Arcadia University Bioinformatics Application Deep Dive

April 3rd, 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.

Arcadia University Bioinformatics Application Deep Dive

Final Report

KINBERCON 2019
Philadelphia Pennsylvania
April 3rd, 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-2001317[1].

[1] https://escholarship.org/uc/item/1196z33x

Contents

Participants and Contributors

Hans Addleman, Indiana University
Zach Bare, KINBER
Matthew Flint, Arcadia University
Vitaly Ford, Arcadia University
Kristy Hamm, KINBER
Leslie Margolis, Arcadia University
Kenneth Miller, The Pennsylvania State University
Jennifer Oxenford, KINBER
Sheryl Reinhard, Duquesne University
John Zottola, Arcadia University
Jason Zurawski, ESnet

Report Editors

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

1. Executive Summary

In April 2019, staff members from the Engagement and Performance Operations Center (EPOC) and the Keystone Initiative for Network Based Education and Research (KINBER) met with researchers in bioinformatics at Arcadia University as part of a training exercise to perform an Application Deep Dive. The goal of this meeting was to help characterize the requirements team for a research team in bioinformatics, and to enable cyberinfrastructure support staff to better understand the needs of the researchers they support. Material for this event includes both the written documentation from the bioinformatics team at Arcadia University but also a writeup of the discussion that took place in person on April 3, 2019.

The case study highlighted the ongoing challenges that the Bioinformatics team has supporting a bioinformatics class that involves accessing data from a remote bioinformatics data source and using remote Galaxy compute resources. The lack of available local compute and storage resources means that they cannot fully demonstrate modern research techniques with students. There is a fair amount of interest from the faculty in exploring the availability of Cloud or other approaches to solving the lack of local compute and data resources.

Arcadia university received an NSF award to help support upgrading the campus network, specifically to include a Science DMZ and monitoring equipment. An update to the state network is also being planned, and details were discussed. As part of the overall review, a clear need was identified to identify and collaborate with regional or national providers for computational resources in addition to any that may exist locally. Additional challenges with securing sensitive data, cybersecurity, and supporting collaborations were also discussed.

Action items from the meeting included:

1) Working with community leaders to explore equipment options for networking, data transfer, measurement, security, computation, and storage at Arcadia University
2) Convening a review of the planned Arcadia University scientific network design with regional and national community leaders.
3) Working to increase availability of processing resources within the campus, region, and national space.
4) Explore advanced training in R&E community standards for facility and staff.
5) Complete the connectivity to PennREN, and explore peering arrangements that can better support scientific use cases.
6) Institute a performance measurement plan to better understand bottlenecks and set expectations with research community.

2. Process Overview and Summary

2.A 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 tool as 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 (online at https://fasterdata.es.net/science-dmz/science-and-network-requirements-review). The EPOC team adapted this approach to work with individual science groups through a set of structured data-centric conversations and questionnaires.

2.B Deep Dive Structure

Deep Dives are basically structured 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 researcher 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 to 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.C Arcadia University Bioinformatics Deep Dive Background

In April 2019, EPOC and KINBER organized a Deep Dive in collaboration with Arcadia University to characterize the requirements for a research team in bioinformatics. The Arcadia University representatives were asked to communicate and document their requirements in a case-study format (see Section 3). The use case for this deep dive was more closely related to educational use of bioinformatics tools than production of a research paper. The three time scales used by the research team also varied from standar- they hose to describe their work in the current form, for 6-9 months, and then beyond 9 months.

This project was one of the science drivers for the successful NSF Campus Cyberinfrastructure award, NSF #1827050, entitled "Transforming Arcadia's Networking Capability, Enhancing for Innovation to Grow Research Leaders in a Technology-driven World". This 3-year award for $352,500 started in 2018. The project is creating a Science DMZ with a data transmission network capable of 10Gbps connectivity (more than 10 times faster than current speeds) to the Keystone Initiative for Network Based Education and Research's (KINBER) PennREN network.

The CC* Project's objectives are to: (1) provide high performance, secure Science DMZ network capabilities for sharing of large datasets and cloud-based education; (2) eliminate technical barriers for faculty engaged in data-intensive projects through a dedicated, friction-free path to Internet2, PennREN, and other high performance computing and data resources; (3) leverage authentication and authorization mechanisms to support our faculty through the InCommon

Federation; and (4) enable future scientific possibilities and unleash innovation for students and faculty researchers. A list of science drivers is given in Appendix A.

The face-to-face meeting took place at KINBERCon on April 3, 2019 (see discussion in Section 4). We document next steps in Section 5.

2.D 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 the state'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.

The Keystone Initiative for Network Based Education and Research (KINBER), a Pennsylvania based non-profit corporation, provides broadband connectivity, fosters collaboration, and promotes the innovative use of digital technologies throughout the state. KINBER's PennREN network provides essential internet and research and education networking capabilities including connectivity to PennREN, Commodity Internet, Internet2, KINBER Peering and Caching Services, KINBER Member Exchange (KMEX), and more. As Pennsylvania's only statewide research, education, and community network, KINBER provides network-based connectivity and services to over 80 organizations and programming to many more, including higher education, K12, healthcare, communities, libraries, public media, museums, government, non-profit organizations, as well as commercial organizations consistent with its mission.

Founded in 1853, Arcadia University is Philadelphia's global university and a pioneer in international education. It is a top-ranked private university offering bachelor's, master's and doctoral degrees. In addition to six colleges and schools in the United States, it supports ten centers and offices around the world. Every year, faculty and staff enrich the lives of the 3,700 current students, 3,000 study abroad participants, and more than 26,000 alumni.

3. Bioinformatics Research and Education at Arcadia University Case Study

3.A Science Background

Arcadia University supports a Bioinformatics educational and research program, which was selected to focus of the application deep dive. The core researchers in this area are Drs. Carlos Ortiz, Kathy Macropol, and Sheryl Smith. This research focus area was one of the motivating applications for the successful NSF CC* proposal (#1827050), which is in part supporting the cyberinfrastructure upgrades needed by the university researchers.

The Computer Science and Math (CSMA) Department supports a 4-credit Bioinformatics course that is typically co-taught by two Arcadia faculty members: one from CSMA and another from the Biology Department. Participating undergraduates learn to take raw sequence data and through a process of analysis produce a high quality finished sequence, and then how to annotate genes and other features. This leads to using these techniques to address a specific question in genomics.

For a use case for the Deep Dive, the researchers chose to highlight how the students currently work with the national infrastructure as part of this course as it emphasized several of the ongoing challenges they experienced in their day to day research needs. At the current time, this analysis is performed by transferring a collection of genome data from the National Center for Biotechnology Information (NCBI), a sub-facility of the U.S. National Library of Medicine (NLM), to a public Galaxy (https://usegalaxy.org) processing center, where the experiments are run remotely. Results are then transferred back to Arcadia.

3.B Collaborators

The Arcadia faculty and students are members of the Genomics Education Partnership (GEP) program (https://gep.wustl.edu) that provides opportunities for undergraduate students to participate in genomics research in the classroom and for the purpose of capstone projects. GEP facilitates collaboration between a growing number of primarily undergraduate institutions around the country. There are numerous universities and colleges associated with GEP, and the Bioinformatics faculty at Arcadia frequently exchange research data with a subset of them. GEP helps to facilitate research projects and provides training / collaboration workshops for community college, college, and university faculty and their teaching assistants. There are over 100 members of GEP, shown in Figure 1, are located around the country and the world. Each member can participate in a variety of ways, such as donating time, resources, or expertise to advance the mission.

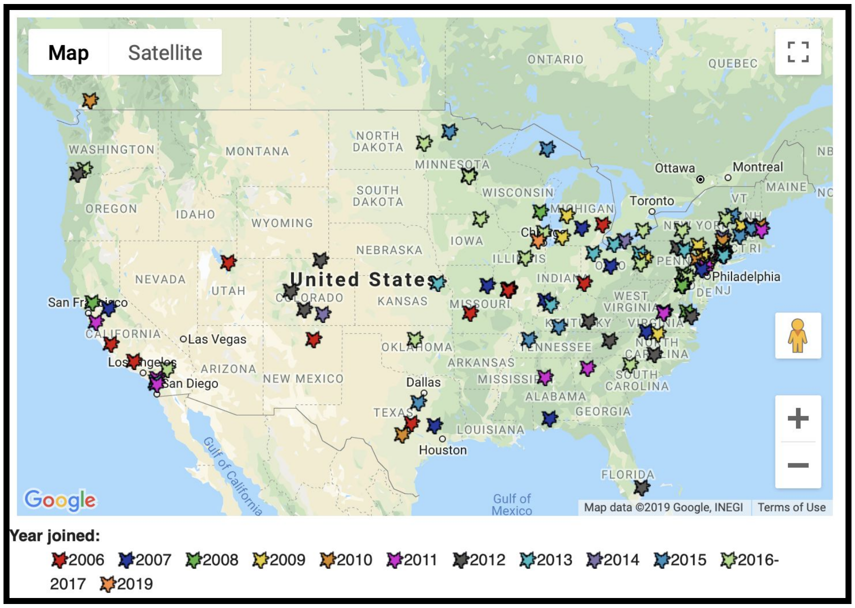


Figure 1: This map shows all members of the Genomics Education Partnership (GEP) program as of 2019.

Arcadia University is currently starting a project to establish network connectivity with KINBER's PennREN that has the potential to better support additional collaboration with other GEP members in Pennsylvania. Possible collaborations with GEP members that can be facilitated by KINBER within the state of Pennsylvania include:

- Geneva College
- University of Pittsburgh
- Washington and Jefferson College
- Lock Haven University of Pennsylvania
- Bucknell University
- The Pennsylvania State University
- Wilkes University
- Muhlenberg College
- Moravian College
- Widener University
- Eastern University
- Saint Joseph's University

3.C Instruments and Facilities

Arcadia University has not traditionally supported advanced resources for instruction. To date, individual faculty have supported their own advanced technology needs for R&E activities or they have used resources at external facilities.

