Title

Baylor University Campus-Wide Deep Dive

Permalink

https://escholarship.org/uc/item/5x9234mv

Authors

Zurawski, Jason
Schopf, Jennifer

Publication Date

2021-01-29

Peer reviewed

Baylor University Campus-Wide Deep Dive

January 6-7, 2020

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.

Baylor University Campus-Wide Deep Dive

Final Report

Waco, TX
January 6-7, 2020

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-2001384[1].

[1] https://escholarship.org/uc/item/5x9234mv

Participants & Contributors

Christopher Becker, Baylor Mass Spectrometry Center
Erik Blair, Department of Electrical & Computer Engineering
Scott Day, Baylor ITS
Sal Ghani, LEARN
Leigh Greathouse, Department of Nutrition Sciences
Michael Hand, Baylor ITS
Bob Hartland, Baylor ITS
Kenichi Hatakeyama, Department of Physics
Byron Hicks, LEARN
Mike Hutcheson, Baylor ITS
Amy Santanta, LEARN
Keith Schubert Department of Electrical & Computer Engineering
Kevin Shuford, Department of Chemistry and Biochemistry
Chad Talbert, Baylor ITS
Jeff Wilson, Baylor ITS
Tim Woodbridge, LEARN
Jason Zurawski, ESnet

Report Editors

Dr. Jennifer M Schopf, Indiana University: jmschopf@indiana.edu
Jason Zurawski, ESnet: zurawski@es.net

Contents

1 - Executive Summary

In January 2020, staff members from the Engagement and Performance Operations Center (EPOC) and the Lonestar Education And Research Network (LEARN) met with researchers and staff at Baylor 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 scientific use cases included:

- Experimental High Energy Physics (HEP)
- Proton Computed Tomography (pCT)
- Nutrition and Relation to Digestive Microbiome
- Baylor University Core Research Facilities
- Molecular Quantum-dot Cellular Automata (QCA), and Material Science of Quantum Computing
- Modeling and Simulation of Low-Dimensional and Nano-Structured Materials
- Computational Fluid Dynamics

Material for this event included the written documentation from each of the research areas at Baylor 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 Baylor 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:

- Tradeoffs for network/software security, and usability of the resulting infrastructure. Better communication to set expectations and understand realities is required.
- Computation use on campus is widespread and healthy. While no major problems were uncovered, upgrades to maintain current usage patterns and encourage growth will be required.
- Storage is a critical need for enterprise use cases and research. In particular, a campus wide 'storage architecture' to support research use cases (e.g. instruments, data sharing) is required in the 2-5 year time window.
- Instrumentation on campus is healthy and expanding. Technology must scale with this in the form of computation and storage.
- Working with LEARN to upgrade network capacity (in multiples of 10G, or upgrades to 100G) will be required in the 1-3 year time frame.
- Network monitoring and visibility will help to establish external science use cases.
- Data sharing via portal systems is not currently a critical need, but growing in scope. EPOC can assist Baylor with options.

Recommendations from the meeting included:

1) Baylor ITS will investigate ways to set expectations with the scientific user community regarding IT processes and timelines. This includes, but is not limited to, reviews for necessary software, procedures for accessing secured computing environments, ways to integrate research use cases to existing computational and storage infrastructure, ways to integrate new research use cases, and relationships within the state (via LEARN) to allow for off-site back-ups and resource sharing.

2) Baylor ITS will investigate upgrades to the research computing and storage infrastructure in the coming years to meet and exceed researcher demand. Options include upgrades to existing infrastructure (nodes, networking, GPUs, condos, etc.) to improve performance and increase access, as well as adding new services such as a CUI (controlled unclassified information) environment.

3) LEARN and Baylor ITS will explore regional and campus upgrades to meet the networking demand of scientific users.

4) Baylor ITS will develop new, and improve existing, researcher-focused services such as integrating storage into the workflow for scientific instruments, establishing new Globus endpoints and portals, investigating site-license or local deployments of popular services (e.g. git), and offering consulting to ensure that researchers have IT support for their scientific missions.

5) Baylor ITS will make campus network upgrades in the coming years to meet capacity demands, as well as performing deeper analysis on traffic patterns using tools like perfSONAR, SNMP, and sFlow monitoring.

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 aims to understand the full science pipeline for research teams and suggest alternative approaches for the scientists, local IT support, and national networking partners as relevant to achieve the long-term research goals via workflow analysis, storage/computational tuning, identification of network bottlenecks, etc.

The Deep Dive process is based on an almost 10-year practice used by ESnet to understand the growth requirements of 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 Baylor University Campus-Wide Deep Dive Background

In January 2020, EPOC and Lonestar Education And Research Network (LEARN) organized a Campus-Wide Deep Dive in collaboration with Baylor University to characterize the requirements for several key science drivers. The Baylor University representatives were asked to communicate and document their requirements in a case-study format (see Section 3: Baylor University Case Studies). These included:

- 3.1 Campus Overview
- 3.2 Experimental High Energy Physics (HEP)
- 3.3 Proton Computed Tomography (pCT)
- 3.4 Nutrition and Relation to Digestive Microbiome
- 3.5 Baylor University Core Research Facilities
- 3.6 Molecular Quantum-dot Cellular Automata (QCA), and Material Science of Quantum Computing
- 3.7 Modeling and Simulation of Low-Dimensional and Nano-Structured Materials
- 3.8 Computational Fluid Dynamics

A face-to-face meeting took place at Baylor University in Waco, TX on January 6th-7th, 2020 (see discussion in Section 4 Discussion Summary). We document next steps in Section 5 Recommendations for Review .

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 43 organizations throughout Texas that includes public and private institutions of higher education, community colleges, the National Weather Service, and K–12 public schools. The consortium, organized as a 501(c)(3) non-profit organization, connects its members and over 300 affiliated organizations through high performance optical and IP network services to support their research, education, healthcare, and public service missions. LEARN is also a leading member of a national community of 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.

Baylor University Baylor University in Waco, Texas, is a private Christian university and a nationally ranked research institution. The mission of Baylor University is to educate men and women for worldwide leadership and service by integrating academic excellence and Christian commitment within a caring community. Chartered in 1845 by the Republic of Texas through the efforts of Baptist pioneers, Baylor is the oldest continually operating university in Texas. Located in Waco, Baylor welcomes students from all 50 states, the District of Columbia, and 89 countries to study a broad range of degrees among its 12 nationally recognized academic divisions.

3 - Baylor University Case Studies

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

- 3.1 Campus Overview
- 3.2 Experimental High Energy Physics (HEP)
- 3.3 Proton Computed Tomography (pCT)
- 3.4 Nutrition and Relation to Digestive Microbiome
- 3.5 Baylor University Core Research Facilities
- 3.6 Molecular Quantum-dot Cellular Automata (QCA), and Material Science of Quantum Computing
- 3.7 Modeling and Simulation of Low-Dimensional and Nano-Structured Materials
- 3.8 Computational Fluid Dynamics

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.

Baylor 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

Content in this section authored by Scott Day, Michael Hand, Chad Talbert, Mike Hutcheson from Baylor Information Technology Services (ITS)

3.1.1 Institutional Background

There are 137 Baylor buildings housing university operations that are supported by the Baylor University Information Technology Services (ITS) group. Network connectivity can be divided into three categories:

- 115 Baylor-owned buildings on Waco main-campus. Network connectivity to these buildings is accomplished through physical pathways and fiber-optic feeders controlled and owned by Baylor.
- 8 buildings (some Baylor-owned, others leased) outside of the main-campus, but maintaining an internal connection to the campus-network. Network connectivity to these buildings is handled via external ISP layer-2/transparent-LAN-services and connected to campus. The largest of these buildings is the School of Nursing located in Dallas.
- 14 buildings (some Baylor-owned, others leased) that connect back to campus over public internet.

To accomplish the physical connectivity, there are 4 main fiber distribution-hubs for the main-campus.

The primary ITS facility that houses university technology is a data center environment. This building was constructed in 2003, and features 2900 ft^2 of space, 4 AC units that can support 133 tons of cooling, and 400 kVA power (with UPS backup). The building has on-campus replica capability (not geo-redundant at the time or writing), but is working to establish alternatives.

3.1.2 Collaborators

ITS provides service to all faculty, staff, and students on the campus. External relationships include The Lonestar Education And Research Network (LEARN); a consortium of 43 organizations throughout Texas that includes public and private institutions of higher education, community colleges, the National Weather Service, and K–12 public schools.

3.1.3 Instruments and Facilities

ITS provides a number of network infrastructure options and services to the campus environment. These were created to manage security and availability requirements, and are designed to scale to the needs of the community for future needs.

Datacenter Network: The campus datacenter is divided into multiple security enclaves based on usage.

The datacenter management of resources relies heavily on VMWare clusters to commission and deploy server resources. The clusters have various forms of redundancy, backups and access to centralized storage.

Servers in each datacenter network use a firewall as their default gateway and policy enforcement point. There are separate backup and management networks in the datacenter for server administration.

Baylor University Research Network (BURN): This is the Baylor implementation of the Science DMZ paradigm. In accordance with best common practices for this type of network, Baylor provides a rich set of policies and mitigations to protect the resources. The connection into the BURN is currently rate-limited to 3Gb in order to share resources with other components of the campus.

Faculty-Staff Network: The faculty-staff network supports wired and wireless network connectivity in faculty and administrative offices as well as classrooms. Separate VLANs (and separate VRFs) support data and VOIP services. There are 12,700 unique MAC addresses per month on Faculty-Staff networks, and an additional 800 VOIP hand-sets.

Command-and-Control Network: The command-and-control network provides a physically segmented network for building control and public safety systems (including fire alarm panels, in-building voice evac panels and security cameras) within main campus buildings. Access to these networks is very restricted on-campus and off-campus.

RESNET Network: RESNET supports wired and wireless network access in residence hall rooms (an average of 900 unique clients are currently connected)

Campus WIFI Networks: There are 2800 wireless access points installed throughout campus buildings. AIRBEAR is a campus wide WIFI network using 802.1x authentication for Baylor faculty, staff, and students. BU-GUEST is a campus wide open SSID, providing guest internet access to guests or conference attendees that are sponsored by a Baylor faculty/staff person. BU-DEVICE is an SSID deployed in the residence halls to help students connect devices that do not support 802.1x. The university advertises *eduroam* on campus access points in academic and administrative spaces. Baylor faculty, staff, and students can connect to *eduroam* networks while traveling abroad as well. The Baylor Research and Innovation Collaborative (BRIC), an off-campus facility that functions as a technology incubation center, has dedicated wireless controllers to help centrally manage and provision one-off wireless network services for faculty, staff, students, and affiliated Baylor guests and collaborators.

R&E Connectivity: Access to R&E networks is available through our connection to LEARN.

perfSONAR: There are two perfSONAR instances on campus. One is administered by Baylor, located in the campus datacenter. LEARN has a perfSONAR instance installed with their on campus router.

3.1.4 Software Infrastructure

ITS maintains enterprise software for the campus, e.g. licensing and deployment of operating systems (Microsoft, Linux), and productivity software (Microsoft Office). Research software is typically maintained by individual groups, but must be purchased, vetted, and approved by an ITS process. Research groups can request longer-term support (e.g. maintenance and operation) of software packages upon request.

Baylor utilizes Globus as an endpoint, and has purchased a standard subscription to the service. The addition of other services (e.g. connectors to cloud providers, etc.) is being explored as an option, once needs and capabilities are discovered as a part of the EPOC review.

3.1.5 Network and Data Architecture

The Baylor network is shown in Figure 1. The design features two border-routers connecting ISP services to two core routers. There is a research router that connects the BURN network to the two border routers. The core routers connect to pairs of redundant firewalls servicing various campus networks. The datacenter is behind a pair of firewalls. The faculty-staff and command-and-control networks are behind another pair of firewalls. RESNET and the campus wireless networks are behind a dedicated pair of firewalls.

There is a VPN server for use by Baylor faculty/staff (and students as requested). The VPN server also supports vendor VPN accounts for those not directly affiliated with Baylor. The VPN server runs on the faculty-staff firewall.

