Title

St. Mary’s University Requirements Analysis Report

Permalink

Authors

Zurawski, Jason
Schopf, Jennifer
Southworth, Douglas
et al.

Publication Date

2023-01-16

Copyright Information

St. Mary’s University Requirements Analysis Report

January 16th, 2023

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 Regents of the University of Texas System, 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 Regents of the University of Texas System. 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 Regents of the University of Texas System.

St. Mary's University Requirements Analysis Report

January 16th, 2023

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-2001503[1].

[1] https://escholarship.org/uc/item/5d65w10j

Participants & Contributors

Mehran Aminian, St. Mary's University
David "Troy" Christman, St. Mary's University
Gopalakrishnan Easwaran, St. Mary's University
Austin Gamble, LEARN
Todd Hanneken, St. Mary's University
Byron Hicks, LEARN
Joseph Longo, St. Mary's University
Maria “Isa” Lopez, St. Mary's University
Wenbin Luo, St. Mary's University
Frank Niewierski, St. Mary's University
Gary Ogden, St. Mary's University
Leticia Romero, St. Mary's University
Jennifer Schopf, TACC
Amy Schultz, LEARN
Doug Southworth, TACC
Ajaya Swain, St. Mary's University
Vahid Vemamian, St. Mary's University
Curtis White, St. Mary's University
Jason Zurawski, ESnet

Report Editors

Austin Gamble, LEARN: austin.gamble@tx-learn.net
Byron Hicks, LEARN: byron.hicks@tx-learn.net
Jennifer Schopf, TACC: jms@tacc.utexas.edu
Amy Schultz, LEARN: amy.schultz@tx-learn.net
Doug Southworth, TACC: dsouthworth@tacc.utexas.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 September and December 2022, staff members from the Engagement and Performance Operations Center (EPOC) met with researchers and staff from LEARN and St. Mary's University for the purpose of a Deep Dive into scientific and research drivers. The goal of this activity was to help characterize the requirements for a number of campus use cases, and to enable cyberinfrastructure support staff to better understand the needs of the researchers within the community.

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

- Jubilees Palimpsest Project
- Industrial Engineering
- Information Services

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 St. Mary's 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:

- St. Mary's University Information Services will have several opportunities to work with members of the research community on improving scientific workflows. This includes:
  - Helping to design next generation computation and storage systems
  - Convert software to take advantage of containerization strategies
  - Lleverage the use of remote resources.

- St. Mary's University could benefit from partnership with R&E computational providers, like TACC, and leverage the LEARN network to reach these resources

- St. Mary's University and LEARN can collaborate on ways to support cloud computing and strategies via network peering.

- St. Mary's University Information Services will review policies on handling sensitive aspects of research data.

- St. Mary's University Information Services, LEARN, and EPOC will start a conversation regarding support for perfSONAR testing, policy for the use of Science DMZ resources, architectural considerations for the Science DMZ, and future CC* proposal options.

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 St. Mary's 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:

- St. Mary's University Information Services should start a conversation with the Jubilees Palimpsest Project regarding ways that the research activity can be better supported when at remote field sites. This may take the form of portable computation and storage, along with containerized workflows, that can facilitate more productivity when not on campus.
    - Work has begun on the next generation machine. The Fractal Node 202 can fit inside a carry-on and pack a desktop-class CPU. This features an ITX motherboard (greater variety of ports), runs hotter than a laptop but also has a desktop (vs. mobile) CPU. A gaming laptop typically has more GPU capabilities, which are not needed. The compilation still arries from carrying on a plane, vs shipping.

- St. Mary's University Information Services should work with the Jubilees Palimpsest Project on ways to containerize the analysis workflow so that it may be run in more locations than just the core computation in the campus data center.

- The Jubilees Palimpsest Project and TACC can discuss options for improving computation and storage allocations.

- St. Mary's University Information Services and LEARN should discuss options to support peering with cloud providers, as well as paths to support workflows at TACC.

- St. Mary's University Information Services should work with the Department of Engineering to better understand the sensitive aspects of research data, and offer assistance in storage, computation, and safe-handling procedures. This could include new tools to facilitate sharing, versus the use of offline media.

- St. Mary's University Information Services should continue their relationship with the sponsored research office on campus to understand the impacts of funding on technology needs.

- St. Mary's University Information Services, LEARN, and EPOC will start a conversation regarding support for perfSONAR testing.

- St. Mary's University Information Services, LEARN, and EPOC will help to develop policy for the use of Science DMZ resources

- St. Mary's University Information Services, LEARN, and EPOC will discuss architectural considerations for the Science DMZ

- St. Mary's University Information Services and LEARN can investigate future CC* proposal options.

