Title

Purdue University Application Deep Dive

Authors

Zurawski, Jason
Addleman, Hans
Chevalier, Scott
et al.

Publication Date

2019-11-01

Peer reviewed

Purdue University Application Deep Dive

May 31st, 2019

Disclaimer

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor The Trustees of Indiana University, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California or The Trustees of Indiana University. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California, or The Trustees of Indiana University.

Purdue University Application Deep Dive

Final Report

Purdue University
West Lafayette, IN
May 31st, 2019

The Engagement and Performance Operations Center (EPOC) is supported by the National Science Foundation under Grant No. 1826994.

ESnet is funded by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research. Benjamin Brown is the ESnet Program Manager.

ESnet is operated by Lawrence Berkeley National Laboratory, which is operated by the University of California for the U.S. Department of Energy under contract DE-AC02-05CH11231.

This is a University of California, Publication Management System report number LBNL-2001318[1].

[1] https://escholarship.org/uc/item/0x07247x

Participants and Contributors

Bilal Abughali, Purdue University
Eric Adams, Purdue University
Hans Addleman, Indiana University
Larry Biehl, Purdue University
Emily Bonem, Purdue University
Scott Chevalier, Indiana University
Natalie Chin, Purdue University
Bradley Clayton, Purdue University
Crystal Dombkowski, Purdue University
Jarrod Doucette, Purdue University
Daniel Fan, Purdue University
Clayton Gentilcore, Purdue University
Lev Gorenstein, Purdue University
Marco Hadisurya, Purdue University
Maralee Hayworth, Purdue University
Petr Kazarin, Purdue University
Yansi Keim, Purdue University
Steve Kelley, Purdue University
Katie Kitamura, Purdue University
Geoffrey Lentner, Purdue University
Joe Levell, Purdue University
Mark Linvill, Purdue University
Xing Liu, Purdue University
Amiya Maji, Purdue University
Kyle Purple, Purdue University
Joshua Randall, Purdue University
Singh Saluja, Purdue University
Shangyuan Sang, Purdue University
Pankaj Sharma, Purdue University
Charlie Smith, Purdue University
Garth Simpson, Purdue University
Doug Southworth, Indiana University
Mark Sullivan, Purdue University
Greg Veldman, Purdue University
Steve Wilson, Purdue University
Michael Witt, Purdue University
Frank Wolf, Purdue University
Yukai Zou, Purdue University
Jason Zurawski, ESnet

Report Editors

Hans Addleman, Indiana University: addlema@iu.edu
Scott Chevalier, Indiana University: schevali@iu.edu
Doug Southworth, Indiana University: dojosout@iu.edu
Dr. Jennifer M Schopf, Indiana University: jmschopf@indiana.edu
Jason Zurawski, ESnet: zurawski@es.net

Contents

1 Executive Summary

In May 2019, staff members from the Engagement and Performance Operations Center (EPOC) met with researchers at Purdue University for the purpose of a campus-wide Application Deep Dive. The goal of this meeting was to help characterize the requirements for two research labs on campus, and to enable cyberinfrastructure support staff to better understand the needs of the researchers they support. Material for this report includes both written documentation from the 2nd Annual Symposium for Research Data Technology at Purdue University, but also a writeup of the discussion that took place in person on May 31st, 2019. Profiled use cases include:

- 3.1 Purdue University Network and Computational Infrastructure Case Study
- 3.2 Tao Lab Case Study
- 3.3 Liu Lab Case Study

The Case Studies highlighted the ongoing challenges that Purdue University has in supporting aspects of the research community that are relatively new to the available technology support structures. Both Case Studies mentioned unique challenges which that summarized into common needs, including:

- Enabling access to local high-performance and high-throughput computation
- Enabling access to local persistent storage
- Use of high-performance data transfer tools (e.g. Globus) that integrate with Cloud storage (i.e. Box, Dropbox, etc.)
- Upgraded local network connectivity to foster collaboration
- Ability to share data with research staff and remote collaborators
- Develop new methods to "on-board" new faculty with regards to the available technology options that the campus, and departments, offer

Purdue received an NSF award[2] to help support upgrading the campus network, specifically to include a Science DMZ and monitoring equipment. While these upgrades were beneficial to certain users, the campus will work to define new methods to raise awareness and encourage adoption of technology.

Action items from the meeting included:

1) The Purdue Central IT Organization (ITaP) will work to educate the Tao and Liu lab staff members about the available storage options offered on campus, both in terms of what is available and the potential benefits. Time will also be spent integrating available storage options into current workflows.
2) When the Globus to Box interface is available, ITaP will work with both labs to take advantage of this approach.
3) Purdue Research Computing and the College of Agriculture will work to integrate institutional storage into the Tao Lab workflow.

[2] NSF Award #1827184, "Integrating Big Data Instrumentation into Campus Cyberinfrastructure"

4) Purdue Research Computing and the College of Agriculture will work to convert some of the Tao Lab workflow steps to use available HPC resources.
5) Encourage adoption of Splashtop as a departmental Remote Desktop solution instead of non University approved solutions such as Team View or Windows Remote Desktop. This is required to help protect departmental resources.
6) Current storage practices in the labs are to primarily store work on local machines, which prevents easy search and curation. Centralized storage is needed to facilitate collaboration between members of the group. Purdue Research Computing and the College of Agriculture will work to integrate this for the Liu Lab.
7) Purdue Research Computing and the College of Agriculture will work to integrate instruments that are hosted by the department more closely with institutional or national storage solutions.
8) Purdue Research Computing and the College of Agriculture will work to integrate the Globus interface to Box to facilitate data from certain instruments that are known to produce large amounts of data. The "Typhoon Trio Variable Mode Imager System" will be targeted initially, this is a variety of imager hardware that produces digital images of radioactive, fluorescent, or chemiluminescent samples.
9) ITaP, along with IT staff working for the College of Agriculture, will continue to work with both labs to enable central campus storage solutions.
10) The Liu Lab utilizes a shared facility with many instruments that need to be able to access portions of the central Science DMZ infrastructure.
11) Purdue University research support staff to investigate integrating Globus Boxconnect with Data Depot.

2 Process Overview and Summary

2.1 Deep Dive Background

Over the last decade, the scientific community has experienced an unprecedented shift in the way research is performed and how discoveries are made. Highly sophisticated experimental instruments are creating massive datasets for diverse scientific communities and hold the potential for new insights that will have long-lasting impacts on society. However, scientists cannot make effective use of this data if they are unable to move, store, and analyze it.

The Engagement and Performance Operations Center (EPOC) uses Application Deep Dives as an essential part of a holistic approach to understand end-to-end data use. By considering the full end-to-end data movement pipeline, EPOC is uniquely able to support collaborative science, allowing researchers to make the most effective use of shared data, computing, and storage resources to accelerate the discovery process.

EPOC supports five main activities

- Roadside Assistance via a coordinated Operations Center to resolve network performance problems with end-to-end data transfers reactively;
- Application Deep Dives to work more closely with application communities to understand full workflows for diverse research teams in order to evaluate bottlenecks and potential capacity issues;
- Network Analysis enabled by the NetSage monitoring suite to proactively discover and resolve performance issues;
- Provision of managed services via support through the IU GlobalNOC and our Regional Network Partners;
- Coordinated Training to ensure effective use of network tools and science support.

Whereas the Roadside Assistance portion of EPOC can be likened to calling someone for help when a car breaks down, Deep Dives offer an opportunity for broader understanding of the longer term needs of a researcher. Deep Dives aim to understand the full science pipeline for research teams and suggest alternative approaches for the scientists, local IT support, and national networking partners as relevant to achieve the long-term research goals via workflow analysis, storage/computational tuning, identification of network bottlenecks, etc.

The Deep Dive approach is based on an almost 10-year practice used by ESnet to understand the growth requirements of DOE facilities[3]. The EPOC team adapted this approach to work with individual science groups through a set of structured data-centric conversations and questionnaires.

[3] https://fasterdata.es.net/science-dmz/science-and-network-requirements-review

2.2 Deep Dive Structure

Deep Dives are structured, base-level conversations between a research group and relevant IT professionals to understand at a broad level the goals of the research team and how their infrastructure needs are changing over time.

The research team representatives are asked to communicate and document their requirements in a case-study format that includes a data-centric narrative describing the science, instruments, and facilities currently used or anticipated for future programs; the advanced technology services needed; and how they can be used. Participants considered three timescales on the topics enumerated below: the near-term (immediately and up to two years in the future); the medium-term (two to five years in the future); and the long-term (greater than five years in the future).

The Case Study document includes:

- ***Science Background—***an overview description of the site, facility, or collaboration described in the Case Study.
- ***Collaborators—***a list or description of key collaborators for the science or facility described in the Case Study (the list need not be exhaustive).
- ***Instruments and Facilities***—a description of the network, compute, instruments, and storage resources used for the science collaboration/program/project, or a description of the resources made available to the facility users, or resources that users deploy at the facility.
- ***Process of Science***—a description of the way the instruments and facilities are used for knowledge discovery. Examples might include workflows, data analysis, data reduction, integration of experimental data with simulation data, etc.
- ***Remote Science Activities—***a description of any remote instruments or collaborations, and how this work does or may have an impact on your network traffic.
- ***Software Infrastructure—***a discussion focused on the software used in daily activities of the scientific process including tools that are used locally or remotely to manage data resources, facilitate the transfer of data sets from or to remote collaborators, or process the raw results into final and intermediate formats.
- ***Network and Data Architecture—***description of the network and/or data architecture for the science or facility. This is meant to understand how data moves in and out of the facility or laboratory focusing on local infrastructure configuration, bandwidth speed(s), hardware, etc.
- ***Cloud Services***—discussion around how cloud services may be used for data analysis, data storage, computing, or other purposes. The Case Studies included an open-ended section asking for any unresolved issues, comments or concerns to catch all remaining requirements that may be addressed by ESnet.

