South Dakota Region Scientific Deep Dive

November 2, 2021

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.

South Dakota Region Scientific Deep Dive

Final Report

November 2, 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-2001427[1].

[1] https://escholarship.org/uc/item/3s07q0d8

Participants & Contributors

Hans Addleman, Indiana University
Cynthia Anderson, Black Hills State University
Robert Anderson, South Dakota School of Mines and Technology
Kenneth Benjamin, South Dakota School of Mines and Technology
Kevin Brandt, South Dakota State University
Debbi Bumpous, Northern State University
Eduardo Callegari, University of South Dakota
William Capehart, South Dakota School of Mines and Technology
Jodi Casanova, Northern State University
Susan Citrak, Northern State University
Bill Conn, University of South Dakota
Omar El-Gayar, Dakota State University
Jason Erickson, South Dakota School of Mines and Technology
Scott Francis, South Dakota Board of Regents
Barbara Goodman, University of South Dakota
Kyle Gruhn, University of South Dakota
Neal Hodges, South Dakota School of Mines and Technology
Eric Holm, Dakota State University
Jeremy Iverson, Northern State University
Ranjeet John, University of South Dakota
Ryan Johnson, University of South Dakota
Chad Julius, South Dakota State University
Paul Kern, South Dakota Board of Regents
Alyssa Kiesow, Northern State University
John Long, Northern State University
Jeff Mahlum, South Dakota State University
Steve Millage, Dakota State University
Ken Miller, ESnet
Alycia Oye, University of South Dakota
Maria Daniela Paez, University of South Dakota
Scott Paulsen, Dakota State University
Jon Schaff, Northern State University
Doug Southworth, Indiana University
Garrett Stevens, Black Hills State University
Reinaldo Tonkoski, South Dakota State University
Brent Van Aartsen, Dakota State University
Courtney Waid-Lindberg, Northern State University
Shengjie Xu, Dakota State University
Jason Zurawski, ESnet

Report Editors

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 use cases and 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 March 2021 and September 2021, staff members from the Engagement and Performance Operations Center (EPOC) met with researchers and staff from Black Hills State University (BHSU), Dakota State University (DSU), Northern State University (NSU), The South Dakota Board of Regents (SDBOR), The South Dakota School of Mines & Technology (South Dakota Mines), South Dakota State University (SDSU) and The University of South Dakota (USD) convened for the purpose of a regional Deep Dive into scientific and research drivers. The goal of this meeting was to help characterize the requirements for a number of campuses and regional use cases, and to enable cyberinfrastructure support staff to better understand the needs of the researchers within the South Dakota community.

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

- BHSU: WestCore; The Western South Dakota Nucleic Acid Sequencing and Genotyping Core Facility
- DSU: Campus Technology Profile
- DSU: South Dakota Center for SMART Power Systems
- NSU: Connecting the Social Sciences across the Great Plains
- NSU: Toward A Greater Dissemination of Social Science Data and Information for Better Civic Engagement
- SD Mines: Realtime Weather Regional-Scale Weather Forecasting
- SDSU: Research Cyberinfrastructure Center
- USD: South Dakota Biomedical Research Infrastructure Network (SD BRIN) Proteomics Core Facility
- USD: Department of Biology & Department of Sustainability

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 the member schools, and the regional network, have 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:

- The South Dakota region is encouraged to pursue a regional CC* grant during the next round of the solicitation (October, 2021). This proposal should describe ways to unify CI capabilities at the various institutions, utilizing the REED regional network as a backplane, to create a state-wide circulatory system for research and education use cases.

- Establish state-wide working groups to address ongoing regional needs:
  - Networking upgrades and strategies to manage and measure resources
  - Ways to link computational resources on a regional basis
  - Ways to leverage storage to facilitate regional use cases
  - Mechanisms to support data mobility
  - Ways to share and increase CI expertise, particularly related to software integration for research workflows
  - Unifying security policies and implementations

- In parallel to the Regional CC* proposal, it is recommended that a number "pilot" efforts be created to experiment with the linkage of regional resources between collaborating partners. Examples include:
    - SDSMT Weather Analysis
    - Sequencing / Genotyping Superfacility
    - Historical Archiving
    - Interactive Computing & Analysis:

- SDBOR should perform an initial, and then regular, audit of the commercial cloud peering that are invaluable to participants to ensure that accessibility remains possible and optimal. A number of research use cases in the region will leverage commercial clouds in the coming years, and these will be critical conduits for scientific innovation.

- Addressing the gap in CI expertise by cultivating knowledgeable staff that can be called upon to handle networking, server support, computational knowledge, software development and implementation, and other core functions.

- Performing routine reviews of campus and regional data architectures with national centers of excellence (e.g., EPOC)

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 South Dakota Regional 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 region:

- The South Dakota region is encouraged to pursue a regional CC* grant during the next round of the solicitation (October, 2021). This proposal should:
    - Propose a goal of linking together CI capabilities at the various institutions, utilizing the REED regional network as a backplane, to create a state-wide circulatory system for research and education use cases. This proposal should include linkages of computation, storage, instrumentation, and CI expertise. This scaling will allow a greater distribution of resources and advance the broader impacts of the capabilities for smaller institutions.
    - Participants should include those profiled in this Deep Dive activity, as well as other institutions that can benefit or contribute to the overall success.
    - Given the October deadline, it is recommended that a working group of participants be established, and the process to begin the writing of the proposal starts immediately.
    - The proposal should seek funding for:
        - Networking upgrades for facilities that require it
        - Computational resources that may be deployed on a regional basis that fill gaps that have been identified
        - Storage to facilitate regional use cases (e.g., bioinformatics, etc)
        - CI expertise, particularly related to software integration for research workflows
    - The proposal should attempt to leverage existing investments in:
        - Computing and storage resources that are available, and could be shared through the implementation of federated identity services
        - Facilities that offer shared instrumentation (e.g., SD BRIN) and can accept a greater population of users
        - Data Mobility approaches for LAN, MAN, and WAN sharing
        - Existing CI expertise that can assist in automating and accelerating scientific workflows
        - Use of the REED infrastructure to implement a regional Science DMZ that links critical resources, while affording a layer of performance and protection

- In parallel to the Regional CC* proposal, it is recommended that a number of "pilot" efforts be created to experiment with the linkage of regional resources between collaborating partners. These efforts can be written up as science drivers for the proposal and will serve as blueprints for other use cases that could be implemented in the future. Examples include:

  - ***SDSMT Weather Analysis***: The SDSMT weather use case is run fully on site. A way to leverage regional resources would be to implement mechanisms that would a) migrate the workflow to a containerized approach that would facilitate operation at other partner locations b) implement a scheduling approach that would reserve computing time to facilitate running c) create a web presence/portal that has ample storage to store and share research data
  - ***Sequencing / Genotyping Superfacility***: Building on the success of the regional sequencing and genotyping centers, move to a model that integrates more computational and storage capabilities provided by partners. This would facilitate more analysis capabilities, but also leverage investments in regional networking and data movement capabilities.
  - ***Historical Archiving***: Work with NSU to create an instrument workflow for digital archiving. This has two parts: a) An 'on site' portion that links NSU instruments to local and regional computation and storage resources through automated software approaches and b) A 'traveling' portion that would work when staff must operate instruments from remote locations and have to rely on portable computation and storage. Both can be integrated into a portal for storage and sharing capabilities.
  - ***Interactive Computing & Analysis***: A number of regional use cases rely on interactive computing use cases: not only is advanced and high performance computing needed, but the latency and responsiveness must remain high due to the distributed nature. A protected logical path between contributed resources and users could be created to allow for a higher quality of service across network segment , and inked directly to modify local resources for visualization and interaction.

- SDBOR should perform an initial, and then a regular, audit of the commercial cloud peering that are invaluable to participants to ensure that accessibility remains possible and optimal (e.g., through peering relationships like I2PX). A number of research use cases in the region will leverage commercial clouds in the coming years, and these will be critical conduits for scientific innovation. At this time, it is not expected that direct connections to any clouds are required, but SDBOR should regularly evaluate this use case as traffic needs increase.

- A critical gap for all participants in the Deep Dive was a lack of CI expertise: knowledgeable staff that can be called upon to handle networking, server support, computational knowledge, software development and implementation, and other core functions. This need will grow over the coming years, and attrition and retirement will exacerbate the situation. It is

recommended that resources, when they can be found, created, or curated, be shared widely within the region. Critical needs in the coming years will be:
    - Software integration (e.g., containerization, workflow analysis and improvement)
    - Network and Information Security
    - Performance Testing (e.g., networking, but also computation and storage)
    - Data Mobility

- It is recommended that a regional working group be formed to discuss and draft sensible policies for data access, retention, and sharing of research results. As the regional capabilities for these use cases proliferate, a baseline approach to these policies that can be adopted by the partners is recommended to streamline the ability to create, house, and share data becomes more common. This working group can address:
    - The current state of individual campus policies
    - A baseline set of policies that can be adopted on a regional level for cooperating use cases
    - Implementation strategies and approaches
    - Ways to onboard new users and use cases, while still evaluating and monitoring current users and use cases
    - Technical details regarding a regional science DMZ to address use cases of instrument use, computation and storage, data mobility, and external users and use cases.

- EPOC will collaborate with sites that have requested a review of network architectures. These could include SDSU and DSU to start - but others are welcome to request this. This review can focus on:
    - Data Mobility (hardware and software)
    - Performance testing (active via perfSONAR, or passive via sFlow/Netflow)
    - Security posture for data and devices
    - Performance engineering for critical paths
    - Integration with regional projects and use cases

- EPOC will collaborate with all participants on participation in the Data Mobility Exhibition

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 a 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[2]. The EPOC team adapted this approach to work with individual science groups through a set of structured data-centric conversations and questionnaires.

[2] 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 South Dakota Regional Deep Dive Background

Between March 2021 and September 2021, EPOC organized a Regional Deep Dive in collaboration with Black Hills State University (BHSU), Dakota State University (DSU), Northern State University (NSU), The South Dakota Board of Regents (SDBOR), The South Dakota School of Mines & Technology (South Dakota Mines), South Dakota State University (SDSU) and The University of South Dakota (USD) to characterize the requirements for several key science drivers. The representatives from each campus were asked to communicate and document their requirements in a case-study format. These included:

- BHSU: WestCore; The Western South Dakota Nucleic Acid Sequencing and Genotyping Core Facility
- DSU: Campus Technology Profile
- DSU: South Dakota Center for SMART Power Systems
- NSU: Connecting the Social Sciences across the Great Plains
- NSU: Toward A Greater Dissemination of Social Science Data and Information for Better Civic Engagement
- SD Mines: Realtime Weather Regional-Scale Weather Forecasting
- SDSU: Research Cyberinfrastructure Center
- USD: South Dakota Biomedical Research Infrastructure Network (SD BRIN) Proteomics Core Facility
- USD: Department of Biology & Department of Sustainability

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.

Black Hills State University (BHSU) Black Hills State University is a regional, comprehensive, public institution that provides access to higher education for aspiring students. BHSU offers a generous number of baccalaureate and select master's degrees, generates new knowledge, promotes excellence in teaching and public engagement, and serves as a regional economic leader. Graduates make significant contributions to the workforce and the betterment of their community. Black Hills State University will innovate to provide cutting-edge education, promote student success, be a sustainable campus, and serve as an economic engine for western South Dakota.

Dakota State University (DSU) In 1881, Dakota State University was founded as a school for teacher education, in the Dakota Territory. Though we have since shifted our primary focus of education to the cyber world, Dakota State remains one of the few universities in the nation, who provides students with the technology needed to excel in any career.

Northern State University (NSU) Since 1901, Northern has been committed to academic and extracurricular excellence. Today, we offer nationally accredited academic programs in arts and sciences, business, fine arts and teacher education.

The South Dakota Board of Regents (SDBOR) The South Dakota Board of Regents is a governing board that controls six public universities in the U.S. state of South Dakota. These include Black Hills State University, Dakota State University, Northern State University, South Dakota School of Mines and Technology, South Dakota State University, and the University of South Dakota. The Board also governs the South Dakota School for the Blind and Visually Impaired and the South Dakota School for the Deaf. The system's primary goal is to provide high quality, diverse educational opportunities and services to the people of South Dakota through the effective use of resources entrusted to it. History has demonstrated that this goal can be met more effectively through an integrated system approach. In this manner, the various campuses complement one another and remain fully responsive to the central authority of the Board of Regents through the presidents and superintendents.

The South Dakota School of Mines & Technology (South Dakota Mines) is committed to excellence in science and engineering academics and research, and to developing the next generation of leaders and problem-solvers. The university offers a wide array of bachelors, masters and doctoral degrees. Founded in 1885 to provide instruction in the region's primary industry, mining, today South Dakota Mines has evolved into one of the leading science and engineering universities in the region.

South Dakota State University (SDSU). Founded in 1881, SDSU is a premier land-grant university that offers a rich academic experience in an environment of inclusion and access through inspired, student-centered education, creative activities and research, innovation, and engagement that improve the quality of life in South Dakota, the region, the nation, and the world. SDSU is recognized as one of fifty-four institutions in the United States as an Innovation and Economic Prosperity University by the Association of Public and Land-grant Universities in 2014.

The university's total research funding and expenditures in FY20 are greater than $61 million, the largest in the state. SDSU is the state's first High Research Activity institution as classified by the Carnegie Foundation for the Advancement of Teaching. Its employees embrace the university's core values of being people-centered, expanding knowledge through creativity, embracing organizational and personal integrity, commitment to diversity, and excellence through continuous improvement.

The University of South Dakota (USD) The University of South Dakota offers undergraduate, graduate and professional programs within the South Dakota System of Higher Education. As the oldest university in the state, the University of South Dakota serves as the flagship and the only public liberal arts university in the state. The University of South Dakota is regionally acclaimed and nationally recognized as a high-quality public liberal arts university with South Dakota's only schools of law, medicine and business. It is recognized for the quality of its faculty, and their excellent teaching, effective service and innovative research. Its faculty are dedicated, experts in their fields, and accessible to their students. USD educates leaders of communities, states and nations.

4 South Dakota Regional Case Studies

Black Hills State University (BHSU), Dakota State University (DSU), Northern State University (NSU), The South Dakota Board of Regents (SDBOR), The South Dakota School of Mines & Technology (South Dakota Mines), South Dakota State University (SDSU) and The University of South Dakota (USD) presented a number use cases during this review. These are as follows:

- BHSU: WestCore; The Western South Dakota Nucleic Acid Sequencing and Genotyping Core Facility
- DSU: Campus Technology Profile
- DSU: South Dakota Center for SMART Power Systems
- NSU: Connecting the Social Sciences across the Great Plains
- NSU: Toward A Greater Dissemination of Social Science Data and Information for Better Civic Engagement
- SD Mines: Realtime Weather Regional-Scale Weather Forecasting
- SDSU: Research Cyberinfrastructure Center
- USD: South Dakota Biomedical Research Infrastructure Network (SD BRIN) Proteomics Core Facility
- USD: Department of Biology & Department of Sustainability

Each of these Case Studies provides a glance at research activities for the members of the region, 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.

The South Dakota Region is committed to supporting these use cases through technology advancements, and is actively pursuing grant solicitations. The landscape of support will change rapidly in the coming years, and these use cases will take full advantage of campus improvements as they become available.

4.1 BHSU: WestCore; The Western South Dakota Nucleic Acid Sequencing and Genotyping Core Facility

Content in this section authored by Cynthia Anderson and Garrett Stevens, Black Hills State University

4.1.1 Science Background

WestCore was established in 2004, and is supported in part by funding from the National Institutes of Health's IDeA Program for Biomedical Research Excellence[3] to the South Dakota Biomedical Research Infrastructure Network (SD BRIN). WestCore's purpose is to provide critical research infrastructure and research support in the area of nucleic acid sequencing and genotyping to researchers throughout SD BRIN. Over the past 15 years, the services provided by WestCore have extended beyond SD BRIN researchers to faculty, students, and research personnel at other higher education institutions in the state, as well as institutes and agencies with research interests in SD and other western states.

WestCore provides expertise in several areas of genetic and genomic analysis, including, but not limited to, genetic, genomic and transcriptomic library development, DNA sequencing and genotyping, quantitative real-time PCR, and NextGen nucleic acid sequencing. These services support a wide array of scientific research questions in various ecosystems and organisms. Research interests supported by WestCore include such things as:

- population genetics and evolutionary studies using amplification and analysis of simple sequence repeats (SSRs),
- Sanger sequencing of individual genes or genetic regions of phenotypic or physiological interest,
- studies of microbiomes, microbial genomes and transcriptomes using NextGen sequencing applications,
- studies of the expression of smaller numbers of genes or the detection of single nucleotide polymorphisms (SNPs) using qPCR techniques.

In the past two years, WestCore leadership has emphasized the expansion of the NextGen sequencing services we are able to provide. Funding through SD BRIN has allowed us to work with SD BRIN faculty and students' projects that have enabled us to establish best practices for the procedures involved in the following types of NextGen services:

- Bacterial community profiling using 16S;
- T-cell receptor repertoire (TCR) sequencing;
- Transcriptome sequencing in bacteria and fungi;
- Eukaryotic community profiling;
- Transcriptome sequencing in eukaryotic microbes with small genomes; and
- Small RNA sequencing.

[3] NIH INBRE grant number 5P20GM103443

4.1.2 Collaborators

WestCore provides services and collaborates on research with a number of faculty and staff as described above. Yearly, an average of 50 individuals from approximately fifteen (15) universities/research centers/other institutions utilize WestCore facilities and services. A critical collaboration is with the Bioinformatics Core at USD. This collaboration provides access to personnel skilled in the analysis of NextGen sequencing data as well as several other data mining and data analysis methods that support researchers generating data via WestCore. More importantly, the Bioinformatics Core at USD provides computational resources on which software critical to NextGen sequencing analysis is housed.

4.1.3 Instruments and Facilities

The WestCore laboratory is a 2500 sq ft research facility located in the Kathryn Johnson Life Sciences Laboratory building, and a second 800 square foot laboratory located in Jonas Science provides supplemental space for WestCore including an RNA clean room, and space devoted to cell culture activities. WestCore has a quality management plan (QMP) in place to ensure the highest levels of quality assurance are followed. In addition to being fully equipped and staffed to perform Sanger sequencing and genotyping projects, WestCore is also able to perform whole genome and transcriptome sequencing, TCR repertoire sequencing, 16S microbiome profiling, 18S/ITS fungal community profiling.

