Title
Kennesaw State University Scientific Deep Dive

Authors
Zurawski, Jason
Addleman, Hans

Publication Date
2021-11-03

Peer reviewed

Kennesaw State University Scientific Deep Dive

November 3, 2021

Disclaimer

Kennesaw State University Scientific Deep Dive

Final Report

November 3, 2021

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. Carol Hawk 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-2001435[1].

[1] https://escholarship.org/uc/item/77k892k5

Participants & Contributors

Hans Addleman, Indiana University
Ramazan Aygun, Kennesaw State University
Jeremy Gulley, Kennesaw State University
Marco Guzzi, Kennesaw State University
Martina Kaledin, Kennesaw State University
Jungkyu Park, Kennesaw State University
Todd Pierson, Kennesaw State University
MinJae Woo, Kennesaw State University
Jason Zurawski, ESnet

Report Editors

Hans Addleman, Indiana University: addlema@iu.edu
Jason Zurawski, ESnet: zurawski@es.net

Contents

1 Executive Summary

Deep Dive Review Purpose and Process

EPOC uses the Deep Dive process to discuss and analyze current and planned science, research, or education activities and the anticipated data output of a particular use case, site, or project to help inform the strategic planning of a campus or regional networking environment. This includes understanding future needs related to network operations, network capacity upgrades, and other technological service investments. A Deep Dive comprehensively surveys major research stakeholders' plans and processes in order to investigate data management requirements over the next 5–10 years. Questions crafted to explore this space include the following:

- How, and where, will new data be analyzed and used?
- How will the process of doing science change over the next 5–10 years?
- How will changes to the underlying hardware and software technologies influence scientific discovery?

Deep Dives help ensure that key stakeholders have a common understanding of the issues and the actions that a campus or regional network may need to undertake to offer solutions. The EPOC team leads the effort and relies on collaboration with the hosting site or network, and other affiliated entities that participate in the process. EPOC organizes, convenes, executes, and shares the outcomes of the review with all stakeholders.

This Review

Between August 2021 and October 2021, staff members from the Engagement and Performance Operations Center (EPOC) met with researchers and staff from Kennesaw State University (KSU) for the purpose of a Deep Dive into scientific and research drivers. The goal of this activity was to help characterize the requirements for a number of campus use cases, and to enable cyberinfrastructure support staff to better understand the needs of the researchers within the KSU community.

This review includes case studies from the following campus stakeholder groups:

- Research Computing
- Department of Physics: High-power Ultrashort Laser-pulse Propagation
- Department of Physics: Theoretical High Energy Physics, Quantum Field Theory, and Quantum Chromodynamics (QCD)
- Department of Chemistry & Biochemistry: Computational Chemistry
- Department of Mechanical Engineering: Nanostructured Materials
- Department of Ecology, Evolution, and Organismal Biology: Conservation of Salamanders
- Department of Statistics and Data Science: Healthcare Data Science

Material for this event included the written documentation from each of the profiled research areas, documentation about the current state of technology support, and a write-up of the discussion that took place via e-mail and video conferencing.

The case studies highlighted the ongoing challenges and opportunities that KSU has in supporting a cross-section of established and emerging research use cases. Each case study mentioned unique challenges which were summarized into common needs.

The review produced several important findings and recommendations from the case studies and subsequent virtual conversations:

- Data volumes will continue to increase for many research cases. These will produce data locally, as well as involve external sources that must be transferred.
- Exploring ways to upgrade and augment existing KSU HPC and storage resources, along with strategies to support data mobility to external partners.
- A desire to have local staff that can assist researchers in converting traditional workflows to use HPC approaches.
- Ways that KSU can leverage commercial clouds for research activities.
- Possible integration of KSU HPC/HTC resources into larger national efforts (e.g., Open Science Grid) for certain use cases.
- Improving LAN networking to support real-time use cases, such as visualization.
- Understanding and supporting research use cases that may involve a requirement to address data sensitivity (e.g., HIPAA, CUI, etc.) and considering dedicated infrastructure and staff to manage this.
- Desire to explore specific technology approaches for research, including upgraded GPUs and quantum computing.
- Working to improve regional connectivity arrangements, perhaps via the PeachNet network, to ensure access to other facilities in GA and nationally.

