Trinity University Campus-Wide Deep Dive

May 29, 2019

Disclaimer

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor The Trustees of Indiana University, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California or The Trustees of Indiana University. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California, or The Trustees of Indiana University.

Trinity University Campus-Wide Deep Dive

Final Report

San Antonio, TX
May 29, 2019

The Engagement and Performance Operations Center (EPOC) is supported by the National Science Foundation under Grant No. 1826994.

ESnet is funded by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research. Benjamin Brown is the ESnet Program Manager.

ESnet is operated by Lawrence Berkeley National Laboratory, which is operated by the University of California for the U.S. Department of Energy under contract DE-AC02-05CH11231.

This is a University of California, Publication Management System report number LBNL-2001319[1].

[1] https://escholarship.org/uc/item/3db1k0hf

Participants & Contributors

Jim Bradley, Trinity University – CIO
Dr. Kwan (Kelvin) Cheng, Trinity University - Physics & Astronomy
Courtney Cunningham, Trinity University - IT
Gerardo Guzman, Trinity University - IT
Dr. Nicolle Hirschfeld, Trinity University - Classical Studies
Akbar Kara, LEARN
Neal Pape, Trinity University - IT
Gerno Reinard, Trinity University - IT
Jared Pack, Trinity University - IT
Dr. David Ribble, Trinity University - Academic Affairs
Amy Santana, LEARN
Dr. Jason Shearer, Trinity University - Chemistry
Peggy Sundermeyer, Trinity University - Academic Affairs
Dr. Ben Surpless, Trinity University – Geosciences
Dr. Yu Zhang, Trinity University - Computer Science
Jason Zurawski, ESnet

Report Editors

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

Contents

1 Executive Summary

In May 2019, staff members from the Engagement and Performance Operations Center (EPOC) and the Lonestar Education And Research Network (LEARN) met with researchers and staff at Trinity University for the purpose of a Campus-Wide Deep Dive into research drivers. The goal of this meeting was to help characterize the requirements for five campus research use cases and to enable cyberinfrastructure support staff to better understand the needs of the researchers they support. Profiled use cases included:

- Physics and Neuroscience
- Computer Science and Engineering
- Classical Studies and Archeology
- Biological Sciences
- Geosciences

Material for this event included the written documentation from each of the research areas at Trinity University, documentation about the current state of technology support, and a write-up of the discussion that took place in person.

The Case Studies highlighted the ongoing challenges that Trinity 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. These included:

- Need for upgrades to the current campus HPC system
- Lack of access to remote high-performance and high-throughput computational resources
- Lock of availability of local persistent storage
- Need for better approaches to facilitate data transfer to large facilities around the country
- Need for additional network connectivity to foster collaborations with external parties, beyond the campus boundary
- Lack of ability to share data with externally located collaborators

Trinity University and LEARN applied for an NSF award to help support upgrading the regional and campus networks, specifically to include a Science DMZ and monitoring equipment. An update to the state network is also underway, and details were discussed. As part of the overall review, it was determined that there was a need to identify and collaborate with regional or national providers for computational resources. Additional challenges were identified relevant to securing sensitive data, general cybersecurity approaches, and support for collaborations.

Action items from the meeting included:

1) LEARN, Trinity University, and EPOC will continue a discussion regarding network architectural needs if the NSF proposal is accepted. This consultation would involve specification of hardware requirements as well as best practices for operational soundness.
2) LEARN, Trinity University, and EPOC will begin a discussion about research storage, and ways this can be integrated into the scientific workflows of the Case Study researchers.
3) LEARN and Trinity University will finalize plans for the updates to the LEARN connectivity and peering arrangements (expected Summer 2019).

2 Process Overview and Summary

2.1 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 data use. By considering the full end-to-end data movement pipeline, EPOC is uniquely able to support collaborative science, allowing researchers to make the most effective use of shared data, computing, and storage resources to accelerate the discovery process.

EPOC supports five main activities

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

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

The Deep Dive process is based on an almost 10-year practice used by ESnet to understand the growth requirements of DOE facilities (online at https://fasterdata.es.net/science-dmz/science-and-network-requirements-review). The EPOC team adapted this approach to work with individual science groups through a set of structured data-centric conversations and questionnaires.

2.2 Campus-Wide Deep Dive Structure

The Deep Dive process involves structured conversations between a research group and relevant IT professionals to understand at a broad level the goals of the research team and how their infrastructure needs are changing over time.

The researcher team representatives are asked to communicate and document their requirements in a case-study format that includes a data-centric narrative describing the science, instruments, and facilities currently used or anticipated for future programs; the advanced technology services needed; and how they can be used. Participants considered three timescales on the topics enumerated below: the near-term (immediately and up to two years in the future); the medium-term (two to five years in the future); and the long-term (greater than five years in the future).

The Case Study document includes:

- ***Science Background—***an overview description of the site, facility, or collaboration described in the Case Study.
- ***Collaborators—***a list or description of key collaborators for the science or facility described in the Case Study (the list need not be exhaustive).
- ***Instruments and Facilities***—a description of the network, compute, instruments, and storage resources used for the science collaboration/program/project, or a description of the resources made available to the facility users, or resources that users deploy at the facility.
- ***Process of Science***—a description of the way the instruments and facilities are used for knowledge discovery. Examples might include workflows, data analysis, data reduction, integration of experimental data with simulation data, etc.
- ***Remote Science Activities—***a description of any remote instruments or collaborations, and how this work does or may have an impact on your network traffic.
- ***Software Infrastructure—***a discussion focused on the software used in daily activities of the scientific process including tools that are used to locally or remotely to manage data resources, facilitate the transfer of data sets from or to remote collaborators, or process the raw results into final and intermediate formats.
- ***Network and Data Architecture—***description of the network and/or data architecture for the science or facility. This is meant to understand how data moves in and out of the facility or laboratory focusing on local infrastructure configuration, bandwidth speed(s), hardware, etc.
- ***Cloud Services***—discussion around how cloud services may be used for data analysis, data storage, computing, or other purposes. The case studies included an open-ended section asking for any unresolved issues, comments or concerns to catch all remaining requirements that may be addressed by ESnet.

- ***Resource Constraints*—**non-exhaustive list of factors (external or internal) that will constrain scientific progress. This can be related to funding, personnel, technology, or process.
- ***Outstanding Issues*—**Final listing of problems, questions, concerns, or comments not addressed in the aforementioned sections.

At an in-person meeting, this document is walked through with the research team (and usually cyberinfrastructure or IT representatives for the organization or region), and an additional discussion takes place that may range beyond the scope of the original document. At the end of the interaction with the research team, the goal is to ensure that EPOC and the associated CI/IT staff have a solid understanding of the research, data movement, who's using what pieces, dependencies, and time frames involved in the Case Study, as well as additional related cyberinfrastructure needs and concerns at the organization.. This enables the teams to identify possible bottlenecks or areas that may not scale in the coming years, and to pair research teams with existing resources that can be leveraged to more effectively reach their goals.

2.3 Trinity University Campus-Wide Deep Dive Background

In May 2019, EPOC and Lonestar Education And Research Network (LEARN) organized a Campus-Wide Deep Dive in collaboration with Trinity University to characterize the requirements for several key science drivers. The Trinity University representatives were asked to communicate and document their requirements in a case-study format (see Section 3: Trinity University Case Studies). These included:

- 3.1 Campus Overview
- 3.2 Physics and Neuroscience Case Study
- 3.3 Computer Science Case Study
- 3.4 Classics & Archeology Case Study
- 3.5 Chemistry Case Study
- 3.6 Geosciences Case Study

A face-to-face meeting took place at Trinity University in San Antonio, TX on May 29th, 2019 (see discussion in Section 4 Discussion Summary). We document next steps in Section 5 Action Items.

2.4 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.

The Lonestar Education And Research Network (LEARN) is a consortium of 41 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 500 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 advanced research networks, providing Texas connectivity to national and international research and education networks, and enabling cutting-edge research that is increasingly dependent upon sharing large volumes of electronic data.

Trinity University was established in Tehuacana, Texas, in 1869, and moved to San Antonio at the invitation of the city in 1942. Founded on the vision of a few Texas pioneers who believed in the transformational powers of higher education, today Trinity serves students from around the world and empowers them with the tools of "a University of the highest order." Trinity University is one of the nation's top private undergraduate institutions, noted for its rigorous academic program, distinguished faculty, and beautiful residential campus near the heart of San Antonio.

3 Trinity University Case Studies

Trinity University presented five scientific use cases, and one campus technology overview, during this review. These are as follows:

- 3.1 Campus Overview
- 3.2 Physics and Neuroscience Case Study
- 3.3 Computer Science Case Study
- 3.4 Classics & Archeology Case Study
- 3.5 Chemistry Case Study
- 3.6 Geosciences Case Study

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

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

3.1 Campus Overview

3.1.1 Campus Background

Trinity University is a unique private, residential liberal arts institution in San Antonio, Texas. The student body includes approximately 2,300 undergraduates and 200 graduate students. It offers 42 majors and 57 minors among six degree programs. It has an endowment of $1.24 billion, the 85th largest in the country and unusual for such a small school. Forty percent of its students attend graduate school immediately after undergraduate school and 65% attend within five years. STEM fields, student research, and computing across the curriculum are emphasized. It was ranked as a "Best College 2019: Regional Universities" in the *U.S. & News Report* in November 2018. It is ranked 8th nationally for its science and lab facilities and has a recently built Center for Sciences and Innovation that focuses resources on student-faculty research partnerships.

3.1.2 Instruments and Facilities

Trinity University was awarded NSF grant funds[2] to purchase a Penguin Computing 45 node (1,512 CPU core) cluster to support large-scale, high-performance computing (HPC). Students and faculty sponsors are able to perform intensive computing tasks for a broad range of scientific research efforts spanning data science, mathematics, computer science, chemistry, geology, biology, and physics. Student researchers in these departments have access to considerable computational power that can be used to analyze large datasets generated through traditional laboratory experiments or through robust computational simulations.

This resource consists of 36 CPU nodes, 5 GPU nodes with 2 NVidia Tesla K80m GPUs, 1 memory node, 1 login node, 1 management node, and 1 data node with 211 Tb of attached storage. This resource enables Trinity University to effectively train and educate students on tangible aspects of large-scale computational projects and research endeavors.

3.1.2.1 Server / Network Data Center

Trinity University features two data center facilities: these are referred to as the *North Data Center* (which contains the majority of the campus infrastructure), and the *South Data Center*, which opened in October of 2019. This newly construstructed South Data Center is rated as a Tier 2 class data center, meaning that it has redundancy built in to accommodate for power outages. In particular the facility features online double conversion UPS and diesel generator secondary source.

The equipment in the older North Data Center is housed in a self-contained Liebert "Smart Row" with redundant air-conditioning and a fire suppression system. The

[2] https://www.nsf.gov/awardsearch/showAward?AWD_ID=1531594

network core room terminates the campus fiber networks and private fiber networks. External connectivity to the North Data Center consists of 1 Gb/s capacity available via a connection to LEARN, and is burstable up to 10 Gb/s. A second connection for "commodity" Internet is also available in the North Data Center via CenturyLink with a data rate of 2 Gb/s. The various buildings on campus are connected via a fiber optic backbone with a 2 Gb/s or more bandwidth to each building. Computers in each building have access to wired connections at 1 Gb/s.

3.1.2.2 Access Controls

Card swipe access and unsupervised 24/7 access to the Trinity University Data Center will only be given to individuals with an approved and demonstrated business need. Those individuals requiring infrequent or temporary access will be granted escorted access as needed.

3.1.2.3 Security Controls

Security is controlled via a proximity card reader system. All areas of the Trinity University Data Center and the network core room are under video surveillance 24/7 both internally and externally. The video feed is monitored by Trinity University Police Department.

