San Diego State University Requirements Analysis Report

August 8th, 2022

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.

San Diego State University Requirements Analysis Report

August 8th, 2022

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-2001474[1].

[1] https://escholarship.org/uc/item/82d7b7b5

Participants & Contributors

Matt Brown, San Diego State University
Michael Farley, San Diego State University
Brian Lenz, San Diego State University
Christopher Leong, San Diego State University
Jerry Sheehan, San Diego State University
Jennifer Schopf, Indiana University
Doug Southworth, Indiana University
Faramarz Valafar, San Diego State University
Jason Zurawski, ESnet

Report Editors

Doug Southworth, Indiana University: dojosout@iu.edu
Jason Zurawski, ESnet: zurawski@es.net

Contents

1 Executive Summary

Deep Dive Review Purpose and Process

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

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

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

This Review

Between December 2021 and April 2022 staff members from the Engagement and Performance Operations Center (EPOC) met with researchers and staff from San Diego State University (SDSU) the purpose of a Deep Dive into scientific and research drivers. The goal of this activity was to help characterize the requirements for a number of campus use cases, and to enable cyberinfrastructure support staff to better understand the needs of the researchers within the community.

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

- School of Public Health, Laboratory for Pathogenesis of Clinical Drug Resistance and Persistence
- The College of Health and Human Services Information Technology
- College of Sciences Technology Services
- San Diego State University Research Technology

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 San Diego State University has in supporting a cross-section of established and emerging research use cases. Each case study mentioned unique challenges which were summarized into common needs.

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

- SDSU Research Technology should consider the role, and use, of secure computing environments to support sensitive data use cases.

- SDSU Research Technology should pursue the creation of a local area networking environment that facilitates fast data migration between instruments located on campus, and centralized storage and computing.

- SDSU Research Technology should investigate an approach to campus storage that integrates different layers of performance and capacity, while supporting intra-campus transfer between instruments as well as inter-campus use cases when sharing with collaborators.

- SDSU Research Technology should investigate a unification of wide area network connections, and the strategies in which they are delivered.

- SDSU Research Technology should consider consolidation of computation and storage options that exist on campus, and in clouds, to a well-defined set that can be easily supported.

- SDSU Research Technology should investigate a data-backup strategy with CENIC and member schools to facilitate DR possibilities within the state.

- SDSU Research Technology should create a dedicated role within the team to engage with researchers on a regular basis, adopting the EPOC approach to finding requirements and creating technical solutions.

- SDSU Research Technology should create a faculty council to help advise on Science DMZ strategy and policy, to ensure researcher buy-in.

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 from the San Diego State University Deep Dive that summarize important information gathered during the discussions surrounding case studies, and possible ways that could improve the CI support posture for the campus as well as recommendations for future activities:

- SDSU Research Technology should pursue the creation of a secure computing enclave to support sensitive data use cases. The location of this enclave could be on-site to SDSU, hosted by a partner (private or public), or could be a hybrid approach. It is recommended that this environment consider the following technological and policy specific aspects:
    - A computation and storage environment that meets current NIST guidelines
    - Adoption of a data transportation medium (e.g., similar to the Globus Sensitive data (HIPPA) capabilities) capabilities on dedicated Data Transfer Node (DTN) resource to support data ingress and egress
    - The creation of a new Science DMZ to specifically support this use case, potentially using a dedicated CENIC VLAN resource to isolate traffic
    - Creation of specific VRF/peering arrangements with collaborators to further address data and information security concerns. This may involve collaborators that are located in California (e.g., UCSD, SDSC, UCLA), as well as those located around the country (e.g., BYU, University of Minnesota) or across the world in other countries.

- SDSU Research Technology should pursue the creation of a local area networking environment (e.g., Science DMZ 'LAN') that facilitates fast data migration between instruments located on campus, and centralized storage and computing. It is recommended that this environment consider the following technological and policy specific aspects:
    - Connectivity between the SDSU Alvarado Research Park, and the SDSU Data Center
    - Creation of secured VLANs, that feature dedicated 10G/100G switching
    - Creating a policy that allows for the use of specific machines only (e.g., pre-ordained workstations and data movement tools that facilitate access to the computing cluster and storage)
    - Implementation of network QoS for latency concerns
    - The ability to "proxy" all other internet services on the outside world, which would negate having to manage "dual" networking configurations, and exposing services to unmitigated risk.
    - Regular audits for compliance and mitigation of existing and new risk profiles.

- SDSU Research Technology should investigate an approach to campus storage that integrates different layers of performance and capacity, while supporting intra-campus transfer between instruments as well as inter-campus use cases when sharing with collaborators. It is recommended that this environment consider the following technological and policy specific aspects:
    - A "pyramid" design where the smallest and fastest storage (e.g., DTNs with limited fast SSD capabilities) feed into an active storage area network (e.g., commercial storage built on top of parallel filesystem technologies), and back ended by slower archival options (e.g., tape libraries).
    - The DTN hardware must run a number of data transfer tools (e.g., Globus) to facilitate sharing
    - The entire environment must be mountable by instruments as well as data transfer hardware
    - A policy that facilitates allocations for instruments and people
    - Mechanisms to integrate cloud resources not as a replacement, but as an augmentation, to existing on-campus storage.