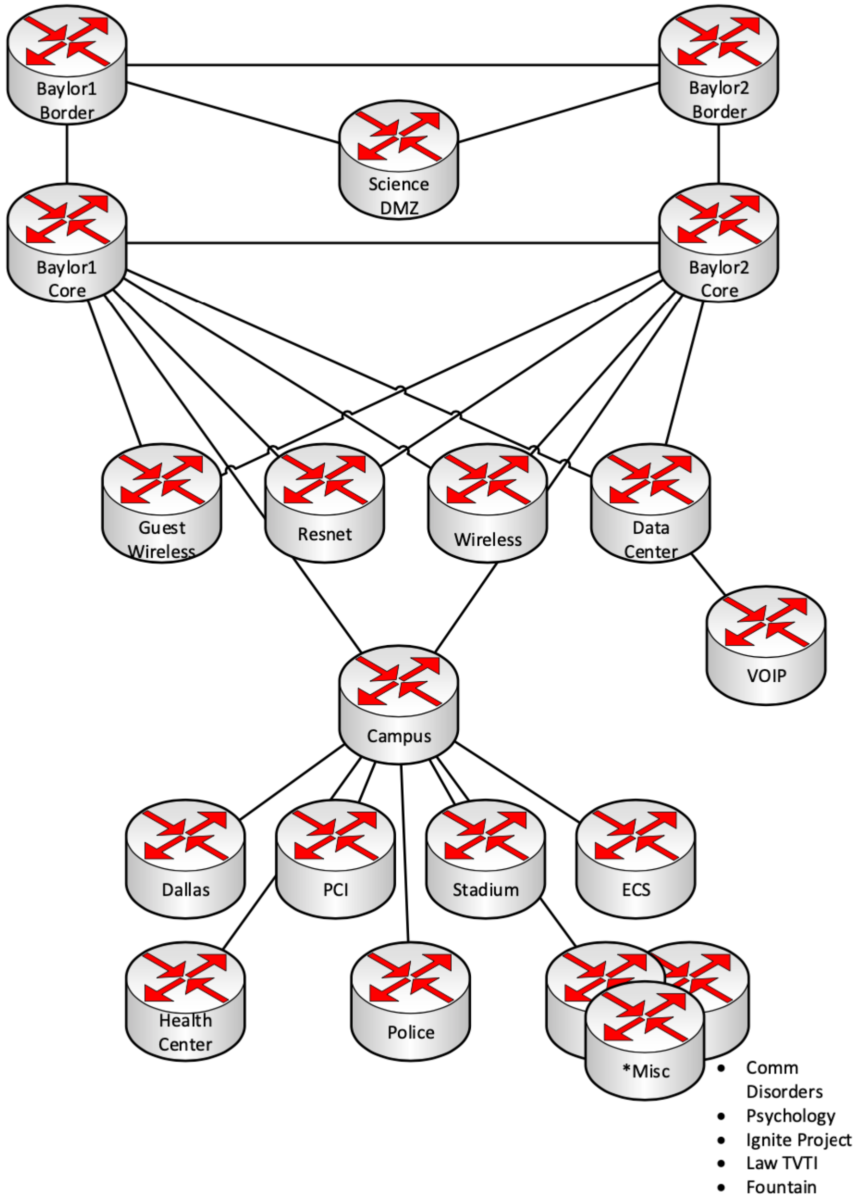

Figure 1 - Baylor Network

LEARN has two POP locations (Baylor1 and Baylor2) on campus interconnected by Baylor fiber. A Grande IRU provides fiber from the Baylor1-POP to LEARN's Waco-Airbase-Rd-POP. A Grande IRU provides fiber from the Baylor2-POP to LEARN's Waco-Clay-St-POP. This provides redundant access to the LEARN network and its resources. LEARN has a router in the Baylor2-POP and a layer 2 presence in the Baylor1-POP. LEARN provides a 2x10Gb connection from their router to a Baylor switch in the Baylor2-POP. LEARN provides an additional 10Gb connection in the Baylor2-POP for additional services and capacity.

LEARN provides peerings for commodity Internet, caching and peering services, Internet2 and DDOS service. Network services to Baylor's School of Nursing (located in Dallas) are provisioned as a VLAN from the School of Nursing, transported over LEARN fiber to the faculty-staff network in Waco.

Grande provides 5Gb of commodity Internet service at Baylor1-POP. Grande also provides two Transparent LAN services (TLS) to campus. The first provides a point-to-point 4Gb connection between our campus datacenter and a remote datacenter location across town to facilitate offsite storage of server backups. The second TLS provides layer 2 access between five off-campus buildings and our faculty-staff network.

Datacenter Network: The campus datacenter is divided into multiple security enclaves. The datacenter network is set up in a three tier architecture with the firewall acting as the core router for datacenter networks. There is a set of distribution switches that forward traffic to top-of-rack switches providing connections to servers.

The datacenter firewall has a 10Gb uplink to the core-routers and a 10Gb port to the more critical networks; remaining interfaces servicing datacenter networks operate at 1Gb. The distribution switches provide 10Gb connections to top-of-rack switches. Top-of-rack switches provide 1/10Gb service to servers. The switches that support server backup and management utilize 1Gb ports.

Baylor University Research Network (BURN): The BURN network is set up in a three tier architecture with the Research-Router at the core. There is a distribution switch that provides a 10Gb connection between the Research-Router and a top-of-rack switch in the HPC rack area. The top-of-rack switch is capable of providing 1/10Gb connections.

Faculty-Staff and Command-and-Control Networks: These networks share the same basic three tier architecture with a campus router, distribution switching that aggregates traffic from buildings, and edge switches within buildings to service end stations. A redundant set of firewalls connects these networks to the main core networks for campus.

The campus router for the Command-and-Control Network connects to the Faculty-Staff campus router providing 1Gb interface to buildings. Dedicated fiber and switching hardware is used to help isolate Command-and-Control traffic. There are separate vlans for the fire alarm, voice evac, and security cameras.

The Faculty-Staff network has redundant campus routers – offering 10Gbs connections to distribution switches (which provide 1Gb connections to campus buildings). Large distribution switches exist at the main Fiber Hubs – aggregating traffic for campus buildings. Some larger buildings have their own distribution switches connecting back to the campus routers. The faculty-staff campus router has two VRFs – one for routing VOIP vlans and one for campus vlans.

RESNET & Campus WIFI Networks: These networks share the same basic three tier architecture with a campus router, a distribution switch that aggregates traffic from buildings, and edge switches within buildings to service end stations. A redundant set of firewalls connects these networks to the main core networks for campus.
The campus WIFI network utilizes routing at the distribution layer. It also has separate networks for management and client traffic.
The majority of the edge switches in the RESNET network are non-POE and are a mix of 100Mb and 1Gb ports to clients.

Present-2 Year Goals:

1. Increase border and core network links from 10Gb to 100Gb.
2. Upgrade data center firewalls.
3. Evaluate use of an overlay network technology across the core/distribution layers of campus to help virtualize network services and deploy network services more quickly and efficiently.
4. Increase network capacities to campus buildings from 1Gb to minimum of 10Gb (preferable 2x10Gb, or 40Gb if needed for higher bandwidth buildings).

3.1.6 Cloud Services

Cloud services are available via providers such as OVH and Azure and can be reached through VPN access. These are available for research use cases, but usage is primarily targeted to enterprise use cases (e.g. SaaS, document storage, backup, etc.).

3.1.7 Known Resource Constraints

There are three known constraints that will impact Baylor ITS in the future:

1. ***Network Connectivity***: Baylor is not at a critical point for connectivity, but is identifying ways to increase capacity to support research and enterprise use cases. Current thoughts are:
    a. Augmenting existing 10G connectivity to LEARN. Hardware can support this today (on both the LEARN and Baylor end), and could be implemented in weeks.
    b. Upgrading equipment to support 100G connectivity. This will involve new hardware on both the LEARN side (Layer 2 and Layer 3), as well as the Baylor side (line card replacement for border devices, and campus upgrades. Upgrade would require multiple months to execute.
2. ***Datacenter space, power, cooling***: There is currently enough space to meet current demands, but strategic planning is needed for identifying ways to meet these needs in the future, particularly as more research disciplines require processing and storage resources.
3. ***Storage***: Faculty, staff, and students all require storage. Cloud options are available, but are not a sufficient way forward. An RFI is underway to

evaluate enterprise storage, with options on how to integrate into the research use cases.

3.1.8 Affiliated Organizations

LEARN provides physical connections in Waco to the main campus and in Dallas to the Nursing School. LEARN has two POP locations on campus (Baylor1 and Baylor2). The LEARN-Baylor2- POP has a router that provides two lagged 10Gbs connections to Baylor. Over this connection we peer with Internet2, caching/peering services, and commodity Internet services. Baylor also utilizes a DDOS scrubbing service from LEARN. The LEARN-Baylor1-POP is an optical site with layer 2 capabilities.

3.2 Experimental High Energy Physics (HEP) Case Study

Content in this section authored by Kenichi Hatakeyama, Jay Dittmann, and Andrew Brinkerhoff from the Department of Physics

3.2.1 Science Background

The Baylor High Energy Physics (HEP) group performs research on elementary particle physics by utilizing data (e.g. proton-proton collision results) obtained from the CMS detector at CERN's Large Hadron Collider (LHC) instrument. The Baylor group's specific work involves searches of new physics principles beyond the Standard Model, as well as the precision of the measurements involving the Higgs Boson or other top Quarks. This research commenced in 2009, during the previous run schedule of the LHC. The schedule for LHC operations will vary in the coming years as the experiment is upgraded to support an era of higher luminosity (e.g. energy intensity). The schedule is as follows:

- 2015-2018: Run 2
- 2019-2021: Long Shutdown 2
- 2022-2024: Run 3
- 2025-mid 2027: Long Shutdown 3
- 2027 and beyond: HL-LHC era

Data from the experiment adheres to a regimented workflow for data distribution. After initial processing at CERN, data sets are distributed to a number of designated facilities that are geographically distributed throughout the world (WLCG: Worldwide LHC Computing Grid). These 'Tier 1s' are well connected to networks, and feature large amounts of computation and storage resources. Further distribution to a large number of 'Tier 2' facilities also occurs, each featuring similar technology profiles, but to a lesser extent. 'Tier 3' facilities, of which Baylor is a part, contribute resources to the overall process of analysis and simulation but are not directly funded to provide resources. This ecosystem facilitates the major parts of the LHC workflow: distributed analysis, storage, and creation of simulation data using a common software framework.

Baylor participates in the creation of simulation data, the analysis of experimental data, and other R&D efforts including use of experimental data in the development of advanced techniques that utilize Machine Learning (ML) to improve the data collection and analysis process. As a part of this process, 200-300TB of data may be resident on Baylor resources at any given time, delivered via software packages that include Open Science Grid (OSG), and the "xrootd" package

3.2.2 Collaborators

The collaboration space for CMS (and the LHC experiment in general) is large. It is estimated that more than 2000 researchers are involved in CMS, spread across approximately 200 sites worldwide. Work on specific aspects of CMS may be smaller, e.g. groups of approximately 50 at a smaller number of institutions may collaborate on single publications that are focused on investigating specific findings from the experimental data.

Data relationships are strongest based along geographical boundaries, as the software packages are designed to “pull” information from the closest resources that may have a copy of the requested information. Within the United States, this typically implies:

- Fermilab, the CMS Tier1 located in Illinois
- Massachusetts Institute of Technology (MIT), a CMS Tier2 located in Massachusetts
- The University of Florida, a CMS Tier2 located in Florida
- The University of Wisconsin, a CMS Tier2 located in Wisconsin
- The University of Nebraska Lincoln, a CMS Tier2 located in Nebraska
- Vanderbilt University, a CMS Tier2 located in Tennessee
- California Institute of Technology (Caltech), a CMS Tier2 located in California
- The University of California San Diego, a CMS Tier2 located in California
- Purdue University, a CMS Tier2 located in Indiana

3.2.3 Instruments and Facilities

The main instrumentation utilized for the Baylor HEP group is data produced at LHC, and captured by the CMS detector. Given the remote and shared nature of this resource, Baylor does not have direct control over operational methods or schedules. Data is consumed when it is available, and simulations are produced as required.

Experimental data undergoes initial analysis at CERN before being widely shared, and is reformed into a well-defined format that software packages understand: Analysis Object Data (AOD) which is a form of the “root” (https://root.cern.ch/) data format. A new format (smaller file sizes to facilitate easier sharing) of this analysis data is primarily used by Baylor researchers: nano-AOD (https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookNanoAOD).

Typical usage at Baylor involves study over multiple data sets collected over many runs. For the recent analysis of data collected over three years (2016, 2017, and 2018), Baylor used 240 TB of data for 1.7 million files (~100MB/file). This analysis was used to further novel Machine Learning (ML) code development, an R&D effort that will create improvements to aspects of the LHC software stack.

Current computation and storage resources at Baylor are scaling to the immediate needs of the research group. However, future upgrades to the LHC will result in data size increases: file sizes, data sets, and experimental run times are all expected to increase after scheduled shutdown and upgrade cycles. Data collection rates are expected to increase by a factor of ~5, requiring a similar growth expectation for storage and computation requirements. As a Tier 3 site, Baylor is not funded by the experiment for upgrades, but will closely monitor the output of activities such as the

Coordinating Panel for Advanced Detectors (CPAD) workshop (https://wp.physics.wisc.edu/cpad2019/).

Beyond the data growth, there is significant discussion within the LHC community regarding alterations to the computational model. “Heterogeneous computing”, e.g. the simultaneous use of CPU, GPU and/or FPGA, has been explored via R&D efforts as a way to utilize a broader set of resources available within the WLCG. It is unknown at this time how this will impact Baylor’s contributions to the experiment.