2 Deep Dive Findings & Recommendations

The deep dive process helps to identify important facts and opportunities from the profiled use cases. The following outlines a set of findings and recommendations from the Kennesaw State University Deep Dive that summarize important information gathered during the discussions surrounding case studies, and possible ways that could improve the CI support posture for the campus:

- Access to HPC resources will be critical for future research goals - this can be local to KSU, or at remote locations such as a national provider (e.g., XSEDE).
- KSU research community can benefit from assistance in converting some aspects of the software workflow to use HPC-centric resources (e.g., converting serial code to be more parallel, etc.)
- The ability to leverage local storage resources will be a limiting factor for research in the future. Most profiled projects anticipate increases to data volume over time, requiring access to temporary and long-term storage.
- Upgrading the local KSU HPC will be required to support a number of use cases, unless other ways to share and use regional or national computational resources become available.
- The amount of available main memory on KSU HPC systems will be a factor in future years, and should be targeted for upgrade to support several use cases.
- The ability to perform data movement activities between KSU and collaborators could benefit from use of a higher performance tool (e.g., Globus) to exchange data between KSU and other locations (home connections, other national facilities).
- Participation in ESnet's Data Mobility Exhibition[2] is recommended to understand the baseline transfer performance.
- Understanding the roles and limitations of cloud based resources use will be important for KSU, as more use cases look to understand if their use cases can benefit from those resources.
- Security and Export controls at certain facilities may require incorporation of specialty infrastructure at KSU to support these collaborators.
- LAN networking at KSU can be improved to emphasize the use case of real-time, and remote visualization between facilities on campus. This could include removing friction from paths, increasing bandwidth, or prioritizing traffic patterns.
- Software costs for critical tools can be prohibitive, which hampers the progress of some research efforts. .
- It is recommended that KSU work with regional networking providers (e.g., the PeachNet network), to ensure access to other facilities in GA and nationally.
- As genomics sequencing output becomes more precise, data volumes will grow. If done offsite, this means more data downloads. If done on site (through potentially local instrumentation), storage resources may be required.

[2] https://fasterdata.es.net/performance-testing/2019-2020-data-mobility-workshop-and-exhibition/

- The acquisition, use, and storage of HIPAA data will require careful thought about ways to leverage KSU resources. This may mean construction of a special CUI infrastructure just to support this (and other) use cases with similar requirements, as well as dedicating staff to the review and management of sensitive data enclaves.
- Access to GPU resources could assist with some aspects of the scientific workflow for KSU researchers. Particularly those that may have access to large amounts of local memory.
- University sponsored cloud resources (e.g., OneDrive) often perform slower than some personal cloud options (e.g., Google Drive, DropBox).
- Desire to explore quantum computing, either locally or through affiliated national efforts.

3 Process Overview and Summary

3.1 Campus-Wide 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 the Deep Dives process as an essential tool as part of a holistic approach to understand end-to-end research data use. By considering the full end-to-end research 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 Indiana University (IU) GlobalNOC and our Regional Network Partners; and
- 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, the Deep Dive process offers an opportunity for broader understanding of the longer term needs of a researcher. The Deep Dive process aims 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 process is based on an almost 10-year practice used by ESnet to understand the growth requirements of Department of Energy (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

3.2 Campus-Wide Deep Dive Structure

The Deep Dive process involves 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.
- ***Non-local Resources***—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.
- ***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.

3.3 Kennesaw State University Deep Dive Background

Between August 2021 and October 2021, EPOC organized a Deep Dive in collaboration with Kennesaw State University to characterize the requirements for several key science drivers. The representatives from each use case were asked to communicate and document their requirements in a case-study format. These included:

- Research Computing
- Department of Physics: High-power Ultrashort Laser-pulse Propagation
- Department of Physics: Theoretical High Energy Physics, Quantum Field Theory, and Quantum Chromodynamics (QCD)
- Department of Chemistry & Biochemistry: Computational Chemistry
- Department of Mechanical Engineering: Nanostructured Materials
- Department of Ecology, Evolution, and Organismal Biology: Conservation of Salamanders
- Department of Statistics and Data Science: Healthcare Data Science

3.4 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 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.

Kennesaw State University (KSU) A leader in innovative teaching and learning, Kennesaw State University offers undergraduate, graduate and doctoral degrees to its nearly 43,000 students. With 11 colleges on two metro Atlanta campuses, Kennesaw State is a member of the University System of Georgia and the second-largest university in the state. The university's vibrant campus culture, diverse population, strong global ties and entrepreneurial spirit draw students from throughout the region and from 126 countries across the globe. Kennesaw State is a Carnegie-designated doctoral research institution (R2), placing it among an elite group of only 6 percent of U.S. colleges and universities with an R1 or R2 status.