- SDSU Research Technology should investigate a unification of wide area network connections, and the strategies in which they are delivered. It is recommended that this environment consider the following technological and policy specific aspects:
    - Management of all CENIC connections to a single policy and control entity
    - Splitting primary and backup connections amongst different physical devices
    - Creating a Science DMZ WAN via CENIC that can be accessed from well-defined locations on the Science DMZ LAN
    - Creating access policy that enables onboarding to Science DMZ services
    - Routing policy that prioritizes CENIC connections over commodity

- SDSU Research Technology should consider consolidation of computation and storage options that exist on campus, and in clouds, to a well-defined set that can be easily supported. It is recommended that this environment consider the following technological and policy specific aspects:
    - On campus clusters and storage could be consolidated to a single facility. This may be run as a "condo model" that is shared and can be allocated based on percentages, or a more bespoke "concierge" model that is private, but may have higher monetary and human costs to manage for individual groups. In either case, the models should be articulated to current and new faculty and staff so they are aware of the purpose, costs, and operational styles.

- Cloud computing options (e.g., R&E models like chameleon, or commercial models like Azure or AWS) should be consolidated to a set number of possibilities. Each of these should be brokered by a central entity, and documentation for use, development, and cost models should be explained to possible users.
    - Leveraging partnerships with existing R&E inroads to cloud providers (e.g., Internet2, CENIC) to facilitate peering and contracts should be investigated.
    - A team managed by SDSU Research Technology should be created to create a set of workflow elements (e.g., containers, BCPs) that can be used researchers: either on local resources or in the cloud.

- SDSU Research Technology should investigate a data-backup strategy with CENIC and member schools to facilitate DR possibilities within the state.

- SDSU Research Technology should create a dedicated role within the team to engage with researchers on a regular basis, adopting the EPOC approach to finding requirements and creating technical solutions. Possible questions to investigate:
    - The available options for computation and storage
    - How the workflows can be integrated to campus or cloud resources
    - The costs and benefits to centralized hosting of resources, versus running independently.

- SDSU Research Technology should create a faculty council to help advise on Science DMZ strategy and policy, to ensure researcher buy-in.

3 Process Overview and Summary

3.1 Campus-Wide Deep Dive Background

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

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

EPOC supports five main activities

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

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

The Deep Dive process is based on an almost 15-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 process tries to answer essential questions about the following aspects of a workflow:

- ***Research & Scientific 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: Local & Non-Local***—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 or use at partner facilities.
- ***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.
- ***Computation & Storage Infrastructure: Local & Non-Local***—The infrastructure that is used to support analysis of research workflow needs: this may be local storage and computation, it may be private, it may be shared, or it may be public (commercial or non—commercial).
- ***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.
- ***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***—Listing of any additional problems, questions, concerns, or comments not addressed in the aforementioned sections.

At a physical or virtual meeting, this documentation 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 San Diego State University Deep Dive Background

Between December 2021 and April 2022, EPOC organized a Deep Dive in collaboration with SDSU to characterize the requirements for several key science drivers. The representatives from each use case were asked to communicate and document their requirements in a case-study format. These included:

- School of Public Health, Laboratory for Pathogenesis of Clinical Drug Resistance and Persistence
- College of Health and Human Services Information Technology
- College of Sciences Technology Services
- San Diego State University Research Technology

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.

San Diego State University (SDSU) SDSU is the oldest higher education institution in San Diego. Our community is fully committed to excellent teaching, meaningful research and service to our regional community and others we serve throughout the state, across the nation and internally. Since its founding in 1897, the university has grown to become a leading public research university, and a federally-designated Hispanic-serving Institution. Each year, SDSU provides more than 36,000 students with the opportunity to participate in an academic curriculum distinguished by direct contact with faculty and an international emphasis that prepares them for a global future.

4 San Diego State University Case Studies

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

- School of Public Health, Laboratory for Pathogenesis of Clinical Drug Resistance and Persistence
- College of Health and Human Services Information Technology
- College of Sciences Technology Services
- San Diego State University Research Technology

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

4.1 San Diego State University School of Public Health and the Laboratory for Pathogenesis of Clinical Drug Resistance and Persistence

Content in this section authored by Dr. Faramarz Valafar, SDSU School of Public Health and the Laboratory for Pathogenesis of Clinical Drug Resistance and Persistence. Additional content provided by Brian Lenz, SDSU College of Health and Human Services Information Technology and Christopher Leong, SDSU College of Sciences Technology Services.

4.1.1 Use Case Summary

Dr. Faramarz Valafar, SDSU School of Public Health and the Laboratory for Pathogenesis of Clinical Drug Resistance and Persistence, studies antibiotic resistance in bacteria, using genomics, epigenetics, phylogenomics, and systems biology. The research uses existing knowledge to understand the process of emerging antibiotics, and uses the knowledge in developing novel diagnostics and prognostic devices. The primary pathogen of study is *M. tuberculosis*. In this process, researchers have also developed novel sensitive sequencing techniques to directly sequence a lesion or tumor without culturing or amplification.