- ***Resource Constraints*****—**non-exhaustive list of factors (external or internal) that will constrain scientific progress. This can be related to funding, personnel, technology, or process.
- ***Parent Organization*****—**overview of the sources of funding and cooperation that facilitate the process of science and technology support.
- ***Outstanding Issues*****—**Final listing of problems, questions, concerns, or comments not addressed in the aforementioned sections.

At an in-person meeting, this document is walked through with the research team (and usually cyberinfrastructure or IT representatives for the organization or region), and an additional discussion takes place that may range beyond the scope of the original document. At the end of the interaction with the research team, the goal is to ensure that EPOC and the associated CI/IT staff have a solid understanding of the research, data movement, who's using what pieces, dependencies, and time frames involved in the Case Study, as well as additional related cyberinfrastructure needs and concerns at the organization.. This enables the teams to identify possible bottlenecks or areas that may not scale in the coming years, and to pair research teams with existing resources that can be leveraged to more effectively reach their goals.

2.3 Purdue University Deep Dive Background

In May 2019, EPOC and Purdue University organized a Deep Dive to characterize the requirements for two uses cases on campus, along with a review of campus technology options:

- 3.1 Purdue University Network and Computational Infrastructure Case Study
- 3.2 Tao Lab Case Study
- 3.3 Liu Lab Case Study

The Purdue representatives were asked to communicate and document their requirements in a case-study format (see 3 Purdue University Case Studies). Each Case Study offers a unique view into requirements that the campus can provide, on a number of different time scales ranging from immediate to future needs.

This exercise is a follow-on to the successful NSF Campus Cyberinfrastructure award, NSF #1827184, entitled "Integrating Big Data Instrumentation into Campus Cyberinfrastructure". This two-year award began in July of 2018 and focuses on adding high speed Science DMZ connections to five big data facilities on the Purdue campus. The facilities support accelerated research in new materials, understanding brain functions and viruses, monitoring lower atmospheric weather, employing geospatial data in teaching and public engagement, and secure computing. These connections enabled high-volume, high-velocity, or interactive science data flows to both the campus research cyberinfrastructure and off-campus facilities.

The face-to-face meeting took place on the Purdue campus on May 31st, 2019 (see Section 4 Discussion Summary). We document next steps in Section 5 Action Items.

2.4 Organizations Involved

The Engagement and Performance Operations Center (EPOC) was established in 2018 as a collaborative focal point for operational expertise and analysis and is jointly led by Indiana University (IU) and the Energy Sciences Network (ESnet). EPOC provides researchers with a holistic set of tools and services needed to debug performance issues and enable reliable and robust data transfers. By considering the full end-to-end data movement pipeline, EPOC is uniquely able to support collaborative science, allowing researchers to make the most effective use of shared data, computing, and storage resources to accelerate the discovery process.

The Energy Sciences Network (ESnet) is the primary provider of network connectivity for the U.S. Department of Energy (DOE) Office of Science (SC), the single largest supporter of basic research in the physical sciences in the United States. In support of the Office of Science programs, ESnet regularly updates and refreshes its understanding of the networking requirements of the instruments, facilities, scientists, and science programs that it serves. This focus has helped ESnet to be a highly successful enabler of scientific discovery for over 25 years.

Indiana University (IU) was founded in 1820 and is one of Indiana's leading research and educational institutions. Indiana University includes two main research campuses and six regional (primarily teaching) campuses. The Indiana University Office of the Vice President for Information Technology (OVPIT) and University Information Technology Services (UITS) are responsible for delivery of core information technology and cyberinfrastructure services and support.

Purdue University was founded in 1869 as Indiana's only land grant university. Purdue has nearly 75,000 students across four traditional campuses, a statewide technology program, extension centers and continuing education programs, as well as another 30,000 students enrolled in an online university.

3 Purdue University Case Studies

Purdue University provided a campus technology overview and two scientific use cases as follows:

- 3.1 Purdue University Network and Computational Infrastructure Case Study
- 3.2 Tao Lab Case Study
- 3.3 Liu Lab Case Study

Each of the Case Studies provides a glance at research activities for the university, the use of experimental methods and devices, the reliance on technology, and the scope of collaborations. It is important to note that these views are limited to current needs, with occasional views into the event horizon for specific projects and needs into the future. Estimates on data volumes, technology needs, and external drivers are discussed where relevant.

Purdue University is committed to supporting these use cases through technology advancements, and is actively executing, and pursuing, grant solicitations to deliver on these initiatives. The landscape of support is likely to change rapidly in the coming years, and the intention is that the Use Cases will be able to take full advantage of campus improvements as they become available.

3.1 Purdue University Network and Computational Infrastructure Case Study

3.1.A Infrastructure Background

Information Technology at Purdue (ITaP) operates all centrally-maintained research computing resources at Purdue University in West Lafayette, Indiana. ITaP's Research Computing Division operated under the name the Rosen Center for Advanced Computing (RCAC) for many years, in memory of Saul Rosen who served as director of Purdue's Computing Center from 1968 to 1987 and who helped to establish Purdue as a pioneering academic institution in high-performance computing.

ITaP Research Computing Division provides advanced computational resources and services to support Purdue faculty and staff researchers. ITaP staff also conduct research and development to enhance the capabilities of these resources. ITaP Research Computing staff provides access to leading-edge computational and data storage systems as well as expertise in a broad range of high-performance computing activities. Specifically, they:

- Evaluate, deploy, and support hardware and software for large-scale scientific computing;
- Promote the effective use of Purdue research computing systems and application software through training and education, consultation, and documentation;
- Contribute to the discovery process through algorithm design and development of effective computing techniques; and
- Partner with researchers to develop grant proposals by providing expertise in the assessment of hardware (e.g. computation, storage, and networking) as well as software (e.g. data movement, workflow) requirements.

3.1.B Computational Facilities and Capabilities

ITaP maintains several different resources for use by campus researchers. Several of them are part of Purdue's Community Cluster Program. Resources include:

- ***Gilbreth*** is a system that has been designed specifically for applications that are able to take advantage of GPU accelerators, and a part of the Purdue Community Cluster. While applications must be specially tuned to use GPUs, a GPU-enabled application can often run many times faster than the same application could on general-purpose CPUs. Due to the increased cost of GPU-equipped nodes, Gilbreth is being offered as an annual subscription to a shared queue at a lower price point than the full cost of a node.
- The ***Scholar cluster*** is open to Purdue instructors from any field whose classes include assignments that could make use of supercomputing, from high-end graphics rendering and weather modeling to simulating millions of molecules or exploring masses of data to understand the dynamics of social networks.
- The ***Data Workbench*** is an interactive compute environment for non-batch big data analysis and simulation, and is a part of Purdue's Community Cluster

Program. The Data Workbench consists of 6 HP compute nodes each with two 24-core AMD EPYC 7401P processors (24 cores per node), and 512 GB of memory per node. All nodes are interconnected with 10 Gigabit Ethernet.

- ***Brown*** is a system that has been optimized for communities running traditional, tightly-coupled applications, and also part of Purdue's Community Cluster Program. Brown was built through a partnership with Dell and Intel in October 2017. Brown consists of Dell compute nodes with two 12-core Intel Xeon Gold "Sky Lake" processors (24 cores per node) and 96 GB of memory. All nodes have 100 Gbps EDR Infiniband interconnect.
- ***Halstead*** is optimized for Purdue's communities running traditional, tightly-coupled applications, and also part of Purdue's Community Cluster Program. Halstead was built through a partnership with HP and Intel in November 2016. Halstead consists of HP compute nodes with two 10-core Intel Xeon-E5 processors (20 cores per node) and 128 GB of memory. All nodes have 100 Gbps EDR Infiniband interconnect.
- ***Rice*** is a system that has been optimized for Purdue's communities running traditional, tightly-coupled applications, and also part of Purdue's Community Cluster Program. Rice was built through a partnership with HP and Intel in April 2015. Rice consists of HP compute nodes with two 10-core Intel Xeon-E5 processors (20 cores per node) and 64 GB of memory. All nodes have 56 Gbps FDR Infiniband interconnect.
- ***Snyder*** is a Purdue Community Cluster that is continually expanded and refreshed, and has been optimized for data intensive applications requiring large amounts of shared memory per node, such as in life sciences. Snyder was originally built through a partnership with HP and Intel in April 2015, though it has been most recently expanded with nodes from Dell. Snyder consists of a variety of compute node configurations, and all nodes have 40 Gbps Ethernet connections. Snyder is expanded annually, with each year's purchase of nodes to remain in production for 5 years from their initial purchase.
- ***Hammer*** is a system that has been optimized for Purdue's communities utilizing loosely-coupled, high-throughput computing. Hammer was initially built through a partnership with HP and Intel in April 2015, and expanded every year thereafter; most recently with nodes from Advanced HPC. Nodes will remain in production for 5 years from their initial purchase.
- ***Research Ecosystem for Encumbered Data (REED) GovCloud*** is designed for working with data that is encumbered with federal security regulations. It is implemented as an expandable set of instances within Amazon's AWS GovCloud facility with access portals in the Purdue academic domain.