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 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 St. Mary's University Deep Dive Background

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

- Jubilees Palimpsest Project
- Industrial Engineering
- Information Services

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 the Texas Advanced Computing Center (TACC) 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.

The Texas Advanced Computing Center (TACC) at the University of Texas at Austin designs and deploys the world's most powerful advanced computing technologies and innovative software solutions to enable researchers to answer complex questions to help them gain insights and make discoveries that change the world. TACC's environment includes a comprehensive cyberinfrastructure ecosystem of leading-edge resources in high performance computing (HPC), visualization, data analysis, storage, archive, cloud, data-driven computing, connectivity, tools, APIs, algorithms, consulting, and software.

Lonestar Education And Research Network (LEARN) is a consortium of 43 organizations throughout Texas that includes public and private institutions of higher education, community colleges, the National Weather Service, and K–12 public schools. The consortium, organized as a 501(c)(3) non-profit organization, connects its members and over 300 affiliated organizations through high performance optical and IP network services to support their research, education, healthcare and public service missions. LEARN is also a leading member of a national community of advance research networks, providing Texas connectivity to national and international research and education networks, enabling cutting- edge research that is increasingly dependent upon sharing large volumes of electronic data.

St. Mary’s University in San Antonio is a private liberal arts school and the oldest Catholic university in Texas. Through small classes, close student-faculty relationships and an engaged community, teaching and learning flourish at St. Mary’s. St. Mary’s offers a variety of academic programs in humanities, sciences and business, and options to pursue graduate, doctoral and law degrees. Our students engage in learning and social opportunities through undergraduate research, internships and community engagement.

4 St. Mary's University Case Studies

St. Mary's University presented a number use cases during this review. These are as follows:

- Jubilees Palimpsest Project
- Industrial Engineering
- Information Services

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 Jubilees Palimpsest Project

Content in this section authored by Todd Hanneken, from St. Mary's University and the Jubilees Palimpsest Project

4.1.1 Use Case Summary

The St. Mary's University Jubilees Palimpsest Project operates within the intersection of study in the digital humanities, multispectral imaging, remote sensing, image processing, and cultural heritage imaging. Technology support is a major requirement to advancing this field of study.

Palimpsests are manuscripts that were erased so that the parchment could be reused. The erased text is generally illegible to the human eye, but can be recovered with multispectral image capture and processing. Following capture, a significant amount of processing takes place at the origin of the palimpsests (e.g., a remote facility such as a museum), which amplifies the amount of data ingested. Additional processing takes place on the servers affiliated with the project (located at St. Mary's University) and is copied off site for redundancy to cloud instances (currently located in AWS).

4.1.2 Collaboration Space

The Jubilees Palimpsest Project is the primary uploader of image data that is captured at remote locations when studying artifacts. On occasion some other collaborators may contribute images as well. The standard approach to uploading is to leverage common tools (e.g., SFTP, FileZilla).

There are two main classes of collaborators that may be looking to download or view data from the Jubilees Palimpsest Project: common internet users and more sophisticated image-processing scientists. Common internet users may browse the images (typically they are not downloading files for study), which results in a lower-resolution delivery of content (e.g., compression of images, but does require server-side processing). Image processing scientists will download the raw data for their own processing, so they can improve their own analysis approaches. The Rochester Institute of Technology (RIT), and the Early Manuscripts Electronic Library, are both collaborators in the image processing space.

4.1.3 Instruments & Facilities

The project currently has two servers (besides backups). One is an AWS instance with only the storage required for the general public viewer, not the data archive. It does not now, but formerly did, utilize AWS elasticity. It does use AWS Cloudfront, which can be useful because the audience is significantly international.

The other server is bare metal (not a virtual machine) in the St. Mary's University data center. It mirrors all the functionality of the AWS server plus the complete data archive, image processing, and development projects.

4.1.4 Data Narrative

The Jubilees Palimpsest Project[3] has three basic workflows:

- Data acquisition
- Data processing
- Data sharing

Figure 1 shows how these three use cases can be logically described, and all have a relationship to the core data but can be performed independently of each other.

The data acquisition workflow involves remote science work in locations that house palimpsest artifacts. Due to the remote nature of this work, it is done infrequently and relies on travel, and short timelines to accomplish basic goals. The workflow can be described as follows:

1. Team travels to remote location with remote imaging equipment provided by Megavision (e.g., a private company)
2. During field study, source palimpsest artifacts are captured
    a. Initial processing of images is required to be sure that the camera is calibrated correctly – this requires availability of on-site computing and storage, or access to a network to support remote computing and storage
    b. After calibration, multiple images are captured
3. Data acquired during field study may be uploaded to cloud storage, or St. Mary's University, if local networking supports this operation
4. Data acquired is also physically carried back on removable media

