ESnet
ENERGY SCIENCES NETWORK

Nuclear Physics Network Requirements Review Report

May 8–9, 2019

BERKELEY LAB

U.S. DEPARTMENT OF ENERGY
Office of Science

ESnet
ENERGY SCIENCES NETWORK

Nuclear Physics
Network Requirements Review Report

May 8–9, 2019

Office of Nuclear Physics, DOE Office of Science
Energy Sciences Network (ESnet)
Gaithersburg, Maryland

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 work was supported by the Directors of the Office of Science, Office of Advanced Scientific Computing Research, Facilities Division, and the Office of Nuclear Physics.

This is a University of California, Publication Management System report number LBNL-2001281.

Cover Images: (**top left**) ALICE image, courtesy of CERN; (**top right**) JLab Hall image, courtesy of JLab; (**bottom left**) GRETA image, courtesy of Mario Cromaz, GRETA Project Team; (**bottom right**) PHENIX image, courtesy of Brookhaven National Laboratory.

Participants and Contributors

Whitney Armstrong, *Argonne National Laboratory*
John Arrington, *Argonne National Laboratory*
Ed Balas, *Lawrence Berkeley National Laboratory, ESnet*
Debbie Bard, *Lawrence Berkeley National Laboratory, NERSC*
Steve Beher, *Michigan State University / Facility for Rare Isotope Beams*
Laura Biven, *Department of Energy Office of Science, ASCR*
Vincent Bonafede, *Brookhaven National Laboratory*
Ben Brown, *Department of Energy Office of Science, ASCR*
Rich Carlson, *Department of Energy Office of Science, ASCR*
Mike Carpenter, *Argonne National Laboratory*
Taylor Childers, *Argonne National Laboratory*
Jody Crisp, *ORISE*
Mario Cromaz, *Lawrence Berkeley National Laboratory*
Eli Dart, *Lawrence Berkeley National Laboratory, ESnet*
Brent Draney, *Lawrence Berkeley National Laboratory, NERSC*
Robert Edwards, *Thomas Jefferson National Accelerator Facility*
Timothy Hallman, *Department of Energy Office of Science, NP*
Barbara Helland, *Department of Energy Office of Science, ASCR*
Bryan Hess, *Thomas Jefferson National Accelerator Facility*
Graham Heyes, *Thomas Jefferson National Accelerator Facility*
Susan Hicks, *Oak Ridge National Laboratory*
Scott Kampel, *Princeton Plasma Physics Laboratory*
Eric Lancon, *Brookhaven National Laboratory*
Jerome Lauret, *Brookhaven National Laboratory*
Carolyn Lauzon, *Department of Energy Office of Science, ASCR*
Steven Lee, *Department of Energy Office of Science, ASCR*
Sean Liddick, *Michigan State University, Facility for Rare Isotope Beams*
Paul Mantica, *Michigan State University, Facility for Rare Isotope Beams*
Andrew Melo, *Vanderbilt University*
Jonathan Menard, *Princeton Plasma Physics Laboratory*
Inder Monga, *Lawrence Berkeley National Laboratory, ESnet*
Brent Morris, *Thomas Jefferson National Accelerator Facility*
Thomas Ndousse-Fetter, *Department of Energy Office of Science, ASCR*
Chris Pinkenburg, *Brookhaven National Laboratory*
Robinson Pino, *Department of Energy Office of Science, ASCR*
Jeff Porter, *Lawrence Berkeley National Laboratory, NERSC*
Gulshan Rai, *Department of Energy Office of Science, NP*
Thomas Rockwell, *Michigan State University, Facility for Rare Isotope Beams*
Lauren Rotman, *Lawrence Berkeley National Laboratory, ESnet*
Arjun Shankar, *Oak Ridge National Laboratory*
Adam Slagell, *Lawrence Berkeley National Laboratory, ESnet*
James Sowinski, *Department of Energy Office of Science, NP*
Tom Uram, *Argonne National Laboratory*
Chip Watson, *Thomas Jefferson National Accelerator Facility*
Paul Wefel, *Lawrence Berkeley National Laboratory, ESnet*
Linda Winkler, *Argonne National Laboratory*
Jason Zurawski, *Lawrence Berkeley National Laboratory, ESnet*

Report Editors

Ben Brown, *Department of Energy Office of Science:* Benjamin.Brown@science.doe.gov
Eli Dart, *ESnet:* dart@es.net
Gulshan Rai, *Department of Energy Office of Science:* Gulshan.Rai@science.doe.gov
Lauren Rotman, *ESnet:* lauren@es.net
Paul Wefel, *ESnet:* pwefel@es.net
Jason Zurawski, *ESnet:* zurawski@es.net

Table of Contents

1 Executive Summary

About ESnet

The Energy Sciences Network (ESnet) is the Office of Science's high-performance network user facility, delivering highly reliable data transport capabilities optimized for the requirements of data-intensive science. In essence, ESnet is the circulatory system that enables the U.S. Department of Energy (DOE) science mission by connecting each and every DOE lab and its user facilities. ESnet is funded and stewarded by the Advanced Scientific Computing Research (ASCR) Program, and managed and operated by the Scientific Networking Division at Lawrence Berkeley National Laboratory (LBNL). ESnet is widely regarded as a global leader in the research and education networking community.

ESnet connects DOE national laboratories, user facilities, and major experiments so scientists can use remote instruments and computing resources as well as share data with collaborators, transfer large data sets, and access distributed data repositories. While ESnet provides network connectivity, it cannot be characterized as an internet service provider as it is specifically built to provide a range of network services that are tailored to meet the unique requirements of DOE's data-intensive science.

In short, ESnet's mission is to enable and accelerate scientific discovery by delivering unparalleled network infrastructure, capabilities, and tools. ESnet's vision is summarized by these three points.

1. Scientific progress will be completely unconstrained by the physical location of instruments, people, computational resources, or data.
2. Collaborations at every scale, in every domain, will have the information and tools they need to achieve maximum benefit from scientific facilities, global networks, and emerging network capabilities.
3. ESnet will foster the partnerships and pioneer the technologies necessary to ensure that these transformations occur.

Requirements Review Purpose and Process

ESnet and ASCR utilize requirements reviews to discuss and analyze current and planned science use cases and anticipated data output of a particular program, user facility, or project to inform ESnet's strategic planning, including network operations, capacity upgrades, and other service investments. A review surveys major stakeholders' plans and processes in order to investigate data management requirements over the next 5–10 years. Questions crafted to explore this space include:

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

Requirements reviews help ensure that key stakeholders have a common understanding of the issues and the actions that ESnet may need to undertake to offer solutions. The ESnet Science Engagement Team meets with each individual program office within the Office of Science every three years. Through a collaboration with ESnet staff and leadership alongside the relevant program officers to identify the appropriate principal investigators (PIs) and their information technology (IT) partners to participate in the review, ESnet organizes, convenes, and executes the review.

This Review

In May 2019, ESnet and the Office of Nuclear Physics (NP) of the DOE Office of Science organized an ESnet requirements review of NP-supported activities. Preparation for this event included identification of key stakeholders to the process: program and facility management, research groups, technology providers, and a number of external observers. These individuals were asked to prepare formal case study documents about their relationship to the NP program to build a complete understanding of the current, near-term, and long-term status, expectations, and processes that will support the science going forward. A series of pre-planning meetings better prepared case study authors for this task and facilitated a smooth transition to the in-person review.

The mission of the NP program is to discover, explore, and understand all forms of nuclear matter. Nuclear science began by studying the structure and properties of atomic nuclei as assemblages of protons and neutrons. At first, research focused on nuclear reactions, the nature of radioactivity, and the synthesis of new isotopes and new elements heavier than uranium. Today, the reach of nuclear science extends from the quarks and gluons that form the substructure of protons and neutrons, once viewed as elementary particles, to the most dramatic of cosmic events: supernovae.

At its heart, NP attempts to understand the composition, structure, and properties of atomic nuclei; discover new forms of nuclear matter, including that of the early universe; measure the quark structure of the proton and neutron; and study the mysterious and important neutrino. Rapid advances in large-scale integration electronics, computing, and superconducting technologies have enabled the construction of powerful accelerator, detector, and computing facilities. These provide the experimental and theoretical means to investigate nuclear systems ranging from tiny nucleons to stars and supernovae. NP also supports the production, distribution, and development of production techniques for radioactive and stable isotopes that are in short supply and critical to the nation.

The DOE Office of NP provides most of the federal support for NP research in the United States. About 1,620 scientists, including 880 graduate students and postdoctoral research associates, receive support from NP. In addition, the program supports three national scientific user facilities. Other agencies use these NP facilities for their own research. Notable is the use by semiconductor manufacturers that develop and test radiation-hardened components for Earth satellites to be able to withstand cosmic-ray bombardment and by the National Aeronautics and Space Administration's Space Radiation Laboratory (NSRL) established at BNL's RHIC facility to study radiobiological effects using beams that simulate the cosmic rays found in space.

The NP program helps the United States maintain a leading role in NP research, which has been central to the development of various technologies, including nuclear energy, nuclear medicine, space exploration, and the nuclear stockpile. The program produces highly trained scientists who help to ensure that DOE and the United States have a sustained pipeline of highly skilled and diverse science, technology, engineering, and mathematics (STEM) workers who are knowledgeable in nuclear science.

This review included case studies from the following NP stakeholder groups:

- Thomas Jefferson National Accelerator Facility (JLab): Facilities
- JLab: Measurement of a Lepton-Lepton Electroweak Reaction (MOLLER)
- JLab: The Solenoidal Large Intensity Device (SoLID)
- JLab: Theory Group & Lattice Quantum Chromodynamics (LQCD)
- JLab: Electron-Ion Collider (EIC)
- Facility for Rare Isotope Beams (FRIB)
- Gamma-Ray Energy Tracking Array (GRETA)

- Argonne National Laboratory (ANL): Gammasphere / Argonne Tandem Linear Accelerator System (ATLAS)
- ANL: CLAS12 / EIC
- Brookhaven National Laboratory (BNL): The RHIC and ATLAS Computing Facility (RACF)
- BNL: The Solenoidal Tracker At RHIC (STAR)
- BNL: Pioneering High-Energy Nuclear Interaction eXperiment (PHENIX) / sPHENIX
- Compact Muon Solenoid (CMS) Heavy Ion Experimentation
- ALICE (A Large Ion Collider Experiment) Project and ALICE-USA Computing

The review participants spanned the following roles:

- Subject matter experts from the NP activities listed previously.
- ESnet Site Coordinators Committee members from NP activity host institutions, including the following DOE labs: ANL, BNL, JLab, LBNL and its National Energy Research Scientific Computing Center (NERSC), and Oak Ridge National Laboratory (ORNL).
- Networking and/or science engagement leads from the ASCR High-Performance Computing (HPC) Facilities.
- DOE Office of Science staff spanning both ASCR and NP.
- Observers from other DOE Office of Science programs and facilities.
- ESnet staff supporting positions related to facility leadership, scientific engagement, networking, software development, and cybersecurity.

Key Findings

Several key findings emerged from the review:

1. A number of new NP facilities and detectors are coming online in the next three to seven years, which will change the way in which scientific results are produced, curated, and collaborated.
2. There are increased and newly emerging usage patterns of ASCR HPC facilities by multiple NP experiments and facilities. This is part of a larger structural change within DOE/SC and indeed within the scientific community at large.
3. Network path diversity and capacity are critical issues for major NP facilities as the experimental data volumes increase and the need for external computation in the workflow changes.
4. Today, many of the facilities utilize local cluster computing to analyze data. As facilities upgrade and produce larger sets of data, such a model will not be able to manage the computing needs. To this end, scientists are challenged to identify resources within DOE ASCR and computing resources of the Open Science Grid (OSG). To fully adopt these resources, workflows will require significant changes to software and network architecture.
5. For workflows to become more mobile, there will need to be a series of steps taken to reduce areas of friction. These include upgrading software to facilitate streaming, increasing and improving network connectivity options, and altering computational models to allow for both local and remote data access.

6. A common access methodology (perhaps an application programming interface [API]) that spans ASCR computational facilities will greatly assist in the development of workflows that can adopt and regularly use these resources.
7. Data portals, typically used to share and disseminate experimental results, are either not widely available or are aging for a number of collaborations. Adoption of high-performance techniques in this software and hardware space is needed by several collaborating groups.
8. Network-based workflows that span facilities rely heavily on performance monitoring software, such as perfSONAR, to set expectations and debug critical transmission problems. This software is now widely available at all profiled facilities.

Recommended Actions

Based on the key findings, the review identified several actions for NP, ASCR, and ESnet to pursue.

ESnet will:

- Start a discussion between ESnet engineering and experimental representatives that are interested in sharing ESnet 6 telemetry data.
- Consider the creation of Large Hadron Collider Open Network Environment- (LHCONE) like overlay networks for certain use cases.
- Continue discussions with ORNL regarding wide-area connectivity options and amounts.
- Continue discussions with JLab; the Mid-Atlantic Broadband Cooperative; COX communications/Old Dominion University Virginia regional network (ELITE); and the Mid-Atlantic Research Infrastructure Alliance (MARIA) regarding wide-area connectivity options and amounts.
- Publish findings from GRETA work as a guide for future experimental design.
- Assist groups looking to measure and understand wide-area performance expectations with tools such as perfSONAR.
- Facilitate discussions with groups looking to adopt the Modern Research Data Portal design pattern.
- Facilitate peering with commercial clouds, as needed, for experiments that are looking into pilot efforts.
- Support efforts to obtain cycles among the DOE supercomputing sites as needed, in particular ensuring the end-to-end path is optimized with the appropriate tools and services.
- Build a data dashboard to easily visualize areas of growth within NP, comparisons to other program offices, and other easy to understand charts or graphs to characterize unique aspects of NP and other program offices.

The NP community will:

- Continue discussions with ASCR facilities on the new and expanded usage of computational and storage resources for certain experimental workflows.
- Further refine the data format (size, quantity) to facilitate more efficient mechanisms for data sharing as long-local workflows are adopted as two major eras begin: EIC for NP experimentation and the Exascale Computing Project for ASCR computation.

- Help ESnet estimate the predicted growth in NP science data production and use over the next 2–10 years. Specific numbers are desired to ensure appropriate investments are made towards network services that will meet these needs in each phase of the data growth.

The ASCR community will:

- Collectively work towards building infrastructure that better supports streaming workflows, along with developing a more uniform interface to the DOE/ASCR HPC facilities.
- Work with NP experiments to further explore the development of advanced portals to share research.

2 Review Findings

Below are the findings for the NP and ESnet Requirements Review held May 8 and 9, 2019. These points summarize important information gathered during the review discussions, in surrounding case studies, and from the NP program in general. These findings are organized by topic area for simplicity and follow common themes:

- Technology support (e.g., networking, computation) within and external to major NP experimentation sites.
- Changing workflow requirements, and the impacts this will have on software and hardware development.
- Data set growth and usage expectations now and into the future.
- Discussion specific to ESnet architecture and support.

2.1 Technology Support

- In order for the ASCR HPC facilities to integrate well with the larger physics experiments, the supercomputers need to support outbound connections from compute jobs. There was also mention of wanting to stream experiment data directly to compute memory. This is not viewed as an immediate need, but a future-looking direction to explore.
- Multiple NP experiments (including STAR and ALICE), expressed a desire for a more uniform interface to the DOE/ASCR HPC facilities. Also, the superfacility model would be easier for experiments to adopt if there were a common interface/API for all three ASCR HPC facilities.
- The Science DMZ model has been adopted by numerous NP facilities and experiments and has caused an increased reliance on the network as a critical component of scientific workflow.
- Opportunistic computation, via efforts like the OSG, are attractive to experiments with flexibility for aspects of their workflow (reconstruction, simulation, analysis). Allocation of computing resources (e.g., Argonne Leadership Computing Facility [ALCF], NERSC, Oak Ridge Leadership Computing Facility [OLCF], and NSF centers) are also gaining in popularity due to the regular nature of the availability and performance capabilities.
- Commercial cloud use has been experimented with but usage remains low for scientific use. The costs of converting (e.g., software development, workflow adaptation) only makes sense in "bursting" scenarios when there is a critical need and cost to transfer data/compute hits a critical inflection point. It is expected that groups will stage pilots in this area in the coming years, but never fully adopt the mechanism except for certain specific tasks.
- PerfSONAR monitoring of network resources has been a critical help to high-energy physics (HEP) collaborations (e.g., the Large Hadron Collider [LHC]) and is being investigated by others with wide-area needs (e.g., ALICE, JLab to BNL/NERSC workflows).

2.2 Scientific Workflow

- Planned new detectors at JLab (MOLLER/SoLID) are early in the design process and have the opportunity to adopt some computational and networking designs developed by others (e.g., GRETA).
- HPC centers have a set of strategic architectural issues to solve with regards to the streaming workflow, e.g., the ability for worker nodes to have external network access or fetch data from a cache that will stage data for them.

- Software design for a given workflow is inherently tied to specific systems, which limits flexibility: locations where things can run, along with environmental factors (e.g., having full access to networking to support streaming versus use of local staged data copies). Adding flexibility into the software with regards to type of computational technology (e.g., central processing unit [CPU] versus graphics processing unit [GPU]), as well as locality of data, would facilitate a richer set of possible run-time environments.
- GRETA makes some design assumptions that can be adapted to future use cases that include separation of the control and data channels, ability to use local or remote computation, and ability to function with or without external networking capabilities (for storage and computation).
- Some aspects of scientific workflows (reconstruction, simulation, analysis) can occur at remote computational facilities if there are sufficient bandwidth and storage resources available between source and destination to facilitate bulk transmission of experimental data sets. Software that supports streaming assists in facilitating this workflow.
- Some aspects of experimental workflows (e.g., calibration) are better done locally, unless increased resources are put into low-latency/high-bandwidth paths that can meet real-time requirements.

