

Tier2 Site Report

HEPiX Fall 2016

Caltech



High Energy Physics
hep.caltech.edu/cms



CMS Computing Team

Harvey Newman
Principal Investigator

Maria Spiropulu
Principal Investigator

Jean-Roch Vlimant
Postdoc, Development

Justas Balcas
Computing, Development

Dorian Kcira
Computing, Tier2

Azher Mughal
Networking, Tier2

Wayne Hendricks
Computing, Tier2

Josh Bendavid
Postdoc, Development

Caltech CMS Group's Hybrid Physics/Engineering Team

- A team with expertise in multiple areas, focused on R&D leading to production improvements in data and network operations
 - For the LHC experiments now and in future LHC Runs
 - Feeding into advanced planning for the next Runs, into the HL LHC era
 - New concepts: intelligent network-integrated systems; new modes of operation for DOE's Leadership HPC facilities with petabyte datasets
- ***Areas of Expertise:***

- | | |
|---|--|
| <ol style="list-style-type: none">1. State of the art long distance high throughput data transfers2. Pervasive real-time monitoring of networks and end-systems,<ul style="list-style-type: none">▪ Real-time monitoring of hundreds of XrootD data servers used by the LHC experiments supported MonALISA system3. Autonomous steering and control of large distributed systems using robust agent-based architectures4. Software defined networks with end-to-end flow control | <ol style="list-style-type: none">5. Integration of network awareness and control into the experiments' data and workflow management as in Caltech's OliMPS, ANSE and SDN NGenIA projects6. Exploration of Named Data Networking as a possible architecture replacing the current Internet, with leading NDN groups in climate science including Colorado State that has deployed an NDN testbed7. Development (as in US LHCNet) of software driven multilayer dynamic circuits and programmable optical patch panels. A virtualized connection service applicable to megadata centers and cloud computing |
|---|--|

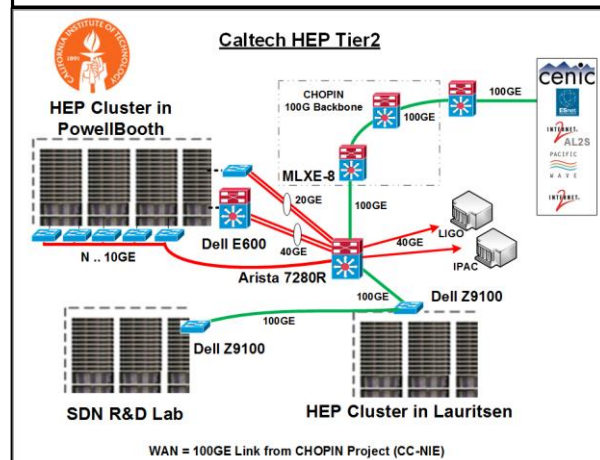
Computing and Networking at Caltech

LHC Data Production and Analysis + Next-Gen Developments

- The Caltech CMS Group, which designed and developed the 1st (MONARC) LHC Computing Model and the Tier2 Center in 1998-2001
 - Operates, maintains and develops several production- and R&D- oriented computing, storage and network facilities on campus:
 - ★ The Caltech Tier2 facility at Powell-Booth and Lauritsen Lab (the 1st of 160+ today) provides substantial and reliable computing resources to US CMS and CMS (67 kHS06 CPU, 4.8 petabyte storage) in the worldwide LHC Grid
 - ★ The group operates and develops the campus' 100G connections to nat'l and international networks through the NSF "CHOPIN" project
 - ★ Leading edge software defined network (SDN) and Data Transfer Node testbeds linked to CENIC, ESnet, Internet2, Starlight and key LHC sites