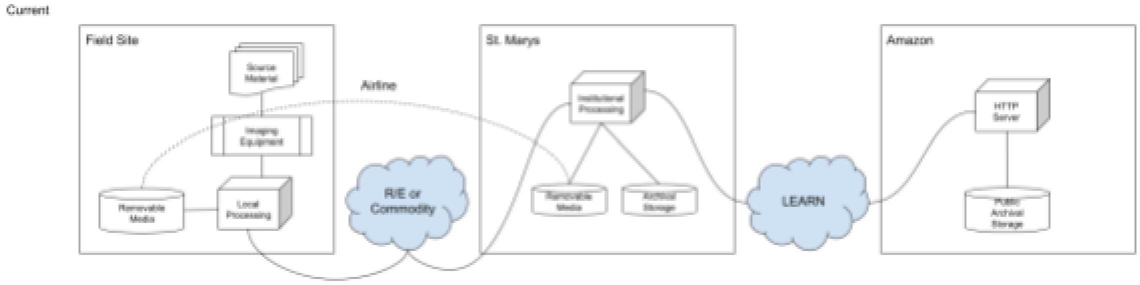

Figure 1: Jubilees Palimpsest Project Workflow

The data analysis workflow involves computation and storage technologies to perform analysis on the previously acquired palimpsest images. This work is currently done using a server at St. Mary's University, and involves running several custom image processing scripts. Results are then disseminated. The workflow can be described as follows:

1. Palimpsest images are processed using scripts on hardware that is located at St. Mary's University.
2. Analysis focuses on trying to find evidence of previous writing (e.g., edge detection)

[3] https://jubilees.stmarytx.edu/ (AWS) and https://palimpsest.stmarytx.edu/ (St. Mary's Campus)

3. Regions of interest are identified, and may be re-processed over time (as technology improves)

Lastly, the data sharing workflow involves the use of networking, and physical shipping, to share the image results with collaborators and the general public. The workflow can be described as follows:

1. Palimpsest images are uploaded to AWS where the portal system will make them available for viewing through Mirador software. Images are displayed as lower quality JPG images through this software to reduce download times and bandwidth requirements.
2. Collaborators may also download high resolution copies (using tools like SFTP) or from the campus server directly.
3. Collaborators with less connectivity may also request physical copies to be sent via mail.

For this work, knowledge preservation and access are fundamental to discovery, especially collaborative discovery. The availability of RAID and error checking are also fundamental, as is serving content over Apache to the collaborators. The analysis workflow processing is done using python, and mostly maintained by researchers at St. Mary's University with the exception of some downstream libraries. Processing produces the relatively small sets of data, that can be visualized by screens and human eyes.

4.1.4.1 Data Volume & Frequency Analysis

The data that is produced by the Jubilees Palimpsest Project can be classified as Terabytes (TB) on a yearly basis. This consists of newly captured raw palimpsest images, as well as the analysis products that are produced over time. A breakdown of the data is as follows:

- A single raw image (50 megapixels, 16 bits per pixel) can be around 100MB after initial capture, in the most uncompressed format
- Over a hundred images may be captured for a single page of a manuscript. A manuscript may be hundreds of pages, and multiple manuscripts may be imaged. Data capture is limited by number of days on site (5-20) and images per day (~2000). Twenty days of capture could produce 4 TB of data. Processing produces tenfold derivative data, not all of which needs to be preserved.
- A single year can produce TB of data, and consists of multiple folios for a given manuscript, and several manuscripts could be captured.

4.1.4.2 Data Sensitivity

There are no sensitive aspects to the data used in this research.

4.1.4.3 Future Data Volume & Frequency Analysis

The data that is produced by the Jubilees Palimpsest Project is not expected to increase by any orders of magnitude. With faster processing there may be more derived data products, and it may be possible to analyze more frequently, but the process is still gated on how many new palimpsest source images can be generated.

4.1.5 Technology Support

4.1.5.1 Software Infrastructure

A number of software packages are used during the analysis phase:

- Kakadu[4] is a non-free software is for jpeg2000 compression and decompression.
- Ubuntu and CentOS are used for operating environments
- Custom python scripts are used for analysis
- IIP Image Server[5]
- Apache
- Previously have used SimpleAnnotationServer[6]
- The main visualization software is Mirador[7], with some use of Leaflet[8].

4.1.5.2 Network Infrastructure

Networking between the office and campus data center was recently upgraded from 10Mbps, and is sufficient for campus networking needs. The only times this can be a challenge is when uploading multiple GB to TB of data after initial acquisition, where it may take hours.

4.1.5.3 Computation and Storage Infrastructure

There are two major computational and storage environments:

- St. Mary's University Computation and Storage
- Amazon Cloud Infrastructure (AWS)

The on campus resource contains 64 GB of ECC RAM, and 64 TB raw storage on 8 spindles. The AWS instance is designed to be smaller in terms of computation, and has less storage (300GB: only the compressed final products ready for use by the general public). The primary use case is disseminating images to the public and is designed to serve images forever.