BHSU DNA Sequencing & Genotyping Core Major equipment:

- LabChip GX (Caliper Life Sciences) for microfluidic quality analysis of nucleic acids;
- Illumina MiSeq next generation nucleic acid sequencer;
- 3130 Genetic Analyzer (16 capillary) (Applied Biosystems) for Sanger sequencing and genotyping applications;
- 3500 Genetic Analyzer (8 capillary) (Applied Biosystems) for Sanger sequencing and genotyping applications;
- two ABI 7500 Real-Time PCR Systems (Applied Biosystems)
- UVP ChemStudio Plus Touch Imaging System Molecular
- Devices SpectraMax M2 plate reader and fluorescence spectrometer

USD Computational Resources housing CLC Bio Genomics Workbench Networked Licensing Software and Associated Data Analysis Modules are housed on USD's "Lawrence" HPC cluster. Please see Appendix B for more information.

Present -2 years

Equipment maintenance is covered primarily by BHSU departmental budgets, and WestCore user fees.

Next 2-5 years

The equipment used regularly to provide core services will begin to require upgrades in the next 4-5 years. While functionality of the current equipment is not expected to decline, improvements in the technologies to generate sequence, fragment analysis and gene expression data/sequence detection are expected to facilitate faster and perhaps even more expansive data acquisition.

Beyond 5 years

WestCore recognized the need to be proactive in planning for the upgrade of equipment and the improvement of the nucleotide sequencing capabilities of the core for both research and teaching purposes. We regularly seek funding to upgrade equipment and software through grants, especially infrastructure building programs such as SD BRIN (NIH) and SD EPSCoR (NSF) as well as through user fees generated from core services.

4.1.4 Process of Science

WestCore's processes vary a bit based on the research methods being used for a project, but follows a general life cycle as outlined below:

- Depending on the methods employed for each project, data generated is handled differently.
    - Analyzed on the equipment then downloaded for transfer to the researcher
    - Downloaded by the researcher for analysis using appropriate software (either using WestCore provided software resources, or the researcher's own resources)
- Data Stored by WestCore for 1 year – up to the researcher to make arrangements for long-term archival. There is long-term archival available to SD BRIN participating researchers through USD's Bioinformatics IT resources.

Each major method employed generates data specific to the methodology. In all cases the earliest steps of the process have QC data such as nucleic acid quantification (Nanodrop spectrometry or Qbit fluorometry), qualitative assessment via electrophoresis with transilluminator images captured (UVP ChemStudio Plus).

Different types of data are generated by each methodology employed:

- Sanger sequencing (3130 or 3500 Genetic Analyzer)
    - QC checked by research associate or trained user prior to downloading
    - analysis in downstream pipelines
        - Sequencher
        - CLC Bio Genomics Workbench
        - Other software (licensed or open source)
- Genotyping (3130 or 3500 Genetic Analyzer)
    - QC checked by research associate or trained user prior to downloading
    - analysis in downstream pipelines
        - GeneMapper

        - Other software (licensed or open source)
- NextGen Sequencing (Illumina MiSeq)
    - Critical data prior to NextGen sequencing run includes
        - Quantification of library (DNA or cDNA) (Qbit or LabChip GX)
        - Qualitative assessment – may be done at multiple steps, especially in RNA-based experiments (LabChip GX, and/or qPCR (ABI 7500 Real-Time PCR System, or ABI Quant Studio 3)
        - Library pre-run for optimization of multiplexed runs (MiSeq)
    - Post run QC checked by research associate then downloaded
        - to local drive for transfer to researcher
        - to local temporary (1 year) backup drive
        - analysis in downstream pipelines
            - CLC Bio Genomics Workbench
            - Other software (licensed or open source)
- Quantitative PCR (qPCR) for gene expression, genetic variation, mutation detection, and other applications (ABI 7500 Real-Time PCR System, or ABI Quant Studio 3)
    - ABI 7500 Real-Time PCR System Software or Quant Studio Design & Analysis software
    - Statistical software (R, Excel, or other licensed or open source)

The size of the data produced by the equipment and the downstream analysis ranges depending on the application and data analysis. Raw Sanger sequence and fragment analysis files generate 100 kb, and the analyzed data at about 500 kb to 5 Gb depending on the size of the project. The largest files are generated by the NextGen sequencing applications. Raw data from this application range from the 3 – 26 Gb depending on the chemistry used and the read-lengths generated. Analysis of the NextGen data using CLC Bio can be 1-5 terabytes depending on the analysis applications employed. On average we run about 3-5 projects per month. Some faculty choose to analyze their data using open source software, in which case the ownership of the raw data files are passed to the researchers using cloud based file sharing programs such as Illumina's Base Space, Google Drive, Microsoft One Drive, or Dropbox. Researchers who would like to use WestCore's licenses for CLC Bio are currently able to download the Genomics Workbench software and the associated modules, and borrow a CLC licenses through the License Server resources installed on the Lawrence Computer at USD's Bioinformatics Core.

Present -2 years

WestCore services continue to grow, especially with respect to the number of NextGen sequencing projects. This will increase the need for bioinformatics computational resources to be available to faculty state-wide. There will be a need to require researchers to run their analyses on the grid nodes in order to free up licenses for use by other researchers.

Next 2-5 years & Beyond

It is anticipated that the technology for generating whole genome, transcriptome, microbiome and other such data will continue to improve. We anticipate the need to upgrade equipment to keep up with the capabilities. Furthermore, the need for bioinformatics computational support will increase. Our licenses currently expire in June 2025 and funds will need to be procured to renew the licenses for these resources, or we will need to identify reliable open source software to meet researcher needs.

4.1.5 Non-local Resources

USD's Bioinformatics Core resources provide supercomputing capabilities to house networked software for NextGen Sequencing data analysis applications for data generated by WestCore. Please see Appendix B for more information.

4.1.6 Software Infrastructure

The following is list of commercial licensed software used:

- CLC Genomics Server Software, located on Lawrence Computer at USD
  - Remotely manages remote access to the software
  - Allows admin to set-up user groups and data access permissions
- CLC Genomics Premium Server Extension
- CLC Genomics Microbial Genomics Server Extension
- CLC Bio Genomics Workbench, network server license and 2 networked licenses accessible to faculty statewide with permissions to borrow license
  - Allows jobs to be run on the job nodes on the Lawrence computer instead of on researchers local computer
- CLC Genomics Premium Modules
- CLC Microbial Genomics Module
- CLC Genome Finishing Module Server
- CLC Genome Finishing module
- CLC Genomics Grid Worker (allows jobs to run on grid nodes instead of on user's local computer)
- LMX License Server Management Software for CLC Bio Genomics Workbench and related software listed above – allows us to monitor and manage borrowed licenses
- Sequencher
- ABI Sequence Analysis
- ABI GeneMapper
- ABI 7500 Real-Time PCR System Software
- Quant Studio Design & Analysis software

4.1.7 Network and Data Architecture

Please see Appendix A for information on the regional connectivity for BHSU and WestCore

4.1.8 Cloud Services

Some data shared via Illumina's Base Space, Google Drive, Dropbox, or Microsoft One Drive

4.1.9 Known Resource Constraints

The availability of the Lawrence Computer at USD offers the much needed ability to make the most use of a limited number of licenses for this very expensive specialized software. The largest constraint to growth and research support in this area is the lack of WestCore personnel time to troubleshoot problems using the networked software, and provide effective training to faculty wishing to use the resources:

- CLC Bio Genomics Workbench and associated modules require regular updates to ensure functionality of the system. Due to budget limitations, WestCore does not have personnel dedicated to data analysis support. As a result, there is a constraint on the time available for the WestCore personnel to manage software updates and troubleshoot shoot problems with researchers and USD IT staff. This will be especially important over the next few years as the applications of NextGen sequencing technology are being employed by a much larger number of faculty in South Dakota.
- This process of sharing data with researchers and allowing them to access CLC Bio Genomics Workbench and associated module licenses is working acceptably at this time but as the user base increases, and during times of heavy use, it would be really helpful to train faculty to run jobs on the Grid nodes (supported by CLC Bio Grid worker software also currently on the Lawrence Computer). It has also been noticed recently that while regular Workbench jobs can be sent to the grid nodes for analysis, the modules are not able to be run on the grid nodes. WestCore personnel need to identify the modules with problems and work with USD IT to troubleshoot.

4.1.10 Outstanding Issues

None to report at this time.

4.2 DSU: Campus Technology Profile

Content in this section authored by Brent Van Aartsen, Scott Paulsen, Steve Millage and Eric Holm from Dakota State University

4.2.1 Science Background

In 1881, Dakota State University was founded as a school for teacher education, in the Dakota Territory (Madison, SD). In 1984, the governor of SD shifted our primary focus of education to the cyber world. Dakota State is still one of the few universities in the nation that provides students with the technology needed to excel in any career.

Enrollment

- Total Headcount - 3,186
- Total Graduate Headcount - 446
- Internet-only students - 1759
- Dual Credit (high school students only) - 239

Degree Offerings

- Ph.D. Degrees - 3
- Masters Degrees - 8
- Bachelors Degrees - 45

Faculty and Staff

- Faculty/adjuncts, graduate assistants - 157
- Full-time & Part-time Staff - 234
- Student Employees - 168

Budget

- Operating Budget - $54.2 million for FY21

Research

- FY20 Expenditures from Sponsored Grants and Contracts - $8.7 million
- Madison Cyber Labs facility opened in October 2019

4.2.2 Collaborators

The Information Technology Services (ITS) department at DSU has a good relationship with other campus administrators. Both the CTO and CIO, co-heads of ITS, serve as members of the President's Cabinet. We work closely with the college deans as well as the Office of Research and Economic Development.

The relationship between ITS and researchers is a mixture of positives and negatives. While ITS does its best not to impede the work of the researchers, we also have to weigh the security implications of the data being worked with. A challenge we have is not having a data classification policy. This policy is in the works from our system level but the process has been in the works for several years. Having the data classification policy will allow ITS to build the appropriate protections for the research environments based on what is defined within the policy. For the researchers, the policy will allow them to better classify their data to ensure they are storing it in the proper environment with the proper protections.

Another challenge we face, much like all other IT operations, is funding both for equipment and personnel. Personnel resources is one of our bigger challenges right now. Madison's close proximity to Sioux Falls, the largest metropolitan in South Dakota, means we have to compete with that job market for IT talent. From the equipment side, we have not pursued any CC* grants at this time.

4.2.3 Instruments and Facilities

4.2.3.1 The MadLabs Research Environment & Network (MADREN)

Dakota State University's (DSU) MadLabs Research Environment & Network (MADREN) is an extensive technology infrastructure dedicated to cyber security research. The MADREN includes a server cluster comprised of 10 Lenovo SR630s servers, each with dual Intel Xeon Gold 5118 Processors, for a total of 240 cores @ 2.3 GHz. The cluster is supported by 2.56TB of TruDDR4 @ 2666MHz RAM available and a 126TB HPE Nimble Storage Adaptive Flash Array. These resources are accessible through virtualization via VMware Horizon and VMware Cloud Director.

The MADREN also contains a large cluster accessible through VMware Horizon and VMware Cloud Director. It includes 5 Lenovo SR670s servers, each with dual Intel Xeon Gold 6242 Processors, for a total of 160 Cores @ 2.8 GHz each, and 1.92TB of TruDDR4 Performance+ RAM @ 2933MHz. The cluster has 40 NVIDIA Tesla T4 16GB cards, with 12,800 Turing Tensor Cores and 102,400 CUDA Cores. The total GPU capacity represents 324 teraFLOPS (Single-Precision), 2.6 petaFLOPS (Mixed-Precision), 5,200 TOPS (INT8), or 10,400 TOPS (INT4).

All MADREN resources have access to both Internet1 (I1 or Commodity Internet) and Internet2 (I2) via the South Dakota REED Network. I1 bandwidth is a shared resource with the DSU Production Network with a max throughput of 3 Gbps. I2 bandwidth has a max data transfer of 100 Gbps (30Gbps aggregate firewall traffic with full security protections).

4.2.3.2 The IA Lab

The Problem

Technology education is inherently hands-on by nature. It's a major principle of constructivist learning. Much like a biology or chemistry lab, a great deal of setup goes into creating a hands on lab for technology labs. Multiple computers are required, plus networking gear to connect them together, plus any additional accessories such as a firewall, router, cellular telephones, etc. Once all of these are properly configured, the hardware setup must be duplicated several times for utility by multiple students. This process is effective, but is very time consuming and outright prevents online students from participating in hands-on labs.

The Need
Several needs exist for an effective lab implementation within the technology realm.

Extendibility
The need within the technology program is a system that can provide the same user experience to students, online or on-campus. The default tends to be that online students are second-class students, unable to participate in physical labs.

Efficiency
The creation of labs can be considerably time consuming for a single on-class, upwards of 8-10 hours per lab. Coupling this time requirement with having to replicate the lab many times over for both on-campus and online populations, it simply isn't possible for a single faculty member to manage their own labs with courses of 40-90 students.

Versatility
Any lab environment for use in the technology area needs to support all areas of technology. Being restricted to a single platform (such as Microsoft Windows) creates restrictions that are impossible to overcome. The lab solution needs to support any/all technology platforms.

Safe
Teaching cybersecurity fundamentals can have grave consequences with beginners. A simple typo can make the difference between a basic lab exercise and launching a real-world cyber-attack against another organization. Any lab environment used must protect the learners from themselves.

The Solution
DSU's information Assurance Lab is our custom designed solution to the problems of technology education. Our lab was designed and implemented in 2009 and its use has continually grown ever since with the additions of new classes plus growing enrollment.

The IA Lab allows for an instructor to focus their time on creating and testing their lab. Once their lab is created, it can be cloned for testing in a matter of minutes. Once the lab is finalized, the lab administrator can copy unique instances of the lab to all students within the class. This process takes approximately 20 minutes total, depending on the size of the class.

The lab has the ability to run any platform (Windows, MacOS, FreeBSD, or Linux), in addition to popular firewall and router platforms as well as GSM cellular base stations. These labs are all safely contained so that students are safe when practicing any cybersecurity concepts.

Users

Due to the self-service nature of our lab implementation, it can be used for projects far beyond the classroom. The IA Lab hosts research projects for undergraduate and graduate students, in addition to housing research projects for faculty members. Due to the safe/secure nature of the lab, it also houses DSU's High Performance Computing/Hadoop environment.

The labs users vary from semester to semester, but largely include students from the following programs:

- Information Systems
- Cyber Operations
- Computer Science
- Network Security Administration
- Computer Game Design
- Digital Arts and Design
- Mathematics for Information Systems
- Computer Education
- DSc Cyber Security
- DSC Information Systems
- MS Analytics
- MS Applied Computer Science
- MS Information Assurance
- MS Information Systems
- MSE Educational Technology

Technology

In order to facilitate the large lab environment, enterprise grade hardware is required. This is the type of hardware that would be found in any large-scale corporate IT environment and includes:

- Virtualization software that's both custom-created for our unique needs coupled with software from VMWare
- Wireless/Cellular/Mobile modules to create GSM base stations
- Enterprise-Grade servers
    - Large memory capacity per server, in excess of 128 GB
    - High network throughput, in excess of 4X gigabit interfaces
    - Storage Area Network connectivity, dual 8GB fibre channel
- Large-scale networking equipment (from Juniper Networks)
- Fibre Channel Networking equipment providing dual connections to every server (From Cisco)
- Large storage capacity for storing student/staff labs and research from HP/3PAR

Accomplishments

Our lab design is very unique to our own needs. We use a great deal of software and hardware and run it to its limits to accomplish what we do, but we have achieved many accomplishments:

- Several entities have modeled their own labs based off of our own design

    - National Security Agency
    - Naval Post Graduate School
    - Northeastern University
    - University of Cincinnati
- We've worked closely with VMware to improve their own software, discovered and reported bugs, and helped shape the direction of some of their products
- Our online programs have a significant competitive advantage over the online programs of other universities across the nation
- Our Center of Academic Excellence in Cyber Operations hinged significantly on our ability to focus on hands-on education of our graduates

Present -2 years

- Complete Perf Sonar deployment for MADREN
- Purchase and Deploy DTN for MADREN
- Continue to refine and enhance current systems to meet researchers needs in both the MADREN and IALab.

Next 2-5 years

- Grow AI/ML capabilities and capacities for both teaching and research
- Replacement/Enhancement of current compute cluster and GPU cluster equipment
- Renewal of support contracts on 100G firewall and border routers
- Expand IALab to accommodate delivery of statewide K12 Cyber Academy program

Beyond 5 years

- Replacement of 100G infrastructure
- Evaluation of 400G infrastructure

4.2.4 Process of Science

x

Present -2 years

x

Next 2-5 years

x

Beyond 5 years

x

4.2.5 Non-local Resources

x

Present -2 years
x

Next 2-5 years
x

Beyond 5 years
x

4.2.6 Software Infrastructure

- Locally or remotely management of data resources (e.g. portals):
  - None
- Facilitates the transfer of data sets from or to remote collaborators (e.g. Aspera, Globus, FTP, SCP, etc.):
  - None
- Specialty software managed for research groups:
  - Atlas.ti by CleverBridge
  - SAS
  - Stata
  - Statistics & Machine Learning Tools (ST) by The Mathworks, Inc.
  - Spartan by WaveFunction, Inc.
  - SigmaPlot by Systat Software Inc
  - VMWare Cloud Director
  - VMWare Horizon

Present -2 years

- Most likely looking at adding Globus as a service offering for our researchers
- Deploy InCommon SSO

Next 2-5 years
Design, build, deploy compute and/or gpu Kubernetes based cluster for researchers to make use of

Beyond 5 years
Unknown at this time.

4.2.7 Network and Data Architecture

WAN
DSU is connected to the South Dakota Research, Education and Economic Development (REED) The REED Network is a 100Gbps network that connects South Dakota's six public universities to one another and also to the EROS Data Center and the Sanford Lab at the Homestake Mine. REED also connects to Internet2 and ESnet by partnering with two neighboring RENs, Great Plains Network (GPN) and Northern Tier Network Consortium (NTNC).