Present

Laboratory Instrumentation

Genomic research equipment such as sequencers, microscopes, or other elements of wet-lab research is not all currently available on site. Because of this, instead of creating their own data, researchers have been using other existing data sets that were produced and curated in their research and education activities.

Computational Resources

At this time, most computational resources are not centrally maintained by Arcadia University staff. Most research faculty use a heterogeneous collection of limited local and remote options (e.g. cloud computing). Faculty have access to university provided workstations, that can be used to perform a small amount of computational work. The computer lab that can be used by general researchers is a collection of workstations that are reset daily to a known good state, and therefore cannot be used for storage or permanent software installations. One member of the faculty has set up their own small cluster, however this is fully used for a targeted research project and is not available for other uses.

Storage Resources

Currently, there is not a centrally-maintained storage resource for use by Arcadia staff. Research faculty use a collection of local and remote options (e.g. cloud storage). These efforts are not coordinated by the University.

Network Resources

Networking is provided by a 1Gbps commodity internet connection that is shared by the entire campus and supports all aspects of network connectivity, including both education, research, and administrative uses. Arcadia is working with KINBER to procure an additional 10Gbps of capacity to the PennRen network.

Next 6-9 Months

Laboratory Instrumentation

Research equipment is expected to grow on campus. The Bioinformatics faculty have discussed the purchase of more genomics sequencing equipment to facilitate advanced research. While the exact equipment has not yet been specified, current generations of this type of hardware can create 100s of GB of data per sequencing job, with data sizes compounding regularly. The current levels of Arcadia University computation and storage will not meet the expected needs of this scientific work.

Computational Resources

Arcadia University faculty and staff are discussing options of a purchase of a cluster to facilitate R&E activities on campus. This resource would be a general-purpose resource able to be used by many different user groups across campus.

Storage Resources

Storage has been identified as a concern for the future. Additional storage resources are required as data sets grow for a variety of research use cases. Campus and research groups will discuss ways that scalable storage can be added and upgraded.

Network Resources
Arcadia will be upgrading the networking equipment and growing the available capacity to a 10Gbps capable connection to KINBER's PennREN. A Science DMZ will be established to segment network traffic with an enterprise use case, from that of the emerging science use cases with additional support from NSF Grant No. 1827050, .

Beyond 9 Months
Data science education programs at Arcadia will continue to increase, and, as a result of such initiatives, we expect to work on hundreds of gigabytes of data every month. It is planned that this will be supported in part by the advanced connectivity enabled by the Science DMZ, along with access to other facilities via KINBER.

3.D Process of Science

The Arcadia University Bioinformatics program features a practical scientific use case: analysis and processing of a curated data set. All data that is currently being explored, is public and open, and not subject to any security or policy controls.

Present
A student team during the Fall 2017 semester were working with genomes from all of the influenza viruses and were trying to identify all the palindromes with lengths between 4 and 36. In order to accomplish this, approximately 3 million RNA sequences had to be downloaded and analyzed for palindromes. The team created a list with all possible coordinates of occurrences and identified the sequence where they occurred and the nature of the palindrome. Their program ran for more than two days on local machines and generated 15 GB of data, exhausting the available computational and storage resources.

Figure 2 shows the setup on local resources used in this experiment. Installing a local version of Galaxy on the Arcadia lab computer resources was not successful in practice due to several factors:

- Downloading the reference data sets from NIH took days on the legacy commodity network connection due to the lack of capacity and advanced connectivity
- The available storage was insufficient for the size of the requested data.
- The lab resources are 'reset' to a known good state daily, which means that they cannot offer persistent storage or non-standard software installs.
- The lab resources can take a day or longer to process the data sets, and in most cases were not able to complete before the laboratory environment needed to be reset for other users.

- Galaxy can be checkpointed to a specific state and resumed in cases where the processing exceeds a predefined limit. However, the checkpoint files are large, and there was not sufficient storage capacity available at Arcadia to store the checkpoint files.

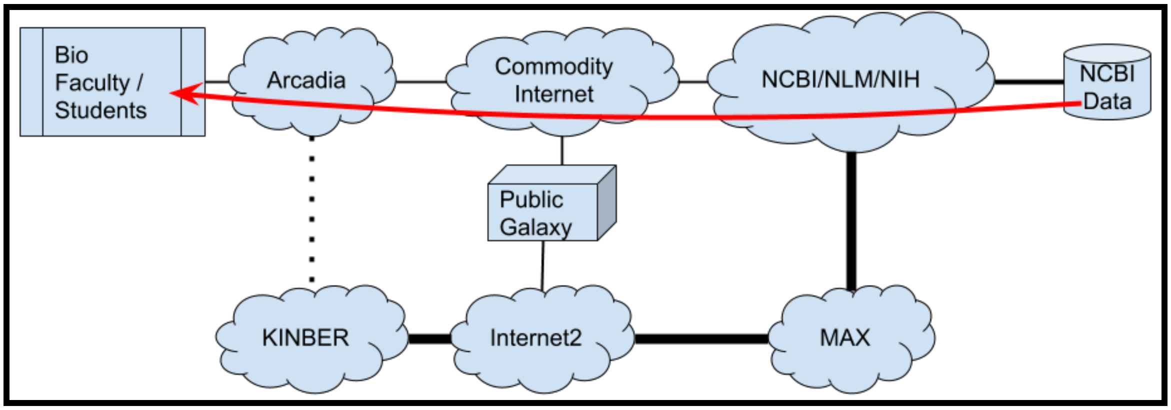


Figure 2: Attempted workflow using local Arcadia lab computing resources.

Because of the limitations in the locally available resources, a more common approach used currently is shown in Figure 3. This includes transferring a collection of genome data from the NCBI to a Galaxy processing center. The Galaxy processing center is operated by the Galaxy program as a service to users that do not support their own processing capabilities. This workflow relies on two factors outside of the control of Arcadia University, the availability of storage capacity at the Galaxy site to accept the curated data set and the availability of processing capacity at the Galaxy site to operate on the curated data set. Because of these constraints, the process can often take multiple days making these resources unusable for academic activities.

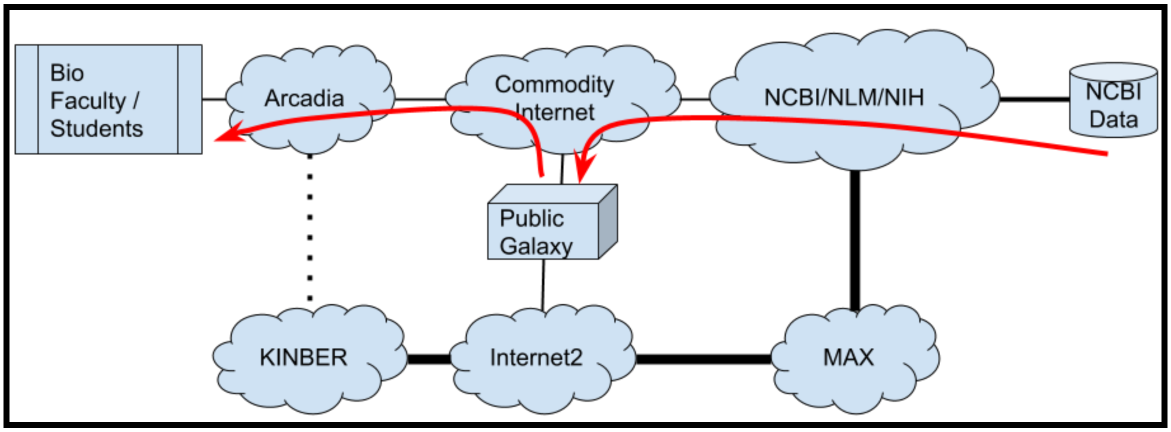


Figure 3: Current workflow for NCBI data to Arcadia using external resources at a Galaxy site.

Next 6-9 Months

Establishing the connectivity to KINBER will have a large impact on the network aspect of the workflow. This workflow, shown in Figure 4, includes access to a 10Gbps regional network, along with 100Gbps national connections, to ensure a fast path to data sets and computation.

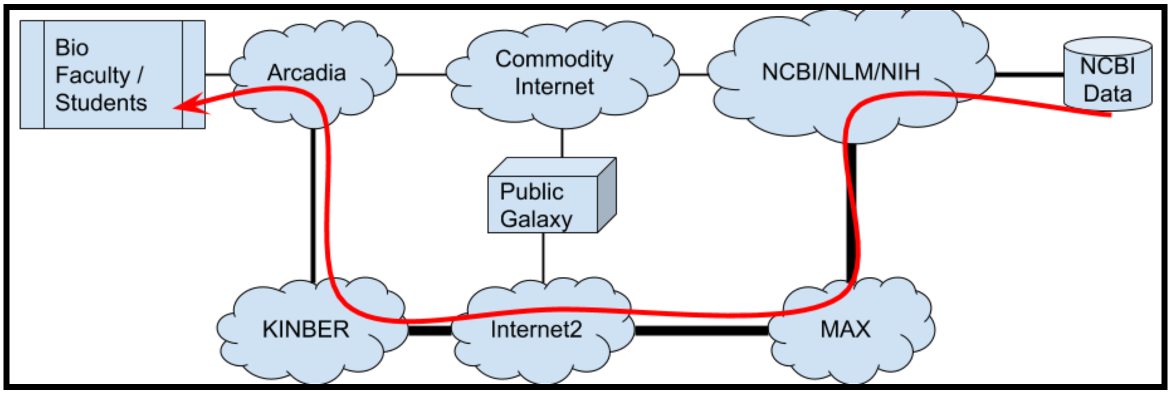


Figure 4: Bioinformatics workflow for NCBI data to Arcadia using external resources at a Galaxy site when KINBER connectivity is enabled.