3.1.C Storage Facilities and Capabilities

ITaP maintains several different storage resources to accompany computational systems. These include:

- Several data storage resources are offered by Purdue outside of Research Computing. One such offering is Purdue University Research Repository

(PURR), a research collaboration and data management solution for Purdue researchers and their collaborators. Purdue Libraries offers a broader overview of all storage options available to students, faculty, and staff at Purdue.

- ***Globus*** is a third-party supported set of services to support file transfers. It works within ITaP's various research storage systems, can be used to share data between ITaP and remote research sites running Globus, and can also share data between Purdue research systems and personal systems. In the near future, the Globus connector for BOX will be available at Purdue.
- The ***Data Depot*** is a high-capacity, fast, reliable and secure data storage service that was designed for the needs of Purdue researchers in any field. It enables data sharing with both on-campus and off-campus collaborators.
- The ***Home Directory*** for all ITaP research resources is provided by a DDN GS7KX filesystem appliance. Home is the primary space used to permanently hold files for a given user.
- Each Purdue-supported compute cluster is assigned a default Lustre or GPFS parallel ***scratch*** filesystem. The parallel file systems provide work-area storage that has been optimized for a wide variety of job types and designed to perform well with data-intensive computations, while scaling well to large numbers of simultaneous connections.
- The ***Fortress*** system is a large, long-term, multi-tiered file caching and storage system that utilizes both online disk and a robotic tape library with a capacity of over 10PB. ITaP upgraded Fortress from DXUL to HPSS in October of 2011.
- A ***REED Folder*** is a managed storage solution built on top of the Box Cloud platform for research projects requiring compliance with regulations or heightened security.
- A ***Box Research Lab folder*** is a managed storage solution built on top of the Box Cloud platform for research labs to share and collaborate within the lab as well as with outside collaborators.

3.1.D Network and Data Architecture

The Purdue Research Data Network, shown in Figure 1, is a high-speed network infrastructure designed to facilitate the transfer of the large quantities of data produced by and analyzed on Purdue's high-performance computing systems. Based on the Energy Sciences Network (ESNet)'s Science DMZ Model[4], the research network connects to statewide and national research network infrastructures including iLight[5] and Internet2[6].

Some facts about the Research Data Network:

- 100 Gb/second connection to Internet2 via iLight.
- 160 Gb/second of bandwidth to the Purdue Central Research Data Depot storage service
- 160 Gb/second of bandwidth to each Purdue-supported computational system

[4] https://fasterdata.es.net/science-dmz/

[5] https://ilight.net

[6] https://www.internet2.edu

Labs and instruments with requirements for high-bandwidth connections to research storage or computing resources are eligible to directly peer to the research network.

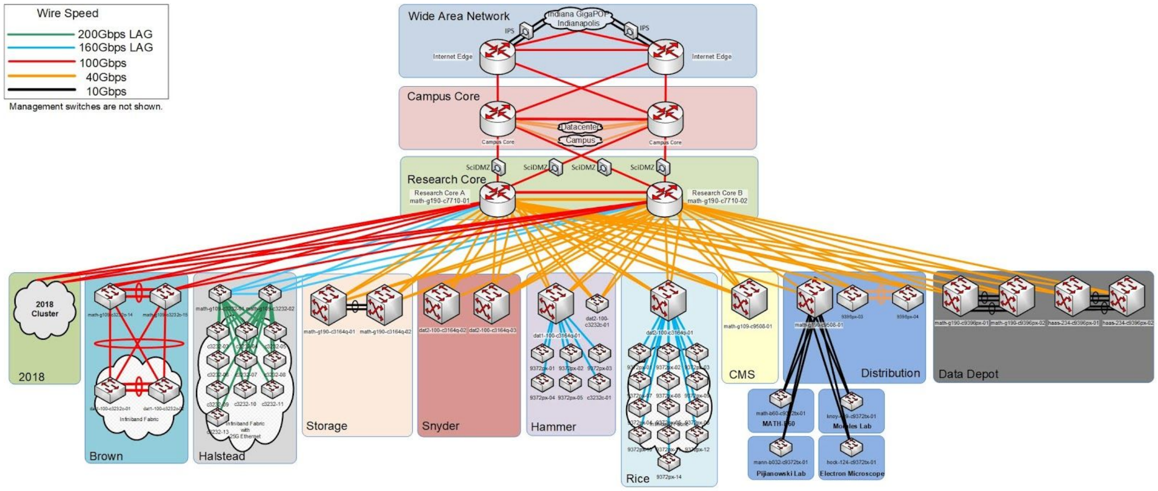


Figure 1 - Purdue Research Data Network

3.1.E Cloud Services

Purdue features a rich ecosystem of cloud resources including support for Amazon and Box.

3.1.E.1 REED GovCloud

The Research Ecosystem for Encumbered Data (REED) GovCloud is a managed computational environment that was designed for use with data whose handling requires compliance with various security protocols beyond those generally used within an open academic setting. REED GovCloud is implemented within Amazon's GovCloud service and has access portals and control points positioned within Purdue's Research Computing Infrastructure.

REED+ Secured Research is Purdue’s ecosystem for sensitive research data. REED+ helps researchers meet compliance for their HIPAA, FERPA, ITAR, and EAR data needs. As a managed research ecosystem with sufficient storage, high speed computing capability and security, REED+ strives to efficiently and cost effectively handle Purdue’s controlled research data and processing needs in a manner compliant with the highest level of cybersecurity applicable to Unclassified data possessed by Purdue University and Purdue University researchers. Development of the REED+ ecosystem is supported by the National Science Foundation under Grant 1840043[7].

3.1.E.2 Box

[7] https://www.nsf.gov/awardsearch/showAward?AWD_ID=1840043&HistoricalAwards=false

Box Secure Store is another storage solution that can meet higher-level compliance requirements, It offers accessibility from anywhere with an internet connection without having to use a VPN. The management of collaborators is straight forward. It automatically implements file versioning, which enables files to be rolled back to previous versions quickly and easily as needed. This system is also verified compliant with HIPAA, FERPA, and NIST 800-171 regulations for data storage.

3.1.F Parent & Affiliated Organizational Cooperation

Purdue University is a resource provider to the NSF-funded national high-end computational cyberinfrastructure XSEDE (http://xsede.org) and a member of the XD Service Provider Forum (https://www.xsede.org/web/sp-forum/spf-membership). Purdue's research computing systems are connected to the XSEDE resources via 100 Gbps network links. Purdue's Condor pool, a distributed computation resource for high-throughput computing, is available to XSEDE users via the Open Science Grid. Purdue's DiaGrid, a HUBzero web-enabled platform with online scientific applications backed up by high-performance and high-throughput computing resources, supports XSEDE researchers through hosting of their scientific applications and gateways. DiaGrid is also available to the broader research community, educators, and students.

As a partner to the XSEDE project, Purdue staff provides in-depth advanced consulting services to help researchers nationwide to effectively utilize the XSEDE resources. Their expertise areas include optimization, scientific application development, and science gateway development and operations. Purdue staff also participate in the XSEDE Campus Champion program, which connect campus research needs to advanced digital resources on XSEDE and other national cyberinfrastructures.

Purdue is also a resource provider to the Open Science Grid with its Condor pool for high-throughput computing applications. Purdue provides computing and storage resources to the OSG, supporting the CMS (Compact Muon Solenoid) project as a Tier-2 site as well as other OSG virtual organizations.

3.2 Tao Lab Case Study

Dr. Weiguo (Andy) Tao, Professor of Biochemistry, Chemical Biology, and Analytical Chemistry at Purdue University, and his research group works expand how mass spec technology can be used in a clinical setting by investigating novel proteomic biomarkers for several diseases, such as Alzheimer's disease, breast cancer, bladder cancer, and kidney cancer.

3.2.A Science Background

The lab focuses on the development and application of biological mass spectrometry for functional proteomics. Mass spectrometry-based proteomics is highly interdisciplinary, bringing together work in biology, chemistry, instrumentation, statistics, and bioinformatics. Proteomics holds significant promise for the discovery of diagnostic or prognostic protein markers for the detection of new therapeutic targets and as a powerful tool to further the understanding of basic biological processes and mechanisms. The realization of these expectations relies on the development of novel approaches to chemistry research and instrumentation.

The Tao lab develops novel strategies and reagents to efficiently target and discover proteins of important biological relevance as potential biomarkers. Such proteins are typically low in abundance, dynamically expressed, and post-translationally modified. The targeted proteomics approach involves the integration of a number of technologies including the selective targeting of proteins with activities of interest, multi-step sample preparation, and mass spectrometry.