4.1.2 Collaboration Space

There are extensive collaborators for this research that include:

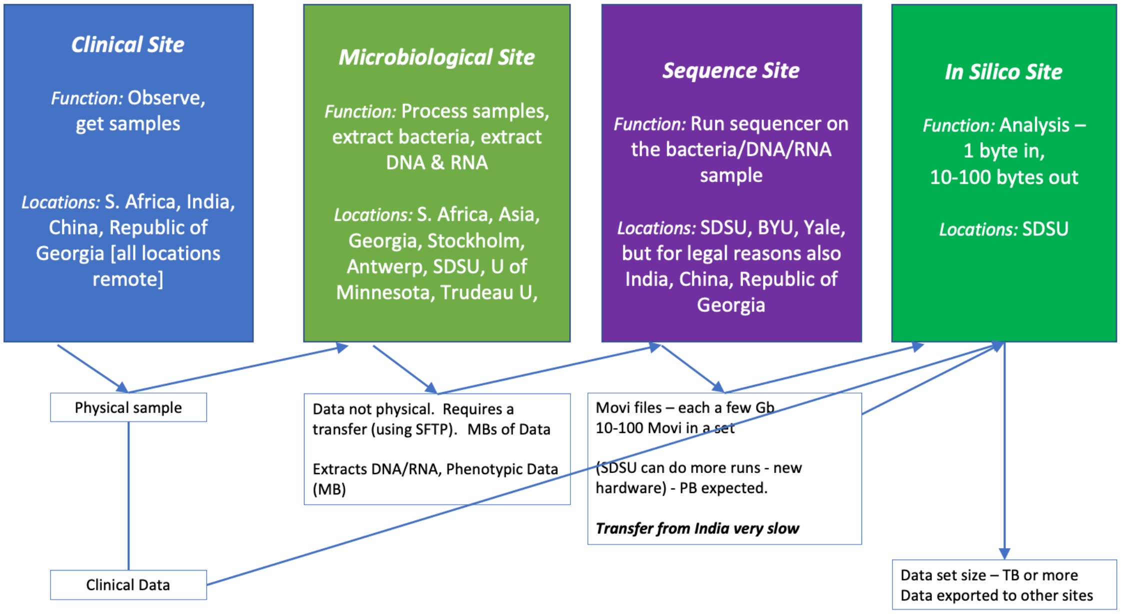
- The Public Health Agency of Sweden
- University of Stellenbosch in South Africa
- Republican Center for Tuberculosis and Pulmonary Disease in Minsk, Belarus
- The WHO Supranational TB Reference Laboratory in Antwerp, Belgium
- The WHO Supranational TB Reference Laboratory in Netherlands
- Hinduja Hospital in Mumbai, India
- University of Minnesota
- Albert Einstein School of Medicine
- UCLA School of Medicine
- UCSD School of Medicine
- BYU Genomic Center

4.1.3 Instruments & Facilities

The research uses BSL2 (Bio-Safety Level 2) and BSL3 (Bio-Safety Level 3) facilities across the collaboration sites to identify isolates of interest. There is also research activity performed with the sequencing center at BYU, or sequence data generated by other SDSU facilities.

4.1.4 Data Narrative

Some laboratory work is performed overseas, or locally in San Diego. This work can include culturing, DNA extraction, and engineering laboratory mutants. The extracted DNA and RNA is then shipped to BYU, or to facilities at SDSU, for sequencing. Sequencing data from BYU and SDSU are then stored at an SDSU computing cluster where it undergoes significant amount of in silico analysis and secondary/tertiary data generation. In silico analysis included de novo and reference-based genome assembly, variant calling, including structural variations, DNA methylomic analysis including methylomic variation detection, genome annotation, phenotypic consequence prediction for all detected variants, building   regulatory and metabolic models for tertiary prediction experiments. These steps are described in Figure 1.



Clinical Site
Function: Observe, get samples
Locations: S. Africa, India, China, Republic of Georgia [all locations remote]
Microbiological Site
Function: Process samples, extract bacteria, extract DNA & RNA
Locations: S. Africa, Asia, Georgia, Stockholm, Antwerp, SDSU, U of Minnesota, Trudeau U,
Sequence Site
Function: Run sequencer on the bacteria/DNA/RNA sample
Locations: SDSU, BYU, Yale, but for legal reasons also India, China, Republic of Georgia
In Silico Site
Function: Analysis – 1 byte in, 10-100 bytes out
Locations: SDSU
Physical sample
Data not physical. Requires a transfer (using SFTP). MBs of Data
Extracts DNA/RNA, Phenotypic Data (MB)
Movi files – each a few Gb
10-100 Movi in a set
(SDSU can do more runs - new hardware) - PB expected.
Transfer from India very slow
Clinical Data
Data set size – TB or more
Data exported to other sites

Figure 1: Scientific Workflow

4.1.4.1 Data Volume & Frequency Analysis

Data volume can be characterized as being in the Terabyte (TB) range currently, with raw data being produced on a daily basis during times of research. Processed data can be reduced to Gigabyte (GB) size.

