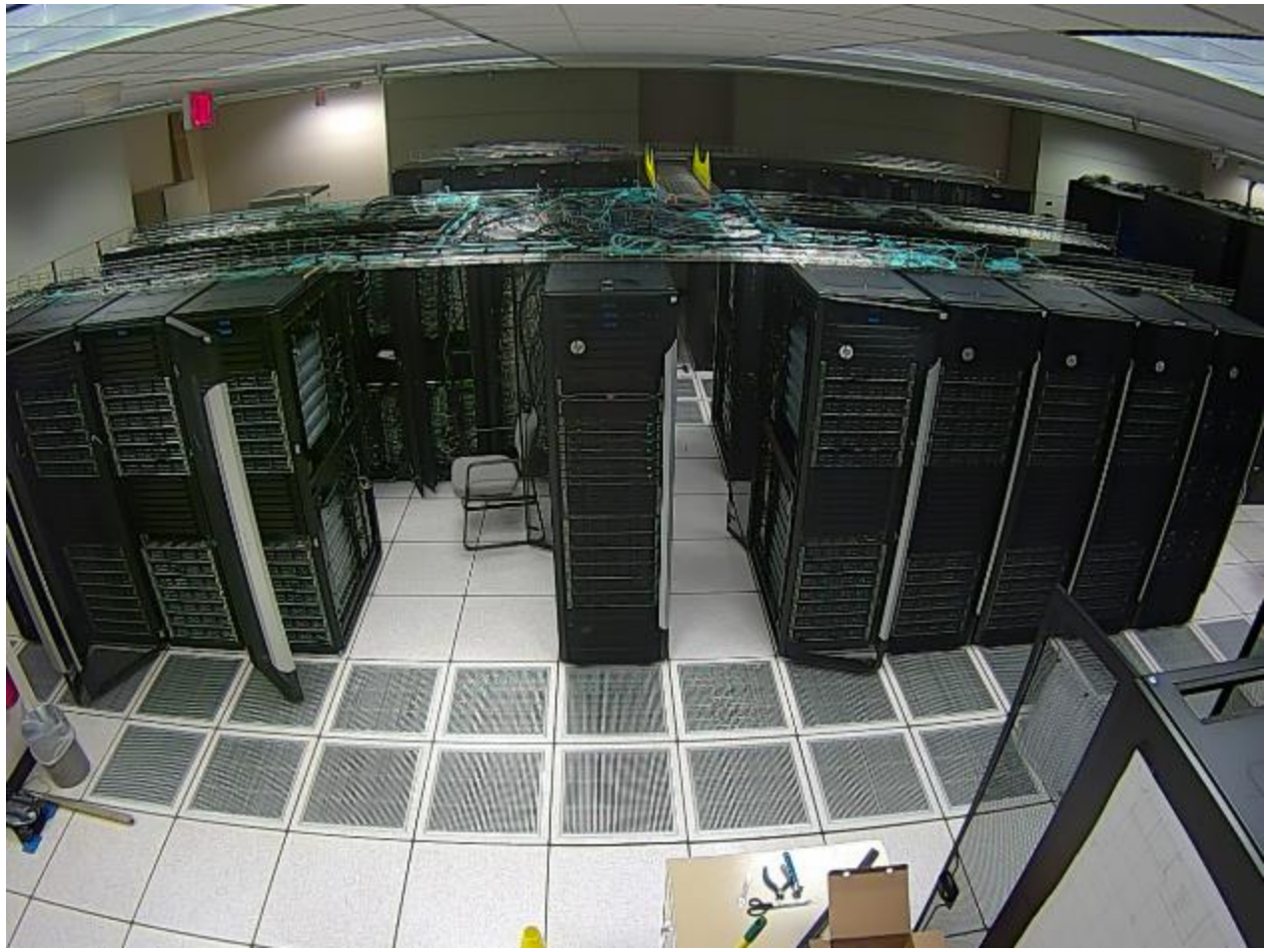
Client Performance



Credits and Contacts

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

Thank You!



PSC
PITTSBURGH SUPERCOMPUTING CENTER