There is a strong interest in using commercially-provided cloud resources (e.g. AWS, Azure) to run Galaxy and the Basic Local Alignment Search Tool (BLAST-https://blast.ncbi.nlm.nih.gov) algorithms , and there has already been some experimentation by the Arcadia faculty on this front This work is in the nascent stage and is motivated by the lack of a clear solution for storage and computation. Faculty are evaluating a variety of solutions in this space, including several commercial clouds (e.g. AWS, Microsoft Azure, Google Cloud), as well as those provided by R&E providers (CloudLab, Jetstream, Chameleon).   Some options for workflows with these resources are shown in Figure 5.

It is expected that access to external computation and storage will result in changes to the overall workflow. In particular, faculty and staff are expected to explore options to create/curate their own data sets locally, implying that the network will become a critical portion of the pipeline to transfer data in/out of the facility.

As the faculty shifts their computational jobs into the cloud, they will need better networking capabilities than Arcadia currently provides. Faculty and students work on multiple projects in the bioinformatics major, generating tens of gigabytes of data in the classroom for research projects that will need to be transferred to and from the cloud.

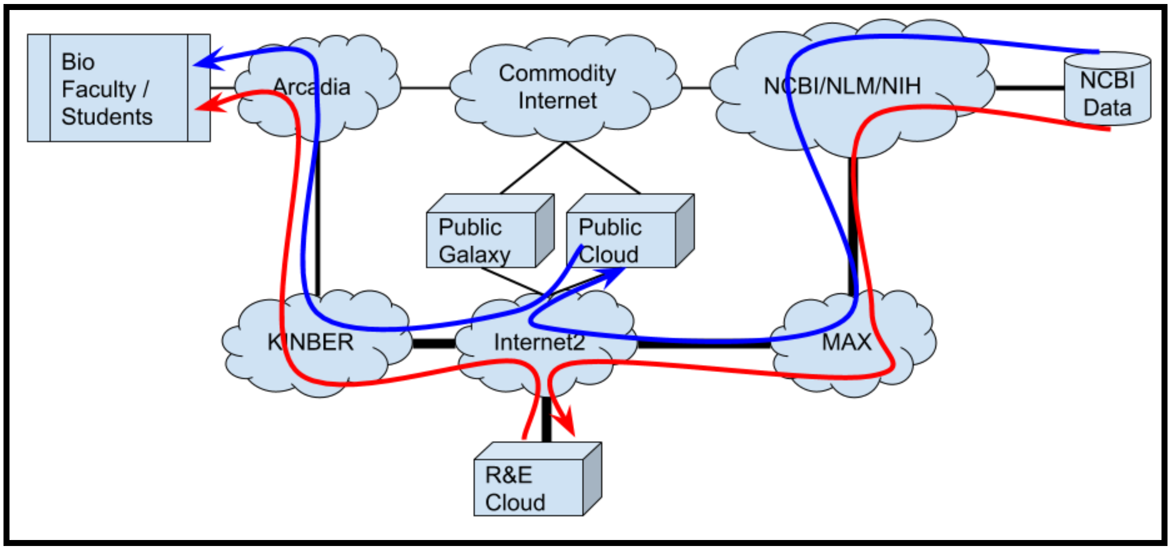


Figure 5: Bioinformatics workflow for NCBI data to Arcadia using Cloud resources.

Beyond 9 Months

The precision and quantity of data will increase in this area, placing more and more demands on the requirements of computation and storage. It is expected that the number of students and faculty going through this program will increase, adding a multiplier to resource requirements.

At this time it is unknown if the location of data sets, computation, and storage are local or remote. Collaborations with external research and industry entities will increase and rely heavily on services provided by KINBER.

3.E Remote Science Activities

At the current time Arcadia has limited external scientific drivers. The prior sections alluded to the current use cases and desires; this section will codify those in more detail. It is expected that the installation of the Science DMZ, and facilitation of more computational and storage resources, will usher in an era were external collaboration is easier to achieve and sustain.

Present

There are two primary external drivers for the Bioinformatics faculty at Arcadia University:

- Data sets are currently stored at NCBI
- Galaxy processing resources

For the current use cases, shown in Figures 2 and 3, the bottleneck is not the Arcadia local network. Because local storage resources are lacking, and the network to Arcadia is limited, the approach in Figure 3 is acceptable at this time for R&E activities.

Next 6-9 Months

Arcadia currently has plans to set up a Science DMZ to fulfil the funding awarded by the NSF. When that takes place, the workflows shown in Figures 4 and 5 will foster a faster and friction-free networking environment. This will facilitate easier access to external resources (e.g. data sets, cloud providers), and encourage more innovation from other scientific groups.

It is also expected that the adoption of local instrumentation will result in a data mobility requirement external to the campus, namely migration of produced data to external computational resources and the retrieval of results, as shown in Figure 6. This workflow is still in the nascent stages of planning and relies heavily on faculty getting access to local science instrumentation as well as storage resources to facilitate the workflow requirements.

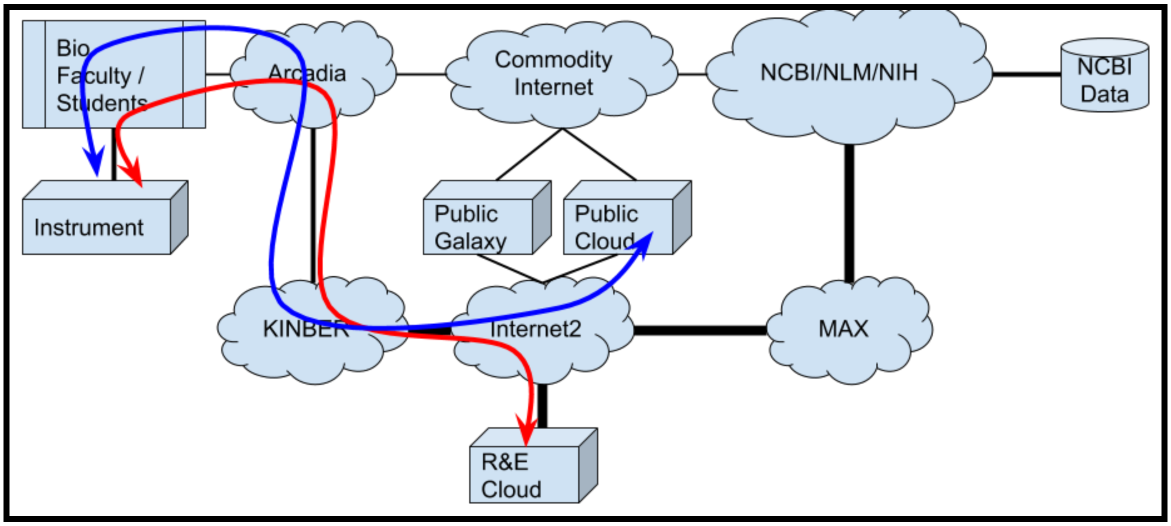


Figure 6: Bioinformatics workflow using local science resources and remote computation.

Beyond 9 Months

The requirement on networking to accelerate workflows will increase. As the data sizes increase, the reliance on external storage and compute will as well. It is infeasible for Arcadia to support all computation and storage locally, thus the PennREN connectivity to other locations in the state, and to national resources, will increase and become more critical to daily operations.

3.F Software Infrastructure

The Arcadia University Bioinformatics program uses a specific set of software components to achieve research goals. Most of these components come in the form of scripts and helper programs designed to interact with existing tools developed by others.

Present

The primary operating environments are Windows and Linux. Faculty workstations are administered by the resource owners, laboratory resources are administered by

Arcadia technology staff. Helper scripts and programs written in Python, Orange, Weka, Anaconda, etc, are used to manipulate and process data. The aforementioned Galaxy project (and software) are used, along with BLAST.

Next 6-9 Months

There is an established plan to adopt software CI components including Globus for data transfer and perfSONAR for network monitoring. Both will be integrated as a part of the Science DMZ infrastructure build.

There is a strong desire to bring Galaxy resources local, which would require building and maintaining a software/hardware infrastructure to support this. At the present time, students and faculty use publicly available Galaxy servers which are often overloaded and, as a result, the data processing jobs are extremely slow and sometimes halt with errors.

Beyond 9 Months

As workflows move toward cloud computation and storage, the number of locally generated "helper codes" will increase and require curation. Local faculty and staff are committed to storing and releasing most of these via mechanisms like Gitlab to facilitate development and sharing.

3.G Network and Data Architecture

Our goal is to connect researchers and students to a high capacity network that will allow them to utilize cloud computing, transfer large amounts of data, generate/store data locally, and share it with collaborators.

Present

The current infrastructure at Arcadia, shown in Figure 7, is in the process of being completely restructured. This includes upgrading outdated networking equipment, changing networking architecture, establishing and implementing new security policies and procedures, adding network storage, creating a Science DMZ, and increasing network bandwidth and redundancy. Some of the connectivity, shown in Figure 7, will not be present in the new version of network design as providers change and capacities are upgraded.