Specifically, an entire sequencing run of multiple results is TB scale where this could be between 10s and 100s of TBs on a "per patient" or ("per isolate") could be 10-50GB each.

Sequencing runs occur between 4-7 times a year (when not operating in pandemic mode), and produce 100GB of raw data for each isolate. After analysis, 10-100x of the raw size is reduced to the summary data.

4.1.4.2 Data Sensitivity

Certain aspects of the research workflow could include sensitive data: namely Personally Identifiable Information (PII) associated with some of the sequencing data used for medical research.

Specifically, no patient specific data exists now as the research is focusing on virus/bacteria results only. In the future it may be a factor if some of the things are funded.

4.1.4.3 Future Data Volume & Frequency Analysis

Longer term data volumes are expected to be in the Petabyte (PB) range, with raw data being produced on a daily basis during times of research. Processed data can be reduced to Terabyte (TB) size range.

Specifically, an entire sequencing run will be PB scale (e.g., could be 10s to 100s of PB); the "per patient" data size will approach 1TB as new technology is adopted.

4.1.5 Technology Support

The School of Public Health and the Laboratory for Pathogenesis of Clinical Drug Resistance and Persistence relies on a mixture of infrastructure dedicated to the research activities, and institutional resources described in Section 4.2.

4.1.5.1 Software Infrastructure

A list of tools that are used during the analyses process can be found here: https://gitlab.com/LPCDRP. A number of these are proprietary, and developed by our own group.

4.1.5.2 Network Infrastructure

A large quantity of data is downloaded to the research systems from external sources. The data that is shared with collaborators is much smaller in size, and is usually in the form of results of the analyses.

4.1.5.3 Computation and Storage Infrastructure

The current plan for this research is to commission a new cluster store all data, develop tools, and perform data analytics activities.

4.1.5.4 Data Transfer Capabilities

Data transfer is done routinely, in the form of downloading external data using sftp. Each raw sequencing movie file is generally in the order of GBs. For partners in the US and Europe, the transfer process meets expectations, but transfers to collaborators in India have repeatedly failed when file sizes reach the GB level.

The other regular form of data transfer is internal to SDSU: data migrates between workstations and cluster resources, which reside in different locations. The current cluster is on campus, while the workstations and the lab are off campus. A genome assembly that takes 20 minutes locally, takes over two hours when data is fetched from our cluster NFS.

4.1.6 Internal & External Funding Sources

Dr. Valafar is supported by NIH Grant #R01-AI105185-05A1[3], "NIAID Evolutionary and Functional Significance of Novel Mutations in MDR/XDR TB".

4.1.7 Resource Constraints

The following limitations exist:

[3] https://grantome.com/grant/NIH/R01-AI105185-05A1

- Limitations of storage, computing, and networking of the cluster and workstation infrastructure is severely reducing productivity as well as the range of experiments that are attempted.
- The workstations and cluster environment are currently over 10 years old, and in need a refresh.
- The latency that is witnessed, during frequent communication between the cluster and workstations, makes real-time analysis activities hard.

4.1.8 Ideal Data Architecture

The research requires:

- A well-defined and available set of IT support for hardware, software, and networking
- Upgraded computational hardware
- An increase in storage speeds, and capacities
- A more accessible backup system, with components that will work on and off-site.
- An increase in the communication capacities and speeds between the cluster and workstations used for research
- Additional support for software integration, as well as packaging, for external sharing would definitely accelerate discovery.

4.1.9 Outstanding Issues

Aforementioned data transfer issues to international collaborators.

4.2 San Diego State University Technology Overview

Content in this section authored by Matt Brown, Michael Farley, and Jerry Sheehan, SDSU Research Technology, Brian Lenz, SDSU College of Health and Human Services Information Technology and Christopher Leong, SDSU College of Sciences Technology Services.

4.2.1 Use Case Summary

The context for this use case is supporting the research described in Section 4.1. The lab's current setup includes a number of outdated, 10 year old hardware components consisting of Linux-based workstations and Linux-based servers used for storage and computational analysis.

Support for the hardware, both desktops and servers, has been provided by the college, currently College of Health and Human Services, IT support staff. Central IT provides network connectivity and operates the campus data center, although the current servers are house in a retired Library data center.

The server infrastructure that supports the research described in Section 4.1 is dedicated to this specific use case. The current Compute Cluster is composed of a mix of Rack Mount Servers and Desktop systems. There is a mix of Debian and Ubuntu Operating systems along with some Windows VM's for access to Window's only applications. The primary Cluster is located at a temporary data center in the campus Library. Primary user desktop systems are located at Dr. Valafar's Alvarado Research Offices.

There are dedicated VLAN's to secure and segment the Lab's network communications. The systems at the Library data center and Linux Desktops (also compute nodes) located at the Alvarado Lab use a secured VLAN. There is one server, Alborz, that has external access to outside world. NeoAlborz is the primary head node and LDAP server for user accounts, although Alborz may also have a legacy/secondary LDAP server as it was the original primary head node. NeoAlborz's primary network connection is within the secured VLAN, however it does have a secondary network connection to the outside world for proxy services.