A primary goal of this work is to ensure operation as staffing changes over time (including project leadership). The project would like to move toward a model that utilizes containers or VMs for essential operations that can be ported to resources (cloud or local to university). This will make the analysis and sharing workflows more fungible, and able to be operated from wherever there are resources available.

Digital Humanities projects have a big problem, ""What happens when the PI dies?"" My project's hope for immortality is to have a very small, simple, secure, portable VM that

[4] https://kakadusoftware.com

[5] https://iipimage.sourceforge.io/documentation/server/

[6] https://github.com/GlenRobson/SimpleAnnotationServer

[7] https://projectmirador.org

[8] https://leafletjs.com

gives posterity what it needs. I don't care whether it is on AWS or campus. Most users are off campus, but that server does not use much bandwidth.

4.1.5.4 Data Transfer Capabilities

Years ago (2017) there was a need to upload 8TB of data from removable media to the data center on campus. The experience (at the time) did not go well, but the cause was never fully identified. The student worker tasked with doing the activity could have been using wireless, or the wrong transfer tools. Since this time, there have been improvements to the campus network, but there has not been a need to migrate that much data in a single session.

4.1.6 Internal & External Funding Sources

The project has received support from the NEH, with the last grant spanning 2016-2019.

4.1.7 Resource Constraints

Camera technology has improved dramatically since the start of the project. In 2017 a camera used for research was rated at 50 megapixel per image; new sensors are 3x better can producing 3x more data as a result. To cope with the increase in data, the project is deploying strategies to ensure that easily derived data (e.g., gamma corrections, simple analysis products, etc.) are not stored to save space.

4.1.8 Ideal Data Architecture

The data architecture described in Section 4.1.5.3 could use improvements to make it more portable. Possible suggestions are:

- Porting portions of the workflow to easily deployed and maintained containers that can be run anywhere
- Simplifying some portions of the workflow to take advantage of some automation
- Experimenting with other forms of local and remote processing and storage approaches

4.1.9 Outstanding Issues

None to report.

4.2 Industrial Engineering

Content in this section authored by Gopalakrishnan Easwaran from St. Mary's University, Engineering Department

4.2.1 Use Case Summary

The Engineering department at St. Mary's University features programs that focus on Industrial Engineering, Engineering Management, and Data Analytics. As a part of this, there is currently an effort to study large-scale optimization problems in supply chain management, logistics and production/service planning and operations. This process involves faculty and students from the engineering department using HPC Cluster resources, the Machine Learning Cluster, and a number of Windows servers. Some of the data that is involved in this research is confidential data that resides on a faculty member's computer, and must be transferred to the servers for computational experimentations and analytics. The results of the research are archived in faculty computers or external storage devices.

4.2.2 Collaboration Space

The main collaboration space for this work comes from:

- Real-time data from partner companies that is shared under NDA with faculty
- Simulation data, that is generated at St. Mary's University
- Online libraries containing data sets for experimentation

All of the experimentation and research involved in this project reside within St. Mary's domains. Data is shared through external hard storage. Future plans may involve more online forms of data sharing via cloud storage.

4.2.3 Instruments & Facilities

This research involves the use of St. Mary's University institutional resources: HPC Cluster, Deep Learning Cluster, and Windows Server.

4.2.4 Data Narrative

The research follows a workflow structure:

1. Data from industrial partners is collected via offline collection mechanisms (e.g., external media)
2. Simulation data from online sources is downloaded
3. Researchers at St. Mary's University utilize HPC resources to generate additional simulation
4. Training data is fed to AI/ML analysis tools on GPU cluster, and then AI/ML tools are applied to tested against simulations and real-world observations from industrial partners.
5. Results are analyzed, and additional analysis may be done as required.

Instruments are used for both research (knowledge discovery) and teaching courses (senior undergraduate and graduate) within Engineering.

4.2.4.1 Data Volume & Frequency Analysis

The data that is produced by this project can be classified as Gigabytes (GB) on a monthly basis.

4.2.4.2 Data Sensitivity

Yes, there are sensitive aspects to the use case's data: the use of NDA controlled supply chain data from partners in industry requires a level of protection.

4.2.4.3 Future Data Volume & Frequency Analysis

The data that is produced by this project may grow to the level of Gigabytes (GB) on a weekly basis. If new approaches to processing are used, that can increase the number of simulations. Additionally, reviewing more real-world data sets can provide additional insight and opportunities.