3.2.4 Process of Science

There are three primary workflows within CMS:

- Analysis
- Simulation
- Re-processing

The analysis workflow is the process of downloading selections of the pre-processed experimental data for a given run, and executing analysis code. The outcome of this process is to determine if any ‘events’ of interest were observed, e.g. the presence of particles of interest, and their relative behavior during the collision. The WLCG performs this step constantly during an active run, and returns results back to the central collection mechanism. Specific actions during this process involve the use of Baylor computing resources (e.g. the Kodiak cluster), running the OSG software stack, to download nano-AOD formatted data from the nearest available resource (e.g. the Tier1 or Tier2 centers in the US). The act of data ‘download’ has flexibility; the software supports two major modes of operation that include bulk download along with the option to stream data remotely during processing. Baylor utilizes both, but often prefers the former since streaming can sometimes be slower due to occasional performance abnormalities (discussed in 3.2.11 Outstanding Issues).

The simulation workflow has two major components: production and analysis. Production is the creation of Monte Carlo (MC) data that ‘simulates’ experimental data. The data set size is meant to simulate what would be captured during a live experiment. This simulated data is then used (locally, as well as shared more widely) to validate and improve the software stack used for analysis. MC production and validation is controlled centrally, thus Baylor’s role in this process is controlled by the CMS experiment and designed to utilize spare resources when they become available.

Lastly, the act of re-processing is similar to that of analysis, but involves older data from previously run experimental runs. Re-processing actively occurred during LHC downtime to re-analyze the entire data set to ensure that no experimental result was left out of findings, along with validating improvements to the software stack.

CMS data stored locally resides on the research storage of the Kodiak cluster, and currently is less than 300TB total. Due to expected data growth in the 2-5 year time frame, the size requirements will easily double, and could reach 5x if the predictions of data growth prove true.

3.2.5 Remote Science Activities

The nature of the LHC workflow pipeline implies that all core analysis for the experiment at a whole is done in a distributed fashion, thus the process is deeply integrated to capabilities and performance of the network. It is typical to either perform a bulk-download, or live streaming, on data sets stored at major facilities (CERN, T1s, or T2s) using local computational resources. Results are then stored back into the global computing infrastructure (e.g. Worldwide LHC Computational Grid [WLCG]).

Most analysis work that Baylor performs is done using local resources, after the download (e.g. bulk or streaming) of the research data from the collaborating sites. Options to utilize computational allocations at the Tier1 center (Fermilab) exist, but do not constitute a significant computational or storage resource used in the process of research.

3.2.6 Software Infrastructure

There are two primary use cases for research software:

- Analysis of research data, which consists of downloading and processing results from prior LHC runs
- R&D activities to explore Machine Learning (ML) approaches to data analysis

The primary software package used for analysis activities is developed and maintained by the Open Science Grid (OSG) effort. The Baylor University HEP group uses components of this to manage aspects of the workflow:

- Data location, download, and curation: PhEDEx, or manual transfers using Grid Community Toolkit (formerly GridFTP). Future data transfer requirements will be migrated to a new tool: Rucio. This tool has not fully been adopted throughout the collaboration, but is expected to replace existing approaches in 2020.
- HPC processing of data and execution of analysis codes

The R&D activities the Baylor University HEP group performs involve the development of novel methods to use Machine Learning (ML) for the process of data analysis. These software components are developed in C++ or Python, and managed via ‘Git’ repositories maintained by Baylor faculty, staff, and students.

3.2.7 Network and Data Architecture

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

The Baylor University HEP group utilizes institutional HPC and storage resources for the majority of the analysis workflow. This is primarily Kodiak, the institutional cluster.

Remote computation and storage is also available via a small allocation of resources that is made available at the CMS Tier1 (Fermilab). Due to the size of the allocation, it is infeasible to use this as a primary resource, thus it is often used only if local resources are not available.

3.2.8 Cloud Services

The CMS experiment as a whole has R&D efforts to explore the viability of commercial cloud resources used for analysis, but the Baylor University HEP group is not involved in this effort. There are no current plans to utilize commercial clouds for portions of the analysis workflow, or software development process.

3.2.9 Known Resource Constraints

Networking is a significant portion of the workflow, and the efficient download of data will impact productivity in the future. A description of this problem is available in 3.2.11 Outstanding Issues.

Storage is also a significant portion of the workflow, and as data sets grow in size the requirements for long term storage will increase.

3.2.10 Parent Organization(s)

All analysis is facilitated via support from the Baylor's High Performance and Research Computing Services group (HPRCS). The Baylor HEP group does not maintain any additional HPC resources to support the operation of the Tier3 center.

3.2.11 Outstanding Issues

On occasion there have been challenges in receiving research data from the CMS collaboration. This is not a routine situation, but has impacted productivity. For instance, when copying even a small amount of data (e.g. several GB) from Fermilab, and using the gfal-copy tool via xrootd protocol, it may take ~5 minutes (e.g. 25 Mb/s).

The problem can be summarized as:

- Automated tools (such as PhEDEx and XRootD) locate where data sets are located at other participating CMS sites
- During the bulk download or streaming process to retrieve data, networking problems cause significant slowdowns (e.g. Mbps of throughput)
- When the problem is reported/investigated, the automated tools are typically able to find other copies of the same data set that can be downloaded 'faster', and thus re-configure the process dynamically
- Investigation into cause never proceeds to identify if the problem is:

- Local to Baylor HPC
- Remote to source of data

An investigation into the cause of the problem is requested.

3.3 Proton Computed Tomography (pCT) Case Study

Content in this section authored by Keith Schubert from the Department of Electrical & Computer Engineering

3.3.1 Science Background

Proton Computed Tomography (pCT) is working on developing a new medical imaging modality through the use of proton (or ion) computed tomography methods. Ions can be used to image the body with only a few percent of the radiation damage of a normal x-ray image, but with greater accuracy for treatment with ions, since the relevant quantities (relative stopping and scattering powers) are directly measured. The major reason for pCT is this latter case, the planning and verification of proton or ion radiation treatments.

3.3.2 Collaborators

Currently imaging data sets are being produced at a few locations around the world:

- Chicago's Northwestern Medicine Chicago Proton Center
- Heidelberg
- Loma Linda University

Monte Carlo Simulations (used for calibration) are also produced locally, as well as the locations above.

The data sets (real and fabricated) typically involve sets that describe the "tracking" (direction, speed, etc), timing, and energy measurements for 360 million to 2 billion protons/ions, usually of at least 9 total data elements. The compressed data is stored locally, and used to produce image reconstruction software, improve the generation of simulated data sets, and contribute to the work on various treatment planning methods.

There are dozens of universities that cooperate in the process, but not to the levels of data movement as the groups above. Active participants include:

- Loma Linda University (LLU)
- University of California Santa Cruz (UCSC)
- Northern Illinois University (NIU)

Others that have participated in the past include:

- Stanford
- University of California San Francisco (UCSF)
- University of Haifa
- University of New South Wales (UNSW)
- Ludwig Maximilian University
- University of Manchester
- State University of New York (SUNY) Stonybrook
- City University of New York (CUNY)

3.3.3 Instruments and Facilities

Imaging requires an accelerator (cyclotron or synchrotron) and our trackers and detectors. Data is gathered on a local machine (typically in a hospital) and preprocessing is done there and a compressed file sent to Baylor for storage.

Often storage is a removable hard drive at the hospital facility, that is then 'removed' and transported to a partnering local university, where it is uploaded to Baylor. This is done nominally because the IT infrastructure of hospitals is not as sophisticated, or too secure, to facilitate data sharing. For example:

- NIU handles the Chicago region
- Ludwig-Maximillian University handles the Heidelberg region
- Loma Linda University handles LLUMC

Connections are slow in some cases (that do not have a partnering local university) and it often takes a couple months before some data sets are completely uploaded. In rare instances, Baylor staff will travel to partnering institutions to retrieve data versus having to wait longer periods of time for a bulk upload.

Many collaborators log into our systems to download data, and a few collaborate by downloading data sets. Data sizes (after triggering and compression) are currently 10-100GB. Increased numbers of ions, or longer experimental runs, have the potential to increase the data set sizes. Due to the nature of the collected data, it is highly compressible which facilitates easier storage and transfer.

3.3.4 Process of Science

Data is produced at a proton/ion treatment center or on a simulator, and transferred to Baylor, where it is stored and made available to collaborators. Reconstructions are done at the site or here, and from there the individual process can vary.

Currently data must be shared manually (exchange of hard drives) due to performance problems of hospital / clinical network infrastructures. The Baylor team is developing hardware/software solutions to facilitate easier data exchange from the collaborators.

3.3.5 Remote Science Activities

The generation of real data is all remote, using the instrumentation that is designed and built at Baylor.

Getting data back to Baylor can be challenging, due to the nature of the remote networking infrastructure. E.g. a hospital network is designed for protection of sensitive info first and foremost: expedient transfer of large data sets doesn't fit this profile typically. Baylor University staff have designed the data collection infrastructure to facilitate data reduction when applicable, but sometimes removal and shipment of hard drives may be required (or worst case – physical visits to retrieve data). In some cases, hard drives from a clinical environment can be removed and shared ‘locally’, e.g. a clinic can be nearby a well-connected university. Once the hard drive arrives at a more capable location, remote data access is possible.

3.3.6 Software Infrastructure

We use several different Monte Carlo simulators including Geant4, and specialized medical simulators. We generate our own in-house reconstructions. Treatment plans are typically done remotely.

3.3.7 Network and Data Architecture

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

3.3.8 Cloud Services

There are no sensitive aspects to the data purposefully: it is de-identified immediately. It is a project goal to avoid use of sensitive data above all else. This being said, a migration to the cloud is possible, but is being avoided in favor of workflows that are local and processed by Baylor directly.

3.3.9 Known Resource Constraints

Passing data from original data sources are the main problems, then distribution to various collaborators.

3.3.10 Parent Organization(s)

Support from Baylor ITS has been helpful in using HPC resources.

3.3.11 Outstanding Issues

Data retrieval may take months for particularly large data sets at poorly equipped facilities.

3.4 Nutrition and Relation to Digestive Microbiome Case Study

Content in this section authored by Leigh Greathouse from the Robbins College of Health and Human Sciences, Department of Nutrition Sciences and Biology and Health and Human Performance

3.4.1 Science Background

The research focus is to identify biomarkers and mechanisms within gut microbiome and diet that can be used to reduce incidence of and improve survival among individuals diagnosed with colon and lung cancer. The main source of data is generated from sequencing of human (and some mammalian) tissues and fecal samples. In addition, it is routine to download and store large population microbiome datasets from similar work, from which to conduct meta-analyses for biomarkers as leads for potential hypotheses and mechanisms for further study.

As the collaboration space has grown, the use of a custom "data pipeline" created for the analysis of large microbiome sequencing datasets has grown. The workflow involves mechanisms to pre-process, analyze, and post-process data sets in an automated fashion using Baylor computational resources (e.g. Kodiak). The workflow is able to process locally created/curated data sets (sequenced off-site, but using samples initiated by Baylor), as well as the downloaded external datasets.

Processing involves bursts of activity every 3-6 months, depending on the project status. Collaborators are given permission to use our data for a period of time until the project is completed, but access directly to the data is controlled locally.

3.4.2 Collaborators

Collaborators/Contractors:

- James White (contractor; Baltimore, MD) – bioinformatician that has created and update the bioinformatics pipeline, interface with collaborators to conduct microbial sequence analysis, and download large sequencing datasets as needed from public databases or from other collaborators
- Garth Ehrlich/Josh Mell (collaborators; Drexel University, Philadelphia, PA) – microbial geneticists that conduct microbial sequencing and processing; they provide raw sequence datasets that get transferred to storage on Baylors's Kodiak cluster for analysis
- Joseph Petrosino (collaborator; Baylor College of Medicine, Houston, TX) – virologist and director of the microbiome sequencing center at Baylor College of Medicine (BCM). This group processes samples for sequencing and returns both raw and processed sequences for analysis. These are usually stored at a BCM database, but can be downloaded and transferred to Kodiak for further processing as needed
- Philip Abbosh (collaborator; Fox Chase Cancer Center, Philadelphia, PA) – Urologist that submits both public and self-created datasets for processing on the data pipeline using Kodiak; storage of these data are brief and only for analysis

- Nick Chia/Jun Chen (collaborators; Mayo Clinic, Minneapolis, MN) – microbiologist, biostatistician, and director of the microbiome research center; share processed sequencing data for further statistical analysis
- Aadil Sheik (graduate student) – utilizes bioinformatics pipeline on Kodiak to process microbial sequences and conduct analysis as needed

3.4.3 Instruments and Facilities

Most, if not all, research utilizes the Kodiak compute facility to store and analyze data sets (locally produced, or downloaded). Dataset sizes vary between 2GB to 10TB currently. Input data comes from public resources of collaborators, or the output of sequencing/analysis that are performed offsite (there is currently no mechanism to sequence locally). Sequencing occurs at Baylor College of Medicine, Drexel University, and Mayo Clinic primarily (other sites/collaborators are always possible). There are no upgrades or new facilities planned.