Current areas of focus include:

1. Using chemical proteomics to understand virus and bacteria infections, to identify receptors, and to identify protein interacting partners
2. Developing phosphoproteins in extracellular vesicles as disease biomarkers
3. To use phosphoproteomics for molecular signaling in cancer cells:

This Case Study profiles the liquid chromatography–mass spectrometry (LC-MS) process, which uses instruments that analyze protein sequence, structure, modifications, interactions, and functions. All samples undergo the same LC-MS method, and generate raw data that undergos the same process (bioinformatic and statistical analysis):

- Treatments will be treated to viruses and bacteria. Then, proteins from viruses and bacteria will be extracted. These proteins will be digested using trypsin. The digested proteins (peptides) will be loaded into LC-MS.
- Extracellular vesicles (EVs) will be isolated from plasma or urine. Proteins from EVs will be extracted. These proteins will be digested using trypsin. The digested proteins will be called peptides. 2% of peptides will be loaded to LC-MS. 98% peptides will undergo phosphopeptide (phosphorylated peptides) enrichment. Then, the enriched phosphopeptides will be loaded into LC-MS.

- This area focuses on the identification of specific tyrosine-kinase substrates and their phosphorylation sites. Tyrosine is one of the amino acids which are the building blocks of proteins. Kinases are the enzymes that add phosphate groups to proteins. After the treatment of proteins using particular kinase, the proteins will be digested using trypsin. The digested proteins will be called peptides. 2% of peptides will be loaded to LC-MS. 98% peptides will undergo phosphopeptide (phosphorylated peptides) enrichment. Then, the enriched phosphopeptides will be loaded into LC-MS.

By analyzing the protein sequence, structure, modifications, interactions, and functions, it will ultimately help to understand how those proteins work and communicate with other proteins. In short, it will help to understand the biochemistry of proteins. The short-term goal of our lab is to be able to identify known proteomic biomarkers from literature for detecting disease using mass spec. The long-term goal of our lab is to be able to identify some novel proteomic biomarkers for diseases by utilizing mass spectrometry.

3.2.B Collaborators

The Tao Lab collaborates with a small number of primary groups, but these interactions can span great distances. In addition to collaborators at Purdue, there are domestic interactions with researchers at:

- Indiana University
- Carnegie Institution for Science
- University of Illinois, Urbana-Champaign

On occasion there are interactions with collaborators within China, including several former alumni of the Tao Lab. Most collaborations do not involve significant (e.g. > 1GB) data transfers when summaries are used, but may involve the remote use of instrumentation delivered via remote desktop interactions. In the event that raw output is requested, the data size may exceed 1GB in size.

3.2.C Instruments and Facilities

The primary instrument used by the lab is a high-resolution mass spectrometer, the LTQ Orbitrap Velos ETD from Thermo Fisher Scientific[8].

Present

The mass spectrometer is used heavily and is a focal point of the technology used in the lab. The technology within the device is limited. It uses an internal control PC that must be controlled via remote control connection (e.g. using software such as VNC[9]) from other PC resources. Additionally, all data collected by the machine must be exported, via a network connection, to external file systems of nearby PC resources.

[8] https://www.thermofisher.com/order/catalog/product/IQLAAEGAAPFADBMARX
[9] https://www.realvnc.com/en/

The Tao Lab currently uses three different PC workstations for data capture, storage, and analysis,:

- PC1 manages the mass spectrometer.
- PC2 and PC3 are used to analyze the raw data.
- Drobo[10] storage devices can be accessed through the local network.
- Shared storage through a mapped “F” Drive is accessible to all of the machines.

The internal LAN (connected at 1Gbps) enables data transfers between the three machines. Software packages, including MaxQuant and Proteome Discoverer),are installed on both PC2 and PC3 to facilitate higher throughput for the overall analysis workflow including the conversion of raw data into usable data sets.

Additional analysis and statistical modeling using Perseus or R can also be performed on PC2 and PC3, or on student or staff personal computers that are able to connect to the LAN environment. Email and USB flash drives are typically used to transfer data between “Bring Your Own Device” resources (e.g. “BYOD”), and the Lab PCs, as shown in Figure 2.

[10] https://www.drobo.com/storage-products/

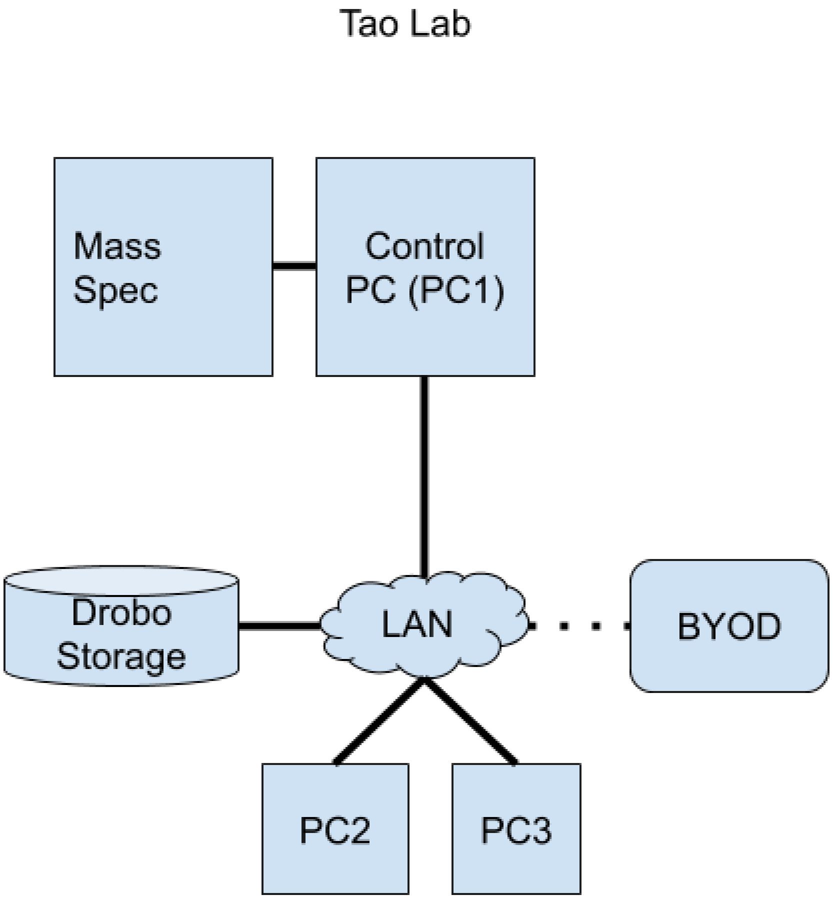


Figure 2 - A diagram of the data workflow used by the Tao Lab that includes personal (e.g. "BYOD") Devices.

2-5 Years

There are no plans to upgrade the mass spectrometer in the next five years, although it is anticipated that computing and storage resources will be upgraded. The available storage within the lab is reaching critical levels and additional sampling cannot be performed without migration or deletion of historic data.

Research computing at Purdue has both computation and storage resources available, thus the integration of institutional computation and storage is a natural progression of this use case.

Beyond 5 Years

Advancements in mass spectrometry will allow for cheaper machines that produce more data. It is anticipated that the Tao Lab will upgrade instrumentation in the longer time frame, which will require significant upgrades in terms of storage and processing. The use of Purdue central resources will be required at this stage, and a possible architectural diagram of this use case that was developed jointly with ITaP during the discussion portion of the Deep Dive is shown in FIgure 3.

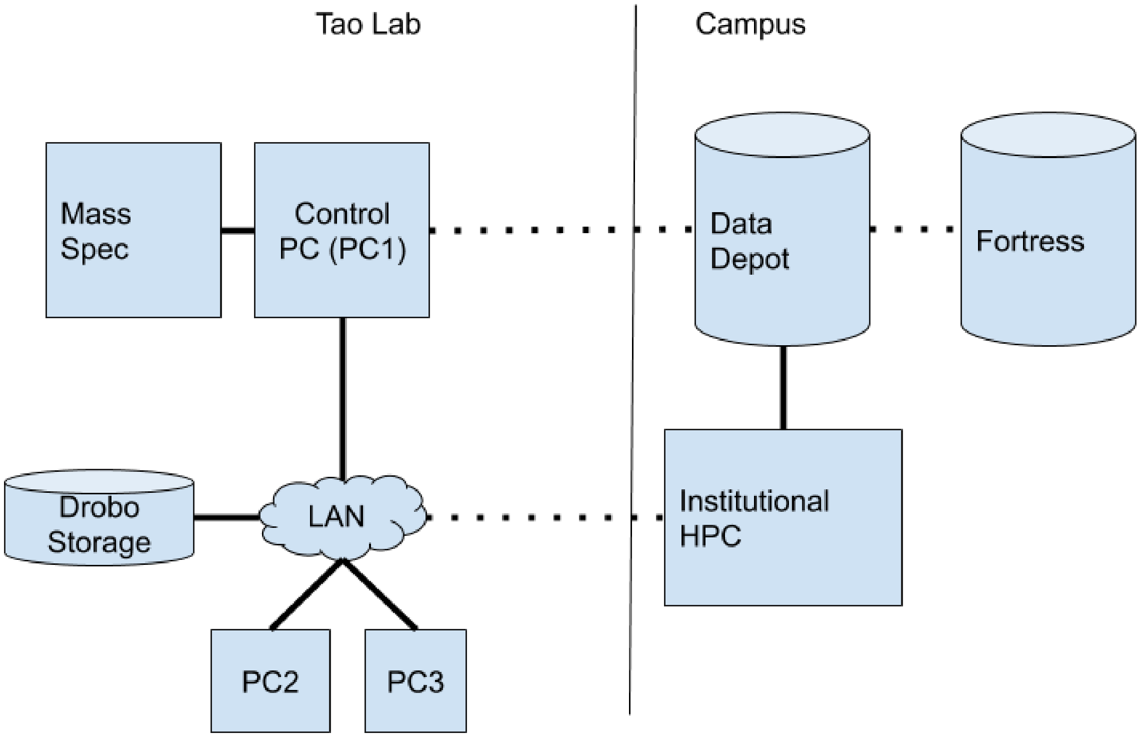


Figure 3 – Architectural diagram of the Tao Lab use case in the 5+ year time frame with integrated use of institutional resources

3.2.D Process of Science

Mass spectrometers (MS), the primary instrument used for the research goals of the Tao Lab, measure the mass to charge ratio (m/z) of one or more molecules present in a sample. The results are typically presented as a mass spectrum, which is a plot of intensity as a function of the mass-to-charge ratio.