DSU Bandwidth:

- I1 - 3 Gbps
- I2 - 100 Gbps

LAN

The DSU Production network is the primary network for DSU users. The network utilizes a 20 Gbps backbone and layer 3 routing to connect campus buildings to the main datacenter. At the access layer, users may connect to the network via 1 Gbps ethernet connections or the wireless network. DSU's wireless network provides wireless coverage in all DSU facilities, primarily utilizing 802.11ac Wave 2. Wifi 6 (802.11ax) is beginning to be deployed throughout campus and will be available in one residence hall and one academic building by the start of the Fall 2021 semester.

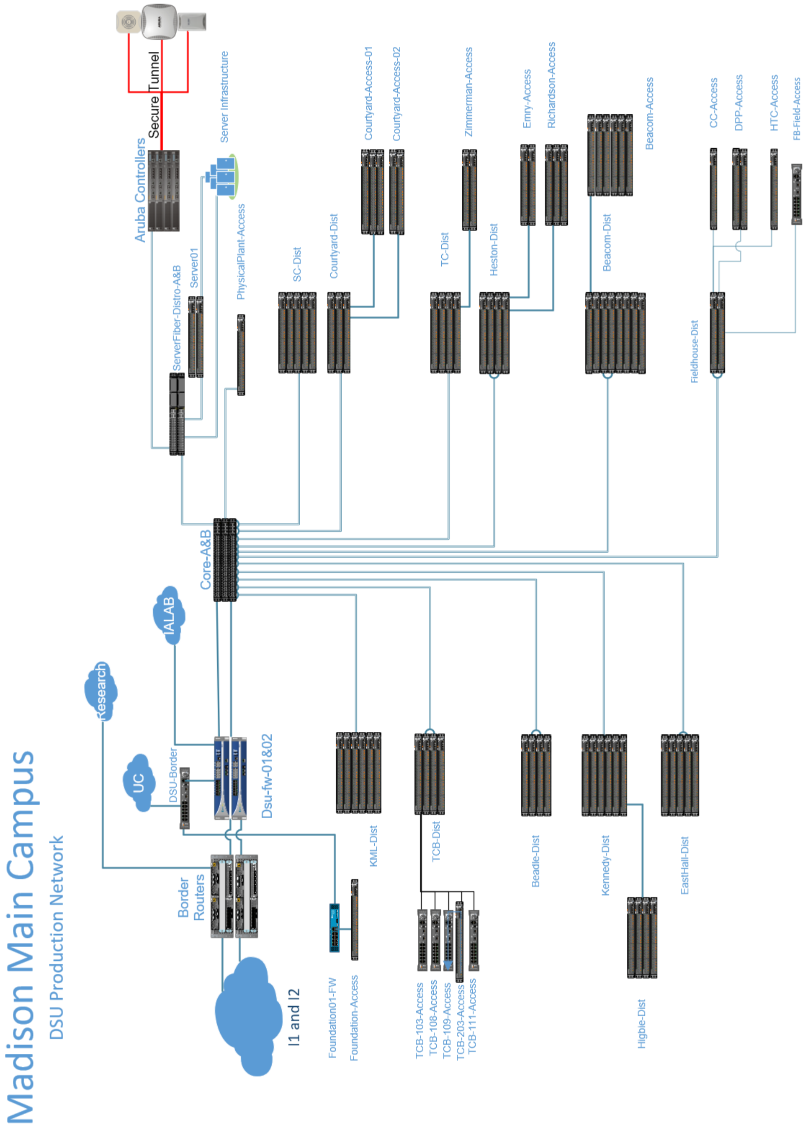


Figure 4.2.7.1 – Production Network

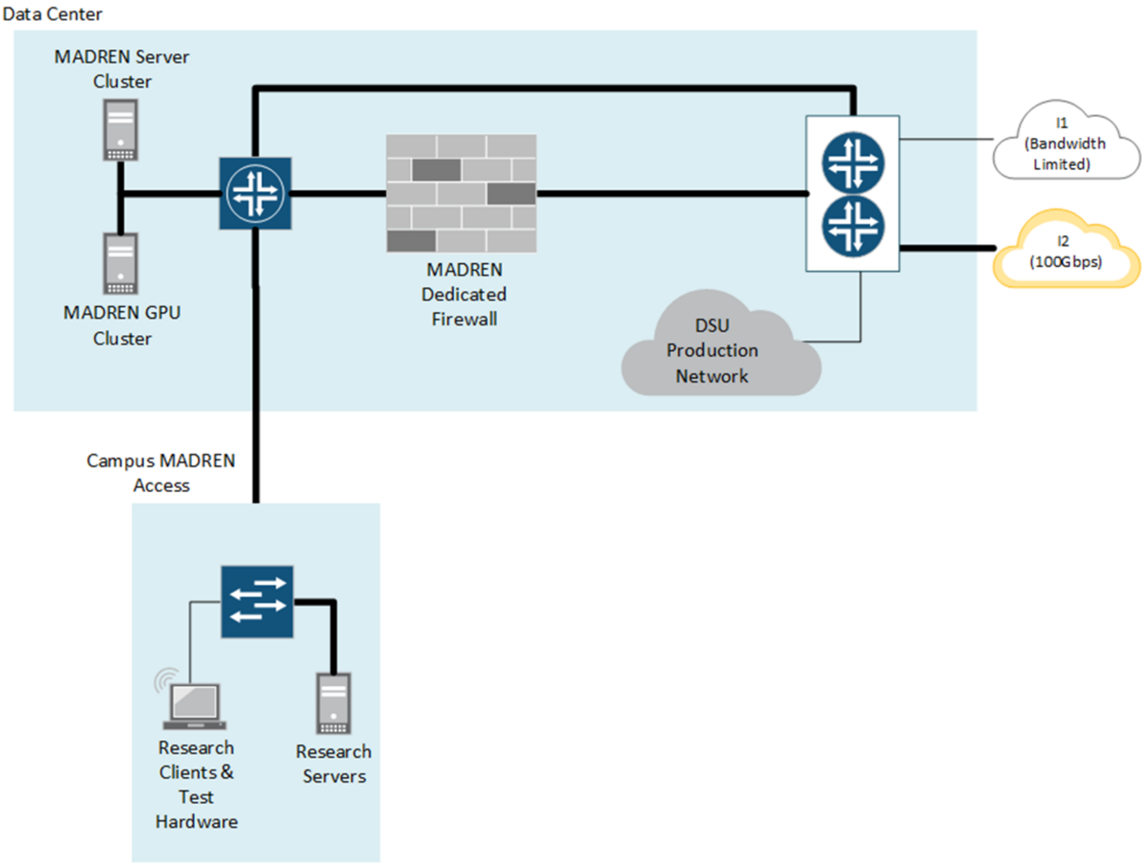


Figure 4.2.7.2 – MADREN

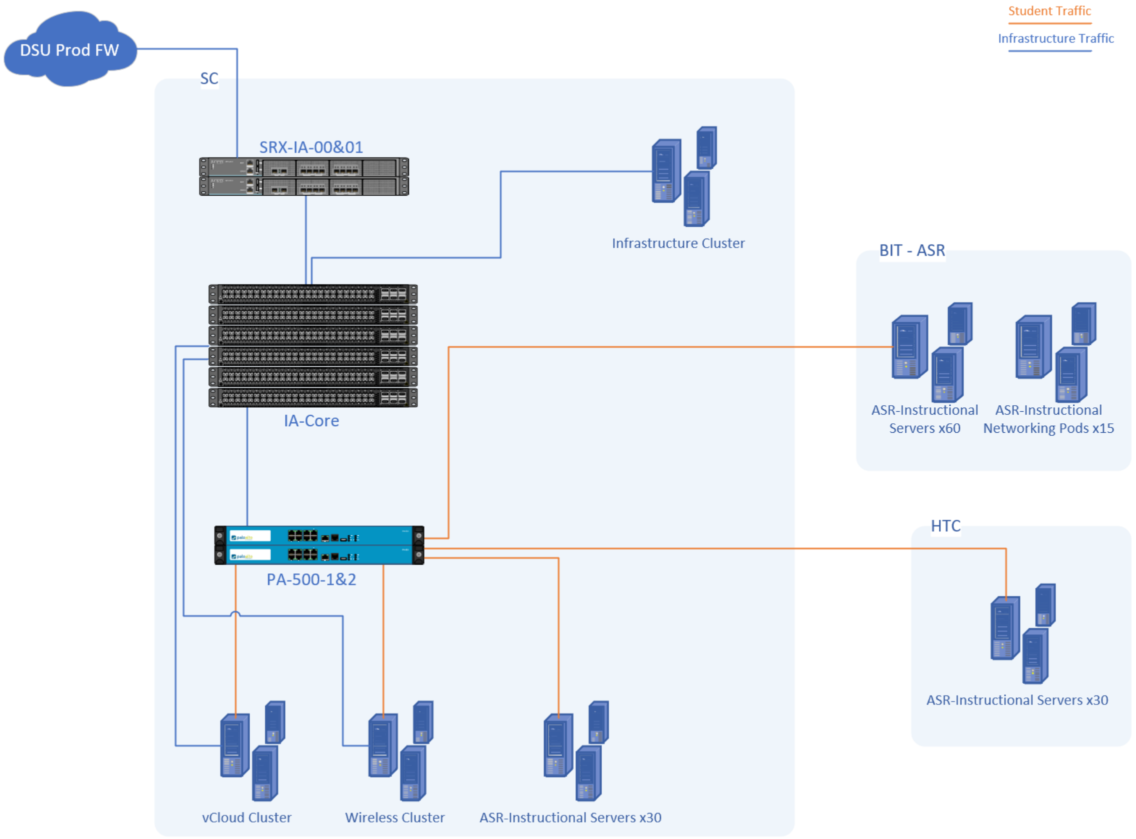


Figure 4.2.7.3 – IALab

At present, no high-performance data transfer technologies have been deployed in any environment. The MADREN has been deployed following the principles of a traditional science DMZ but tailored to meet the needs of DSU's researchers. We do have a perfSONAR node running but it is currently running tests only within the REED. The other perfSONAR nodes within the REED are only 10 Gbps capable so we do not have test data on our 100 Gbps throughput.

Present -2 years

- Expanding IALab to match continued growth of student population
- Planning and Deploying Wifi 6 to remainder of campus

Next 2-5 years

Expanding MADREN to external sites

Beyond 5 years

Unknown at this time.

4.2.8 Cloud Services

Present -2 years

- Working on creating a Government Cloud Azure Tenant for MADREN. This will give our researchers access to Microsoft service offerings such as Azure Cloud and Office 365. The added benefit is that being hosted within the Government Cloud area of Azure by default adds the security controls needed to work with controlled unclassified information, which a number of our researchers do work with.
- Explore Chameleon and how to help facilitate its use for our researchers.

Next 2-5 years
Unknown at this time.

Beyond 5 years
Unknown at this time.

4.2.9 Known Resource Constraints

Present -2 years

- At present, our main data center is nearing max capacity for both power and cooling. While there is additional room for more equipment physically in the space. Over the next year we will be working with datacenter engineers to complete an assessment of our data center to build a plan for increasing our capacity. Without fixing our capacity issues, we cannot expand our current systems.
- With the recent acquisition of a facility adjacent to campus, we now have a location we can use for a secondary datacenter. However, that space, while designed to be a datacenter, was not built with adequate cooling or power. We will also need to have the datacenter engineers review that space as well.
- Research IT Knowledge is another area where we have challenges. Our Information Technology Services department is fairly new to the world of research IT. While they are a group of highly skilled IT professionals, there is a lot of technology in the research IT world that is new to them which adds a learning curve and additional time to research solutions.

Next 2-5 years
Unknown at this time.

Beyond 5 years
Unknown at this time.

4.2.10 Outstanding Issues

None to report at this time.

4.3 DSU: South Dakota Center for SMART Power Systems

Content in this section authored by Dr. Shengjie Xu, Dakota State University and Dr. Reinaldo Tonkoski, South Dakota State University.

4.3.1 Science Background

The South Dakota (SD) Center for SMART (Secure Machine Learning and AI for ResilienT) Power Systems will uniquely strengthen the research capacity in SD by facilitating collaboration among experts with advanced skills in artificial intelligence (AI), communications, control, cyber security, data science (DS), and machine learning (ML) for the next-generation smart electric industry.

Types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project. The data produced in the course of this research project will be software codes, simulation models and PHIL output, technical papers, and presentations. No physical collections are included in this project. The amount of data produced for the period of performance is expected to be within 1Tb, and it will be updated periodically. None of the data is expected to be sensitive in nature.

The standards to be used for data and metadata format and content. We will follow the IEEE common data format for the power systems simulation data that will be generated in this project. For the other data such as technical papers and presentations, team members will follow the respective journal or conference proceedings' template as the format. Other generated simulation data will use standard and well-documented formats as applicable. Where existing standards are absent or deemed inadequate, the team will document this along with any proposed solutions or remedies.

Policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements. The team will host all the software codes with detailed documentation on a website hosted on Dakota State University's and South Dakota State Univh ersity's servers. We post all technical reports developed as part of this project on the team members' websites (as applicable) and on the project site to be created. We will be cognizant of the various privacy, confidentiality, security, intellectual property rights, and copyrights of these posted articles and will include appropriate disclaimers where required.

The availability of the data sets will be communicated to research peers at conferences and referenced in publications. The materials will be made available via persistent URLs. The team will endeavor to share results, including software, simulation environments, and other artifacts, with the research community. Codes will be made available as open source and hosted on publicly available repositories, such as, IEEE DataPort and GitHub.

Policies and provisions for re-use, re-distribution, and the production of derivatives. The PIs will allow the data for re-use, re-distribution, and the production of derivatives when permission is sought by the requesting entity. The PI(s) will specify conditions for data use and attribution in the specific Creative Commons license[4].

Plans for archiving data, samples, and other research products, and for preservation of/access to them. All data will be retained for at least five years after the end of this project or at least five years after publication, whichever is greater. Data related to graduate student research will be retained for at least five years after the degree is awarded. Data related to undergraduate student research will be retained for at least five years after the end of this project. Data that support patents will be retained for the entire term of the patent. Longer periods will apply when questions arise from inquiries or investigations with respect to research. In aggregate, these materials are less than or equal to 200 GB in size and will be stored, preserved, and made accessible at no cost to the project. The PI(s) will evaluate, distill, and select the data for preservation, and supply traditional and contextual metadata that will be reviewed and augmented as appropriate by librarians who are experts in digital repositories, including data and metadata organization and management. The data sets, associated publications, and metadata will be deposited and preserved in the digital repository for a period of at least three years after project completion.

The contributions of this work are not only from synthetic data, but also the refined methods discovered from the work. All of the produced data will be made available through open source software and data repositories.

4.3.2 Collaborators

The involved CS faculty in Dakota State University, and other involved faculty in universities such as South Dakota State University (Brookings, SD) and South Dakota Mines (Rapid City, SD).

Breadth and depth of the collaboration space:

- Number of users: 50 (including participating faculty members, students, and staff at DSU and SDSU)
- Number of participating facilities: 14; List: Sandia National Laboratories, Idaho National Laboratory, National Renewable Energy Laboratory; Pacific Northwest National Laboratory, National Rural Electric Cooperative Association (NRECA), Missouri River Energy Services, Heartland Consumers Power District, East River Electric, Brookings Municipal Utilities, NorthWestern Energy, BluWave AI, SBS CyberSecurity, BURNS McDonnell, GENPRO Energy Solutions

[4] http://creativecommons.org/

4.3.3 Instruments and Facilities

Dakota State University's (DSU) MadLabs Research Environment & Network (MADREN) is an extensive technology infrastructure dedicated to cyber security research. The MADREN includes a server cluster comprised of 10 Lenovo SR630s servers, each with dual Intel Xeon Gold 5118 Processors, for a total of 240 cores @ 2.3 GHz. The cluster is supported by 2.56TB of TruDDR4 @ 2666MHz RAM available and a 126TB HPE Nimble Storage Adaptive Flash Array. The MADREN also contains a large GPU cluster accessible through VMware Horizon and VMware Cloud Director. The total GPU capacity represents 324 teraFLOPS (Single-Precision), 2.6 petaFLOPS (Mixed-Precision), 5,200 TOPS (INT8), or 10,400 TOPS (INT4). These resources are accessible through virtualization via VMware Horizon and VMware Cloud Director.

SDSU Research Cyberinfrastructure (RCi) services within the SDSU Division of Technology and Security (DTS) manage an HPC cluster ("Roaring Thunder" or "RT"). RT has a general compute pool containing 56 compute nodes (2x Intel Gold 6148, 20 Core 2.4 GHz 40 core processors, 192GB RAM per system), 5 high memory nodes (2x Intel Gold 6148 40 core processors, 3TB RAM each system), and 4 - NVIDIA GPU nodes with 724GB of RAM for each system. Three GPU nodes have dual P100 16Gb RAM GPU's and one node has dual 16GB V100 GPUs. Cluster storage consists of a general parallel file system (GPFS) containing 1,500TB of high-speed storage space for data-intensive compute jobs. The cluster leverages three networks consisting of 100Gbps InfiniBand (IB) (Cluster data application processing), 10 Gbps Ethernet (science data transfers), and 1Gbps (cluster management). The cluster will be operational and available for the research team for the duration of the project. SDSU has an NSF MRI under review that may further expand these resources with AI- and Big Data-focused computer hardware for use during the project.

The aforementioned resources are used and customized for research projects with efficient computational requirements. The challenge & task will be to create virtual machines/templates with pre-installed scientific computing tools & libraries (e.g., Scikit-Learn, PyTorch, TensorFlow, etc.).

Present -2 years

Computing capabilities are sufficient for this time period, and the overall goal will be to link virtual machines to the available GPUs.

Next 2-5 years

Computing capabilities should be sufficient for this time period, but may trend upward as usage increases.

Beyond 5 years

It is unknown for this timescale.

4.3.4 Process of Science

The data produced in the course of this research project will be software codes, simulation models and PHIL output, technical papers, and presentations. No physical collections are included in this project. The amount of data produced for the period of performance is expected to be within 1Tb, and it will be updated periodically. None of the data is expected to be sensitive in nature.

4.3.5 Non-local Resources

Virtual machines that are organized by VMware Horizon and VMware Cloud Director at MADREN are used.

4.3.6 Software Infrastructure

Data Management and Analytics Tools:

- Commercial tool: MATLAB;
- Open source tools: Anaconda, RStudio/RStudio Server

Processes raw data into final and intermediate formats or data products:

- PCAP files, csv spreadsheets, image files (PNG, JPEG)

4.3.7 Network and Data Architecture

See Section 4.2.7 for more information.