4.2.5 Technology Support

4.2.5.1 Software Infrastructure

The following software packages are used in the process of research:

- C++ and python on both Linux and Microsoft Visual Studio. Also includes a number of machine learning libraries for python.
- Microsoft Access and My SQL Server on Windows
- Microsoft Excel
- Tableau[9]
- Power BI[10]
- Alteryx[11] for data analytics/visualization
- ILOG CPLEX[12]
- Gurobi Optimization Solver[13]

4.2.5.2 Network Infrastructure

No additional information on network connectivity was provided, and Section 4.3 should provide institutional information.

4.2.5.3 Computation and Storage Infrastructure

Current requirements require that multi-TB external storage devices be used for storing data that is under NDA. Future requirements may facilitate the use of cloud sharing to store and exchange this data.

4.2.5.4 Data Transfer Capabilities

It is common that during the course of research, which is within the university domain, there are multiple GB of data that must be exchanged between servers and faculty devices. The data upload/download typically takes hours to complete.

[9] https://www.tableau.com

[10] https://powerbi.microsoft.com/en-us/

[11] https://www.alteryx.com

[12] https://www.ibm.com/products/ilog-cplex-optimization-studio

[13] https://www.gurobi.com

4.2.6 Internal & External Funding Sources

No sources of funding were reported.

4.2.7 Resource Constraints

During the course of research, the slower network speeds, limited network capacity, and dropped connections can limit the productivity. This is typically seen when sending data from external hard drives to institutional storage.

4.2.8 Ideal Data Architecture

The research requirements to support AI and ML continue to increase, and having access to a Hadoop[14]-based cloud data architecture would assist with managing "big data" research and teaching activities.

4.2.9 Outstanding Issues

No outstanding issues were reported.

[14] https://hadoop.apache.org

4.3 Information Services

Content in this section authored by Joseph Longo from St. Mary's University, Information Services

4.3.1 Use Case Summary

The technology profile is being prepared by the Information Services group, and covers the enterprise network, along with research networks for the High Performance Cluster and Deep Learning Cluster. The enterprise network is operated by the Infrastructure and Enterprise Systems group within Information Services.

The servers are operated and maintained by the Science, Engineering and Technology school on campus. The faculty utilize these systems to perform instruction and research.

4.3.2 Collaboration Space

No additional collaborations are reported at this time.

4.3.3 Capabilities & Special Facilities

Currently all the services are all based on the peering arrangements and research networks through the grant facilitated by the NSF and LEARN. There is limited cloud infrastructure today, and the site operates mostly as an on-premises data center.

4.3.4 Technology Narrative

4.3.4.1 Network Infrastructure

St. Mary’s currently has a 10Gbps symmetrical connection through the TX-LEARN network. The last mile for this connection is provided by Spectrum Enterprise. There is also a 10Gbps BGP connection provided by AT&T for fault tolerance. The internal campus network has a mix of both 1Gbps and 10Gbps network segments to the edge.

4.3.4.2 Computation and Storage Infrastructure

The University has a few different environments and computational services. Within SET there is an HPC Array used by the faculty. SET also has an AMAX Deep Learning Cluster used for machine learning and analytics to support research efforts.

The central IT data center is a traditional data center with minimal iron/physical servers and a significant virtual server infrastructure.

4.3.4.3 Network & Information Security

Currently the security architecture is using Trinity Cyber to monitor and protect bi-directional internet traffic outside the perimeter. This service sits inline in the Equinox data center. This possible because of the TX-LEARN connectivity. This is done at layer 2.

For boundary protection, St. Mary’s University utilizes Fortinet Next-Generation firewalls in a HA pair configuration; these are currently 1800D devices. An additional HA Pair of Fortinet 1500D Next Generation firewalls that will be configured in front of the server infrastructure.

At the host level, Information Services is using Carbon Black EDR, but are going to make a transition to Microsoft Defender in the upcoming months.

4.3.4.4 Monitoring Infrastructure

Information Services uses Solar Winds, and perfSONAR capabilities to monitor network performance.

4.3.4.5 Software Infrastructure

No additional software packages are enumerated – faculty can request assistance but typically operate their own software as required.

4.3.5 Organizational Structures & Engagement Strategies

4.3.5.1 Organizational Structure

Currently, St. Mary's University operates a model of Centralized IT support. Curtis White is the VP of Information Services and he has 5 direct reports:

- Joseph Longo (Information Security)
- Troy Christman (Infrastructure & Enterprise Services)
- Frank Niewierski (Client & Systems Support Services)
- Jeff Schomburg (Academic Technology Services)
- Felicia Maldonado (Library Services).

Some of our Faculty operate and maintain server infrastructure to support the research and teaching.

4.3.5.2 Engagement Strategies

Information Services tries to work with our Science, Engineering and Technology school as frequently as possible. When a request comes in, attempts are made to facilitate and help provide solutions to encourage and support the research and educational requirements.