3.1.3 Software Infrastructure

The Trinity University Information Technology Services division also manages and supports multiple student computer labs. There are 711 computer workstations in academic labs, 30 computer workstations in dormitory labs, and 9 virtual desktop images (VDI) available for student researchers. These workstations support either Windows or Mac OS clients, and run a broad set of data analytics and research software including:

- Statistical Analysis System (SAS)[3]
- Statistical Package for the Social Sciences (SPSS)[4]
- The R Project for Statistical Computing (R)[5]
- NVivo 11 Pro[6]
- Tableau[7]

3.1.4 Network and Data Architecture

The campus network, shown in Figure 1, includes external connectivity via two circuits: a 10Gbps link to a commodity provider (that is artificially limited to 2Gbps, with the potential to burst higher) and a 10Gbps link to LEARN. The commodity connection lands at the edge via a Campus operated Cisco 1000 series router, and the LEARN connection is handled by a LEARN provided Juniper that is then linked to the same Cisco 1000 series router. A set of Cisco 6500s provide the switching core,

[3] https://www.sas.com/en_us/software/stat.html

[4] https://www.ibm.com/products/spss-statistics

[5] https://www.r-project.org

[6] https://www.qsrinternational.com/nvivo/what-is-nvivo

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

and the network distribution layer (to individual buildings, floors, offices, etc.) is handled by Cisco 9300s and 3850s.

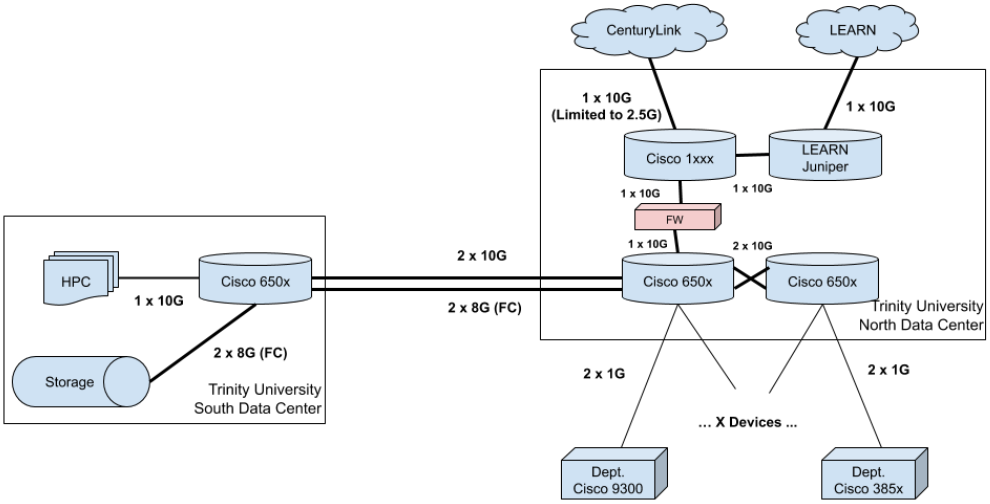

Figure 1 - Architectural diagram of the current Trinity University network.

3.1.5 Cloud Services

Trinity University maintains a small number of physical servers and hosts the Enterprise Resource Planning (ERP) system on premise. However, most enterprise services are now virtual Software as a Service(SaaS) with plans to migrate the majority of what remains to the cloud in coming years.

3.1.6 Known Resource Constraints

The IT division has a staff of 48, which includes a team of six network and server infrastructure experts. This small technical team manages two on premise data centers, the campus network, and research technology support services for Trinity University. Staffing numbers, staff skills and expertise in specific subject areas, and budget are known resource constraints.

3.2 Physics and Neuroscience Case Study

3.2.1 Science Background

Dr. Kwan (Kelvin) Cheng is the Williams Endowed Professor in Interdisciplinary Physics, Physics, and Astronomy at Trinity University. Most of his work centers on understanding the underlying physical mechanisms that cause neurodegenerative diseases, for example Alzheimer's. He also investigates gene-editing RNA/protein complexes through the combined use of experimental and simulation techniques, such as single-molecule imaging, cell imaging, molecular spectroscopy and multiscale molecular dynamics simulations. A part of this work involves extensive simulation on computational resources at Trinity, regional facilities at Texas Tech University (TTU), and national facilities including the Texas Advanced Computing Center (TACC).

The simulation models try to predict the dynamics and energetics of biomolecules, and to develop new understanding of the interactions between molecules. Input data for the models includes the output from X-ray Diffraction, Nuclear Magnetic Resonance (NMR), and cryogenic electron microscopy (cryo-EM) in the form of coordinates of atoms within the molecules that are deposited in the database, Protein Data Bank (PDB)[8]. PDB was established as the first open access digital data resource for biology and medicine, and is a leading global resource for experimental data. The outputs of the simulations are the predicted structures in PDB formats that are shared with the community for use by others working in the same biological system.

3.2.2 Collaborators

There is extensive collaboration done with other facilities that have computational and instrumentation resources available, along with other institutions that host similar research interests.

Dr. Alan Sill at TTU has assisted with the provision of HPC resources at Texas Tech University (TTU) on occasion, and this site has become widely used due to the availability of computation time.

TACC has numerous machines (highly parallelizable MPI-capable infrastructure) that fit the workflow profile for this research. The use of Lonestar was routine, although has not been primary due to long queue wait times for resources. The newer TACC resources have not been explored to date, but will be in the future via the connection with TTU.

University of Texas Austin also has several resources that are available for use, including several GPU-specific facilities. It is anticipated that GPU work will increase in the future, but for now is treated primarily as a prototyping platform.

[8] http://www.rcsb.org

This work is a collaboration with Dr. Pengyu Ren, a UT faculty member, who specializes in computational biomedical engineering.

Additional analysis work is being performed with Dr. Sara Y. Cheng at the University of California Berkeley. This does not involve exchange of large amounts of data.

Lastly, there is cross-highway collaboration performed with a research group at the University of the Incarnate Word in the field of molecular docking of multi-purpose drugs on various targets for drug discovery in Alzheimer's. Typically, the files for this collaboration are small enough that they can be shared via email.

3.2.3 Instruments and Facilities

The department has dedicated access to three resources specifically for GPU-analysis that have a total of 10 total GPUs available. The AMBER package, (described in Section 3.2.6 Software Infrastructure), can use GPUs and is being experimented with locally.

In addition to the GPU-based resources, the departmental hosts include traditional personal computer resources, each with multiple CPUs (12 in most cases). Most parallel work that can be run on the Trinity HPC system is first prototyped on the departmental resources before migration to larger resources. This development pipeline prevents wasting of HPC resources, either local or remote. During the prototyping process, checkpointing is routinely employed to prevent any complications due to a fault in the computation.

Lastly, there are some older-model iMac Pro workstations that are used for visual design and simulation work. The iMac machines are excellent for analysis of molecular structures, but are limited by a lack of main memory (RAM) resources and could use an upgrade.

The department makes extensive use of the Trinity University HPC resources when they are available. The main software package, GROMACS (described in Section 3.2.6 Software Infrastructure),uses MPI to run on multiple hundreds of processors, if available. This scalability can increase research productivity. Many of the codes within GROMACS are extremely aggressive with regards to parallelism, and have caused machine halts on occasion. The use of checkpoints has assisted on these occasions.

In the general case, the use of local resources is preferred to simplify workflows. The use of remote resources is typically only done when there is a benefit to the scalability, or there is a requirement to use more than what can be sourced locally. Use of the local resources has resulted in several publications.

3.2.4 Process of Science

Computational simulations that are performed in this research mainly involve the modeling of molecular dynamics and other computationally expensive operations of several macromolecules, .e.g, human beta-amyloid aggregates and RNA-editing spliceosomes from yeast. At present, we are actively investigating the disruptive behavior of beta-amyloid aggregates on lipid nanodomains, mimicking the cholesterol-enriched lipid rafts founds in cell membranes. Also, we are studying the bonding behaviors and binding affinity of a key regulatory protein (Dib1) in a large RNA/protein complex, representing the spliceosome at the pre-catalytic stage. In both cases, the input data includes PDB files that are the outputs from X-ray Diffraction, NMR or cryogenic electron microscopy (cryo-EM) experiments. It is routine to use summaries of this data, and on occasion compare and validate the experimental data on neuronal cell damage by neurotoxic proteins, beta-amyloid aggregates, and explain the stability and triggering process among molecules in large spliceosome complex, a gene editing/splicing machine consisting of RNA and protein molecules in yeast from ongoing molecular biology experiments.

Future goals include simulating and predicting new molecular structures of beta-amyloids, as well as gene-splicing RNA/complexes at other RNA-splicing stages in the splicing pathway. The simulation results may lead to new understanding of the process of the Alzheimer's disease and future therapeutic treatments. In addition, they hope to create new approaches to gene editing and splicing, which can be critical to understanding how genes are manipulated in cells in normal and pathogenic processes.

There are four stages of the pipeline include:

1. ***Design***: This typically involves pre-processing the input datasets to verify completeness. This may include steps to rebuild molecules or structural repair, for example, for broken RNA or proteins. For example, the un-resolved RNA or protein components from experimentally derived beta-amyloid and spliceosome complexes need to be rebuilt and reconstructed using molecular design tools found in AMBER and Chimera programs, and home-built energy minimization scripts
2. ***Simulation***: This is done on the local computational resources to start, and then on larger facilities as needed. For now CPUs are used. The investigators see growth via the use of GPUs, and are exploring possible migration of more of their infrastructure to use GPU resources. Currently, both the pre-production runs of beta-amyloid and spliceosome simulations are performed in local clusters in the physics computational facilities. Both atomistic and coarse-grained (CG) MD simulations, or multiscale MD simulations, will be performed. Due to the reduction in the number of atoms (~ four heavy atoms to one CG atom) as well as the reduced interaction forces among CG atoms, simulations up to a few microseconds are possible in our CG MD simulations, compared with only a few hundred nanoseconds in the atomistic MD simulations. Using the Trinity cluster, we have achieved ~ 20 ns/day for the

CG but only ~ 1 ns/day for atomistic simulations using parallel and multicore computing algorithms. With CG simulations, we can achieve better sampling, up to microsecond time scales, of the translational and rotational conformations of the interacting molecules in large macromolecular complexes, such as spliceosomes. However, detailed secondary structures, necessary for characterizing detailed bonding behavior, e.g., hydrogen bonding, cannot be resolved from the structures produced using our current CG model. Therefore, we propose using a reverse-mapping algorithm to convert a CG structure back to an atomistic structure. Further atomistic MD simulations, or relaxation of the atomistic structures, will allow us to obtain the secondary structures of interacting molecules. This new and advanced multiscale (CG simulations - to - reverse-mapping - to - atomistic relaxation) MD simulation protocol has successfully been tested and implemented in my lab for the large spliceosomes. We therefore aim at obtaining more than 2 microseconds of simulation data for each spliceosome complex.