4.3.8 Cloud Services

Chameleon, Google Colab Lab, and RStudio Cloud are all being explored.

4.3.9 Known Resource Constraints

A GPU licensing issue with each virtual machine in the MADREN does exist, and is being worked on.

4.3.10 Outstanding Issues

None to report at this time.

4.4 NSU: Connecting the Social Sciences across the Great Plains

Content in this section authored by Dr. Courtney Waid-Lindberg, Associate Professor of Sociology at Northern State University.

4.4.1 Science Background

Research takes different shapes; one is review and dissemination of data. Per the organization's website, "The Great Plains Sociological Association (GPSA) is a regional organization consisting of members from the Great Plains, particularly Minnesota, North Dakota, and South Dakota." The GPSA provides a mechanism for scientists in- and outside the sociology field to discuss and share their scholarship in a welcoming environment that promotes academic freedom and intellectual diversity. As part of doing so, the GPSA sponsors an annual meeting as well as the journal titled the "Great Plains Sociologist."

Starting in the fall, Northern State University (NSU) will host the journal via an online platform to build connections among sociologists in the region. Part of doing so requires an online "host mechanism," e.g., server, with the capacity to manage and grow capacities in data publication and dissemination. NSU is expected to host this journal, and potentially virtual workshops/conferences, now and into the future, though there is currently no mechanism in place to do so. Thus, an online platform via a cloud server or something of the sort is a necessary element of the infrastructure. All information will be managed and maintained by NSU, once the online/hosting infrastructure is in place.

4.4.2 Collaborators

The GPSA reaches across the northern Great Plains, such as Minnesota, North Dakota and South Dakota, but has potential to expand much farther across the Great Plains into states south of the Dakotas. Such reach could extend to Nebraska, Oklahoma, and Texas, allowing for data collection, publication and dissemination to expand beyond the current audience. Through a well-designed online platform and server, the GPSA can build a great journal resource through the Great Plains Sociologist while also expanding into other areas of data dissemination/research distribution through outreach activities, such as virtual workshops or speaker series via Zoom or Teams. In order to do so, the virtual infrastructure needs to be in place, e.g., servers, bandwidth, etc. Through the growth of the journal and potential workshops, it is expected that at least 15 institutions and 450 faculty/staff may participate each given year, whether through dissemination of data in the journal or through virtual discussions/workshops/speakers.

4.4.3 Instruments and Facilities

To host the journal, there is a need for a server and website adequate to do so. This may require a separate resource tied to both the GPSA and Northern State University. Northern State University faculty will manage the site and ensure that data are reviewed and disseminated appropriately. Further, any virtual discussions/workshops/speaker will require the appropriate platform, e.g., Zoom

Webinar or other such resource, in order to adequately reach a wide audience for data or research discussion.

Currently, there is no mechanism to host the journal at Northern State University nor host any virtual discussions through workshops/speakers. The foundation of such activities is required to promote these endeavors. Thereafter, the GPSA will build its capacity through the journal, workshops, etc.

Present -2 years
Initial start-up of host site for journal and potential workshops, discussions, or speakers.

Next 2-5 years
Build capacity of the journal, and expand in outreach activities.

Beyond 5 years
Maintenance of the journal and outreach activities.

4.4.4 Process of Science

The purpose of the GPSA is to connect a network of scientists across the region, whether through the journal, Great Plains Sociologist, or via conferences/workshops/discussions. Data produced by researchers will be disseminated via an online platform through the journal or virtual discussions. Designing an online platform with a server and the bandwidth large enough to handle hosting an online journal and collegial discussions will help promote communication among scientists and build capacities in scholarship throughout the region.

Present -2 years
Initial start-up of host site for journal and potential workshops, discussions, or speakers.

Next 2-5 years
Build capacity of the journal, and expand in outreach activities.

Beyond 5 years
Maintenance of the journal and outreach activities.

4.4.5 Non-local Resources

None to report at this time.

4.4.6 Software Infrastructure

None to report at this time.

4.4.7 Network and Data Architecture

None to report at this time.

4.4.8 Cloud Services

Cloud services may be used for storing publication data and allowing communication among colleagues. Regarding the platform for online discussions, Zoom and Microsoft Teams are preferred. However, we support the service that best serves the needs of online communication among colleagues.

Present -2 years

Initial start-up of host site for journal and potential workshops, discussions, or speakers.

Next 2-5 years

Build capacity of the journal, and expand in outreach activities.

Beyond 5 years

Maintenance of the journal and outreach activities.

4.4.9 Known Resource Constraints

We are limited in server space to support the journal, and cloud space to support the online synchronous communication. Thus, storage and speed are important for data dissemination and collegial communication. Currently, there is no "host" mechanism for the journal, and no "cloud" space designated to Zoom-related communications.

Present -2 years

Initial start-up of host site for journal and potential workshops, discussions, or speakers.

Next 2-5 years

Build capacity of the journal, and expand in outreach activities.

Beyond 5 years

Maintenance of the journal and outreach activities.

4.4.10 Outstanding Issues

None to report at this time.

4.5 NSU: Toward A Greater Dissemination of Social Science Data and Information for Better Civic Engagement

Content in this section authored by Jon D. Schaff, PhD, Director Center for Public History and Civic Engagement at Northern State University

4.4.1 Science Background

The Center for Public History and Civic Engagement (hereafter, "The Center"), beginning operation in Fall 2021, collects, analyzes, transfers, shares and stores data related to local and regional history and civic engagement. The Center promotes the study of social science as it relates American foundations and ideals in K-12 and post-secondary education and in the general public of South Dakota.

The Center has as part of its duties the transmission for general public usage of data and information related to public history and civic engagement. Northern faculty and students, particularly in the History and Political Science departments, are both engaged in a wide range of local and regional historical research. NSU currently offers two Public History programs: a 12-credit certificate and an 18-credit minor. Since 2012, 41 students have earned credentials in public history.

The NSU Williams Library operates and maintains a variety of digitization tools, including digital audio recording equipment, two tabloid size scanners, a slide scanner, a microfilm scanner, and an oversize scanning setup. The Betterlight scanning system allows for extremely high quality scans for items as large as 4'x 6'. This scanner has been used to digitize large volumes using a book cradle along with many oversize maps and art pieces.

The Library, as part of NSU and the BOR regental systems has access to substantial data storage resources via the ContentDM storage program, with access paid for yearly via OCLC. Both the library and the center are supported by the NSU technical services umbrella.

In the next 2-5 years, NSU and the Williams Library seek to substantially grow their regional history data sets. That region is defined as NSU, Aberdeen, Brown County, north-central South Dakota, and the Upper Great Plains. Local newspapers are a major source of data and a significant resource in local, regional, and national history. The addition of a large-format cradle scanner will allow the digitization of local newspapers and bound volumes of other local records. This data will complete existing data sets and provide increased digital access to a wide range of present and future researchers.

Beyond 5 years, NSU and the Williams Library seek to become a definitive repository of regional history. These collections will build up existing strengths in local history and regional history data sets including:

- Northern State University;
- Germans from Russia collections including extensive oral histories;

- Regional Native American history, especially the local Lake Traverse Indian Reservation;
- Aberdeen area history (including the Fischer Quintuplets and Frank L. Baum/Wizard of Oz materials);
- Upper Great Plains regional history.

4.4.2 Collaborators

The Center has a close partnership with the Beullah Williams Library on campus. It has as a goal to partner with K-12 schools across the state to disseminate relevant data and knowledge regarding public history and civic engagement. South Dakota contains 150 K-12 school districts. Other intended collaborators are state and local governments.

4.4.3 Instruments and Facilities

The Center needs enhanced audio/visual technology and increased bandwidth to achieve the goal of recording and/or broadcasting various Center events across the state in a high quality professional manner. Additionally, the Center currently lacks resources to provide stipends to leading social scientists and historians to share their research with the university, South Dakota K-12 communities, and state and local government.

Present goals include establishing the Center and beginning programming. Mid-term (two-five years) goals include establishing programming, including a speaker series, to disseminate social science and historical knowledge across the state. The long-term (beyond five years) goal is to have an endowed speaker series that allows us to perpetuate said programming and speaker series across the state. Currently the state lacks any sort of systematic programming in the social sciences and civic engagement.

The 108 volumes of the Aberdeen American News to be digitized are tightly bound in bound volumes and cannot be easily removed for flat-bed scanners. The Bookeye 4V1A scanner uses a book cradle and specialized software to digitize bound volumes. The physical condition of some of the newspapers is poor and it will be important to handle these materials as little as possible.

The goal of this project is to digitize approximately 108 bound volumes of the Aberdeen American News from 1936-1950 for which only poor quality microfilm images exist. These volumes also include Sunday editions which are not always present in microfilmed newspaper collections.

The DPM also has a number of copies of small, super-local papers that may or may not have ever been microfilmed. They date to the 1880s-1910s and include some rare titles.

4.4.4 Process of Science

Dissemination of civic and social science knowledge via experts. Knowledge is discovered by the university community and the various collaborators/stakeholders in the Center (e.g., K-12 education, state and local government, etc.).

The method of delivery of information will be via Web and other online broadcasting software. The Center will develop a websithat ich will archive various public presentations.

In the short term, the Center will purchase and become proficient in the use of professional quality audio/visual technology. In the mid-term (2-5 years) a speaker and programming series is to be established. In the long-term (beyond 5 years) the Center will have endowed a speaker series that will host major programming for the university and across the state to relevant collaborators.

4.4.5 Non-local Resources

None to report at this time.

4.4.6 Software Infrastructure

None to report at this time.

4.4.7 Network and Data Architecture

None to report at this time.

4.4.8 Cloud Services

The Center must archive recordings of various public events. This requires either substantial server space or cloud access.

4.4.9 Known Resource Constraints

None to report at this time.

4.4.10 Outstanding Issues

The Center is a new entity recognized by Northern State University. To that extent the Center has equal access to the limited resources of the university and some resources provided by the State of South Dakota. This includes IT resources. As the Center is new it is only now beginning its seeking of grant resources.

The University and state currently have no systematic programming related to social science or civic engagement activities. We are starting from scratch with limited short-term funds. The Center has no long-term funding.

The transportation of archival materials to the university for digitization will be a challenge. The physical condition of some of the newspapers to be digitized is poor and it will be important to handle these materials as little as possible.

Similarly, the bound volumes of the Aberdeen American News will need to be unbound to be digitized and then re-bound. The availability of these services in Aberdeen are uncertain and the Dakota Prairie Museum doesn’t want the newspapers removed from the city.

4.6 SD Mines: Realtime Weather Regional-Scale Weather Forecasting

Content in this section authored by Bill Capehart, South Dakota School of Mines

4.6.1 Science Background

SD Mines has provided real-time weather forecasts for western South Dakota for several years using the NCAR Weather Research and Forecast Model (WRF)[5]. These forecasts are typically done twice or four times a day given the available computing resources and create two forecast spaces, the first being a 9-km grid-spaced environment covering most of the Northern Great Plains, and a smaller domain 3-km grid-spaced domain covering western South Dakota. Earlier versions of the model consisted of tests with a third 1-km domain covering the SD Black Hills. The current forecast products are disseminated by a simple web page interface[6].

Earlier versions of this endeavor also employed a limited forecast ensemble in which the representations of cloud physics, convection, and boundary layer turbulence varied between the ensemble members. This was later leveraged in a multi-institutional NSF project, The Big Weather Web Project,[7] in which WRF forecast ensembles were performed across multiple organizations, with the results collated at a central repository where they were accessible by common frameworks used in the meteorology community by which to share and remotely process data (See Section 4.6.4).

Currently, our forecast capacity is limited by the need for dedicated machines that must be available four times a day (0300, 0900, 1500, 2100 UTC) for a period of approximately four wall-clock hours for our current capacity. The availability of high-performance computing resources, storage, and access would enable us to perform these forecasts faster, increasing the ability to deliver guidance to stakeholders. It would also allow us the ability to resume ensemble forecasts over the region as well resume use of the high-resolution Black Hills domain.

Additionally, a better framework for storing and processing data on-demand will give us the opportunity to provide a more customized dissemination of forecast products including more interactive graphics and raw forecast data for "power users" such as the National Weather Service and SD Wildland Fire.

4.6.2 Collaborators

Our real-time forecasting efforts leverage our relationship with the UCAR's Unidata Community[8] who provide a framework for member universities (including SD Mines) to share and access real-time weather observation and national- and global-scale forecast products which are critical for initializing forecast models. UCAR's

[5] https://www2.mmm.ucar.edu/wrf/users/

[6] http://kyrill.ias.sdsmt.edu/firemet/wrf_rap/d02.php

[7] https://bigweatherweb.org/Big_Weather_Web/Home/Home.html

[8] https://www.unidata.ucar.edu

National Center for Atmospheric Research also provides the support hub for community development and use of the Weather Research and Forecast Model that is used in our forecasting efforts.

Our earlier Big Weather Web initiatives involved a number of universities across the US and also benefited from support from Amazon Web Services to store and serve the data as well as XSEDE for long-term storage. (Our operational forecasts are not part of that repository).

Our users include the National Weather Service (specifically the Rapid City Forecast Office) and South Dakota Wildland Fire. In both user groups the effectiveness of our forecast products is reduced as product latency (or "staleness") increases.

SD Mines Atmospheric and Environmental Sciences has just been awarded a Unidata Equipment Grant to purchase three servers to replace aging parts of our program's cyberinfrastructure . While we may be committed to the contracted purchases, we are interested in optimizing the networking and shared resource potentials that collaborating with USD could bring.

The regional research institutions in SD have been given a grant to hire local specialists to help support CI.

4.6.3 Instruments and Facilities

The acquisition of data needed to run our forecast system relies on existing frameworks from NOAA who serve global- and continental-scale forecasts and surface and upper air data from UNIDATA.

Locally our models are running on two aging workstations connected through MPICH. Output is stored on local RAID drives with a limited capacity meaning that older forecasts must be discarded for new forecasts.

The amount of storage used for a given forecast member is currently 4 GB. Each forecast comprises the following inventories:

- 1 NetCDF gridded file per domain per forecast.
- 1-WMO Gridded Binary Edition 2 [GRIB2] file domain per forecast.
- 20 NetCDF weather station time-series files per forecast retained for the 9-km domain. If we were to expand to the Black Hills Domain this would add 5-6 additional files time series files.
- A time-series of image graphics for the web-based forecast display (stored in a directory which can be cleared with each forecast).

If we were to expand our domains to include the third Black Hills Domain then this would increase storage expectations to 5-6 GB per forecast.

Currently the forecasts are running once every 12 hours (0300 & 1500 UTC) but we would like to return a 6-hr forecasting cycle. Storing data for an extended period would be desirable for teaching and possible research. It is also important to collect an extended period of forecasts in order to collect the background error and model biases in model validation studies.

It would be desirable to have active access to 30-90-days of forecast in active storage (which is our current limit). We are willing to discuss the use of long-term "cold storage" for extended archiving of forecast model data that could be retrieved for case studies and periodic model validation and quality control.

Present -2 years
Expanding to four forecasts per day for three model domains (5-6 GB per forecast.

Next 2-5 years
Adding Forecast Ensembles (which would increase storage capacity) As we get closer to the 5-year mark, we may be seeing enough changes in the state-of-the-art numerical weather prediction practices to render the current or revised system obsolete.

Beyond 5 years
Beyond 5-years, we cannot project due to expected advances in numerical weather prediction models, the life-cycle of the WRF model, etc.

4.6.4 Process of Science

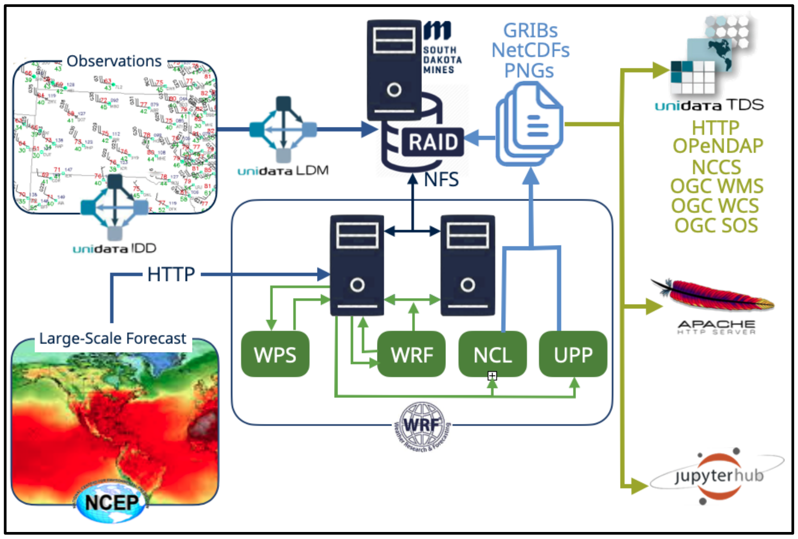


Figure 4.6.4.1 - Workflow

The current typical process of generating our forecast process is as follows:

- Preconditions to be able to execute a forecast:
    - NOAA produces a global- or continental-scale forecast.

    - Observations are collected by the Unidata Network and acquired through our Unidata Local Data Management (LDM)[9] system from the Unidata Internet Data Distribution (IDD)[10] Service's access to the global WMO Global Telecommunications System (GTS)[11] feed.
- Local Generation of Forecast
    - Acquire a global forecast by actively connecting to NOAA Operational Model Archive and Distribution System (NOMADS)[12] using axel (current method).
    - Preprocessing of NOAA/NOMADS and UNIDATA with the WRF Preprocessing System (WPS).
    - Execute WRF (currently using two machines joined through MPICH with a run-time of approximately 3-hrs - we'd like HPC resources to reduce the runtime)
- Post Processing of Forecast
    - Native WRF NetCDF output files in are renamed for long-term storage
    - Native WRF NetCDF output files are converted to WMO GRIB Format for easier standardized access using the Universal Post Processor[13]
    - Native WRF ASCII time series access files are converted into standardized and community-compliant NetCDF Time Series files.
    - Graphics for the web page are created (currently uses NCAR Command Language [NCL][14] which is approaching end-of-life in favor of using Python-oriented resources).
    - Cleaning up Temporary Files needed to generate the forecast and wait for the next forecast cycle.
- User Access
    - Users can currently access the current models output from a simple graphical web page and can access the archived GRIB data, and NetCDF time series data from a Unidata Thematic Real-time Environmental Distributed Data Services (THREDDS)[15] server.
        - Current URL: http://kyrill.ias.sdsmt.edu:8080/thredds/

We are using a framework that has not been updated in several years due to limited infrastructure. During this period there have been a number of changes in technology used for sharing and processing meteorological data. This is an ideal time to move our system forward to accommodate these changes.