2.3 Data Set Growth and Management

- Data volumes from experimental sources are increasing across the board as old detectors age out, new are built, and facilities upgrade capabilities.
- The EIC era will see increased data volumes and computational needs.
- Computation and storage needs are growing beyond the local capacity of many experiments. Software is available to facilitate new models (e.g., streaming, remote data analysis) but will require support of networks, as well as computational centers to support the needs.
- The development of advanced portals (built on the Modern Research Data Portal design pattern) is attractive for some collaborations to share aspects of research (e.g., models for simulation, data outputs for user analysis) with collaborators and the wider scientific community.
- Use of advanced data transfer tools (XRootD, Globus, etc.) is now common, as this is the interface that enables high-speed transfer to national HPC facilities).
- Some experiments continue to refine their data formats (size, quantity) to facilitate more efficient mechanisms for data sharing. This will greatly assist analysis activities at locations that are not as well connected to the global internet infrastructure.

2.4 ESnet Specific

- ORNL projects will need multiple 400G connections to ESnet after 2020.
- JLab will require upgraded capacity, preferably on diverse paths, to ESnet. Current needs are already taxing the current infrastructure.
- Some experimental groups would like to explore network mechanisms (similar to LHCONE) to facilitate remote workflow needs that must cross a network.
- There is interest from some scientific communities (LHC, STAR) in receiving ESnet 6 telemetry data for use in possible machine-learning (ML) activities to assist with computational/network/storage workflow needs.

3 Review Action Items

ESnet recorded a set of action items from the NP-ESnet Requirements Review, continuing the ongoing support of collaborations funded by the NP program. Based on the key findings, the review identified several actions for NP, ASCR, and ESnet to pursue.

ESnet will:

- Start a discussion between ESnet engineering and experimental representatives that are interested in sharing ESnet 6 telemetry data.
- Consider the creation of LHCONE-like overlay networks for certain use cases.
- Continue discussions with ORNL regarding wide-area connectivity options and amounts.
- Continue discussions with JLab / ELITE / MARIA regarding wide-area connectivity options and amounts.
- Publish findings from GRETA work as a guide for future experimental design.
- Assist groups looking to measure and understand wide-area performance expectations with tools such as perfSONAR.
- Facilitate discussions with groups looking to adopt Modern Research Data Portal design considerations.
- Facilitate peering with commercial clouds, as needed, for experiments that are looking into pilot efforts.

The NP community will:

- Continue discussions with ASCR facilities with regard to the new and expanded usage of computational and storage resources for certain experimental workflows.
- Quantify the increasing data needs as two major eras begin: the EIC for NP experimentation and Exascale Computing at ASCR HPC facilities.
- Further refine the data formats (size, quantity) produced by experiments to facilitate more efficient mechanisms for data sharing as multi-facility workflows are adopted.

The ASCR community will:

- Collectively work towards building infrastructure that supports a streaming workflow, e.g., the ability for worker nodes to have external network access, and for remote data sources to stream data directly to compute resources.
- Begin discussions about ways to offer a more uniform interface to the DOE/ASCR HPC facilities.
- Work with NP experiments to further explore the development of advanced portals (built on the Modern Research Data Portal design pattern) to share aspects of research (e.g., models for simulation, data outputs for user analysis).

4 Requirements Review Structure

Requirements reviews are a critical part of a process to understand and analyze current and planned science use cases across the DOE Office of Science. This is done by eliciting and documenting the anticipated data outputs and workflows of a particular program, user facility, or project to better inform strategic planning activities. These include, but are not limited to, network operations, capacity upgrades, and other service investments for ESnet as well as a complete and holistic understanding of science drivers and requirements for the program offices.

The requirements review is an in-person event. It is by design a highly conversational process through which all participants gain shared insight into the salient data management challenges of the subject program/facility/project. Requirements reviews help ensure that key stakeholders have a common understanding of the issues and the potential actions that can be taken in the coming years.

4.1 Background

Through a case study methodology, the review provides ESnet with information about:

- Existing and planned data-intensive science experiments and/or user facilities, including the geographical locations of experimental site(s), computing resource(s), data storage, and research collaborator(s).
- For each experiment/facility project, a description of the "process of science," including the goals of the project, how experiments are performed, and/or how the facility is used. This description includes information on the systems and tools used to analyze, transfer, and store the data that are produced.
- Current and anticipated data output on near- and long-term time scales.
- Timeline(s) for building, operating, and decommissioning of experiments, to the degree these are known.
- Existing and planned network resources, usage, "pain points," or bottlenecks in transferring or productively using the data produced by the science.

4.2 Case Study Methodology

The case study template and methodology are designed to provide stakeholders with the following information:

- Identification and analysis of any data management gaps and/or network bottlenecks that are barriers to achieving the scientific goals.
- A forecast of capacity/bandwidth needs by area of science, particularly in geographic regions where data production/consumption is anticipated to increase or decrease.
- A survey of the data management needs, challenges, and capability gaps that could inform strategic investments in solutions.

The case study format seeks a network-centric narrative describing the science, instruments, and facilities currently used or anticipated for future programs; the network services needed; and how the network will be used over three timescales: **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 template has the following sections:

Science Background provides a brief description of the scientific research performed or supported: the high-level context, goals, stakeholders, and outcomes. It summarizes the data life cycle and how project components are involved.

Collaborators captures the breadth of the science collaborations involved in an experiment or facility focusing on geographic locations and how data sets are created, shared, computed, and stored.

Instruments and Facilities includes a description of the instruments and facilities used, including any plans for major upgrades, new facilities, or similar changes. When applicable, descriptions of the instrument or facility's compute, storage, and network capabilities are included. The section offers an overview of the composition of the data sets produced by the instrument or facility (e.g., file size, number of files, number of directories, total data set size).

Process of Science includes documentation on the way in which the instruments and facilities are and will be used for knowledge discovery, emphasizing the role of networking in enabling the science, where applicable. This should include descriptions of the science workflows; methods for data analysis and data reduction; the integration of experimental data with simulation data; or other use cases.

Remote Science Activities details the use of any remote instruments or resources used in the process of science and how this work affects or may affect the network. This could include any connections to or between instruments, facilities, people, or data at different sites.

Software Infrastructure discusses the tools that perform tasks such as data source management (local and remote), data sharing infrastructure, data movement tools, processing pipelines, collaboration software, etc.

Network and Data Architecture includes a description of the network architecture and bandwidth for your facility and/or laboratory and/or campus. It includes detailed descriptions of the various network layers — the Local Area Network (LAN), Metro Area Network (MAN), and Wide-Area Network (WAN) — capabilities that connect your science experiment/facility/data source to external resources and collaborators.

Cloud Services, if applicable, covers cloud services that are in use or planned for use in data analysis, storage, computing, or other purposes.

Data-Related Resource Constraints reviews any current or anticipated future constraints that affect productivity, such as insufficient data-transfer performance, insufficient storage system space or performance, difficulty finding or accessing data in community data repositories, or unmet computing needs.

Outstanding Issues is an open-ended section where any relevant discussion on challenges, barriers, or concerns that are not discussed elsewhere in the case study can be addressed by ESnet.

5 Nuclear Physics Case Studies

The case studies presented in this document are a written record of the current state of scientific process and technology integration for a subset of the projects, facilities, and PIs funded by the Office of NP of the DOE Office of Science. These case studies were discussed fully in person, with additional conversation on certain topics driving the in-person review held on May 8 and 9, 2019.

The case studies were presented, and are organized in this report, to provide an overview of critical resources needed for operation from the level of individual experiments, up to the level of larger facility and overall site-specific needs. The case studies presented include:

- JLab: Facilities
- JLab: MOLLER
- JLab: SoLID
- JLab: Theory Group and LQCD
- JLab: EIC
- FRIB
- GRETA
- ANL: Gammasphere / ATLAS
- ANL: CLAS12 / EIC
- BNL: RHIC and RACF
- BNL: STAR
- BNL: Pioneering High-Energy Nuclear Interaction eXperiment (PHENIX) / sPHENIX
- CMS Heavy Ion Experimentation
- ALICE Project and ALICE-USA Computing

Each of these documents contains a complete set of answers to the questions posed by the organizers:

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

A summary of each is presented prior to the case study document, along with a "Discussion Summary" that highlights key areas of conversation from authors and attendees. These brief write-ups are not meant to replace a full review of the case study but provide a snapshot of the discussion and focus during the in-person review.

5.1 Thomas Jefferson National Accelerator Facility (JLab)

The Thomas Jefferson National Accelerator Facility (JLab) operates the Continuous Electron Beam Accelerator Facility (CEBAF) accelerator, a polarized electron source and injector along with a pair of superconducting radio-frequency linear accelerators, that provides the core source of research activity. Recent facility upgrades have increased the scientific output, posing new challenges for the technology support strategy. The need for networking, storage, and computation remains high, and the facility is working to address these needs through a set of scalable solutions that are local, regional, and national in nature. Experimental workflows that used to rely on local resources are experimenting with the use of remote computation via ESnet and allocation ASCR computing facilities. Future upgrades to detectors, and the installation of new experiments, will require more technology support that is currently being researched and implemented.

5.1.1 Discussion Summary

JLab featured five case study elements:

- Facility Overview, including aspects that related to all experimental networking and computing.
- MOLLER, a planned detector fo Hall A.
- SoLID, a planned detector for after MOLLER.
- EIC, an Electron Ion Collider facility for the future.
- LQCD / Theory, a wide-ranging effort with distributed collaborators.

The discussion items in the subsequent section are separated by area but share many common aspects due to the shared facility nature.

5.1.2 JLab — Facility Notes

- The JLab facility upgrade from 6 GeV to 12 GeV has increased capabilities of experiments/ detectors. As a result of this, greater data volumes are now a regular occurrence. JLab experimentation (within halls A, B, C, and D) runs concurrently throughout the year. Downtimes for any specific experiment do not affect the others.
- JLab data volumes vary by experimental location:
  - Halls A and C are the highest intensity parts of the facility and are capable of producing 10–100 terabytes (TB) per experimental run. The produced data does not require a large amount of computation.
  - Halls B and D house GlueX and CLAS12, the current area where data production will grow to be the largest facility-wide in the coming years.
    - JLab's CLAS12 generates approximately 13KB per event with an average data rate of 300MBps (600MBps max). This runs one-third of the year and produces about 150–200 TB per year.
    - JLab's GlueX currently operates in low-intensity mode, which is about 700MBps or around 2 petabytes (PB) of data per year. When in high-intensity mode, it can produce 1.5 to 2GBps, or 6PB of raw data per year. Trigger refinements are being improved to shrink some of these. Limiting factors (e.g., power) prevent GlueX from running at a higher rate.
- JLab has four computational workflow activities: simulation, calibration, reconstruction, and data analysis. Currently not all of these workflows can be supported internally. External sources (OSG for GlueX, NERSC for GlueX and CLAS12) are used on occasion for certain parts of the workflow. This balance was struck to prevent much of the computational infrastructure from sitting idle, given that there is not always a pressing need for the max capacity. It is typical that demands are not level across experimental campaigns. Compute demand is typically highest in early workflow stages, such as calibration, and then again at the end when reconstruction is needed.

- JLab calibration workflows are ~5-10% of a raw data set size and are used to prepare before further computation. These are done at JLab using local computation and storage resources and typically as an experiment is running, and may produce GB of output that is stored in local databases. This particular part of the workflow requires many "reads" of the data, which makes it storage/computationally expensive.
- JLab reconstruction workflows involve taking an entire bulk data set and converting from raw observation to processed formats. Due to the need to access the entire data set, WAN bandwidth is a limiting factor. This is typically done locally as a result, due to the need to move several PB. Output size is smaller, typically 10% of input size (TB or GB). This step can start only after the calibration step. This particular part of the workflow does not require that many "reads" of the data, which makes it less storage/computationally expensive.
- JLab would like to experiment with a streaming workflow for reconstruction, wherein time on a large computing resource (NERSC, OSG, etc.) is used instead of local computation. Given there is a single/small number of reads for the data, this will translate well to the environment if enough WAN bandwidth exists. GlueX reconstruction using remote resources (OSG/NERSC) is possible, but it can completely overwhelm local networking capabilities (current 10Gbps wide-area connectivity). It is routine to have 2Gbps to 6Gbps "spikes" during a data movement activity off-site, which has the potential to affect other facility network needs through primary connection. This frees local compute but does consume local network resources.
- JLab statistical analysis workflows are used to process reconstructed data to find events. Since these statistical analysis workflows make use of reduced data sets versus the larger raw datasets gathered from the detector, work can often be performed remotely.
- JLab simulation workflows are derived data sets from modeling programs and input from actual observation. These are less tied to a running experiment and are smaller in size. Thus, they can be run remotely without a large networking component. Multiple simulations can be run with the same base set of data by changing small-sized input parameters. The heaviest need on the network is returning a simulated data set to JLab.
- JLab storage uses Lustre with ZFS storage nodes. These storage nodes can sometimes struggle to keep up with workflow needs, which are characterized by many small parallel computational jobs requiring read/write operations. An effort is underway to find a storage resource with higher speed solid-state drive storage to serve as a gateway to the WAN transfers. This could even be a set of data-transfer computers with fast local storage connected to the SAN.
- Experimentation that utilizes JLab resources (instrumentation, computation, storage, networking) must be internally reviewed and scheduled against institutional priorities. Given the limited nature, experimental wait times could exceed several years (e.g. as of mid 2019 the list of experiments represents a backlog of between seven and nine years).
- JLab works with the ELITE regional network in coastal Virginia that connects to MARIA (in northern Virginia) and then connects to ESnet. Currently the entire JLab facility operates off a 10Gbps data connection; efforts are being made to create a second 10Gbps connection. Ideally the lab would like to connect at a higher rate, if the local regional network can support it or if agreements can be made to bring higher speed connectivity directly into the JLab facility.
- JLab "business traffic" shares the network with research traffic and relies on cloud-provided services (e.g., mail, etc.). If the network is being overwhelmed with science traffic, this has the potential to affect lab operations.

- JLab uses a Science DMZ architecture with policy-based routing (PBR) to re-direct packets through parts of the network that could slow science traffic.
- NERSC and JLab will continue conversations about how to integrate more JLab use of computation resources. NERSC has worked with other single instruments (e.g., the National Center for Electron Microscopy) and may be able to do the same for JLab provided a clean and efficient network path can be established via ESnet. Latency is not that much of a factor for JLab. New NERSC computation resources (Cray Slingshot, etc.) can assist with this connection between computation, network, and storage.
- JLab's use of external resources (NERSC, OSG) is largely left to individual experiments. Some trust external computation and have made the effort to convert workflows since computational resources locally are not always available. Others are willing to wait for local computation. As experiments have greater data needs/computational timelines, it is expected more will compute off-site and will need a fast network to facilitate.
- JLab is not exploring remote control of experiments fully at this time. Some automation is possible, but many manual aspects to experimentation (e.g., a human needing to manipulate parts of the equipment) remain. Some control systems must be physically segregated from networks.
- JLab is planning ahead for 40G/100G networking and will explore data-transfer nodes (DTNs) / perfSONAR nodes that can connect at those speeds in the future.

5.1.3 JLab — Measurement of a Lepton-Lepton Electroweak Reaction (MOLLER) Notes

- JLab's MOLLER is a planned detector for Hall A.
- The MOLLER experiment will acquire about 4 PB of raw data over a two- to three-year period (e.g., GlueX acquires 6 PB of raw data per year).
- The MOLLER experiment will produce an intermediate step in the analysis and generate about 20PB of data during this phase. This intermediate step is expected to be stored and computed entirely at JLab.
- Over the course of the MOLLER experiment and analysis (four-plus years), about 1PB of data will be transported off-site. These data would consist of small subsets of the raw data and data that have been highly reduced in size.
- MOLLER simulation work is and will be carried out at collaborator institutions.

5.1.4 JLab — Solenoidal Large Intensity Device (SoLID) Notes

- SoLID is a planned detector for Hall A and will not be deployed until after the MOLLER work. This is expected in six to seven years. Both MOLLER and SoLID are stepping stones to the EIC.
- SoLID has a large number of potential collaborators: 300 at over 70 institutions.
- SoLID will require simulation, data taking, and data analysis. Most of this will occur at JLab on computational resources there but could occur at other locations. The workflows are still being designed along with the instrument.
- The SoLID detector has two operational modes:
  - In Semi-Inclusive Deep Inelastic Scattering (SIDIS) J/Psi mode, it will produce data at a rate of 3–4GB/sec. Over the course of three years, this will result in 100 PB in total of raw data.
  - In Parity-Violating Deep Inelastic Scattering (PVDIS) mode, the raw data rate off the detector will be 6GB/s, which would add 175 PB of raw data over the same three-year period.

- SoLID is considering a streaming data acquisition (DAQ) system that would result in much higher data rates coming from the detector and increase the numbers listed previously, which assume a more standard capture system based on current detectors. In either case, the WAN requirements will not change for SoLID.
- Most data processing for SoLID will occur using internal computation resources. Some other JLab experiments are now using external computational resources such as NERSC, OSG, and SoLID may also do so in future. It is likely that highly reduced and pre-processed SoLID data sets eventually will be exported off-site for such use of external resources. An initial estimate for off-site data movement for external SoLID computation is approximately 25 PB (a very rough estimate) of raw or reduced data over the next six years of data generation.