Research data is primarily stored on Kodiak, which has roughly 1-12 TB at any given point in time. Compression is utilized to save space, and condense sets that are not in active use. At most 20TB could be required during busy research periods. This volume may increase by up to 2 times in the coming years as more ensemble data sets become available for meta-analysis.

Since all sequenced data is produced externally to Baylor, this storage is critical and often the only copy available. If sequencing instrumentation is purchased/operated locally, more storage will be required to keep up with demand. The group expects to have a dedicated sequencing machine within the next 2-3 years that is shared between laboratories; this would imply that local sequencing of multiple small samples over the year (#5-10/month) and potentially large sample sets (#100-200) once every 2-3 years would be routine. Total data set sizes will scale to be 10s - 100s of TBs of persistent local storage required.

3.4.4 Process of Science

We expect to generate two types of microbial sequencing data over the next 2-5 years;

1. microbial sequencing from humans with colon cancer throughout the course of treatment and follow up for discovery of diet-microbiome interactions in predicting response to therapy, and
2. microbial sequencing of Bacteroides fragilis spp. for discovery of small RNAs that are used to communicate with host cells to change signaling and pathogenicity.

To conduct #1, we will collaborate with Joe Petrosino at BCM to send stool samples (N=600) for sequencing and bioinformatics processing, which will be transferred to Kodiak for storage and additional processing by our group. We will also work with the ColoCare Trial consortium to transfer their data to Kodiak for analysis and integration with our data to improve machine learning capabilities and prediction.

To conduct #2, we will collaborate with Garth Ehrlich and Josh Mell at Drexel University, and will send them 12 RNA samples for RNA-seq for two different projects over the next year. They will conduct RNA-seq, process the data and conduct the analysis. They will also share the raw data with us during this analysis, which will be stored on Kodiak.

The majority of data analysis for 16S rRNA sequencing (#1) will undergo standard read trimming and primer extraction, followed by filtering to human reads, and QC for read quality; followed by alignment to reference microbial genome databases (SILVA, Greengenes); then assignment of reads to microbial taxa. These assigned reads will then be used to conduct biostatistics and microbial ecology testing (e.g. alpha diversity, beta diversity) in conjunction with clinical metadata (e.g. age, sex, ethnicity). These data will also be integrated with data from a larger similar clinical study (ColoCare) to increase the power of our statistical analysis.

The RNA-seq data will be processed and analyzed by Josh Mell at Drexel University using a variety of free tools, including the STAR alignment tool. The reads will be QC'd, trimmed and filtered, then aligned to reference genomes. We will then characterize the RNA species found and expressed in each sample for discovery of small RNAs that could contribute to host communication.

Will use mostly sequencing and processing facilities/resources at BCM for #1, and Drexel University for #2, with anticipated storage of sequencing data on Kodiak from both (~10TB)

Will likely continue to utilize these two collaborations and their facilities for the majority of our sequencing and processing needs as they are very well positioned for these types of studies and have the resources to stay up with all of the sequencing and bioinformatics tools. Beyond 5 years, if Baylor University has the sequencing and bioinformatics resources, we will likely move towards sequencing and processing of our own data in house.

3.4.5 Remote Science Activities

Currently we only use genome sequencers at other facilities via collaborations – those listed above at Drexel University and Baylor College of Medicine (BCM). We send them samples, they are processed and stored on their servers, and they share the raw sequences with us as needed for further analysis. The use of external resources is still inexpensive compared to the costs of local purchase/maintenance of sequencing hardware. Once it becomes cost effective (and data volumes increase), a move to local control will be required.

We will likely continue to use both Drexel and BCM for our remote sequencing needs through these collaborations. They are highly efficient and have the ability to keep up with currently technology changes financially.

3.4.6 Software Infrastructure

Data transfer occurs using both Globus and SCP, depending on the endpoints involved.

Analysis is performed using a combination of free/publicly available bioinformatics tools including:

- Mothur[2]
- QIIME2[3]
- Dada2[4]
- Deblur[5]
- Picrust[6]

We will likely upgrade our bioinformatics pipeline within the next 2-3 years to replace with newer versions, or similar products that are Free/Open Source Software (FOSS), as the field advances.

3.4.7 Network and Data Architecture

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

3.4.8 Cloud Services

We don't intend to use any cloud services at the present or in the future beyond enterprise use cases (e.g. use of Box/Drive for sharing small (~GB) sized attachments that may not be possible to transmit via mail.

3.4.9 Known Resource Constraints

The major constraints we foresee will be the following:

- ***Storage*** – as our datasets grow we will likely need to have more storage, upwards of 50TB within the next 5 years
- ***Bioinformatics*** – it would be very helpful to not only ourselves but also other faculty/labs to have dedicated bioinformatics staff that can help process large datasets and maintain bioinformatics sequencing pipelines as needed
- ***Data transfers*** – we will need to be able to more efficiently and rapidly transfer datasets between our compute facility and that of our collaborators that can handle large datasets >1TB.

The group frequently applies for grant applications to upgrade core-capabilities, including the analysis pipeline. Upgrades to increase speed/efficiency, integrate

[2] https://mothur.org
[3] https://qiime2.org
[4] https://benjjneb.github.io/dada2/
[5] https://github.com/biocore/deblur
[6] http://picrust.github.io/picrust/

new tools, and create a dedicated data movement capability are planned, but not funded at this time.

3.4.10 Parent Organization(s)

Currently, we are using about 60TB worth of storage on Kodiak. Unfortunately, only a few faculty are capable of the type of bioinformatics work we need assistance with, but they are already dedicated to their own projects and cannot really collaborate with us at this time due to their own labor/time constraints. However, we are trying to pursue the University to fund a Translational Microbiome Research Institute, which would integrate computational facilities, dedicated bioinformaticians, and statisticians for managing microbiome studies among Biology, Biochemistry, Environmental Sciences, Nutrition, and Public Health.

3.4.11 Outstanding Issues

We do use high performance data transfer tools including Globus once or twice a year to transfer or download large datasets. The main issue we have been running into is the ability to transfer large datasets from the International Cancer Genome Consortium data storage facilities that house large sequencing data sets of cancer cases, which is not a problem with the The Cancer Genome Atlas house at the NIH. Thus having tools that can easily integrate with our computational facility to handle large (>10TB) dataset download and transfers in the future would be very beneficial.

3.5 Baylor University Core Research Facilities Case Study

Content in this section authored by Christopher Becker, Director of Baylor Sciences Building and Baylor Mass Spectrometry Center

3.5.1 Science Background

Baylor University features five "core research" facilities that serve the college of Arts & Sciences:

- The Baylor Mass Spectrometry Facility
- Animal Vivarium
- Biomolecular Sciences Center,
- Center for Microscopy and Imaging
- Nuclear Magnetic Resonance Center

Together these cores make up a significant portion of the scientific instrumentation for the campus. The generated data supports federal, industrial, and startup research primarily in A&S supporting the Departments. of Biology, Chemistry & Biochem., Environmental Sciences, Geosciences, Physics, and Psychology and Neuroscience.

Currently each core has a different mechanism of handling data. Some instruments facilitate direct sharing of data via intelligent controllers (e.g. an exposed file system mount or a file transfer mechanism). Others may feature access ports (e.g. USB) for connection of removable storage. A small number are not as sophisticated, and feature software that simply mails results after completion. This lack of a uniform mechanism has hampered efforts to standardize and centralize an approach to data curation and mobility.

In most cores there is an (unenforced) policy that that no data should be stored on the instruments, to prevent liability should something be deleted; this also has the benefit of requiring less local storage resources. Despite this, it is routine to offload stored data to external storage as it fills up over time. Ideally the "chain of custody" for data stays with the end user at all times and they transfer data out of the core PC when done with an analysis, but typically that data is transferred as a copy with the original left behind at the instrument PC.

3.5.2 Collaborators

We directly collaborate with the Veteran's Administration (VA) and have active grants and other projects with collaborators elsewhere in Texas and in Europe. We regularly need to interface with instrument vendors across the US and in England/Germany and enable them to remotely access our equipment to provide software and instrumentation support.

Our 5 Core Research Facilities aserve approximately 60 faculty, and 200-300 end-users, which frequently need to transfer data to their home labs in the building and to other collaborators across the country.

On several occasions we have needed to have out of state collaborators interface with Baylor researchers and an instrument during an experiment in progress. On these occasions we have typically used the remote desktop application Team Viewer [7], or Skype[8], while monitoring the activity.

3.5.3 Instruments and Facilities

The five facilities are:

- The Baylor Mass Spectrometry Facility
- Animal Vivarium
- Biomolecular Sciences Center,
- Center for Microscopy and Imaging
- Nuclear Magnetic Resonance Center

In addition, the campus features three other[9] shared instrumentation labs for physical, biophysical and analytical instrumentation. These labs are supported by PhD level staff that offer training and consultation for all researchers (graduate level or higher) on any of the equipment we have.

Instrumentation includes:

- 5 NMR (Nuclear Magnetic Resonance) spectrometers @ 300-600 MHz
- 4 High Resolution microscopes
- 10 chromatography mass spectrometers
- Flow cytometry and cell sorting instrumentation
- numerous light spectroscopy instruments
- Circular Dichroism (CD) spectrometer
- Calorimetry instruments
- High-Performance Liquid Chromatography (HPLC) instruments
- X-Ray instruments
- Electron Paramagnetic Resonance (EPR) instruments
- Other minor instrumentation and equipment supporting the above areas of interest

Instrumentation is upgraded as base funding and grants allow. Current work includes major renovations to the Mass Spectrometry Center, the Center for Microscopy and Imaging, and the Molecular Biosciences Center. An NSF MRI grant application has been submitted for a new 2.7M Transmission Electron Microscope in the Center for Microscopy and Imaging.

[7] https://www.teamviewer.com/en-us/

[8] https://www.skype.com/en/

[9] https://www.baylor.edu/bsb/index.php?id=929341

3.5.4 Process of Science

Most of our instruments are scheduled (and usage is logged/invoiced) using the Baylor Facility Online Manager (FOM) scheduling tool. All usage is recorded to better gauge patterns, requirements, and justify funding for future upgrades.

Networking is critical to the facilities, despite not all instruments being fully integrated into an automated workflow. Some instruments are fully networked; others have no network connectivity due to security reasons (e.g. OS updates not being possible to accommodate specific instrument hardware/software requirements). Networking plays a large role in scheduling and collaboration, as well as in receiving support when an instrument is down. In the past one of our primary pain points has been getting remote support for important instruments with active projects that are not permitted to be connected to the internet due to security concerns.

In addition to the instrument PCs affiliated with the control of each device, there are individual processing PCs dedicated to specific analyses and specific software programs for data processing. The facility has limited local storage and processing power for any calibration-level computations. Longer term computational/storage usage cases (e.g. simulation or analysis) must be done offsite.

File and data set size varies by instrument. At the upper end we have instruments under heavy use that will generate up to 20 2GB files in a day. Users typically are able to manage these datasets with portable/removable media or through local file transfer to other parts of Baylor (e.g. Kodiak). Future instruments (e.g. protein reconstruction on a transmission electron microscope [TEM]) have the potential to generate up to several terabytes of data in a 2-3 day period.

3.5.5 Remote Science Activities

As described above, we frequently need vendor support for our instruments and on occasion collaborators need to be able to access ongoing experiments to make real-time judgement calls on moving forward. End-users also want to remotely access instrument PCs overnight to check on runs and/or look at data, which can cause disruption if someone else is currently using the instrument. The use of remote access software is not routine, but does occur in these rare cases. There is strong desire from the user community to allow fully automated remote use, but this is not a high priority given the security and communications implications it will require.

A core facilities goal over the next 5 years will be to replace/convert a majority of instrumentation to be connected to at least the local university network, be capable of receiving remote vendor support, and in many cases also have access to the full internet to facilitate real-time research, collaboration, and data transfer with other institutions/collaborators. Overcoming complications between internet access and updates that sometimes breaks an instrument could be a significant hurdle.

3.5.6 Software Infrastructure

Software usage depends heavily on the instrument. Typical patterns include:

- Proprietary vendor software that uses SMTP and the Baylor MS Exchange server to automatically email data files directly to users
- Embedded software systems on control PCs that can (but often shouldn't) be connected to the network
- Custom analysis packages installed on processing PCs
- The aforementioned use of collaboration tools: Team Viewer, Skype, and other phone/video conference options.

Given the lack of a uniform instrument layer or process to handle data, removable media and email dominate the data transfer space. Some users are sophisticated enough to utilize enterprise cloud storage (e.g. Box) when available. Others can take advantage of shared local file storage (e.g. mounting a shared drive via the Baylor LAN).

This later option, file share on the local network for a few instruments, has not gone smoothly in the past. Frequently users will not retrieve or backup data sets, and performance via this sharing mechanism can be slow. It is desired to have a mechanism to easily share data back to user laboratories without multiple connection, log-in, and authentication steps and without monopolizing instrument time for data transfer would be very beneficial.