4.2.2 Collaboration Space

Compute and storage resources on campus are very distributed. Some colleges, such as Engineering, have compute clusters for their needs. The Computational Science Research Center (CSRC)[4] support some researchers, but is limited in scope and primarily serves as a center to support a PhD program. SDSU has received NSF funds to support the Pacific Research Platform (PRP)[5] and does host some hardware for this in the CSRC data center. Efforts are underway to convert faculty containerized and running on PRP. Common use cases include utilizing Jupyter Notebooks and GPU resources as part of the PRP.

The Computational Science Research Center does support paid/sponsored HPC Resources and also manages the Sciences Research Network (100Gb). Other Colleges and Units also have Research HPC Clusters in their areas such as Engineering and

[4] http://www.csrc.sdsu.edu/

[5] https://pacificresearchplatform.org/

Management Information Systems in the Fowler College of Business. External resources are also used such as AWS, systems at National Labs, SDSC, etc.

4.2.3 Capabilities & Special Facilities

SDSU's main Campus Data Center offers hosting services along with 1GbE and 10GbE. The new HPC Server Infrastructure is planned to be deployed there along other clusters. The data center offers redundant power, generator backup power, security, and required cooling.

SDSU's main campus data center houses systems for business and academic purposes and was recently updated in 2020-2021 to increase capacity, efficiency, and security. The 5,000 square foot facility offers secure colocation services for departments and colleges. Cooling is provided by four Liebert units with built in capacity for redundancy and expansion as well as environmental monitoring. Racks are supplied with two power sources backed by two MGE UPS systems as well as a diesel generator for extended power outages. The data center also has a Halon fire suppression system. The data center includes several security measures including biometric palm readers, secure key box checkout as well as security camera coverage

Data center networking includes 10 GbE top of rack network for main campus networking. The Science DMZ also has a presence in the campus data center.

4.2.4 Technology Narrative

This overview combines responses from SDSU Research Technology, the SDSU College of Health and Human Services Information Technology, and the SDSU College of Sciences Technology Services.

4.2.4.1 Network Infrastructure

SDSU has two connections to SDSU's Internet Service Provider, CENIC. The primary connection is a dark fiber that directly connects to the CENIC POP at the San Diego Supercomputer Center (SDSC) located at University of California, San Diego. The fiber is connected with an Enhanced Wave-Division Multiplexer (EWDM). This allows SDSU to provide high-speed connectivity to the campus/administrative network and the Research Network. The campus network has a 20 Gbps connection to CENIC CalREN-DC ("Digital California") network. The campus/administrative network will be upgraded Q1 2022 to a can easily be upgraded to 100 Gbps connection, if needed. The campus network also has a redundant 20 Gbps connection over AT&T infrastructure to CENIC/Riverside. The EWDM allows the research network to support a 100Gbps connection on the DC network and another 100 Gbps connection to the CENIC CalREN-HPR ("High-Performance Research") network. The research described in Section 4.1 is located at Alvarado and connected to campus via a 10 Gb connection.

SDSU research environment has a Science DMZ ("SDMZ") which was funded through the NSF Office of CyberInfrastructure[6]. The network consists of Alcatel-Lucent OS10K

[6] http://iotlab.sdsu.edu/index.php/science-dmz/

and Brocade MLXe-4 routing and switching infrastructure. Connectivity is achieved through two independent 10 Gbps uplinks to the CENIC CalREN-DC ("Digital California") and CalREN-HPR ("High-Performance Research") networks, respectively, and a 100 Gbps uplink to the HPR network. The SDMZ is run by researchers. The Science DMZ is available to researchers upon request and has a physical presence in several buildings on campus, including the campus data center.

Figure 3, in Appendix A, features a breakdown of this design.

4.2.4.2 Computation and Storage Infrastructure

The CSRC does have Petabyte BGFS storage capabilities also at recharge.

Campus provides Google Drive storage to faculty and staff for free, but does not offer other services geared towards computational/storage intense workloads. AWS and Azure are available for use under a charge back arrangement, and architectural guidance is available for these platforms.

Virtual machine hosting services are available via charge back, but generally are for smaller workloads not geared towards HPC/HTC computing or large amounts of data.

4.2.4.3 Network & Information Security

Boarder and departmental firewalls are in place and make use of deep packet inspection, application detection and signatures, along with standard port filtering and traffic monitoring. These systems are actively monitored by the IT Network and Infrastructure Team in conjunction with the IT Security Office.

VLANs and associated Firewall Rules are used to segment and further secure access for Colleges, departments, instructional labs, and research labs. MAC addresses can also be restricted to specific ports for certain types of access.

The campus utilizes Palo Alto firewalls in three locations: border, data center, and departmental. The campus utilizes VLANs as security boundaries. GlobalProtect is used for VPN services for faculty, staff, and students.

4.2.4.4 Monitoring Infrastructure

SDSU utilizes PRTG[7] to monitor its network and compute resources. All campus network equipment is monitored and alerts are sent to staff to investigate issues. In addition, perfSONAR is deployed on the Science DMZ and monitored by researchers.

4.2.4.5 Software Infrastructure

A variety of Linux-based systems are used for research workflows, and many scripts are written in Python.