The Caltech Tier2+3 at Powell Booth

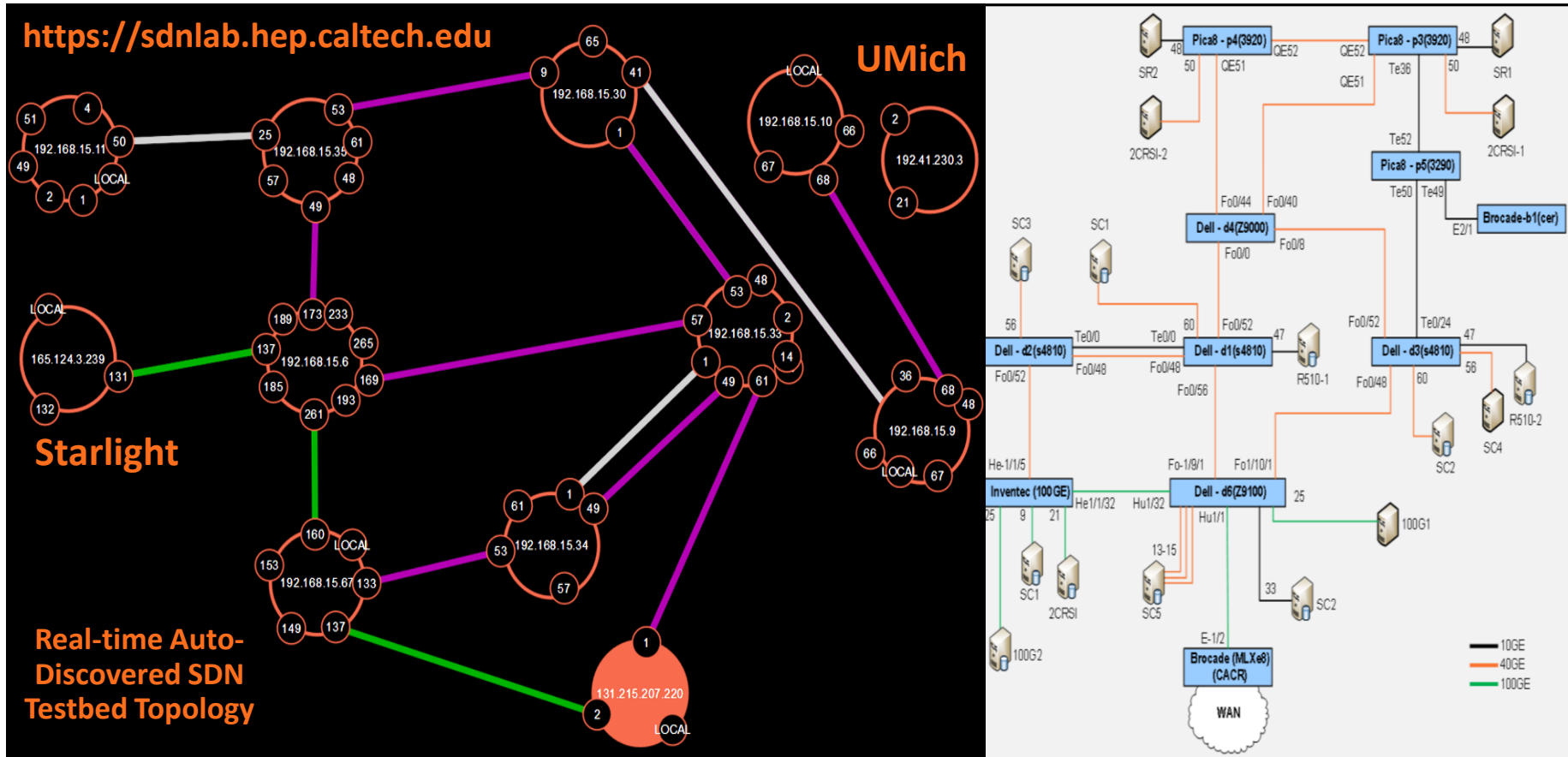


The Caltech Tier2 was the first to deploy a 100G uplink and meet the 20G+ throughput milestone; helped other US CMS groups achieve the milestone

SDN State of the Art Development Testbed

Caltech, Fermilab, StarLight, Michigan, UNESP; + CERN, Amsterdam, Korea

- 13+ Openflow switches: Dell, Pica8, Inventec, Brocade, Arista; Huawei
- Many 40G, N X 40G, 100G Servers: Dell, Supermicro, 2CRSI, Echostreams; and 40G and 100G Network Interfaces: Mellanox, QLogic
- Caltech Equipment funded through the NSF DYNES, ANSE, CHOPIN projects, and vendor donations