5.1.5 JLab — Electron-Ion Collider (EIC) Notes

- At the time that this review was held, the DOE Office of Nuclear Physics was in the process to select accelerator facility conceptual designs. In January 2020, BNL was chosen by the EIC program office.
- The planned EIC requires an accelerator facility with a high luminosity and a versatile range of beam energies (center of mass range ~20 to ~100 GeV, upgradable to ~140 GeV), beam polarizations (longitudinal, transverse, tensor polarization of at least 70%), and beam species (proton as well as ion beam from D to heaviest stable nuclear). Two pre-conceptual designs are being proposed for the EIC, eRHIC at BNL and the Jefferson Lab Electron-Ion Collider (JLEIC).
- The EIC User Group currently consists of more than 864 physicists from over 184 laboratories and universities in 30 countries.
- When the EIC begins operation, it will produce about 250GB/sec of data. The trigger is expected to achieve data reduction of 30x to 40x, but the degree of reduction is unknown.
- EIC computational and WAN bandwidth needs are unknown. It is known that BNL and JLab will collaborate heavily and could benefit from increased facility to facility capabilities on the network.
- Scientists doing simulation work on the EIC need to move simulation data sets between sites, and those data sets are 10TB to 100TB in size.

5.1.6 JLab — Lattice Quantum Chromodynamics (LQCD) / Theory Notes

- The U.S. Lattice Quantum Chromodynamics (USQCD) Collaboration is a consortium of about 160 people at about 50 institutions, including universities and national labs. Main computing facilities include JLab, Fermilab (FNAL), and BNL.
- Computational calculations are coordinated and occur at DOE and National Science Foundation (NSF) computing facilities, including ORNL (OLCF), ANL, and NERSC, which are used for the parts of the calculations. Results from those calculations are brought back to the coordinating facilities for further analysis on local computing facilities.
- Members of the USQCD project teams apply for computational time at computing facilities. Some data may be kept for an extended period on the leadership systems, such as NERSC. In general, though, the member labs provide long-term storage.
- JLab LQCD/Theory features two steps to the process:
  - Configuration generation: step to create the configuration needed to perform the calculation. This step is resource intensive but required only one time. The output of this process requires storage resources and the ability to share with others (final location can be U.S.-based or international).

  - Calculation: use of the configurations to do a scientific calculation. A large pool of participants from around the world participate in this step, which involves taking a configuration, performing work, and sharing results. A portal system would be useful to this work, either centrally managed at a facility or tied to multiple facilities. NERSC has such a system in place, which has caused a large use in NERSC resources due to proximity.
- The NERSC QCD portal (http://qcd.nersc.gov/) is a good candidate for development along the lines of the Modern Research Data Portal design pattern.
- LQCD data sets vary in size, location, and usage pattern:
  - Data sets are typically about 10 TB in size spread over about 1000 files. Secondary calculations are 100 TB spread over about (500K) files. All data sets are transferred to JLab for further analysis over the period of a yearly allocation. In total, a few hundred TBs are produced off-site and transferred to JLab.
  - Data sets produced at HPC facilities are combined with other data sets, on order 200 TBs are generated at JLab and consumed there. These additional data sets generally do not leave JLab.
  - Global analysis data sets are a few GB. These might be transferred off-site, but the network demands are low.
- ORNL/Summit and NERSC/Perlmutter may use the same workflows in future, and future five to ten factor increases in datasets may result in likewise increased data transfers back to JLab.
- Exascale use has targeted about 50 times improvement on their benchmarks; thus, they expect to see about 10 times more data produced.
- Typically, about 30 projects are allocated to the JLab LQCD facility, and in total 1PB disk storage is allocated to the projects with 5.5PB of tape used. These projects produce a few 10 to 100 TB of data over a yearly allocation, and some fraction of this is transferred off-site.

5.1.7 Facility Case Study

5.1.7.1 Background

Thomas Jefferson National Accelerator Facility (at JLab) is funded by the Office of Science for the DOE. As a user facility for scientists worldwide, its primary mission is to conduct basic research of the atom's nucleus at the quark level.

As a center for both basic and applied research, JLab also reaches out to help educate the next generation in science and technology. JLab is a user facility offering capabilities that are unique worldwide for an international community of nearly 1,600 active users. One-third of all PhDs granted in NP in the United States are based on JLab research, with 608 PhDs granted to date and 211 in progress.

Complementary to the NP experimental program, JLab also hosts a computation and theory center that is an active participant in LQCD theory. As part of this program, JLab hosts an LQCD Computing Facility.

Raw data from experiments are created at JLab and stored locally on tape. A backup tape copy of the raw data is also kept at JLab. As tape technologies evolve, previously archived raw data are copied onto new media. Combined with the increase in capacity of media, this means that currently all of the raw data taken in the lifetime of the laboratory are still stored in the tape library at the lab.

How the raw data are processed to derive publishable physics results is described in a later section.

5.1.7.2 Collaborators

The more than 1,600 JLab users are from 278 institutions in 39 countries (see **Figure 1**).

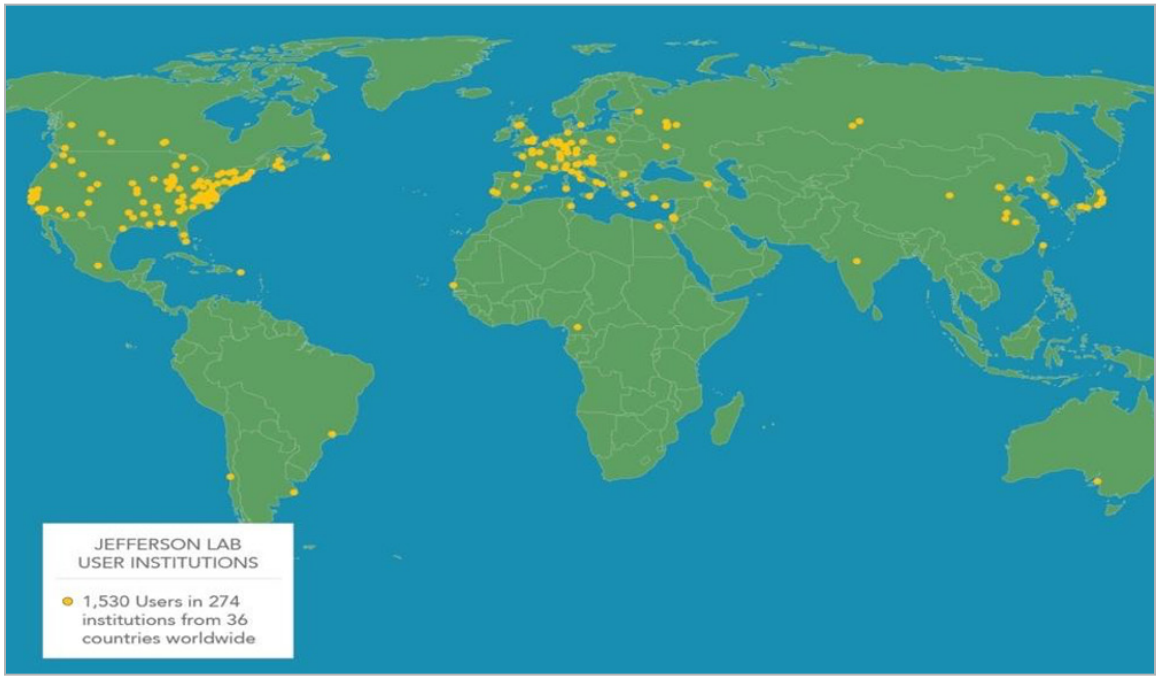

Figure 1. Geographic distribution of JLab collaborators.

The number of users varies depending on the institution. At any time, only a fraction of the users are physically present on the JLab site. Most users use JLab resources remotely, either by logging into JLab computing resources directly, by using remote conferencing capabilities, or by moving data to/from JLab. **Figure 2** shows the distribution of non-U.S. JLab collaborators by country.

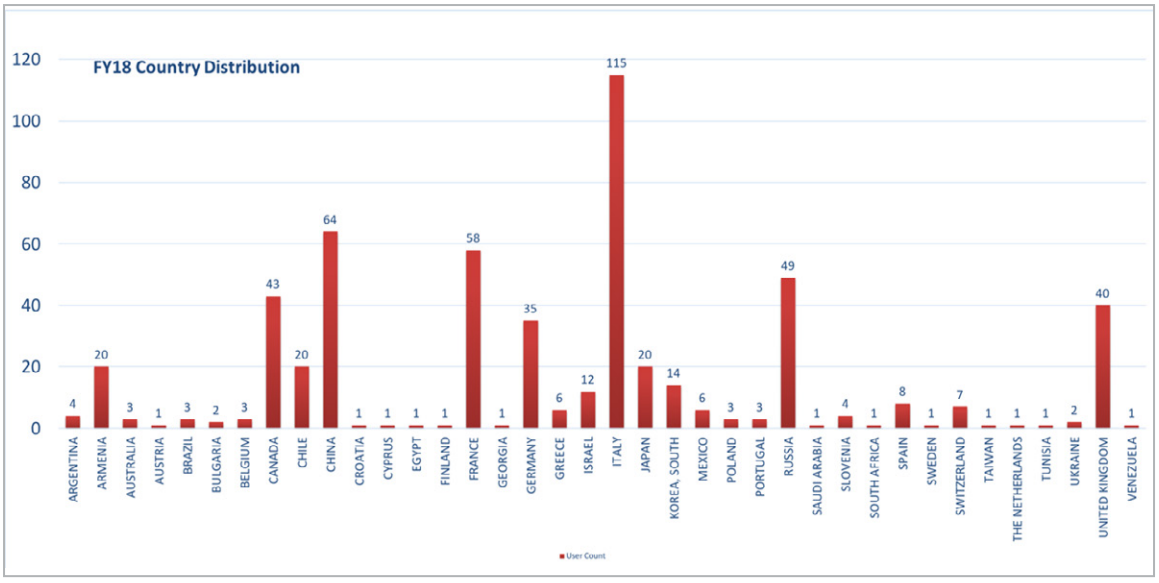

Figure 2. Distribution of non-U.S. JLab collaborators.

5.1.7.3 Instruments and Facilities

JLab's primary scientific instrument is CEBAF. CEBAF is a high-intensity electron accelerator with unique capabilities to probe the nuclear structure of matter at the quark level. Experiments are housed in one of four areas — hall A, B, C, or D — and CEBAF is able to deliver beams to all four halls simultaneously. The four halls are instrumented with particle detectors and ancillary equipment that allow experiments to study different aspects of nuclear science. In all four halls, the science is studied in a similar way: either an electron or photon beam (depending on the hall) is directed onto a target. In all four halls, a similar basic science pattern of measurement exists:

1. Either an electron or photon beam (depending on the hall) is directed onto a target. Targets can be solid, liquid, or high-pressure gas. In certain materials at low temperatures, the spins of nuclei in the target can be aligned to create polarized targets. In addition, the spin of electrons or photons in the beam can be polarized with respect to the beam direction. Electrons or photons striking the target can interact with atomic nuclei in the target material.
2. An array of detectors then measures the properties of the particles created or scattered in the interaction. A mix of commercial and custom electronics converts the analog signals from the detector into a digital representation. The data generated by the detectors in response to a single interaction are known as an event.
3. Frequently only a fraction of the interactions is of interest to a particular experiment. The data rate to storage can be considerably reduced by using some of the detectors to identify interesting interactions and trigger readout of the corresponding event. JLab facilities allow a range of experiments where beam type, energy, luminosity, and polarization can be varied along with target material, target polarization, and detector packages.

5.1.7.3.1 Hall A

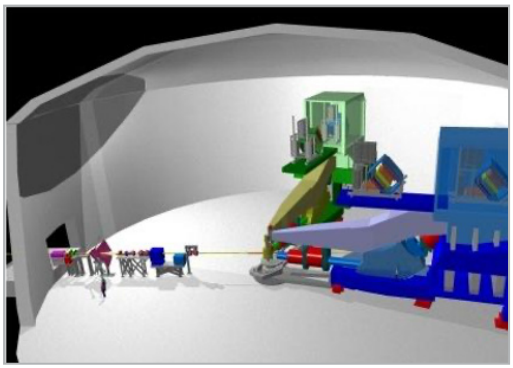

Figure 3. JLab Hall A & CEBAF.

Hall A (see **Figure 3**) is physically the largest of the four halls. It contains two precision high-momentum spectrometers that are movable to various angles. These are used to study interactions of the electron beam with a target at the pivot point of the spectrometers. This configuration limits the solid angle coverage of the detector packages, which are mounted in shielded huts at the end of the spectrometer arms. Consequently, the amount of data generated per interaction is only a few kilobytes. However, the high intensity of the beam provided by CEBAF produces an interaction rate up to hundreds of kilohertz. Typical Hall A experiments are short lived, weeks or months on the floor. The experiments differ by beam conditions, such as energy, intensity, and polarization; by type of target; and by the addition of equipment to supplement the base installation. Since Hall A has a large floor area, it is ideal for custom installations that are not part of the base equipment. Two examples are the proposed SoLID and MOLLER detectors.

5.1.7.3.2 Hall B

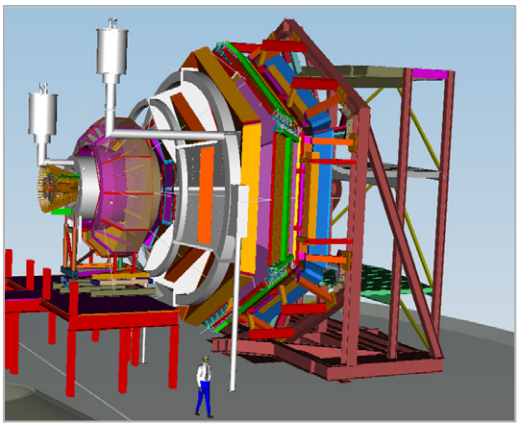

Figure 4. CLAS12 Detector.

Hall B is currently instrumented with the CLAS12 detector (see **Figure 4**). CLAS12 is composed of many individual detector packages that are used together in various combinations to study a wide variety of physics. The momentum of the incoming beam causes particles from the interaction to predominantly travel in the beam direction. To take advantage of this, the majority of the detectors are positioned downstream of the target and present a large solid angle to the interaction products.

Hall B can operate in two modes, either using the CEBAF electron beam directly or using a beam of high-energy photons derived from the electron beam. In photon-beam mode, the rate of interactions is much less than the electron beam rate, and the data rate from the detector is correspondingly smaller. The CLAS12 detector is designed to simultaneously run multiple experiments that have the same target and beam condition requirements. For example, Run Group A that took data in the fall of 2018 produced a data set that will be shared by 13 different studies. In its highest rate mode, CLAS12 takes data at ~300 MB/s, about 20 kB per event, ~2-3 PB/yr.

5.1.7.3.3 Hall C

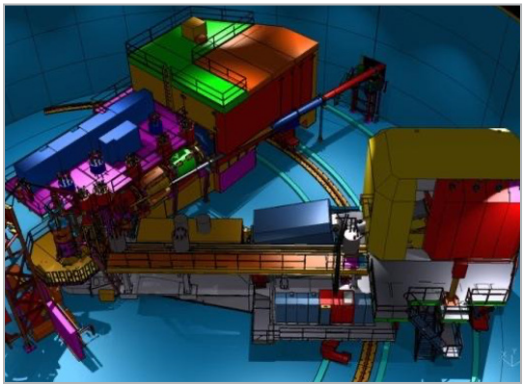

Figure 5. JLab Hall C 2 Armed Spectrometer.

Hall C is similar to Hall A as the main instrument is a two-arm spectrometer (see **Figure 5**). The principle difference is that the arms are asymmetric. Also, in common with Hall A, the experiment program in Hall C consists of a series of relatively quickly performed scientific workflows. Unlike Hall B, Halls A and C typically do not run simultaneous experiments that share data sets. Although Hall C is used to perform experiments requiring high-beam energies and luminosities, typical data amounts are fairly small, and the rate of data generation is low.

5.1.7.3.4 Hall D

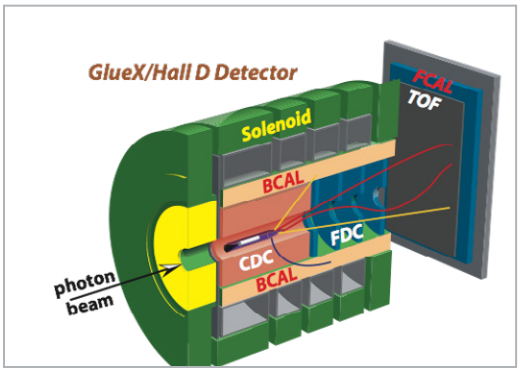

Figure 6. JLab Hall D / GlueX Detector.

Hall D is the newest of the four halls and was constructed as part of a recent accelerator upgrade. The main instrument in Hall D is the GlueX detector, which is shown in **Figure 6.** GlueX is designed to operate with a photon beam generated using the electron beam from CEBAF. One of the main features of Hall D is the availability of a high-intensity beam of polarized photons. The GlueX target is surrounded by the detector, which is itself inside a solenoid magnet. The detector is asymmetric in the direction of the beam since most of the particles coming from interactions in the target travel in that direction. GlueX has recently completed GlueX Phase I operation using a relatively low-luminosity photon beam,

which generates 500 MB/s and 43 kHz event rate. In the fall of 2019, operation recommenced with a higher luminosity photon beam, 90 kHz event rate, and 1.5GB/s data rate, ~ 6PB/yr. Phase II is scheduled to continue for five years. The GlueX collaboration has more than 125 members.

The properties of the detectors in the four halls are summarized in **Table 1.**

Hall D	Hall B	Hall C	Hall A
Polarized photons	Electron luminosity 10^{35}	Electron luminosity up to 10^{38}	
Excellent hermeticity	Good hermeticity	Precision spectrometers	
Photon energies of ~8.5-9 GeV		11 GeV beamline	
Photon luminosity 10^{8} photons/s		Target flexibility	
Good momentum/angle resolution		Excellent momentum resolution	
High multiplicity reconstruction		Energy reach	Custom installations

Table 1. Properties of JLab detectors.