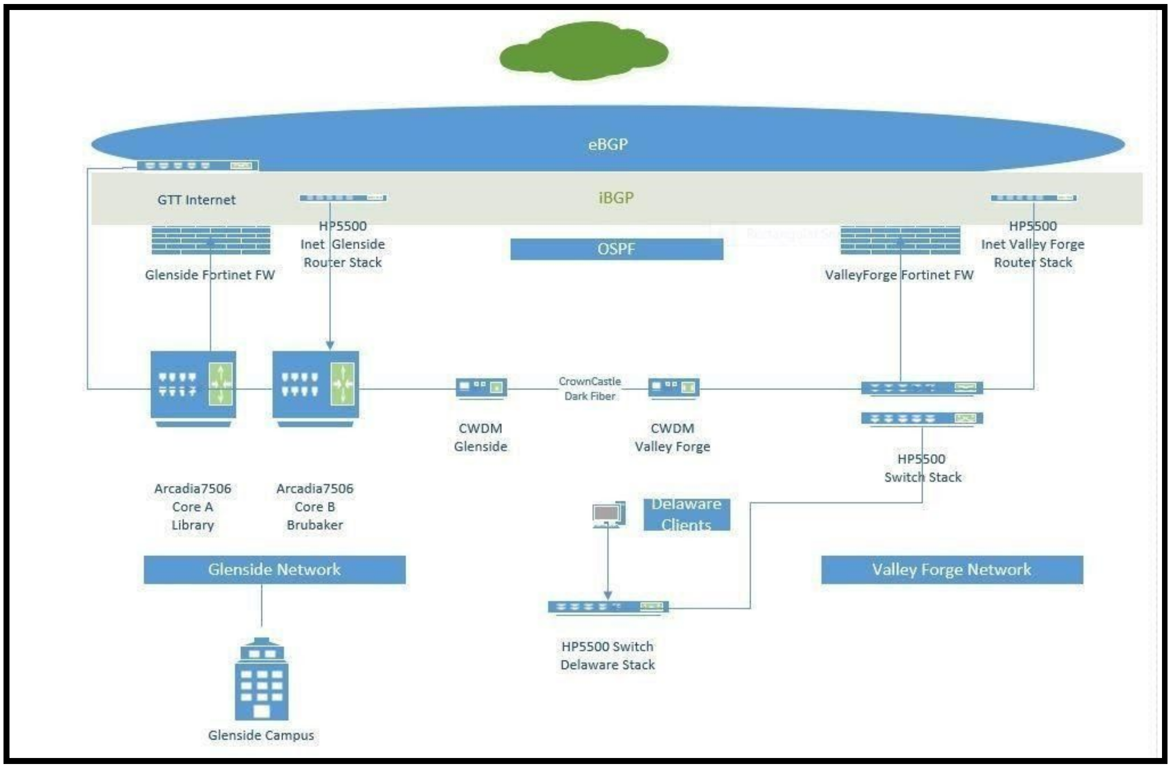


Figure 7: A diagram of the Arcadia infrastructure in early 2019.

Next 6-9 Months

Arcadia has been awarded an NSF Campus Cyberinfrastructure award (NSF Award #1827050) which will focus on creating a Science DMZ to enable a data transmission network capable of 10Gbps connectivity, which is more than 10 times faster than current speeds. This will also include:

- Upgrading external connectivity to KINBER
- IPv6 peering
- Updating data transfer hardware/software/storage
- Updating network monitoring hardware/software

Discussions are ongoing to settle on a design implementation that meets current needs and scales to future use cases. One possible architecture is shown in Figure 8.

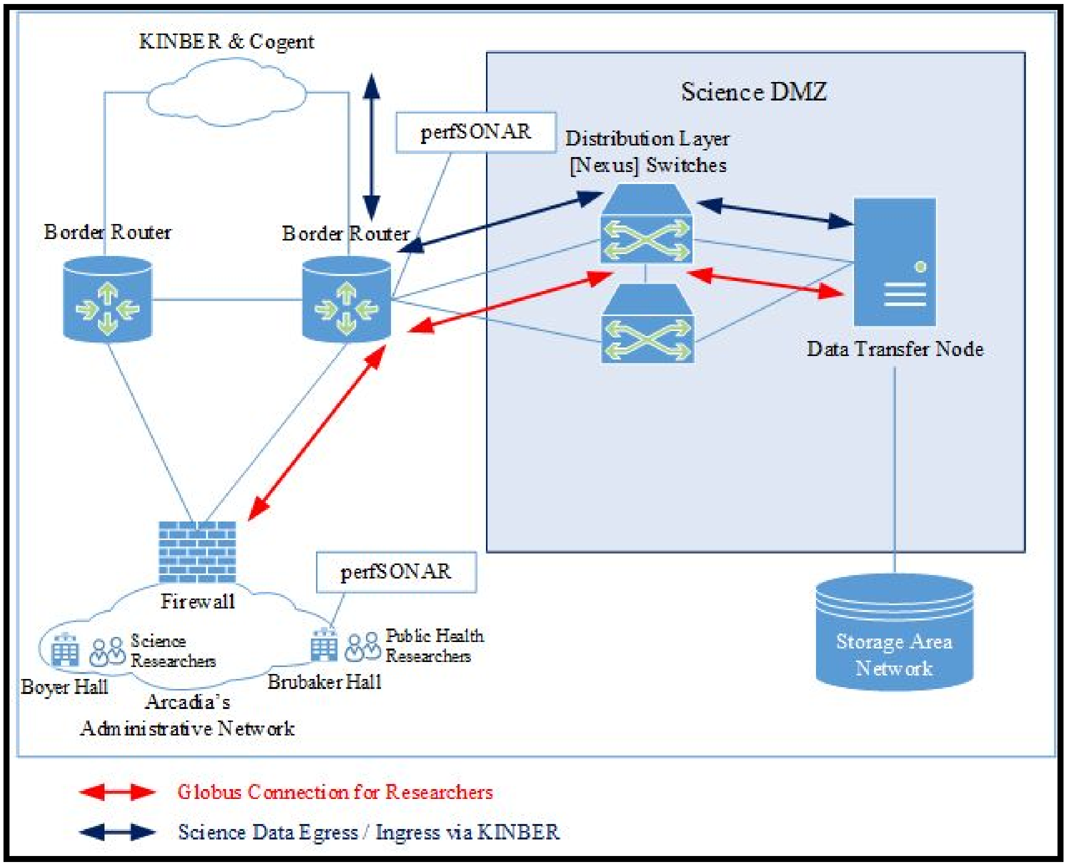


Figure 8: A diagram of the planned Arcadia science DMZ that is expected to be implemented in 2019.

Beyond 9 Months

The growth and usage of the network will be closely monitored, using such tools as perfSONAR, and plans for upgrades will track against research use cases. The 10Gbps connectivity to KINBER's PennREN is expected to scale for the short term, but could grow as needs dictate.

3.H Cloud Services

Present

The Galaxy use mentioned above can be classified as a 'cloud', albeit private and singly focused.

About 1.5 years ago, Arcadia joined the AWS Educate program. We have previously conducted training for moving research and teaching constituents into the Cloud and will continue doing so in the future. When the Science DMZ network is complete, there will be a more controlled network environment available to facilitate the exchange of research data to remote computational resources.

From the IT management and infrastructure perspectives, most Cloud services are still too costly. Arcadia does utilize some cloud services such as Panopto to facilitate online learning and lecture recording, but that is one of a few services that requires Cloud connectivity at this time.

Next 6-9 Months

Use of cloud resources is expected to increase, as costs drop and compute and storage requirements continue to increase, especially for undergraduate capstone and faculty research projects that rely on the availability of computational capabilities. Arcadia is exploring the use of R&E Cloud providers such as CloudLab, Jetstream, and Chameleon, along with shared resources from efforts like the Eastern Regional Network (ERN - http://ern.hpc.rutgers.edu/) is one path forward. Migration to new cloud infrastructures is not an easy task - and it is expected that faculty, staff, and student time will be used to migrate workflows and create software components.

Beyond 9 Months

The desire to use external resources such as more advanced Cloud services remains high, but concrete plans have not been established. Arcadia is exploring options for educational and research activities with external clouds (private, public, and research-focused), and these will become more clear in the coming years.

As workflows move toward cloud computation and storage, the process of science will change considerably. Local faculty and staff are committed to understanding these impacts.

3.I Known Resource Constraints

Present

With the recent award of NSF funding to facilitate the construction of a Science DMZ, there is considerable effort being utilized to understand requirements, specify and procure equipment, and physically prepare the campus and faculty. These discussions are fluid, and reflect both the original scope of the proposal, as well as community discussions led by KINBER. Storage and computation are identified needs, but may not be directly addressed by local solutions in this timeframe.

Next 6-9 Months

NSF Grant No. 1827050 will expire slightly beyond this time frame, barring any extensions, leaving behind viable options to support and accelerate scientific use cases. Arcadia expects to leverage other infrastructure grant opportunities for computation and storage, as well as support the individual grants of researchers that pursue them.

Beyond 9 Months

Arcadia faculty, staff, and student populations will be growing, and resources to support them will be a requirement.

3.J Parent & Affiliated Organizational Cooperation

Arcadia University leadership, faculty, and staff collaborated extensively on the NSF award. In addition to local cooperation, the KINBER regional network has provided technical and policy resources to prepare for the next stage of technology advancement through high speed research networking.

3.K Outstanding Issues

Currently there is no dedicated science network, and the current enterprise/business network experiences multiple reported issues throughout the year. These include, but are not limited to, equipment failures and performance slowdowns. It is believed that establishing the Science DMZ network will solve this major data transfer challenge.

As documented in the previous sections, Bioinformatics has identified clear use cases that will require more computing and storage resources to support existing and future projects. Additional drivers, including those identified in Appendix A, will also be able to take advantage of technology improvements related to networking and computation.

Because of the current state of the network, groups like Physics, Chemistry, and Computer Science, have been known to transfer data manually by physically copying data to hard drives and sharing/transporting it both locally and remotely. Many of these groups desire the ability to automate downloads, perform backups, or engage in other network-centric tasks, are not able to do so due to the lack of a stable and capable local network infrastructure.

Data security is an identified challenge for researchers on campus. There are multiple faculty who work with sensitive data and thus they require a certain level of security to comply with federal standards and procedures. Arcadia has recently hired a Senior Security Analyst and is in the process of re-evaluating their overall security posture.

4. Discussion Summary

On April 3, members of the EPOC team and staff from KINBER met with representatives from the Bioinformatics research area at Arcadia University. This review was held in Philadelphia, Pennsylvania, as part of the KINBER annual Meeting, KINBERCON.