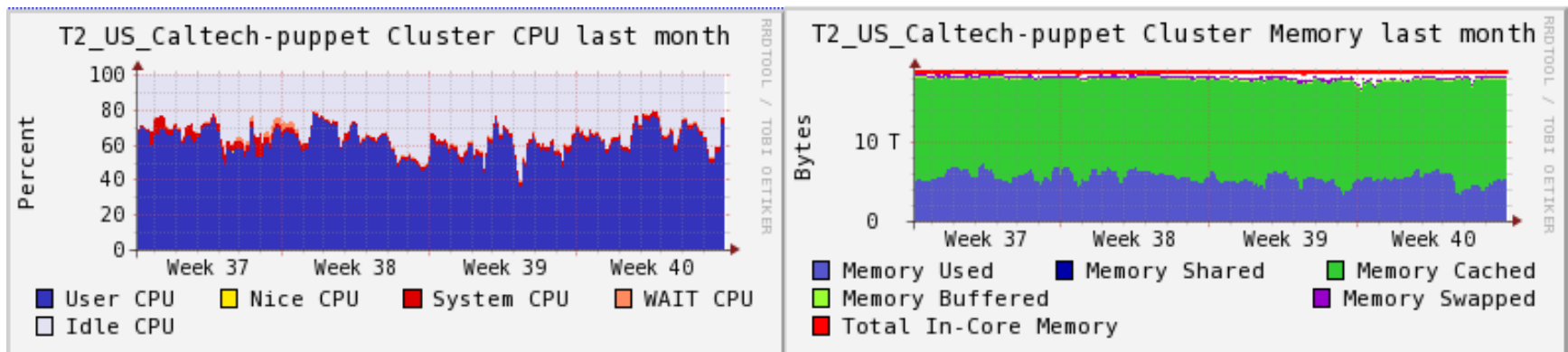


Tier2 Testbed: Targeting Improved Support for Data Federations and Data Ops

- ❑ A focus is to develop and test the Use of HTTP data federations analogous to today's AAA production infrastructure.
 - ❑ As a first step, we have built server systems that produce and ingest network data transfers up to 80 Gbps
 - ❑ In collaboration with the AAA project, HTTP support has been enabled at the US redirector and the Caltech testbed.
 - ❑ As part of the project, the Caltech group has developed a plugin for CMSSW based on Davix library for HTTP[S] access, and performance measurements tested between different protocols. More details:
 - ❑ https://cms-mgt-conferences.web.cern.ch/cms-mgt-conferences/conferences/pres_display.aspx?cid=1764&pid=13512
 - ❑ The group also is targeting improved file system performance, using well-engineered file system instances of the
 - ❑ Hadoop Distributed File System (HDFS)
 - ❑ Clarify and improve HDFS throughput
 - ❑ CEPH File System: Known for high performance; engineered with SSDs + SAS3 Storage Arrays + 40G, 100G network interfaces
 - ❑ Partnering with Samsung, HGST, 2CRSI
- ❑ We are setting up multi-GPU systems for machine learning with support from Orange Labs Silicon Valley and Echostreams

CMS Tier2 Stats

- Around 7000 cores/threads contributing to CMS
- Around 300 compute nodes and 20 storage datanodes in 18 racks
- Compute and storage is mostly combined
- Most network connections are 1Gb, with a portion of large datanodes on 10Gb
- Disk sizes range from 1TB to 8TB with a mix of enterprise and commodity drives
- Use of virtual machines limited to testing and experimental hosts
- 100Gb link between Powell-Booth Lab and Lauritsen Lab



Current Hardware Deployed

Most of our current hardware is in these two configurations. We also have some older 1U machines due to be retired in the next few months.

Supermicro Twinpro
Combined Compute and Storage



Supermicro 4U Storage Chassis



Tier2 Improvements

- The move to CentOS7 is in progress
- A Ganeti cluster will be implemented to virtualize most core services
- Bestman will be retired and replaced by LVS to balance CentOS7 gridftps
- Replacing Ganglia with Telegraf/InfluxDB to make metrics exportable and compatible with other tools
- Working with UCSD to implement an XrootD cache system for hot datasets between Caltech & UCSD
- HTTP as a Data Access Protocol for CMS data over XrootD
- CephFS build and testing is underway

Current Tier2 Challenges

- There are drawbacks to combining compute and storage at scale
- Distributed filesystems such as HDFS work best when most nodes in the cluster have similar transfer capabilities
- Upgrading network or storage capabilities for the entire cluster can be expensive or impossible with combined hardware
- A split into two classes of nodes makes sense for us - each focused on a specific operation, fast data transfer or compute