Present -2 years

- Move to the latest edition of WRF (this does not change the workflow)

[9] https://www.unidata.ucar.edu/software/ldm/

[10] https://www.unidata.ucar.edu/projects/idd/iddams.html?query_float=IDD

[11] https://community.wmo.int/activity-areas/global-telecommunication-system-gts

[12] https://nomads.ncep.noaa.gov

[13] https://ral.ucar.edu/solutions/products/unified-post-processor-upp

[14] https://www.ncl.ucar.edu

[15] https://www.unidata.ucar.edu/software/tds/

- Improve speed access to NCEP products needed to initialize the WRF model. This is one of the major networking bottlenecks to the forecast product.
- Move the WRF execution to a faster computing platform than the current one.
- Migrate to a Python front and backend for processing for managing workflow. This includes leaving NCL as the graphics generator for Python resources. This also could include a JupyterHub[16] type framework by which a user can access remote computing resources and process data at the point of storage rather than retrieving data before accessing it.
- Create a customizable end-user interface to replace the static design of the current one.

Next 2-5 years

Given the constant state of change and updates to best practices in the operational meteorology community, we cannot project changes in workflow beyond the current framework. Some of the WRF preprocessing framework uses legacy approaches dating back to the 1990's (!) and is likely due for a major update. Likewise WRF may be approaching an end-of-life in the coming years in favor of newer modeling systems, e.g., MPAS[17]. These would likely involve a complete reinventing of our real-time modeling frameworks.

Beyond 5 years

Again, this is "past our headlights"

4.6.5 Non-local Resources

Operational Numerical Weather Prediction relies heavily on being able to access remote resources in real time. The operational meteorological community has created a number of resources to facilitate data access and delivery over the years. Many of them are currently leverated in the current workflow and process. Others listed here can be leveraged for this application.

- NOAA Operational Model Archive and Distribution System (NOMADS)[18]: The operational repositories of numerical weather prediction forecasts can be accessed and downloaded through the NOMADS services. Here forecast model output can be accessed as complete downloads (e.g., with wget, axel, or other resources). NOMADS also has more advanced means for users to retrieve data from its website, such as using OPEnDAP by which a client can request a specific field and grid-elements of data from a given file through web services rather than by downloading the entire file. However, the

[16] https://jupyter.org/hub

[17] https://ncar.ucar.edu/what-we-offer/models/model-prediction-across-scales-mpas

[18] https://nomads.ncep.noaa.gov

current framework for initializing WRF and bringing in the large-scale model data into the WRF framework does not permit this more efficient approach.

- Unidata Internet Data Distribution (IDD)[19] Service: Despite its age, IDD continues to be a resource that connects universities and similar organizations to the live real-time GTS (Global Telecommunication System)[20] that distributes meteorological data across the globe. For universities in the Unidata network, the means by which we harvest the data moving through the GTS feed is by using. Unidata's Local Data manager (LDM)[21] which monitors the feed and intercepts specific weather data messages, including station, ship and buoy observations, weather balloon data, and forecast model output. We currently use this in our real-time efforts to capture weather stations and upper air observations. Gridded data needed for WRF could be accessed through the UNIDATA IDD-CONDUIT[22] feeds but past experience has shown that this may result in data loss during model data transition.
- Unidata Thematic Real-time Environmental Distributed Data Services (THREDDS)[23]: THREDDS data servers are used at SD Mines to serve complex datasets and leverage data sharing protocols such as OPeNDAP, OGC WCS, OGC WMS, and HTTP. Atmospheric Sciences software has been designed to leverage OPeNDAP in its standardised datasets archived as NetCDF and GRIB. ArcGIS can also access data through OPeNDAP services. SD Mines current server[24]. While we can use data served on THREDDS to share the final forecast products to "power users," we cannot use the resource to initialize WRF beyond using the HTTP services from THREDDS services on the aforementioned NOAA NOMADS servers.
- JupyterHub Notebook Servers[25]. Jupyter Notebooks[26] (and now JupyterLab[27]) are increasingly being used in computer science and in particular data science in the weather and climate sciences. While Jupyter Notebooks and JupyterLab are typically used on a local machine, including a users laptop,
- Google Colab[28]. Google Colab is a web-based means to serve and share Jupyter notebooks. We've tried it here. It made my students cry, and one almost rage-quit, since they are running on virtual machines that require specific software installs that may be above the paygrade of students (and faculty). Our community prefers the JupyterHub approach.

[19] https://www.unidata.ucar.edu/projects/idd/iddams.html?query_float=IDD

[20] https://www.wmo.int/pages/prog/www/TEM/GTS/index_en.html

[21] https://www.unidata.ucar.edu/software/ldm/

[22] https://www.unidata.ucar.edu/data/conduit/index.html

[23] https://www.unidata.ucar.edu/software/tds/

[24] http://kyrill.ias.sdsmt.edu:8080/thredds/catalog/catalog.html

[25] https://jupyter.org/hub

[26] https://jupyter.org

[27] https://blog.jupyter.org/jupyterlab-is-ready-for-users-5a6f039b8906?gi=eeda36d7d07f

[28] https://research.google.com/colaboratory/faq.html

Present -2 years
Improved and faster access to NCEP-generated output

Next 2-5 years
The aforementioned Headlights Problem.

Beyond 5 years
A Bigger Headlights Problem.

4.6.6 Software Infrastructure
Local resources involve the following:

- Software
    - Fortran, and C-family compilers (GCC and Portland Group -- note that the BOR rejected license renewals for Portland Group Fortran compilers due to 2009's HB 1022).
    - MPICH
    - NCAR Command Language for graphics and data processing
    - IDL for Scripting
    - migrating to Python 3 from NCAR Command Language to IDL
- Hardware
    - Unix Network. Currently this network is aging. We have just received a Unidata Equipment Grant to replace some of our aging machines.
    - RAID Storage Services (currently at 63TB total local available storage) We also have some data distributed across a number of workations on hard disks that likely need to be placed back under RAID storage.

Present -2 years
- Implement UNIDATA Equipment Grant
- Upgrade networking and accomodations for a campus level admin (which may include integrating some of our resources in to a campus-level position)

Next 2-5 years
Maintain upgrades and continue migrating to best practices for a modern university shared network.

Beyond 5 years
"Past our headlights"

4.6.7 Network and Data Architecture
SDSMT campus is a "STAR" network with 100G to each of the "research" buildings with 10-40-100 to other buildings. By August 2021 we will be capable of 100G to most buildings and into research labs. We have one building with asbestos issues

that will be torn down and replaced in the next few years. By August 2021 SDSMT will be connected at 100G to the REED network and through to GPN and Internet2.

4.6.8 Cloud Services

We have previously used AWS and XSEDE resources on other projects. Currently we are not using cloud resources. We are curious as to how these can be used. However, note that we will be needing to have computing resources on demand in order to produce timely forecast guidance.

Present -2 years

- Working with Pacific Research Platform (PRP) FIONA and Nautilus nodes.
- With the GPARGO grant we will be expanding our involvement with PRP and XSEDE.

Next 2-5 years

Working with our researchers and their needs, especially in XSEDE, PRP FIONA and Nautilus. Beyond that it is unknown.

Beyond 5 years

Unknown

4.6.9 Known Resource Constraints

The primary issue here is timeliness of forecasts as well as being able to serve and store recent forecasts to stakeholders. We also are indeed of local specialized technology support.

Present -2 years

Mines will be hiring a scientific computing specialist based

Next 2-5 years

Securing Long Term Funding for support

Beyond 5 years

“Past our headlights”

4.6.10 Outstanding Issues

No additional issues to report at this time.

4.7 SDSU Research Cyberinfrastructure Center

Content in this section authored by Kevin Brandt, Chad Julius, and Jeff Mahlum from South Dakota State University: Division of Technology and Security.

4.7.1 Science Background

As the state's 1862 Morrill Act land-grant institution, the research work of SDSU is carried out on its main resident campus in Brookings, at sites in Sioux Falls, Pierre, Rapid City, Aberdeen and through Extension offices and Agricultural Experiment Station research sites across South Dakota.

In addition to the Agricultural Experiment Stations, SDSU's other Research Centers include the: E.A. Martin Program in Human Nutrition, Geospatial Science Center of Excellence, Mountain-Plains Consortium, North Central Regional Sun Grant Institute, South Dakota Water Resources Institute and the Water and Environmental Engineering Research Center and also collaboration with the SD is involved in the EPSCoR 2DBEST Center. SDSU also hosts a Functional Genomics Core Facility, Core Mass Spectrometry Facility, Animal Disease Research and Diagnostic Laboratory, Materials Evaluation and Testing Laboratory, SDSU Seed Testing Laboratory, and a Genomics Sequencing Facility. SDSU also hosts the South Dakota State Climate Office, which supplies climate and drought information and records weather data from many stations located throughout the state. The Division of Technology Services (DTS) provides critical research cyberinfrastructure services to these and other research laboratories/groups within the university.

4.7.2 Collaborators

SDSU Research Cyberinfrastructure(RCi) and University Networking Services within the DTS work closely with the South Dakota Board of Regents - Regent Information Systems, who coordinate Wide Area Network (WAN) research connectivity services (Research, Education and Economic Development (REED) network), through the South Dakota(SD) State Bureau of Information and Telecommunications (BIT). Regional Connectivity to Internet 2 is provided to the REED network by the Great Plains Regional Network (GPN) who also coordinates the Great Plains Research Platform as an important Cyberinfrastructure service to its member institutions.

State BOR state higher education institutes are active members of the Northern Tier Network. This consortium provides coordination activities related to backup out-of-state research wide-area network connections to Internet2.

Great Plains Network Institutional Collaboration:

- NSF CC* Team: Great Plains Regional CyberTeam was awarded by the National Science Foundation. Award #1925681, (July 1, 2019, to June 30, 2022)
- NSF CC* Compute : GP-ARGO: The Great Plains Augmented Regional Gateway to the Open Science Grid - National Science Foundation. Award # 2018766, (July 1, 2020, to June 30, 2023)

In-State Institutional Collaboration, South Dakota Board Of Regents Research & Development Innovation grants:

- 2018: Project Titled: “Building the Next Generation of the South Dakota Research, Education, and Economic Development Network” Established 100 GBPS connectivity to the SD REED/GPN/I2 network(s) for USD, SDSU and DSU (August 01, 2018, to June 30, 2021).
- 2021: Proposal Titled: "The South Dakota Research Cyberinfrastructure Initiative: Expanding Research Cyberinfrastructure Resources Enabling Next Generation Research and Education Programs within South Dakota." Enhance CI services/resources at SDSU, USD, and the SDSMT. (April 15, 2021, to April 14, 2022).

4.7.3 Instruments and Facilities

Hardware Cyberinfrastructure

SDSU Research Cyberinfrastructure (RCi) manages the largest HPC cluster instrument within state public higher education ("Roaring Thunder" or "RT"). RT has a high-performance CPU intensive, large memory, and NVIDIA P and V series GPU cluster computing pool(s). Cluster storage consists of a parallel file system (PFS) containing high-speed storage space to support data-intensive cluster compute jobs. The cluster leverages three networks consisting of 100Gbps InfiniBand (IB) (Cluster data application processing), 10 Gbps Ethernet (science data transfers), and 1Gbps (cluster management). Project funding came from an NSF grant (MRI: Acquisition of a High-Performance Cluster to Enable Advanced Bioscience and Engineering Research), NSF 15-504, MRI Project Reference:1726946 Grant Award: $796,359 Term: 10-01-17 thru 12-31-18).

SDSU RCi provides access to over sixty single compute server systems to address applications and research needs outside of SDSU's cluster computing environment. These servers include a VMWare cluster providing an HPC virtual environment, multiple high-memory (1-3TB RAM) systems for genomics, and a mix of other systems providing tools to host department-specific applications and other service functions to support the research community. GPFS storage from RT can be exported to these systems over NFS or through direct InfiniBand (IB) 100Gbps connections (Protocol Nodes) to lessen the need for extensive data set movements.

Research Network

In 2014, the South Dakota State University (SDSU) Division of Technology and Security (DTS) were awarded an NFS CC* Cyberinfrastructure Program grant: CC*IIE Networking Infrastructure: Building a Science DMZ and Enhancing Science Data Movement to Support Data-Intensive Computational Research at South Dakota State University, NSF 14-521 CC*IIE Project Reference: 1440622 Grant Award: $494,520.51. Term: 12-01-14 thru 11-30-16

Research Storage/Archival Services

The DTS manages a parallel file storage system and various block storage systems. These include a new multi-petabyte block storage system with various NAS (Network Attached Storage) devices for research laboratory use. These systems provide enterprise-class storage for single instance servers and cluster systems. Raw data, metadata, and research products are archived and made publicly available within SDSU's OpenPrairie, a product contracted between SDSU and the Be press corporation (Digital Commons). The SDSU DTS manages multi-petabyte cold storage data repositories for science data archival purposes.

Data Center Facilities

The SDSU DTS houses all servers, storage, and central networking equipment for the Brookings, SD (South Dakota) campus within the Morrill Hall building, rooms 112 and 114. This area is configured in a hot-and-cold aisle layout with all cabling overhead in existing trays.

Present -2 years

- Complete the implementation of a new block and file storage to support the growth of research data capacity needs within sponsored research.
- Upgrade it's single node compute infrastructure to include a mix of large memory and high-performance CPU systems. The typical use for these systems primarily supports research involving Genomics, Engineering Simulation, and other discipline areas. They are designed to connect with 10/100 Gbps Ethernet and 100 Gbps InfiniBand for direct cluster storage (Parallel File System) access.
- Deploy an additional HPC (High-Performance Computing) cluster, and a parallel file system to support growth within institutional research and education.

Next 2-5 years

- Continue response to funding opportunities for expanding cyberinfrastructure services that support research at SDSU and other state public higher-education institutes .

Beyond 5 years

- Harder to determine but will be aligned with the strategic goals and objectives of the institution and the SDBOR.

4.7.4 Process of Science

This initiative involves a broad examination of research data flows across security equipment within SDSU's research network.

4.7.5 Non-local Resources

This section is not relevant to the SDSU technology support team.

4.7.6 Software Infrastructure

SDSU's Research Cyberinfrastructure and University Networking Services currently provides software / file-sharing resources to research staff using:

- Box.Net (Commercial Cloud) - Secure cloud data/file management and collaboration tool used to store or share data.
- Microsoft Office 365 (Commercial Cloud) - Secure data storage and collaboration tools.
- File Servers- Microsoft Windows Server 2012-2019 (SMB) based file sharing secured using NTFS as part of Active Directory domain services available on the campus LAN and through an authorized VPN connection.
- Current data transfer and management tools used to transfer data sets to internal and external resources include:
  - Globus (Commercial) - SaaS tool providing a data management endpoint and desktop client that ties campus storage together for internal and external data movement.
  - SCP (Open Source) - Primarily used for secure copy over SSH between Linux systems (WinSCP for Windows) on the campus LAN or through VPN services.
  - Cloud Sync Agents (Commercial/Open Source) - Desktop tools that allow simple data syncing and data movement to their respective cloud platforms.
- Over three hundred cluster software application modules supporting research / scientific pipelines.
- Research applications that are broadly used that involve data flow / can require high network throughput and large storage capacity.
  - MathWorks Matlab - site license (full module suite)
  - Mathematica - network license
  - ESRI ArcGIS Desktop / Server application suite
  - SAS statistical applications
  - R, R and RStudio statistical applications

Present -2 years

Current data transfer tools such as FTP, SCP, and Cloud Agents should not change much, if at all, during this period. Globus will continue to be developed and deployed to other research resources. For example, the current Globus managed endpoint ties directly to cluster storage through a 100Gbps InfiniBand connection. With the development and/or purchase of new equipment (cluster computing etc.), there will be further development of Globus file transfer services.

Next 2-5 years

Cloud applications may change with heavier use of Microsoft Azure services. Globus software will most likely continue to be the primary data transfer tool of choice for scientific data transfers.

Beyond 5 years

Driven by future research needs, software tools for data movement and management like FTP, SCP, and others will more than likely still be a part of a scientific pipeline. Innovative technologies may be groundbreaking when it comes to this far out in the timeframe.

4.7.7 Network and Data Architecture

The current network infrastructure (Figure 4.7.7.1), shows the current local and Science DMZ network and the path out to I1 and I2. It should be noted that the server named *FIONA*, is a Globus managed endpoint that connects at 100Gbps. The core router routes all faculty/staff network traffic. The core router routes all traffic through the university firewall that in turn routes everything to the university Edge router. From that Edge, all routes in BGP route to I2 through the South Dakota BIT router, all routes not in the BGP table, route through SD BIT to I1.