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

- Arcadia University, in support of NSF Award #1827050, requested assistance from EPOC and KINBER to help facilitate the construction of a scientific network infrastructure and support cyberinfrastructure use cases.
    - Network equipment that is capable of facilitating scientific use cases (e.g. Wide Area, Local Area) will be a critical requirement to facilitate connectivity to the KINBER PennREN network.
    - Data transfer hardware and software, integrated with scientific workflows, will be used by a large population of faculty and students.
    - Guidance on the specification and use of measurement and monitoring infrastructure, provided by community standard solutions, that can provide visibility into performance implications.
    - Access to high performance computational equipment, either locally placed or available from national or regional providers, will accelerate the research output of faculty and students.
- Arcadia University is upgrading campus connectivity to include a second connection to the regional network KINBER.
    - Training in the proper configuration and management strategies for network traffic between KINBER's PennREN and a legacy provider will be required.
    - Traffic management is needed to ensure research prioritization, as well as access to commercial needs.
- Arcadia University has a clear need to identify sources of computational resources from regional and national providers, in addition to any that may exist locally.
    - They will work with EPOC, KINBER, and others to learn about and integrate solutions that explore R&E and commercial offerings.
    - They will continue the conversation with researchers on the needs and drivers of computational effort.
- Cybersecurity needs came up in several different contexts.
    - Arcadia University has requested facilitation of a meeting between local security staff and national experts to understand, prepare for, and support the research network needs.

    - Partnerships with organizations such as IU's Center for Applied Cybersecurity Research (CACR) will be critical in understanding and preparing for advanced threats.
- The Pennsylvania State University can assist Arcadia University in a research collaboration to provide computational resources for biological research via Galaxy.
- KINBER will explore possibilities of network peering arrangements to better serve Arcadia University use cases related to sensitive medical research
    - Medical research around the state of Pennsylvania, and surrounding regions, is increasing.
    - Facilitating network peering arrangements with public, private, non-profit, and for-profit entities will enable collaborative. opportunities for faculty and staff at Arcadia University as well as other KINBER member institutions.
- Arcadia University will work with KINBER to explore collaborative opportunities with the Eastern Research Network (ERN).
    - Sharing computational resources with larger schools (The Pennsylvania State University, Rutgers) may give Arcadia University faculty and staff assistance in handling scientific workflows.
    - New research opportunities through other ERN members will be explored.

5. Action Items

EPOC and KINBER recorded a set of action items from the Arcadia University Bioinformatics Application Deep Dive, continuing the ongoing support and collaboration. These are a reflection of the case study report, in person discussion, and items specifically related to the execution of the NSF award.

1. EPOC and KINBER will convene a review of Arcadia University scientific network design to provide feedback on the plan before the RFP begins.
2. EPOC and KINBER will assist in the specification of cyberinfrastructure support equipment for networking, data transfer, measurement, security, computation, and storage.
3. EPOC and KINBER will facilitate conversations with computational providers within the regional and national environments (including the ERN effort, GALAXY clusters at Penn State University, and potentially others).
4. Arcadia and KINBER will explore training opportunities for Arcadia University faculty and staff.
5. Arcadia and KINBER will complete peering arrangements with KINBER to facilitate faster R&E connectivity to the PennREN network.
6. Arcadia and KINBER will institute a performance measurement plan between Arcadia University, KINBER, and regional/national collaboration sources.

Appendix A –Science Drivers from NSF Award #1827050

CC* Network Design: Transforming Arcadia's Networking Capability, Enhancing for Innovation to Grow Research Leaders in a Technology-driven World

Principal Investigator: Leslie Margolis
Co - Principal Investigator: Vitaly Ford
Project dates: July 1, 2018 - June 30, 2020 (Estimated)
Awarded Amount: $352,500
https://www.nsf.gov/awardsearch/showAward?AWD_ID=1827050

Abstract:

As a small private university, Arcadia's existing computing infrastructure constrains the productivity of faculty in Bioinformatics, Computer Science, Chemistry, and Physics who are conducting data-intensive research. Specifically, the current infrastructure impedes researchers' ability to efficiently and securely access, share, or analyze large-data sets with collaborators at other institutions. To address these research and education needs, a collaborative team representing key university faculty and technologists at Arcadia is creating a Science DMZ with a data transmission network capable of 10Gbps connectivity (more than 10 times faster than current speeds) to the Keystone Initiative for Network Based Education and Research's (KINBER) PennREN network.

Project's objectives are to: (1) provide high performance, secure Science DMZ network capabilities for sharing of large datasets and cloud-based education; (2) eliminate technical barriers for faculty engaged in data-intensive projects through a dedicated, friction-free path to Internet2, PennREN, and other high performance computing and data resources; (3) leverage authentication and authorization mechanisms to support our faculty through the InCommon Federation; and (4) enable future scientific possibilities and unleash innovation for students and faculty researchers.

Arcadia is currently considering incorporating a data analytics requirement into its general curriculum and leverage the newly developed cyberinfrastructure to enable cloud-based opportunities, distance learning and researching on a global scale. This opportunity is supporting greater faculty and student analytical scholarship by forming a frictionless environment built to innovate and thrive in our technology-drive world.

Summary of Science Drivers and Network Requirements

Although a predominantly teaching institution, the majority of faculty at Arcadia University maintain active research programs to both contribute knowledge to their field and to integrate research into their educational activities. Faculty identified as *Science Drivers* are those with research programs in NSF-supported disciplines that are data or computationally intensive. Additional faculty who will benefit from increased networking capacity and capabilities are presented in the *Broader Impacts* section. These other researchers are affiliated with Arcadia's professional degree programs; their research is supported directly or by pass-through funds originating from the National Institutes of Health and other biomedical institutes. The scientific and educational projects have been the main drivers for this proposal. A summary Table 1 describes the needs and potential impact this project would have on *Science Drivers*.

Topic/Field	**Bioinformatics**	**Computer Science**	**Chemistry/Physics**
Need	Research and education	Research and education	Research
Researchers	Carlos Ortiz Kathy Macropol Sheryl Smith	Kathy Macropol Vitaly Ford Yanxia Jia	Emanuele Curotto Michael Wilson Tatjana Miletic
Resources	Genomics data for classroom assignments and research projects: 100 GB for storage.	Video, audio, image, and cloud data: 1 TB for storing and downloading.	Computational chemistry data generated from simulations: 1 TB for storage.
Current Limitations	Slow rate of genome data exchange with partner universities; limited network capacity for working with the cloud and storing the data locally.	Loss of instruction time due to slow download rate; constrained storage capacity.	Storage and data transfer rate requirements are not sufficient.
Expected Impact	Access to bigger datasets with faster access time; more scaled experiments; opportunity to transfer and securely store data to and from the cloud.	Opportunity to analyze and work with large datasets; more productive research collaborations; technical support for innovating around data analytics as part of Arcadia's general education outcome requirements.	Opportunity to access and securely store data for faculty and students to research; broader collaborations through more frequent data exchange.

Table 5 - Science Drivers' Summary

Bioinformatics

In the bioinformatics courses, faculty and students work closely with the Genomics Education Partnership (GEP) providing opportunities for undergraduate students to participate in genomics research in the classroom and capstone projects (research is conducted and courses are offered by Drs. Carlos Ortiz, Kathy Macropol, and Sheryl Smith). There are about 60 universities and colleges associated with GEP, and the bioinformatics faculty frequently exchange data with them.

Currently, genome data is transferred from the National Center for Biotechnology Information to a Galaxy Server processing center and experiments are run there. The bioinformatics faculty dream is to set up their own servers on the cloud for Galaxy and Basic Local Alignment Search Tool (under consideration are CloudLab or Amazon AWS) with the purpose of performing data-intensive computing on the data sets generated by their labs at Arcadia. The following is an example of the impact of the current network on research and education from the experience of a student team working in fall 2017 semester. Students were trying to identify all the palindromes with lengths between 4 and 36 in the genomes of all influenza viruses. Approximately 3 million RNA sequences had to be downloaded and analyzed for palindromes. The team created a list with all possible coordinates of occurrences and identified the sequence where they occurred and the nature of the palindrome. Their program ran for more than two days on a local machine and generated 15 GB of data, exhausting the available computational resources.

As the faculty shifts their computational jobs into the cloud, they will need better networking capabilities than Arcadia currently provides. Faculty and students work on multiple projects in the bioinformatics major, generating tens of Gigabytes of data in the classroom for research projects that will need to be transferred to and from the cloud.

Computer Science

The Computer Science and Math Department is in the process of integrating cloud-based technologies and a data analytics program into its curriculum. The department conducts faculty-student research in the area of machine learning and data mining. Examples of the data they work with include audio data for voice recognition, image data for image recognition, social media data from twitter and twitch, streaming data collected from camera and sensors attached to “Internet of Things” devices, such as Raspberry Pis and mobile phones. These projects require downloading tens of Gigabytes of data (e.g., images and audio files), crawling data (e.g., social media) from the Internet, and uploading data (e.g., sensor data and camera image and video data) to the cloud-based servers as well as storing data on local servers. Capstone computer science student projects often build web and mobile applications stored on the cloud.

Students in data mining courses and research capstones (taught by Dr. Kathy Macropol) regularly seek to develop solutions that require downloading portions of big data from such resources as Yahoo News Feed Dataset (13.5 TB), Wikipedia content/edit history (10 TB), Link Click Prediction Dataset (1 TB), and Google N-Grams (863 GB). Also, students in the database class regularly work with several Gigabytes of data at a time that they need to download for the assignments in the classroom. Arcadia does not have sufficient capacity to transfer and store large amounts of data. However, this proposed project will address these network requirements.

Major assignments in the operating systems and network security courses also require that students individually download between 15–20 GB of different operating systems and software to set up individual virtual infrastructures on their workstations. Currently, it takes between 3 to 5 one-hour long classes to finish downloading everything. Some computer science research projects work with large datasets from the Pecan Street Dataport[19] and Teradata (research by Dr. Vitaly Ford). All of the above projects demand strong infrastructure support to provide sufficient network bandwidth and large capacity for data storage. Additionally, the department has been restructuring its undergraduate curriculum by integrating Amazon AWS cloud assignments into the courses. In fall 2017, Arcadia established an education contract with AWS and all faculty and students can utilize their resources under the Amazon Educate program (students receive $75 credits and faculty receive $200 credits per year).

