HEPiX 2018 - Purdue Tier-2 Site Report

May 14, 2018

Stefan Piperov

spiperov@purdue.edu

for the Purdue Tier-2 Team:
Majid Arabgol, Erik Gough,
Thomas Hacker, Norbert Neumeister

PURDUE UNIVERSITY | Research Computing INFORMATION TECHNOLOGY

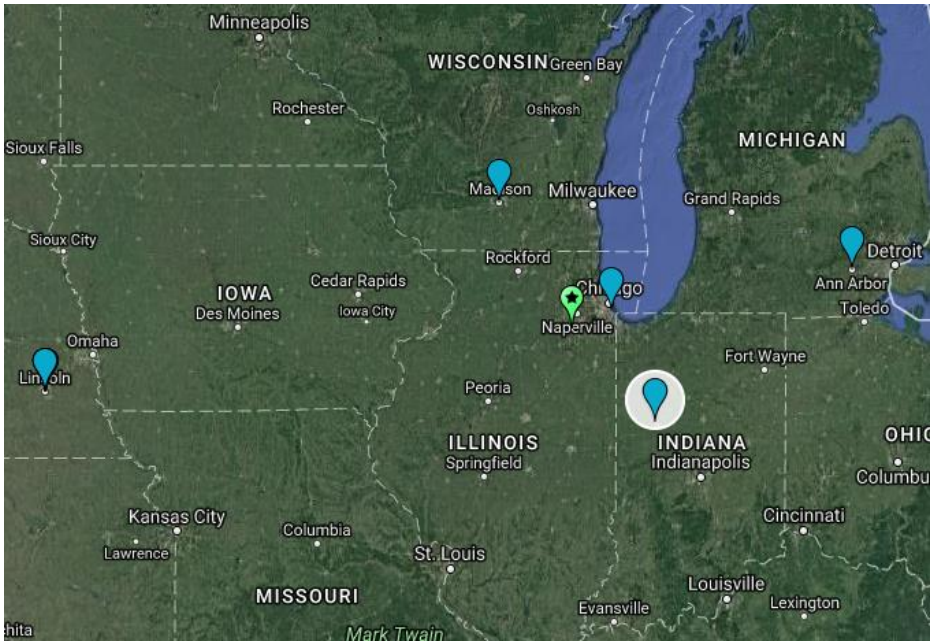
OUTLINE

- **Purdue CMS T2 Overview**
- **Current Status of Resources**
  - Compute
  - Storage
  - Networking
- **2017 Activities**
- **2018 and Beyond, Looking to the Future**

Purdue CMS T2 overview

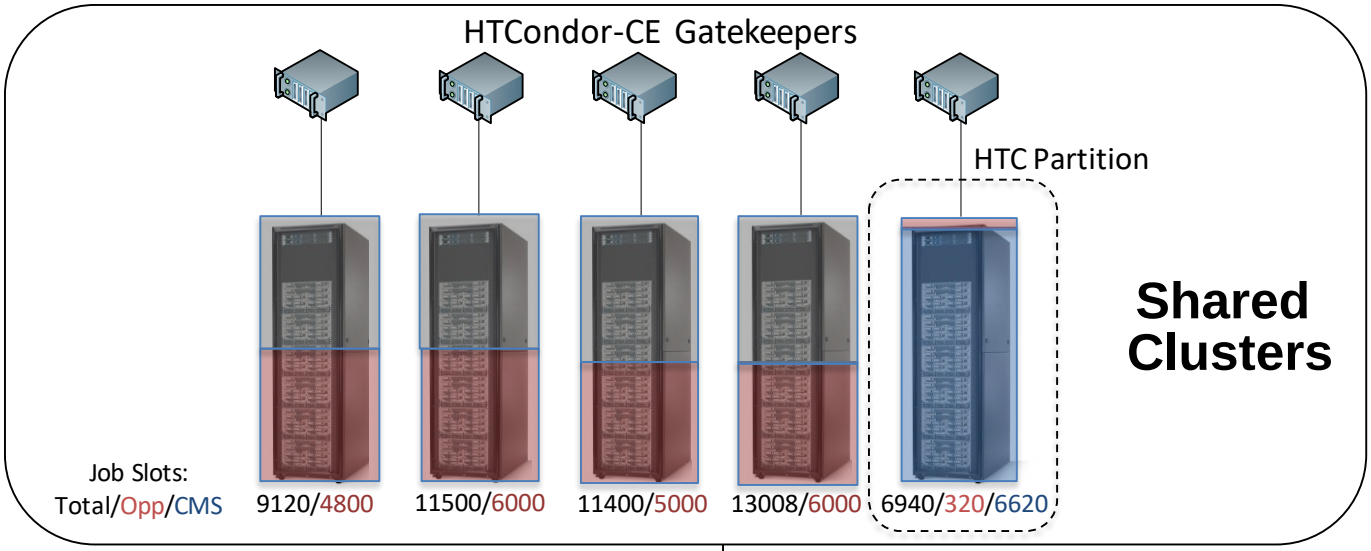
- **Purdue University hosts one of the two WLCG Tier-2 centers in Indiana.**