3. ***Analysis***: The output data sets from the simulations can be large and vary in size from 2-50 TB in the current form. Due to the data volume, the use of local resources has been preferred since it is simple to transfer data locally to lab resources for the visualization steps. For example, in our spliceosome simulations, based on the structures obtained from our microsecond multiscale simulations, we will examine the binding affinity among molecules in Dib1 and its neighbors. Here, information rich binding energy, residue-contact map and hydrogen-bonding map will be used to quantify the binding affinity. To calculate binding energies, we will compute the pairwise interactions between Dib1 and its neighboring RNA and protein chains. Both long-range and short-range intermolecular electrostatic and non-electrostatic, or van der Waals, interactions will be separately determined. The time-evolution of binding energy will provide the kinetics of interactions and the equilibrated interaction energy between Dib1 and each neighbor. For residue-contact map analysis, time-averaged minimum distances between each amino acid residue of Dib1 and each neighboring amino acid residue of protein or nucleotide residue of RNA will be computed to generate a 3D residue-contact map. In the 3D residue-contact map, the minimum distance (z-axis) will be color coded, with the residue location of each partner of Dib1 (y-axis) versus the residue location of Dib1 (x-axis). Similarly, the 3D time-averaged hydrogen-bonding map can also be generated. Here, the probability or propensity of hydrogen bonding between each atom in Dib1 and each atom in each Dib1 neighbor will be calculated based on distance and angle thresholds between the hydrogen donor and acceptor atoms. Again, we will plot this on on 3D map with the hydrogen bonding propensity (pseudo color coded as z-axis), and the residue location of each partner of Dib1 (y-axis) versus the residue location of Dib1 (x-axis). Note that that hydrogen-bonding map is a subset of the contact-map since not all residue-residue contact pairs form hydrogen-bonds. The hydrogen-bonding contains rich information about the chemical specificity of the residues and involves formation of multiple

hydrogen bonds, e.g., single donor to multiple acceptors, in each Dib1-neighbor pair. We will use Molecular Mechanics Poisson-Boltzmann Surface Area (MMPBSA) method to determine the "free energy" of binding (involving both enthalpic, or potential energy, and entropic, or available conformations, contributions) between Dib1 and its neighbors. This new tool was originally intended for studying small molecule, e.g., drug compound, binding to protein targets for drug screening and discovery. Recently, this MMPBSA method has also been applied to study large protein-protein and protein-polymer interactions. We have installed and setup the necessary algorithms and libraries of functions to our local servers and also to our Trinity Cluster.

4. ***Visualization***: 3D models of the molecular structures are created from the simulation and analysis steps. Due to the need for low-latency, the large data sizes, and the higher-performance graphics requirements, it is not anticipated that cloud computing can be used for this step. Currently, we are using VMD, Pymol and Chimera as the major molecular visualization programs to study the interactions among atoms and molecules, in addition to conventional 2D and 3D plots of analyzed data. For example, for the case of beta-amyloid interactions with lipid nanodomains, we need to visualize the structure and dynamics of specific protein amino-acid residues with the liquid-ordered (Lo) , liquid-disordered (Ld) or Lo/Ld comain boundaries using VMD or Pymol. For the case of spliceosomes, rendering of interactions among RNA-bases and protein amino-acid residues, particularly the dynamic breaking and forming of hydrogen-bonds are routinely examined using Pymol and VMD.

The process is easily repeatable once the basic pipeline has been established and can be replicated for multiple runs.

3.2.5 Remote Science Activities

Research activities are primary performed internal to Trinity. On occasion the use of externally located resources may occur in the form of remote computation, or downloading remote data sets. Computation has been used at the following facilities in the past, and can continue when there is a need to run larger jobs that the local HPC at Trinity cannot support:

- Texas Advanced Computing Center (TACC)
- Texas Tech University
- University of Texas Austin

Remote data access, in the form of reference information that can be compared against research products that Trinity develops, are retrieved from the Protein Data Bank (PDB). This involves the use of an online portal (e.g. HTTP downloads), and is not a common activity.

3.2.6 Software Infrastructure

The group focuses on the use of “off the shelf” software when applicable, as there are no local experts that want to spend time writing MPI codes manually.

GROMACS[9] is a computational package for molecular dynamics that simulates the Newtonian equations for motion for systems with hundreds to millions of particles. It is primarily designed to be used to model biochemical molecules like proteins, lipids and nucleic acids that have many complicated bonded interactions, but since GROMACS is extremely fast at calculating the nonbonded interactions (that usually dominate simulations) many researchers are also using it for research on non-biological systems, such as polymers.

Assisted Model Building and Energy Refinement (AMBER[10]) refers both to a set of molecular mechanical force fields for the simulation of biomolecules and a package of molecular simulation programs.

GROMACS and AMBER are used extensively for research purposes. Having assistance from central IT to manage multiple versions would be welcomed.

The visualization of the modeling results is done using several tools:

- UCSF Chimera[11]: an extensible program for interactive visualization and analysis of molecular structures and related data, including density maps, supramolecular assemblies, sequence alignments, docking results, trajectories, and conformational ensembles.
- PyMOL[12]: a molecular visualization system. This is a commercial product, but is used via an educational license program.
- Visual Molecular Dynamics (VMD)[13]: a molecular visualization program for displaying, animating, and analyzing large biomolecular systems using 3-D graphics and built-in scripting.

All of these tools require a high-performance workstation, with enhanced graphical capabilities. This can be a higher-end Apple PC, such as an iMac Pro, or a Windows or Linux machine that is outfitted to support more commodity use cases but with a graphics card used to support video gaming. The internals of the latter are more scalable and support hardware upgrades over time, and traditionally cost less. All of the software at analysis software at Trinity is run on iMac Pro hardware, thus there are concerns about migration to Mirosoft-based operating systems; further testing of this would be required.

[9] http://www.gromacs.org

[10] http://www.gromacs.org/Documentation/Terminology/Force_Fields/AMBER

[11] https://www.cgl.ucsf.edu/chimera/

[12] https://pymol.org

[13] https://www.ks.uiuc.edu/Research/vmd/

3.2.7 Network and Data Architecture

For the main components of this section, please see Section 3.1.4 Network and Data Architecture

3.2.8 Cloud Services

Several approaches to use Cloud services for data storage have taken place. An attempt was made to use Google Drive to facilitate sharing, but it lacked several key features, including a detailed timestamp record for file changes was problematic. Likewise, the interface and usability of Dropbox did not scale for some data sets.

Given these challenges, local storage has been preferred. This is done using a combination of external hard drives and a collection of network-connected resources that share a common filesystem. It is anticipated that between 4 and 40TB of storage will be required in the near term, and the current approach will focus on the use of portable external hard drives.

3.2.9 Known Resource Constraints

The current university computational resources are sufficient for the research activity described above, however there are long-term needs to consider. Some work will naturally scale toward GPU computing, but local GPU resources are currently limited. However, GPU resources could be purchased/installed into local analysis machines, or a larger GPU cluster could be considered.

Storage is not currently a critical need, but will be soon. The department is using a set of external hard drives (10 total, of around 4TB each) to manage data sets. This is not a scalable solution, and the cost/complexity of managing this approach is likely to break down in the future.

There are no current pressing external-facing data needs, primarily because storage and processing are done locally. As users on campus gravitate toward the use of external resources (local & national computation centers, cloud resources, etc.) the network will become more critical. Prior use of resources at TACC did not run into transmission issues due to the small data transfer requirements, however the data set sizes are expected to grow in the coming years.

3.2.11 Outstanding Issues

The availability of HPC resources on the Trinity University campus scales to the current needs of a wide variety of researchers but are not predicted to be sufficient for future needs of the Physics and Neurology Departments. There are plans to augment the current computational resources, and to explore a wider deployment of GPUs.

Availability of campus storage resources that are both connected to the computational resources and available for internal/external data transfer capabilities could greatly assist with the management of research data: both internal tot the campus, and between the campus and external entities.

3.3 Computer Science Case Study

3.3.1 Science Background

Computer vision is an interdisciplinary scientific field that deals with how computers can be made to gain high-level understanding from various sources of information. Dr. Yu Zhang's team uses digital images and videos to conduct research in computer vision and also provides software and services to multiple departments and disciplines in image processing and image recognition projects.

A particular aspect of Zhang's research involves the study of Adaptive Filter Pruning of Convolutions (Ad-PC), a new inference algorithm based on convolutional neural networks (CNN). The algorithm selectively prunes filters in order to minimize the computational cost of using deep neural networks. This work is able to speed up computational time and reduce the need for extensive processing of the full set of input data.

3.3.2 Collaborators

Collaboration is primarily done within Trinity at this time.

3.3.3 Instruments and Facilities

Research is performed on a mixture of local computational resources (e.g. PCs and workstations) as well as the Trinity University campus HPC resources. This use consists of acquiring time on various nodes, loading in data, and running computational codes across the data set. The Ad-PC research uses multiple datasets for training, including:

- CIFAR-10 consists of 60,000 images broken into 10 classes. Each class contains 6,000 images that are 32x32. 50,000 images are used for training and 10,000 images are used for testing.
- CIFAR-100 is very similar to CIFAR-10, but has 100 classes. There are 500 training images for each class and 100 test images for each class.
- ImageNet-12 contains over 1.2 million images from 1000 classes. The validation set is 50,000 images. The images in this dataset are 224x224 and are generally more challenging to classify than those in the CIFAR-10 and CIFAR-100 datasets.

Depending on the size of the data, multiple variations of different CNN architectures are used. For the CIFAR-10 and CIFAR-100 datasets, we employ the simplest network, Visual Geometry Group (VGG)-9, a commonly used deep learning neural network model. For ImageNet-12, we use AlexNet, which has 5 convolutional layers and 3 fully-connected layers. This is the first convolutional network that performed very well on ImageNet. It was originally distributed across two GPUs, but we currently only use it on one. We also use ResNet-18 with the ImageNet-12 data. ResNet-18 is the largest network architecture and employs wide residual blocks that allow the network to be very deep. It is broken into blocks that contain multiple layers. Blocks have skip connections between them that add a block's input to its output.

3.3.4 Process of Science

CNNs have become commonplace in computer vision applications, including image classification, instance segmentation, pedestrian and car detection, and object localization. Over the last few years, the networks for all of these applications have grown deeper, with a significant increase in parameters and convolution operations. Such large networks have significant inference costs, especially when used with embedded sensors or mobile devices where computational and memory resources may be limited. For these devices, computational efficiency is a critical enabling factor. In fact, any device or service that has time constraints could potentially benefit from a small reduction in accuracy if it comes with a significant improvement in inference time.

The aforementioned projects that Ad-PC is able to analyze is typically 1000-10,000 images, or 30GB-300GB of data. Such analyses requires access to HPC frameworks due to computational expectations. Therefore, data and results will be transferred between local machines in the lab and the campus HPC resource. At this time, there is no work to explore the use of remote HPC resources.

3.3.5 Remote Science Activities

At this time, all research work is done local to Trinity University.

3.3.6 Software Infrastructure

Software that is developed as a result of this research is curated and managed by the department.

HPC codes are run on university computing resources, and normally includes use of MPI APIs.

3.3.7 Network and Data Architecture

The existing campus network and technology support are sufficient for the research needs of the Computer Science department. For the main components of this section, please see Section 3.1.4 Network and Data Architecture

3.3.8 Cloud Services

At this time, there is no use of cloud-based computation or storage. Future needs may see exploration, but the cost of these resources vs. the use of free (albeit limited) local resources is more attractive.

3.3.9 Known Resource Constraints

With fast development of computer vision technologies and the increase in the number of images, the demand for a fast and reliable network connection has been increasing steadily. Data sets (e.g. those that are produced/curated by others) must be brought locally for processing work.

In addition, due to the complicated data structures and lack of a uniform analytical approach, multiple methods and CNN structures will be needed for data analyses and different CNN settings will be tested for evaluation and validation purposes. Local data transfer, from different desktop/laptop computers in different departments and the HPC, will be routinely conducted throughout the study periods demanding fast and stable local network connections.

The existing HPC resources (which are less than a teraflop of computing power), are becoming seriously taxed due to the increasing use by local user groups. Collectively, this points to a need to augment existing HPC to keep up with demand. In some cases, students in Computer Science, who require access to machinery to perform thesis work, must wait behind other users that may use the majority of the resource over a long period of time. This has created a wait list that can be as long as days or weeks.