4 Kennesaw State University Case Studies

Kennesaw State University presented a number use cases during this review. These are as follows:

- Research Computing
- Department of Physics: High-power Ultrashort Laser-pulse Propagation
- Department of Physics: Theoretical High Energy Physics, Quantum Field Theory, and Quantum Chromodynamics (QCD)
- Department of Chemistry & Biochemistry: Computational Chemistry
- Department of Mechanical Engineering: Nanostructured Materials
- Department of Ecology, Evolution, and Organismal Biology: Conservation of Salamanders
- Department of Statistics and Data Science: Healthcare Data Science

Each of these Case Studies provides a glance at research activities, 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 primarily limited to current needs, with only 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.

4.1 Research Computing

Content in this section authored by Ramazan Aygun, Kennesaw State University

The HPC systems at KSU consist of infrastructure serving two different populations: students, as well as the faculty and research staff. Two different login nodes handle the authentication for students and faculty. The different HPC environments use the same queue on a MOAB instance (e.g., a workload and resource orchestration platform that automates the scheduling, managing, monitoring, and reporting of HPC workloads) for processing data. However, they are primarily separated at the data storage level by providing separate storage systems between the two populations.

4.1.1 Compute and Job Queues

Three different job queues are available to the users. The primary queue with most of the resources is called "Batch." This queue is for general compute processing and is for CPU intensive research. The batch queue contains 58 servers containing 130 CPUs and 1736 cores. The second queue is called "GPUQ," which includes two servers with 2 CPUs, 28 cores, and 4 Nvidia P100 GPUs in each. The last queue is called "HIMEM," which processes jobs requiring a large amount of memory. This queue contains one server which has 4 CPUs, 48 cores, and 1536 GB of memory. Most of the systems connect to the InfiniBand network via either a 56 Gb or 100 Gb connection. In addition, they all connect to a private network using 1 Gb ethernet. The entire cluster follows a "condo" computing model, wherein individual machines may be contributed by a specific faculty member, allowing them the ability to have guaranteed access when needed, and still contribute to the greater computing power of the university when not needed.

4.1.1 Storage

There are two similar storage solutions that serve faculty and students. The storage solution serving the students contains a scratch tier backed by solid state drives (SSD) and a data tier using high performance hard disk drives (HDD). The faculty storage solution has a scratch tier utilizing NVMe storage and a data tier using HDDs. Both storage solutions connect to an InfiniBand fabric utilizing multiple connections. The student storage solution uses multiple 56 Gb and 40 Gb paths to connect to the storage network. The faculty storage solution uses multiple 100 Gb and 56 Gb connections for connecting to the storage network. Both storage solutions utilize GPFS as the filesystem hosting the scratch and data volumes.

4.1.1 Network and Security

All of the compute and management nodes are connected to a non-routable private network. This network handles all of the traffic going between the compute nodes, job scheduler, login nodes, and management nodes. Systems connect to this network via 1 Gb ethernet. The only nodes that connect to the routable network are the login nodes, the job scheduler, and NSD05/06 for backups. Users will have to connect via a VPN to access the HPC service as logins are authenticated against LDAP.

4.1.1 Future Plans

Some infrastructure upgrades are coming within the next year for the HPC service. KSU is upgrading some of the compute nodes, including the GPU servers, with newer systems. In addition, the InfiniBand network is going to be upgraded to an ethernet-based network. All servers will be connected to the new storage network using 25 Gb ethernet. The SANs will use multiple 25 Gbe connections for multipathing and redundancy.

4.2 Department of Physics: High-power Ultrashort Laser-pulse Propagation

Content in this section authored by Jeremy Gulley, Kennesaw State University

4.2.1 Science, Research, or Educational Use Case Summary

The research area focus involves laser light propagation and how this interacts with solids. Research is performed through simulation. Research heavily relies on computation to represent laser behavior, and then test this against theoretical models, results from experiments, and ways to validate the two.

Simulation is done at resources located at KSU, and with collaborators at the University of Texas where TACC resources can be leveraged. Data mostly stored at KSU, or with affiliated cloud storage. Data sizes are variable and can range from 10s of MB to as much as 10s of GB (compressed). Research staff does save important results to cloud storage, and deletes older resources to conserve resources. SCP is used for most local data movement (HPC to personal computers), and then web-interfaces to cloud storage.

