Campus Cyberinfrastructure Plan

Background

Since the 1960s, Purdue University has operated central computing infrastructure in support of research. In 2001, this research computing organization was combined with administrative IT to form Information Technology at Purdue (ITaP). Reporting to the Purdue System CIO, IT Research Computing has operated large-scale research supercomputers, Purdue’s “community clusters,” since 2004.

Today, ITaP operates community clusters, research storage, and high-speed networks, and provides cyberinfrastructure expertise for Purdue researchers.

Vision

- To be the one-stop provider of choice for research computing and data services at Purdue.
- To deliver powerful, reliable, easy-to-use, service-oriented computing to Purdue researchers.

General Principles

- Regular investment in current-generation cyberinfrastructure.
- Empower faculty to control their resources (through interface development and deployment).

Computation

For 10 years, Purdue has operated a world-class Community Cluster Program, each year deploying an HPC system on the order of 10,000 cores, with approximately 50 faculty groups investing in every system. Purdue was awarded the 2010 Campus Technology Innovators Awards\(^1\) for its community cluster program.

In total, over 200 faculty groups have invested in the program, with over 1,000 active users. ITaP and faculty are partners, with ITaP centrally funding administrative and support staff, infrastructure for the clusters, networking, and storage. Conte, the fall 2013 system, currently ranks #190 on the TOP500 list and debuted at #28. The Community Cluster Program continues to be the core of Purdue’s cyberinfrastructure strategy.

Beginning in 2015, the model has evolved to one oriented around the key research communities at Purdue, rather than a single, one-size-fits-all system. Each community cluster deployment targets a specific community:

- High-Performance Computing for Physical Sciences and Engineering with emphasis on parallel computation (Rice, Halstead, Brown)
- Data Intensive, Large Memory for Life Sciences (Snyder)
- Machine Learning and Accelerated Computing (Halstead-GPU, Brown-GPU)
- Interactive, Non-Batch Computing (Data Workbench)
- Secure Computing for ITAR, CUI, and classified research (EXRC, REED)

This model also will accommodate partnerships with communities whose work may not be geared to traditional Linux-based clusters, such as non-UNIX users, data analytics, rendering, or reconfigurable clouds.

Finally, control of a faculty member’s community cluster resources is fully in their hands. ITaP engineers provide easy-to-use Web interfaces to purchase capacity, allocate access, and report on usage, all with the click of a button.

Data

Since the 1990s, Purdue has provided the Fortress archive system to all researchers at the University at no cost. This large-scale archive has grown to over 3 PB of research data.

Beginning in 2013-2014, Purdue’s cyberinfrastructure focus has been on data. Each HPC system includes a parallel scratch filesystem of at least 1.5PB, with initial per-user quotas of 100 TB per user. The 2017 Brown Cluster provides a parallel filesystem with over 3 PB of capacity, and 200 TB per user quotas.

In fall 2014, Purdue’s Research Data Depot entered production, providing a highly redundant, highly reliable 2.5 PB of storage for purchase. Building upon this institutional investment, faculty can purchase storage per TB per year for actively used and shared datasets and applications, and other uses. Over 460 researchers are partners in the Research Data Depot.

The Research Data Depot is well suited as a storage target for instruments, or for collaborative data sharing using the resource’s Globus DTN.

In this data-driven research era, new demands for accessing and working with data arise regularly. To help meet this challenge, Purdue provides a freely available data analytics platform based on Hadoop, MapReduce, Spark, and various NoSQL engines.

In 2017, ITaP piloted a database engine-driven platform to support Big Data mining. This system will allow faculty in management, economics, hospitality, and other disciplines easy access to database systems to research and teach students techniques in real-world data analytics problems.

Finally, ITaP and the Purdue University Libraries jointly develop and maintain the Purdue University Research Repository (PURR) for creating data management plans, sharing data with collaborators, and publishing and describing finished datasets. ITaP and Libraries personnel are engaged with Purdue researchers to train them on best practices for managing and working with their research data.

Networking

As a founding partner in iLight, the state of Indiana’s high-speed optical fiber network, Purdue has WAN connections of 100 gigabits to Indianapolis. A Purdue-funded 2014 network upgrade saw improvements of research networking capabilities up to 160 Gb to the Research Data Depot, and 400Gb to each computing resource.