4.3.6 Internal & External Funding Sources

St. Mary's University and LEARN are collaborating on NSF award number 1925553, to support the construction of a Science DMZ on campus.

4.3.7 Resource Constraints

No resource constraints were reported.

4.3.8 Outstanding Issues

No outstanding issues were reported.

5 Case Study Discussion & Campus Planning

On November 30th 2022, staff from St. Mary's University, LEARN, and EPOC participated in a discussion on the use cases and potential next steps to develop a set of sustainable approaches to provide technological support. Notes from this set of discussions appear in the following sections.

5.1 Jubilees Palimpsest Project

Much of the discussion for this case study involved understanding, and creating, the workflow seen in Figure 1. After understanding how it works currently, some discussion was devoted to ways that it may be improved in the future. The focus was put on several of the potential areas of bottleneck:

- Lack of on-site computation (or network) to support analysis during the image capture process during field study
- Scalability of computing resources to perform the analysis workflow
- Scalability of storage and networking resources to perform the sharing workflow

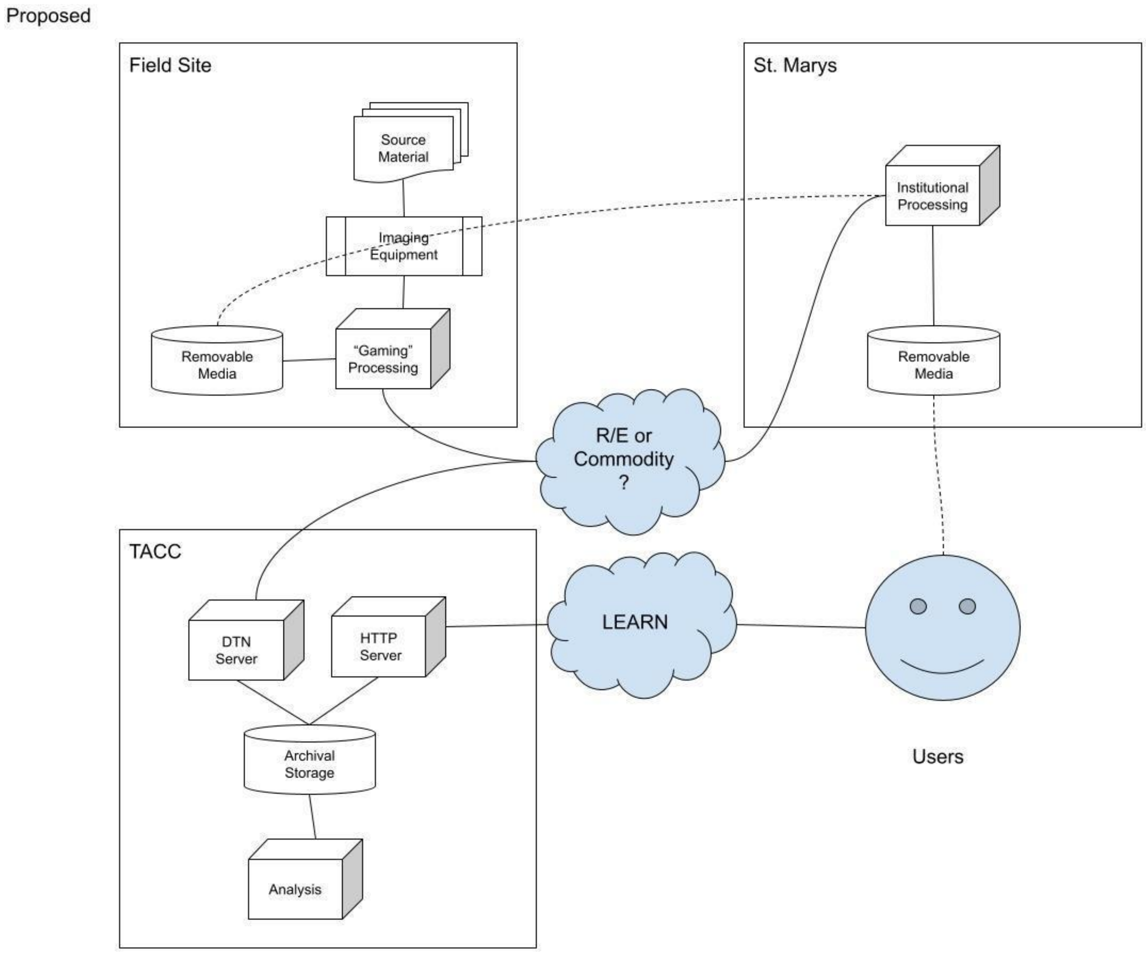

Figure 2: Proposed Workflow Changes

To address some of these concerns, the workflow in Figure 2 was created. Some of the changes include:

- Development of a “portable” system that can be taken to remote locations when imaging. This would consist of a higher-specification analysis laptop (e.g. increased CPU cores and processor, memory, storage, and GPU), potentially based off of a “gaming” device, as well as software that would facilitate data transfer back to either St. Mary’s or an alternate location (e.g., cloud resources, or allocations at Texas Advanced Computing Center [TACC]). This system would facilitate calibration runs on site, limited analysis work, and also would start the data transfer process so that longer-term analysis could proceed in parallel.
- Porting the analysis workflow to containers, such that it could be run anywhere (e.g., cloud resources, or allocations at TACC). This would facilitate more locations for processing and re-processing.
- Applying for allocations at TACC, so that analysis, storage, and sharing workflows could be supported at an R&E connected site.
- Working with LEARN to ensure the engineered path to cloud resources is working efficiently

Additional discussion also centered on the role of AI/ML approaches during the analysis stage, and if these would help speed up the research effort. These have not been tried previously, and may work, but would require some assistance from researchers to ensure the training (and subsequent identification) were working correctly.

5.2 Industrial Engineering

The case study review had several discussion points to consider for the future:

- How is sensitive data (even on external drive) managed today? In particular is the project PI following requirements for the NDA’ed data, and is there any form of audit or access control that is required by external parties? Some concerns exist about who is responsible, or held accountable, during a potential data breach event.
- If the sensitive data were to be accessed via cloud share, how would that change the campus technology requirements?
- How large are the relative data sets (e.g., shared partner data, simulations), how long are they relevant for the research, and are they saved indefinitely? Data storage over time could be a factor, particularly if there are security controls required.

5.3 Information Services

Most of the discussion on the technology support use case focused on ways the group could support research needs going forward. Some highlights included:

- Working more closely with the research community to assist in the management of resources. This includes clusters, software packages, and sensitive data sets. Ideally the IS group could unburden researchers from administrative and operational tasks, and make the infrastructure more secure during that process.

- Understand the areas of friction in data transfer for both “east-west” (e.g., within a campus) and “north-south” (e.g., from campus to external resources). The best way to do this will be via tools like perfSONAR with LEARN and EPOC can assist with.
- How to grow computing resources. Right now research groups have their own, but there is no centrally managed HPC or GPU-based resources that anyone can use. To address this, there are some options:
    - Operating “private” clusters for a larger population, and even augmenting thing capabilities (e.g., condo model)
    - Partnering with others in Texas, via LEARN, to operate community computing
    - Working with TACC to steer those that need computing resources
    - Building cloud capabilities (e.g., standard workflow configurations) that can be deployed as needed
- Continuing to work with the grant office to understand when IT needs will be included in submitted applications for funding.
- Developing a model where new IT resources can be centrally managed (for education or research) vs. faculty creating and managing their own
- Exploring what it means to have a sensitive data environment: either on campus or through partnerships with others in R&E or industry
- The general struggles of upgrading and maintaining infrastructure around campus

Appendix A – The Lonestar Education And Research Network (LEARN)

Introduction

The Lonestar Education And Research Network (LEARN) is a consortium of 43 organizations throughout Texas that includes public and private institutions of higher education, community colleges, the National Weather Service, and K–12 public schools. The consortium, organized as a 501(c)(3) non-profit organization, connects its members and over 300 affiliated organizations through high performance optical and IP network services to support their research, education, healthcare and public service missions. LEARN is also a leading member of a national community of advance research networks, providing Texas connectivity to national and international research and education networks, enabling cutting- edge research that is increasingly dependent upon sharing large volumes of electronic data.

LEARN's Mission

Empower non-profit communities to execute their missions through technology and collaboration.

LEARN's Vision

LEARN will be the most efficient and effective enabler of research, education, healthcare, and public service communities in Texas using technology and shared services.

Network Services

Members are entitled to appoint an individual to the Board of Directors and to acquire network services from LEARN at member rates. Network services are designed and provisioned based on the needs of individual members through collaboration between those members and the LEARN staff.

Network services, which are funded by the members who consume the services at rates which are set by the Board, sustain current and future network requirements including capital refresh at periodic intervals to keep the network state-of-the-art.

Network services include:

- Layer 1 Dedicated Transport Services Between LEARN Points-of-Presence (POPs),
- Layer 2 IP/MPLS Transport Services,
- Service Level Agreement (SLA) based Layer 2 connections to Cloud Service Providers (AWS, Google, & Azure),
- Routed Layer 3 IP Services,
- Connection Gateways to the National Research and Education Networks (Internet2 and Energy Sciences Network, and on 100G ramps to reach Pacific Wave International Exchanges),
- Seamless access to on-net data centers,
- Inter-POP Port aggregation & Co-location Services

- Commodity Internet Services (100G burst capacity spread across 4 POPs),
- Low-Latency High-Capacity Access to Content and Application Providers (Peering and Caching Services),
- DDoS Mitigation Service,
- Managed Network Service and Consultation, and
- Unmetered Network Service.