Storage resources are currently sufficient for the data sets that are explored, but will not scale into the future. Investigation of group storage, shared in the department, would greatly assist research efforts.

3.3.11 Outstanding Issues

The need for additional campus computational resources is critical. The Office of Research is exploring augmentation of the existing cluster. The current HPC resource was funded in 2015 by NSF, and based on recent growth, Trinity's capital development fund supported five additional GPU nodes in 2019. It is expected that an additional update will be needed in about 3 years based on this growth pattern.

Other computational needs, for example use of GPUs, are being explored by affiliated user groups, such as the Department of Physics. GPU resources could be leveraged by the Computer Science Department, after some initial exploration in running the neural network software.

3.4 Classics & Archeology Case Study

3.4.1 Science Background

Dr. Nicolle Hirschfeld is a professor of Classical Studies Department at Trinity University. With her colleagues from inArkansas, Indiana, Toronto, Istanbul, and Sydney, her team is studying objects raised from a shipwreck that sank on the southwestern coast of Turkey sometime in the late 13th or 12th centuries BCE. The ship was first excavated in 1960, again in the 1980s and 90s, and most recently in 2010. The ship and its crew originated and sailed within the Levantine/southern Anatolian orbit, for the purpose of dealing in copper and tin and recycled bronze.

Excavation on site and study of the artifacts after recovery have focused on:

- Mapping and documentation of the ship wreckage
- Removal of artifacts from the site for conservation and research
- Documentation and photography of all of the artifacts
- 3D modeling and analysis of selected artifacts
- Microscopy, bulk composition analysis, and isotopic analyses of the metal artifacts to determine their sources and methods of production
- Petrographic and neutron activation analyses of ceramics

The goal of this study is to identify the origins of the cargo and to better understand the production processes and recycling of metals and metal objects, their trade patterns, and how these types of artifacts related to the development of cultural systems.

Due to the inherent location of the research, close coordination with the government of Turkey is required. This involves various permissions to get access to the artifacts for documentation, study, and analyses, especially when sharing across international borders.

3.4.2 Collaborators

Dr. George Bass, University of Pennsylvania and Texas A&M University, was the original PI for the work through the 1990s, but has since retired from active oversight. Trinity University, through Dr. Hirschfeld, continues to work with the Institute of Nautical Archaeology at Texas A&M University.

The current research team consists of:

- Joseph Lehner, University of Sydney, and Emre Kuruçayırlı, Boğaziçi University: metallography (study of the physical structure and components of metals through the use of microscopy and compositional analyses) of the ingots and ingot fragments
- Nicholas Blackwell, Indiana University Bloomington: bronze artefacts
- Bartlomiej Lis, Fitch Laboratory, British School at Athens: ceramics
- Joan Aruz, The Metropolitan Museum of Art, New York: seals
- Lucas Proctor, University of Connecticut: organics

The excavation and conservation team during the lifespan of the project includes:

- George Bass, University of Pennsylvania and the Institute of Nautical Archaeology, Texas A&M University
- Cemal Pulak, Institute of Nautical Archaeology, Texas A&M University
- Harun Özdaṣ, Dokuz Eylūl University, Turkey
- Tuba Ekmekçi, Asu Selen Özcan, and Esra Altınanıt Biçer: Bodrum Research Center, Institute of Nautical Archaeology

Many scientists from various laboratories and institutes in Turkey and outside the country have participated or are actively participating in analytical studies:

- Radiocarbon Laboratory, University of Pennsylvania (publ. 1967): C-14 date
- Archaeological Research Laboratories, Oxford (publ. 1967): spectrographic analyses of ceramics
- Zofia Stos, Istotrace Laboratory, Oxford (publ. 2009): lead-isotope analyses
- Halford Haskell, Southwestern University, TX with Peter Day, University of Sheffield (publ. 2011): petrographic analysis
- Yuval Goren, Ben Gurion University of the Negev (2006 - present): petrographic analysis
- Lina Kassianidou and Lente van Brempt, Archaeological Research Unit, University of Cyprus (2013): initial evaluation and pXRF
- Gülsu Ṣimṣek and Barıṣ Yağcı, KUYTAM (Surface and Science Technology Center), Koç University, Istanbul (2017- present): Argilent 7700x ICP Mass-Spectrometer
- Moritz Jansen (University of Pennsylvania & Deutsches Bergbau Museum, Bochum, Germany) and Sabine Klein (Goethe University, Frankfurt, Germany), 2018-present: lead and copper isotope ratio research

Creation and refinement of 3D models of ingots and ingot fragments, begun in 2017 and continuing to this date, are being performed by collaborators at the University of Toronto, the University of Arkansas, the University of Sydney, and (in 2017-18) the University of Central Florida. Members of this collaboration perform this modeling locally at their facilities, using data such as photos, videos, and measurements that were shared from on-site observations.

3.4.3 Instruments and Facilities

Mapping: Maps produced for the 1960 excavation were based on a photographic montage of the seabed. Objects were located on the general plan by means of triangulation from set benchmarks using measuring tapes and plumb bobs.

In 2010, the team began using the program Site Recorder[14], which can account for variability in vertical distance as well as the distortion at the edges of images. This same program also allowed direct tape measurements to be taken to calculate the

[14] http://www.3hconsulting.com/ProductsRecorderMain.html

relative positions of data points, eliminating the need for plumb bobs and right angles, which are difficult to use in the strong bottom currents. Once the items were drawn or photographed and measured, the individual lots were plotted relative to the datum points using Photo Modeler[15], which extracts 3D measurements from photographs taken with an ordinary camera. Finally, Rhino[16], a computer graphics and computer-aided design (CAD) application, is used to consolidate photomosaic, datum points, and lot distributions into a single map-image. The result is a 2D map showing the relative distribution of artifacts and topographical elements on a horizontal plane only. The current workflow involves the use of personal (e.g. laptop) computers by research staff.

Artefacts: The drawings and photographs produced for the 1967 publication of the original excavation have all been digitized. Drawings and digital photographs of the objects recovered in the 1980s, 90s, and 2010 are currently being completed. Efforts to capture digital video and still pictures are continuing with the aim of creating 3D models of the ingots, selected ingot fragments, and perhaps the bronze objects. The processing of these images is done external to Trinity University, but ideally the data itself will be available within a repository hosted by Trinity in the future.

3.4.4 Process of Science

This research uses three primary scientific work-flows, as outlined below. Of interest are the aspects of:

- Field Work: Physical travel to the site for in-situ research activities
- Sample Analysis: Analysis of material samples via external partners
- 3D Imaging: Creation of artifact 3D models using point clouds developed using photogrammetry and structured light scanning

So, for example, an ingot fragment currently stored in the conservation lab of the Bodrum Research Center (Turkey) is conserved, catalogued, photographed, scanned, and sampled on site. Samples are sent to labs and Istanbul and Bochum (Germany) for metallographic analyses. 3D imaging is completed by team members working in Toronto, Arkansas, and Australia.

The field team uses a series of software packages to assist during research activities, namely to organize the forms of research data. Currently FileMaker[17] is used for this process, although there are limitations to the platform that are causing usage issues for the group. Available disk space, and a fear that a proprietary system is primarily responsible for housing years of work, are challenges to address. In the general case, a Trinity University managed resource that would facilitate storage, a web portal, and access for internal and external collaborators) would be useful for research needs. Currently, due to the complexity of sharing of large amounts (currently two terabytes; estimated five by end of project) of research data, hard drives are physically carried or shipped to collaboration sites. Attempts to use

[15] https://www.photomodeler.com/index.html
[16] https://www.rhino3d.com/
[17] https://www.filemaker.com

Google Drive or Dropbox for data management and sharing have proven unwieldy (protracted file transfer time) or unsuccessful (data corruption).

3.4.4.1 Field Work

The underwater excavation aspect of this project was completed in 2010. Current fieldwork takes place where the objects are currently stored: in the Bodrum Research Center of the Institute of Nautical Archaeology and in the Museum of Underwater Archaeology, also located in Bodrum, Turkey. Due to the laws surrounding the curation of antiquities in Turkey, it is not possible to export any of the artifacts. Thus trips to Bodrum are a necessity, though they are resource-intensive. Typically 4-6 members of the research team come to Bodrum every summer and stay on site for up to two months. It is expected that there are still several more years of site visits to perform.

Research teams make periodic trips to Bodrum to perform a variety of activities on the physical components, including:

- Photography and video
- Measurements and observations
- Collecting samples for analysis

3.4.4.2 Sample Analysis

Samples from the copper, bronze, ceramic, and organic artifacts are routinely subject to analysis to identify their composition or structure. Typically, the workflow is as follows:

1. Samples collected at physical site
2. Samples sent to analysis facilities, which involves government approval is they leave Turkey
3. Analysis performed
4. Results sent back to collaborators, typically as reports or graphs

Due to the commoditized nature of this work, the group does not maintain their own equipment and typically works with external sites for this analysis. Raw data from the instruments is typically not needed and the data volume of the reports is low.

3.4.4.3 3D Imaging

The on-site research team has been experimenting with a new pipeline for research based on photography and videography, namely creating digital models of the artifacts. The typical workflow for this process is:

1. On-site research team takes photographs and videos of the artifacts to be modeled.
2. Due to the size and complexity of sharing these electronic resources, physical media is mailed to collaborators, for example at the University of Toronto and University of Arkansas.
3. The research teams make wire frames and point models based on the digital images and videos.

4. The resulting models are shared back with Trinity University and other collaborators via Google Drive or by shipping hard drives.

Due to the data intensive nature of this workflow, it is highly desirable to make the data available via mechanisms that expose storage and are capable of high-speed data transfer. This would facilitate more collaborators and an easier pipeline for the work to be performed.

3.4.5 Remote Science Activities

Please see Section 3.4.4.

3.4.6 Software Infrastructure

There are a number of software packages that are utilized during the process of research. These include:

- Site Recorder[18]: Used for archeology
- FileMaker[19]: Used for archiving and accessing data
- Microsoft Word and Excel
- Adobe Photoshop and Illustrator
- 3D Modeling:
    - Artec Studio 11 and 12[20]
    - X-Rite ColorChecker[21]
    - Agisoft PhotoScan (now known as Agisoft Metashape)[22]
    - MeshLab[23]
    - CloudCompare[24]
    - GeoMagic Design X[25]
    - Autodesk 3D Max[26]
    - Photo Modeler[27]
    - Rhino3D[28]

Most of these software packages are run on personal computers (laptops, etc.) when in the field. The remainder is done using machines local to Trinity, or at partner institutions; analysis software for 3D imaging and photo processing falls into this category.

[18] http://www.3hconsulting.com/ProductsRecorderMain.html
[19] https://www.filemaker.com
[20] https://www.artec3d.com/3d-software/artec-studio
[21] https://xritephoto.com/colorchecker-classic
[22] https://www.agisoft.com
[23] http://www.meshlab.net
[24] https://www.danielgm.net/cc/
[25] https://www.3dsystems.com/software/geomagic-design-x
[26] https://www.autodesk.com/products/3ds-max/overview
[27] https://www.photomodeler.com/index.html
[28] https://www.rhino3d.com/

3.4.7 Network and Data Architecture

The work presented in this case study does not utilize special network infrastructure beyond what is provided on the campus. The primary mechanism for data sharing remains physical shipment of hard drives to collaborators. Sharing of data via cloud services has not worked in previous attempts (see Section 3.4.8 Cloud Services).

3.4.8 Cloud Services

Cloud services for data storage and sharing are not used frequently.
Google drive has been problematic due to the download and upload times, as well as corruption or data loss when compressing files. In particular, several GB image may take hours to upload when at an international site. Downloading, when located at a University in the U.S. goes faster, but is challenging for collaborators at international locations. Investigation has shown that the location of cloud infrastructure impacts performance considerably. Lastly, managing identities of collaborators (either personal, or those affiliated with a university) is problematic.