Present

A typical workflow is:

1. Samples are loaded into the LC-MS instrument.
2. Processing is initiated via the control PC (PC1).
3. The LC-MS generates raw data output in a proprietary format (Xcalibur), and stores these results on PC1. This 800 MB file contains data for all of the peaks that represent different ions with distinct mass to charge ratios (m/z).
4. Using local networking, the raw data is transferred to either PC2 or PC3, where the data is converted into a more readable format using the Xcalibur software.
5. After conversion, MaxQuant and Proteome Discoverer are used for analysis.
6. Further statistical work to produce data and figures for publications using Perseus or R can also be performed on PC2 and PC3.
7. Results are stored in the shared "F" drive and Drobo storage.

The most time consuming parts of this process are:

- The analysis of the raw data set, which generally takes between 40 and 120 minutes and is set by the researcher. The time to analyze the raw data is in part limited by the fact that the proprietary Xcalibur software can only run on a single PC and cannot be parallelized. It is possible this stage could be expedited if there was additional memory or cores on PC1.
- The time to process the raw data set into other formats, which can be parallelized, although that has not yet been done. This can take several hours to several days depending on the size of the raw data. For example, if there are three data categories, such as control, kidney cancer type 1, and kidney cancer type 2, and each has three biological replicates, that means there are 9 sets of raw data to analyze for one experiment which may take more than a day.

2-5 Years

The workflow itself is not expected to undergo significant changes during this timeframe. MS technology will still produce raw data that requires additional processing, and it is expected that most MS hardware will utilize proprietary formats, and thus proprietary software,. Advancements are expected to be made in the software that is capable of converting raw files into more usable formats during this time both in terms of usability as well as the time required through the use of parallel processing.

During this time window the Tao Lab will likely require additional resources:

- ***Storage***: additional storage connected to the pipeline between the control and analysis resources for steps 3, 5, and 7. The raw data (800 MB - 1 GB), Maxquant data (1-5 GB), Proteome Discoverer data (1-5 GB), Perseus data (10-200 MB), and R data (less than 1 MB) must be stored as a part of the overall workflow.
- ***Processing***: availability of multi-process or multi-core local machines, or institutional HPC for step 5 of the workflow. The Maxquant and proteome discoverer software packages require significant CPU availability.

Beyond 5 Years

Technology trends indicate that there will be significant advancements in mass spectrometry that will allow for cheaper machines that produce more data. With the use of intuitional storage and processing, a new mass spectrometry instrument will be able to adjust to these new needs.

3.2.E Remote Science Activities

The Tao Lab does not currently rely on external instrumentation, storage, or processing. However, there are external users that often utilize the mass spectrometers to perform operations remotely.

Present

Remote access is typically accomplished via the use of control software:

- Team Viewer[11]
- VNC[12]
- Splashtop Business[13]

Of these applications, Team Viewer is discouraged by ITaP as well as local College of Agriculture IT, and is routinely blocked or removed from lab resources due to security concerns. Splashtop Business is the approved mechanism for remote use. VNC is used for communication between the Mass Spectrometer and the local PCs, but has been blocked for wide area use.

2-5 Years

The Tao Lab does not anticipate any changes to operational workflow during this time. Remote access by external parties could increase as the number of collaborators increases.

Beyond 5 Years

As mass spectrometers become more common, it is anticipated that remote access of the Tao Lab machine will decrease.

3.2.F Software Infrastructure

There are several software packages used by the Tao Lab for different functions of the workflow. Some of these are proprietary software packages that come with instruments, others are open source:

- ***XCalibur***[14]: Thermo Fisher provided software that facilitates setup, data acquisition, data processing, and reporting
- ***MaxQuant***[15]: Quantitative proteomics software package designed for analyzing large mass-spectrometric data sets
- ***Proteome Discoverer***[16]: Identifies and quantifies proteins in complex biological samples
- ***Perseus***[17]: Supports biological and biomedical researchers in interpreting protein quantification, interaction and post-translational modification data
- ***R***[18]: Software environment for statistical computing and graphics

Present

[11] https://www.teamviewer.com/en-us/
[12] https://www.realvnc.com/en/connect/download/viewer/
[13] https://www.splashtop.com
[14] https://www.thermofisher.com/order/catalog/product/OPTON-30487
[15] https://www.maxquant.org
[16] https://www.thermofisher.com/order/catalog/product/OPTON-30795
[17] https://maxquant.net/perseus/
[18] https://www.r-project.org

XCalibur is used to run the instrument and generates the raw data. This software must be used due to the proprietary nature of the data format. This stage is the highest time consumption for the use case, and it cannot be accelerated through parallel processing methods.

MaxQuant and Proteome Discoverer are used to analyze the raw data and convert the XCaliber format into more readable files. MaxQuant is open source and could be parallelized across HPC resources by college of Agriculture IT staff. These two pieces of software, since they are exposed via PC2 and PC3 in the Tao Lab, can be operated remotely using remote access software.

Statistical analysis using Perseus and R can be performed on lab and personal machines after the data has been converted to more readable formats.

2-5 Years

Software and data formats are not expected to change in a significant way. New releases may offer incremental improvements to process and output. Significant changes to speed up operation through parallelization are possible.

Beyond 5 Years

Data formats and software may change as hardware within the Tao Lab changes. It is unknown how this will impact workflows at this time.

3.2.G Cloud Services

Present

At this time the Tao Lab does not anticipate any significant use of cloud storage or computing.

2-5 Years

The use of cloud resources to facilitate storage and sharing of data may become more important as the number of collaborators grows.

Beyond 5 Years

It is unknown at this time if significant use of cloud storage or computing will be a factor.

3.2.H Known Resource Constraints

The lack of storage is a current and future constraint for the Tao Lab. The current availability of storage is approximately 15TB and is almost completely utilized. Due to this limitation, the Lab is routinely deleting older data sets to make room for new research products. Data set sizes will only grow, as the frequency of operation of the LC-MS increases and the precision of the instrument grows.

Certain software packages have limitations in operation:

- Requirement to use 32-bit architecture
- Lack of support for multi-core/multi-processor operation

Given these limitations, many workflow operations are inherently CPU bound.

3.2.I Outstanding Issues

The sharing of Tao Lab resources with external collaborators has been a significant source of friction due to the use of certain software packages. Team Viewer has significant security vulnerabilities, and thus has been blocked and removed from lab resources on several occasions after new staff unaware of the security posture re-install the package. Splashtop Business, an approved and licensed piece of software, has been recommended for use instead. Additionally, VNC, which is used for control of the LC-MS instrument, can only be used locally.

The availability of storage resources is a significant challenge for the Tao Lab and is forcing deletion of data sets from primary storage on a routine basis. Long-term storage, either in the form of external drives or tape archives, is a critical need.

3.3 Liu Lab Case Study

Dr. Xing Liu is an Assistant Professor in Biochemistry, in the College of Agriculture, at Purdue University. Her laboratory is interested in exploring the mechanisms governing the dynamics of Cullin-RING Ligases (CRLs), from perspectives including biochemistry, proteomics, mathematical modeling, and molecular genetics, using both human cells and plants as the experimental systems.

3.3.A Science Background

The primary goal of the researchers in the Liu lab is to understand the mechanism of protein ubiquitination (attaching ubiquitin to target protein) and how this post-translational modification determines proper cellular and organismal functions. To achieve this goal, human and plant cells are used as experimental systems. Further, multidisciplinary approaches are applied, such as biophysical assays, genetics, genome editing, quantitative mass spectrometry, and computational modeling.

The profiled research action will focus on the dynamics of Skp1•Cul1•F-box protein (SCF) ubiquitin E3 ligases. This group of enzymes play critical roles in eukaryotic biology, ranging from cell division to organ formation. Misregulation of these enzymes often leads to human diseases such as cancer, and growth defects in plants including important crops. Each SCF has a core Cul1 protein that assembles with a family of F-box proteins to form different enzyme complexes that target distinct substrates in cells. The genome of human and the model plant Arabidopsis encodes 69 and ~700 F-box proteins, respectively, each of which needs to assemble with Cul1 for the degradation of its substrates. Therefore, timely recruiting the right type of F-box protein is essential for proper SCF function. We aim to understand how the cellular repertoire of the SCF is dynamically controlled to meet the needs from a changing pool of cellular substrates.

3.3.B Collaborators

This emerging area of research does not involve any significant external collaborators and is done within the confines of Purdue University. All instrumentation, storage, and processing is handled locally using existing resources.

3.3.C Instruments and Facilities

A number of scientific instruments are available to faculty performing research in the Department of Biochemistry.