3.5.7 Network and Data Architecture

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

Specifically, the centers are located in the Baylor Sciences Building on the southeast side of campus.

3.5.8 Cloud Services

We use scheduling software (FOMS) that is hosted on a local Baylor server. We also utilize enterprise cloud storage (e.g. Box) for data transfer/storage, and store our chemical inventory on CiSPRO cloud. We do not have any specific plans for additional cloud services, though we have discussed the possibility of using AWS or other private clouds for future data storage and/or backups. Cloud-based processing is not being actively explored, or asked for by users.

3.5.9 Known Resource Constraints

Our primary constraint is keeping instrument computers networked in the face of an update cycle that can either

A. disrupt in-progress experiments (i.e., with forced reboots shutting the pc down or substantially damaging data sets because of processor time that was shared with update download and/or install processes) or
B. render an instrument unusable because an update breaks the software that controls the machine or the PC itself breaks.

This seems to be a common problem with our industry, and Baylor IT has been working with us to identify a solution. In the past, non-networked PCs were less of a concern, but collaboration, automatic backup requirements of grants, data sharing, and online-instrument-support has made an enormous push to “full connectivity” in the last 5 years and that effort is still growing. Networked computers are also needed to effectively implement the billing infrastructure that the university is mandating for our instruments (we will need at least internal networking maintained for most of this equipment to function with our FOMS calendar and billing software).

Related to the above is our need for improved management of data transfer, data storage, and data backup. A solution that will seamlessly make data accessible to end users without tying up instrumentation resources with excessive analysis and transfer times will be an important component of instrument-time management as our userbases continue to grow. Another issue is that almost all of our resources require manual backups and storage of data. We need an automated process so that backup and storage can be a reliable safeguard to the enormous amount of funding and research that goes into the collection of that data.

3.5.10 Parent Organization(s)

Our office reports to the Associate Dean of Research for Arts & Sciences and collaborates closely with the Assistant Vice Provost of Research Facilities. We are not partnered with grant-funded facility upgrades.

3.5.11 Outstanding Issues

We have had concerns over the timeline for approval of new technologies (i.e. the turnaround time on security reviews over both custom and commercially available software packages). The approval process is not always transparent, and can on occasion be onerous for users. Communication and expectation management should be improved.

3.6 Molecular Quantum-dot Cellular Automata (QCA), and Material Science of Quantum Computing Case Study

Content in this section authored by Erik Blair from the Electrical and Computer Engineering Department

3.6.1 Science Background

This case study covers two interrelated areas of research, both areas are heavily dependent on simulation and thus are strongly tied to the computational and software environments provided by Baylor University:

- Molecular quantum-dot cellular automata (QCA)
- The material science of quantum computing

Molecular QCA (mQCA) is a general-purpose, classical computing paradigm for the post-Moore's law era. It is designed to provide energy-efficient, high-speed general-purpose computing. The research group works to develop models and theories related to quantum phenomena pertaining to mQCA devices and circuits. This work is typically performed using the MATLAB software package to model dynamic quantum processes in circuits and devices. Emerging work also involves the use of ab initio modeling of molecules at the atomic scale to explore the design of candidate QCA molecules. This latter research is deeply tied to high-performance computing , and the use of software such as Q-Chem, NWChem, and Gaussian.

The second research area involves the material science of quantum computing. The research group performs the aforementioned ab initio modeling of point defects in semiconductors. Doing so involves using HPC resources to run the Quantum ESPRESSO package currently, and will shift to use VASP (Vienna Ab Initio Simulation Program) in early 2020. The overall goal of the research is to determine conditions under which it is possible to create stable, point defects in ZnSe for quantum information processing applications.

3.6.2 Collaborators

At the current time, these research activities do not involve any external collaborations groups to Baylor University. The primary team is located on site, and consists of the PI, and graduate student researchers (3 Ph.D. candidates, and 1 MS student). The student researchers are the primary users and developers of software for use on the HPC resources.

3.6.3 Instruments and Facilities

The process of science is heavily dependent on simulation. This requires access to a variety of HPC resources on campus that can run certain programs that the group uses to model the simulations.

At current time, there are two computational clusters at Baylor:

- The PolarBear Cluster

    - This is a liquid-cooled tank the research group purchased directly but is managed by the Baylor University Research IT team. The infrastructure was designed originally to support MATLAB development/simulation, which is more bound to high-memory than to high computational requirements.
    - The infrastructure consists of:
        - Four (4) Intel(R) Xeon(R) Gold 6244 CPUs
            - 3.60GHz
            - 8 cores each
        - Four (4) Intel(R) Xeon(R) CPU E5-2690 v2
            - 3.00GHz
            - 20 cores each
- The Kodiak Cluster
    - The group uses this resource for quantum material science calculations for jobs that exceed the capabilities of PolarBear.
    - Many of the calculations can scale easily to higher core count/machine count resources and are not memory bound. For instance, running Quantum ESPRESSO or custom developed Python scripts, it is possible to consume 200 or more cores.

The research group anticipates outgrowing resources at Baylor for larger simulation activities, and will look to apply for time at larger facilities (e.g. TACC). Frontera and other systems there support 100s of cores/machines, and can run the same software stack that Baylor supports.

The workload of simulation is not input-data intensive. Minor variations in input set may produce different results, but at this time it is not anticipated that the bottleneck to research will be data related. Access to computation will continue to be the largest challenge. As such, storage is not a primary concern. Checkpoints from long-running jobs are periodically saved, but not saved for long periods of time. These can be several MB to a GB in size. The output of simulation has the potential to grow in the coming years to be 100s of GB, particularly when analysis is used to create visualizations. The produced data is not subject to any privacy concerns.

3.6.4 Process of Science

Present – 2 years

In all cases, our data is the result of model calculations of quantum systems at the atomic level or at the device level:

- MATLAB models of quantum phenomena
    - Typically, we use the MATLAB parallel computing toolbox (PCT) and the PARFOR construct
    - Often these calculations do not parallelize well
        - across nodes because they involve high memory usage
        - across time steps because they are interdependent
- Quantum Chemistry/Quantum Physics calculations. These parallelize well

- Quantum ESPRESSO
    - This is open source software. We model the electronic structure of crystals and crystal defects at the atomic scale
    - Calculations can be very large, and we would like to be able to use a large number of cores here.
- VASP
    - This software is presently under procurement
    - Calculations are similar to Quantum ESPRESSO calculations
- Gaussian, Q-Chem
    - This is parallelizable
    - We use a site license for Gaussian and a research-group-only license for Q-Chem

The types of calculations for the current technology horizon and (next 2-5 years) and strategic planning (5 years and beyond) remain the same as outlined above.

3.6.5 Remote Science Activities

We may seek accounts on the Texas Advanced Computing Center (TACC) as the need for parallel resources arises beyond those available locally at our institution. There is not a need (now or in the future) to utilize remote instrumentation beyond computational capabilities.

3.6.6 Software Infrastructure

Data sharing between members of the research group can be done in a number of ways:

- File transfer (typically using FTP or SCP). Programs like FileZilla or CyberDuck have also been used to transfer files (locally, and remotely).
- Sneakernet (e.g. USB key between local resources)
- Cloud sharing (e.g. Google Drive)
- Github (used for software version control primarily)

Future collaboration will involve access to known remote data resources, such as the MaterialsProject.org[10]. Data sets from this resource will be downloaded and used to train Machine Learning (ML) infrastructure that will be developed to automate some aspects of mQCA research. Downloading data sets from this infrastructure is expected to use web-portal software.

Currently the steps of postprocessing use custom-build scripts in Python or MATLAB. Developed software is stored/tracked within Github (tied to projects that are maintained by PI and/or students).

[10] https://materialsproject.org

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 use is currently restricted to enterprise use cases (sharing between collaborators, or storage/backup purposes). There are no plans to explore cloud-based computation.

3.6.9 Known Resource Constraints

There are none at this time - the scope of work fits the available technology resources.

3.6.10 Parent Organization(s)

This research relies heavily on Baylor ITS / high-performance computing staff, who install and troubleshoot software and assist with the management and maintenance of our experimental liquid-cooled cluster.

3.6.11 Outstanding Issues

Given the highly-parallelizable nature of this work, the 'Condo' model of computing is appealing. Adding more resources to Polar Bear or Kodiak to support other research groups, while allowing our needs to burst to available resources, is desirable. HPC requirements will increase in the coming years as the number and complexity of simulations increases. Many of the software packages that are used can scale as the number of cores/CPUs are added to a simulation. The research group will explore options (depending on funding streams)

An open area of research is the use of GPUs. Baylor maintains a small GPU-based infrastructure, but it is not used for this research currently. It is expected that some future work will explore using that hardware for either simulation, or ML research.

3.7 Modeling and Simulation of Low-Dimensional and Nano-Structured Materials Case Study

Content in this section authored by Kevin Shuford from the Department of Chemistry and Biochemistry

3.7.1 Science Background

The Shuford Group is a theoretical/computational research team that investigates interdisciplinary topics spanning chemistry, physics, materials science, and engineering. Our primary research interests are modeling and simulation of fundamental processes in low-dimensional and nano-structured materials. Established research areas include ultrafast quantum dynamics of molecules and semiconductors, nano-optics, and plasmonics. A current group focus is Sustainability – specifically renewable energy generation and storage. We are exploring new materials and unique designs to enhance light capture and conversion efficiency in solar applications as well as boost energy and power density in electrical energy storage devices. These research topics provide numerous opportunities for interdepartmental collaboration, theoretical method development, and the advancement of both fundamental and applied science.

3.7.2 Collaborators

We collaborate with several other groups on campus at Baylor, other domestic partners, as well as international collaborations with groups in South Korea at Sungkyunkwan University (SKKU). These collaborations are not data intensive.

3.7.3 Instruments and Facilities

The primary research performed is modeling, meaning the main facilities are linux workstations, mac desktops and HPC clusters that run custom modeling applications as well as simulations using other forms of software. These components are typically replaced on 3-year times scales, or as funding allows.

Most data is kept on campus, with some rare instances of data sharing with the aforementioned collaboration groups. Simulation inputs are relatively small, but outputs (and checkpoints) are GBs to TB in size.

3.7.4 Process of Science

Simulations are paralyzable - thus the computational needs are fast (3Ghz+) processing utilizing many cores/processors and large (100s of GB) of main memory. Interconnection speeds of 10G or greater (ethernet or infiniband) between cluster members are required.

Efficient file transfer and the ability to share data offsite are things to be considering in the near term., but are not required currently.

3.7.5 Remote Science Activities

It is desirable to run simulations ‘remotely’, e.g. logging in to clusters from locations external to the University. Baylor ITS has enabled ways to do this securely.

3.7.6 Software Infrastructure

We use both commercial and open source software, as well as in-house development of tools and scripts. The primary tools for our group are:

- Molecular Quantum Chemistry Codes:
    - Gaussian[11]
    - Gamess[12]
    - Qchem[13]
- Plane Wave DFT:
    - VASP[14]
    - Quantum Expresso[15]
- Molecular Dynamics Simulators:
    - LAMMPS[16]
    - NAMD[17]
    - Gromacs[18]

3.7.7 Network and Data Architecture

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

3.7.8 Cloud Services

We have access to numerous enterprise cloud storage platforms. The most relevant for our group is Box, which is used to share files that all members need access to.

3.7.9 Known Resource Constraints

As the University grows, the current HPC cluster will need to be expanded dramatically to keep up with anticipated usage. All aspects of computation, storage, and networking will need to be scaled up.

3.7.10 Outstanding Issues

Our main problems in the near term will be related to resource limitations as the number of users grow on campus, this research is a significant user and thus has

[11] https://gaussian.com
[12] https://www.msg.chem.iastate.edu/gamess/
[13] https://www.q-chem.com
[14] https://www.vasp.at
[15] https://www.quantum-espresso.org
[16] https://lammps.sandia.gov
[17] https://www.ks.uiuc.edu/Research/namd/
[18] http://www.gromacs.org

scaled our work to utilize all that can be made available. HPC sustainability will need to be addressed in a fair and equitable manner.

3.8 Computational Fluid Dynamics Case Study

Content in this section authored by Scott James from the Department of Geosciences and School of Engineering and Computer Science

3.8.1 Science Background

The focus of this research is reactive flow and transport models in environmental systems (sub-fields of computational fluid dynamics). Projects range from simulating subsurface radionuclide fate and transport to surface water flow including sediment dynamics and water-quality components (e.g., micro- and macroalgae growth kinetics).

