

Purdue University Campus Cyberinfrastructure Plan

Comments in the NSF “CI2030: Future Advanced Cyberinfrastructure” report describe “*An integrated cyberinfrastructure that reaches from university and college campuses to the national centers is needed; this will require coordinated investments by all of the stakeholders....*”

*... the vast majority of capacity-class computing activities will be carried out **on campuses**.*

*... communication and collaboration across research silos offers the possibility of building and deploying an **integrated cyberinfrastructure** that effectively and efficiently supports a broad range of scientific and engineering research.*

Research reveals that “*modest, but **consistent investment** in HPC results in measurable increases in research competitiveness, and a corresponding increase in research funding and publication counts.*” and “*investment in high performance computing as measured by entries on the Top 500 list is a good predictor of research competitiveness*”

In this context, Purdue University operates a campus cyberinfrastructure under the following vision and principles:

Vision

- To be the one-stop provider of choice for research computing and data services at Purdue
 - Delivering cost-effective, powerful, reliable, easy-to-use computing and data services to Purdue researchers.
 - “*High Performance Computing at the Highest Proven Value*”

General Principles

- Governed jointly with the faculty – “*A coalition of the willing*”:
- **Consistent investment** annually in current-generation cyberinfrastructure resources
- Delivering a broadly-usable cyberinfrastructure to a wide cross-section of scientific disciplines.
- With self-service interfaces to empower faculty to control their resources

Institutional Cyberinfrastructure 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 ITaP (Information Technology at Purdue). Today, as part of what is now named **Purdue IT** and reporting to the Purdue System CIO, Purdue’s Rosen Center for Advanced Computing (RCAC) has operated large-scale “community clusters” since 2004.

Today, RCAC operates community clusters, research storage, science gateways, a visualization facility, high-speed networks, and provides cyberinfrastructure expertise for Purdue researchers.

Governance

Research computing at Purdue is governed by RCAC leadership and the faculty partners. Twice annually, governance meetings are held – one meeting open to the entirety of the cluster community, and one smaller meeting of the governance committee consisting of staff from the Rosen Center, and the Office of Research, academic IT, and top faculty partners. These meetings are used to present the state of the program, discuss architecture decisions and technology deployments, and to gather requirements for upcoming research computing needs by the campus.

Areas of Focus (2021-2024)

- Computing to support Data Science, AI, and Machine Learning
 - Robust accelerated HPC platforms
 - Composable computing infrastructures
 - Create an upgrade path for growing from laptop to HPC
- Support Application of Quantum Information Science (QIS)
- Provide robust instructional resources for HPC and Data Science education
- Grow Research Software Engineering offerings to enable more Purdue research
- Secure computing to support regulated research (export control and HIPAA), national defense, and aerospace

Computation

The core of Purdue's cyberinfrastructure strategy, Purdue has operated since 2004 a world-class community cluster program - each year deploying an HPC system on the order of 550 nodes, with approximately 70 faculty groups investing in each system. In total, over 200 faculty groups from 60 academic departments and every academic college invest funds in the program, with over 1200 active users. Purdue IT and the faculty **share the capital costs** of the program, with faculty purchasing compute nodes with grant or institutional funds; and Purdue IT centrally funding staff, shared network and storage infrastructure.

Data

Since the 1990s, Purdue has provided the Fortress archive system to all researchers at Purdue, at no cost to the individual researcher. This large-scale archive has grown to hold over 40PB of research data as of Summer 2023.

In fall 2014, the Research Data Depot entered production, providing a highly-redundant, highly-reliable 7 PB storage service available for purchase by research labs. Building upon an institutional investment, faculty can invest funds to purchase TBs of storage per year, and utilize high-capability storage for actively-used, shared datasets, applications, or other uses. Over 600 labs have purchased over 4 PB of storage in the Data Depot. Providing an on-ramp to computing for data-only researchers remains a strategic goal.

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

Science Gateways

Science gateways are online platforms that provide access to advanced resources for researchers, educators, and students. *“Through streamlined, online, user-friendly interfaces, gateways combine a variety of [cyberinfrastructure](#) (CI) components in support of a community-specific set of tools, applications, and data collections.* Purdue’s RCAC provides several services to support science gateways, including hosting gateways on appropriate cyberinfrastructure, designing and developing gateways, and providing technical support and training to gateway developers and users.

MyGeoHub (<http://mygeohub.org>) is a geospatial science gateway that specializes in support for geospatial data access, processing, analysis, and visualization. It was established to support integrative, geospatial, data-driven research activities and broad dissemination of the outcomes, and currently hosts more than 10 research projects and more than 10,000 annual users. MyGeoHub currently supports 360 accessible resources including courses and training materials, 55 tools, and more than 15,500 simulations run annually. Students and instructors from more than 30 U.S. and international higher education and research institutions have used resources on this site in their classes in the past 12 months.

The NanoHUB (award # 1227110) provides a library of 3,000 learning resources to 195,000 users worldwide, each year. Its 232 simulation tools, free from the limitations of running software locally, are used in the cloud by over 10,800 users annually.

Visualization

The Envision Center for Data Perceptualization provides novel solutions to effectively communicate complex research concepts through computer graphics, advanced visualization, auditory (sound), haptic (touch), and multimodal interaction integrated with state-of-the-art high-performance computation. The Envision Center is staffed by a project manager, a technical lead, developers, and students to assist researchers, instructors, and leaders in their quest for new knowledge and innovative products. Recent and ongoing research supported by center staff includes projects in the life sciences, audiology, engineering, communications, theatre, computer science, geology, and management. The Envision center facility provides facilities for motion capture, augmented and virtual reality, and a large visualization wall for interacting with large datasets.