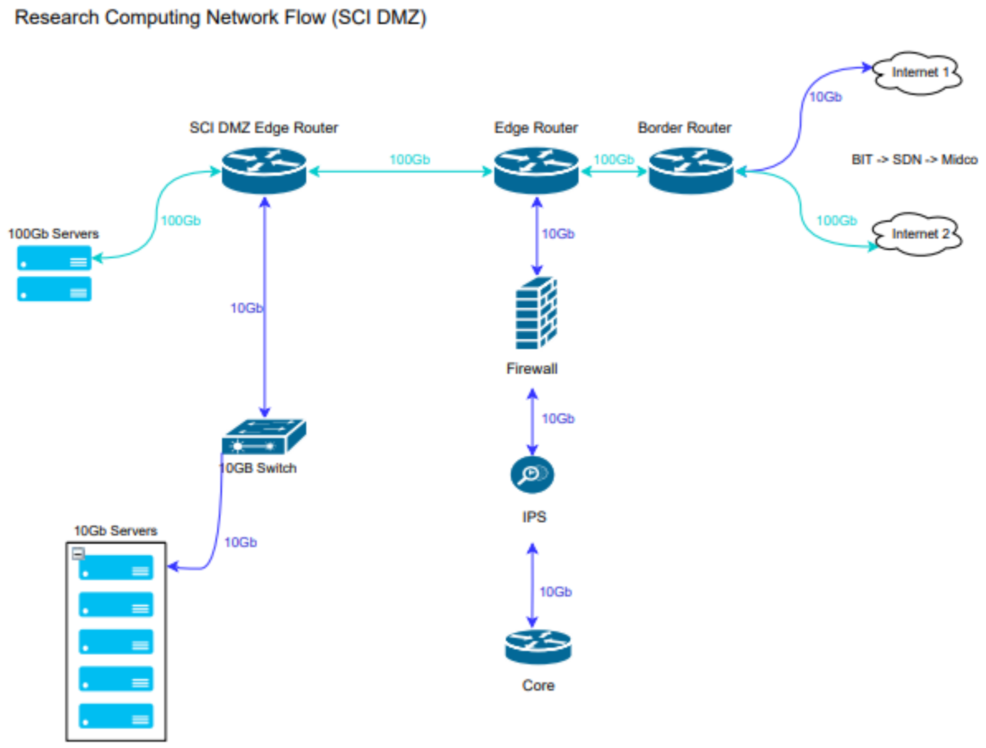


Figure 4.7.7.1 – SDSU Network Architecture

Attached to the Core router(s) are sub-core switches connected at 40Gbps (Figure 4.7.7.1). Building switch stacks are connected to the sub-core switches at either 1Gbps or 10 Gbps connectivity, depending on building use needs. Individual port connections for faculty, staff and equipment are at speeds of 1Gbps, with the ability to push mGig (1/2.5/ 5/ 10) on one ASIC per switch.

Present -2 years

- Replace the university edge router(s).

- Add two PerfSonar nodes, one on the university edge router and one on the WAN border router.

Next 2-5 years

- Continue to leverage any future expansion of REED. GPN and Internet 2 services.

Beyond 5 years

- A phased approach to an internal network rebuild.

4.7.8 Cloud Services

While there are some capabilities to burst into the cloud (AWS, Azure) within the RT cluster, this feature has not been fully implemented yet. University Networking has a footprint within Azure for external replication of services such as DNS, domain controller and some web applications.

Present -2 years

- Continued use of XSEDE cloud / related services.
- Investigate commercial cloud compute solutions.
- Research an alternative long-term cloud cold storage solution.

Next 2-5 years

This will be aligned with the needs of the SDSU and its collaborators.

Beyond 5 years

This will be aligned with the needs of the SDSU and its collaborators.

4.7.9 Known Resource Constraints

Problem: All internal-external research data traffic traverses SDSU's Intrusion Detection Protection system and campus firewall before reaching the SDSU Science DMZ (or campus DMZ). At times, this path can adversely impact the rate of external research data transfers.

Need: The DTS seek expert knowledge on an alternative approach to reduce data flow friction between research labs/instruments, cluster, and the SDSU Science DMZ. Such a solution should not adversely compromise the system and data security.

Present -2 years

- The use of AI technologies within multiple disciplines (i.e., Precision Ag, Genomics, Engineering) will require transferring and computing massive amounts of data. Moving that data from the field to campus researcher storage will strain current network connectivity.

- Data/System Security Needs: Balance security with data transfer process usability (Researcher request(s): easy-of-use for data transfers).
- With the launch of Landsat 9, SDSU's Image Processing Lab will be transferring and computing data from high-resolution images requiring low network latency and high network throughput.
- Address a knowledge gap in CI network performance tuning to assure high-speed data transfers of research data.
- To support and foster growth in research innovation, there is a need to significantly increase compute and cluster storage capacities.

Next 2-5 years

- Increase university research network backbone connectivity.
- The continued growth of HPC resources focusing on higher education needs within SDSU, the SDBOR, and its collaborators.

Beyond 5 years

Continued growth aligned with the strategic plans of SDSU and the SDBOR.

4.7.10 Outstanding Issues

No additional issues to report at the time of this study.

4.8 USD: South Dakota Biomedical Research Infrastructure Network (SD BRIN) Proteomics Core Facility

Content in this section authored by Dr. Eduardo Callegari, Dr. Barbara Goodman, Maria Daniela Paez, from the University of South Dakota, Sanford School of Medicine, Division of Basic Biomedical Sciences

4.8.1 Science Background

The Proteomics Core Facility (PCF) is a collaborative research enterprise established in 2002 and funded by SD BRIN NIGMS-NIH (2002 to date). The facility has been providing protein identification, characterization and quantitation (proteomics analysis as a service and collaboration for researchers throughout South Dakota, nationwide, and globally). The data are generated from the Liquid Chromatography- Mass Spectrometry instrumentation during the proteins and peptides analysis, mainly RAW files and other formats such as MGF, XML, PKL, Dat, mzML, etc.

Examples of the projects in which PCF has been collaborating:

- Biomedical Sciences (breast, prostate and pancreatic cancer, regenerative medicine, neurodegenerative diseases, Autoimmune diseases, bacterial and viral diseases, etc.),
- Veterinary and Agriculture (plants, animals from farms and diseases associated with that)
- Biotechnology (2nd generation of biofuels)
- Bioremediation (the use of microorganisms in the process of the contaminant, effluents from factories and ecology)
- Paleoproteomics (analysis of food from archaeological samples), and other emergent fields.

The original data coming from different projects are backed up and only the copy is processed for the analysis. All the information collected is under the custody of the PCF. The results generated can become:

- a hypothesis for grant submission,
- inclusion in an honor's, master's, or Ph.D. project
- an abstract for scientific meetings and peer-reviewed journals.

4.8.2 Collaborators

The PCF users are:

- The scientific community from the University of South Dakota
- The SD BRIN partner sites (six primary undergraduate institutions or PUIs)
- Other universities such as:
  - South Dakota State University
  - North Dakota State University
  - Florida International University
  - McGill University
  - Collaborators in South America (Chile, Colombia, Peru, and Argentina)
  - Collaborators in Europe (Spain and Germany)

The Facility since its inception, has analyzed more than 20,000 samples corresponding to 130 research projects from approximately 110 users.

Facilities that collaborate with PCF are:

- WestCore (nucleic acid sequencing core facility at Black Hills State University)
- Mass Spectrometry Facility from North Dakota State University
- University of North Dakota-School of Medicine and Health Sciences
- SD-BRIN Bioinformatics Core and University of South Dakota IT Research Computing Group

Also, the PCF works with the other IDeA states in the network of IDeA National Resource for Proteomics (the main leaders are the University of Arkansas Medical School, University of Oklahoma Health Sciences Center and Oklahoma Medical Research Foundation).

4.8.3 Instruments and Facilities

The following list of instruments are used to provide Proteomics analysis (protein identification, characterization and relative quantification):

- ***Typhoon 9410 multi-imaging system*** (GE Healthc are)
- ***Proteome Works spot cutter*** (Bio-Rad)
- ***NanoAcquity Ultra Performance Liquid Chromatography*** (UPLC) in 1D and 2D configuration (2D-nanoUPLC)
- ***Ultimate 3000 RS nanoUHPLC with nano***, capillary as well as microflow options
- ***Easy nLC 1200 UHPLC chromatography system*** (Thermo Scientific)
- ***Quadrupole Time of Flight mass spectrometer*** (Waters Synapt G1 HDMS) with Ion Mobility and MSe technology
- ***Orbitrap with high resolution/accuracy mass (HRAM) mass spectrometer*** (QExactive Plus, Thermo Scientific) for MS, MS/MS, DIA and PRM protocols with enhanced resolution up to 280.000, allowing to work in Protein Mode for intact protein analysis
- ***Gelfree 8100 Fractionator station*** (Expedeon) for protein fractionation in solution from complexes mixtures without the necessity of band or spot cutting
- ***MassPrep work station*** (automatic liquid handling system for protein digestion).

The facility is in the process of upgrading the Waters Synapt G1 HDMS mass spectrometer to an orbitrap with High Resolution/ Accurate Mass (HRAM) mass spectrometer. The PCF recently obtained grant funds to allow this purchase.

Acquisition of a new nitrogen generator that can support the nitrogen gas necessary for the correct operation of the two mass spectrometers.

Resources available for the users

Most of the resources available to the users at the PCF, are described above. The high complex instruments are indirectly available for the users because of the experience, time, and expertise that require their operation. The facility operates those instruments and provides the analysis and final report to the users.
Other resources such as tools for protein analysis are available for the users and also the PCF personnel provide training, consulting and advice (see software and hardware infrastructure).

Composition of the data sets

A standard RAW file generated at the sample analysis through the mass spectrometer per hour has an average size between 300 to 500 Mb and sometimes 900 MB or more, depending on the complexity of the samples analyzed. The instruments operate 24/7, 365 days per year. The daily number of files included in a directory or project per day depends on the number of samples available for services. We can estimate 15 files per day.

The PCF utilizes space on multiple storage appliances, including the South Dakota Data Store (SDDS), and NSF-funded dual-tier platform consisting of a disk-based "Sharing" tier and tape-based "Archival" tier. There are immediate plans to consolidate all PCF data to SDDS (i.e., migrate off of older storage systems), and the SDDS platform is greatly expandible and should serve the PCF's storage capacity needs for the foreseeable future.

USD operates The South Dakota Data Store ("SDDS," funded by NSF award ACI-1659282), housed at the USD Main Campus data center in Vermillion, SD and the USD Community College for Sioux Falls data center in Sioux Falls, SD. SDDS is accessible via the cloud-based Globus data management platform (globus.org). SDDS includes a high-capacity disk-based Sharing Tier for data sharing as well as an Archival Tier hosted on magnetic tape. The Sharing Tier has a current capacity of over 400 TB (expandable to 1 PB) and the Archival Tier's current capacity is over 1 PB (expandable to 4.5 PB). SDDS also directly supports data curation and publication through Globus, including popular metadata standards like the Dublin Core.

4.8.4 Process of Science

The process proceeds as follows:

1. The Facility meets with the users and discuss the experimental design, different approaches and protocols to collect, process and analyze the samples.
2. The PCF receives the samples, organizes them, and creates an internal protocol with the characteristics and source of the samples.
3. The samples are processed and analyzed through the Liquid Chromatography-tandem mass spectrometry.

4. The RAW file generated at the step before is moved into another PC workstation and two extra backups are made. The file is processed for the mass or peak list extraction.
5. Those files are exported into Bioinformatics tools to perform the "translation" of the data from numbers (mass to charge ratio or m/z) into peptide sequences and later in proteins ID.
6. The list of proteins identified is organized, analyzed, and visualized using post-mass spectrometry tools.
7. The following step depends on what the users require.
    a. If they want to analyze by itself, the facility exports the data in different formats according to the preferences of the user (can be either MGF, XML, Dat, mxML, MSF, etc.).
    b. If the users request the facility's help, a meeting is set up to discuss the data organization, analysis, visualization and if certain formats are required for either publication or grant application.
8. Finally, before submitting the data for publication, the users and the facility have a meeting and discuss the contribution of the facility personnel to decide if it should be included as co-author or at the acknowledgments. At the beginning of the project, the Proteomics Core provides to each user recommended guidelines for authorship on manuscripts from the Association of Biomolecular Resource Facility (ABRF), as example of policy to be use in the criteria to decide when personnel should be include as co-author and when only at the acknowledgments.

4.8.5 Non-local Resources

We do not have any remote instrument or resources used in the process of sciences.

4.8.6 Software Infrastructure

Data acquisition from instrumentation:

- MassLynx software (Waters) v4.1 with MaxEnt 1 and 3, Transformation, Biolynx and MassSeq keys.
- Xcalibur software and Thermo Foundation (Thermo Scientific)

Data processing and peak list extraction (RAW file into peak list format):

- Mascot Distiller v2.6.2.0 (Matrix Science, license)
- ProteoWizard (MS convert and See MS)
- Proteome Discoverer v2.2 (Thermo Scientific)
- ProteinLynx Global Server v3.0 (PLGS 3.0) (Waters Corp)
- MaxQuant software (Max Planck Institute of Biochemistry, Germany)
- Skyline (MacCoss lab, University of Washington)

Bioinformatics analysis (specific for Proteomics)

- ProteinLynx Global Server v3.0 (PLGS 3.0) "Expression Analysis" (Waters Corp)
- Mascot server v2.7 and Daemon toolbox with license (Matrix Science)

- Proteome Discoverer v2.2* (Thermo Scientific).
- Protein Deconvolution (Thermo Scientific).
- Compound Discoverer software (Thermo Scientific).
- The Global Proteome Machine open source (GPM) (X! Hunter and X! Hunter) (Winnipeg, CA).
- ISB/SPC Trans Proteomic Pipeline, System Biology (Seattle Proteome Center), open source
- Scaffold 5, Q+S* (www.proteomesoftware.com)
- PTM (www.proteomesoftware.com)
- ProteoIQ software* (www.premierbiosoft.com)
- Programs also used to perform protein relative quantitation using label "free" as well as isotopic label and reporter for data coming from mass spectrometry analysis
- Open source tools from Expasy (www.expasy.org), String-DB, KEGG GO and pathways, FunRich, The Reactome as well as repository and validation site such as ProteomeXchange from PRIDE.

Processes raw data into final and intermediate formats or data products.
The RAW files (Thermo Scientific and Waters) generated are converted into formats such as MGF, XML, PKL, Dat, mzML, excel, TIFF and JEPG, etc. The main data files (RAW) generated have an average of 400MB to 1.1 GB per hour. The PCF analyzes around 15-20 GB daily. The RAW file after post-processing can produce an extra peak or mass list (text standard as MGF, DTA, XML formats, etc.) of 20 to 100 MB each one. The peak list file will be used to search against specific biological databases through certain bioinformatics tools such as Mascot (license), Proteome Discoverer (license), Scaffold (license) or, The GPM X! tandem (open source) or MaxQuant (open source) search engines.

Hardware
3 Workstations:

1) HP Z820 Intel Xeon CPU E5-2650 0 @2.00GHz 2.00 GHz (2 processors), 96 GB of RAM and 2 hard drive of 10 TB
2) HP Z840 Intel Xeon CPU E5-2650 0 @2.40GHz 2.40 GHz (2 processors), 96 GB of RAM and 2 hard drive of 10 TB
3) Lenovo workstation Intel Platinum 8160T CPU @ 2.10GHz 2.10GHZ (2 processors), 192 GB of RAM, 1 hard drive of 5TB and SSD of 1TB.

Mascot Head Node

- Budget for replacement in 2022-2023
- Specs TBD

Mascot Search Node

- Replaced 2019
- Specs TBD

South Dakota Data Store (SDDS)

- Sharing Tier (Disk Based)
    - Spectralogic BP-NAS (formerly “Verde”)
    - Array of 8TB spinning disks
- Archival Tier
    - Spectralogic T380 Magnetic Tape Library
- Connected to Globus for external sharing

4.8.7 Network and Data Architecture

Present -2 years

- The LAN consists of 1 gigabit ports connecting at 10 gigs to the building core, and uplinked to the campus backbone at 10 gigs to the campus core
- The intermediate data frame (IDF) has redundant 10 gig connections to the main data frame (MDF) for the building and the MDF has redundant 10 gig connections to the campus core
- The campus core is connected at 4.25 gig to the public Internet and is also attached at 100 gig to I2.
- Combination of Cisco 3750 and 3850 switches
- Wireless is provided by Cisco 9130 access points. The 9130s are currently not able to operate at full wif if 6 capabilities until the switch infrastructure is updated
- Building level connectivity is currently on a Cisco Nexus 93180.

Next 2-5 years

- Building connected to the backbone at 40 gig with the redundant connection via Catalyst 9500 to campus network core
- LAN connections will consist of 1 gig connections to a redundantly connected 10 gig switch stack consisting of Cisco 9300 UN switches. These switches provide Multigig (Up to 5 gig over a single port) connectivity to Access points and 60 watts of power via UPOE.

Science DMZ:

USD operates a Science DMZ network to support bulk research data flows. The Science DMZ consists of a separate network enclave, isolated from the TCP congestion often associated with traditional enterprise network traffic. To support high-speed, unencumbered scientific data movement, the Science DMZ employs a Data Transfer Node (DTN) connected to the HPC cluster network and other research data hubs on the Vermillion campus. The DTN hosts USD's Globus server providing high speed data transfer and publication capabilities based on the GridFTP technology.

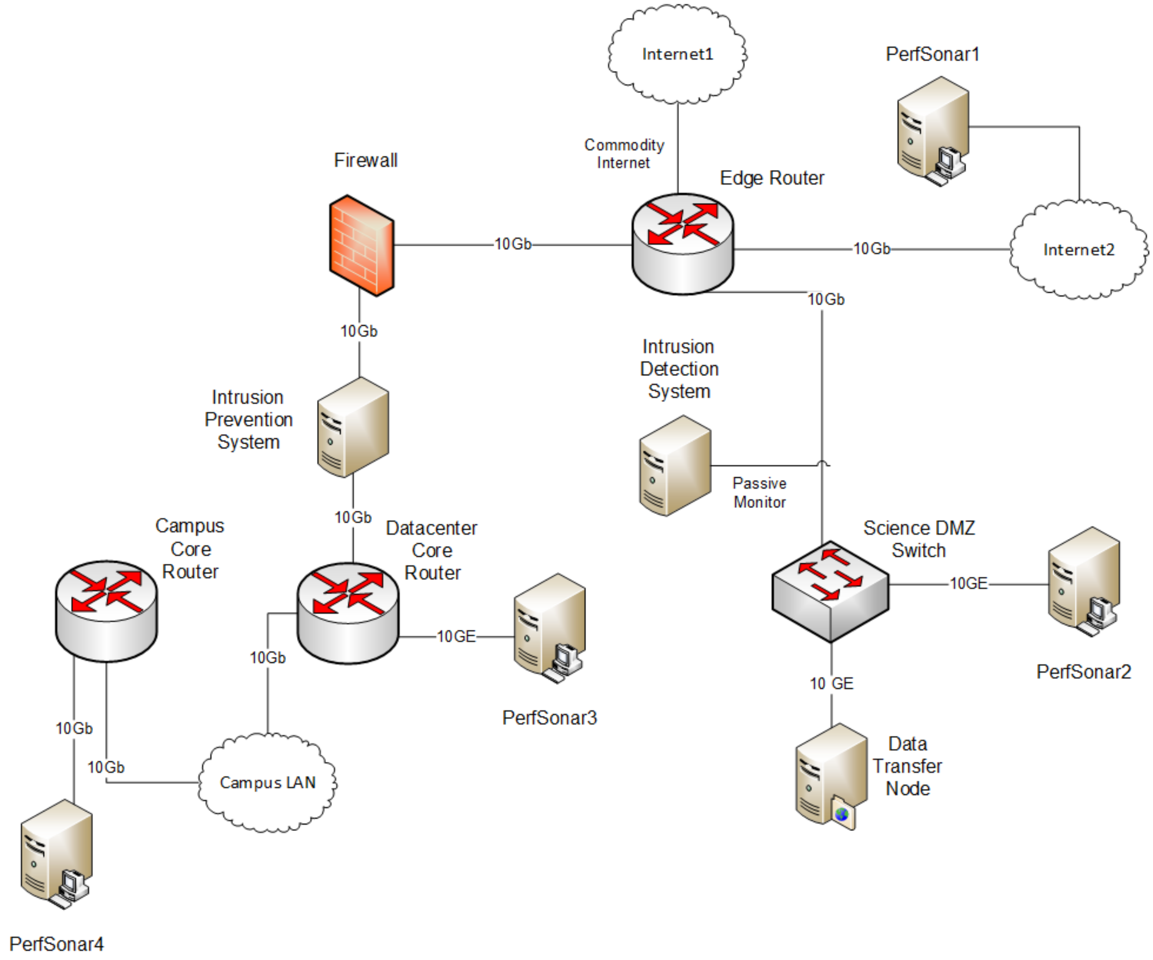


Figure 4.8.7.1 – USD Science DMZ

4.8.8 Cloud Services

Nothing currently, but will consider cloud services as applicable.