[7] https://www.paessler.com/prtg

4.2.5 Organizational Structures & Engagement Strategies

4.2.5.1 Organizational Structure

The SDSU Information Technology Division organizational chart is picture in Figure 2. More information about the structure and mission of this division can be found at: https://it.sdsu.edu/about.

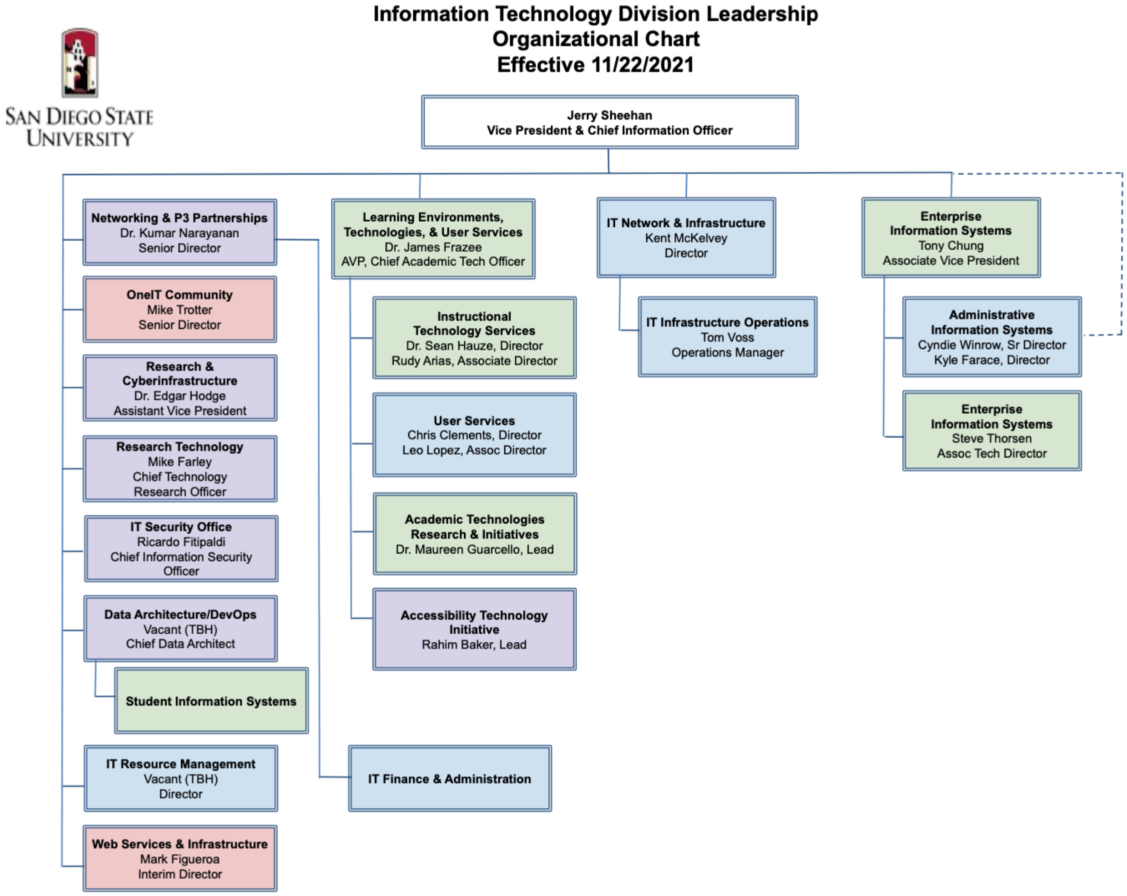

Information Technology Division Leadership
Organizational Chart
Effective 11/22/2021
San Diego State University
Jerry Sheehan
Vice President & Chief Information Officer
Networking & P3 Partnerships
Dr. Kumar Narayanan
Senior Director
OneIT Community
Mike Trotter
Senior Director
Research & Cyberinfrastructure
Dr. Edgar Hodge
Assistant Vice President
Research Technology
Mike Farley
Chief Technology Research Officer
IT Security Office
Ricardo Fitipaldi
Chief Information Security Officer
Data Architecture/DevOps
Vacant (TBH)
Chief Data Architect
Student Information Systems
IT Resource Management
Vacant (TBH)
Director
Web Services & Infrastructure
Mark Figueroa
Interim Director
Learning Environments, Technologies, & User Services
Dr. James Frazee
AVP, Chief Academic Tech Officer
Instructional Technology Services
Dr. Sean Hauze, Director
Rudy Arias, Associate Director
User Services
Chris Clements, Director
Leo Lopez, Assoc Director
Academic Technologies Research & Initiatives
Dr. Maureen Guarcello, Lead
Accessibility Technology Initiative
Rahim Baker, Lead
IT Finance & Administration
IT Network & Infrastructure
Kent McKelvey
Director
IT Infrastructure Operations
Tom Voss
Operations Manager
Enterprise Information Systems
Tony Chung
Associate Vice President
Administrative Information Systems
Cyndie Winrow, Sr Director
Kyle Farace, Director
Enterprise Information Systems
Steve Thorsen
Assoc Tech Director

Figure 2: SDSU IT Division Organizational Chart[8]

4.2.5.2 Engagement Strategies

The campus provides outreach with Pacific Research Platform (PRP) on a regular basis.