Specific projects include:

1. Simulating radionuclide and chlorinated solvent transport at the Santa Susana Field Laboratory for the Department of Energy (funded through CDM Smith, Inc.). The model is built using the commercial software, FEFLOW, although custom interface modules have been written to facilitate model calibration and uncertainty quantification.
2. Writing code amendments to the Delft3D surface water flow, sediment dynamics, and water quality code to simulate the effects of marine hydrokinetic and current-energy-capture devices (turbines). This work is sponsored by Sandia National Laboratories.
3. Simulating Sargassum (kelp) growth in the Gulf of Mexico for the Advanced Research Projects Agency – Energy (ARPA-E) to investigate the potential of macroalgal biofuels. I also write custom growth-kinetics subroutines for the HYCOM flow and Lagrangian particle tracking software.
4. Simulation of flow through aquaculture systems using the Environmental Fluid Dynamics Code (EFDC) software. This work is in collaboration with IBM Research in Dublin, Ireland.
5. Multiphase, multicomponent, flow and thermal modeling to simulate thermally and chemically enhanced oil recovery. This work was done with support from Canadian oil company RII International.

Environmental flow and transport modeling is typically used for decision making (e.g., site remediation or monitored natural attenuation at the Santa Susana Field Laboratory) or for communicating a common language between developers, researchers, and regulators (e.g., simulating the environmental effects of arrays of marine hydrokinetic turbines).

Other areas of research are the application of machine learning to the field of geoscience. This involves collaboration with IBM Research, Saudi Aramco, Stantec Engineering, and Los Alamos National Laboratory. Projects include:

1. Using long short-term memory (LSTM) networks for geologic facies identification using borehole wireline logs.
2. Forecasting ocean wave conditions using a multi-layer perceptron (MLP) model in conjunction with various machine-learning regression techniques.

3. Forecasting Chlorophyll-a concentrations as proxies for oceanic algal blooms using an autoregressive MLP model. Features include satellite multispectral Chl-a data, sea-surface temperatures, sunlight intensity, and day of year.
4. Development of a hybrid MLP/LSTM to forecast soil moisture across the United States. Real-time estimates of soil moisture are important for flood-risk assessment and crop viability.
5. Use of Non-negative Tensor Factorization with k-means clustering to classify and emulate computationally expensive reactive-transport simulations (e.g., subsurface contamination). This work is in collaboration with Los Alamos National Laboratory.
6. Identification of surface water features from high-resolution orthoimages. Applications include identifying water bodies including lakes, rivers, and ephemeral streams for environmental impact assessments required when new oil and gas pipelines are proposed. This work was sponsored by Stantec Engineering.
7. Simulation of harmful algal blooms in lakes using the EFDC software. This effort is part of a large research project funded by the National Institute of Health (with center headquarters at the University of South Carolina).

Machine learning modeling is advancing the state of the art in geoscience applications (reactive-transport emulators) with the potential to provide real-time forecasts for soil moisture, ocean-wave conditions, and Chl-a.

3.8.2 Collaborators

Collaborators include:

1. Los Alamos National Laboratory (Los Alamos, NM). We use Baylor's Box or Los Alamos' equivalent to share data.
2. IBM Research (Dublin, Ireland). We use Baylor's Box or IBM's Dropbox to share data.
3. Stantec Engineering, Inc. (Walnut Creek, CA). We use their private file-sharing website and Baylor's Box to exchange data.
4. Department of Energy (by way of CDM Smith, Inc. in Denver, CO). We use their private file-sharing website and Baylor's Box to exchange data.
5. RII Inc. (Alberta Canada). We share data through Baylor's Box.
6. Sandia National Laboratories (Albuquerque, NM). We share data with Baylor's Box or Sandia's file sharing site.

3.8.3 Instruments and Facilities

This research is computational. In addition to the Baylor Kodiak Cluster, and the dedicated GPU node self-purchased and maintained, the group uses Windows PC resources for development and simulation:

- two 8-core machines
- a 12-core machine
- a 28-core machine
- a 36-core machines (with an NVIDIA Quadro GV100 GPU)

- and two 44-core machines

These computational resources along with Kodiak are sufficient for my group for the next 5 years.

3.8.4 Remote Science Activities

Data sets from governmental entities (e.g., NOAA, NASA, and USGS) and private agencies (e.g., The Weather Channel [TWC]) can be used as input to simulations. At current time, only local to Baylor computation is used, or being explored.

3.8.5 Software Infrastructure

Software use varies depending on project. Some examples include:

- Commercial software
  - FEFLOW[19]
  - CMG-STARS[20]
  - COMSOL[21]
- Open-source codes
  - EFDC[22]
  - Delft3D[23]
  - MODFLOW[24]
  - PFLOTRAN[25]
- Python libraries for machine learning
  - PyTorch[26]
  - Keras[27]
  - TensorFlow[28]
  - Scikit-Learn[29]
  - Talos[30]

At the current time, data management/transfer is not utilized or required.

3.8.6 Network and Data Architecture

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

[19] http://www.feflow.info/fileadmin/FEFLOW/template_new/template.html
[20] https://www.cmgl.ca/stars
[21] https://www.comsol.com
[22] https://www.epa.gov/ceam/environmental-fluid-dynamics-code-efdc
[23] https://oss.deltares.nl/web/delft3d
[24] https://www.usgs.gov/software/software-modflow
[25] https://www.pflotran.org
[26] https://pytorch.org
[27] https://keras.io
[28] https://www.tensorflow.org
[29] https://scikit-learn.org/stable/
[30] https://pypi.org/project/talos/#description

3.8.7 Cloud Services

Beyond enterprise use cases that involve file sharing (e.g. Box), there is no usage of cloud resources. Computation in the cloud is an option for the future, but not actively being pursued.

4 - Discussion Summary

On January 6-7th 2020, members of the EPOC team and staff from LEARN met with representatives from Baylor University. This review was held in Waco, TX.

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

- 4.1.1 Enterprise IT
- 4.1.2 Campus Networking / Research Computing
- 4.2 Experimental High Energy Physics (HEP)
- 4.3 Proton Computed Tomography (pCT)
- 4.4 Nutrition and Relation to Digestive Microbiome
- 4.5 Baylor University Core Research Facilities
- 4.6 Molecular Quantum-dot Cellular Automata (QCA), and Material Science of Quantum Computing
- 4.7 Modeling and Simulation of Low-Dimensional and Nano-Structured Materials

The following Case Studies were not discussed in person, but the text submitted by the researchers is also included in the report for reference:

- 4.8 Computational Fluid Dynamics

4.1.1 Enterprise IT

Baylor Enterprise IT maintains the majority of the campus infrastructure, with the exception of research computing pieces (e.g. the Kodiak and related cluster resources). The server and software components are managed almost entirely out of the main data center (2900 sq ft, 4 x AC units to support 133 tons of cooling, 400kva UPS). Current growth areas for the data center group include working to increase backup capabilities (currently provided by 1 offsite non-geo redundant), and work to increase service redundancy (currently a mixture of n, and n+1). The majority of network-connected devices function at 1Gbps.

Software is varied around the university. Operating System and Office products (e.g. MS Office, MS Server, Red Hat Server) are numerous. Other products (e.g. those needed to support administration or research groups) are not as widely deployed. A process does exist to review new software before deployment.

Enterprise storage is available to support local backups, but is not sufficient to handle large volumes on a per-user basis.

Discussion Summary:

- The security review process was the primary point of discussion. This process is not defined very clearly to set expectations and timelines to research groups. If researchers are responsive, it can go fast (days). If they aren't, it may take longer (weeks, months). Many of the policies are not clearly articulated, and are passed on via verbal interactions. A goal is to streamline this to clearly state what is needed, and how long things may take in the future.
- VM hosting has scaled well for institutional needs.
- Backup remains an area of growth and concern. Working out future storage needs based on this.
- Research data vs. enterprise data was an area of discussion. Primarily if it makes sense to combine these into a single system. There are benefits to doing so – but this can also restrict potential use cases (e.g. sensitive data and non-sensitive on the same framework, latency requirements, access requirements, etc).

Presenting Staff:

- Michael Hand
- Chad Talbert

4.1.2 Campus Networking / Research Computing

Campus networking provides service (wired and wireless) to 137 buildings. Overall there are 3 categories:

1. Main Baylor Campus in Waco (115 buildings) via campus owned/operated fiber optic infrastructure.
2. Baylor Facilities off Campus (8 buildings) that use ISP/transport to extend the Baylor network.
3. Baylor and leased facilities off-campus (14 buildings) that are not on the main network, but connect back to Baylor over the public internet.

There are 4 major networking sites on campus. Each tie back fiber distribution, as well as offers various layers of network service and redundancy. The networks of campus are divided by use case:

1. ***Datacenter Network***: This is further subdivided into subnets with different use cases for security and performance reasons. Most central IT resources and campus services utilize this.
2. ***Baylor University Research Network (BURN)***: Designed to facilitate a faster path to the external world. Features a stricter security posture (e.g. only certain machines/services exposed). This is currently rate limited to 3Gbps, and is populated with DTNs/perfSONAR, and a small number of other services.
3. ***Faculty/Staff Network***: Wired connections to offices and classrooms.
4. ***Command and Control Network***: Building control, health and safety, etc.
5. ***RESNET***: Wired dormitory connections.
6. ***Wireless***: Campus wireless via 2800 wireless access points.

The WAN is provided via a cluster of Cisco ASR9K routers and connects to LEARN and commodity connections. LEARN provides 2x10Gbps currently. All networks (with the exception of BURN) must traverse campus firewall infrastructure.

Discussion Summary:

- The internal wired infrastructure is aging, and some time was devoted to discussing if it makes sense to deliver > 10Gbps connectivity in certain locations. As a distribution technology, 10Gbps is sufficient (and can scale to multiples via ECMP/LAG). Moving to 40G or greater may not be required unless a compelling use case is found.
- Visibility into campus traffic (via tools like Netflow/sFlow) would be desirable to understand usage patterns. The network core can handle this, but a software package (Inmon, Arbor, etc) would be required.
- The core chassis are capable of supporting 100G cards, but it was noted that certain cards function better than others (e.g. verify the backplane is native 100G, and not 12x10G).
- Moving to 100G is not trivial, and will require:
    - LEARN to upgrade Layer 2 and Layer 3 capabilities

    - Campus to get 100G capable hardware
    - It is recommended that augmentation occur in 10G increments (e.g. a < 1 month operation for both parties) as needed.
- Some discussion involved the use of SDN to better segment/manage network use cases. This will require more investigation, as the current set of profiles and divisions has scaled to the campus requirements.
- BURN will require a more clearly defined AUP and set of policy documents to govern user behavior.
- Using monitoring tools (to examine flow) could help offload heavy use cases from other networks.
- A pilot effort to use perfSONAR to validate performance behind the BURN 'Cisco Nexus Sandwich' was considered. Would utilize test points from Baylor and LEARN to external locations.
- LEARN's member portal can be used to understand wide area traffic patterns.

Presenting Staff:

- Scott Day

4.2 Experimental High Energy Physics (HEP)

The Baylor High Energy Physics (HEP) group performs research on elementary particle physics by utilizing data (e.g. proton-proton collision results) obtained from the CMS detector at CERN's Large Hadron Collider (LHC) instrument.

Science Summary:

The Baylor group's specific work involves searches of new physics principles beyond the Standard Model, as well as the precision of the measurements involving the Higgs Boson or other top Quarks.

Data from the experiment adheres to a regimented workflow for data distribution. After initial processing at CERN, data sets are distributed to a number of designated facilities that are geographically distributed throughout the world (WLCG: Worldwide LHC Computing Grid). 'Tier 3' facilities, of which Baylor is a part, contribute resources to the overall process of analysis and simulation but are not directly funded to provide resources. This ecosystem facilitates the major parts of the LHC workflow: distributed analysis, storage, and creation of simulation data using a common software framework.

Baylor participates in the creation of simulation data, the analysis of experimental data, and other R&D efforts including use of experimental data in the development of advanced techniques that utilize Machine Learning (ML) to improve the data collection and analysis process. As a part of this process, 200-300TB of data may be resident on Baylor resources at any given time, delivered via software packages that include Open Science Grid (OSG), and the "xrootd" package. The data requirements are expected to grow by as much as 5x in the coming years due to changes in the underlying technology, software, and process of science.

Discussion Summary:

- Details on the nature of the physics and research were provided.
    - The LHC is a large particle accelerator that records the results of a proton-proton collision via a 'detectors' that convert signals into results that can be analyzed via computational resources.
    - The results of a single experimental run are stored locally at CERN, processed by local computational resources, and prepared for distribution to more than 200 other sites worldwide
    - Using software designed to curate, find, download, and process the results, researchers are able to perform a number of analysis tasks on the data
    - In addition to analysis of experimental data, a large number of simulations are produced using the same computational framework to help develop software.
    - Baylor uses experimental data to develop novel Machine Learning (ML) codes that will influence future software used by the LHC

    - Even though the LHC is currently in shutdown, there is still active work being done to reprocess old data, and simulate new datasets.
- Primary collaborators in the US are the Tier 1 site (Fermilab), and the Tier 2 sites at a number of large/well connected universities. These collaborations involve a push/pull of data.
- Kodiak is the primary computational and storage framework
- Open Science Grid software is used for data transfer and analysis
- As a Tier 3 site, Baylor receives no funding for equipment upgrades.
- It is expected that the data volumes (currently around 200TB) will increase by 5X over the coming years. This is due to file size increases, and required volume of files that are needed to remain on site
- The act of data 'download' has flexibility. The software supports two major modes of operation that include bulk download along with the option to stream data remotely during processing. Baylor utilizes both, but often prefers the former since streaming can sometimes be slower due to occasional performance abnormalities
- The current HPC resources on Kodiak (CPU-based) scale to the LHC needs. Upgrades will eventually be required to support more computational time, as well as storage. GPU codes are being explored, but not widely deployed. LHC software developers will make choices on CPU/GPU adoption in the coming years.