4.8.9 Known Resource Constraints

According to the USD Network policies, the users of the facility, who are outside of USD campus and/or not affiliated with USD, cannot open the links of the reports generated at the facility. The only way that PCF can share that is in static (PDF) or other formats such as Excel and Word, making it impossible for the researchers to entirely access the data.

4.8.10 Outstanding Issues

Nothing else to report.

4.9 USD: Department of Biology & Department of Sustainability

Content in this section authored by Ranjeet John, Department of Biology & Department of Sustainability at the University of South Dakota

4.9.1 Science Background

I am an assistant professor in the Department of Biology, University of South Dakota, with a joint appointment in the Department of Sustainability. My first research plan is to study land cover/use change, food-water nexus and socio-ecological systems across grassland ecosystems in Kazakhstan and Mongolia. This portion of my work was recently funded Interdependent dynamics of food, energy and water in Kazakhstan and Mongolia: Connecting LULCC to the transitional socioecological systems as Co-I, which was approved for the period 2020-2022 by the NASA Land Cover Land Use Change Program (Co-investigator). This study is proposed to examine the interconnectivity of food, energy and water systems (FEWS), as well as their interdependent dynamics under rapid changes in climate and intensified land use in Kazakhstan and Mongolia over a 40-year period (1981-2020). The research objectives are to study grassland ecosystem dynamics and extreme climate events-grazing/disturbance interactions in these post-Soviet Union transitionary economies.This project supports one graduate research assistant.

The second research area in my group focuses on land cover/use change across the Northern Great Plains. I am currently working and planning the following papers; 1) Evaluation of satellite-derived gross primary production products across managed Midwest landscapes and 2) Modeling satellite-derived evapotranspiration and primary productivity in South Dakota to track western creep of irrigated cornbelt vs eastern creep of 100 degree meridian; 3) Linking land cover transitions, demographic change and trends in biophysical properties: A case study in Eastern SD and 4) Multiscale monitoring of Invasive Yellow Sweet Clover using Harmonized Landsat/Sentinel 2, UAV and Machine Learning. My plan is to hire a second PhD through Department of Biology Teaching Assistantship funding (Spring/Fall 2021). I submitted an EPSCoR Research Fellows (RII Track---4) proposal this May. If funded, I anticipate that I will receive summer funding for one PhD/graduate student for 2 years. The new graduate student will work on land cover/use change, invasive species and ecosystem ecology-related research in South Dakota.

The data I store on SDDS are long term data records (LTDRs); mostly satellite data-derived raster files of varying cell sizes and resolutions including 250m, 500m, 1km TERRA MODIS sensor, 30m Landsat OLI/ETM+/TM sensor, 10/20m Sentinel MSI and UAV imagery for validation. In addition, I also have meteorological   data both in excel files and gridded raster files. Finally, I have socio-economic data (population and demography, urban vs rural,, occupation/herder households, livestock and type population, etc.) both in excel files and spatially explicit shapefiles .This data enables me to study Grassland degradation, semiarid and agro-ecosystems, primary productivity, evapotranspiration, disturbance, species richness, climate change, drought, UAV applications, remote sensing, land cover land use change, climate-land interaction. My study areas are grassland ecosystems of the Mongolian Plateau and

Kazakhstan, ecosystems in the Upper MidWest (for e.g. kalamazoo watershed), and grassland/agroecosystems of the Northern Great Plains.
My lab synthesizes, visualizes and analyses this disparate data, performs statistical modeling, asks research questions to address knowledge/data gaps. The satellite and meteorological data originate from Federal or university data portals (NASA Earthdata, USGS LPDAAC, Daymet, PRISM). The data is owned by either me (on SDDS) or members on my lab (local desktop). Of late my students have started working on processing geospatial data on the cloud, specifically, Google Earth Engine (GEE). We hope to create image composites from various satellite datasets, (large raster arrays), download to our server and process using HPC

4.9.2 Collaborators

Our major collaborators include USGS EROS center (4 users) in Sioux Falls, SD; USGS Northern Prairie Wildlife Research Center (1 user), Rapid City; as well as in Michigan State University (4)East Lansing , MI; University of New Hampshire (1),Durham, NH; University of Bari, Bari, Italy (2), Montana State University(2), Bozeman, MT; University of Wisconsin-Madison (1)Madison, WI, Stanford University(1), Stanford-Palo Alto,CA and Oklahoma State University (1), Stillwater, OK among others.

4.9.3 Instruments and Facilities

As described earlier we download satellite and meteorological data from several NASA, USGS, NOAA, USDA, and DOE data portals to our server. We also have several field instrumentations to measure leaf area index, soil moisture, and UAV-mounted sensors to measure surface reflectance. These field observations are used to validate the satellite derived data. Analysis takes place either on workstations or on the Cloud (google earth engine).

4.9.4 Process of Science

We hope to create monthly, seasonal, annual, image composites from various satellite datasets, (large raster arrays, ~30 GB each), download to our server and process using HPC. The process flow involves the use of R & RStudio to analyze trends in satellite data over 30 years (i.e., 30gb data layers x 30). The results on HPC along with the R code used for processing will be shared with collaborators via github.

4.9.5 Non-local Resources

We are currently trying to move our workflow involving extremely large raster data to HPC.

4.9.6 Software Infrastructure

R and RStudio at present, Python for machine learning (matplotlib, scikitlearn and numpy in the future).

4.9.7 Network and Data Architecture

USD operates a Science DMZ network to support bulk research data flows. The Science DMZ consists of a separate network enclave, isolated from the TCP congestion often associated with traditional enterprise network traffic. To support high-speed, unencumbered scientific data movement, the Science DMZ employs a Data Transfer Node (DTN) connected to the HPC cluster network and other research data hubs on the Vermillion campus. The DTN hosts USD's Globus server providing high speed data transfer and publication capabilities based on the GridFTP technology.

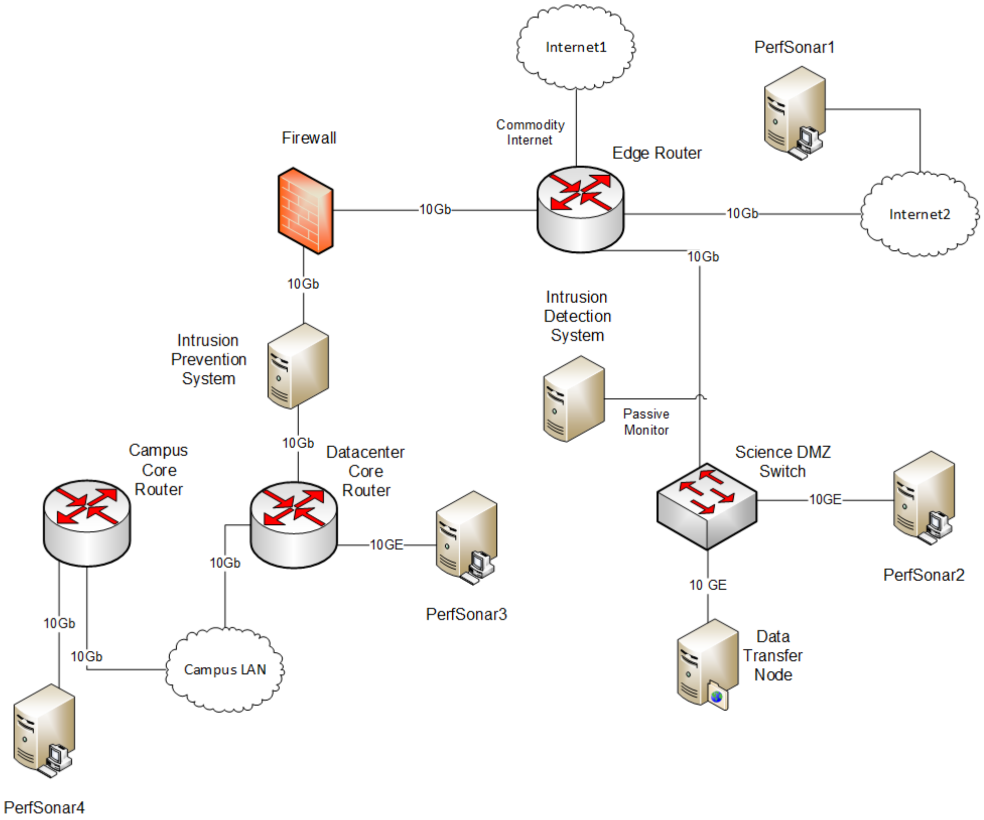


Figure 4.9.1: USD Science DMZ Architecture

4.9.8 Cloud Services

At present we are using Google Earth Engine and our personal google drives. However, we might transition to AWS in order to access the USGS Landsat archive there in the next 2-5 years.

4.9.9 Known Resource Constraints

None to report at this time.

4.9.10 Outstanding Issues

None to report at this time.

5 Discussion Summary

During discussion, with member of the South Dakota region, the following points were emphasized:

- Data volumes for a number of use cases are growing faster than available storage. While this is being considered, it is recommended that the region consider pooling resources to facilitate both scientific use cases and backup arrangements to ensure that there is an ample amount of storage resource available. This will be exacerbated in future years as more data-intensive use cases are identified.

- Computational resources within the region are available, of various varieties (HPC, HTC, Cloud, GPU). As the number of potential use cases grow, it is recommended that the region come together to define some policies and frameworks that would facilitate arrangements that enable sharing of these resources across campuses.

- The requirements to characterize and understand network performance within and outside of the region

- The requirements to support data mobility (e.g., bulk data movement, streaming) within and outside of the region

- The costs and challenges in supporting research software, and the ways it can be made easier to acquire and support.

- The need for CI "human" resources, e.g., staff that can assist with the network, hardware, and software engineering that must be present to support scientific use cases.

- A fully featured "data architecture" for the region is recommended. This is a combinational CI component that accomplishes the following:
    - Capable networking that links the facilities via a regional network.
    - Possible traffic isolation to allow for better segmentation of risk and application of security policy
    - Data mobility through fully featured Data Transfer Nodes (DTNs) and capable data movement tools (e.g., Globus)
    - Network measurement and monitoring through a shared infrastructure (e.g., perfSONAR)
    - Targeted network security policies that allow for high-performance and maximum protection
    - Ability to easily on-board users and use cases

- A working group to study the security policies surrounding a regional DMZ. These include considerations for:

  - Onboarding policy for end users and their affiliated use cases
  - AUP for use and operation of the infrastructure
  - Policy for in-region and out-region data mobility
  - Affiliate accounts and access for non-regional members
  - Considerations for CUI requirements

- As commercial clouds become more viable, campuses and REED to should ensure there are adequate resources available to reach and leverage these distributed needs.

5.1 BHSU: WestCore; The Western South Dakota Nucleic Acid Sequencing and Genotyping Core Facility

Science Summary:

The Western South Dakota Nucleic Acid Sequencing and Genotyping Core Facility (i.e., WestCore) was established in 2004 and provides research infrastructure and support in the area of nucleic acid sequencing and genotyping. WestCore provides expertise in several areas of genetic and genomic analysis, including, but not limited to, genetic, genomic and transcriptomic library development, DNA sequencing and genotyping, quantitative real-time PCR, and NextGen nucleic acid sequencing. Yearly, an average of 50 individuals from approximately fifteen (15) universities/research centers/other institutions utilize WestCore facilities and services.

The facility performs a number of services, and has numerous instruments that are capable of generating GBs of data per run and subsequent analysis data sets that may reach TB in size. In the coming years equipment upgrades will lead to increasing data outputs, along with increased volume of use. Analysis is a key component of the process of science, and can be done in different ways depending on researcher needs:

- Data sets can be made available to the researchers for external analysis (via electronic transfer, or use of removable media)
- Local software and hardware resources can be used to perform analysis activities

The latter can be a challenge to support, as the number of licenses for key software packages is limited (thus it is not possible to allow for extensive parallel analysis efforts). There are also limits to what is available with the computational and storage resources, as well as trained IT staff to assist users as they perform analysis activities.

Discussion Summary, Findings, & Actions:

- The process of science for sequencing and genotyping activities is well understood, and well supported for the WestCore use case. The linkage between the instrumentation and technology is sound (e.g., there is well defined workflow between the act of sequencing, storage, and computation).

- Limitations in the workflow are related to common areas of friction:
    - ***Data Storage***: increases in instrument precision will gradually cause storage resources to be exhausted over time. Mechanisms to improve this could involve purchasing more storage, but also revisiting data retention policies (e.g., reducing time allowed) or integrating non-local (either regional or cloud based) resources for backups.
    - ***Data Sharing:*** Data sharing through well-established tools (e.g., the Modern Research Data Portal based on Globus) is recommended as

    the primary mechanism to expose data to researchers using the facility.
  - ***Software:*** Specialized analysis software, in the absence of an open source alternative, will be required for the process of science. Figuring out mechanisms to share and manage license resources on the analysis framework is a top priority, along with exploring other software options.
  - ***Computational Abilities***: Allowing and encouraging more use of scalable grid resources, instead of use of personal machines, could assist with the aforementioned software scalability issues
  - ***CI Support Staff***: Lack of dedicated IT support for core functions (e.g., analysis) will impact the ability to deliver efficient use of computational resources.

Possible Next Steps:

- Documentation of use case for regional CC* grant

- Leveraging regional CI support to streamline/improve the workflow so that it is more portable

- Experimentation with using more local resources for storage and computation, coupled to the instrument workflow

5.2 DSU: Campus Technology Profile

Science Summary:

Dakota State University offers the MadLabs Research Environment & Network (MADREN) - an extensive technology infrastructure dedicated to cyber security research. This multi-node cluster allows for the creation of virtual environments that can be used for research and education purposes. Experiments and lessons can be designed once, and deployed/reused offering a way to scale educational opportunities and personalize experience. There are a number of use cases that span computing sciences, engineering, and other educational and research activities.

Upgrades are expected in the coming years to keep pace with technology (e.g., network capacity upgrades), and add more functionality (e.g. DTNs, perfSONAR, data movement tools, InCommon). It is also expected that AI/ML use cases will grow, implying a need to install new functionalities to support. As in other locations in the region, CI support staff are in short supply. Having dedicated and knowledgeable resources to assist with research workflows would be valuable. Data and infrastructure security also add a layer of friction to use cases at times. Reviewing and evaluating the risks and mitigations would lead to better outcomes.

Discussion Summary, Findings, & Actions:

- DSU has a rich set of CI technologies available for the research community, including computation and storage, and is working to add more data mobility.

- There are numerous use cases utilizing this now, and more are expected in the future.

- Upgrades are planned to keep pace with research community needs.

- DSU is requesting assistance from EPOC to understand data mobility performance in/out of campus via new perfSONAR nodes. This, coupled with an architectural review, will assist DSU as they plot upgrade paths.

Possible Next Steps:

- EPOC Architectural review of DSU environment

- Participation in the DME

- Documentation of resources that could be shared for regional CC* grant

- Portal software to disseminate research results should be investigated as collaboration space increases

- Review security and data policies for the research environment

5.3 DSU: South Dakota Center for SMART Power Systems

Science Summary:

The South Dakota Center for SMART (Secure Machine Learning and AI for ResilienT) Power Systems is a proposed project that would allow for R&D activities surrounding the next generation of energy generation and operation. Emphasis on emerging technologies such as AI/ML, cybersecurity, and advanced communication and control are proposed.

A significant data generation task will be the creation of simulations: both those that are downloaded and shared from other resources, and those that may be created locally and shared to collaborators. Simulations could be TB sized in the coming years, and will require storage and data mobility resources available to support the use case.

CI assistance, particular in developing the operating environment for some of the technologies, is requested. It is expected that the DSU MADREN can be used for some aspects of this work, in addition to possibly other regional and national frameworks.

Discussion Summary, Findings, & Actions:

- The DSU SMART center use case, if funded, would require the ability to handle ingress and egress of significant simulation data sets

- SMART operation will not generate much data, but has the potential to involve a large population of users and will require protections during operation

- A knowledge CI resource to assist with the creation of workflow and supporting technology will be required.

Possible Next Steps:

- If funded, DSU should consider setting up a data mobility platform (e.g. DTNs, Portal, storage) to handle the ingress and egress of simulation data.

- Documentation of use case for regional CC* grant

- Review security and data policies regarding data sharing.

- Integration with other regional compute and storage integrations, to better scale processing capabilities.

5.4 NSU: Connecting the Social Sciences across the Great Plains

Science Summary:

The Great Plains Sociological Association (GPSA) is a regional organization consisting of members from the Great Plains: it provides a mechanism for social scientists to discuss and share their scholarship. As a part of this, the GPSA sponsors an annual meeting and journal titled the "Great Plains Sociologist". It is expected that at least 15 institutions and 450 faculty/staff could participate in this effort.