Improved cyberinfrastructure will also facilitate Arcadia's global and distant learning collaborations. The Computer Science and Mathematics Department has a strong academic and student exchange partnership with Jiangsu University, China, started in 2011. Jiangsu University is a highly ranked and prestigious doctoral research university. The program allows students to earn a Bachelor of Science in Mathematics from both Arcadia and Jiangsu Universities by completing three years at Jiangsu and one year at Arcadia. In 2015-2016 academic year, Arcadia welcomed 11 students from Jiangsu. In fall 2017, a new cohort of 15 undergraduates arrived for study at Arcadia. The Department is currently developing a new data analysis program for Arcadia and Jiangsu University students. The program will concentrate on big data analysis in different business sectors and study the applications of machine learning techniques. Additionally, enhanced networking capabilities will allow Arcadia to provide improved distance learning options for Jiangsu students and open a window for new opportunities to collaborate with Jiangsu through the China Education and Research Network (CERNET).

Chemistry/Physics

A number of faculty in the Chemistry Department are conducting computationally intensive research; specifically, they are carrying out two main types of simulations of gas phase and condensed phase matter at the atomic level. Quantum Monte Carlo simulations of gas phase and adsorbed matter are performed with the goal of establishing the importance of nuclear quantum effects on physical properties of interest. The main motivation for these efforts is to enhance the basic knowledge needed to engineer better energy storage devices. Two examples are: (1) the simulation of hydrogen clusters, and their isotopomers both in the gas phase and adsorbed on typical reticular metallo–organic frameworks to determine nuclear ground state properties; and (2) the simulation of lithium ions dissolved in a mixture of dipolar organic solvents typically found in lithium ion batteries to establish the nuclear quantum effects of the solute on the solvent response function and other dynamic properties.

The second type of simulation is electronic structure computations aimed at developing spectroscopically accurate potential energy surfaces for weakly bound gas phase species. For example, Dr. Emanuele Curotto is currently optimizing radial neural networks using supervised learning to improve current multi-state empirical valence both models for protonated water clusters and resolve outstanding controversies around the assignment of features in the infrared spectrum of these. Dr. Michael Wilson is making use of electronic structure computations to identify plausible mechanisms for several reactions of phenanthrene and its derivatives. Collaborators for their projects are local, e.g. Temple University and the University of Pittsburgh, national, e.g. Yale University, and international, e.g. Universita` degli studi dell'insubria, Como, Italy. Storage requirements are around a Terabyte, whereas the data transfer from our own servers and cloud–based computing centers can require 10 GB at a time. By improving our cyberinfrastructure, these faculty and their colleagues would be able to extend existing and develop new scientific collaborations enabled by exchanging data and results without delays.

Arcadia faculty enjoy well-established research relationships with collaborators at the University of Pittsburgh, Drexel University, Indiana University of Pennsylvania, etc.

Eleven faculty members of the College of Arts and Sciences brought in nearly half a million dollars in funding from various funding agencies. This amount represents fifteen funded projects during the FY2013-FY2017 time period.

Dr. Emanuele Curotto, who is serving as a collaborator on this NSF CC* project, has a currently active research project titled “Quantum Simulations of Lithium Ion Solvation Dynamics in Mixed Stockmayer Clusters.” with funding from the American Chemical Society, Petroleum Research Fund.

Dr. Tatjana Miletic, another one of our collaborators had received funding from the National Science Foundation as a sub-award from Drexel University. Her project was titled “Systems for Precise Neutrino Detection with the Double Chooz Detectors”.

We also highlight Dr. Tatjana Miletic, a member of the Double Chooz Collaboration and the DarkSide collaboration, who conducts her research via these collaborations and at Drexel University. She is supported by a subaward from Drexel for an NSF funded project *Systems for Precise Neutrino Detection with the Double Chooz Detectors*. Her experience in computationally intensive, large-scale collaborative projects will be beneficial to the project team and other faculty at Arcadia.

With current (and anticipated future) collaborations among Arcadia faculty and colleagues at institutions such as Temple University, University of Pittsburgh, Yale University, and the Universita` degli studi dell'insubria in Como, Italy, demand for improved cyberinfrastructure and specifically networking and storage capacity is essential. Storage requirements are around a Terabyte, whereas the data transfer from our own servers, and cloud-based computing centers can require 10 GB at a time. By improving our cyberinfrastructure, these faculty and their colleagues would be able to extend existing and develop new scientific collaborations enabled by exchanging data and results without delays.

Broader Impacts

This project's outcomes are aligned with two themes and their associated goals in Arcadia's strategic plan: (1) Enhancing Academic Excellence and (2) Improving the University's Resources and Infrastructure. One of the goals of Enhancing Academic Excellence theme is to "institute a comprehensive professional development program for faculty and increase organizational support and resources that enhance faculty and student research, scholarship and creative activities within and beyond the University and the nation". One of the goals of Improving the University's Resources and Infrastructure theme is to "enhance the learning and technological infrastructure required for a vibrant academic environment and student life".

Arcadia University has built and supported interest in data analytics in faculty research, in joint student and faculty research, and in its continued dedication to collaboration with other universities and scholars. Within each department's hiring processes, Arcadia actively considers scholars committed to data innovation in all content areas, in order to both support student interest and learning and to creatively enhance our general educations curriculum's Quantitative Reasoning (QR) requirements. As a result of Arcadia's focus, the total number of analytic "QR" courses offered each semester has risen in the last several years as the use of data-driven research has become embedded in our programs.

In taking the next steps in support of such innovation, the improved cyberinfrastructure will support greater faculty and student analytical scholarship as the infrastructure further blossoms, and will provide new opportunities in which both faculty and students will take advantage of collaboration with each other and with students and scholars across the US. Likewise, the infrastructure developed will help to build opportunities for Arcadia to not only generate new partnerships through research, but also innovate around data analytics as part of our general education outcome requirements.

Beyond Arcadia's academic strategic plan, the project will impact other research and instructional activities that leverage the cyberinfrastructure and open cloud-based opportunities for STEM and liberal arts majors, enabling distant learning and research with global engagement, facilitating faculty and student retention, and allowing for more efficient scientific collaborations with support from KINBER. In addition to those faculty identified as the *Science Drivers* motivating this project, additional faculty research, student capstone experiences and student research projects in other disciplines that use large datasets would become possible, for example those in the College of Health Sciences. These faculty enjoy fruitful collaborations with colleagues from the University of Maryland, Baltimore County, University of Colorado Denver, University of Oregon, and University of Massachusetts, Worcester among others and their research is supported through subawards from federal prime awards to their colleagues. Since 2013, fifteen funded projects have been supported by $1.5 million. For example, public health researchers at Arcadia use data systems/structures, but our existing infrastructure limits the ability to ask questions that cut across countries and continents. Specifically, enhancing the capacity for data transfer from the Demographic Health Surveys online data warehouse as well as Arcadia's ability to store and share this data internally would enable students to explore and answer questions that may be more globally relevant and impactful.

Other work is conducted in collaboration with researchers at Fox Chase Cancer Center (FCCC)/Temple Health to integrate and support family caregivers in cancer care requires transferring study data and communication according to FCCC IRB and HIPAA requirements. Collecting patient and caregiver data on preferences for involving caregivers and an IRB protocol is in the development to explore system changes to involve caregivers. This collaboration requires secure data transfer and storage (once transferred). This work with the National Alliance for Caregiving, National Cancer Institute (NCI), and Greenwald and Associates market research firm uses a large de-identified data set of caregivers in the U.S. and was recently awarded NCI support (to FCCC). This will involve data-related communication with the market research firm. A new project with the National Alliance for Caregiving and other non-profit organizations to explore palliative cancer care will involve data sharing among the organizations as well as student engagement in the data analysis. Additional upcoming projects include the collection of population-level data via the Internet. To be successful in these projects will require advanced network capacity to display content and field surveys as well as secure data storage and transfer.

The Physical Therapy program also has extensive data transfer needs. "Muscle Mechanisms Underlying Recovery of Function after Hip Fracture" is an ancillary study to a larger federally funded trial. This project would benefit from increased network capacity as subject video files are too large to upload to the cloud and require additional security to protect identities. Therefore, there is a need for a secure and privacy-preserving way to transfer and analyze the data in the cloud.

One of Arcadia's Physical Therapy researchers collects heart rate data for the home health physical therapy research. Heart rate is sampled every second in collaboration with the University of Colorado, Denver. In an hour-long session, they have 3,600 data points for session for one person. In the trials they are conducting, there are 12 sessions of exercise for 150 people yielding 6.48 million points for a single study. The ability to download and analyze that data is not possible at this time because of capacity and privacy concerns with human subject research.

Currently, Arcadia's researcher has to travel to Denver to see data and she has no way for colleagues at Arcadia to help analyze it.

In conclusion, improved network capacity at Arcadia University will facilitate our faculties research and instructional activities, and so contribute to the knowledge base in their respective fields and contribute to the development of students' data analytic skills to benefit the future US workforce.

Appendix B – Arcadia University Cyberinfrastructure Plan[2]

Introduction

Information Technology (IT) at Arcadia University provides services to faculty, staff, and students across two schools and 3 colleges including Arts and Sciences, Health Sciences, Education, Global Business, and Global Studies. Additionally, a large study abroad program is offered through its College of Global Studies that operates in 13 countries.

Arcadia's main campus location is located in Glenside, PA and a satellite office operates in Christiana, DE (refer to Fig. 1). The Glenside campus has two server rooms and a disaster recovery site in a data center 20 miles west of Glenside in Valley Forge, PA. The Delaware satellite has a small server room/network closet that contains all equipment to support that site.