Presenting Researchers:

- Kenichi Hatakeyama: Kenichi_Hatakeyama@baylor.edu
- Jay Dittmann: Jay_Dittmann@baylor.edu
- Andrew Brinkerhoff: Andrew_Brinkerhoff@baylor.edu

4.3 Proton Computed Tomography (pCT)

The pCT collaboration is an effort to develop a new medical imaging modality, proton (or ion) computed tomography (pCT). These instruments can be used to direct radiation to a specific area (size, depth) in an effort to offer precise treatment options.

Science Summary:

The pCT collaboration is working to develop new medical imaging modality based on proton (or ion) computed tomography (pCT). Ions can be used to image the body with only a few percent of the radiation damage of a normal X-Ray image, but with greater accuracy for treatment with ions, since the relevant quantities (relative stopping and scattering powers) are directly measured. The major reason for pCT is this latter case, the planning and verification of proton or ion radiation treatments.

This work involves a number of collaborators that have deployed these instruments and share the measured results during use. Telemetry gathered during the process involves collecting electronic output of thousands to billions of ions that may pass through the sample after being emitted by the instrument. These data sets typically involve tracking, timing, and energy measurements for 360 million to 2 billion protons/ions, usually of at least 9 data elements. Data sizes (after triggering and compression) are currently 10-100GB. Increased numbers of ions, or longer experimental runs, have the potential to increase the data set sizes. Due to the nature of the collected data, it is highly compressible which facilitates easier storage and transfer.

Currently data must be shared manually (exchange of hard drives) due to performance problems of hospital / clinical network infrastructures. The Baylor team is developing hardware/software solutions to facilitate easier data exchange from the collaborators.

Discussion Summary:

- Proton (or ion) computed tomography (pCT) is the use of a precision medical instrument similar to an X-Ray. The sample (tissue, etc) is hit with quantities of protons (or ions) but the accuracy (area, depth) can be finely controlled. This precision facilitates more directed treatment options, particularly combating forms of cancer that may be deep within the body (bones, organs, etc).
- Instrumentation is developed locally. Instrumentation is deployed locally, and at collaborators BU, LLU, UCSC, NIU, Stanford, UCSF, University of Haifa, UNSW, Ludwig Maximillian U, U of Manchester, SUNY Stonybrook, CUNY, and potentially others.

- Collaboration involves use of instruments, and data sharing relationships so that the Baylor team can capture measurements/research data on the functionality over time.
- Measurements of interest include the relative telemetry of each of the protons/ions that are sent toward a sample (sometimes in the billions): speed, direction, relative power, etc.
- Data sets when raw can be in the TB range, but the instruments, computational hardware, and software perform initial 'triggering' of the data to reduce size. Further reduction through compression is possible – resulting in data sets in the 10-100GB range. This range can be attributed to the number of protons/ions sent (e.g. more for larger/deeper samples) and length of time sampled.
- Getting data back to Baylor can be challenging, due to the nature of the remote networking infrastructure. E.g. a hospital network is designed for protection of sensitive info first and foremost: expedient transfer of large data sets doesn't fit this profile typically. Baylor University staff have designed the data collection infrastructure to facilitate data reduction when applicable, but sometimes removal and shipment of hard drives may be required (or worst case – physical visits to retrieve data). In some cases, hard drives from a clinical environment can be removed and shared 'locally', e.g. a clinic can be nearby a well-connected university. Once the hard drive arrives at a more capable location, remote data access is possible.
- Data retrieval may take months for particularly large data sets at poorly equipped facilities.
- There are no sensitive aspects to the data, it is de-identified immediately. It is a project goal to avoid use of sensitive data.
- Reconstruction of data (once retrieved) is done at Baylor using a variety of resources, including the Kodiak cluster.
- Simulation is also a part of the workflow (to test models/software), but is not widely practiced at clinical sites. Doctors testing the tools are more likely to want to use real data than simulated data.

Presenting Researchers:

- Keith Schubert

4.4 Nutrition and Relation to Digestive Microbiome

The core research involves identification of biomarkers and mechanisms within the gut microbiome, and how diet can be used to reduce incidence of and improve survival among individuals diagnosed with colon and lung cancer. The work is heavily dependent on the creation and curation of experimental samples, and the analysis of these samples.

Science Summary:

The main source of data is generated from sequencing of human/mammalian tissues and fecal samples. Known large population microbiome datasets are also downloaded and from similar studies from which it is possible to conduct meta-analyses for biomarkers as leads for potential hypotheses and mechanisms for further study.

The process of science typically involves the collection of samples from donors, physical shipment of samples to a processing facility, return of data sets to Baylor, processing of data sets and analysis against others to discover relationships and findings. To facilitate the process of science, significant time, resources, and effort has produced a "pipeline" for the analysis of large microbiome sequencing data sets (locally produced and pulled from other sources).

The group typically experiences a large burst of intense server usage every 3-6 months depending on the project status. Collaborators are given permission to use data for a period of time until the project is completed, but access directly to the data is controlled through Baylor staff.

Discussion Summary:

- The Baylor team is not currently able to sequence samples locally, so relies heavily on the instrumentation of collaborators. Samples are gathered/curated, and physically shipped to locations or sequencing.
- Sequencing produced a data set that is then exchanged back with Baylor staff. Data sets vary widely in size – several GB to TB in size depending on the sample complexity. Data can be transmitted using technology, or via more physical methods (e.g. mailing storage).
- It may take weeks to months to go from collected samples to computable data, and most of the work can be done in bursts.
- Local compute resources are leveraged to process the raw output of a sequenced data set. The software pipeline developed by the group assists in this process to process and analyze these data sets.
- Local data storage is heavily used to store data sets when not in use. Total data sets reside in around 20TB of space.
- Growth in this group will depend on external funding sources. It is possible they may invest in an on-site sequencer (to be maintained within the core Baylor Facility) which could accelerate the speed and quantity of sampling

efforts. With more samples will come a requirement for more compute and storage.

- Previously it has been challenging to manage the electronic transmission of certain data sets due to performance problems. One is being researched that involves an international data transfer.

Presenting Researchers:

- Leigh Greathouse

4.5 Baylor University Core Research Facilities

Baylor features 5 core research facilities in the college of Arts & Sciences: the Baylor Mass Spectrometry Facility, the Animal Vivarium, Biomolecular Sciences Center, Center for Microscopy and Imaging, and the Nuclear Magnetic Resonance center. All of these are located in the 500,000 sq. ft. Baylor Sciences Building.

Science Summary:

Data from these facilities supports federal, industrial, and startup research primarily in A&S supporting the departments of Biology, Chemistry & Biochemistry, Environmental Sciences, Geosciences, Physics, and Psychology and Neuroscience (resources are equally available to all Baylor researchers and participation from several other colleges and departments across campus has been growing over the last several years). The 5 core research facilities serve~60 faculty and 200-300 end-users

Despite a common facility, each core has a different mechanism of handling data. For some instruments there is a possibility to utilize automated sharing of data across a network. Others must rely on less efficient means (e.g. mailing files). Lastly, some utilize USB hard drives to transmit data. Most cores do maintain policy that data cannot reside permanently on a resource, but this is not enforced unless space becomes an issue.

The facility as a whole will resist becoming a curation home due to a lack of main storage capabilities. Ideally the "chain of custody" for data stays with the end user at all times and they transfer data out when done with an analysis, but typically that data is transferred as a copy with the original left behind at the instrument PC.

Data can vary depending on instrument – some can produce GB data sets and are used multiple times per day. Future devices could generate TB data sets of a several day run.

Discussion Summary:

- The core facilities are a menagerie of different scientific capabilities. Each instrument features a different operational style (e.g. local control, remote control), data output (KB to near TB sizes), and way to communicate to the outside world (data movement via software, email, or sharing via removable media).
- Due to the ages, usage patterns, and user community, developing best practices and procedures facility wide is challenging. Newer resources that feature some more technological support are easier than other older ones.
- New technology (e.g. Cryo-em) will force some change in that local storage, networking, and potentially more computation may need to be present. It may also be possible to integrate an experimental pipeline to the data center to utilize research storage/compute on Kodiak

- The use of remote-control software (e.g. Team Viewer) is becoming a bit more common – although this is still used in a semi ‘local’ fashion. E.g. users on campus versus users across the country.
- File sizes for some machines are still small enough to handle with email or cloud storage. Others must be sent with data transfer tools, or removable media.
- Security of the infrastructure is challenging. Sometimes software is several revisions out of date to support features, which causes risk versus productivity.
- Software use requires an approval process, which can be onerous at times. In particular the approval for some software packages can take months. For grants with a limited time window, this can cause significant delays that impact productivity.

Presenting Researchers:

- Christopher Becker - Director of Baylor Sciences Building and Baylor Mass Spectrometry Center

4.6 Molecular Quantum-dot Cellular Automata (QCA), and Material Science of Quantum Computing

The work centers on two core areas: molecular quantum-dot cellular automata (QCA), and the material science of quantum computing. This is the intersection between computational and materials science.

Science Summary:

Molecular QCA (mQCA) is a general-purpose, classical computing paradigm for the post-Moore's law era. It is designed to provide energy-efficient, high-speed general-purpose computing. The research group develops models and theories related to quantum phenomena pertaining to mQCA devices and circuits. This work is generally done in MATLAB for modeling dynamic quantum processes in circuits and devices, along with emerging capabilities to use ab initio modeling of molecules at the atomic scale to explore the design of candidate QCA molecules. This involves high-performance computing in programs such as Q-Chem, NWChem, and Gaussian.

The material science of quantum computing involves performing ab initio modeling of point defects in semiconductors. The research group is in the process of acquiring VASP (Vienna Ab Initio Simulation Program) to run on the Kodiak and Polar Bear clusters, and also uses Quantum ESPRESSO. The goal is to determine conditions under which we can create stable, point defects in ZnSe for quantum information processing applications.

Discussion Summary:

- Research is heavily based on simulation: e.g. running of software packages on HPC resources that simulate experimental situations/conditions.
- Can run for minutes, hours, weeks (depending on inputs, complexity, and resources allocated).
- Data resulting from simulation can be large (GBs), particularly if the output is visual (video). Simulations may produce 100GB of data a week during extremely busy periods.
- Checkpoints on running code are possible to save the state of execution. For long-running code on shared resources this is often a routine behavior. Checkpoints can be large in size (10-100s MBs) but are deleted after execution completes. These are stored locally (on machine of execution).
- Data is not currently shared out of Baylor University, but a possibility of collaboration via existing mechanisms (e.g. https://materialsproject.org) would create data sharing conduits.
- GitHub is utilized for the storing of source code used in this research. Team members (e.g. graduate students, professors, etc.) regularly use this for backups and tracking project status.
- HPC resources consist of those local to the lab (workstations, a purpose built HPC cluster named 'Polar Bear' that is liquid cooled GPUs), local to the university (Kodiak), and regional (TACC). The later has not been used to date,

but could be for a simulation that scales well and is beyond the capabilities of the campus (e.g. > 215 cores).

- Some portions of the simulation workload is highly parallelizable, and could scale to large numbers of cores/GPUs. Other parts (e.g. those developed via MATLAB) do not due to higher memory usage and time interdependence.
- Research data has no PII affiliated (input or output)
- Cloud computing is not being explored, as the computation local, on campus, and regionally (e.g. TACC) fit the needs of this research. Scalability to more resources is also bound on staff resources, which are not expected to grow significantly.
- As funding for the group becomes available (e.g. grants), the condo computing model is appealing. Using funds to augment the condo at Baylor campus could facilitate a greater availability of resources for this and other groups.

Presenting Researchers:

- Erik Blair

4.7 Modeling and Simulation of Low-Dimensional and Nano-Structured Materials

The Shuford Group is a theoretical/computational research team that investigates interdisciplinary topics spanning chemistry, physics, materials science, and engineering.

Science Summary:

The primary research interest for the group involves modeling and simulation of fundamental processes in low-dimensional and nano-structured materials. Established research areas include ultrafast quantum dynamics of molecules and semiconductors, nano-optics, and plasmonics.

A current group focus is sustainability – specifically renewable energy generation and storage. We are exploring new materials and unique designs to enhance light capture and conversion efficiency in solar applications as well as boost energy and power density in electrical energy storage devices. These research topics provide numerous opportunities for interdepartmental collaboration, theoretical method development, and the advancement of both fundamental and applied science.

We use a combination of desktop workstations and HPC resources provided by Baylor. The group generates a sizable amount of data that is shared between members and occasionally collaborators.