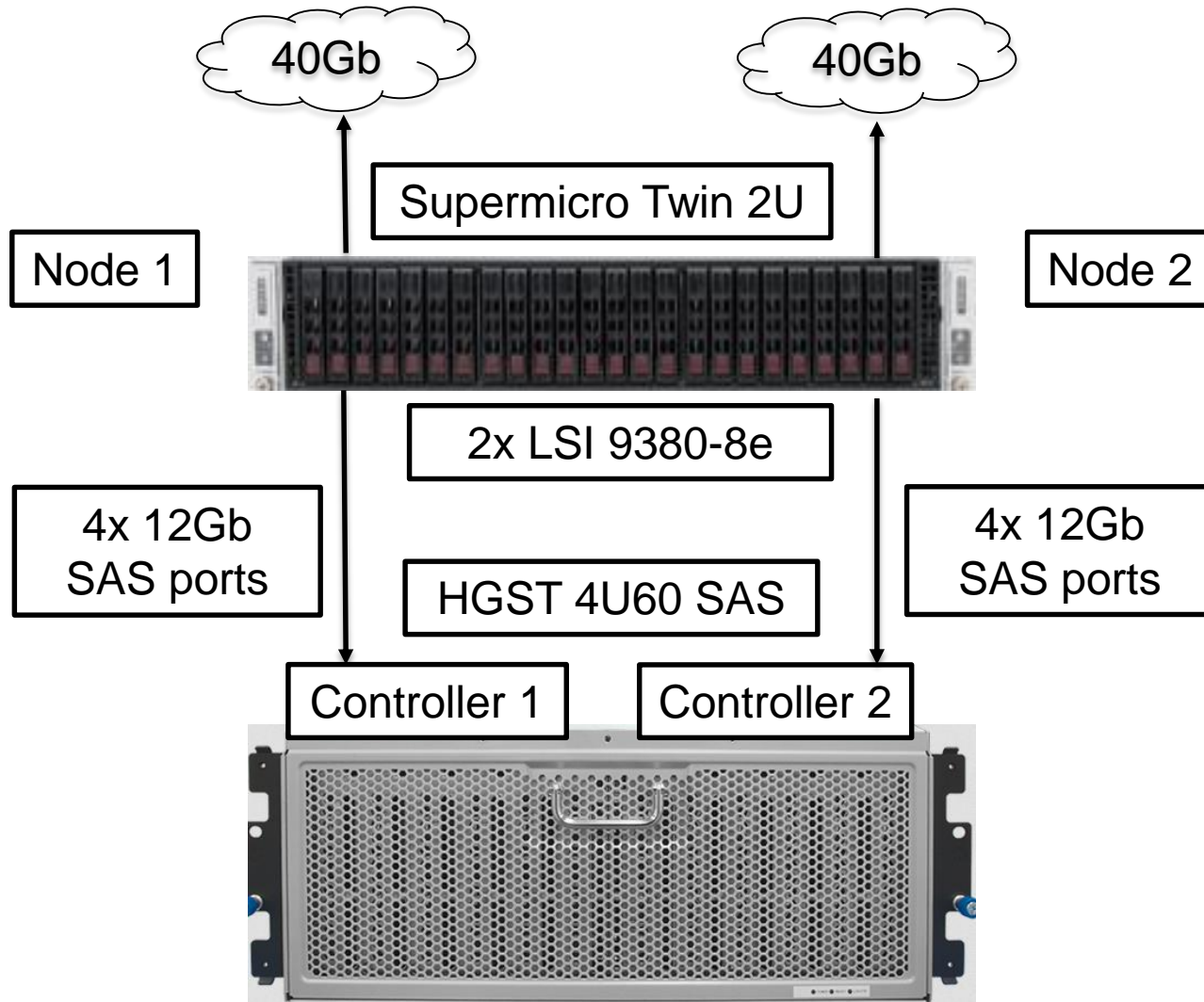
Benefits of Host Classes

- Two classes of hosts will allow us to allocate funds directly for our needs
- No risk of compute jobs or data transfer processes interfering with each other
- We can tune each class of host specifically for the operation it performs
- Newer blade systems and dedicated storage systems require less power and rackspace than our current hardware
- Network cabling can be reduced with only high speed links between chassis

Storage

- Commodity disks may be cheaper initially, but various quality issues we have experienced negate this in the long run
- Cheaper backplanes also suffer from the same issues with quality and reliability
- HGST has an attractively priced 4U60 backplane with 60 fast reliable disks
- 5 year warranty on disks and backplane
- Can connect to two hosts to two onboard controllers and each serve half the disks
- Once a critical mass of high performance datanodes is reached, we will be able to retire all of the remaining commodity disks
- We will retire the oldest enterprise disks gradually and move completely to the dedicated storage chassis