However, the continual use of hard drives is unsatisfactory because they are prone to failure and must be shipped. In addition, a shared central archive would help with the distributed nature of this collaboration. A centrally located storage location is desired, but is unclear if the cloud space can provide this. There are plans to test this connectivity and usability with several cloud providers (e.g. Amazon S3 and Microsoft Azure).

3.4.9 Known Resource Constraints

The most challenging aspect of the current research involves collaboration around shared data. In particular:

- There is not a central location for data storage, thus it becomes stored with individuals in an offline fashion.
- There is not a central tool used for data storage and curation, thus everyone gravitates to a different solution.
- Wide area data transfer is problematic, thus people rely on physical shipment of research data .
- Collaboration space is large in terms of distance and number of collaborators, thus a federated identity system would assist with sharing between necessary parties.
- Backups of data are not routinely performed.

This research work would benefit most from a single, centralized storage location with the following attributes:

- “Open” format for sharing, not locked to a proprietary format.
- Scalable in terms of size.
- Availability to share location with a growing set of collaborators.
- Ability to transfer/receive data with high performance tools.

All of the above items would be desirable for future funding solicitation responses, as entities like the NSF require a sound data management plan.

3.4.11 Outstanding Issues

To summarize, the largest area of friction involves the growing volume of research data, and the challenges in storing, sharing, and curating. As a whole, these challenges are not different from other departments, and could benefit from a Trinity University solution that scales to other parties:

- Centralized storage, with allocations for faculty. Scalable as needs grow.
- Data transfer hardware and software, to integrate with other remote collaborators, for example Globus
- Integrated with campus-wide federated identity to facilitate sharing with affiliated collaborators
- Development of a portal to share research data with external entities

With these items in place, the research group could sunset the following activities:

- Reliance on FileMaker as source of truth for research data
- Eliminate mailing physical media to collaborators
- Create a central repository of project research data, and eliminate copies being curated by individuals

3.4.12 Contributing Authors

- Nicolle Hirschfeld – Professor, Classical Studies, Trinity University
- Joseph W. Lehner – Australian Research Council Discovery Early Career Award Fellow, University of Sydney
- Samuel Martin – PhD candidate Environmental Dynamics, University of Arkansas
- Dominique Langis-Barsetti – PhD candidate, Near and Middle Eastern Civilizations, University of Toronto

3.5 Chemistry Case Study

3.5.1 Science Background

Research performed by the Shearer group centers on understanding the interplay between the metal-ligand coordination environment, geometric and electronic structure, and reactivity in metal containing systems of biological and industrial importance. We are primarily interested in systems containing late first-row transition metals (Co, Ni, and Cu) coordinated by thiolate ligands.

This work features four main components:

- Creation of models via simulation
- Experimental work using remote X-ray spectroscopy facilities
- Experimental work using locally placed optical spectroscopy resources
- Analysis of experimental results using local computation

The "Creation" component is conducted using a mixture of computational and software resources that are local to Trinity University. Both "Experimental" components may involve travel to a variety of experimental facilities that provide "white light" X-ray sources - some of these are located domestically, others are international. Local spectroscopy tools located at Trinity, thus their data is produced and managed locally. Both Experimental components do not require complex or time-consuming data analysis, and are achieved using locally available computation.

The Analysis component can emphasize different technological support strategies for the research group and for Trinity University. Some are readily available, others are critical for support into the future:

- Access to computational resources that are used for the construction and manipulation of electronic structure calculations.
- A data transfer mechanism that can support the ingress and egress of research data from experimental facilities.
- Ability to store, curate, and share current and historic research data.

3.5.2 Collaborators

Collaboration is not formally structured through an established virtual organization, but has grown organically over time. A representative list of known collaborators includes:

- University of Florida
- Cornell University
- Ewha Womans University, Seoul, South Korea

Collaboration typically involves sending samples to Trinity University for spectroscopy work. Trinity researchers will then apply for allocation time at remote X-ray spectroscopy facilities, both domestic and international. During allocated time, samples will be prepared, manipulated, and cataloged. Due to limitations in data transfer capabilities, data results from X-ray spectroscopy will be transported manually using portable storage back to Trinity for analysis. After analysis, reports on samples are shared. The raw data from the analysis process can be shared but is often not requested.

3.5.3 Instruments and Facilities

Trinity researchers apply for allocation time at remote X-ray spectroscopy facilities around the world, including:

- Advanced Photon Source (APS), Argonne National Laboratory, Lemont IL[29]
- Cornell High Energy Synchrotron Source (CHESS), Cornell University, Ithaca NY
- National Synchrotron Light Source II (NSLS2), Brookhaven National Laboratory, Upton NY
- Stanford Synchrotron Radiation Lightsource (SSRL), Stanford Linear Accelerator Laboratory (SLAC), Palo Alto CA
- Canadian Light Source, Saskatchewan, Canada
- Diamond Light Source, Rutherford Appleton Laboratory (RAL), Oxfordshire, United Kingdom
- European Synchrotron Radiation Facility (ESRF), Grenoble, France

The typical workflow for use of any of these facilities are as follows:

1. Apply for research time
2. Prepare samples for transfer
3. Travel to facility
4. Operate during beamline time
5. Collect data via local data transfer to portable media
6. Travel home
7. Perform local processing

Trinity researchers have explored the use of remote data transfer, but have hit a number of barriers, including:

- Lack of facilities at Trinity to accept bulk data transfers from light source facilities (e.g. 100s to 1000s of small files in nested directories). In order to be able to do this, Trinity University would need to deploy a machine with software to facilitate transfer and had significant local storage.
- Security profile required by facilities. Some facilities, for example NSLS2, do not allow network-based transfers of data due to their security policy.
- Need for specific training. Some facilities, for example the Canadian Light Source, require that training be logged before electronic transmission of data is allowed.
- The facilities require only portable media be sued, for example, Rutherford.

[29] https://www.aps.anl.gov

These factors, coupled with the fact that data sets are still plausibly managed via portable media, limit the use of technology in this space. The researchers do note that the use of international facilities can pose a unique challenge with regards to clearing customs when returning. Prior instances of traveling with removable media has aroused suspicions, resulting in additional security and inspection policies being enforced.

After visitation and use of remote facilities, the data is operated on locally. Data sets are typically archived in three locations, on Trinity University Provided Office PC, a PC in the laboratory, and on a home (private) PC.

Data set sizes are not significant, and may be KB to MB consisting of many sets of small files.

3.5.4 Process of Science

The Shearer group utilizes a small-molecule modeling or mimetic approach wherein a small synthetic model compound designed to replicate some aspect of the larger system is prepared, referred to as a Mimic. The Mimic is then subjected to detailed reactivity of spectroscopic studies. By altering attributes of the synthetic system in a logical fashion one can tease out what features of the larger synthetic or industrial catalyst are important to the system as a whole.

Much of this work involves probing the metal-sites in these Mimics by X-ray spectroscopy. Because of the need for a bright "white light" X-ray source, much of this work is performed at external facilities.

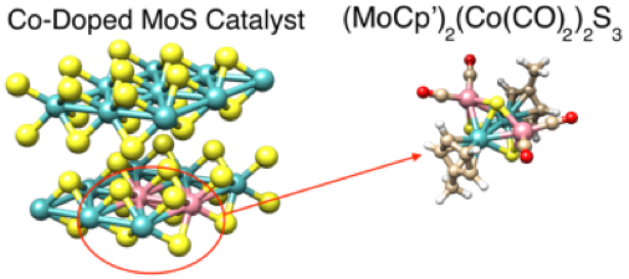

Figure 2 - Structures of cobalt-doped molybdenum sulfide catalyst (left) and model compound (MoCp')2(Co(CO)2)2S3 (referred to as Compound 1) on the right.

In addition to experimental work, the Shearer group also performs a large amount of computational and theoretical studies. Key to this theoretical work is the use of high-level electronic structure calculations. To give an idea of the types of systems the Shearer group examines computationally, Figure 2 shows the metal-sulfide cluster $(MoCp')_2(Co(CO)_2)_2S_3$ (1) .

Compound 1 is a Mimic for cobalt doped MoS, which is an industrial catalyst that removes sulfur from fuel stocks. The Shearer group is using Component 1 to understand how cobalt doping into MoS influences the electronic structure and subsequent reactivity of the catalyst. Key to this effort are high-level electronic structure calculations. Although Component 1 is a relatively small cluster compound, it is fairly complex from the standpoint of an electronic structure. Spectroscopically, Component 1 behaves as a ground state singlet. However, in order to achieve convergence of 1 as a singlet using a hybrid functional with a moderate sized basis set (PBE0/def2-tzvp) required application of a Fermi-like occupation number to the MOs, strongly indicating that Component 1 is highly multiconfigurational in nature. In fact, multi-reference calculations are required to describe Component 1.

A typical workflow for the research, after data collection at the remote facility, is as follows:

1. Local processing consists of PCs located in laboratory
    a. Some Linux, some Microsoft Windows
2. Analysis takes anywhere from 10 minutes to months
    a. Workflows are capable of being parallized, but are exclusively run on machines with a small number of cores or processors. Experience has shown a point of diminishing returns after 16 processors.
3. Analysis is performed with Orca (see 3.5.6 Software Infrastructure) on local resources
    a. Orca is used to deduce and model the electronic structure of a sample
4. Spectrometry results are analyzed using a mixture of homemade scripts written in C and a number of plotting packages
5. Currently, checkpointing is not performed; this is a limitation of the software.

In the general case, data set sizes from the remote resources are KB to small numbers of MB. Data consists of many small files that are used during processing. The electron structure analysis can produce a data set that is greater than 25 GB.

3.5.5 Remote Science Activities

The use of remote instrumentation is the primary science driver. To date there are no significant external collaborations that involve data transfer, and there is no use of external HPC resources. This may change as the group becomes more sophisticated in the need to leverage HPC/HTC resources located around the country.

3.5.6 Software Infrastructure

Beyond typical office/productivity software, the research group uses the following products:

- UCSF Chimera[30]: an extensible program for interactive visualization and analysis of molecular structures and related data, including density maps,

[30] https://orcaforum.kofo.mpg.de/app.php/portal

supramolecular assemblies, sequence alignments, docking results, trajectories, and conformational ensembles

- ORCA[31] (genOmics Research Container Architecture): provides access to a menu of validated bioinformatics software and the power to use these inside a secure and private containerized environment.
- Igor Pro[32]: an interactive software environment for experimentation with scientific and engineering data and for the production of publication-quality graphs and page layouts.
- Homemade scripts in C, plotting languages, etc.

3.5.7 Network and Data Architecture

This case study utilizes existing campus technology and network infrastructure (please see Section 3.1.4 Network and Data Architecture) for all aspects of the data workflow. In particular, most processing of experimental results is done using personal and local compute resources. Some codes have been converted to run on the campus HPC resources, and can do so when time is available.

3.5.8 Cloud Services

Cloud services are not utilized in a formal method by this research group. The use of data storage and sharing (e.g. Google Drive, BOX, Dropbox, etc.) are more of a novelty to share files between local collaborators, and are not a part of the workflow to migrate data from remote facilities back to Trinity. Use of portable media is by far the common use case.

3.5.9 Known Resource Constraints

The most significant impact on the scientific workflow for this use case is exporting data from a number of the experimental facilities. Certain facilities, such as NSLS2 and Diamond, do not make automated data sharing methods available to visiting users - thus the use of removable media is a requirement. This is done for a number of reasons, the main reasoning being that data security for government run facilities is significantly higher for visitors than for staff.