LEARN is currently listed as a telecommunication/Internet service provider with the Universal Service Administration Company (USAC). Becoming a USAC telecommunications/Internet service provider permits LEARN's school, library, and rural healthcare customers to receive significant discounts through the Universal Services Fund.

The Board and the staff are committed to ensuring LEARN remains the trusted and preferred means by which its members obtain network services in Texas. There is a broad consensus among LEARN's members that the organization has a unique role to play in the state in providing highly reliable, cost-effective network services to the higher education, K–12, research institutions, healthcare, city and county governments, libraries and museums, and not-for-profits and public service entities. LEARN is a trusted partner and convener in these communities.

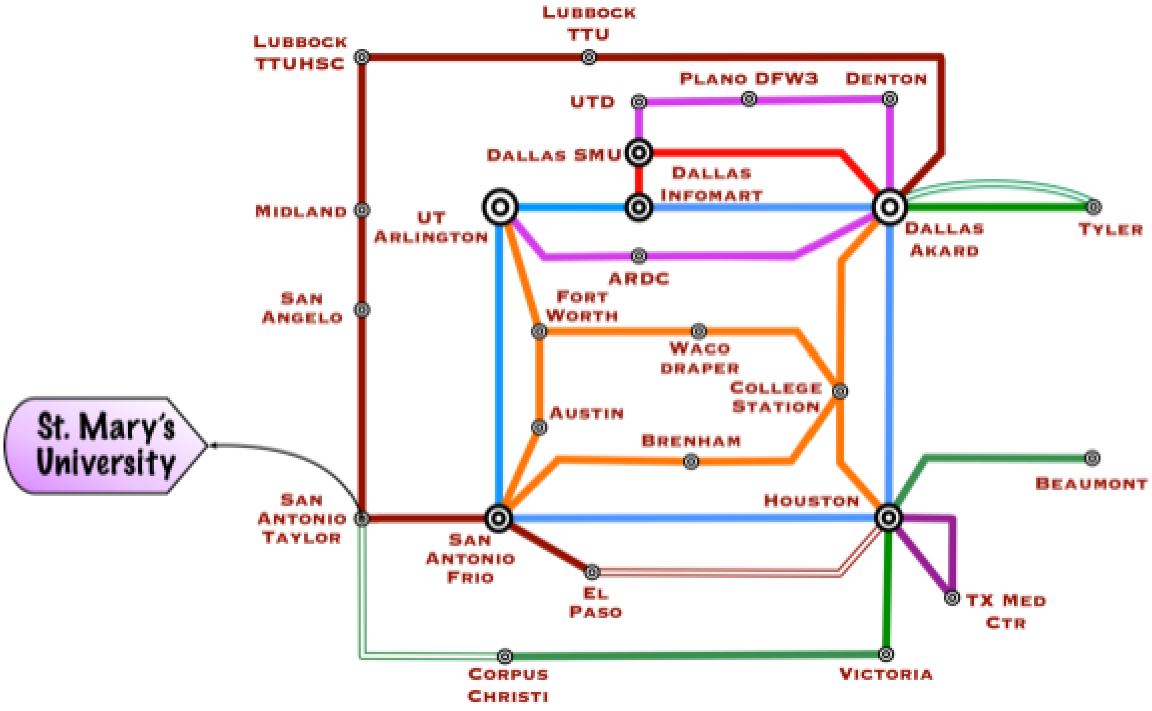

Figure 3: LEARN Connectivity Serving McClennan College

CC* Funding

In 2019, LEARN was awarded NSF Awards #1925553: "CC* Regional: Accelerating Research and Education at Small Colleges in Texas via an Advanced Networking Ecosystem Using a Virtual LEARN Science DMZ".

LEARN is partnering with national organizations in the implementation of this project. Projected impacts include increased opportunities for students to learn about and gain experience in advanced aspects of science, technology, engineering and mathematics (STEM) for which they might not otherwise have had an opportunity, for extension of the project to students and faculty at other campuses in Texas, and for the extension of the LEARN model to other regional networks and smaller campuses throughout the United States.

Objectives:

- Establish a small college collaborative environment within the LEARN community
- Improve network connectivity/services at each college campus for research and education
- Establish a network performance monitoring infrastructure
- Establish a means to facilitate the transfer of large data sets
- Deliver technical training to personnel at each campus
- Develop and implement an outreach program for informing/educating faculty, staff, and students at each college, and develop and disseminate project results