Present

Table 1: Purdue facilities in common use by the Liu Lab.

Facility Name	Services
Life Science Microscopy Facility[19]	Transmission and Scanning Electron Microscopy, Light Microscopy
Purdue Interdepartmental Nuclear Magnetic Resonance (NMR) Spectroscopy Facility	Liquid and Solid State NMR Services
Purdue Genomics Facility[20]	DNA Sequencing, Oligonucleotide Microarray Services
Analytical Mass Spectrometry Facility	General Mass Spectrometry Services
Bindley Flow Cytometry Facility	Flow Cytometry, Cell Sorting
Purdue Proteomics Facility[21]	Protein Mass Spectrometry, Protein Sequencing, 2-D Gel Electrophoresis, Amino Acid Analysis

The Liu lab routinely uses microscopy to examine samples, perform spectral analysis, and sequence samples using the facilities at Purdue listed in Table 1. These instruments are not connected to the Purdue Science DMZ infrastructure, so transmission of the data sets is accomplished using portable media or email.

2-5 Years

It is anticipated that the departmental instruments will be connected to high-speed local storage and computation infrastructure by the Purdue Science DMZ, which will facilitate the use of a stable and predictable pipeline for the research data sets.

Upgrades and augmentations to key components, including the genomics sequencers and microscopy instruments, are expected during this time frame.

Beyond 5 Years

The number of local and remote instruments is expected to grow during this time, as the need to integrate technology into the workflow increases. Storage and computation, along with network connectivity, must keep pace with these demands.

3.3.D Process of Science

As described in Section 3.3.C Instruments and Facilities, multiple data sources and instruments are used for the research. Some data such as the gel images, microscope

[19] https://www.purdue.edu/discoverypark/bioscience/index.php
[20] https://www.purdue.edu/hla/sites/genomics/
[21] https://www.purdue.edu/discoverypark/bioscience/index.php

images, fluorescent spectra are final and do not require additional post-processing or manipulation. Others, such as western blot images (used to detect the presence of a specific protein extracted from either cells or tissue) are analyzed by software (e.g. ImagJ[22]) for quantification. Additional data sets, such as the Sanger sequencing (method of DNA sequencing based on the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during in vitro DNA replication) results are used to confirm the identity of the genes that lab-staff have cloned. In most cases, post-processing analysis involves the use of graphical packages and statistical analysis techniques.

Present

A general workflow is:

1. Obtain experimental materials (eg. cultured human cells, Arabidopsis plants) of the correct genetic background.
2. Grow the experimental materials under desired conditions.
3. When necessary, treat the experimental materials with chemicals or change the growth condition.
4. Collect samples from the experimental materials at different time points following the treatment or the change of growth condition.
5. Process samples to extract DNA, RNA, or protein, for analyses at the molecular level.
6. Alternatively, samples such as plant tissues are processed for microscopic analyses.

The following is a summary of the requirements for the various forms of data used by the Liu Lab:

- ***Regular experimental data that does not rely on special instrumentation***: gel images, western blot images, regular microscope images, quantitative data measured by the researcher. These data sets are usually small in size (< 2 MB each), and are accumulated quickly (~10 items / week). Because every researcher in the lab generates these data sets on a routine basis, they are hard to manage for both backups and information sharing. A method to catalog and share these results is needed.
- ***Data collected from shared local instruments***: confocal imaging files, fluorescence emission spectra, images from fluorescent scanners. These data sets need to be transferred from the computer that controls the instrument to a lab computer or personal computer of a researcher. Some of the raw files require special software for viewing, which may only be installed on the computer controlling the instrument. As a result of the more precise nature, these files usually require more storage space (5-20 MB / experiment). There is not currently a centralized storage mechanism for this data.
- ***Data from university facilities***: sequencing data, proteomics data. Organization of these data sets is important, and a database platform that

[22] https://imagej.nih.gov/ij/

everyone in the lab could use would greatly accelerate the management and curation of these data sets.

- ***Data for reagents stored in the lab***: Excel formatted files are used almost exclusively to collect information about plasmid stocks, frozen cell lines, shared primers, etc. and are shared between lab staff. An easy-to-navigate software package to simplify this is desirable. Such software should allow entering detailed information of each reagent, sorting the existing records, and search for specific items.
- ***External data sources***: The Liu lab uses public online databases to help with their research. One difficulty encountered is that one cannot retrieve bulk data records from some resources. There is not currently a way to automate downloads of data sets, so navigation and curation is time consuming.

The following diagram maps current mechanisms for data migration between Liu lab staff and instrumentation. This diagram features the location of instrumentation, storage (currently cloud resources), and the use of "Bring Your Own Device" (e.g. "BYOD") personal computing resources:

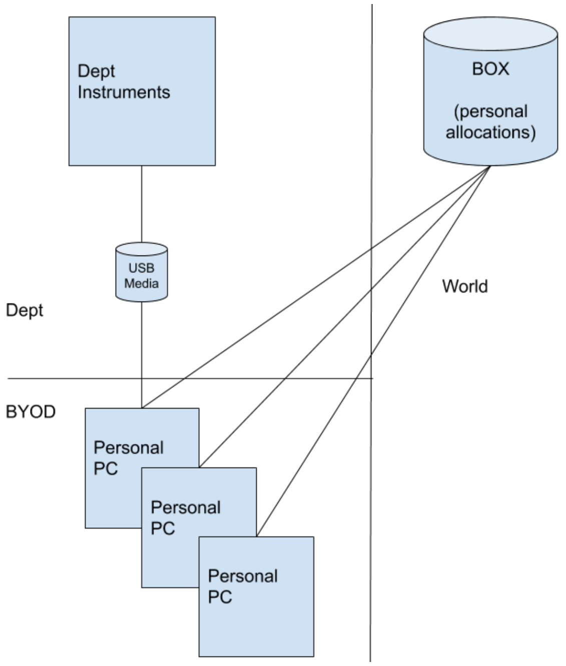


Figure 4 – Liu Lab Workflow

The lack of a clear data migration mechanism between the instruments, cloud storage, and personal computing resources is not a serious problem, but will grow in the coming years.

2-5 Years

During this time frame it is anticipated that there will be minimal changes to the process of science: it is still expected that most wet lab and instrumentation use will remain stable. The use of additional local resources, within the Liu lab or the Purdue campus, will depend on the availability and usability of the resources. External instruments could be utilized if local resources do not meet the needs of the research.

Data storage, curation, search, and organization are all highly important factors for the research. The lack of clear mechanisms to share data, internally, and to external parties, will impact productivity in this time frame. The Liu lab requires help in researching and adopting advanced tools.

Beyond 5 Years

The data set sizes from new instrumentation and changes to the workflow are unknown for this time frame. The same issues as remarked in the 2-5 year window must be addressed before the future, as they will hamper productivity.

3.3.E Remote Science Activities

At the current time, there is not a significant driver or use case for remote scientific use cases.

Present

We currently have one major collaborator who is a physicist at Purdue that is helping to perform computational modeling. The way we communicate and share data is primarily through emails. Regarding the computational model itself, we have not asked for any raw data yet.

2-5 Years

The use of computational modeling will increase, and the use of local computational resources is expected to be the normal operating procedure.

Beyond 5 Years

It is unknown what remote resources will be utilized during this time window.

3.3.F Software Infrastructure

The Liu lab uses a number of software packages to facilitate their workflow. Many are tied to the instruments described in Section 3.3.C Instruments and Facilities to manage the flow of data from the devices directly. Some are analysis focused.

Present

A partial list of software used within the lab includes:

- ***Microsoft Excel***: widely used spreadsheet program that allows for manipulation of numerical data sets

- ***ImagJ***[23]: a public domain Java image processing program inspired by NIH Image[24]. It runs either as an online applet or as a downloadable application on any computer with a Java 1.4 or later virtual machine. It can display, edit, analyze, process, save and print 8-bit, 16-bit and 32-bit images. It can read many image formats including TIFF, GIF, JPEG, BMP, DICOM, FITS and "raw". It supports "stacks", a series of images that share a single window. It is multithreaded, so time-consuming operations such as image file reading can be performed in parallel with other operations.
- ***Scripts***: A set of homemade software packages used to process raw data from instruments. These are typically maintained by each researcher, and not widely shared or maintained beyond the initial use cases they are defined to solve.

2-5 Years

Software and data formats are not expected to change in a significant way, new releases may offer incremental improvements to process and output. Significant changes to speed up operation through parallelization are possible.

Beyond 5 Years

Data formats and software may change as hardware within the Liu Lab changes. It is unknown how this will impact workflows at this time.

3.3.G Cloud Services

Present

At this time the Liu Lab only partially uses cloud resources. Some data exchange (using personal allocations of Box) is possible, but not widely used.

2-5 Years

The use of cloud resources to facilitate storage and sharing of data may become more important as the number of collaborators grows. The Liu lab would like to explore the use of BOX connections to Globus if this becomes a more widely available resource for data movement on campus, as their primary workflow currently relies on transmission of research data from instruments to cloud storage, and then to personal computers of individual researchers.

Beyond 5 Years

It is unknown at this time if significant use of cloud storage or computing will be a factor.

[23] https://imagej.nih.gov/ij/

[24] https://imagej.nih.gov/nih-image/

3.3.H Known Resource Constraints

As the number of staff in the Liu lab increases, the need for stable and predictable collaboration tools will as well.