Discussion Summary:

- The group's research sits at the intersection between science and engineering, and touches on topics that range from chemistry, to physics, to materials science, to computer and electrical engineering.
- The work is heavily based on simulation, thus the use of computational resources (e.g. local workstations, Baylor HPC) is important to the overall process of science.
- Collaboration is limited to a small set of external partners. Sungkyunkwan University (SKKU) in South Korea is a site of regular collaboration. Most other collaborations are internal to the University.
- Computing resources used local to the lab are a collection of Linux and Macintosh workstations. These have sufficient cores, memory, and storage to run the software that is needed for simulation design and research. The Baylor HPC resources (Kodiak) run longer simulations and a collection of software packages including VASP and ESPRESSO
- Future computing needs will require fast processors, interconnection and increased memory (e.g. the workloads are highly parallelizable on many of the software packages). This group will require upgrades in the 2-5 time cycle to the university HPC systems to scale with the number and complexity of simulations that will be required via this research group.

- Remote access (e.g. ability to launch jobs external to the University) is highly desirable for research group members.
- Software utilizes many off the shelf (commercial and open source) products, as well as custom scripts developed and curated by the research group.
- Cloud use is limited to file sharing platforms (e.g. BOX), there is no need to use cloud computing at the current time.
- It was suggested that heavy HPC users be 'taxed' via the grants to facilitate upgrades. This mechanism can support future upgrades and ensure a scalable solution to the resources.

Presenting Researchers:

- Kevin Shuford

4.8 Computational Fluid Dynamics

With a background in computational fluid dynamics, I solve reactive flow and transport models in environmental systems. Projects range from simulating subsurface radionuclide fate and transport to surface water flow including sediment dynamics and water-quality components (e.g., micro- and macroalgae growth kinetics).

My other primary area of research is the application of machine learning to geosciences. I have collaborations with IBM Research, Saudi Aramco, Stantec Engineering, and Los Alamos National Laboratory.

Science Summary:

Specific CFD projects include:

- Simulating radionuclide and chlorinated solvent transport at the Santa Susana Field Laboratory for the Department of Energy (funded through CDM Smith, Inc.). The model is built using the commercial software, FEFLOW, although custom interface modules have been written to facilitate model calibration and uncertainty quantification.
- Writing code amendments to the Delft3D surface water flow, sediment dynamics, and water quality code to simulate the effects of marine hydrokinetic and current-energy-capture devices (turbines). This work is sponsored by Sandia National Laboratories.
- Simulating Sargassum (kelp) growth in the Gulf of Mexico for the Advanced Research Projects Agency – Energy (ARPA-E) to investigate the potential of macroalgal biofuels. I also write custom growth-kinetics subroutines for the HYCOM flow and Lagrangian particle tracking software.
- Simulation of flow through aquaculture systems using the Environmental Fluid Dynamics Code (EFDC) software. This work is in collaboration with IBM Research in Dublin, Ireland.
- Multiphase, multicomponent, flow and thermal modeling to simulate thermally and chemically enhanced oil recovery. This work was done with support from Canadian oil company RII International.

Specific ML Projects include:

- Using long short-term memory (LSTM) networks for geologic facies identification using borehole wireline logs.
- Forecasting ocean wave conditions using a multi-layer perceptron (MLP) model in conjunction with various machine-learning regression techniques.
- Forecasting Chlorophyll-a concentrations as proxies for oceanic algal blooms using an autoregressive MLP model. Features include satellite multispectral Chl-a data, sea-surface temperatures, sunlight intensity, and day of year.
- Development of a hybrid MLP/LSTM to forecast soil moisture across the United States. Real-time estimates of soil moisture are important for flood-risk assessment and crop viability.

- Use of Non-negative Tensor Factorization with k-means clustering to classify and emulate computationally expensive reactive-transport simulations (e.g., subsurface contamination). This work is in collaboration with Los Alamos National Laboratory.
- Identification of surface water features from high-resolution orthoimages. Applications include identifying water bodies including lakes, rivers, and ephemeral streams for environmental impact assessments required when new oil and gas pipelines are proposed. This work was sponsored by Stantec Engineering.
- Simulation of harmful algal blooms in lakes using the EFDC software. This effort is part of a large research project funded by the National Institute of Health (with center headquarters at the University of South Carolina).

Presenting Researchers:

- Scott James

5 - Recommendations for Review

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

- Baylor ITS will evaluate the current process of Software Verification/Security Review and better set expectations with the research community. This will involve better stating timelines, goals, and improving communication.
- Baylor ITS will explore upgrades and augmentations to computation on Kodiak. This may involve horizontal scaling (more nodes) or vertical enhancements (interconnect upgrades, CPU/memory augmentation).
- Baylor ITS will explore upgrades and augmentations to networking on Kodiak. This may involve a new switching/routing architecture, and the addition of data transfer hardware/software.
- Baylor ITS may consider implementing more 'condo' models of computing, as the number of facility/research groups that require on-site (but not necessarily full-time use) of computing increases.
- Baylor ITS will expand a program to make GPU resources available.
- Baylor ITS will continue to explore storage for the campus. This must take the form of enterprise (e.g. students/faculty general storage) as well as for research uses (e.g. affiliation with computation on Kodiak, or the core research facilities).
- Baylor ITS and LEARN will explore relationships in Texas to facilitate regional or statewide 'fate sharing' arrangements on backups.
- Baylor ITS and LEARN will explore network upgrades. This could be addition of 10Gbps connectivity, or upgrades to support 100Gbps.
- Baylor ITS will work to create "onboarding" documentation for use of BURN. This could be technical, but should also contain policy for use, monitoring, and expectations for users.
- Baylor ITS will evaluate if upgrading campus links to beyond 10G is cost effective or necessary, except to critical and/or data intensive use cases.
- Baylor ITS will explore other Globus service options, and consider bringing up more endpoints at key locations (e.g. BRIC, Baylor Science Building, etc). This action should dovetail with efforts to increase and improve research storage options.
- Baylor ITS will consider if a "campus wide" deployment of Git (version control software) makes sense given the wide set of use from the research community.
- Baylor ITS will consider expanding network monitoring to include deeper use of sFlow/netflow as well as perfSONAR.
- Baylor ITS will work with EPOC on ways to better attract researchers, and showcase the 'services' that are available on campus.
- Baylor ITS will work with regional experts (e.g. LEARN, TACC) on ways to implement security compliance mechanisms such as NIST 800.171.

- Baylor ITS will work with the department of Physics on the reported performance abnormality during downloads of LHC data.
- Baylor ITS will work with Keith Schubert/pCT on ways to simplify the workflow of retrieving data from remote sites. This may take the form of purpose-built DTNs that are shipped with scientific instruments.
- Baylor ITS and EPOC will partner to deploy a "Modern Research Data Portal" to assist researchers with external data sharing needs (e.g. Dr. Greathouse).
- Baylor ITS will work with Core Facilities (and other interested researchers) to understand the impact of newly deployment instruments such as sequencers. Ways to mitigate security risks, data volumes, and remote usage are critical to consider.
- Baylor ITS and the Core Facilities will start a conversation about an improved data access layer. This should involve the creation of a mechanism to store research data from instruments to a central or distributed set of storage resources that can be easily reached by the research community, and the computational infrastructure. Use of DTNs for external sharing is also desirable.

Appendix A - LEARN Regional Networking Diagram

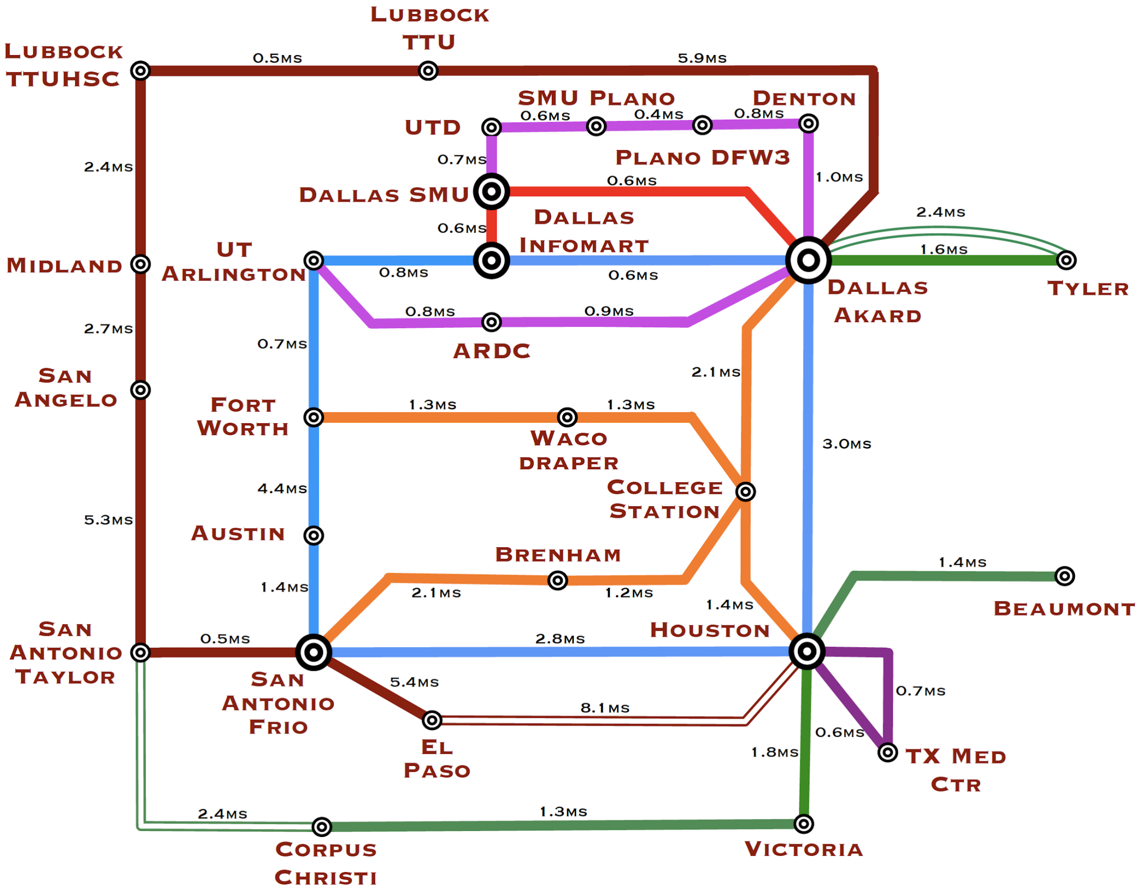

Figure 7 - LEARN Latency Map

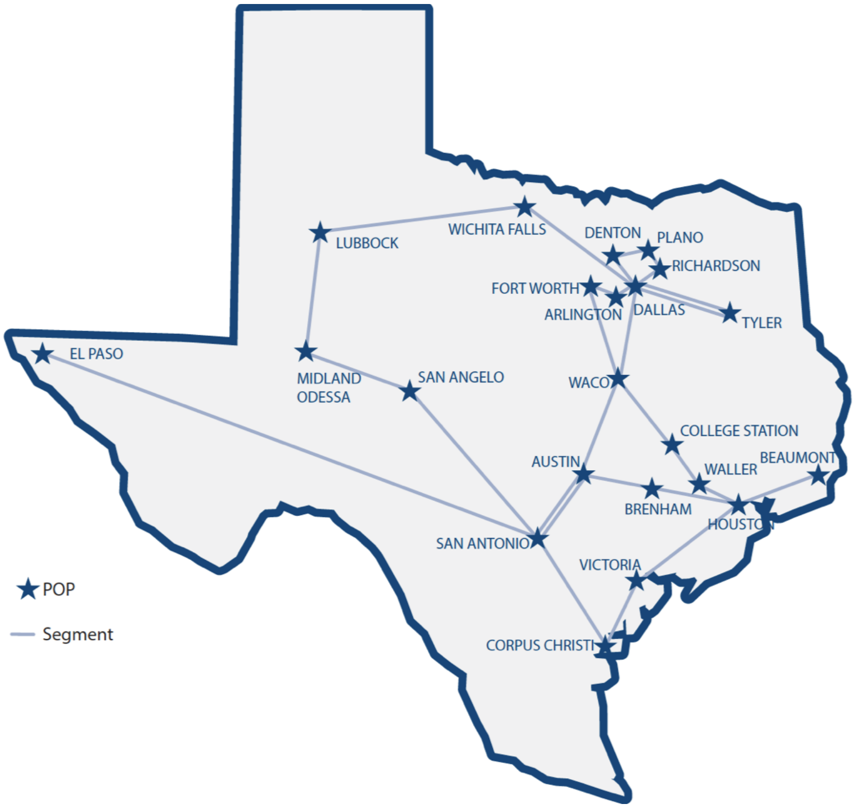

Figure 8 - Schematic of the LEARN Network.

Utilizing LEARN Membership for Research Connectivity & HPC Resources

LEARN provides to its members, a carrier class MPLS Layer 2/3 network built over the advanced 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.