In the experience of the researchers at Trinity, the network security requirements of many of the laboratories are especially stringent; many universities do not meet the security demands required to exchange data sets between the national laboratory and local university networks. Therefore, we have been exchanging data between computers manually (i.e. by physical media such as usb-flash drives), which sometimes requires physically mailing disks.

The research itself is hitting a limitation with regards to computation. Because of the number of heavy atoms coupled with the large active-space, the calculations are very computationally demanding and long. Four processors with 120 GB of

[31] http://www.bcgsc.ca/services/orca
[32] https://www.wavemetrics.com/products/igorpro

memory devoted to each processor are currently used, and that level of main memory is not enough for analysis. The research group notes that memory, and not processing time, is the largest limitation. At a minimum, 256 GB of RAM is a requirement for some of the codes that need to be run.

Even with 480 GB of memory devoted to the calculation, Compound 1 required special truncation of the CASSCF wavefunction to make the calculation tractable. The larger systems we wish to investigate, which would mimic the cluster with higher fidelity, would require more computational resources than we currently have access to. For these, the use of national laboratory supercomputing facilities will be required due to their higher memory per-node design.

The research group notes they have not fully explored the use of Trinity's HPC resources, and are interested in learning if there is time available. Beyond this, the use of regional or national computation will be explored. Current model for computation is done almost entirely on local (PC) resources. Each machine has a specific role in the research process, thus the resources are not fungible and can be oversubscribed.

3.5.11 Outstanding Issues

The use of remote facilities, and the lack of technology enabled workflow acceleration, is the most critical factor to this research. Data mobility via portable media is still possible (and regularly used) due to the size of data sets involved. As the group performs more experimentation, and regularly visits remote facilities, a mechanism to automate the process of data transfer from experimental source to destination is very desirable. If this could be coupled with reliable storage (with backups), and HPC resources, it would be possible to significantly increase productivity.

3.6 Geosciences Case Study

3.6.1 Science Background

Dr. Ben Surpless and his team at Trinity University perform research in structural geology. Simply stated, this is the study of the ways that earth materials deform under natural stress. This work involves large-scale processes, for example the movement of tectonic plates, as well as more localized, temporally-distinct events, such as fracture formation. A specific focus involves the behavior of rock as it fractures due to faults and folding, and the subsequent evolution of the environment that results after these events.

For example, consider the systematic structural features, such as faults or fractures that could occur on an exposed rock outcropping on the Earth's surface. Various measurements and observations of these features can be gathered at that location and be used to hypothesize the behavior of similar rock systems below the surface. These hypotheses can be used to make predictions about more generalized rock behavior using models that are developed after initial observations are analyzed.

Dr. Surpless and team routinely study the environment of southern Utah, and travel to the region to for aspects of this research. The formation and study of fracture networks, including the observations of rock fractures over a large area on cliff faces has been part of a recent study. There are several observational aspects that are of interest:

- Variety, age, and complexity of rock
- Spacing of fractures
- Orientation of fractures on a macro and micro level
- Characteristics of the fractures on different time scales

Due to the complexity of the environment, which includes sheer cliff faces that can be ~200m (~650ft) in height, it is infeasible for research staff to directly observe many aspects of the structures. The use of unmanned aerial vehicle (UAV) technology has been adopted to assist in the process of capturing still and moving images of each region of study. This allows for safe observation, but adds complexity to the research pipeline due to the technology components that are now managing critical research products.

After the data is captured from the UAV, it is processed to build 3D models of the landscape. This rendering process can be time consuming, as all videos and images must be processed and stitched together. After completion, the resulting models can be used for a variety of research projects, such as:

- The study and location of ground water
- Understanding the impacts to the location, production, and storage of fields related to oil or gas energy use
- Understanding underlying soil stability, especially as it relates to earthquakes

The work is still in the nascent stage, and extensive use cases and collaborations are not well defined at this time.

3.6.2 Collaborators

Trinity is a member of the Keck Consortium, along with several other schools:

- Amherst College
- Beloit College
- Carleton College
- Colgate University
- The College of Wooster
- Colorado College
- Franklin & Marshall College
- Macalester College
- Mount Holyoke College
- Oberlin College
- Pomona College
- Smith College
- Union College
- Washington and Lee University
- Wesleyan University
- Whitman College

In particular, students from The College of Wooster and Mount Holyoke College have worked closely with Dr. Surpless on research projects in recent years. Pursuit of NSF grants in the geoscience space has the potential to increase this further.

Prior work has also been performed with members of the Southwest Research Institute. There is not an active project in this space, but collaboration is possible in the future.

Prior conversations with the U.S. Geological Survey (USGS) around the use of UAVs for mapping and imaging purposes did not result in a close collaboration. The USGS does have a similar research effort ongoing, but uses significantly larger and more complicated infrastructure, for example larger vehicles that require a trained pilot and the use of formal airfields. Participation requires significant cost investment, in some cases greater than $10,000 USD per flight, which is not feasible at this time.

There are no formal collaborations between Dr. Surpless and providers of computational infrastructure or storage. All processing is done locally on non-HPC resources. There is sporadic use of cloud resources as a storage infrastructure to facilitate exchange of data between the aforementioned collaborators.

3.6.3 Instruments and Facilities

There are numerous instruments involved in the field work for this research includes:

- Petrographic Microscopy: optical microscope used to identify rocks and minerals in thin sections
- Mass Spectrometry/Isotope-Ratio Mass Spectrometry: ion and isotope analysis
- Schmidt Hammer: testing compressive strength of the material
- Digital Photography
- Digital/Analog Observations (e.g. notes)
- UAV: which features a 4K-capable video camera

The data output of the instruments is reasonably small in size, and is often captured as analog results or as raw output from the device.

The data from the UAV is the largest and hardest to deal with at the current time. For example, a two week field study in 2018 in the Utah desert produced over 380GB of data, including raw and processed video footage, screen captures and stills extracted from the videos, metadata associated with the videos, as well as notes and analysis from in situ work.

Currently, data management is performed on an ad-hoc basis. Original files are stored using a mixture of personal and lab PCs, external hard drives, and backups that are performed to cloud resources. Long term curation of research data is a known problem that must be addressed prior to pursuit of NSF funding sources, as a requirement is publication and dissemination of the resulting data sets. This is still being investigated, as the amount of required storage may vary. For example, The publication of raw video will increase storage demand by 10-100 times as opposed to the publication of results and models, which are significantly smaller.

Analysis is also ad-hoc, and is primarily performed on two licensed, local PC resources. The main software (discussed in Section 3.6.6) for the processing and analysis of the video does not lend itself to remote use cases or HPC by default, and requires that the data being manipulated be in close proximity to the source of processing due to a need for low latency. Prior experimentation to separate the computational pieces and storage resources, using the communal campus storage, resulted in operational complications. A full analysis was not done at the time, but the additional latency, even at the distance between a campus and local resource, resulted in the analysis software operating too slow to be useful. It is believed that the software requires local and fast access to all forms of data involved in analysis to function.

3.6.4 Process of Science

This use case will highlight the recent field trial by Dr. Surpless and his students. This involved physical travel to Orderville in southern Utah, which is remote, and is close to the border of Arizona and several national parks or monuments including Zion NP, Dixie NF, and Grand Staircase-Escalante NM. This work was performed without the assistance of local universities or technology providers.

- The field work was performed over a two to three week period, and followed a similar pattern. Work started early prior to the heat of the day and increasing winds experienced in the southern Utah region. Two days of in situ field work took place to document the preliminary site and capture details about what was physically there, such as the properties of the rock bodies (hardness, color, texture), spacing, and the orientation of accessible fractures. The notes, measurements, and resulting photography would be digitally scanned at the end of the 2-days.
- Every third day, a UAV flight would obtain fracture data from inaccessible rock outcrops above where the prior two-days of field work was performed. The overall goal of a flight operation is to capture and document a specific region that is inaccessible to geologists' ground investigations. This would include video transmitted back to pilot for calibration purposes. The pilot would also narrate what was seen, , take timestamps and make notes.
- At the end of the flight, the video is downloaded and reviewed to make sure observation was successful. There is no modeling work done in the field, so all capture (including re-capture based on simple calibration) must be complete.
- The schedule would repeat.

The process of science involves a very close relationship between the two portions, and is depicted in Figure 3. The main idea being that data can be directly compared between the video capture via UAV and the human observation work. The resulting models (generated later back at Trinity) are then verified for accuracy.

Weeks after the field work, the analysis is conducted. The steps include:

- Data is pulled together from the various sources it was recorded on during the field work and centralized onto the local departmental PCs used for analysis. Due to storage limitations, this may mean deletion or backup of other data from the PC resources.
- Still images are created from video stream using the free video-editing software, VLC media player.
- Agisoft PhotoScan (now known as Agisoft Metashape) is run on the two local PCs to build high-resolution 3D models of the landscapes using GIS data from the still images along with creating realistic textures by layering the original images on top of the mesh models.
- After model creation, georeferencing is done using the field observational data, GIS software, and the created model.
- The models are then exported, via ArcGIS, to further process the results and insert proper spacing, lengths, units, etc. This permits quantitative analysis of the inaccessible rock bodies above where field data were collected, thus

permitting direct comparison between data gathered by drone and data from direct field observations.

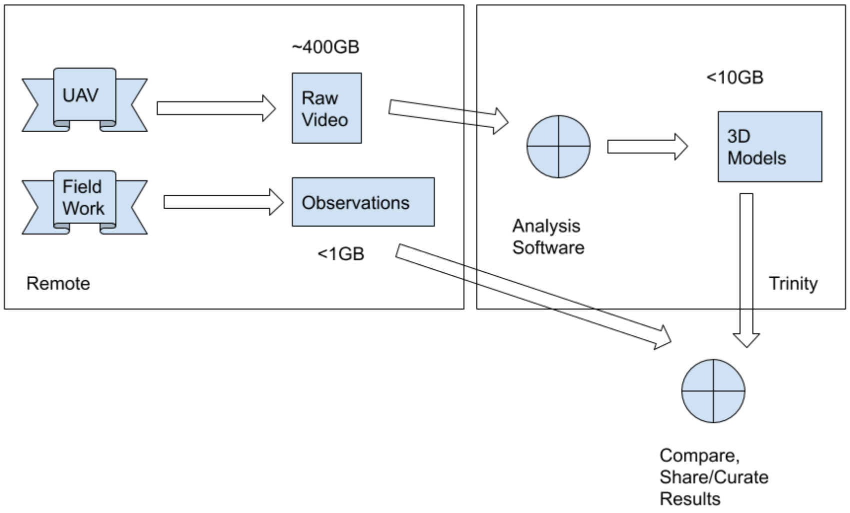

Figure 3 - Example of a Geoscience workflow.

As previously mentioned, pursuit of NSF funding via the Keck Consortium is being planned. This would further field trials and support additional analysis on the data products.

The NSF-funded project (through the Keck Geology Consortium) was already funded and completed by spring 2019. However, there are plans to submit a new, more extensive proposal to NSF (3-year project), tightly linked with the Keck program, to build on what has been accomplished to date

3.6.5 Remote Science Activities

Beyond the field work, there is currently no remote component to this work. All processing is done locally, with only minor use of network infrastructure to facilitate data sharing between occasional collaborators.

This pattern has the potential to change with funding opportunities, as more members of the Keck consortium could become involved in the observation and analysis work.