Present

At the current time, there are no good solutions in place to facilitate aspects of the workflow:

- ***Data Transfer***: Data migration from local or departmental instruments is typically done with removable media or email. While data sets are not large currently (< 200 MB / month), they are accumulating and growing, making this a problem that requires a stable and predictable solution.
- ***Data Storage and Curation***: Each staff member in the Liu lab routinely manages their own data sets. This is an issue as duplication becomes more common. For example, one staff member may have created a data set that is not shared with other lab members, and a second staff member may perform similar analyses to create the same dataset at a later date. A central method to store and catalog what is available would resolve this issue.
- ***Data Storage Backups***: While the data sets used across the lab staff are not large currently (<100 MB / person / month), they are accumulating and growing (can easily reach ~6GB / year). What is needed is a stable and simple means to backup the data, so that each lab personnel does this routinely (or it is done automatically). The goal is that data generated by each person will be accessible even long time after the person has left the lab. this a problem that requires a stable and predictable solution.
- ***Access to HPC/HTC resources***: Most staff members do processing using local resources, either on Lab machines or personal computers.

2-5 Years

As the use of instrumentation increases, the need for integration with technology (storage, processing, collaboration tools) will increase. It is unknown at this time where the areas of friction will be, but the aforementioned problems must be solved by this time window or the risk of productivity drains increases.

Beyond 5 Years

It is unknown at this time what changes or constraints will exist for this time window.

3.3.I Outstanding Issues

Being able to streamline the access to storage and processing is a high priority. Figure 5 addresses some possible solutions in this space. These include the addition of the Globus Connect software enhancements to access cloud storage, as well as integration with Purdue Institutional storage components.

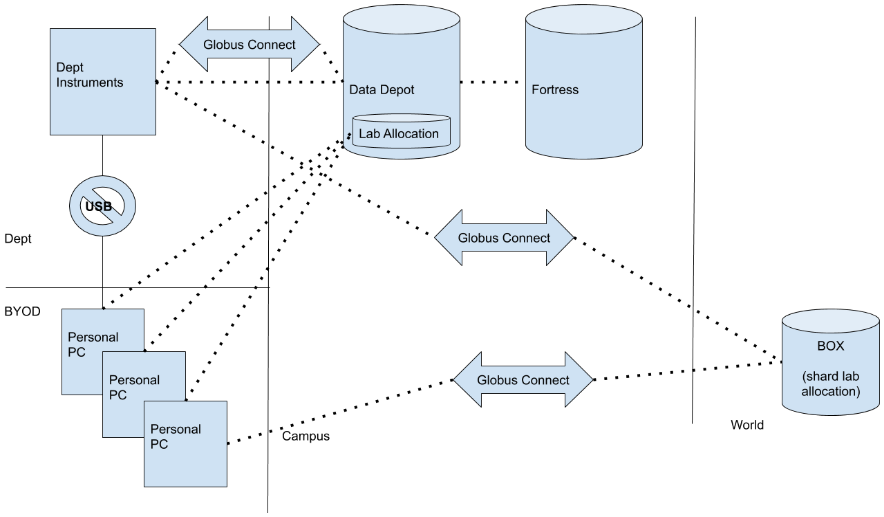


Figure 5 – Liu Lab Improved Workflow

Figure 5 shows several key needs addressed:

- Integration of institutional storage (long term and short term) using departmental allocations in Data Depot and Fortress
- Integration of Globus Connect as a data transfer/curation mechanism between institutional storage, cloud storage, and personal use (Box, PCs)
- Policy that can be used to control allocated group storage

This approach can be scalable, but requires some engineering (physical and social) to be put into practice.

4 Discussion Summary

On May 31, 2019, members of the EPOC team, staff from Purdue IT, representatives from the College of Agriculture IT, and researchers from the College of Agriculture met to discuss the science use cases and technology support options at Purdue University. This review was held in Lafayette, IN.

During the discussion, the following points (outside of clarifications to the Case Study described in Section 3 Purdue University Case Studies) were emphasized.

4.1 Purdue Campus Discussion

The Purdue campus infrastructure features numerous technology offerings and services that appeal to researcher needs. A key gap in adoption is sociology: some research groups have a mistrust of managed solutions and have been slow to adopt mechanisms that they are not familiar with.

Storage was the first and most discussed topic. The two Purdue storage resources that were most relevant for the use cases were: Specifics related to this line of discussion are:

- ***Short-term Storage***: Data Depot is Purdue's high-capacity, fast, reliable, and secure data storage service designed, configured, and operated for the needs of Purdue researchers that is shareable with both on-campus and off-campus collaborators. This resource has a low cost, or in some cases may be free, and is an alternative to the use of private and non-scalable storage solutions.
- ***Long-term Storage***: Fortress is a large, long-term, multi-tiered, file caching and storage system utilizing both online disk and robotic tape drives that is integrated with Data Depot. This solution is ideal for archived research data that is no longer used frequently, but must be retained.

Faculty within the College of Agriculture have been encouraged to migrate toward these resources in the past, but have not done so due to the complexity of training staff, the mistrust for the medium (e.g. being unsure of reliability), and lack of a simple way to integrate these resources into existing workflows. Working with Departmental IT support, it is expected that both Tao and Liu labs will make an effort to use these resources in the near term.

Other discussion involved other technology support items:

- ***Computation***: Purdue Research Computing has numerous resources available for computational work, listed in Section 3.1. A barrier to adoption is the time needed to convert existing codes and workflows, or the lack of a strong use case to justify devoting the time needed to convert the workflows. Research Computing support staff are dedicated to moving some aspects of the Tao workflow where there are software components that can benefit from HPC.

- ***Data Movement Software using Globus***: Purdue has an institutional contract for the use of Globus data transfer services. ITaP staff are actively working to add features including integration to cloud storage. Use of this service is high among other departments, but lower with the College of Agriculture. It is expected that some of the workflow modifications discussed at this review may change this.
- ***Cloud Services***: Purdue maintains an institutional Box service that allocates cloud storage to faculty and staff. ITaP are working with Globus to integrate software that will enable the use of Globus with this cloud storage.
- ***Access to National Resources via R&E Networking***: Most researchers are not aware of networking paths, and as a result are more inclined to do work locally instead of collaborating using R&E networks. As the need to access remote instruments increases, and the tools such as Globus facilitate better sharing of data sets, it is anticipated that network use will rise.

ITaP will continue to work with the colleges and departments of Purdue to raise the visibility and usability of services that are offered, and will also continue engagement efforts to work on integration and build trust.

4.2 Tao Lab Discussion

The Tao Lab's most immediate needs are related to computation and storage:

- ***Computation***: Some analysis work is easily parallelizable and can be run on machines that feature multiple cores or processors. Use of institutional resources could accelerate some of the activities.
- ***Storage***: Current research progress is not possible without deleting older data sets from local storage. Use of institutional resources could remove this storage restrictions and allow for staff to increase the number of research activities without having to worry about storage aspects.

Figure 3 in Section 3.2 summarizes the high level discussions on these topics, and plots a roadmap for working with the group to fix some of the short-term and long-term support issues.

4.3 Liu Lab Discussion

The Liu Lab's most immediate needs are related to storage and workflow management.

- ***Storage***: Current research workflows use existing local storage, but shifting to using institutional resources, such as Data Depot or Fortress, could enable easier data storage and allow for staff to increase the number of research activities without having to worry about storage space.
- ***Bring your own Device (BYOD)***: A number of staff use personal resources for computational work, generally referred to as BYOD or Bring Your Own Device. In the general case is not a problem. However, for the Liu Lab, the

lack of a clear way to exchange info between Purdue resources, Cloud Storage, and BYOD is challenging. The use of Globus and Box, with the necessary connectors, may address the need for easier data sharing, along with migration to the use of Data Depot or Fortress.

Workflow/Data Paths: With the integration of tools such as Globus for data movement into the workflow, the data flow path could be altered to facilitate additional data mobility. Since many of the instruments that they utilize are located in a central departmental facility, it is possible that by integrating technology offerings to meet the use case for the Liu Lab would have the broader effect of benefitting a large number of additional users.

5 General Findings and Action Items

EPOC and Purdue University recorded a set of action items from the event, continuing the ongoing support and collaboration. These are a reflection of the Case Study report and in person discussion.

A. ***General Findings Related to ITaP***:
   1. Location of scientific resources on Purdue campus is mostly centralized which is beneficial to many users. This architecture removes the necessity for researchers to own or maintain their own instruments and allows for ITaP and College of Agriculture IT staff to build services directly beneficial to researchers for storage, computation, and external network use.
   2. College of Agriculture departemental instruments are centralized and can utilize ITaP HPC/Storage support. It is not always the case that users of these resources take advantage of this support, and reply on external means such as USB Drives or email for transmission between instrument locations and personal machines
   3. The concept of shared "group storage" can be done via campus or external cloud providers, but is not a regular part of most workflows. This is due to some people not being aware of the service, or the costs and benefits being unknown.
   4. The availability of centralized HPC resources are available as a condo model, where users may purchase a certain amount for dedicated use and burst beyond purchase amount when resources are available.
   5. The Globus to Box interface to facilitate easy migration of research data is planned for campus, but not readily available at the time of review.