5.1.7.3.5 Compute and Storage

The compute and storage requirements for the experimental program are determined by the groups performing experiments in the halls with the assistance of a per-hall computing coordinator. An overall coordinator for Experimental Nuclear Physics (ENP) coordinates the requirements for the entire program, including all four halls and any other ENP activities. These requirements are reviewed at least twice per year internal to JLab and once per 12 to 18 months by an external review panel. The ENP computing coordinator works with JLab IT to develop a strategy that will meet the requirements in the current to two-year timescale and plan for experiments and equipment upgrades lying further in the future. The most recent external review was November 2018. The largest perturbation to steady state foreseen in the next few years is the switch to GlueX Phase II, which occurred in the fall of 2019. Beyond that, the next planned upgrade will be the MOLLER detector in Hall A, followed by SoLID in the same hall. MOLLER and SoLID have their own short case studies.

The science workflow is described in more detail in the next section of this document, but an outline is needed here to understand the compute and storage requirements. While an experiment is taking data, events are accumulated with an average event rate and event size. The uptime for accelerator and detector, and the need for calibration and testing, means that production quality data are being collected only 60–70% of the time during an experimental effort. Each hall currently has local storage close to the detector that can hold at least 48 hours of data. The most demanding system on-site at the moment is GlueX in Hall D, which in the fall of 2019 took data at an average rate up to 1.5GB/s, which corresponds to an average 90 kHz event rate. Since these are average rates, networking between the detector and the local storage plus the storage itself are provisioned for 2GB/s to allow for higher than average rates. In the case of GlueX, ~580 TB of Redundant Array of Independent Disks (RAID) storage is available. Once the data are staged on the local storage, control is passed to systems operated by IT that copy the data over the LAN to staging storage in the data center. From there the data are archived in an IBM tape library with a current capacity of 27 petabytes over almost 9,000 tapes.

JLab local computing resources are diagrammed in **Figure 7.** As well as the systems described in the previous paragraph, there are clusters used for Theory/LQCD (covered in an additional case study). An approximately 200-node compute cluster is used by ENP to process data from the experiments, as described in the next section. A path from the mass storage system to the WAN is described in more detail in Section 7.

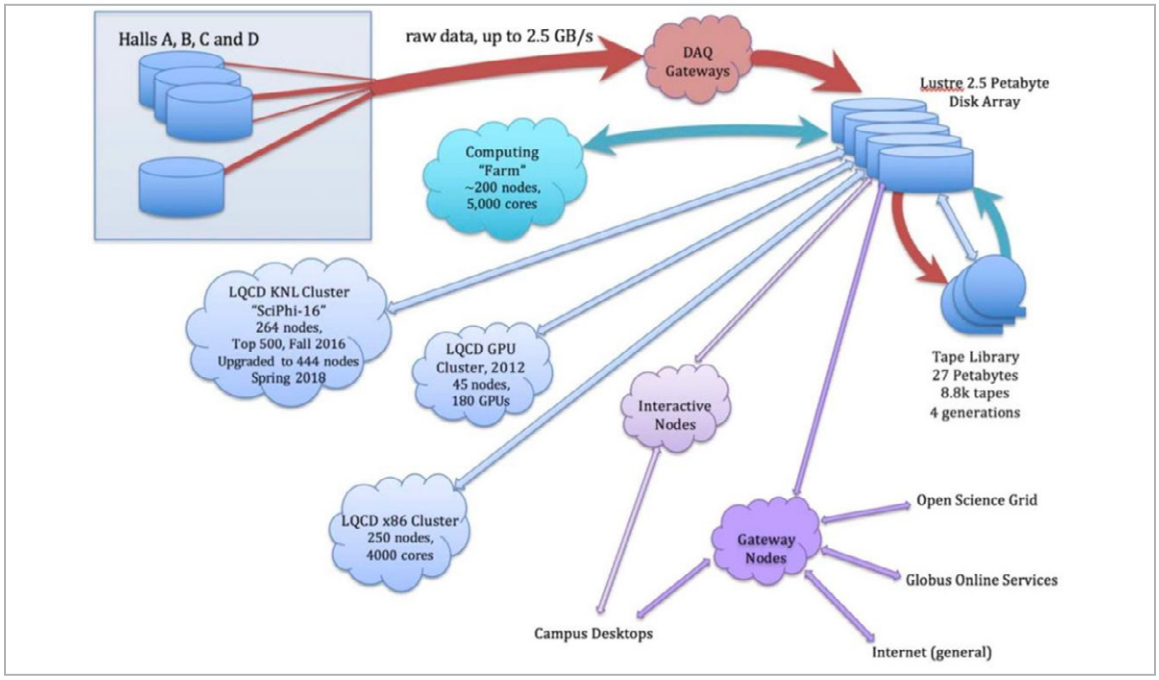

Figure 7. JLab compute and storage.

5.1.7.4 Process of Science

Baseline detector equipment in the four JLab halls was described in the previous section. Science activities begin with individuals or groups of PIs presenting proposals for experiments to a program advisory committee (PAC) tasked with prioritizing experiments and scheduling lab resources. The scheduling unit is the so-called "PAC day," which is equivalent to two calendar days. As of spring 2019, there were 2,650 PAC days of already approved experimental proposals, consisting of more than 50 individual experiments. Since several halls can receive beams simultaneously, one calendar year provides 300 to 370 PAC days of experiment time, and the current list of experiments represents a backlog of between seven and nine years. Each year, the PAC may approve more experiments or re-evaluate the continued importance of those already queued.

Experiments may use the baseline equipment, add to the base, or require entirely new detectors. A review process ensures that only experiments that are ready for beam time are given a firm schedule. When an experiment is running, a team, usually a mix of staff and users, is responsible for operating the detector and the data acquisition software and hardware that digitize, format, and store the data. Part of this process is monitoring of data quality and part is control of aspects of the detector and data acquisition. There has been much interest in providing the ability to remotely monitor and control experiments over the WAN. These are not tasks requiring high bandwidth but concerns of network availability and security currently limit this capability to passive monitoring or remote login via on-call experts from home.

Computational data processing has four components.

5.1.7.4.1 Calibration

A sample of the raw data set, often runs taken under controlled conditions, is used to calculate calibration constants that convert the digitized values from the detectors into derived measurements, such as position, time, energy, momentum, and particle type. The size of the sample is typically 5 to 10% of the total data set, but the calibration calculation is often performed multiple times as scientists try to understand the detector's behavior.

Typically, calibrations are stored in a database since parameters can vary over the lifetime of the experiment and are associated with groups of data-taking runs. The size of the database is typically a few gigabytes.

Calibration is very closely tied to the raw data set and is an important part of the data quality process while an experiment is running. It is also very frequently interactive as the act of trying to calibrate the detector often uncovers variances between the observed and predicted behavior of the detector. As a consequence, JLab provisions local resources to cover the calibration workflow.

5.1.7.4.2 Reconstruction

In the reconstruction phase, the calibration is applied to the bulk of the data set to convert raw data into measured physical parameters. Reconstruction requires access to the entire data set. Since the fall of 2019, GlueX has produced 6 PB/yr. The output of reconstruction is frequently smaller than the input since it now, for example, consists of a few numbers describing a particle's tracks rather than the coordinates of several hits where the track was detected by the detector. For example, GlueX reconstruction output is 10% of the size of the input, and so is about 600 TB/yr. Since the input to reconstruction is the entire data set, the ability to run off-site is restricted by the available WAN bandwidth.

5.1.7.4.3 Analysis

The output from reconstruction contains events that capture the physics being studied. The physics measurements that go for final publication are extracted by statistical analysis of these data. If more than one experiment is sharing the data set, the reconstructed data are first sorted into physics channels of interest to individual experiments. The size of this data set relative to the raw data varies by experiment. For example, GlueX is a single experiment but 80% of the events in the data set do not represent physics of interest to GlueX, so only 20% are included in final stage analysis. Even so, the GlueX raw data set, at ~6 PB/yr, is large enough that the input for analysis is still ~500 TB/yr. For some experiments taking part in the run groups using CLAS12 data only 1% of the total CLAS12 data set is of interest, and the analysis input is only tens of terabytes. In the case of some hall A and C experiments, the analysis input is only gigabytes. For most experiments, off-site transport of the data set used in final stage analysis is not an issue since the data sets involved are relatively small. The computing requirements are modest enough that the choice of local computing or remote resources can be left to the experiment.

5.1.7.4.4 Simulation

A fourth computational process is simulation. Here a model of the detector along with detector calibration and numerical models of physical processes are used to predict the events that will be measured by an experiment. The numerical models are usually provided by third-party packages, such as GEometry ANd Tracking (GEANT) 4, that are well tested and trusted. Simulation is used during the detector design process and to aid the choice of operational parameters prior to taking data. Once data have been taken, the calibration constants derived from real data can be used to improve the accuracy of the simulation, which is then used in the analysis phase to compare measurement with theory. Typically, an experiment simulates a number of events proportional to the number in the experimental data set with a ratio that depends upon the required statistical accuracy. Using an earlier example, the experiment that uses 1% of the CLAS12 data, ~50 TB, as the input to its analysis may require 10 times that number of simulated events, which is a 500-TB simulated data set. These data sets are large but, unlike the experimental data, they can be regenerated by rerunning the simulation. So it is typical to consume simulated data at source with simulation immediately followed by running reconstruction on the simulated data, which results in a factor of 10 decrease in data volume. The input to simulation is small, typically consisting of the detector model, magnetic field maps, and detector calibration databases. This makes simulation an ideal task to run off-site since the data required from JLab are small. However, an experiment may require simulation outputs be returned to JLab for storage or analysis. In such cases, WAN incoming traffic may be similar in scale to outgoing WAN traffic generated by physical result reconstruction.

The JLab scientific computing requirements reached a plateau in the fall of 2019, when GlueX Phase II running began. The next increase in requirements will be in at least five years from the new equipment, MOLLER and SoLID, in Hall A.

A measure of the compute load generated by the experiments is hours of computing time normalized to a standard type of processor core. In the JLab case that is the Intel Broadwell used in the 2018 upgrade to the local ENP compute cluster.

Figure 8 shows the compute requirement per year in million core hours per calendar year. The columns are split to emphasize the simulation (Monte Carlo [MC]) and non-simulation (non-MC) contributions from GlueX and CLAS12. On this timescale halls A and C contribute at a level smaller than the uncertainties on the GlueX and CLAS12 requirements and are ignored in this plot.

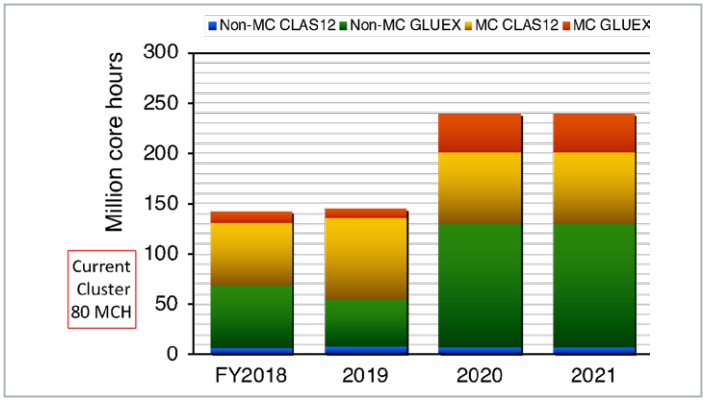

Figure 8. JLab Compute requirements by year.

Although GlueX moved to Phase II running in the fall of 2019, we do not anticipate that the compute load will increase until spring 2020. What is clear from the plot is that the existing JLab ENP cluster, at 80 MCH per year, was a factor of three too small to cover the entire workload after 2019.

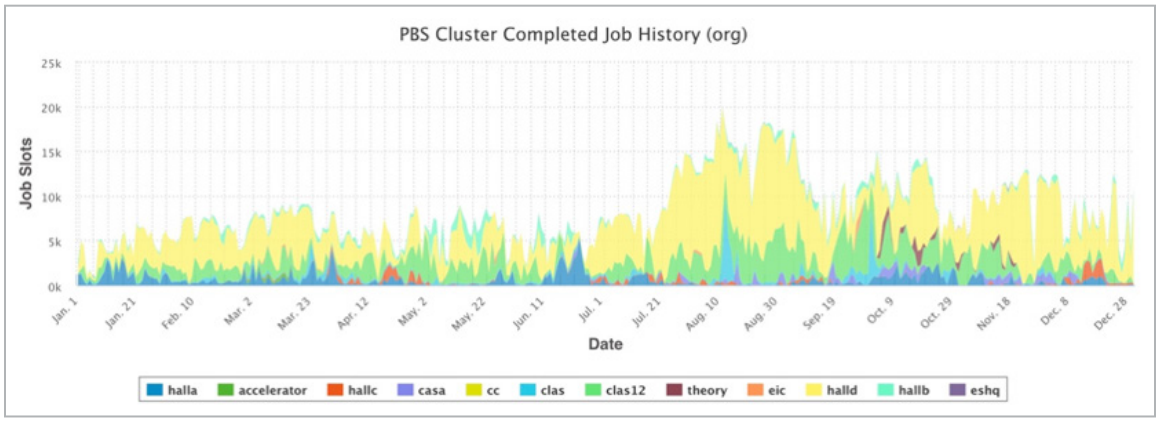

Figure 9. Job history, JLab ENP cluster for 2018.

Figure 9 shows the job history for the JLab ENP cluster for 2018. The chart is color coded according to the top-level groups responsible for the load. The vertical axis is "job slots," the number of simultaneous jobs that the batch system is running. It is clear that the workload is not constant. There are periods when the load on the cluster is high for weeks and periods where the system is almost idle. There are several reasons for this:

- The number of high-priority and high-workload users is low. These are basically the prime "cooking" accounts for GlueX and CLAS12 that run the bulk reconstruction. When they are not running, there are not enough other users with significant volumes of work to fill the cluster.
- Reconstruction can start only when calibration is successfully complete. Experiments may run a calibration, pause to examine the results, and go around the cycle several times before they are ready to run reconstruction on the entire data set. Even when calibration is complete, they reconstruct only moderate-sized batches of data, maybe 10% at a time, then pause to check data quality.
- It is also unreasonable that experiments should wait a whole year for the data to be processed before they can move on to analysis. The most efficient use of user time is to process as quickly as possible for any given computing capacity. Using Figure 9 as an example, GlueX and CLAS12 both have a fare-share allocation of 45% of the cluster, but in the period July to September GlueX used much more than its allocation. This was allowed since at that time CLAS12 had no work to divert resources from GlueX. Once that was done, the GlueX workload dropped off as the staff analyzed the results.

A further consideration is that the computing requirement calculations depend upon the raw data set size in events, the processing time per event, and a scaling factor that considers the details of the workflow. Until 2018, these numbers were based on a combination of assumptions and experience from prior detectors. CLAS12 and GlueX are starting to gain experience with real data, but the inputs to the requirements calculation are still being refined.

Taking all of these factors into consideration, the current JLab computing plan is to provision locally for all of the compute tasks that are most efficiently done, or required to be done, locally along with as much of the reconstruction as possible. The remaining reconstruction and the bulk of the simulation will use off-site resources as described in the next section.

5.1.7.5 Remote Science Activities

As described in the previous section, the JLab compute load is not uniform but consists of bursts of activity on top of an underlying background workload. It is not cost effective for the laboratory to provision for the peaks of the workload, since that would lead to large periods of underutilization. It is also not desirable to force researchers to slow their progress by spreading work over a longer time period. Section 2 of this document outlined the large number of institutions with which JLab collaborates. Many of these are members of the individual GlueX and CLAS12 collaborations and have significant computing resources of their own. To facilitate the use of these facilities, JLab has begun a collaboration with the OSG. GlueX has pioneered this approach at the laboratory and currently has access to resources at several collaborating institutions. For various reasons, some collaborators, such as the Italian National Institute for Nuclear Physics (INFN) in Italy and the Massachusetts Institute of Technology (MIT) in the United States, cannot make resources available to us using OSG. As part of the process to run our software on OSG, we are using containers (Singularity and DOCKER) and the CernVM distributed file system. This same process allows us to give containerized software and data to sites such as INFN and MIT for them to run locally. **Table 2** shows the contributions to GlueX data processing via OSG in the past year.

Site	Contribution
UConn	10 MCH
FSU	5 MCH
Northwestern	2 MCH
Regina	2 MCH
Florida International	2 MCH
Opportunistic	10 MCH

Table 2. GlueX OSG contributions.

GlueX has also pioneered the use of NERSC to process JLab NP data. NERSC preferred that we process raw data rather than simulate, so JLab has run GlueX reconstruction on a mix of CORI I and II at NERSC. In 2018 GlueX requested an allocation of NERSC units equivalent to ~45-90 MCH of computing (depending on whether it is CORI I or II) and received 23 MCH. To view this from a networking perspective, a 10Gb/s link can move enough data off-site to utilize 30 MCH per month. GlueX's goal is 120 MCH per year at NERSC, effectively all of its reconstruction. This assumes that the raw data files can be compressed by a factor of two before being moved over the network.

CLAS12 is following GlueX's lead in terms of use of NERSC and OSG.

5.1.7.6 Software Infrastructure

Prior to GlueX's accelerator upgrade from 6 to 12 GeV, and associated upgrade of CLAS to CLAS12, experimental generation of data was an order of magnitude less than will be the case over the next five years. As a consequence, all of the bulk computing for ENP at JLab was performed using the local cluster. Today a large fraction of the computing is still done locally but, as detailed earlier in this document, we are transitioning to a mode where we expect an increasing percentage of the work will be performed off-site.