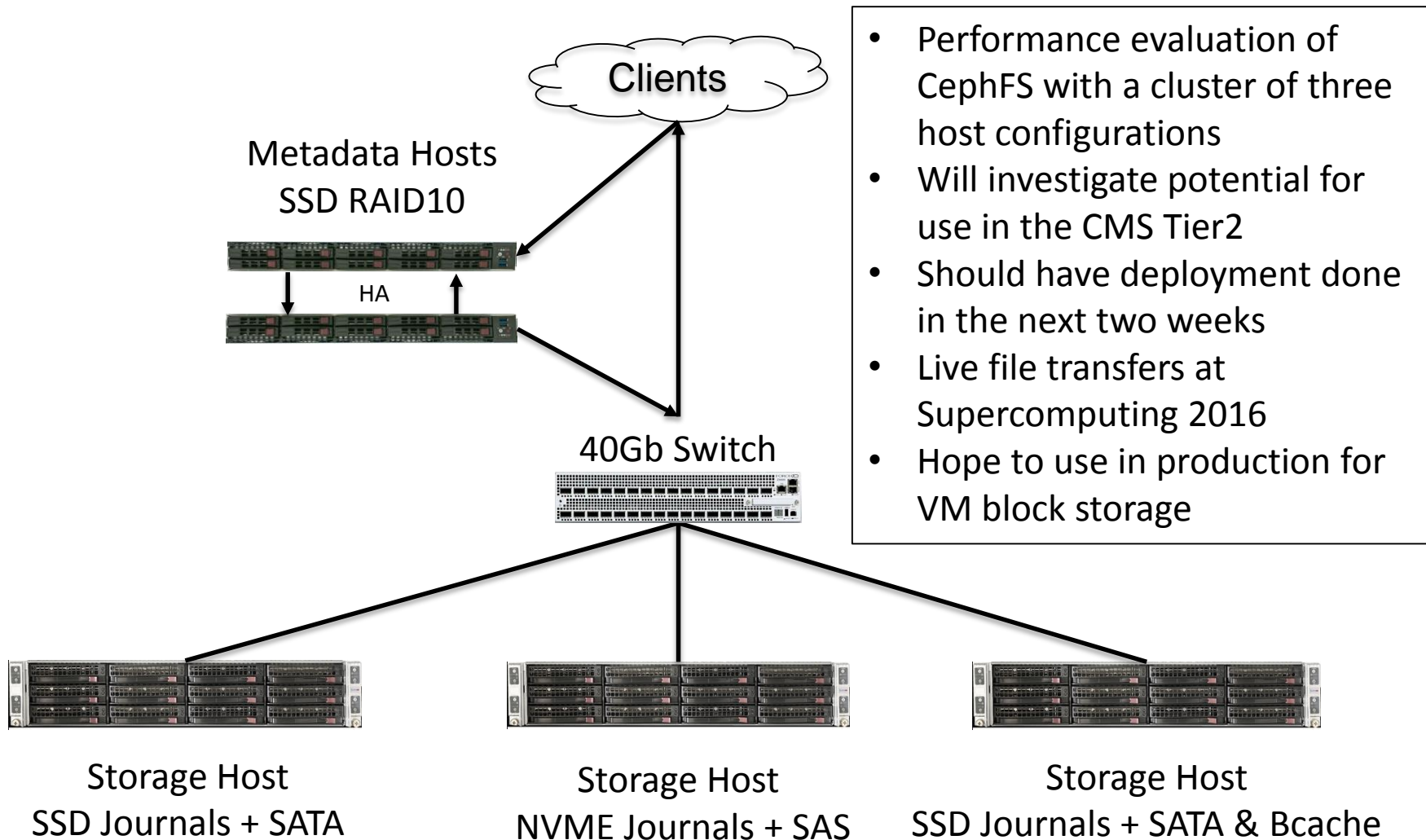
Prototype Storage Nodes



60 6TB SAS disks, 30 served by each node

Tier2 Update - HEPiX Fall 2016

CephFS Test Configuration



Compute

- With the next round of purchasing, all 4 and 8 core nodes will be retired
- By deploying storage separately, we can focus on density and efficiency
- Supermicro has a 6U microblade chassis with 28 nodes (1300 E5-2640v4 threads)
- Redundant 10Gb network links in the backplane for all nodes
- Intel switch module has 4x40Gb uplinks and also has SDN capability
- Each node can support two 2.5” 6Gb SATA SSDs, so local job IO is fast

Compute Chassis

Supermicro X11 Microblade Chassis



Intel Switch Module



Results

- Limited rackspace and power consumption is markedly reduced
- Equivalent Twinpro PSU wattage 21000W, vs 6U Microblade 16000W
- Data transfer speeds between the storage nodes is greatly increased
- Limits our risk of exposure during a hardware failure by ensuring data is replicated very quickly between high speed storage nodes

Stop by our booth at
Supercomputing 2016 next month
to see live high-speed data transfer
demonstrations

Thank You

caltech.edu