Future goals to move toward more parallel forms of computation, but do not anticipate moving to a more instrument based workflow that would involve moving captured data from external partners.

4.2.2 Discussion, Findings, and Actions

The following are recommendations based on the scientific use case:

- Access to HPC resources will be critical for future research goals - local or remote.
- Storage will be a limiting factor for research in the future.
- Understanding the roles and limitations for cloud use will be important for KSU.
- Export controls at certain facilities (ARL) may require incorporation of specialty infrastructure at KSU to support collaborators.

4.2.3 Next Steps

The following are possible next steps for KSU to pursue with the research use case:

- The research use case can benefit from technical assistance in converting software workflow to use more HPC-enabled resources.
- Upgrading local KSU hardware will be required, with potential need to augment with access to other forms of computational power.
- Data movement could benefit from use of a higher performance tools (e.g., Globus) to exchange data between KSU and other locations (home connections, other national facilities).

4.3 Department of Physics: Theoretical High Energy Physics, Quantum Field Theory, and Quantum Chromodynamics (QCD)

Content in this section authored by Marco Guzzi, Kennesaw State University

4.3.1 Science, Research, or Educational Use Case Summary

The research area is quantum field theory and physics of the strong interactions, and hadron collider phenomenology. This research is performed through analytical calculations (implemented in computer codes) that predict the values of physical observables in proton-proton collisions that are measured with very high precision at the CERN Large hadron Collider (LHC). The theoretical nature of this work does not involve collection of experimental data, but does involve calculations and data-vs-theory comparisons for observables. Therefore, there are no requirements to create, collect, transfer, share or store large amounts of data. Experimental data are available online on the CERN LHC website. Local HPC access and usage are a critical part of this workflow.

4.3.2 Discussion, Findings, and Actions

The following are recommendations based on the scientific use case:

- Access to and usage of KSU HPC resources will be critical for future research goals. It is anticipated that local computation will be used almost exclusively, without requirements to leverage national or international resources (e.g., WLCG).
- Understanding the roles and limitations for cloud use will be important for KSU.

4.3.3 Next Steps

The following are possible next steps for KSU to pursue with the research use case:

- LAN networking can be improved to emphasize the use case of remote visualization between facilities on campus. This could be removing friction from paths, increasing bandwidth, or prioritizing traffic patterns.
- Upgrading local hardware will be required in the out years to keep up with CPU and memory requirements.

4.4 Department of Chemistry & Biochemistry: Computational Chemistry

Content in this section authored by Martina Kaledin, Kennesaw State University

4.4.1 Science, Research, or Educational Use Case Summary

Computational chemistry research produces data related to the molecular structure, and properties such as energies, molecular vibrations, and molecular dynamics simulations. The process of science involves the analysis of this data through computation.

4.4.2 Discussion, Findings, and Actions

The following are recommendations based on the scientific use case:

- Access to HPC resources will be critical for future research goals - local being preferred to remote.
- Managing software requirements on HPC (e.g., specific versions of tools and compilers) can sometimes be challenging when using specific packages.
- Storage will be a limiting factor for research in the future.
- Understanding the roles and limitations for cloud use will be important for KSU.
- Software costs for critical tools can be prohibitive.

4.4.3 Next Steps

The following are possible next steps for KSU to pursue with the research use case:

- Upgrading local HPC will be required, unless other ways to share/use regional/national resources become available. Memory on each system will be a factor in future years.
- LAN networking can be improved to emphasize the use case of remote visualization between facilities on campus. This could be removing friction from paths (e.g., understanding the role of VPN use), increasing bandwidth, or prioritizing traffic patterns. Low-latency is a requirement for some visualizations.
- KSU should work out regional networking arrangements, perhaps via PeachNet, to ensure access to other facilities in GA and nationally.
- Data transfer could benefit from use of a higher performance tools (e.g., Globus) to exchange data between KSU and other off-campus locations (residences, or other national facilities).

4.5 Department of Mechanical Engineering: Nanostructured Materials

Content in this section authored by Jungkyu Park, Kennesaw State University

4.5.1 Science, Research, or Educational Use Case Summary

Research into nanostructured materials experimentally and theoretically; specifically related to heat transfer and wearable technologies. Most of the process of science is done using atomic-level simulations on HPC resources.