- **Together with Wisconsin and Nebraska we form the 'mid-west' core which surrounds the Tier-1 at FermiLab, and provides for a large fraction of the US-CMS computing operations.**



PURDUE UNIVERSITY

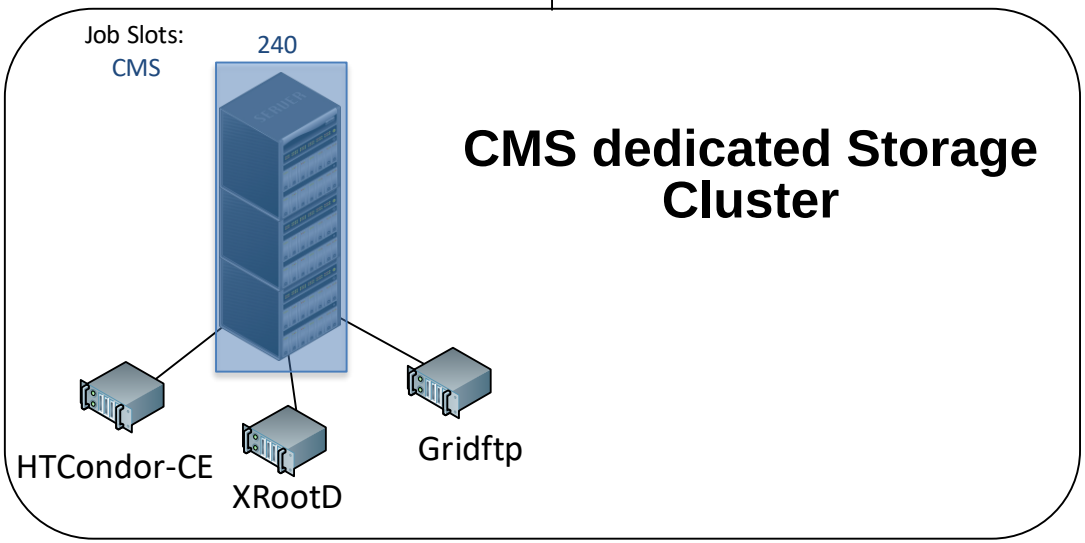
PURDUE T2 ARCHITECTURE



High Speed Network

6,860 CMS dedicated job slots

22,120 Opportunistic job slots



PURDUE UNIVERSITY

SITE RESOURCES - COMPUTE

- **Purdue CMS dedicated compute resources fall into two groups**
  - **CMS dedicated cluster**
    - Provides a limited number of batch slots via HTCondor
    - HTCondor-CE gatekeeper receives pilots with 48h walltime
    - HDFS storage for CMS
    - File access protocols (XRootD/Gridftp)
  - **CMS owned Community Cluster nodes**
    - Provides dedicated batch slots via PBS
    - A single HTCondor-CE gatekeeper receives pilots with 48h walltime for all CMS owned nodes

PURDUE UNIVERSITY

COMMUNITY CLUSTER PROGRAM

- **Purdue Community Cluster Program began in 2008**
  - A Community Cluster resource is built every year
  - Each cluster has a lifespan of 5 years
  - Each cluster runs the PBS scheduler
  - Managed by Purdue central IT department (ITaP)
- **Benefits for Purdue and CMS**
  - Peace of Mind: ITaP system administrators provide system support so faculty and graduate students can focus on research
  - Low Overhead: ITaP provides all infrastructure
  - Cost Effective: Leverage group purchasing power
  - Opportunistic: Compute nodes are shared among partners when nodes are idle

PURDUE UNIVERSITY

DEDICATED COMPUTE RESOURCES

Cluster	Purchase Year	Inter connect	Scheduler	Node Type	#CMS Nodes	Slots/Node	Job Slots	HS06	# Years Left
Hadoop	2013 2014	10Gbps	Condor	SuperMicro Dell	15	16	240	3929	1
Community Cluster	2015	10Gbps	PBS	HP DL60	161	20	3220	57789	2
	2016	25Gbps	PBS	HP XL170	55	40(HT)	2200	26345	4
	2017	25Gbps	PBS	Dell PE640	25	48(HT)	1200	16875	5
						Total Job Slots	**6860**	**104938**	