A few years a Campus wide survey for Research IT Needs was conducted to form a baseline for project needs. This included application licensing, storage needs, backup, security, web support, IT consulting services, HPC, and other topics.

Based on the survey we purchased additional site licenses and also Globus for large and secure file transfers.

[8] https://it.sdsu.edu/about/_files/itd-high-level-org-chart-20211122.pdf

4.2.6 Internal & External Funding Sources

SDSU was awarded NSF Grant 1245312[9] in 2012, "CC-NIE Network Infrastructure: Implementation of a Science DMZ at San Diego State University to Facilitate High-Performance Data Transfer for Scientific Applications", to facilitate construction of a Science DMZ network on campus.

4.2.7 Resource Constraints

Current constraints on the research described in 4.1 are that the systems are more than 10 years old. In addition, storage space is also a constraint limiting samples sequenced, stored, and analyzed. The research group has expressed a desire to increase the number of samples sequenced which is not supported with the current storage system.

Funding, storage, and shared compute resources all remain high priorities for improvement.

4.2.8 Outstanding Issues

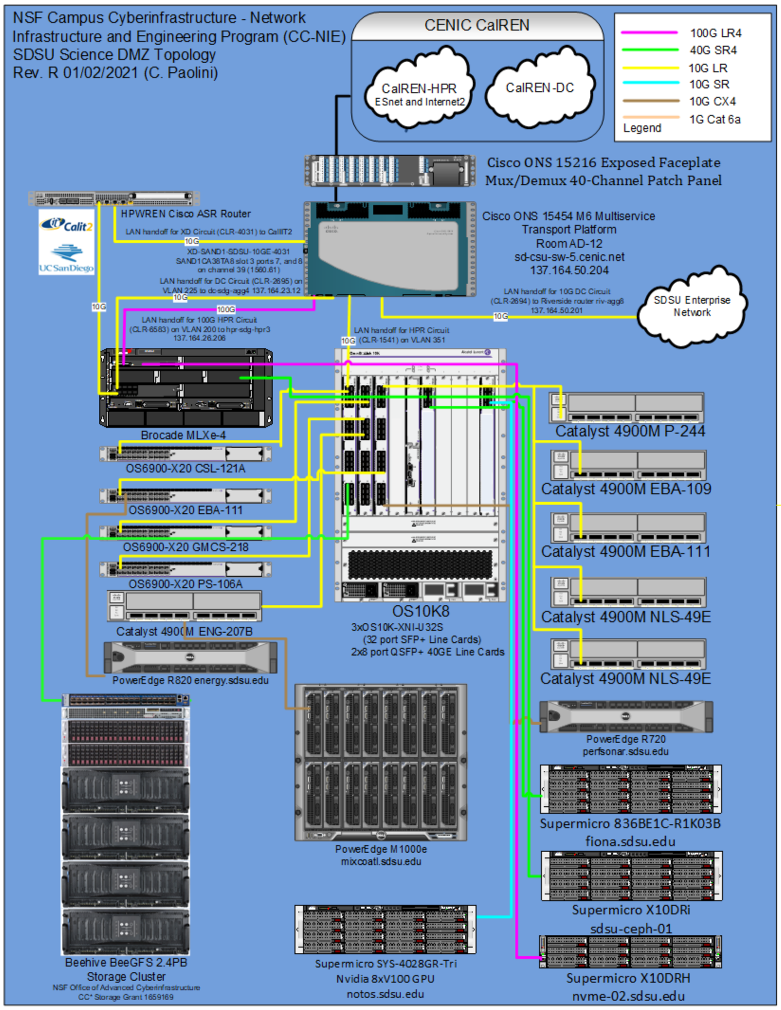
None to report at this time, beyond what was described in Section 4.1. The long term management of dedicated infrastructure will require more staff and expertise.

[9] https://www.nsf.gov/awardsearch/showAward?AWD_ID=1245312

Appendix A – SDSU Network Diagrams

Current SDSU Science DMZ

Funding for the Science DMZ infrastructure was provided through NSF Office of CyberInfrastructure CC-NIE Grant 1245312[10]. The goal of this funding was to design and deploy a Science DMZ to support data intensive science and accommodate specific end-to-end dataflow scenarios. The original designed featured a 10Gbps connection to the CENIC/CalREN-HP network. This upgrade enabled researchers at SDSU to host a number of delay-sensitive applications and datasets to support research projects in the numerical simulation of earthquake rupture and wave propagation, coastal ocean modeling, pulse detonation engines, vortex rings in Bose-Einstein condensates, and large-scale data for proteomics, gene promoter bioinformatics, and microbial metagenomics.