3.6.6 Software Infrastructure

Software used in this research consists of:

- Agisoft PhotoScan - used to build models of landscapes and creating textures by laying the drone photos on top of the mesh models.
  - The primary process is called ‘point matching’; the software is able to line up images that share common features and overlap pieces to form a larger image.
  - Two intermediate forms are created:

    - Sparse Point Cloud - initial attempt to create the large image from source files. Typically involves the user assisting to make decisions about hard areas, edges, etc.
    - Dense Point Cloud - More intricate (larger) image that takes the sparse point cloud and adds additional detail
  - A polygonal mesh is produced by using the dense cloud (e.g. 10-15 million data points) and imagery.
    - This is the most intensive operation for CPU and memory
    - The size of the model impacts the ability to operate in all cases - larger models need more memory and CPU time.
    - Small models can take less than an hour to create, larger ones can take hours.
- ArcGIS to provide access to accurate maps

There is not a need for any additional collaboration tools at this time. Field work is not meant to be interactive for remote collaborators, and most interactions are done asynchronously or local to Trinity during the analysis phase.

3.6.7 Network and Data Architecture

For the main components of this section, please see Section 3.1.4 Network and Data Architecture

3.6.8 Cloud Services

Cloud services are not widely adopted or used. To date, the only cloud resources that could be used are related to collaboration needs (e.g. sharing data) into Dropbox or Google Drive. Performance to these is not a primary concern versus ease of use and available storage space. In all cases, a ‘personal’ account is used (e.g. not one sponsored through Trinity University).

3.6.9 Known Resource Constraints

Storage resources, and mobility of data, are the two largest concerns going forward. In summary:

- Availability of storage local to campus is a high priority
  - Data volumes of UAV videos will increase, particularly if funding is sought for projects.
  - Agisoft PhotoScanrequires data to be local for processing reasons, thus storage on campus (with fast interconnection to processing PCs) will be required.
- Ability to share data (e.g. capacity, mobility, access) is a high priority
  - Keck Collaboration places individuals at other institutions. Use of networks to share UAV data will assist in the education of students, and the further distribution or involvement of other research groups.
- Data backups (onsite or offsite) is a medium priority
  - Backups are performed via cloud services now
  - Local or geographically close backups are desired

- Federated access to storage is a medium priority
    - Being able to share widely with semi-fungible collaborators would be useful.
    - If required for funding, data management, including portals for research data, would be something to explore.

3.6.11 Outstanding Issues

Experimentation with Agisoft PhotoScanusing local-to-campus data storage revealed the strong interdependence between the storage and processing, with the two needing to be very close to each other. As a result of this, all current work (storage and processing) is performed on the same PC resource. This does not scale well, particularly as the time to process and size of data inputs increases. Last experimentation on this was in 2015, thus it is time to attempt to connect high-speed network storage to the processing PCs again after Trinity networking upgrades.

The aforementioned lack of storage will hamper research productivity in the future. The current volume of research data (e.g. < 1TB for a field study) is currently feasible using existing resources.

It is anticipated that early 2020 will be when an NSF grant is applied for. This would imply that Fall of 2019 will be the ideal time to plan network/storage/computation upgrades, with possible implementation (if grant is successful) in the Spring of 2020.

4 Discussion Summary

On May 29th 2019, members of the EPOC team and staff from LEARN met with representatives from Trinity University. This review was held in San Antonio, TX.

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

- Network and data architecture
- Security profile for scientific use cases
- Use of Cloud services
- Local and regional HPC/HTC use
- Storage (local and remote)
- LEARN R&E networking capabilities
- Pending CC* grant proposal to the NSF

4.1 Network and Data Architecture

A full review of the Trinity University network infrastructure was conducted between Trinity University IT staff, LEARN, and EPOC. Figure 1 shows the current campus infrastructure and Figure 4 shows some discussion items that could be considered to better support some of the research use cases at Trinity University.

The design of the current network facilitates a typical enterprise workload, and emphasizes availability and redundancy for key services around campus. Discussion about the construction of a dedicated science network revolved around central themes:

- Upgrade to core and edge hardware
- Increased capacity via LEARN
- Balancing routing/peering between multiple connections
- Handling increased science workflow
- Integrating storage into the final network design

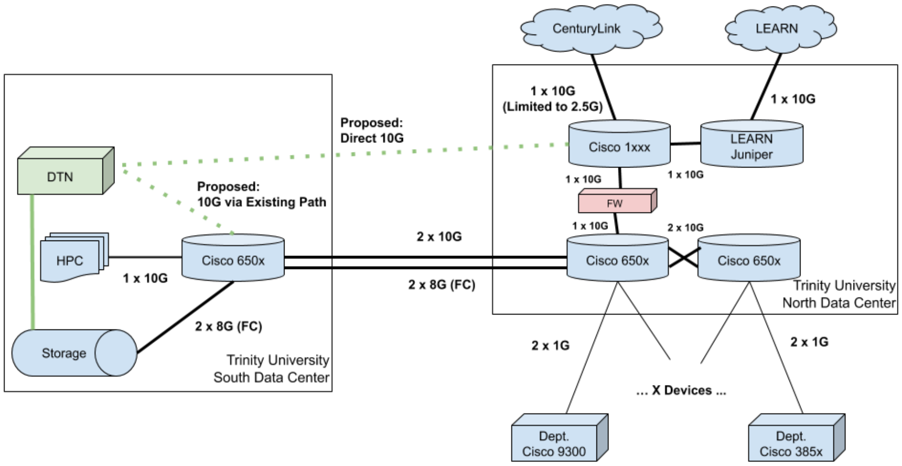

Figure 4 - Architecture of the proposed Trinity University network.

These preliminary discussions are contingent on NSF funding, see Section 4.7 Pending Proposal to the NSF.

4.2 Security Profile for Scientific Use Cases

The use of a Science DMZ infrastructure brought questions regarding the protection and profiles for different classes of network traffic, particularly on portions of the network that may not feature a firewall infrastructure. Discussion topics included:

- Design of the new network, to facilitate an alternate connection/peering point to LEARN for only scientific traffic
  - Multiple BGP peerings, and the local preferences on each
  - Use of a dedicated VRF for only science traffic on LEARN
  - New border devices to handle high-performance use cases
- ACL strategies for scientific use
  - Port to application mappings
  - Inbound vs. outbound strategies
  - Use of routing and/or whitelists to access critical resources
- Protection mechanisms
  - ACLs on core/edge devices
  - IDS (intrusion detection systems) vs. IPS (intrusion prevention systems)
  - Host-based controls that can be automated
  - Measurement/monitoring of netflow/sflow data and analysis tools

4.3 Use of Cloud Services

There is not currently a heavy push, nor an available solution, to integrate cloud storage campus wide. Discussions in the past did center on the use of BOX, Dropbox, or Google Drive as a solution for faculty and staff. In discussions with researchers, it was not clear that many could utilize this system for a number of reasons:

- Security profile at 'far end' locations such as facilities at federal laboratories may not allow the use of these for data sharing .
- International users have performance complications when sharing via these methods.
- The tools that researchers are using do not natively integrate with these solutions at this time.

Cloud services may become more critical in the future, but are not currently a high priority.

4.4 Local and Regional HPC/HTC Use

Several of the Case Studies presented in this Campus-Wide Deep Dive utilize some aspect of HPC use, provided by the Trinity Campus. These include Physics, Chemistry, and Computer Science. Given the use, wait times for resources are growing and have been reported to be as high as days to weeks during busy periods (e.g. end of semester, as students are performing cap-stone research activities). As a result of this, there are efforts to augment the capacity of the existing system by adding more nodes and increasing memory footprint. This will facilitate more users and more cycles into the future.

The use of regional facilities (TTU, TACC, TAMU, etc) is also growing when users are not able to use local resources. Trinity and LEARN will work toward simplifying the network path to facilitate this use case by normalizing peering where necessary, along with considering the use of Data Transfer Node (DTN) hardware and software.

4.5 Storage (Local & Remote)

The need for additional storage to facilitate research use cases was a common theme for all research areas reviewed. Storage is currently handled on an ad-hoc basis by individual groups, which typically means that it is accomplished by the use of removable media internal to a specific project.

Discussion centered on:

- Campus-wide solutions that could centralize storage and give allocation to all users (See Figure 5). Given the need for immediate solutions, some of the presented ideas may take longer to implement.
- Network and system improvements to allow people to access the storage locally and remotely
- Data transfer systems to facilitate external high-speed sharing with collaborators and national facilities
- Scalability into the future
- Integration with cloud providers (lower priority)

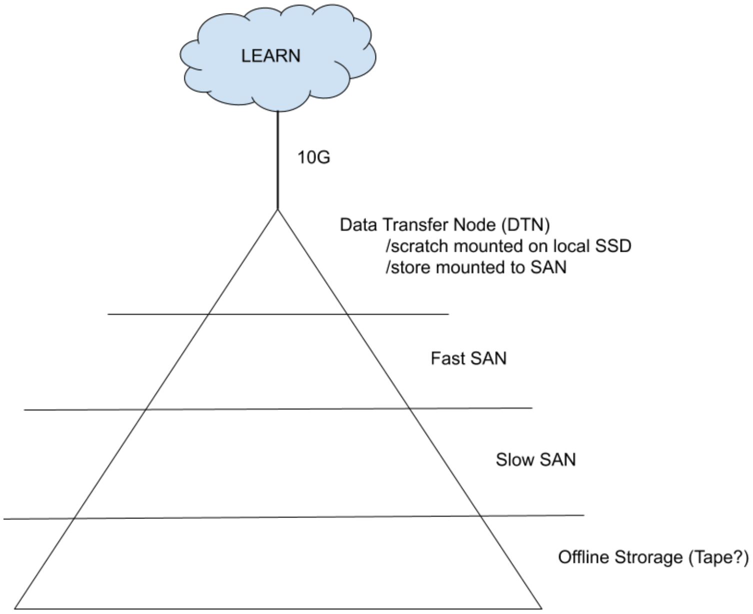

Figure 5 - Diagram of a Data Storage Pyramid.

Trinity University IU staff will devote resources to this problem in the future, but acknowledges that stop-gap solutions such as more local storage may offer immediate relief.

4.6 LEARN R&E Networking Capabilities

LEARN and Trinity are in the process of upgrading a metro ring of networking around San Antonio and are also normalizing the 10Gbps connection to the campus. This work is mostly complete, but will require some additional testing and validation for a short period of time. LEARN is also exploring ways to improve data paths between Trinity and collaborators, and will continue to address peerings and VLAN changes as needed.

If the pending grant proposal to the NSF is accepted (see 4.7 Pending Proposal to the NSF), the architecture of the Science DMZ portion of the network may change to facilitate this new use case.

4.7 Pending Proposal to the NSF

Trinity and LEARN have partnered on a proposal to the NSF to facilitate the installation and operation of a Science DMZ enclave for the campus that would be operated by LEARN on a regional basis. The proposal has not been awarded at the time of this report, but there has been indication from the NSF that it is under consideration. If awarded, Trinity, LEARN, and EPOC will work together to implement some of the policy and technology support. Figure 6 shows one depiction of how this infrastructure could be installed on the Trinity University Campus: additional hardware (funded by the grant, and maintained by LEARN) would be installed to facilitate a Science DMZ network infrastructure.