4.5.2 Discussion, Findings, and Actions

The following are recommendations based on the scientific use case:

- Desire to explore quantum computing, either locally or through affiliated national efforts.
- Access to HPC resources will be critical for future research goals - local being sufficient for most use cases.
- Storage could be a limiting factor for research in the future.
- Understanding the roles and limitations for cloud use will be important for KSU.

4.5.3 Next Steps

The following are possible next steps for KSU to pursue with the research use case:

- LAN networking can be improved to emphasize the use case of remote visualization between facilities on campus. This could be removing friction from paths, increasing bandwidth, or prioritizing traffic patterns.
- Upgrading local HPC is desirable, unless other ways to share/use regional/national resources become available.

4.6 Department of Ecology, Evolution, and Organismal Biology: Conservation of Salamanders

Content in this section authored by Todd Pierson, Kennesaw State University

4.6.1 Science, Research, or Educational Use Case Summary

Research is organismal biology, with a focus on the ecology, evolution, and conservation of salamanders. Primary data that is handled is genomic: generation (via local collection and wet-lab), processing (handled offsite at commercial or R&E facilities), analysis (handled at KSU), and storage (at KSU and within cloud storage). Use of remote instruments (DNA sequencing) to create data (10GB to TB in size). Use of local HPC is critical for processing and analysis. Storage is leveraged locally and within the cloud. Outsourcing the sequencing task is expected to continue - the cost point of ownership is still too high versus the quality and cost of using other resources. Commercial and R&E connected facilities (e.g., GGBC at the University of Georgia) are expected to be used into the future. Other local data tools are not producing large data volumes.

4.6.2 Discussion, Findings, and Actions

The following are recommendations based on the scientific use case:

- Access to HPC resources will be critical for future research goals - local and remote.
- Storage will be a limiting factor for research in the future - local and cloud resources.
- University sponsored cloud resources (e.g., OneDrive) often perform slower than some personal cloud options (e.g., Google Drive, DropBox).
- Understanding the roles and limitations for cloud use will be important for KSU.
- As sequencing output becomes more precise, data volumes will grow. If done offsite, this means more data downloads. If done on site (through local instrumentation), storage resources may be required.

4.6.3 Next Steps

The following are possible next steps for KSU to pursue with the research use case:

- Upgrading local HPC will be required, unless other ways to share/use regional/national resources become available.
- Data movement could benefit from use of a higher performance tools (e.g., Globus) to exchange data between KSU and other locations (home connections, other national facilities).
- KSU should work out regional networking arrangements, perhaps via PeachNet, to ensure access to other facilities in GA and nationally.

4.7 Department of Statistics and Data Science: Healthcare Data Science

Content in this section authored by MinJae Woo, Kennesaw State University

4.7.1 Science, Research, or Educational Use Case Summary

Research involves clinical data concerning patients and care providers (see Table 1 for sizes). Many of the datasets are subject to HIPAA, and as a result use of dedicated infrastructure (supplied by collaborators) to support the compliance requirements. Some datasets include a large volume of medical imaging and clinical notes/records. These data sets are manipulated to develop better clinical decision support tools. This is being done using models that can work on the training data. Computation will play an important role for future iterations of this work.

Type	***Size (DICOM Format)***	***Size (PNG Format)***
Single Image	.5 MB	10MB
CT Scan (~300 Images)	150MB	3GB
Study (~1000 CT scans)	150GB	3TB

Table 1 - Research Data Sizes

4.7.2 Discussion, Findings, and Actions

The following are recommendations based on the scientific use case:

- The acquisition, use, and storage of HIPAA data will require careful thought about ways to leverage KSU resources in the future. Current use cases rely on loaned infrastructure. This may mean construction of a special CUI infrastructure at KSU just to support this (and other) use cases with similar requirements.
- Access to GPU resources could assist with various aspects of the scientific workflow. Particularly those that may have access to large amounts of local memory.
- Access to HPC resources will be critical for future research goals - local or remote.
- Storage will be a limiting factor for research in the future as data sets grow.
- Understanding the roles and limitations for cloud use will be important for KSU.

4.7.3 Next Steps

The following are possible next steps for KSU to pursue with the research use case:

- The research use case can benefit from technical assistance in converting software workflow to use more HPC-enabled resources.
- Upgrading local KSU hardware will be required, with potential need to augment with access to other forms of computational power.
- Data movement could benefit from use of a higher performance tools (e.g., Globus) to exchange data between KSU and other locations (home connections, other national facilities).

Appendix A - Regional Networking Diagram

PeachNet
University System of Georgia
Information Technology Services