Expertise

RCAC staff provide domain expertise, technology solutions, software development to Purdue researchers. Staff offer open office hours around campus, training sessions in data science and cluster usage, and instruct in academic courses. Expert staff are available to partner with faculty on projects, and provide cyberinfrastructure/computational expertise.

Networking

All Purdue centrally managed compute and storage systems are connected to a central research network core in the Purdue Math Building Data Center at 400 Gbps. This central research core connects to Purdue’s campus core at 400 Gbps, which then connect to the internet edge wide-area network routers at 400 Gbps. At the campus peering point for research networks, Purdue operates an intrusion detection system funded by an NSF CICI award (# 1738981). Purdue maintains a 100

Gbps connection to the Indiana GigaPOP in Indianapolis, Indiana, with an additional 100 Gbps path (shared with Indiana University) to Chicago, Illinois. From there, systems can access a variety of research networks, including the NSF-supported ACCESS network, at 100 Gbps.

As a long-time LHC site (a CMS Tier-2), Purdue has infrastructure in place to monitor network performance with PerfSONAR, and supports IPv6 in a dual-stack mode, to selected computing resources and data transfer nodes.

In 2018, Purdue was awarded an NSF CC* networking grant (# 1827184) to add high-speed Science DMZ connections and broaden PerfSONAR to selected big data facilities for large science data flows to both the campus research cyberinfrastructure and off campus facilities.

Physical Infrastructure

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 datacenter space. Most recently, an NSF ARI award provided funding for significant upgrades to power and cooling, providing 2.5 MW of power and 525 tons of cooling capacity.

Some resources, including the secondary site for the Fortress archive and the Research Data Depot are located in a datacenter in Felix Haas Hall. Built in 1908 and remodeled in 1984, Haas hall provides 1985 square feet of space, 400 KW of power, and 60 tons of cooling capacity.

Cybersecurity

Purdue has an established security operations team with defined monitoring and incident response processes and procedures. Logs from across the institution are collected and analyzed automatically to identify potential malicious activity. These alerts are evaluated by an IT security team that responds to events and escalates incidents as appropriate. The Purdue system-wide security team is a founding member of the OmniSOC, a shared SOC for higher education.

The Purdue system security team has an established vulnerability management program for assessing risks. Vulnerabilities are evaluated to confirm their presence on Purdue systems, determine if the systems are internet-facing, understand the ease of exploitation, and correlate against adversary activity. A vulnerability report is distributed to system owners each month for remediation. This approach provides relevant insight into adversary interest as it relates to vulnerability exposure.

HPC systems are configured with physical, administrative, and technical controls based on NIST SP 800-171. System configuration information is maintained centrally under version control. Updates are automatically pushed to all systems and deviations noted. Administrative access is limited to as small a subset of staff as required, and two-factor authentication is required for root-level access to any system.

The Purdue team supports compute resources with higher security requirements, including controlled unclassified information (CUI) and HIPAA. RCAC staff are awardees of NSF CICI grants to deploy a campus research data security framework based on NIST SP 800-171 and to implement an intrusion detection system around the HPC Science DMZ.

Research Software Engineering (RSE)

Purdue's RCAC offers a wide range of research software engineering (RSE) solutions in collaboration with researchers to advance scientific discovery, enable greater research reproducibility, and generate lasting impact beyond project durations. RSEs partner with research teams to create or modify software as well as deploy and support research activities. With professional RSE training and extensive experience, RSEs in RCAC have comprehensive expertise in system and software design, programming languages, and deployment practices, combined with up-to-date knowledge in the latest methods and cyberinfrastructure technologies. RSE scientists and engineers work closely with research teams in defining the product, even helping to make the pitch to funding agencies, understanding at a certain depth the underlying methods, developing iteratively with frequent feedback, and providing technical support and training. RSEs also have experience and material to train students and others in the best practices of software engineering and open-source development.

National Cyberinfrastructure Partner and Resource Provider

Purdue's RCAC is a partner in the NSF-funded advanced cyberinfrastructure ACCESS (access-ci.org). RCAC has been a resource provider to the national advanced cyberinfrastructure ecosystem through NSF supported ACCESS (previously TeraGrid and XSEDE) and Open Science Grid. Its Anvil system, funded by NSF in 2020, was developed in 2021 and entered production operation in February 2022. Anvil integrates a large-capacity high-performance computing (HPC) cluster with a comprehensive ecosystem of software, access interfaces, programming environments, and composable services to form a seamless environment able to support a broad range of current and future science and engineering applications.

Open OnDemand and ThinLinc tools allow users to quickly become productive on Anvil through Linux and Windows desktops or familiar tools through their browser (e.g., Jupyter, RStudio). Complex scientific software environments and application stacks are supported via containers orchestrated within a powerful composable subsystem.. Anvil provides 1 billion CPU cores hours and 500,000 GPU hours each year to US researchers through the NSF allocation process. Purdue is also a resource provider to the Open Science Grid, providing access to community clusters for high-throughput computing applications. Purdue provides computing and storage resources to the OSG to support the CMS (Compact Muon Solenoid) project as a Tier-2 site and other OSG virtual organizations.

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