The local systems consist of clusters of Linux compute nodes that are managed by a batch job submission system, Auger. Auger is the front end for underlying third-party batch software, and JLab is currently in the process of transitioning from utilizing Portable Batch System (PBS) to Slurm. The JLab scientific computing group has developed a workflow tool, Scientific Workflow Indefatigable Factotum (SWIF), that greatly simplifies the use of the batch system for users. In particular, SWIF takes care of mundane tasks, such as resubmitting failed jobs, and managing scheduling jobs that have dependencies, for example input data from tape. Based on feedback from SWIF users and the requirement for managing off-site workflows, a new version, SWIF2, is under development.

Mass storage is managed by Jasmine, which is a collection of user programs and server processes that interface with the JLab tape library and mass storage system. Every file written to tape resides at a designated location within a virtual filesystem. The structure of this filesystem is mirrored in a "stub" directory on centrally managed machines. One can examine the names of files stored on tape by inspecting this directory tree and can obtain basic information about them by examining the contents of the corresponding stub files. In particular, a stub file indicates the size of the actual file, its md5 checksum, creation time, owner, group, permission, and other bits of metadata.

For off-site data transport, JLab offers Globus along with other standard tools, ftp, scp, etc. Currently JLab hosts OSG submission nodes that allow users to submit jobs to resources available on OSG. JLab does not provide services that allow OSG jobs to run on our local clusters, but this may change in the future. As JLab increases both the size of the local ENP batch cluster and makes increased use of off-site OSG resources, JLab may trade some of local compute cycles over a long period of time for the ability to burst into a large off-site OSG resource for a short time.

SWIF2 has already been modified to allow users to run jobs at NERSC. As of this time, only GlueX is using this capability and it has yet to be generally released.

In the case of both OSG and NERSC, the software is containerized using Singularity and DOCKER.

GlueX use the CernVM distributed file system as part of its workflow. This is not a high-bandwidth application, but the availability of a distributed filesystem is important to the ability to run GlueX code off-site. CLAS12 will also adopt the use of CernVM-FS.

A goal for the near future is to integrate Globus, CernVM-FS, and OSG into SWIF2 so that users will be able to submit workflows at JLab and the system will run them at the most appropriate, or requested, site.

5.1.7.7 Network and Data Architecture

JLab's network is anchored by a pair of 100Gbit-capable core routers that act as a pure layer 3 transport area for all the site's various network collections ("network pods," illustrated in **Figure 10**) including internet connectivity. This network core is surrounded by firewalls, each of which serves as a security border for its network pod. Firewalls are bypassed for science traffic in certain cases, as we detail later.

In **Figure 10,** we show each network pod with redundant head-end routers and firewalls. In each case, firewall-router pairs are housed in separate physical locations, connected via multiple, redundant campus fiber links. For example, one of the scientific computing routers is in the data center, but the second is in another building that is used as a second network hub for resiliency with power and fiber path diversity. As a result, even in the case of a data center power outage, routing can be maintained for unaffected services. Similarly, JLab's two internet border routers are in physically diverse locations and make use of distinct entrance facilities for fiber to the campus as well as diverse power substations and generator backup.

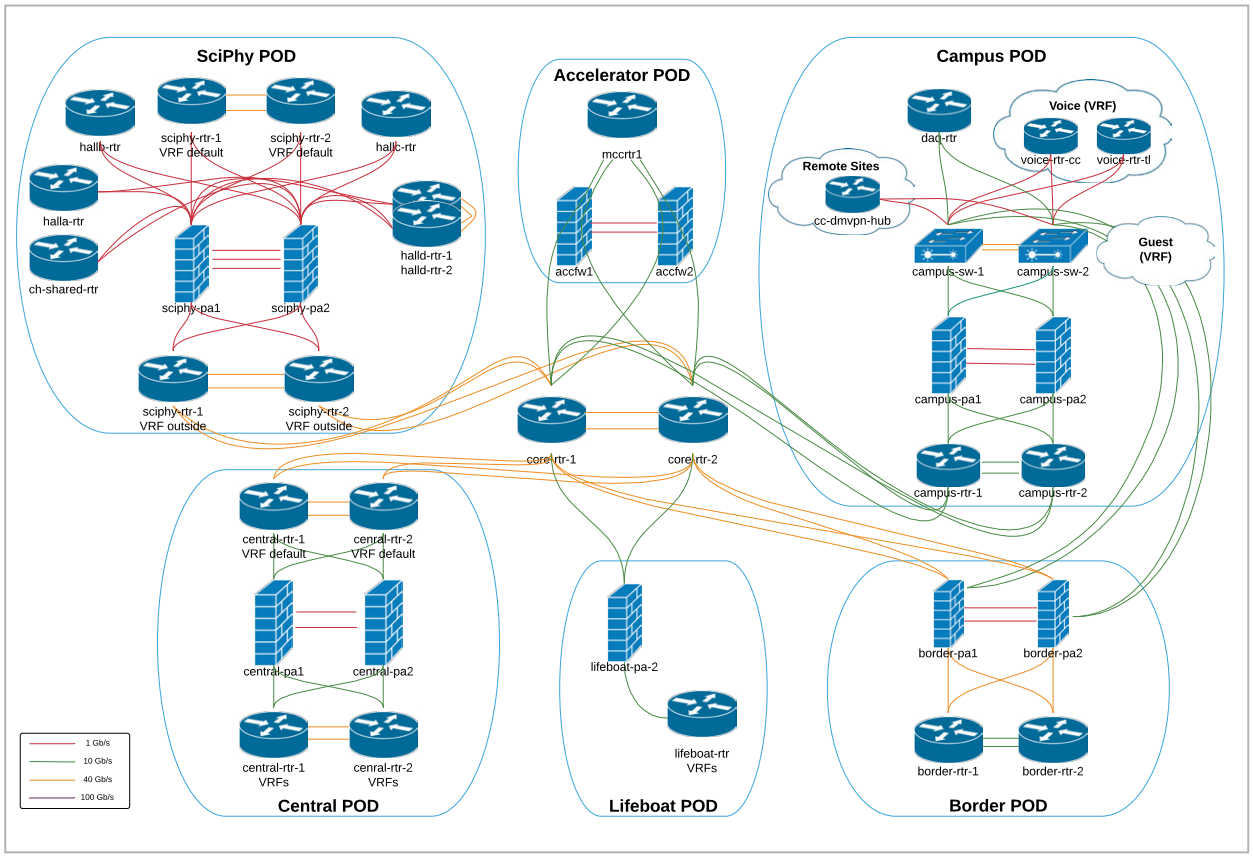

Figure 10. JLab layer 3 design.

Figure 10 shows a high-level overview of the lab's layer 3 design as it exists today. Although the core and border pod routers are 100Gbit-capable Arista units, the diagram reflects their configuration and effective connection performance as of April 2019. The top left pod, for Scientific Computing and Experimental Physics (SciPhy), is where networking for each of the four experimental halls is originated along with the batch farm, LQCD clusters, and mass storage. The border pod, bottom right, is where connectivity to ESnet is managed.

For large science flows, this core+pod design includes provisions to bypass the firewalls using PBR for well-known source/destination pairs. This firewall bypass is used in two cases:

1. To move data from the experimental halls to the tape library for storage.
2. To move data from the tape library and Lustre disk pool for off-site processing.

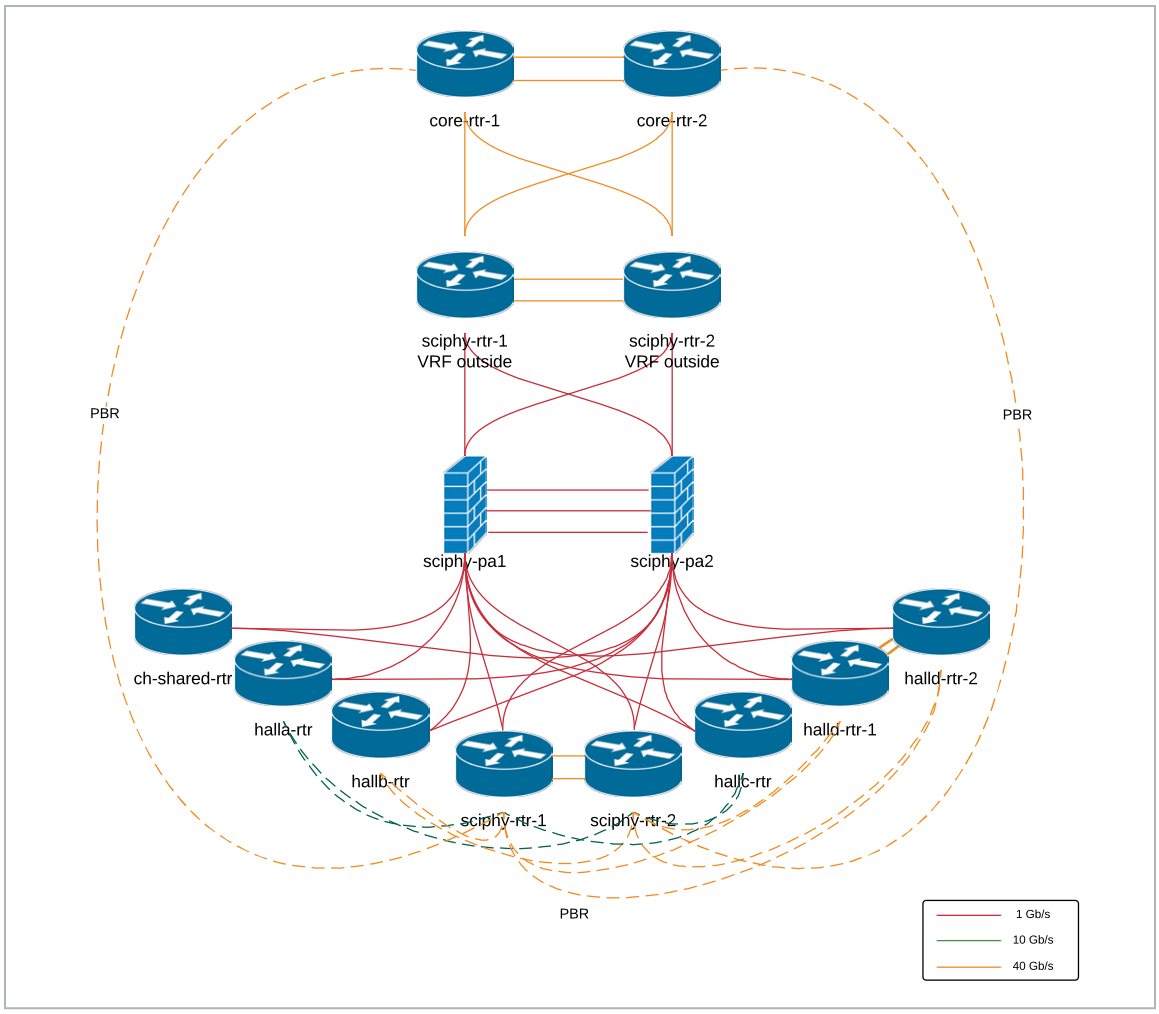

Figure 11. JLab Scientific Computing Network.

Both these cases are illustrated in **Figure 11** for the scientific computing network. The dotted lines represent firewall bypass for science data. The policy-based routes at the bottom represent the data flow from the experimental halls to the tape library; The policy-based routers from SciPhy (sciphy-rtr-1 and -2) to the core routers represent data from storage headed off-site. A similar PBR strategy is used from the core routers to the border routers to bypass the internet border firewalls for flows from DTNs.

Figure 12 shows the connectivity from the network core out to the first layer 3 ESnet hop at Atlanta and Washington, respectively. To reach this first layer 3 hop, traffic from JLab must first traverse ELITE, the Metro area ring, which is operated by Old Dominion University. ELITE Member institutions include Old Dominion University, the College of William and Mary, and the National Oceanic and Atmospheric Administration. As of this writing, ELITE is a 10Gbit/sec fully protected Dense Wavelength Division Multiplexing (DWDM) ring. Two physically diverse ELITE routers provide layer 2 exit from the ring to the wide area. There are two ELITE DWDM optical nodes located at JLab, collocated with the JLab border routers.

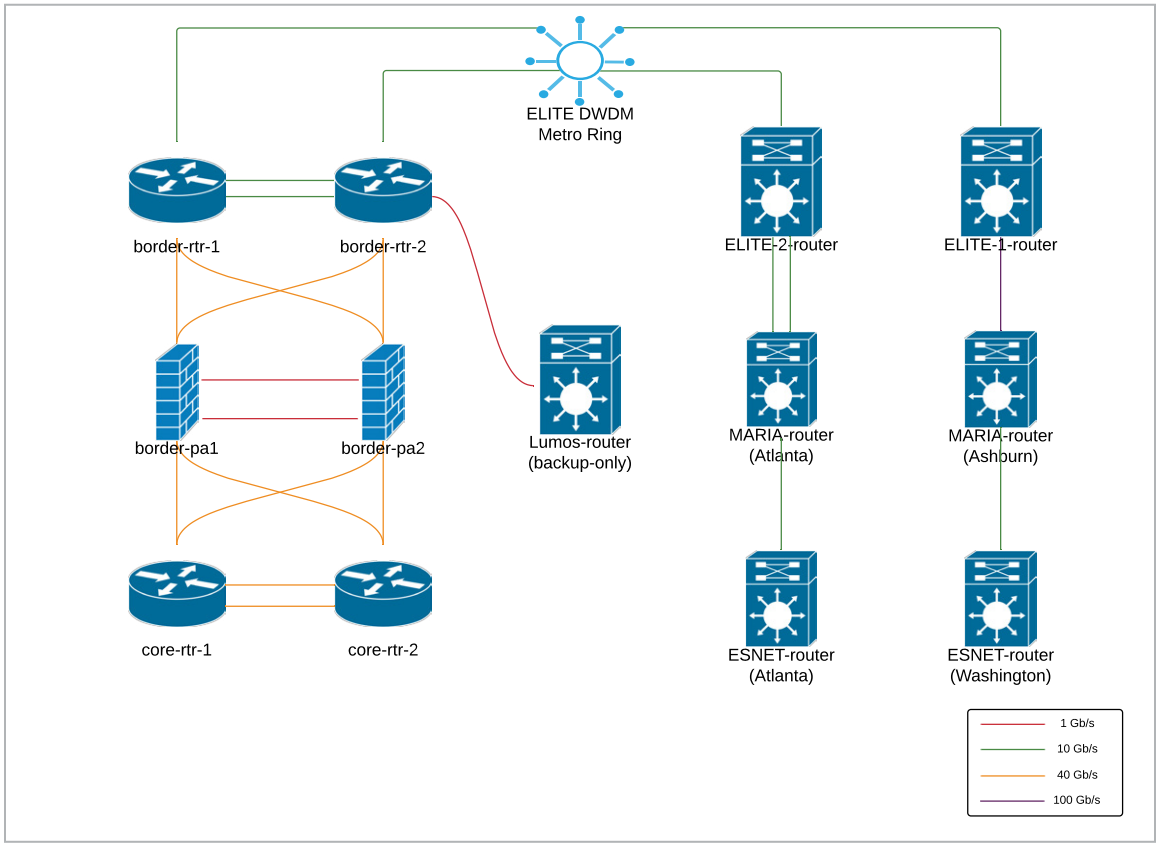

Figure 12. JLab path to ESnet.

The layer 2 path from ELITE to Atlanta and from ELITE to Ashburn is provided by MARIA, which is operated by Virginia Tech for the collaboration. Cross-connects from the two MARIA routers to ESnet routers provide the final step in JLab's connection to ESnet.

The path from the ELITE-1 router to Ashburn has recently been upgraded to a protected 100Gbit/sec circuit. The path from the ELITE-2 router to Atlanta has recently been upgraded to a 20Gbit/second port channel with two 10Gbit members. The bottleneck for JLab traffic is currently the ELITE DWDM ring itself. Although JLab has two optical nodes, they are on the same lambda. A second lambda has been ordered, which will bring the aggregate bandwidth to 20Gbit/sec this summer. The Lumos router is a 1Gbit/sec path of last resort.

Due to the complexity of the layer 2 path from JLab to both Atlanta and Washington, circuit outages have been relatively common. Significant effort has been put into automatic failover at every layer to avoid making these outages user-affecting. As a result of this design, JLab has achieved 100% uptime over the past year despite a host of fiber cuts and equipment problems off-site between JLab and ESnet.

Network upgrades will be funded incrementally over the next few fiscal years. The Networking and Scientific Computing groups have begun procurements to build out a redundant 100Gbit network from the data center to the internet border. Within the current year, JLab will deploy 100Gbit meshed routers and will be fully ready to meet ESnet at 100Gbit on any anticipated ESnet 6 timeline.

In addition to the necessary router and switch upgrades, a project to build out the DTN and storage capability is underway. Recent tests have demonstrated the limits of the existing DTN hardware and software, and design is underway to streamline DTN capabilities and usability. The aim of this project is to build out a robust set of DTNs for data movement to ASCR facilities, the OSG, and collaborator sites with compute capability for JLab experiments.

Network performance monitoring is done using two PerfSONAR nodes, one in the scientific computing network pod on the data-transfer network with the DTNs, and one on the network border outside the firewalls. Both these PerfSONAR nodes are slated for upgrade in the next budget cycle to move them beyond 10Gbit/sec.