As a long-time Large Hadron Collider (LHC) partner (a CMS Tier-2 center\(^2\)), Purdue has infrastructure in place to monitor network performance with perfSONAR, and fully supports IPv6 in a dual-stack mode to selected computing resources and data transfer nodes.

Since 2015, key labs with large-scale data transfer and acquisition needs have been piloting a layer-2 extension of the Research Science DMZ out into the campus, to better enable transferring large amounts of data to the Research Data Depot and community clusters.

Infrastructure

\(^2\) https://www.physics.purdue.edu/Tier2/
Most research computing assets reside in Purdue University’s Mathematical Sciences Building on the West Lafayette campus. The building, built in 1966 and expanded in 1982, has received multiple renovations to its 6,900 square feet of data center space. Most recently, an NSF ARI award provided funding for significant upgrades to power and cooling, providing 2.5 MW of power and 350 tons of cooling capacity. Some resources, including the Fortress archive and the secondary site for the Research Data Depot are located in a data center in Haas Hall, a 4,750 square foot space providing a total of 450 kW of electrical power and 148 tons of cooling capacity.

**Education, Outreach and Training**

On campus, Purdue research computing staff facilitate research by regular engagement with faculty and graduate students. *Coffee Hour Consultation* is a popular open-format office hour for researchers to interact with research computing staff. Research computing computational scientists also deliver a wide variety of training in UNIX, effective use of clusters, parallel programming, visualization, and productivity with data tools. Locally developed instruction is supplemented by material from XSEDE, Software Carpentry, and national CI experts.

Since 2007, Research Computing staff have led teams of undergraduate students in cluster competitions at Supercomputing, ISC, and ASC events. Research Computing staff member Stephen Harrell has served as chair for the SC16 and SC17 cluster competitions.

Purdue instructors regularly use CI resources in support of their courses, from disciplines as varied as Computer Science, Aeronautical Engineering, Animal Science, and Chemistry.

Since 2017, ITaP leads a Women in HPC campus organization, with the objective to expose and encourage women in the Purdue community to pursue research and careers in HPC and technology fields.

To develop Cyberinfrastructure practitioners, ITaP operates a long-running program for developing students from hardware technicians, to junior administrators or facilitators, and eventually hired into full-time roles. Since 2005, the program has seen over 60 students, with 24% continuing into a research computing career, either at Purdue, another institution, or industry.

**Cybersecurity**

The community cluster program supports best-practice cybersecurity practices including two-factor authentication, host firewalls, and storage permission configurations that default to private. ITaP is currently deploying an NSF-funded project to monitor the research network with a high-performance passive intrusion detection system.

ITaP operates a central computing and data storage system built to support research governed by International Traffic in Arms Regulations (ITAR) regulations. This system has supported many faculty research groups’ ITAR computing and storage needs.

To comply with regulations spelled out in DFARS 7012, Purdue has designed and currently operates the "REED" computing environment for controlled unclassified information (CUI). Today, REED supports 5 sponsored projects, and is one of the first university-operated systems built in to support DFARS-7012 research. REED is designed to comply with the controls described in NIST 800-171.

Purdue is a cleared institution, and operates high-performance computing as well as desktop computing resources to support classified research projects.

**Collaboration**
Purdue is active in national communities of campus cyberinfrastructure practitioners. From 2004-2013, Purdue was a Teragrid (and later XSEDE) resource provider, and has been a U.S. CMS Tier-2 site since 2005. Purdue is currently a partner in the NSF XSEDE project. Purdue staff chaired the XSEDE SP forum and leads XSEDE’s Campus Champions program.

Purdue researchers are prominent members of international virtual organizations, like the LHC’s Compact Muon Solenoid (CMS) experiment, and the Large Synoptic Survey Telescope (LSST). Purdue resources are integrated into these and other experiments’ computing and data infrastructures.

Purdue leads the HUBzero Foundation, providing a state-of-the-art, ready-made cyberinfrastructure and online collaboration platform for research. Hubs power such communities as NanoHUB.org, nees.org, PURR, and the Indiana CTSI.

Purdue is a member of InCommon, using the federated identity infrastructure to easily allow Purdue researchers to access national resources like XSEDE, the Open Science Grid (OSG), or Globus.

Purdue is a member of CASC, Software Carpentry, and is a founding partner for the annual HPC syspros workshop at Supercomputing.