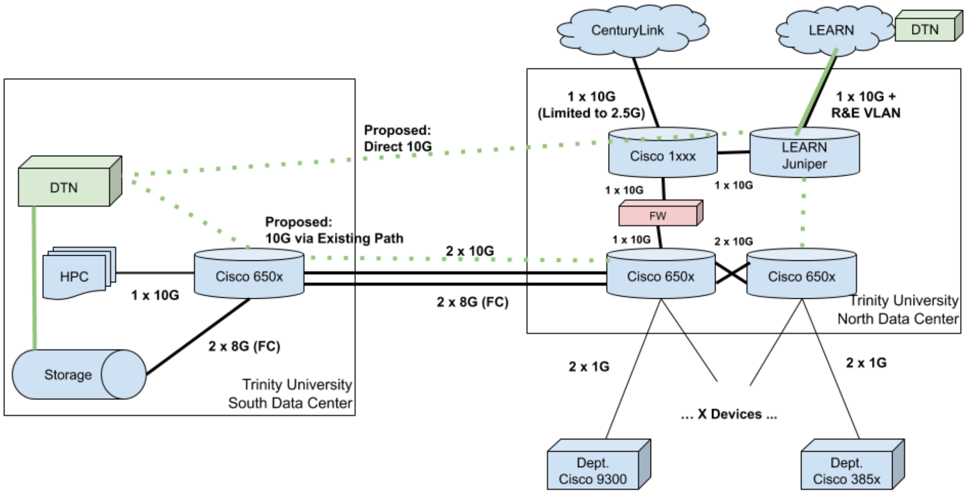

Figure 6 -Architectural diagram of the proposed LEARN DMZ Network.

5 Action Items

EPOC and LEARN recorded a set of action items from the Trinity University Campus-Wide Deep Dive, continuing the ongoing support and collaboration. These are a reflection of the Case Study reports, and in person discussion.

- LEARN, Trinity University, and EPOC will continue a discussion regarding network arhitectural needs if the NSF proposal is accepted.
    - Specification of hardware
    - Best practices for operational soundness
    - Integration of research use cases
    - Operation at the regional and national level
    - Performance testing and monitoring
    - Network Security best practices
- LEARN, Trinity University, and EPOC will begin a discussion about research storage, and ways this can be integrated into scientific workflows
    - Specification of hardware
    - Best practices for operational soundness
    - Integration of research use cases
    - Performance testing and monitoring
- LEARN and Trinity University will finalize plans for LEARN connectivity and peering arrangements.

Appendix A - Trinity University Technology Support Overview

All Trinity faculty and staff have access to the facilities and resources associated with Trinity University in San Antonio, Texas, where they are employed. Trinity is classified as a research university with a high level of research activity. Faculty and staff have access to a full range of academic and research related resources including online survey instruments, Institutional Review Board, and library facilities.

Data Hardware & Software

Trinity University was awarded NSF grant funds to purchase a Penguin Computing 45 node (1,512 CPU core) cluster to support large-scale, high-performance computing (HPC) at Trinity University. Students and faculty sponsors are able to perform intensive computing tasks for a broad range of scientific research efforts spanning data science, mathematics, computer science, chemistry, geology, biology, and physics. Student researchers in these departments have access to considerable computational power that can be used to analyze large datasets generated through traditional laboratory experiments or through robust computational simulations.

The Trinity University HPC cluster consists of 36 CPU nodes, 5 GPU nodes with 2 NVidia Tesla K80m GPUs, 1 memory node, 1 login node, 1 management node, and 1 data node with 211 Tb of attached storage. This resource enables Trinity University to effectively train and educate students on tangible aspects of large-scale computational projects and research endeavors.

The Trinity University Information Technology Services division also manages and supports multiple student computer labs with Windows and Mac OS client workstations. There are 711 computer workstations in academic labs, 30 computer workstations in dormitory labs, and 9 virtual desktop images (VDI) available for student researchers. Data analytics and research software is installed on the physical and virtual workstations (e.g. SAS, SPSS, R, NVivo 11 Pro, Tableau).

Server/Network Data Center

The University data center consists of two facilities, a network core building and a data center. The Trinity University Data Center is a Tier 2 class data center with online double conversion UPS and diesel generator secondary source. The equipment in the network core is housed in a self-contained Liebert "Smart Row" with redundant air-conditioning and a fire suppression system. The network core room terminates the campus fiber networks and private fiber networks.

Plans are in process to move the data center to a new facility on campus. Internet2 connectivity is available via a connection to LEARN (the Lonestar Education and Research Network). A committed data rate of 1 Gb/s is available and is burstable up to 10 Gb/s. A second connection for general Internet is also available via

CenturyLink with a data rate of 2 Gb/s. The various buildings on campus are connected via a fiber optic backbone with a 2 Gb/s or more bandwidth to each building. Computers in each building have access to wired connections at 1 Gb/s.

Access Controls

Card swipe access and unsupervised 24/7 access to the Trinity University Data Center will only be given to individuals with an approved and demonstrated business need. Those individuals requiring infrequent or temporary access to the Data Center will be granted escorted access as needed.

Security Controls

Security is controlled via a proximity card reader system. All areas of the Data Center and the network core room are under video surveillance 24/7 both internally and externally. The video feed is monitored by Trinity University Police Department.

Appendix B - Trinity University Cyberinfrastructure Plan

As the regional networking organization representing R&E community in Texas, LEARN strives to meet the needs of both currently connected members and future membership.

LEARN CI Outreach to Small Campuses Across Texas: It is part of the LEARN CI plan to provide smaller campuses with access to the national R&E network so that faculty, staff and students may have access to and take advantage of research and education opportunities in projects that might require their skill, talents and expertise.

Participating Campuses. Campuses are currently connected to LEARN as follows:

Participating Institution	R&E Current Connectivity via LEARN	R&E Future Connectivity via LEARN	LEARN Membership
Trinity University	1G	1G/10G	Current
Texas Wesleyan University	0G	1G/10G	Future
South Texas College	0G	1G/10G	Future
South Plains College	0G	1G/10G	Future
McLennan College	0G	1G/10G	Future

They have become/seek to become a member of the advanced networking R&E community because of their driving research and education applications as listed in the project description. The efforts described in this proposal will assist them to connect to the LEARN backbone and peer with Internet2, ESnet, PacificWave as well as access valuable services including the virtual Science DMZ and data transfer node. Additionally, they will have access to Texas Advanced Computing Center (TACC) at UT Austin. This enables each of these institutions to use HPC compute and storage resources at TACC. The proposed efforts will also allow participants, as part of the LEARN community, to keep up with the latest developments in advanced networking technology, staff training developments, and more.

Appendix C - LEARN Regional Networking Diagram

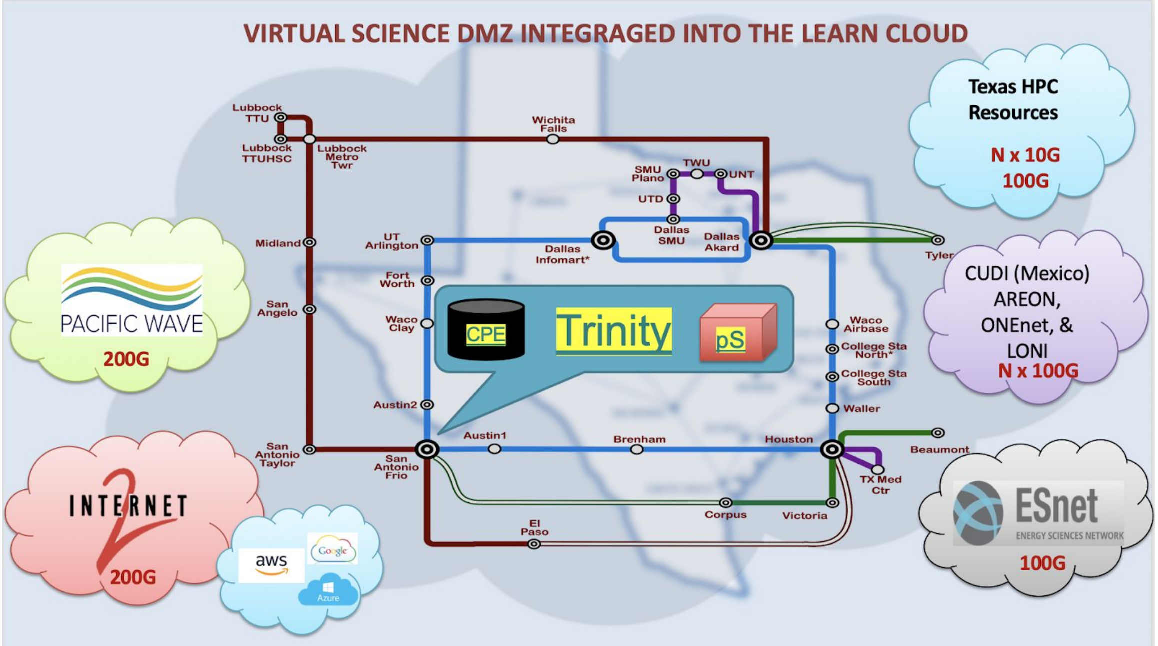

Figure 7 - Proposed LEARN Virtual Science DMZ.

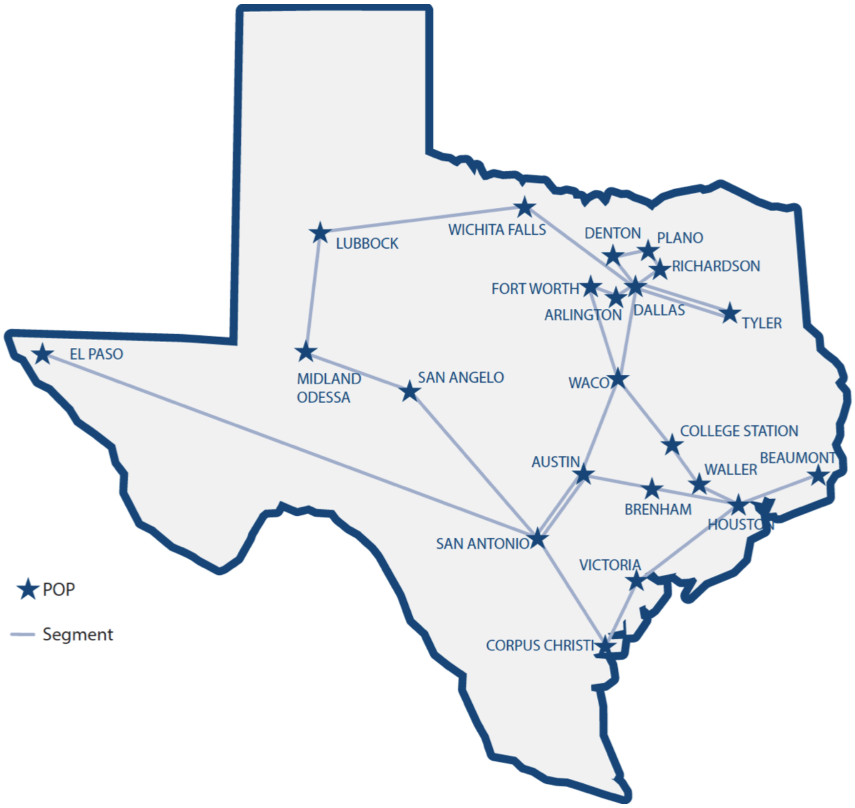

Figure 8 - Schematic of the LEARN Network.

Integrating Virtual Science DMZ with Campus CI and TACC

LEARN is enabling a virtual overlay for Science DMZ for HPC flows. If the NSF grant is funded, the grant will enable a 10G capable WAN CPE and PerfSONAR server at the Trinity University San Antonio Campus, which complements the perfSONAR enabled LEARN infrastructure. The virtual science DMZ on-ramps for the Trinity University campus will be via the San Antonio LEARN POP.

LEARN provides to its members, a carrier class MPLS Layer 2/3 network built over the advance optical Layer 1 and fiber IRU based infrastructure. LEARN connects over 50 campuses including high performance computing centers, such as, The Texas Advanced Computing Center (TACC), which connects to LEARN at 100Gbps.

With LEARN's partnership with Internet2, our researchers at LEARN connected campuses have the option to leverage the layer 2 cloud connectivity via LEARN's 100G port in Houston and 100G port in Dallas. Cloud is playing an increasingly important role in scientific discovery and data sharing.