NSU, as host institution, will attempt to build a technology platform to further this effort. The technology platform should have mechanisms to:

- Store and publish research results
- Facilitate virtual discussions (e.g. video, chat)
- Host virtual events
- Protect intellectual property
- Be available to participants from around the region and world

Implementation details are flexible; there is a strong desire to host this on premises, although use of cloud platforms is also possible provided that control of the platform and content is still allowed to NSU operators.

Discussion Summary, Findings, & Actions:

- The use of online platforms that offer a number of related virtual functions (e.g., literature dissemination, social interaction, host of synchronous online events) continues to grow due to the COVID-19 pandemic, and a general maturation of the technology and use cases.

- There are many commercial offerings in this space that are more fully featured, and affordable, than attempting to leverage free and self-operated tools. E.g., locally running an H.323 capable remote A/V platform to facilitate online communication is nearly extinct versus the ability to purchase services through platforms like Zoom, WebEx, Bluejeans, or Teams. Particularly as the commercial platforms integrate chat, ability to run on multiple platforms, and offer privacy and security considerations well beyond what a local staff can deliver.

- Adopting a commercial platform is recommended, provided that factors such as cost, user population, and ownership of data can be understood and accepted.

Possible Next Steps:

- Evaluate platforms, leveraging regional experts to evaluate the technology

- Working with REED to ensure that there is adequate network support to reach a cloud-based solution.

5.5 NSU: Toward A Greater Dissemination of Social Science Data and Information for Better Civic Engagement

Science Summary:

The Center for Public History and Civic Engagement collects, analyzes, transfers, shares and stores data related to local and regional history and civic engagement. The center utilizes a number of electronic tools (scanners, digitizers) to convert printed media into electronic versions. Some are located on premises; others can be operated remotely from locations that are not able to loan delicate materials.

In the coming years, the center will be making substantial efforts to grow the collection of digital resources - this includes converting printed to digital, but also offering video and audio from oral history projects. Delivering on these goals will require a number of technical capabilities:

- Creating workflows to bind instrumentation to storage and processing infrastructure. Both for centrally located machines, as well as those that may travel
- Creation or use of a repository (storage, web platform for search, categorization and delivery) for content
- Availability of CI staff to assist with the aforementioned tasks

Use of cloud resources, to better scale delivery to users, can also be considered as a way to augment local capabilities.

There are still open questions to this work:

- Data set sizes are unknown at this time, and should be investigated further
- Access methods and protections are also unknown, and should be considered

Discussion Summary, Findings, & Actions:

- The growth of capturing physical history and converting it to digital has increased steadily over the years.

- The subsequent speed with which one can capture, as well as the resulting data sizes for scanned relics, has also increased.

- CPHCE should define a set of workflows for the various use cases, to better integrate technology:
    - Local operations could consist of integrating the instruments that are used to scan with local storage and processing capabilities, along with a conduit to data mobility tools (either a platform that can share results, or tools that can be used to transfer to cloud resources)
    - Remote operations (e.g. visiting off-site locations with an instrument such as a scanner, or A/V equipment) should involve a well-equipped portal computation unit that integrates to the instrument, and has both ample local storage and a way to transmit results to the aforementioned central processing and storage facility

- CPHCE requires CI expertise to create workflow between the hardware and software, along with creating the local or remote processing and storage resources

Possible Next Steps:

- Documentation of use case for regional CC* grant

- Leveraging regional CI support to create and test the workflow (potentially via a pilot) so that it can run locally, using remote resources, or potentially in a cloud environment

- Creation or adoption of a platform to disseminate the results (either using local, regional, or cloud-based resources)

5.6 SD Mines: Realtime Weather Regional-Scale Weather Forecasting

Science Summary:

Researchers at SDSMT utilize the NCAR Weather Research and Forecast Model (WRF) to create weather forecasts 2-4 times per day, depending on the availability of local computing and storage resources. Forecast accuracy depends on the data spacing: 9km and 3km spaced data is used now, there is desire to move to more accurate 1km spaced data. The subsequent weather products can be provided to the NSF Big Weather Project, as well as local resources that find a current and accurate forecast useful.

The research activity is currently gated on several factors:

- Computational availability - processing the data takes several hours on current machines, which limits how many products can be created. Faster and more numerous resources would lead to more accurate and numerous forecasts.
- Storage resource availability - forecasts use a set of sensor data, and more accurate forecasts require more fine-grained (e.g., larger) data sets.
- Ability to download larger data sets via network connections.
- CI staff to help manage the workflow and adapt to new approaches (e.g., containerization for key parts of workflow, migration to other computational resources).

Advancing this science requires several factors be addressed:

- ***Computational:*** Current computing capabilities require several hours for a single run. By either upgrading the local capabilities, or porting to and utilizing other regional resources, it would be possible to run existing workflow faster, and more often. Possibilities to run more complex workflows also exist.
- ***Storage:*** Current storage limits the amount of data that can be downloaded and retained. More local storage would allow a greater history of observations into a computational run, as well as facilitate use of more precise (e.g. larger) data sets.
- ***Workflow:*** The current workflow is closely tied to the local resources. Porting and upgrading the approach to use new hardware, and remote locations, would allow for the use of other resources and the ability to scale to new releases of the data and codebase.

Discussion Summary, Findings, & Actions:

- The process of science is well defined and understood. This can be attributed to using community based tools that are widely adopted and supported.

- Weather data continues to become more precise and available, meaning the upstream sources of experimental data are strong and growing.

- The ability to utilize experimental data depends on availability of storage and computational resources to run the well-understood codes of the workflow.

- Current computational and storage capabilities are lacking to be able to advance the scientific output. Without upgrades to the technology, it will not be possible to produce research products that are more accurate, or keep up with timeliness demands.

- Hardware upgrades through a grant are possible in the coming years.

- A parallel path to pursue involves leveraging community resources in a number of ways:
    - Working with CI experts from the region to containerize the workflow. This would allow the general process of science (e.g., download data, run analysis, post results) the ability to run on both the local machines, as well as other regional resources that have spare cycles.
    - Once the workflow is containerized, it will be possible to upgrade and modernize the approach with the latest analysis code
    - Running more analysis, with possible greater resolutions of data, is possible once additional computational and storage resources become available

Possible Next Steps:

- Pilot effort within the region to containerize the current workflow

- Pilot effort within the region to run the newly containerized workflow on shared computational resources

- Documentation of use case for regional CC* grant

- Experimentation with more precise weather data, more frequently, using regional resources.

5.7 SDSU: Research Cyberinfrastructure Center

Science Summary:

South Dakota State University is the state's largest university, and offers several CI-related technologies to facilitate research. "Roaring Thunder" (RT) is a 56-node general compute cluster, and GPFS storage, that is available for research purposes - a primary use case to support Bioscience use cases such as those profiled in the other case studies. The cluster has the ability to integrate with the campus Science DMZ to facilitate data mobility (and uses several tools including Globus). The ability to integrate with certain cloud storage components is available, although this is not widely used by the research community at this time. SDSU will be making upgrades to core components in the coming years: increasing the storage, adding network measurement through perfSONAR, and upgrading network capabilities to 100Gbps.

This setup still does experience some areas of friction, in particular traversing firewall infrastructure in some cases. SDSU will be seeing help in understanding how to mitigate this in the future as the usage increases. This, coupled with discussion regarding the use of Jumbo frames, and machine tuning, can help to increase the performance profile of some parts of the data movement infrastructure.

Discussion Summary, Findings, & Actions:

- SDSU has a rich set of CI technologies available for the research community, including computation, storage, and data mobility.

- There are numerous use cases utilizing this now, and more are expected in the future.

- Upgrades are planned to keep pace with research community needs.

- SDSU is requesting assistance from EPOC to understand data mobility performance in/out of campus, particularly on a path that includes a firewall. This, coupled with an architectural review, will assist SDSU as they plot upgrade paths.

Possible Next Steps:

- EPOC Architectural review of SDSU campus and computing

- Participation in the DME

- Documentation of resources that could be shared for regional CC* grant

- Portal software to disseminate research results should be investigated as collaboration space increases

5.8 USD: South Dakota Biomedical Research Infrastructure Network (SD BRIN) Proteomics Core Facility

Science Summary:

The South Dakota Biomedical Research Infrastructure Network (SD BRIN) Proteomics Core Facility provides protein identification, characterization, and quantitation services to researchers within the region and around the country. Datasets are generated from the Liquid Chromatography- Mass Spectrometry instrumentation during the proteins and peptides analysis process. The Facility since its inception, has analyzed more than 20,000 samples corresponding to 130 research projects from approximately 110 users.

The facility has a number of high resolution instruments on site that generate data sets that may range in size from 100s of MBs to 10s of GBs for a single run. The facility is capable of running constantly to process samples.

The facilities instruments utilize storage and processing provided by the University of South Dakota, which has sufficient capacity to support current operations. There are also integrations with cloud storage, and data mobility tools, to facilitate sharing information if a researcher desires in formats they can utilize. There are areas of friction in data sharing however: external users (to USD) cannot access data retained at the facility in a straightforward manner. The only way that PCF can share this data is via static (e.g., PDF) mechanisms.

Equipment upgrades in the coming years will produce more precise (e.g., larger) data sets, and most likely do so faster than in prior generations. Computing and storage needs must grow along with these capabilities.

Discussion Summary, Findings, & Actions:

- The PCF has established clear and efficient workflows to go from samples to scientific data. The ability to leverage computation and storage from USD makes this an effective research use case.

- The ability to share data outside of the USD should be viewed as a strong requirement, and the policy to prevent external access should be reviewed. Using tools like Globus or the Modern Research Data Portal, especially for users in the region or around the country with proper credentials, should be sufficient to gain access to the raw data sets.

- Evaluating the workflow to ensure that all instruments are using common shared storage, and the storage to computation pipeline, will ensure efficient operation now and into the future as machines are upgraded.

- Use cases that facilitate remote processing (e.g., that may involve streaming of data from one location to another) should be explored as a way to scale computational power.

- Remote-control/viewing of data on instruments is desirable, and may factor into future usage patterns.

Possible Next Steps:

- Documentation of use case for regional CC* grant
- Review security and data policies regarding data sharing.
- Portal software to disseminate research results should be investigated as collaboration space increases
- Integration of other regional compute and storage integrations, to better scale processing capabilities.

5.9 USD: Department of Biology & Department of Sustainability

Science Summary:

Research conducted in the Department of Biology and the Department of Sustainability involves studying land coverage/use change, food-water nexus, and socio-ecological systems across grassland ecosystems. The research objectives are to study grassland ecosystem dynamics and extreme climate events-grazing/disturbance interactions, and how they relate to the economics of the target regions.

Data used for this research varies:

- Long term data records (LTDRs) of satellite data-derived raster images files of varying cell sizes and resolutions (250m, 500m, 1km TERRA MODIS sensor, 30m Landsat OLI/ETM+/TM sensor, 10/20m Sentinel MSI)
- UAV imagery
- Meteorological data (flat files, as well as gridded raster images)
- Socio-economic data (population and demography, urban vs rural, occupation/herder households, livestock and type population, etc.) as flat files and spatially explicit shapefiles

The satellite and meteorological data originate from Federal or university data portals (NASA Earthdata, USGS LPDAAC, Daymet, PRISM), and socio-economic data comes from other federal and international resources such as the US state department, USDA, and the UN. Data set sizes vary depending on type and resolution - 10s of GBs for some satellite images is possible, with the size growing as the number of layers and precision of data increases.

Researchers synthesize, visualize, and perform analyses on this disparate data using statistical modeling to answer research questions and address knowledge/data gaps. HPC analysis and storage are a critical part of this research workflow, along with software tools (e.g., R, R Studio, Matlab, Python based AI/ML tools). Some of this can be done locally, other use cases can leverage cloud resources.

Discussion Summary, Findings, & Actions:

- Research involves acquiring data from a variety of sources on a regular basis (although there is no need to automate this step).

- Upstream data repositories have a number of different ways to share data, thus speed is highly dependent on their technology platforms and approaches.

- Analysis is performed mostly through interactive sessions (e.g., R Studio) to analyze results. Given this requirement, batch processing cannot typically be used.
- Interactivity implies a low latency/high response time when using the analysis tools.

- Storage for data sets, and produced data products, is critical.

- Disseminating data to research collaborators could leverage data mobility tools (e.g., Modern Research Data Portal, Globus).

- Use of cloud in the future is possible, provided that the tools facilitate the research use case.

Possible Next Steps:

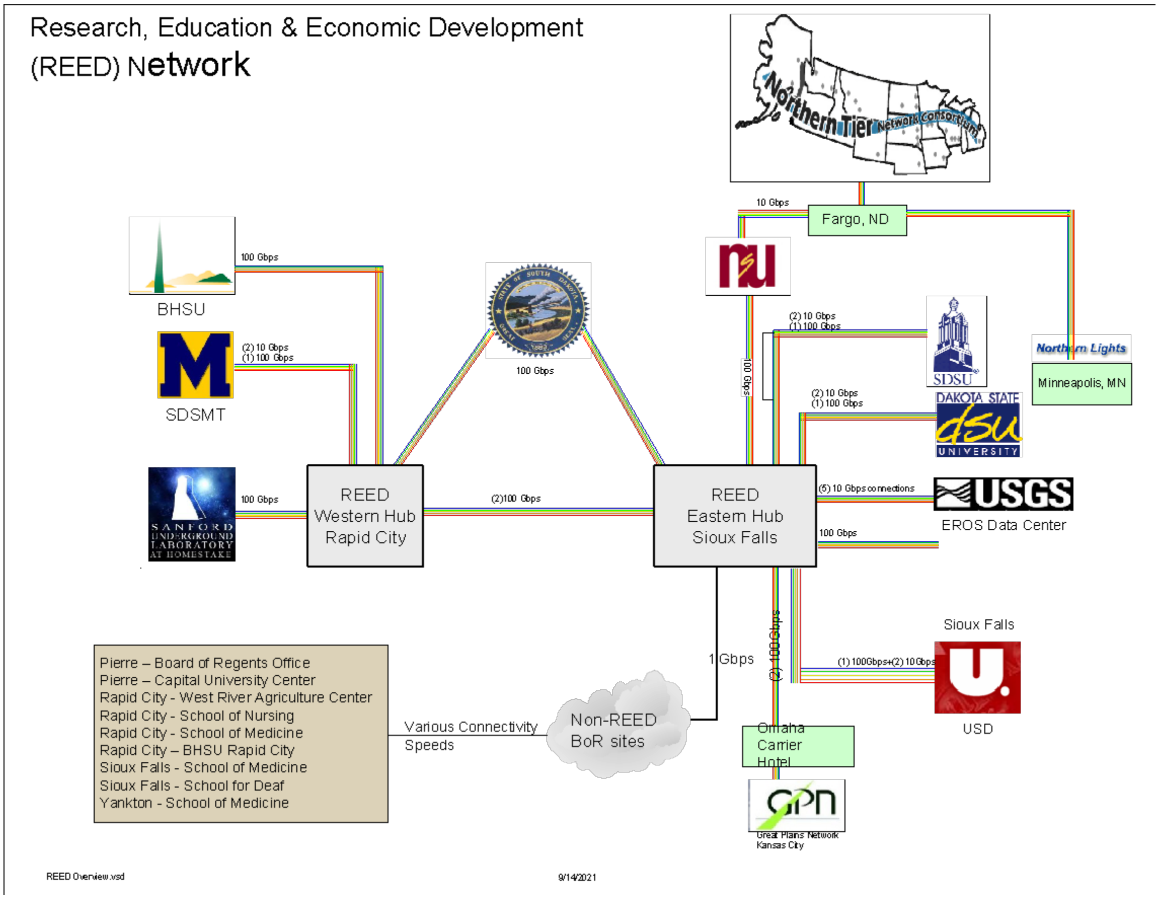
- Documentation of use case for regional CC* grant

- Use of regional computational resources is possible, provided that the primary analysis tool (R Studio) can be run remotely and not suffer any latency-related complications.

- Portal software to disseminate research results should be investigated as collaboration space increases

- Understanding cloud use cases is important, and how the network latency, computational power, and storage relates to what can be done locally or within the region.

Appendix A - Regional Networking Diagram



Appendix B - Research Cyberinfrastructure at The University of South Dakota

The Lawrence Supercomputer was acquired through a combination of state and federal funding: a FY16 SD Board of Regents Research and Development Innovation award, and National Science Foundation Major Research Instrumentation award OAC-1626516. Lawrence runs the CentOS 7 Linux operating system and is made up of over 2,300 CPU cores, including systems with 1.5TB of memory, Nvidia P100 GPU accelerators, and over 400TB of ZFS network storage accessible via a 56Gb FDR Infiniband network. Lawrence has an estimated performance of over 100TFLOPS.[29]

USD operates The South Dakota Data Store ("SDDS," funded by NSF award ACI-1659282), housed at the USD Main Campus data center in Vermillion, SD and the USD Community College for Sioux Falls data center in Sioux Falls, SD. SDDS is accessible via the cloud-based Globus data management platform (globus.org). SDDS includes a high-capacity disk-based Sharing Tier for data sharing as well as an Archival Tier hosted on magnetic tape. The Sharing Tier has a current capacity of over 400 TB (expandable to 1 PB) and the Archival Tier's current capacity is over 1 PB (expandable to 4.5 PB). SDDS also directly supports data curation and publication through Globus, including popular metadata standards like the Dublin Core.

USD operates a Science DMZ network to support bulk research data flows. The Science DMZ consists of a separate network enclave, isolated from the TCP congestion often associated with traditional enterprise network traffic. To support high-speed, unencumbered scientific data movement, the Science DMZ employs a Data Transfer Node (DTN) connected to the HPC cluster network and other research data hubs on the Vermillion campus. The DTN hosts USD's Globus server providing high speed data transfer and publication capabilities based on the GridFTP technology.

In addition to local HPC resources, to address the computational scale-out needs of research faculty, USD provides support for faculty requiring access to national HPC and cloud resources such as XSEDE, Globus, and Amazon Cloud.

All equipment is hosted in environmentally controlled, physically secured data centers. The data centers provide in-rack and ceiling cooling units as well as dedicated fire suppression systems. All equipment is protected by generator-backed uninterruptible power supplies.

[29] https://usdrcg.gitbook.io/docs/lawrence-hpc/about-lawrence