NSF Campus Cyberinfrastructure - Network Infrastructure and Engineering Program (CC-NIE)
SDSU Science DMZ Topology
Rev. R 01/02/2021 (C. Paolini)
CENIC CalREN
CalREN-HPR
ESnet and Internet2
CalREN-DC
100G LR4
40G SR4
10G LR
10G SR
10G CX4
1G Cat 6a
Legend
Cisco ONS 15216 Exposed Faceplate Mux/Demux 40-Channel Patch Panel
Cisco ONS 15454 M6 Multiservice Transport Platform
Room AD-12
sd-csu-sw-5.cenic.net
137.164.50.204
HPWREN Cisco ASR Router
Calit2
UC San Diego
LAN handoff for XD Circuit (CLR-4031) to CalIIT2
10G
XD-SAND1-SDSU-10GE-4031
SAND1CA38TA8 slot 3 ports 7, and 8 on channel 39 (1560.61)
LAN handoff for DC Circuit (CLR-2695) on VLAN 225 to dc-sdg-agg4 137.164.23.12
10G
10G
100G
LAN handoff for 100G HPR Circuit (CLR-6583) on VLAN 200 to hpr-sdg-hpr3 137.164.26.206
LAN handoff for 10G DC Circuit (CLR-2694) to Riverside router riv-agg8 137.164.50.201
10G
SDSU Enterprise Network
LAN handoff for HPR Circuit (CLR-1541) on VLAN 351
10G
Brocade MLXe-4
OS6900-X20 CSL-121A
OS6900-X20 EBA-111
OS6900-X20 GMCS-218
OS6900-X20 PS-106A
Catalyst 4900M ENG-207B
PowerEdge R820 energy.sdsu.edu
OS10K8
3xOS10K-XNI-U32S
(32 port SFP+ Line Cards)
2x8 port QSFP+ 40GE Line Cards
Catalyst 4900M P-244
Catalyst 4900M EBA-109
Catalyst 4900M EBA-111
Catalyst 4900M NLS-49E
Catalyst 4900M NLS-49E
PowerEdge R720
perfsonar.sdsu.edu
Supermicro 836BE1C-R1K03B
fiona.sdsu.edu
Supermicro X10DRi
sdsu-ceph-01
Supermicro X10DRH
nvme-02.sdsu.edu
PowerEdge M1000e
mixcoatl.sdsu.edu
Supermicro SYS-4028GR-Tri
Nvidia 8xV100 GPU
notos.sdsu.edu
Beehive BeeGFS 2.4PB
Storage Cluster
NSF Office of Advanced Cyberinfrastructure
CC* Storage Grant 1659169

Figure 3: Current SDSU Science DMZ

The project directly addressed the needs of faculty and students who encompassed the Computational Science Research Center, which draws participation from all science and

[10] https://www.nsf.gov/awardsearch/showAward?AWD_ID=1245312

engineering departments at SDSU. The Science DMZ has provided researchers with the capability to rapidly exchange large datasets and deploy Web-based science applications that are accessible through a dedicated network that is separate from, and therefore not impacted by, general Internet traffic generated by the campus population. The DMZ has promoted remote usage of computing resources at SDSU, cultivate development of collaborative tools for sharing data with the broader scientific community, establish new research partnerships, and foster new mentorship opportunities between faculty and students engaged in computational science.

Potential SDSU Science DMZ Upgrades

During the April 2022 review of science drivers, SDSU and EPOC collaborated on the diagram featured in Figure 4. This is a rough attempt to improve the internal traffic flows support remote sections of campus that operate scientific instruments, and would benefit from the use of centralized HPC and storage resources.

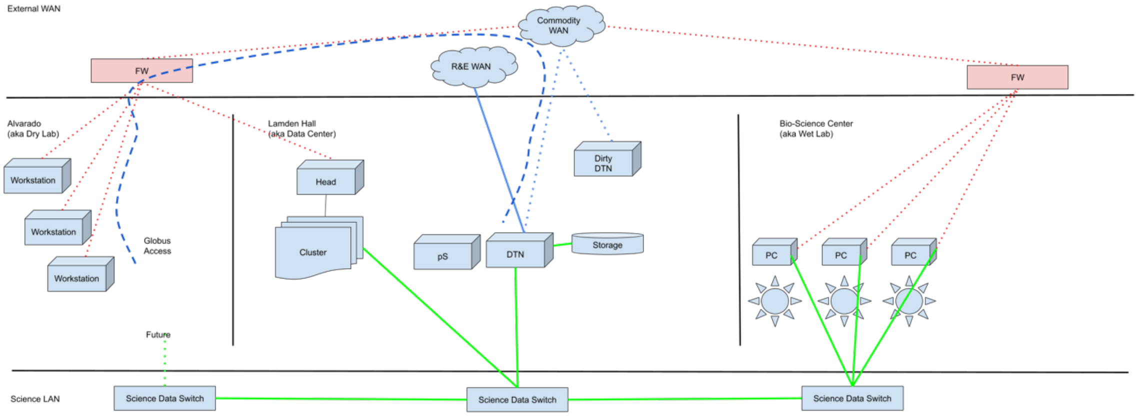


Figure 4: Proposed Internal Science DMZ

This data-architecture allows for a low-latency, high capacity, internal network that would facilitate sharing data from remote instruments, to the SDSU institutional computing and storage capabilities. It also facilities access to special-purpose DTN infrastructure, for sharing data with external collaborators, through the existing Science DMZ connection that could also be upgraded over time.