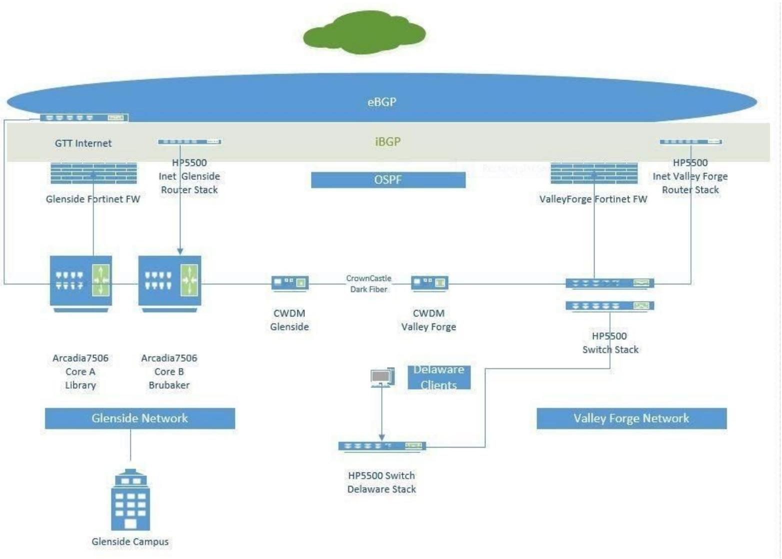


Figure 9 - Arcadia's Enterprise Network

In 2012, the Arcadia contracted R&R Voice and Data, Inc. (R&R) to provide a complete upgrade and overhaul of the Glenside campus' fiber infrastructure. The networking infrastructure specified by Arcadia University deployed a Cisco 3-tier model: Core–Distribution–Access layer switches. In doing so, the university moved from a single core switch, located in Boyer Hall, to a two-core switching environment, one each located in Landman Library and Brubaker Hall. The network was designed with a "no single point of failure" methodology.

[2] As submitted for NSF Award #1827050, see also https://www.nsf.gov/awardsearch/showAward?AWD_ID=1827050

On Arcadia University's main campus there is a 24-pair redundant fiber ring that connects all of the campus buildings together. Currently, only 2 pairs (4 strands) are utilized on the network which has been adequate to support administrative needs. Arcadia has dark fiber that connects the Valley Forge co-location with the Glenside campus and a site-to-site VPN that connects the Delaware office to the Glenside campus. Arcadia provides a 1 Gbps Internet connection for the students, 600 Mbps connection for faculty/staff, and 400 Mbps connection at the Delaware campus.

Current Enterprise Network Environment

The University's LAN environment comprises three levels of switching, covering a typical collapsed core to distribution to edge model. Being a "collapsed" core, all routing for any remote sites is performed within the Core Router/Switch itself for all of the VLANs, (Virtual LAN Interfaces, supporting various routed networks).

The Core consists of two (2) Hewlett Packard 7506 Ethernet/Routing chassis with an assortment of blades and modules to support supervisory functions as well as providing fiber links to down level distribution switches. Both chassis are bonded as a function of HP's Intelligent Resilient Framework (IRF) technology which creates a large IRF fabric from multiple switches to provide data center class availability and scalability. Essentially, this technology bonds each chassis into a single "Virtual" chassis. This bonded function allows for a failover to occur for both routing and switching protecting all of the core functions to provide continuous operation should a failure occur.

Arcadia's IT staff has facilitated links to the distribution switches to be redundant, allowing for two paths in 10 GE (10 Gigabit Ethernet) bonded trunks, or Link Aggregation Control Protocol (LACP). With this in place, should one core suffer a failure, the LACP-bonded interfaces would continue to flow traffic to the other chassis without interruption. Each distribution location consists of a pair of HP 3800 or 5500 Switches, which are also stacked or bonded together to provide a single "Virtual Switch." As in the Core, the bonded HP 3800 or 5500 switch will continue to support their uplinks to the HP 7506 Core as well as their downlinked Edge switches.

Arcadia owns Autonomous System Number (ASN) 30231 and IPv4 and IPv6 blocks from ARIN information below:

```
ARCADIA-UNIVERSITY (NET-74-113-108-0-1)
74.113.108.0 - 74.113.111.255
ARCADIA-UNIVERSITY (NET6-2620-121-4000-1)
2620:121:4000:: - 2620:121:40FF:FFFF:FFFF:FFFF:FFFF:FFFF
```

Wireless Access

In the past year, Arcadia implemented HPE Clearpass in conjunction with joining Eduroam. A complete rebuild of the wireless network was performed. The Clearpass replaced the Bradford Network Access Control that was just on the student network.

Currently, the Clearpass is deployed across that network, handling the Network Access Control for faculty, staff, students, and guests. Secure access for BYOD devices is provided as well. Role-based policies, enterprise-grade AAA with RADIUS/TACACS+, built-in device profiling, Apple Bonjour and DLNA device sharing, and integration with third-party MDM solutions – all through a single web interface.

Firewall and Data Storage

Arcadia has Fortigate 800D Firewalls located in front of each of the Internet Routers. One for faculty/staff and another for students. The Internet Routers are a redundant stacked pair, but the availability of the Firewalls directly affects the availability of the Internet access on campus.

Arcadia has a capacity of 50 TB of an HPE 3PAR Storage Area Network (SAN). The SAN is used for storing faculty and student files. Additionally, all backups on the network are stored in the SAN as well.

Servers

In 2010, the IT department deployed a VMWare ESXi Cluster to host different specialized applications on request.

Identity and Access Management

Current authentication aspects include: Eduroam, G Suite, Self-Service, AD, LDAP, and CAS.

At Arcadia, there is an internal push to start moving to Software as a Service applications (SaaS). The first we will be implementing next month, migrating from internally hosted Client Access Server (CAS) for Single Sign-On over to a cloud identity management platform, OneLogin. All access will be SAML or Shibboleth based in preparation for InCommon Federation.

Phone System and Surveillance Video

In 2010 Arcadia contracted R&R to implement security cameras across campus. This involved connecting the cameras to the network as well as storage for all these 290 cameras. All cameras are POE cameras with static names per location.

In 2015 Arcadia worked with ShorTel, now Mitel, to implement a pilot VoIP system. Currently we have a dozen phones on the network using a VoIP VLAN. At some point in the future we will increase that footprint. All the network configuration is completed and management of system is performed via VMWare ShorTel appliance.

Education Applications

Arcadia outsources educational platforms to Canvas and Google Suite Apps including its email system. In addition, Arcadia provides faculty with access to distant learning tools such as Panopto Recordings, BlueJeans, and BigBlueButton. At the same time, Arcadia hosts Rolodex, Event Management System, and Self-Service on the premise.

Planned Arcadia Science Network Environment

Objectives

The main goal of Arcadia's Cyberinfrastructure Plan is to transform the institution's infrastructure to meet the increasing demands of data-driven research and education conducted by its faculty and students. Arcadia desires the ability to perform high speed data transfers that the researchers currently need and that faculty will introduce into the classroom as part of the academic plan to integrate data analytics skills into all of our undergraduate studies. The new capability will be called Arcadia Science Network. From a University-wide perspective, we are moving toward an intentional enterprise architecture that addresses research, education, and administrative needs. The enterprise architecture adopts a cloud-first approach; includes a science DMZ and network; and aligns our security, services, applications and monitoring. This will shift our design from on-campus tools, storage and perimeter focus to one operating in a hybrid model that moves seamlessly between internal and external capabilities. The project's objectives include the following:

1. Provide high performance, secure ESnet Science DMZ network capabilities for large datasets sharing and cloud-based education.
2. Eliminate the technical barriers for faculty through dedicated, friction free path to Internet2, PennREN, and other high performance computing and data resources.
3. Leverage authentication and authorization mechanisms to support our faculty science drivers and join InCommon Federation.
4. Enable the future scientific possibilities and unleash innovation for student and faculty researchers.

Planned Deliverables for the Proposed Arcadia Science Network

To address the above-mentioned objectives, we will pursue the following deliverables.

- Implement the proposed Science DMZ design.
- Upgrade the backbone connections of the network from 1 to 10 Gbps. We plan to quote different Bandwidth Service Providers to select the best fit for dark fiber and providing a pipeline to Internet2 and PennREN
- Add Cisco 9500 cores to each of the server rooms, Glenside campus, with 40 Gbps interconnect to each building. We plan to connect 2 buildings to the Science DMZ as a Proof of Concept: Brubaker and Boyer Halls. Both of those buildings house data-intensive curriculum and research projects. Boyer houses Computer Science and Math Departments and Brubaker houses Physical Therapy and Public Health Departments.

- Implement perfSONAR, the networking performance measurement tool.
- Complete our path to InCommon Federation and identity management for services and applications.
- Implement IPv6 strategy within the new Arcadia Science Network as the Arcadia Administrative Network is not ready to handle transition from IPv4.
- Implement security mechanisms on the perimeter.

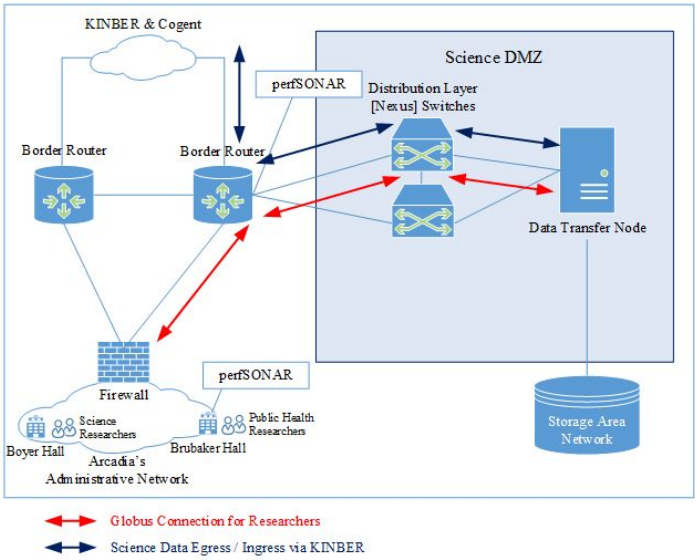


Figure 10 - Proposed Arcadia Science Network

The plan is to bring the proposed Arcadia Science Network (High speed Science DMZ network) up alongside the current administrative network. The new science network will provide 10 Gbps connectivity to two crucial locations on campus that are major hubs of high performance networking research and education needs (Boyer and Brubaker Halls). The new Cisco ASR border routers will replace the current border routers (limited to 1 Gbps) to provide support for 40 Gbps. This will allow our current infrastructure to keep functioning while the new Arcadia Science Network is being built from the ground up. Once the new cyberinfrastructure for Acardia's Science Network is running in a high quality manner, future steps will include providing the enhanced capability to all areas and campus'

Additional Network Monitoring

Each connection to Arcadia's Administrative Network will have a Cisco Firewall before the Cisco ASR Border Routers that have access to the Internet. This adds an extra layer of security between the university and the Arcadia Science Network. We will be able to monitor the activity at the Cisco ASR router level within perfSONAR. The Cisco ASR Border Routers connecting to the Science DMZ will not be passing through the Administrative Network firewall to allow science traffic to remain friction-free. Within the Science DMZ we will have our distribution layers switches and a Data Transfer Node that will have a direct connection to the SAN.