PURDUE UNIVERSITY

OPPORTUNISTIC RESOURCES

- **Purdue CMS has opportunistic access to all Community Clusters via PBS standby queues**
  - Requires four additional HTCondor-CE gatekeepers, each receives 4h pilots with no chance for eviction
  - Opp. Slots: the total number of non-CMS slots on each cluster
  - Max Slots: the maximum number of usable standby slots per cluster for CMS at any given time

Cluster	Enabled	Mem/Slot	Opp. Slots	Max Slots
Conte	6/15/16	4GB	9,120	4800
Rice	8/10/16	3GB	11,500	6000
Hammer	9/7/16	3GB	320	320
Halstead	2/24/17	6GB	11,400	5000
Brown	3/1/18	4GB	13,008	6000
		Totals	45,348	22,120

PURDUE UNIVERSITY

OPPORTUNISTIC RESOURCES

- **Purdue significantly increased its opportunistic utilization in 2017.**
- **We provided more than the yearly compute pledge of a USCMS T2 - for free! (~290 million HS06 hours*)**
- **And we had the potential to provide an additional ~170 million HS06 hours*, (or 65% of a USCMS T2).**

Opportunistic Slots (4h)	2016	2017
Glidein Core Hours	.9 M hours	19 M hours
Payload Core Hours	.3 M hours	7.8 M hours
# Completed Jobs	.17 M jobs	2 M jobs

} 41% CPU efficiency

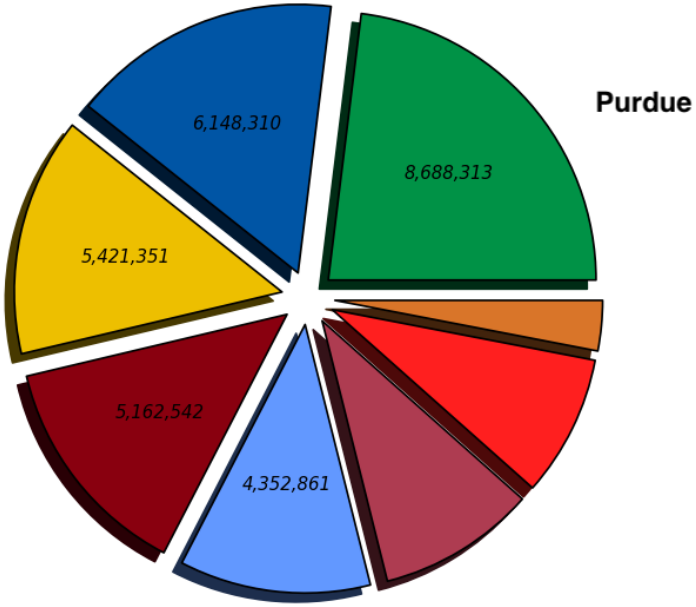
* Calculated using 15.3 HS06/slot

PURDUE UNIVERSITY

2017 – PRODUCTION/ANALYSIS JOBS

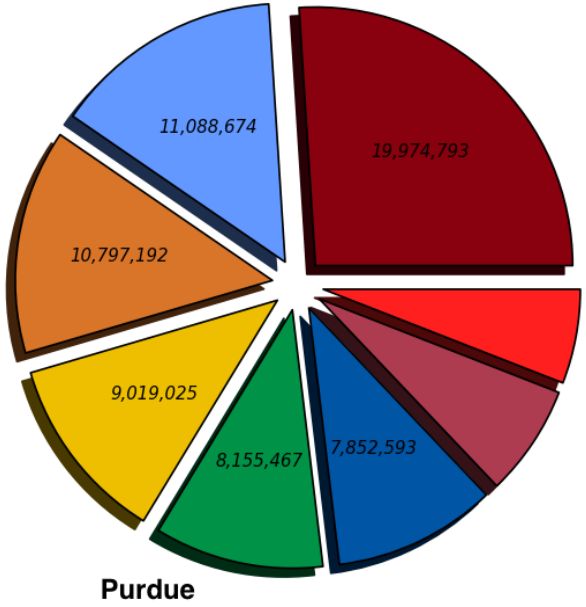
Job counts for US-CMS Tier 2 sites

Production: 23%



- 38.0M total jobs completed
- 8.7M jobs completed by Purdue
- Avg. WallClockHr: 1.65