B. Action Items related to ITaP
   1. ITaP will work to educate the Tao and Liu lab staff members about the available storage options offered on campus, both in terms of what is available and the potential benefits. Time will also be spent integrating available storage options into current workflows.
   2. When the Globus to Box interface is available, ITaP will work with both labs to take advantage of this approach.

C. ***General Findings Related to the Tao Lab***:
   1. The availability of local storage is a critical issue, and the lab must routinely delete data to make space for ongoing research.
   2. The use of institutional storage is available, but not currently implemented. The interactions between instruments and storage resources is not automatic.

3. Some software analysis tasks may be converted to HPC versus the use of local analysis tools.
4. The use of the Remote Desktop tools Team Viewer is discouraged, but others such as Splashtop is permissible.

D. ***Actions Related to the Tao Lab***:

1. Purdue Research Computing and the College of Agriculture will work to integrate institutional storage into theTao Lab workflow.
2. Purdue Research Computing and the College of Agriculture will work to convert some of the workflow steps to use available HPC resources.
3. The use of Remote Desktop tools (e.g. Team Viewer) is discouraged, but use of others (Splashtop) is possible. The College of Agriculture will continue to raise the issue with faculty groups, as well as with new students and staff that are unaware of the policy.

E. ***General Findings Related to the Lui Lab:***

1. The data sizes from current research activities are small, as are the number of files generated. Efforts to improve workflow could result in these numbers growing in the future.
2. The need for additional compute resources is not anticipated at this time, as most analysis is done on personal computers. When data volumes grow, it is expected that this may change.
3. Current storage practices across the group are to primarily use local machines, which prevents easy search and curation. Centralized storage is needed to facilitate collaboration between members of the group.
4. Instruments hosted by the department are not automatically connected to institutional or national storage solutions, or Science DMZ Networking.

F. Actions Items Related to the Lui Lab:

1. Current storage practices across the group are to primarily use disk space on local machines, which prevents easy search and curation. Centralized storage is needed to facilitate collaboration between members of the group. Purdue Research Computing and the College of Agriculture will work to integrate this for the Liu Lab.
2. Purdue Research Computing and the College of Agriculture will work to integrate instruments that are hosted by the department more closely with institutional or national storage solutions.
3. Purdue Research Computing and the College of Agriculture will work to integrate the Globus interface to Box to facilitate data from certain instruments. The "Typhoon Trio Variable Mode Imager System" will be targeted initially, this is an a variety of imager hardware that produces digital images of radioactive, fluorescent, or chemiluminescent samples.

Appendix A - College of Agriculture Facilities Overview

The Purdue College of Agriculture provides support services to researchers via the central support agency named AgIT. Technical services provided in the College of Agriculture include network administration and maintenance; development of network infrastructure; technical operations and maintenance; technical support of faculty and staff; systems design, analysis, and programming; web server development.

AgIT's mission is to develop an organization-wide computing and information system in which people are able to effectively use and apply information technology in fulfilling the mission of the Purdue College of Agriculture.

Our Desktop Support Specialists provide technical support for specific departmental systems and customers. We serve as first-contact people for any technical support issue. Services that AgIT provides include the following:

- Troubleshooting of computer hardware and software in faculty/staff offices and departmental labs
- Software installations and upgrades
- New hardware pre-purchase research and recommendations
- Setup of new hardware
- Hardware repair or coordination of outsourced repair tasks

AgIT also provides Business Relationship Management via their BRM group. BRM’s exists to bring the full potential value of the university IT investment to the College of Agriculture faculty. This is accomplished through targeted consulting, knowledge gathering, and building partnerships with both faculty and IT stakeholders across the entire IT community.

In addition to centralized IT resources each department has a local IT resource. Academic IT Specialists (AITS) primary role is to liaison with the broader Purdue IT community acting on behalf of, or in concert with, faculty, students, and staff. Aligning their needs with the best available IT resources to accelerate scientific research, and support the mission of the department, college, and university. AITS support research instrument connected computers in each PI lab. AITS provides consultation on wide-ranging research and enterprise IT services and solutions. These services include but are not limited to 3D printing, data management, research workflow assessment, IT purchase consultation and facilitation, IOT sensor development and deployment, and where appropriate, broader technical support. AITS also works closely with central Purdue IT organizations to develop IT best practices and IT policy.

Appendix B - Purdue University 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[25] 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)

[25] http://campustechnology.com/articles/2010/08/01/campus-technology-innovators-award-2010.aspx

- 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. In 2018, a Purdue-funded network modernization will deploy over 1000 new access-layer switches and a new campus core.

As a long-time Large Hadron Collider (LHC) partner (a CMS Tier-2 center[26]), 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

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.

[26] https://www.physics.purdue.edu/Tier2/

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

Campus cyberinfrastructure follows the policies, standards, and best practices outlined by SecurePurdue[27].

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.

Research storage services include tools for faculty-driven self-service access and group management, and all storage spaces default to private to the research group only. Tools allow for easy access management and compartmentalization of research projects to authorized users in the group.

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.

[27] https://www.purdue.edu/securepurdue/

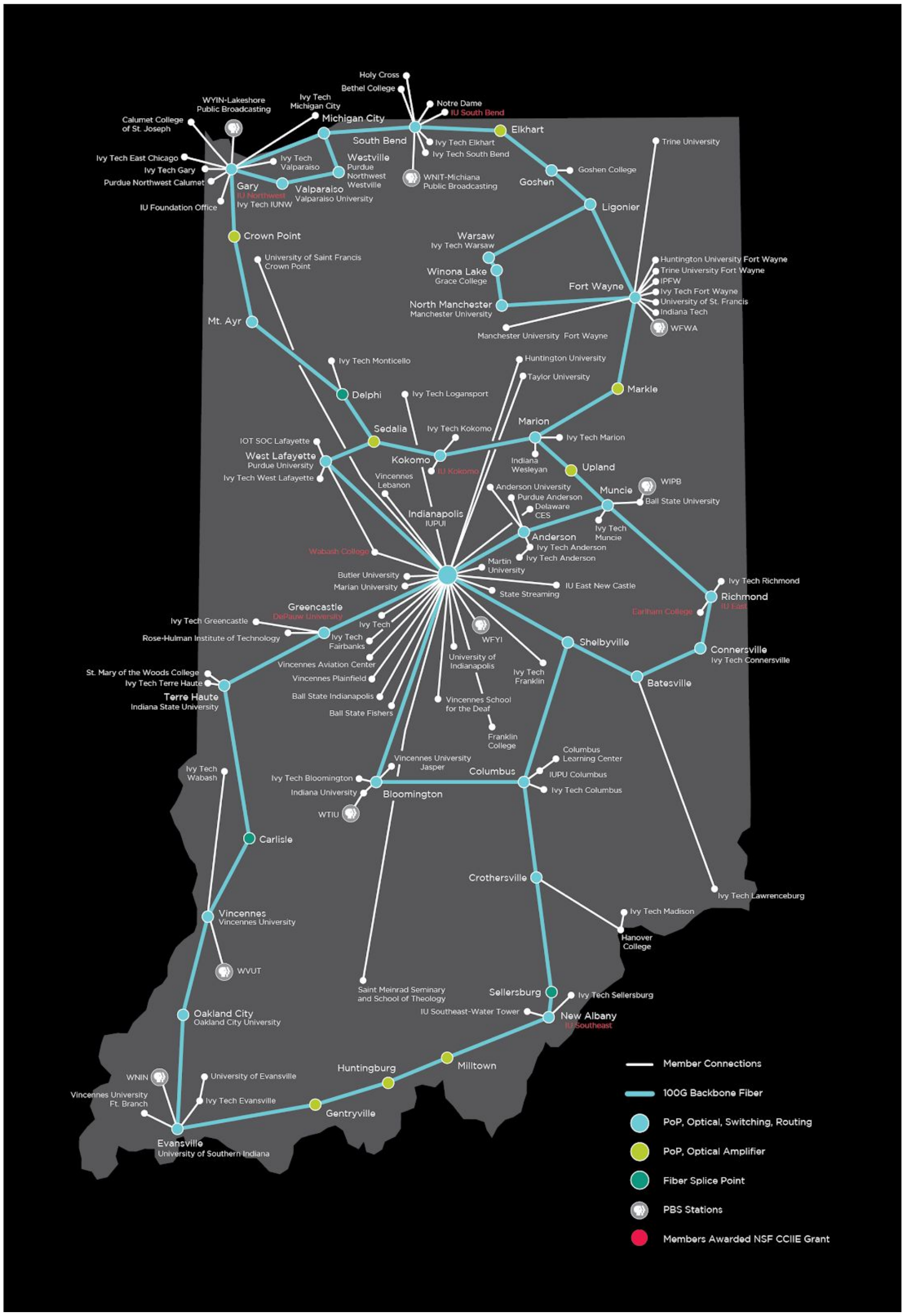
Purdue researchers are prominent members of international virtual organizations, like the LHC's Compact Muon Solenoid (CMS) experiment, and the Large Synoptic Survey Telescope (LSST). Purdue resources are integrated into these and other experiments' computing and data infrastructures.

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.

Appendix C - iLight Regional Networking Diagram



I-Light Network Map [28] [29]

[28] http://docs.globalnoc.iu.edu/uploads/68/3f/683fa39008b53a293991e3c272a6d481/ilightnetwork-v4.png

[29] https://carto.bldc.grnoc.iu.edu/worldview2/?map_type=google&map_set=Indiana%20Regional%20Networks&default_net=I-Light