We plan to utilize:

- perfSONAR will be installed to monitor activity at the distribution level within the Science DMZ.
- Splunk which offers us the ability to search, monitor, and analyze machine-generated big data (including the network traffic) from one centralized web page.
- Snort in conjunction with Cisco networking equipment. Snort is a free open source network Intrusion Detection System (IDS) and Intrusion Prevention System (IPS).
- Rapid7 Nexpose to monitor internal and external address for any security concerns.
- SolarWinds to monitor all of its servers and network devices with the purpose of centralizing its monitoring from the several locations that are currently being used.

Enhanced Identity Management and Security

We are currently implementing One-Login, a cloud-based SSO solution. We have focused on SAML and Shibboleth and are currently authenticating with eduroam.

At a device level, we have to Dell EndPoint Security Suite Enterprise System (ESSE). This included both AntiVirus and encryption products, which will allow for managing authentication and encryption as well as preventing malware via one centralized, remote console.

We will work with KINBER, as our Leadership Institution, and our external consultants to assist us in the effective configuration of our border router's Access Control Lists and join InCommon Federation.

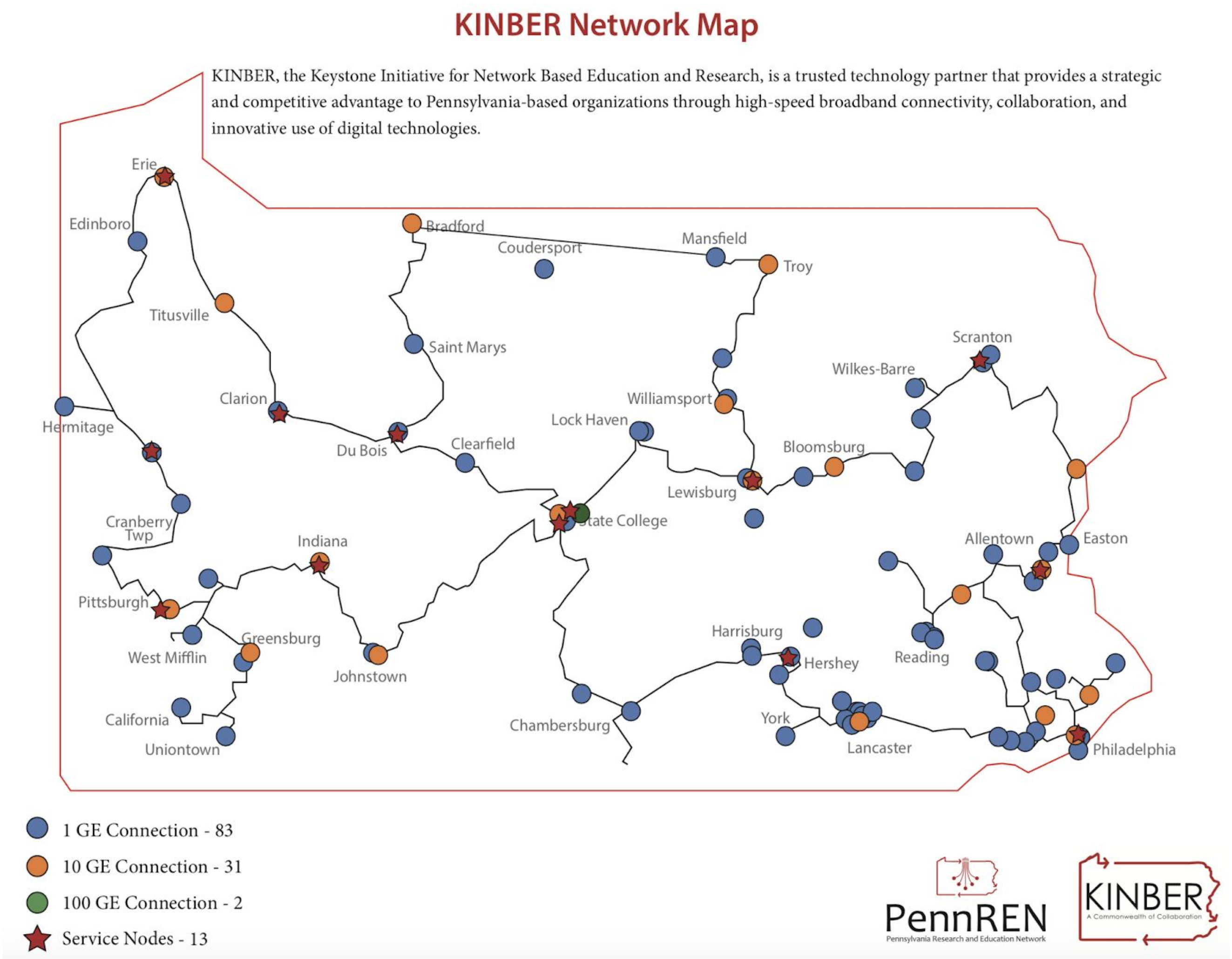
With respect to physical security, the campus is outfitted with key card access and cameras throughout. Data centers and network closets are secured with limited access based on roles.

Finally, we have a third-party provider scanning our environment and as we enhance the network, we intend to contract for periodic intrusion testing.

Sustainability

We feel confident that this grant will position us well for sustainability as we put the cyberinfrastructure in place to manage our growth through a dark fiber option and alignment of our authentication method to InCommon. Please also see additional information in Project Description Section 2f, Sustainability Analysis.

Appendix C – KINBER's PennREN Regional Networking Diagram



KINBER Network Map

KINBER, the Keystone Initiative for Network Based Education and Research, is a trusted technology partner that provides a strategic and competitive advantage to Pennsylvania-based organizations through high-speed broadband connectivity, collaboration, and innovative use of digital technologies.

Appendix D – List of KINBER's PennREN Connected Institutions

- ***Higher Education***
    - Allegheny College
    - Bloomsburg University of Pennsylvania
    - Bucknell University
    - Bucknell University Small Business Development Center
    - Bucks County Community College
    - Butler County Community College
    - California University of Pennsylvania
    - Carnegie Mellon University
    - Cheyney University of Pennsylvania
    - Clarion University of Pennsylvania
    - Community College of Beaver County
    - Community College of Philadelphia
    - Delaware County Community College
    - Duquesne University
    - Drexel University
    - East Stroudsburg University of Pennsylvania
    - Edinboro University of Pennsylvania
    - Franklin and Marshall College
    - Harrisburg Area Community College
    - Harrisburg Area Community College Gettysburg Campus
    - Harrisburg Area Community College Lancaster Campus
    - Harrisburg Area Community College Lebanon Campus
    - Harrisburg Area Community College Midtown Campus
    - Harrisburg Area Community College York Campus
    - Indiana University of Pennsylvania
    - Jefferson
    - Kutztown University of Pennsylvania
    - Lafayette College
    - La Salle University
    - Lehigh Carbon Community College
    - Lehigh University
    - Lock Haven University of Pennsylvania
    - Lock Haven University of Pennsylvania Clearfield
    - Luzerne County Community College
    - Mansfield University of Pennsylvania
    - Millersville University of Pennsylvania
    - Monell Chemical Senses Center
    - Montgomery County Community College
    - Northampton County Community College
    - PASSHE Center City Multi Campus

    - Penn State Abington
    - Penn State Beaver
    - Penn State Behrend
    - Penn State Berks
    - Penn State Brandywine
    - Penn State DuBois
    - Penn State Fayette
    - Penn State Greater Allegheny
    - Penn State Harrisburg
    - Penn State Hazleton
    - Penn State Hershey
    - Penn State Lehigh Valley
    - Penn State Mont Alto
    - Penn State Schuylkill
    - Penn State Shenango
    - Penn State University Park
    - Penn State Wilkes-Barre
    - Penn State Worthington Scranton
    - Penn State York
    - Pennsylvania College of Technology
    - Pennsylvania Highlands Community College
    - Reading Area Community College
    - Shippensburg University of Pennsylvania
    - Slippery Rock University of Pennsylvania
    - Susquehanna University
    - Thaddeus Stevens College of Technology
    - Thiel College
    - University of Pittsburgh Bradford
    - University of Pittsburgh Greensburg
    - University of Pittsburgh Johnstown
    - University of Pittsburgh Pymatuning
    - University of Pittsburgh Titusville
    - University of Scranton
    - Villanova University
    - West Chester University
    - Westmoreland County Community College
- ***K-12***
    - Fannett Metal High School
    - Hempfield School District
    - Penn Manor School District
    - The Hill School
    - Troy Area School District
- ***Hospitals/Healthcare***
    - Butler Memorial Hospital
    - Geisinger Health System

    - Nova Stream/Lancaster General Health
- ***Library/Library System***
    - Centre County Federation of Public Libraries
    - Union County Library System
- ***Public Media***
    - WQED Multimedia
- ***State/Local Government***
    - Area Transportation Authority of North Central PA
    - Lackawanna County
    - Borough of Pottstown
    - Venango County
    - Patton Township
- ***Non-Profit/Cultural***
    - Camp Susque
    - DRIVE
    - Meta Mesh Wireless
    - PSECU
- ***Commercial***
    - Apogee
    - Bopax/Subway
    - Conxx
    - Empire Access
    - Get Wireless
    - MAW Communications
    - New Leaf Initiative
    - The Pajama Factory
    - River Valley Internet
    - Sunesys/Crown Castle
    - TierPoint
    - The WARE Center
    - Zito Media