Analysis: 11%



- 77M total jobs completed
- 8.2M jobs completed by Purdue
- Avg. WallClockHr: 1.40

PURDUE UNIVERSITY

SITE RESOURCES - STORAGE

- Home and group storage (GPFS), provided by Purdue's central IT
- Hadoop Distributed Filesystem (HDFS) - 3.8PB usable
  - Mixture of 2U Compute/Storage and 4U pure storage Hadoop datanodes*.
  - 36 Enterprise class hard drives, 10G NICs, 12Gbps HBAs
  - Over time, the disk size has grown (2TB->3TB->4TB->6TB->8TB)
- Data access services run directly on storage servers
  - Gridftp – 35 servers; XRootD – 40 servers

* Due to the success of the Community Cluster program, our HDFS storage has transitioned from being part of a mixed Compute/Storage module to one of (almost) pure storage

PURDUE UNIVERSITY

SITE RESOURCES - NETWORKING

- **Wide Area Networking**
  - WAN link speed: 100 Gb/sec
  - Purdue Connectivity to FNAL
    - Dedicated 100G link from Purdue to Indiana GigaPop Indianapolis
    - Shared 200G link from Indianapolis to Chicago
    - Shared 200G link on ESNet to FNAL and LHCONE sites
- **Data Center Core and Local Area Networking**
  - Community Clusters to Research Core: 80G to 400G
  - CMS Storage to Research Core: 240G (recently upgraded from 160G)

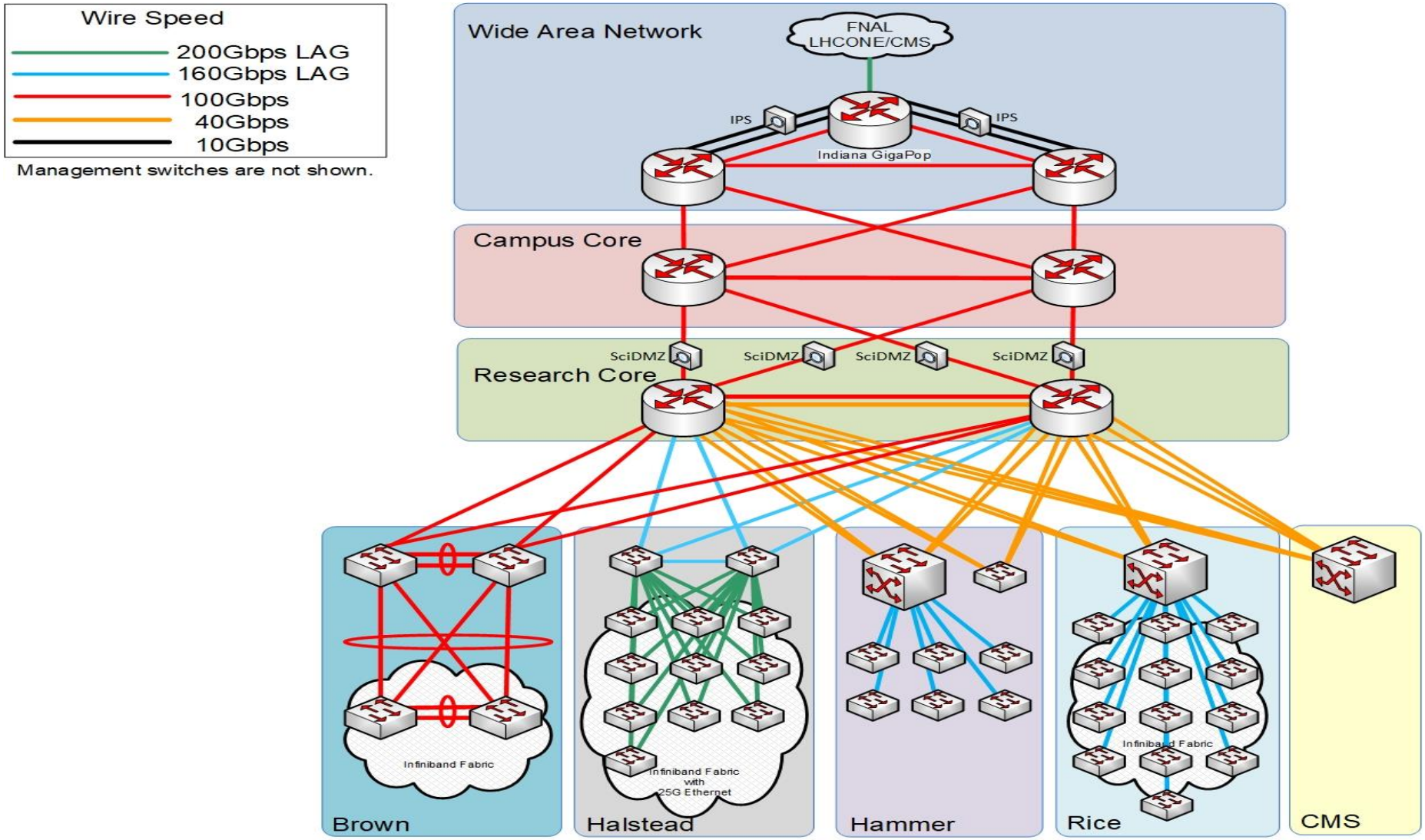
Network	Capacity	Observed Peak Utilization – 2017
WAN	100G*	~85G (CMS LHCONE traffic only)
CMS Storage	240G	160G