5.1.7.8 Cloud Services

Until recently, JLab has made minimal use of cloud services. Email was provided by locally hosted servers and use of collaborative cloud storage was minimal. At the start of 2019, JLab moved to using Microsoft Office 365. This includes email, calendar service, and Microsoft OneDrive cloud storage. Although bandwidth to and from these services is much lower than that required for scientific data, JLab's reliance on cloud services for day-to-day business has made the uptime for at least this part of the WAN traffic of critical importance. **Figure 13,** ESnet Traffic associated with GlueX, in the next section is a screenshot of the ESnet top flows in early April 2019. Although the focus of the figure is the scientific network traffic, it also captures the flow to Microsoft which, although small, is not negligible.

JLab has investigated use of commercial cloud services for data processing, in particular for the case where a rapid turnaround is required that exceeds resources available locally and via the OSG and NERSC. An example use case is an experiment that discovers a problem in their reconstruction code or calibration database late in the reconstruction process and has to rapidly perform a computationally expensive task. Due to the current cost of cloud computing, we expect that use cases will be rare and when used, commercially provided cloud computing will replace one or more of the other resources, so the WAN requirements will not significantly increase.

5.1.7.9 Data-Related Resource Constraints

As discussed in Section 4, the computing requirements for CLAS12 and GlueX exceed the local computing resources. Our current plan is to use JLab internal compute resources as much as possible, while leveraging a variety of remote resources (OSG, NERSC, collaborator compute, commercial cloud services) where cost effective and appropriate. The large, 6 PB/yr, GlueX raw data set plus the 3 PB/yr from CLAS12 require a WAN capacity that exceeds our current 10GB/s link if a significant fraction of the associated compute is performed off-site. **Figure 13** is a screenshot of ESnet traffic from the start of April 2019.

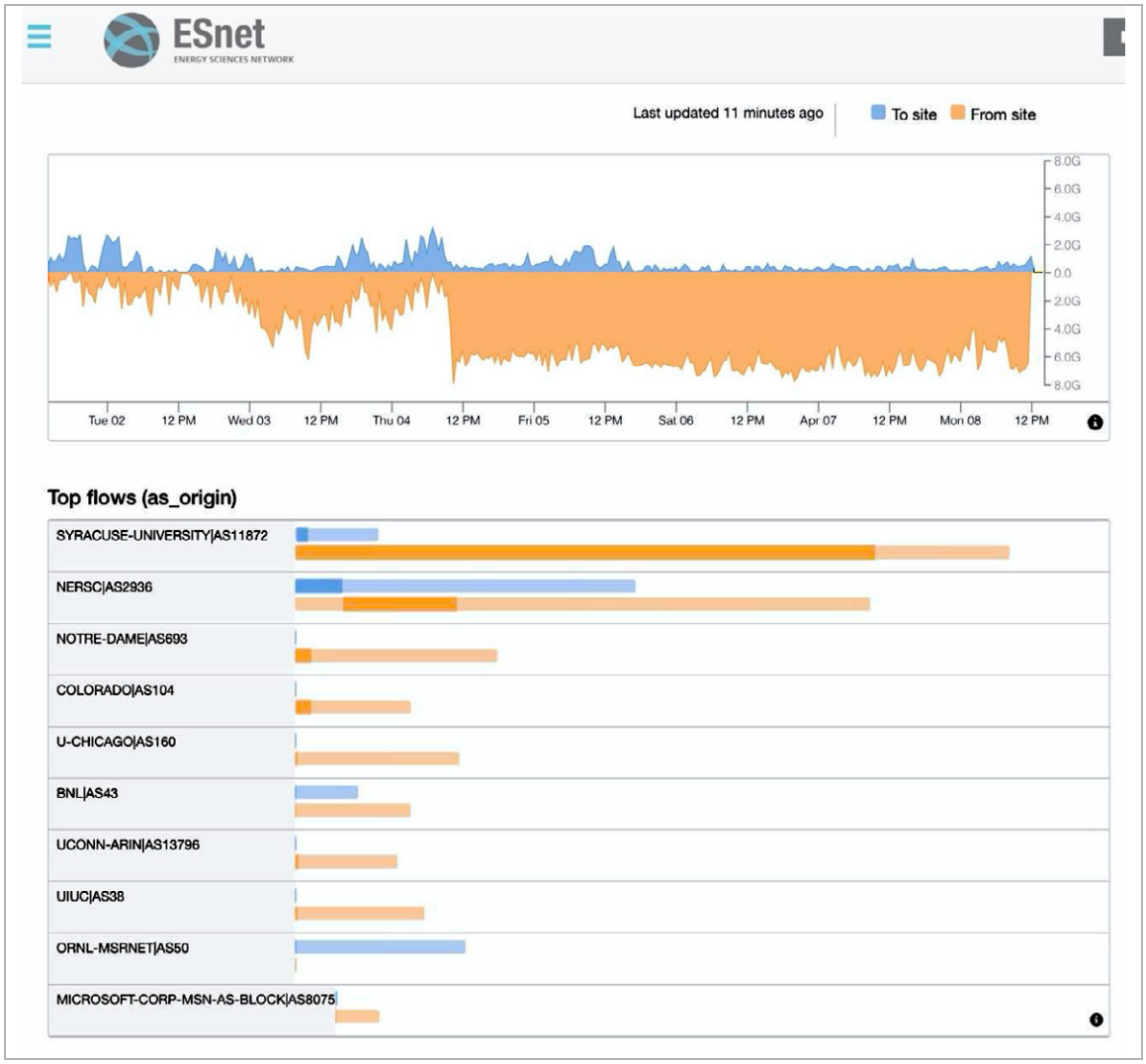

Figure 13. ESnet traffic associated with GlueX experiment.

The sharp increase in traffic at around noon on April 4 was a single GlueX user running at Syracuse University via the OSG. The second-highest user was GlueX running reconstruction at NERSC. The NERSC user was limited by JLab's inability to serve data from our local mass storage at a rate that can match the available NERSC resources, but the OSG user was limited by the network. These are the two main people in GlueX who are developing the capability to use the OSG and NERSC for reconstruction using JLab data. The NERSC and Syracuse GlueX users are using 2018 data sets, which are a factor of two smaller than those JLab expected GlueX to generate after fall 2019 activities with CLAS12 began.

5.1.7.10 Outstanding Issues

The outstanding issues have been touched upon earlier in this document.

Since JLab's local cluster and mass storage system were not designed with our current computing model in mind, two issues need to be resolved. First, the current design pulls raw data from tape onto a Lustre staging disk. From there, several gateway nodes feed data to remote sites via the network. In this design there are a relatively small number of high-bandwidth data sources, the tape data movers, and a small number of consumers, the gateways.

The system was designed to feed data to the local batch farm where a much larger number of data consumers are. JLab has started a project to resolve this problem using a high-bandwidth staging area based around solid-state drives. The second issue is that the system is not set up for users requesting off-site access to JLab resources. Users on the OSG cannot run jobs at JLab, and no easy mechanism exists for an off-site computing resource to request data from our systems.

5.1.7.11 Summary

JLab supports experimental NP and theory / LQCD programs. The experimental program uses equipment in four areas: halls A, B, C, and D.

In the five-year timescale, the two biggest contributors to the JLab computation workload and WAN bandwidth requirements will be GlueX in Hall D and CLAS12 in Hall B. These experiments will be in steady state data production over this period. In comparison, halls A and C are negligible since their contribution is smaller than the uncertainty in the requirements estimates for B and D.

Both GlueX and CLAS12 anticipate making use of off-site computing resources. The current 10Gbit/s WAN link restricts the rate at which data from experiments can move off-site. Although simulated events are smaller than raw events from the experiment, the statistics required may generate an incoming dataflow to JLab equal in size.

The first major upgrade will be the installation of MOLLER in Hall A. This will be followed by SoLID in the same hall. Current estimates are that, assuming approval, MOLLER will be installed in the fall of 2023 and be operational late 2024 to early 2025. SoLID installation in Hall A will follow sometime after MOLLER operation begins. It is not expected that either MOLLER or SoLID will have WAN requirements beyond what would be needed for GlueX on CLAS12.

The JLab site is transitioning to use of cloud services for aspects of day-to-day business. This requires that at least part of the WAN connection has high uptime and requisite quality of service. This would also provide a reliable connection over which, potentially, experiments could be controlled remotely.

5.1.7.12 Case Study Contributors

- **Graham Heyes,** *JLab, ENP,* heyes@jlab.org
- **Amber Boehnlein,** *JLab, CIO,* amber@jlab.org
- **Bryan Hess,** JLab, *IT Scientific Computing,* bhess@jlab.org
- **Brent Morris,** JLab, *IT Networking,* bmorris@jlab.org

5.1.8 Measurement of a Lepton-Lepton Electroweak Reaction (MOLLER)

5.1.8.1 Science Background

The MOLLER facility is a proposed dedicated detector that will be used to carry out the MOLLER experiment. The MOLLER experiment proposes to measure the parity-violating asymmetry in electron-electron (Møller) scattering. The measurement will be carried out at JLab's accelerator by rapidly flipping the longitudinal polarization of electrons that have been accelerated to 11 GeV and observing the resulting fractional difference in the probability of these electrons scattering off atomic electrons in a liquid-hydrogen target. This asymmetry is proportional to the weak charge of the electron, which in turn is a function of the electroweak mixing angle, a fundamental parameter of the electroweak theory. The accuracy of the proposed measurement allows for a low-energy determination of the mixing angle with precision on par with the two best measurements at electron-positron colliders.

5.1.8.2 Collaborators

MOLLER has 120 collaborators from 30 institutions (mostly universities).

5.1.8.3 Instruments and Facilities

The MOLLER facility, if funded by the DOE, will be built by JLab and an international collaboration of university physics laboratories. The effort involved in data taking and data analysis will be shared by JLab and the collaborating institutions and will rely heavily on JLab computing resources.

5.1.8.4 Process of Science

The MOLLER experiment will acquire about 4 PB of raw data over a two- to three-year period. To put this into context, GlueX acquires 6 PB of raw data per year. Raw data analysis will be carried out at JLab. An intermediate step in the analysis will generate about 20PB of data. These intermediate data sets will again be analyzed with compute resources at JLab. The analysis of large sets of data will typically be orchestrated by collaborators off-site. Over the course of the experiment and analysis (four-plus years), about 1PB of data will be transported off-site. These data would consist of small subsets of the raw data and data that have been highly reduced in size. Simulation work is and will be carried out at collaborator institutions. Data transfer will be minimal as simulation results are typically analyzed and interpreted at the same institution in which the simulated data is generated.

5.1.8.5 Remote Science Activities

At the current time, there are no known remote science activities.

5.1.8.6 Software Infrastructure

Please see Section 5.1.7.6, "Software Infrastructure."

5.1.8.7 Network and Data Architecture

Please see Section 5.1.7.7, "Network and Data Architecture."

5.1.8.8 Cloud Services

At present, MOLLER is not making use of commercial cloud services.

5.1.8.9 Data-Related Resource Constraints

Please see Section 5.1.7.9, "Data-Related Resource Constraints."

5.1.8.10 Outstanding Issues

Please see Section 5.1.7.10, "Outstanding Issues."

5.1.8.11 Case Study Contributors

- **Markus Diefenthaler,** *JLab, ENP,* mdiefent@jlab.org
- **Graham Heyes,** *JLab, ENP,* heyes@jlab.org
- **Stephen Wood,** *JLab, ENP,* saw@jlab.org

5.1.9 Solenoidal Large Intensity Device (SoLID)

5.1.9.1 Science Background

SoLID is a proposed large acceptance detector intended for use by at least five experiments covering a variety of physics topics. SoLID is a flexible detector that can be physically configured to be optimal for various types of physics programs.

In one configuration, SoLID will be used for SIDIS on polarized targets to study the transverse momentum structure (of quarks) in the proton and neutron. The large acceptance of the device will allow finely binned data over the multidimensional parameter space spanned by the SIDIS process.

In a second configuration, the detector will be arranged to carry out PVDIS. This makes the SoLID facility well matched to the JLab CEBAF accelerator which can produce excellent parity quality beam. The PVDIS experiment is able to search for interactions beyond the Standard Model and complements the MOLLER experiment.

5.1.9.2 Collaborators

SoLID has 300 collaborators from 72 institutions.

5.1.9.3 Instruments and Facilities

The SoLID detector, if funded, will be built by JLab and an international collaboration of university physics laboratories. The effort involved in simulation, data taking, and data analysis will be shared by JLab and the collaborating institutions but will rely heavily on JLab computing resources. The detector will be installed in Hall A. The timing of installation and operation is contingent not only on approval of funding but also on the scientific program in Hall A. Currently the MOLLER detector is scheduled to be installed in Hall A before SoLID. Although MOLLER is also currently unfunded, the tentative schedule for MOLLER is installation starting late 2023 and lasting 15 months. In that case it is likely that installation of SoLID cannot begin before 2025.

5.1.9.4 Process of Science

In the SIDIS J/Psi mode, the SoLID detector will produce data at a rate of 3–4GB/sec. Over the course of three years, 100 PB in total of raw data from the detector will be taken in this mode. In PVDIS mode the raw data rate off the detector will be 6GB/s, which would add 175 PB of raw data over the same three-year period. In the main JLab case study, it was noted that our tape library has a capacity of 27 PB. The data set sizes outlined would appear to sum to a requirement 10 times our library capacity. These numbers should be treated with caution because they do not consider reductions in data set size due to filtering, online processing before storage, compression algorithms, etc. Also, these rates assume a traditional triggered DAQ system. A streaming DAQ is being considered as a possibility for SoLID. A streaming DAQ architecture would result in much higher data rates coming from the detector. However, if a streaming DAQ architecture is used, additional compute resources would be employed to reduce the size of the detector data sets, and the net impact upon JLab WAN requirements should not change for SoLID experiments. The choice of streaming or traditional DAQ should not change the WAN requirements of experiments using SoLID.

It is assumed that most of the data processing will follow the model, outlined in the main JLab case study, of a mix of local and remote computing resources. The mix of off-site and on-site and the impact on networking cannot be accurately estimated at this stage since the data processing model and software have yet to be developed. It is likely that highly reduced and pre-processed data sets would be exported off-site. JLab staff estimate that 25 PB (a rough estimate) of raw/reduced data would be exported off-site over the next six years of experimental activity.

5.1.9.5 Remote Science Activities

At the current time, there are no known remote science activities.

5.1.9.6 Software Infrastructure

Please see Section 5.1.7.6, "Software Infrastructure."

5.1.9.7 Network and Data Architecture

Please see Section 5.1.7.7, "Network and Data Architecture."

5.1.9.8 Cloud Services

At present, SoLID is not making use of commercial cloud services.

5.1.9.9 Data-Related Resource Constraints

Please see Section 5.1.7.9, "Data-Related Resource Constraints."

5.1.9.10 Outstanding Issues

Please see Section 5.1.7.10, "Outstanding Issues."

5.1.9.11 Case Study Contributors

- **Markus Diefenthaler,** *JLab, ENP,* mdiefent@jlab.org
- **Graham Heyes,** *JLab, ENP,* heyes@jlab.org
- **Stephen Wood,** *JLab, ENP,* saw@jlab.org

5.1.10 Lattice Quantum Chromodynamics / Theory

5.1.10.1 Science Background

The structure of the proton and neutron, and the forces between them, originates from an underlying quantum field theory known as quantum chromodynamics (QCD). This theory governs the interactions of quarks and gluons that are basic constituents of the observable matter in our surrounding environment. QCD has been thoroughly tested by experiments at high energies, giving us insight into nature's workings over distances that are smaller than the size of nucleons (the term used for both protons and neutrons). However, at low energies or larger distances, the theory becomes formidable and efforts to theoretically determine fundamental NP phenomena directly from QCD have been met with less success. A long-standing effort of the DOE's NP program is to determine how QCD in this low-energy regime manifests itself into the observed spectrum of hadrons and the observed nuclear phenomena, and to use QCD to make reliable predictions for processes that cannot be experimentally accessed. These theoretical efforts provide critical support to the DOE's nuclear experimental projects, in particular, those being executed at JLab, Brookhaven's RHIC, and Michigan State's FRIB, as well as the planned EIC.

The Theory Group at JLab is pursuing several methods to discern this structure. One of the large efforts is using LQCD to make predictions that guide as well as confront experiments. These calculations use leadership-class computing facilities at DOE and NSF centers, as well as local computing facilities. As there is usually no long-term retention of data at these facilities, data are transported back to the local computing facilities for long-term storage.

Calculations at JLab, called global analyses, use experimental data to constrain the parameterizations of observables. These calculations use local computing facilities and may also use off-site resources.

5.1.10.2 Collaborators

The Theory Group at JLab is composed of JLab staff members and joint JLab and neighboring universities' staff members. In addition, there are bridge faculty members with universities. This team is pursuing a multifaceted campaign to understand the origin of matter.

JLab is a member of the USQCD Collaboration, a consortium of about 160 people at roughly 50 institutions, including universities and labs. These researchers are members of LQCD groups that typically span a few institutions. The collaboration hosts local computing facilities at JLab, FNAL, and BNL. Calculations are coordinated. DOE and NSF leadership computing facilities at ORNL, ANL, and NERSC are used for the parts of the calculations. Results from those calculations are brought back to the labs for further analysis on local computing facilities supported by FNAL, BNL, and JLab.

Members of the LQCD project teams apply for time on all of these computing facilities. Some data may be kept for an extended period on the leadership systems, such as NERSC. In general, though, the member labs provide long-term storage. JLab will host the long-term storage of data for LQCD projects related to the JLab science program.

The global analysis effort involves experimentalists and theorists at about 10 universities and labs across the country. The computational requirements for these analyses are lower than the LQCD projects.

User/ collaborator and location	Is a primary or second-ary copy of the data stored?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of data set? (re-port in bytes, e.g., 125GB)	Frequency of data transfer or download? (e.g., ad-hoc, daily, weekly, monthly)	Are data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g., difficult tools, slow network)?
R. Briceno J. Dudek R. Edwards B. Joo K. Orginos D. Richards F. Winter JLab	Primary	Lustre file sys-tems and tape storage at JLab. Access via tape and Globus.	Multiple differ-ent data sets of 100TB. Total storage 3PB.	Ad-hoc	Yes. Data sent to OLCF and NERSC.	Globus is sufficient
Other USQCD members, including: BNL, Michigan State, Kentucky, U. Wash., U. Conn.	Primary	Lustre file sys-tems and tape storage at JLab. Access via tape and Globus.	Multiple data sets of 10TB each. Total storage 2 PB total.	Ad-hoc	Yes. Data sent to universities	

Table 3. Lattice Quantum Chromodynamics / Theory Data Summary.

5.1.10.3 Instruments and Facilities

The HPC / LQCD Computing Systems at JLab include a 440 Xeon Phi (Knights Landing) plus Omnipath cluster, a 276 Xeon 16 cores plus InfiniBand cluster, a 45-node quad NVIDIA K20m Kepler GPU plus InfiniBand cluster, and a new 32 eight-way RTX-2080 (256 GPUs) cluster. A 3 PB (petabyte) Lustre parallel distributed file system storage system is shared between Experimental Physics and LQCD programs. The IBM TS3500 Tape Library is shared with the Experimental Physics and LQCD programs. Within a two-year horizon, it is expected that the Lustre file system will undergo upgrades and increase in size. The LQCD GPU system will double in computational capability.

LQCD projects will vary according to allocations on the JLab system. Flagship projects, such as the hadron spectroscopy and hadron structure programs, rely heavily on the JLab systems. These projects use the OLCF and NERSC systems to generate initial data sets. These sets are typically about 10 TB in size spread over about 1000 files. They are used for secondary calculations also at OLCF and NERSC. The size of these data sets is about 100 TB spread over about (500K) files. All data sets are transferred to JLab for further analysis over the period of a yearly allocation. In total, a few hundred TBs per year are transferred from off-site locations to JLab.

These LCF-generated data sets are often combined at JLab with other data sets. This results in an additional 200 TB/year of data. Such combined data sets do not generally leave JLab. These additional data sets generally do not leave the lab.

Global analysis data sets are a few GB. These might be transferred off-site, but the network demands are low.

New computing facilities are appearing. Summit is now in operation at OLCF, and Perlmutter at NERSC should appear in 2020. With these new capabilities, it is expected that the amount of data generated at the LCFs will increase about 5–10 times. The same workflows will be used, thus the amount of data to transfer back to JLab will increase by the same amounts.

In the two- to five-year time frame, JLab will see the appearance of the exascale Aurora system at the ALCF in late 2021. Members of the USQCD collaboration are participating in the ECP, and members are working directly with Intel and ALCF to deploy efficient codes on Aurora. The ECP application projects have benchmarks that are used to mark the progress of the development efforts. JLab anticipates that ECP applications will produce 10 times more data than Summit and Perlmutter. Thus, for LQCD projects, JLab expects to have about 1PB/year generated at the OLCF moved to JLab, most of which will be consumed at JLab.

On the scale of five years and beyond, the Frontier system at OLCF should begin deployment in 2023 and operations. Frontier is expected to provide two to three times the compute capability of Aurora, accordingly data generated and transferred to JLab would scale accordingly.

5.1.10.4 Process of Science

The workflow for JLab LQCD computing was outlined previously. There are typically about 30 projects allocated to the JLab LQCD facility, and in total 1PB disk storage is allocated to the projects with 5.5PB of tape used. These projects produce a few 10s of TB to as much as 100 TB of data over a yearly allocation, and some fraction of this is transferred off-site.

When the deployment of a large system occurs, data generation increases more rapidly than the typical JLab facility infrastructure refresh cycle. Planned upgrades to JLab facility infrastructure over the next five-year cycle include the Intel Knights Landing compute resource and a planned JLab GPU system. These upgrades will support increased data production rates and reduce current instrument bottlenecks.

The USQCD collaboration has records of the number of effective flops for LQCD applications over the last two to three years. It has been found that the number of flops delivered by the LQCD facilities projects is comparable to the number of flops acquired on the LCFs. This trend is expected to continue till the exascale systems deploy. LQCD applications receive something like about 15% of the available cycles on LCFs nation-wide, and it is this fraction that is comparable to USQCD resources. However, the investments of DOE into the exascale systems has accelerated compared to historical levels. It is expected that the exascale nodes will become available as commercial systems, much as single node versions of Summit are available now. Thus, with incremental scaling of local computing facilities, parity with the fraction of exascale computing resources available for LQCD can be expected in the greater than five-year time frame.

In the two- to five-year time frame, it is expected that data transfers into JLab LQCD will scale with these new LCF systems. JLab will be used for long-term storage. As the local computing facilities will not scale as quickly as new LCF deployments, some amount of the analysis work will be carried out at the LCFs. However, the final stage of the analysis workflow is more optimally carried out on JLab local systems. Thus, the imbalance of LCFs to local computing capability is offset.

5.1.10.5 Remote Science Activities

At the current time, there are no known remote science activities.

5.1.10.6 Software Infrastructure

The LQCD projects rely on the tape management utilities provided by JLab. Offset transfers use Globus, with a data-transfer system maintained by the lab.

5.1.10.7 Network and Data Architecture

Please see Section 5.1.7.7, "Network and Data Architecture."

5.1.10.8 Cloud Services

At present, there is no use of commercial cloud services by this project.

5.1.10.9 Data-Related Resource Constraints

Please see Section 5.1.7.9, "Data-Related Resource Constraints."

5.1.10.10 Outstanding Issues

Please see Section 5.1.7.10, "Outstanding Issues."

5.1.10.11 Case Study Contributors

- **Robert Edwards,** *JLab*, edwards@jlab.org

5.1.11 Electron-Ion Collider (EIC)

5.1.11.1 Science Background

The NP community proposes a U.S.-based EIC with high energy and high luminosity, capable of a versatile range of beam energies, polarizations, and ion species to precisely image the quarks and gluons and their interactions and to explore the new QCD frontier of strong color fields in nuclei. The Nuclear Science Advisory Committee recommended to DOE and the NSF in 2015 that the EIC receive highest priority for new facility construction. Subsequently, a National Academy of Sciences panel was charged to review both the scientific opportunities enabled by a U.S.-based EIC and the benefits to other fields of science and society. The National Academy of Sciences report strongly articulates the merit of an EIC:

> ***"In summary, the committee finds a compelling scientific case for such a facility. The science questions [How does the mass of the nucleon arise? How does the spin of the nucleon arise? What are the emergent properties of dense systems of gluons?] that an EIC will answer are central to completing an understanding of atoms as well as being integral to the agenda of NP today. In addition, the development of an EIC would advance accelerator science and technology in nuclear science; it would as well benefit other fields of accelerator-based science and society, from medicine through materials science to elementary particle physics."***

JLab hopes that this positive report will help drive DOE Office of Science funding in future, and the DOE Office of NP is already supporting increased efforts, setting in motion the process towards formal project R&D, engineering and design, and construction. The DOE Office of NP is already supporting increased efforts towards the most critical generic EIC-related accelerator research and design.

5.1.11.2 Collaborators

A multi-national community of physicists supporting the EIC concept has already formed an EIC User Group to support theoretical design of an accelerator and detectors. To date, this EIC User Group consists of 864 members, from 184 laboratories in 30 countries.

5.1.11.3 Instruments and Facilities

The EIC will enable us to embark on a precision study of the nucleon and the nucleus at the scale of sea quarks and gluons, over all of the kinematic ranges of interest. This requires an accelerator facility with a high luminosity and a versatile range of beam energies (center of mass range ~20 to ~100 GeV, upgradable to ~140 GeV), beam polarizations (longitudinal, transverse, tensor polarization of at least 70%), and beam species (proton as well as ion beam from D to heaviest stable nuclear). The EIC will have one or more interaction regions with an integrated detector with high acceptance.

If and when built, the EIC will be a facility in the era of exascale computing. This will affect the interplay of experiment, simulations, and theory profoundly and result in a new computing paradigm that can be applied to other fields of science and industry. The community envisions a Petascale-capable system at the beamline and a computing model with artificial intelligence at the trigger level and an unprecedented compute-detector integration to deliver analysis-ready data from the DAQ system. A similar approach would allow accelerator operations to use real-time simulations and artificial intelligence over operational parameters to tune the machine for performance.

Two pre-conceptual designs are being proposed for the EIC, eRHIC at BNL and JLEIC at JLab. Pre-conceptual design reports have been prepared using lab resources for accelerator and detector simulations. A similar report will be taken for future design work for the EIC. In the next several years, the EIC User Group will play a key role in further defining the science program at the EIC and will take advantage of existing computing resources at laboratories and universities.

5.1.11.4 Process of Science

Critical Decision-0 or Mission Need approval for the EIC by the DOE Office of Science is anticipated to be in FY 2019. The work on the conceptual design report in FY 2020 will result in an increase of accelerator and detector simulations. Part of the Critical Decision-1 process, sometime in the next two to five years, will be a site section and an increase of the already ongoing collaboration between BNL and JLab. Present R&D on the EIC includes simulations and software tests locally and at supercomputer centers. These activities will not significantly increase the WAN requirements on this timescale, but it is important that the other activities at JLab have enough bandwidth available to them that they do not affect the EIC R&D. Beyond five years the landscape is very much dependent on both the choice of site for the EIC and the results of the studies taking place in the next five years. In any case the WAN bandwidth between BNL and JLab will need to be greater than it is now, but that is very much dependent on the JLab end which is currently driven by the 12 GeV CEBAF program. The true bandwidth requirements will depend upon the computing model. We anticipate this being a mix of local and remote resources, but networking requirements will depend strongly on what the data rate is off the detector. A further factor is if, and by how much, the data set can be reduced locally before remote resources come into play.

5.1.11.5 Remote Science Activities

At the current time, there are no known remote science activities.

5.1.11.6 Software Infrastructure

Please see Section 5.1.7.6, "Software Infrastructure."

5.1.11.7 Network and Data Architecture

Please see Section 5.1.7.7, "Network and Data Architecture."

5.1.11.8 Cloud Services

At present, the EIC is not expected to make use of commercial cloud services.

5.1.11.9 Data-Related Resource Constraints

Please see Section 5.1.7.9, "Data-Related Resource Constraints."

5.1.11.10 Outstanding Issues

Please see Section 5.1.7.10, "Outstanding Issues."

5.1.11.11 Case Study Contributors

- **Markus Diefenthaler,** *JLab, ENP,* mdiefent@jlab.org
- **Graham Heyes,** *JLab, ENP,* heyes@jlab.org

5.2 Facility for Rare Isotope Beams (FRIB)

FRIB is a dedicated facility for nuclear science on the campus of Michigan State University (MSU). The facility itself is under active development and is in a unique position to create specific and scalable technology solutions for the scientific needs that have been identified. Early design principles have involved working with approximately 19 scientific working groups to define and document experimental designs and operational needs. These include data volumes and expectations for operation. Due to the size and complexity of the facility, storage and computation are not plentiful, but can be used for local calibration needs during operation. Experiments that populate FRIB will leverage other national and regional resources for storage and computation.

5.2.1 Discussion Summary

- FRIB is located on the campus of MSU.
- Nineteen working groups at FRIB are looking into experimental design and operation.
- Gamma-Ray Energy Tracking In-beam Nuclear Array (GRETINA) (an early iteration of GRETA) runs part time at FRIB and the other part of the time at ANL.
- FRIB as a facility has a wide data range for experiments: 28GB for a week on the low end to a high of 30TB. The average per experiment is around 4TB and increasing.
- FRIB offers only temporary storage abilities. Custodial copies of data will be the responsibility of the experimental spokesperson.
- The common operational method for FRIB for small data volumes (less than a TB) is to produce data locally and have experimental teams travel home with portable storage (tape, disk). Data transfer over the network is becoming more common as data volumes grow.
- The usage pattern for FRIB is that groups do a "one and done" operational mode: come in, run, and leave. There are no do-over opportunities. Thus, reliable and verifiable data collection (and delivery when they leave) is critical.
- Globus is available, but usage is not common. The data are handed off to each experiment's spokesperson as a part of the operational process.
- FRIB can use information on Globus/performance testing.
- Use of the network for data movement is a challenge. In one example a recent experiment totaled nearly 250 TB when complete. Using Globus, it took nearly three months to transfer these data between collaborating institutions.
- FRIB has 2x10G to MSU and utilizes the MSU WAN. FRIB managers would like dedicated higher speed connectivity to the FRIB facility.
- FRIB shares some knowledge with the AGLT2 facility (also at MSU) and could leverage some of that capability to get connected to MiLR in Michigan to assist with WAN needs.
- FRIB is looking for assistance in designing a Science DMZ infrastructure and has taken some basic steps to start this process.
- FRIB's operational pattern does not rely heavily on WAN, but this is expected to change. Local campus HPC resources (MSU iCER) are available but limited in terms of compute and connectivity. FRIB, therefore, relies on external computing resources like the Ohio Supercomputer Center (OSC), ORNL, and NERSC. When used, local compute resources

are employed to support quick turnaround calibration and small sample compute jobs for actively running experiments. Simulation and reconstruction compute jobs are especially time consuming if run on local resources, further promoting use of external resources.

- National Superconducting Cyclotron Laboratory (NSCL) DAQ is the main data acquisition application for producing and handling data flow at FRIB/NSCL. In NSCLDAQ, the process that manages the data pipeline is the ReadoutGUI. The ReadoutGUI constructs the data pipeline and controls the run state of the system (whether the data sources are producing data or not). Source data are managed through the use of ring buffers. Data sinks, or consumers, access the data pipeline via the DAQ-net. Data sinks can include, e.g., storage, scalers, and online analysis.
- FRIB follows a requirements-based approach for personnel and environmental safety, and property protection and information security. FRIB maintains management systems registered by National Sanitation Foundation – International Strategic Relations (NSF-ISR) to the ISO 9001 (Quality Management), ISO 14001 (Environmental Management), ISO 45001 (Integrated Safety and Health Management), and ISO 27001 (Information Security Management) standards.

5.2.2 Science Background

Goals of the science: FRIB was designed, and is being constructed to be, the world's most powerful rare isotope research facility. FRIB will enable researchers to make major advances in our understanding of nature by accessing key rare isotopes that previously only existed in the most violent conditions in the universe. FRIB capabilities will provide unprecedented opportunities to study the origin and stability of nuclear matter. It will be possible to carry out studies of a wide range of nuclei at the very limits of nuclear stability where specific aspects of the nuclear many-body problem can be explored. Specifically, the unique features of FRIB will allow the delineation of the proton or neutron limits of existence to higher masses than other facilities. It will double the number of neutron-rich nuclei that will lead to new information about matter with unusual features, such as halos, skins, and their new collective modes. The accelerator's high power will yield the highest intensity of different isotopes produced anywhere in the world, thereby allowing the possible r-process sites and the respective paths to be determined. It also will be the only place where measurements of most of the key nuclear reactions involved in explosive astrophysical environments can be made. FRIB will provide the U.S. community with a valuable source for production of rare isotopes that are crucial for the exploration of fundamental symmetries and that may benefit society.

Departments and/or laboratories involved: FRIB will be a DOE-SC scientific user facility serving users organized in the FRIB Users Organization (FRIBUO). The FRIBUO has been involved in the development of the science program at FRIB from the beginning. The organization ensured that the optimum facility is being built for enabling world-leading science on day one of operation. At the beginning of January 2020, the FRIBUO has 1,400 members, representing 119 U.S. colleges and universities, 13 national laboratories, and 53 countries. A listing of the membership is available at fribusers.org.

Stakeholders: DOE, FRIB Laboratory, FRIBUO, NSF, MSU, State of Michigan

Data Life Cycle: FRIB will adopt a data management plan similar to that in use at the NSCL, which is the current scientific user facility operating under financial assistance from the NSF. The NSCL user community has accepted and practices against this plan and will largely transition to become the FRIB user community.

The spokesperson (lead investigator on an approved experiment) is responsible for the sharing and preservation of research data and results from an experiment. Scientific personnel (all persons carrying out research at FRIB) record primary data in a rational and customary fashion and produce and maintain any computer files, records, logbook entries, etc. necessary to interpret the research data and validate the results. The spokesperson arranges

for research data storage and dissemination of results obtained from an experiment and ensures that a copy of the research data is provided to FRIB for storage. FRIB's Business Information Technology department facilitates the recording of research data during the running of the experiment, the transfer of the data to long-term storage media by the spokesperson or his/her designee, and will, as a courtesy, keep a duplicate of the recorded research data for a period of two years after completion of an experiment.

The spokesperson is knowledgeable of and complies with relevant data policies, including long-term storage of the research data and records of the data analysis, and responds to research data access requests. The spokesperson divides analysis tasks among the group of scientists working on that experiment and provides appropriate access to the research data as required by collaborators or by federal agency requirements, expeditiously analyzes the data, and subsequently publishes the results.

Other researchers (individuals who are funded to do research in nuclear science or a closely related field, or who are faculty at a university or college) requesting access to research data do so by contacting the spokesperson. Due to the complexity and variety of experiments, it would be extremely difficult for an individual not directly involved in an experiment to reliably independently reanalyze the data. Other researchers are encouraged to collaborate with knowledgeable scientific personnel.

Requests for access to material and analyses related to published works are forwarded to the spokesperson of the pertinent experiment, who is to ensure the availability of processed data to enable the validation of results. There is no requirement to share proprietary data. Privileged or confidential information shall be released only in a form that protects the privacy of individuals and subjects involved.