- **In 2017, we started to see bottlenecks and device failures**
  - WAN
  - CMS storage uplinks

* shared with all research network traffic

PURDUE UNIVERSITY

SITE RESOURCES – NETWORKING 2018



PURDUE UNIVERSITY

UPGRADES AND DEPLOYMENTS

- **CentOS 7 Upgrades**
  - Upgrades are currently underway
    - Storage nodes
    - Compute nodes
    - Service nodes
  - Includes migration to new puppet and xCAT environment
  - Includes switch to lcmaps-voms and GUMS retirement

USER SUPPORT

- **Users get prompt response in case of any problem related to utilization of the resources at our site**
- **User Access**
  - 50 users with local UI access
  - 100 grid users with /store/user access
- **Provide support for our local CMS heavy-ion group**
- **Analysis datasets from different physics groups are managed through Dynamic Data Management. However, some dedicated storage space is allotted for different physics groups at the site.**

OPERATIONS

- **Very few infrastructure related outages in 2017**
  - One power blip caused HDFS storage to go offline for ~3 hours
  - Periodic WAN outages
    - High WAN throughput caused hashing errors on 100G WAN board
    - Currently, WAN is rate limited to 75G with 5G burst
    - Will be resolved during WAN upgrades in 2018
- **Saturated CMS storage uplink bandwidth (160G)**
- **Difficult to find sources of network utilization**
  - Can I easily find out how many local jobs are reading from remote storage? And vice versa
- **Kernel patching is becoming very common**
  - Stack Clash, ELF/PIE, Spectre/Meltdown
  - Most were done via rolling reboots, not much harm
- **Unclaimed glideins on opportunistic resources**

LOOKING TO FUTURE DEMANDS

- **The need for compute and storage will grow significantly after LS2**
- **We took a look at our current infrastructure and facilities and asked, How big can we get?**
- **Datacenter**
  - Purdue T2 resources are part of Purdue's primary research computing datacenter (and PODs)
  - CMS does not pay for power, cooling and rack space!
  - Efficient use of our allocated space is crucial to our growth
- **Networking**
  - CMS historically has not paid for internal networking components
  - Purdue central IT (ITaP) has been excellent at addressing any concerns about internal bandwidth to T2 resources

PURDUE UNIVERSITY

LOOKING TO FUTURE DEMANDS

- **Compute**
  - Purdue compute resources currently take 250U of rack space
  - We could expand our compute another 250U to 500U
  - At 675 HS06/U we could increase our site capacity by an additional 170k HS06
- **Storage**
  - Purdue storage resources currently allocated 11 racks with 36U usable per rack
  - We have 124U available for immediate expansion
  - Using dense storage servers (~1PB/4U) we could grow HDFS storage by a significant amount (~15 PB usable) before exhausting rack space
- **Use of super-computer HPC centers and commercial cloud resources**

PURDUE UNIVERSITY

PURDUE CMS HARDWARE

PURDUE UNIVERSITY

SUMMARY

- **Purdue remains stable and reliable**
  - Availability, Reliability – **99%, 99%***
  - Site Readiness – **97%***
- **We are a cost effective computing resource and will continue to participate in the Community Cluster program**
- **We made a significant contribution to completion of CMS production and analysis jobs in 2017**
- **Opportunistic access to Community Clusters works well, but limited to 4 hour pilots.**
- **Our computing facilities allow for considerable growth as demands for compute and storage increase in the future.**

* cumulative 2017

PURDUE UNIVERSITY