The spokesperson and his/her collaborators are expected to promptly prepare and submit for publication, with authorship that accurately reflects the contributions of those involved, all significant findings of an experiment. The spokesperson ensures the availability of processed data used to generate charts, figures, illustrations, etc. in published works. The spokesperson is encouraged to share research data, upon request from other researchers, unless the sharing would incur an unreasonable burden of cost or time or usurp the scientific results of an experiment. All scientific personnel should encourage and facilitate the sharing of research results.

5.2.3 Collaborators

The FRIBUO's members are interested in conducting scientific research at FRIB. Members of the FRIBUO have formed working groups specializing in specific instruments, facility locations, or scientific topics. Each working group is led by a set of conveners whose affiliations are taken to be representative of the geographical distribution of the community in the table in this section. Experimental research performed at the current NSCL and future FRIB within the context of a working group is informed by the relevant data management plan for the creation, sharing, and storage of data described in the previous section.

The working groups include:

- Astrophysics and SECAR: This group is an umbrella collaboration for various equipment and theory projects in nuclear astrophysics at FRIB. This includes work to construct a recoil Separator for Capture Reactions (SECAR) optimized for measurements of radiative capture reactions with low-energy FRIB radioactive beams.
- Nuclear Data: A group focused on coordinating the efforts of the nuclear data community and the science program foreseen for FRIB physics.
- Data Acquisition: A group established to examine the issues of data acquisition at FRIB and to envision what modern data acquisition would be like in the age of FRIB experiments, starting in the year 2020 and beyond. The scope of our activities covers data acquisition (readout, run control, time synchronization), data movement (buffer transfer, event building, buffer/event storage in files, serving files to experimenters) and data analysis (online and off-line, data display).

- Detectors for Equation of State (EoS) Physics: The main tasks of this group are to identify the resources required to probe the density dependence of the symmetry energy at FRIB.
- High-Resolution In-beam Gamma Spectroscopy: This group is an umbrella collaboration for projects aimed at supporting the advancement of NP through state-of-the-art gamma-ray detector technologies. Leading initiatives include development of the gamma-ray tracking array, GRETINA/GRETA, and support of synergistic activities with Gammasphere.
- HRS: High-Rigidity Spectrometer (HRS), a group to advance the implementation of an HRS for FRIB. The HRS will be the centerpiece experimental tool of the FRIB fast-beam program. Through precise exit-channel selection, the HRS will also increase the scientific discovery potential from other state-of-the-art and community-priority devices, such as the GRETA and the Modular Neutron Array (MoNA-LISA), in addition to other ancillary detectors.
- Ion Traps: This working group is focused on designing, constructing, and utilizing Penning and Paul ion traps for experiments at FRIB.
- ISLA: ReA12 Recoil Separator: The goal of this working group is to define the requirements and characteristics of a device that can filter out unreacted beam particles and separate and characterize the reaction residues of interest for experiments at ReA12, a superconducting linac designed to accelerate rare isotope beams, located at MSU.
- Isotopes and Applications: This working group has concentrated on promoting the various applications that utilize exotic isotopes and developing systems to harvest radioactive isotopes at FRIB.
- Laser Spectroscopy and Neutral Atom Traps: This working group is focused on designing and constructing new laser-based spectroscopy measurements at FRIB, and utilizing future and existing systems.
- Neutron Detection: This working group is focused on designing, constructing, and utilizing detectors such as He-3, plastic and liquid scintillators, and MoNA-LISA for neutron detection at FRIB.
- Radioactive Decay Station: This working group was formed to promote and facilitate the design and construction of experimental apparatus, which will take full advantage of the new and exciting opportunities provided by FRIB using decay spectroscopy. An efficient, state-of-the-art detection station(s) equipped with instruments capable of characterizing various forms of radiation such as gamma rays, conversion electrons, beta particles, protons, alpha particles and neutrons will be required for decay studies at FRIB.
- ReA Energy Upgrade: The NSCL/FRIB ReA energy upgrade working group is focused on defining and supporting the physics associated with the potential energy upgrades of the current ReA3 accelerator at NSCL to higher energies, with the ultimate goal of ReA12 for physics at FRIB.
- Scintillator Arrays: This working group is focused on designing, constructing, and utilizing scintillator arrays employing new high-resolution materials and perhaps optimized for high-energy gamma rays. Moving forward, the conveners welcome broad participation from the community as well as cross-fertilization between other working groups in formulating an agenda for a scintillator array for FRIB.
- Silicon Arrays Solenoid Detectors: This working group is focused on developing charged-particle detector arrays for specific purposes at FRIB, as well as significantly advancing the fundamental characteristics of silicon, investigating the uses of other materials, and working with others to advance data acquisition systems for large channel-count spectroscopy.

- Solenoid Detectors: This working group has focused on designing, constructing, and utilizing a solenoidal spectrometer system for measurement of reactions in inverse kinematics at FRIB. Their activities have culminated in the Solenoidal Spectrometer Apparatus for Reaction Studies, or SOLARIS.
- Target Laboratory: This working group is focused on setting up a target laboratory for the in-house fabrication of thin films, windows, special radioactive sources, and related items needed for experiments at FRIB.
- Time-Projection Chambers: This working group is focused on designing, constructing, and utilizing a time-projection chamber / active-target system at ReA3 / FRIB.

User/collaborator and location	Is a primary or secondary copy of the data stored?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of data set? (report in bytes, e.g., 125GB)	Frequency of data transfer or download? (e.g., ad-hoc, daily, weekly, monthly)	Are data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g., difficult tools, slow network)?
Astrophysics and SECAR (NSCL, MSU, ANL, ORNL, UT, MIT, NCSU)	Primary	Data transfer, tape, hard drive	*	Ad-hoc	No	
Nuclear Data (NNDC, BNL, ANL, TUNL)	Primary and secondary	Data transfer, tape, hard drive	*	Ad-hoc	No	
Data Acquisition (ORNL, ANL, LBNL, MSU, NSCL, URochester)	Primary	Data transfer, tape, hard drive	*	Ad-hoc	No	
Detectors for EoS Physics (MSU, Notre Dame, Ohio State, WU, TAMU)	Primary	Data transfer, tape, hard drive	*	Ad-hoc	No	
High-Resolution In-flight Gamma-Ray Spectroscopy (FSU, ANL, LBNL, NSCl, MSU, ORNL)	Primary	Data transfer, tape, hard drive	*	Ad-hoc	No	
HRS: High Rigidity Spectrometer (NSCL, MSU)	Primary	Data transfer, tape, hard drive	*	Ad-hoc	No	
Ion Traps (MSU, NSCL, UMich, ANL)	Primary	Data transfer, tape, hard drive	*	Ad-hoc	No	
ISLA:ReA12 Recoil Separator (NSCL, MSU, LBNL, ANL, UBucknell, ANL)	Primary	Data transfer, tape, hard drive	*	Ad-hoc	No	

User/collaborator and location	Is a primary or secondary copy of the data stored?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of data set? (report in bytes, e.g., 125GB)	Frequency of data transfer or download? (e.g., ad-hoc, daily, weekly, monthly)	Are data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g., difficult tools, slow network)?
Isotopes and Applications (LLNL, LANL, UAlabama, NSCL, MSU, HopeC)	Primary	Data transfer, tape, hard drive	*	Ad-hoc	No	
Laser Spectroscopy and Neutral Atom Traps (NSCL, ANL)	Primary	Data transfer, tape, hard drive	*	Ad-hoc	No	
Neutron Detection (ORNL, UTK, LSU, NSCL,MSU)	Primary	Data transfer, tape, hard drive	*	Ad-hoc	No	
Radioactive Decay Station (ORNL, UTK, NSCL, MSU, ANL, LLNL)	Primary	Data transfer, tape, hard drive	*	Ad-hoc	No	
NSCL/FRIB Reaccelerator (ReA) Energy Upgrade (MSU, NSCL, ANL, TAMU, UConn)	Primary	Data transfer, tape, hard drive	*	Ad-hoc	No	
Scintillator Arrays (WU, UMass, ANL)	Primary	Data transfer, tape, hard drive	*	Ad-hoc	No	
Silicon Arrays (ORNL, RutgersU, IndianaU, UTK, TAMU, ANL)	Primary	Data transfer, tape, hard drive	*	Ad-hoc	No	
Solenoid Detectors (ANL, UConn)	Primary	Data transfer, tape, hard drive	*	Ad-hoc	No	
Target Laboratory (ANL, Oregon, NSCL, MSU)	Primary	Data transfer, tape, hard drive	*	Ad-hoc	No	
Time-Projection Chambers (NSCL, MSU, Notre Dame, Ohio State, WU, TAMU)	Primary	Data transfer, tape, hard drive	*	Ad-hoc	No	

Table 4. FRIB Data Summary.

Table 4 abbreviations.

HopeC	Hope College	RutgersU	Rutgers University
IndianaU	Indiana University	TUNL	Triangle Universities Nuclear Laboratory
LLNL	Lawrence Livermore National Laboratory	UAlabama	University of Alabama
LSU	Louisiana State University	UConn	University of Connecticut
MIT	Massachusetts Institute of Technology	UMass	University of Massachusetts, Lowell
NCSU	North Carolina State University	UMich	University of Michigan
NNDC	National Nuclear Data Center	URochester	University of Rochester
Notre Dame	University of Notre Dame	UT	University of Tennessee, Knoxville
Ohio State	Ohio State University	WU	Washington University, St. Louis
Oregon	Oregon State University		

*Average data set sizes: For the experiments performed at NSCL during calendar year 2018 where primary data was stored on-site, data sizes ranged from 28 Gbytes to 29 Tbytes with an average size of 3.9 Tbytes.

5.2.4 Instruments and Facilities

Instrument Descriptions

The instruments at FRIB will include existing instruments currently used for science experiments at NSCL and new instruments that will be constructed after FRIB begins user operations in CY22.

Existing/Near Future Instruments (Present to Five Years)

The FRIBUO has 19 working groups for experimental instrumentation that develop plans for using existing devices as well as proposals for new equipment.

1. The A1900 fragment separator is a third-generation projectile fragment separator composed of 40 large diameter superconducting multipole magnets and four 45° dipoles with a maximum magnetic rigidity of 6 Tm. Its length is approximately 22 meters. The A1900 has a solid angle acceptance of 8 msr and a momentum acceptance of 5.5% and can accept over 90% of a large range of projectile fragments produced at the current NSCL. The A1900 is instrumented with position and timing detectors at the intermediate dispersive image and at the final focal plane. Energy-loss and total energy particle detectors, and, in some instances, photon detectors, are also located at the final focal plane. Although the A1900 is used mostly for transmitting separated isotopes to downstream experiments, it can also be used as a stand-alone experimental device or in conjunction with downstream devices for executing an experiment. Typical use-rate of existing instruments in this experimental program is 100%.
2. The S800 spectrograph is a superconducting high-resolution vertically bending magnetic spectrograph that resides in the S3 vault. The spectrograph has energy resolution E/dE=104; maximum rigidity of 4 Tm; momentum acceptance of 5%; and solid angle of 20 msr. The analysis beam line leading down to the target position can be used to dispersion-match the beam, or it can be operated as a second fragment separator with a momentum acceptance of 6%, maximum rigidity 4.9 Tm, momentum resolution of 2000, and solid angle of 6 msr. The focal-plane detector system includes tracking detectors, an ion chamber, plastic scintillators for timing and energy loss, and a 32-segment CsI(Na) hodoscope for particle identification. The target position of the S800 can accommodate several different detector arrangements. One arrangement includes a large but removable multipurpose scattering chamber for charged-particle spectroscopy. Another arrangement is available for gamma-ray spectroscopy, which can be supplemented by the triplex-

plunger device for lifetime measurements. Neutron spectroscopy can also be carried out around the S800 target position. A liquid-hydrogen target can also be installed at the S800 target position. Typical use-rate of existing instruments in this experimental program is 50%.

3. The Sweeper Magnet is a superconducting dipole magnet with a maximum field of 4 T. The bend radius is 1 m with a bend angle of 400. It has a vertical gap of 14 cm which allows for neutron coincidence experiments (with the neutron walls or MoNA-LISA) covering about ±70 mrad. The sweeper also has its own detection system, which can be used to determine the detailed properties of the fragments following the breakup — the charge, mass, angle, velocity, momentum, and energy. By combining this information with the corresponding information about the neutrons, it is possible to reconstruct the properties of the original neutron-rich exotic nucleus. Typical use-rate of existing instruments in this experimental program is 5%.
4. The Low-Energy Beam and Ion Trap experiment (LEBIT) facility consists of a beam transport and manipulation system and a Penning trap based on a 9.4T superconducting magnet. A novel Single-ion Penning trap has been constructed and connected to the LEBIT beam line. Typical use-rate of existing instruments in this experimental program is 20%.
5. The beam-cooler and laser spectroscopy (BECOLA) facility includes a secure laser room with two turn-key lasers (a solid-state Ti:Sapphire ring laser and dye ring laser) and a frequency doubler. The beta-NMR station is used with BECOLA beams to measure ground-state moments of nuclei where the spin polarization is produced in fast fragmentation reactions or via laser optical pumping. A positron polarimeter, which can be placed at the end of the BECOLA beam line, is used for tests of symmetries in beta decay. Typical use-rate of existing instruments in this experimental program is 30%.
6. The Active-Target Time-Projection Chamber (AT-TPC) consists of a 250-liter cylindrical volume filled with a target gas (depending on the goals of the experiment) in which the charged particles emitted when a nuclear reaction takes place are traced in three dimensions. For experiments conducted with low-energy beams from the ReA3 linac, the detector is placed inside a large bore solenoid that can apply a magnetic field up to 2 Tesla parallel to the beam direction. The AT-TPC can also be placed elsewhere in the laboratory, such as in front of the S800 spectrograph, to conduct experiments at higher energies. When coupled with the S800, the sensor plane of the AT-TPC has a hole in its center so that the high-energy beam recoils can escape the gas volume and be collected and analyzed by the S800. Typical use-rate of existing instruments in this experimental program is 10%.
7. The Joint Array for NUclear Structure (JANUS) is composed of two annular double-sided segmented silicon detectors surrounded by the Segmented Germanium Array (SeGA). The silicon detectors are placed so that particles scattering to large angles can be detected with high efficiency. With a solid angle coverage of 29% of 4 pi, and an effective solid angle coverage after projectile reconstruction from the target recoil of 78% of 4 pi, the JANUS system is well-suited for low-energy Coulomb excitation using beams from ReA. Typical use-rate of existing instruments in this experimental program is 5%.
8. The Coincident Fission Fragment Detector consists of four large-area (30 cm x 40 cm) parallel-plate avalanche counters (PPACs), two position-sensitive timing micro-channel plate detectors, and two silicon monitor detectors. The Coincident Fission Fragment Detector is used to measure fusion-fission and quasifission reactions at ReA. From the time-of-flight and position measurements of the PPACs, the velocity of the binary fission-like fragments can be reconstructed, and the mass ratio of the fragments can be deduced. Typical use-rate of existing instruments in this experimental program is 2%.

9. SECAR is a recoil separator device specifically designed for inverse-kinematics experiments with light targets. The device was designed with four sections. The first section captures the particles exiting the target and selects a single charge state. The second section uses a crossed-field device, a velocity filter, to pass particles with the recoil velocity along the axis, effectively rejecting unreacted projectiles and scattered projectiles. The velocity filter is used in combination with a dipole magnet to form a mass focus. The third section has a second combination of dipole magnet and velocity filter, which further enhances the rejection of projectiles. The fourth section consists of a dipole magnet and a drift section to give a final rejection of any scattered beam particles that have made it to this point in the device. At the final focus, a variety of detectors are employed to identify and count the particles transmitted by the separator, including a final discrimination against projectiles that have been scattered into the detector. Typical use-rate of existing instruments in this experimental program is 30%.
10. The SeGA consists of 18 segmented germanium detectors with associated electronics and cryogenic support. SeGA is used in conjunction with the S800 spectrograph for inelastic scattering, charge-exchange, and nucleon-knockout experiments, with the triplex-plunger device for level lifetime measurements, as well as with devices employed for online radioactive decay studies. Typical use-rate of existing instruments in this experimental program is 25%.
11. The Beta Counting system (BCS) relies on implanting the fast ions into segmented silicon or germanium detectors. Fragment implantations are correlated in time and position with subsequent decays on an event-by-event basis, allowing the identification of the species observed to decay and a direct measurement of the decay time. The BCS system has also been used in conjunction with SeGA and the CloverShare Compton-suppressed HPGe array. The BCS is outfitted with a fully digital data acquisition system. Typical use-rate of existing instruments in this experimental program is 15%.
12. The Modular Neutron Array (MoNA) and its companion neutron array (LISA) consist of a total of 288 bars of plastic scintillator. Each of these bars measures 10 cm by 10 cm and 2 m wide. The bars are typically stacked to form two walls that are each 2 m wide and 1.6 m high, but due to its modularity, the array can be configured in other ways as well. The detection efficiency for neutrons with energies up to 100 MeV is about 70%. The position of the light emission along the bar can be determined within a few centimeters. Typical use-rate of existing instruments in this experimental program is 10%.
13. Two neutron time-of-flight walls (2m x 2m, position sensitive in two dimensions, and liquid-scintillator filled) have been used in conjunction with a removable 53" thin-walled reaction chamber to study proton/neutron emission ratios from intermediate-energy reactions. Typical use-rate of existing instruments in this experimental program is 5%.
14. The Neutron Emission Ratio Observer (NERO) is a low-energy neutron detector composed of three concentric rings of 3He and BF3 proportional counters embedded in a polyethylene matrix. NERO detects neutrons ranging in energy from 1 keV to 5 MeV with an efficiency of approximately 30% to 40%. Typical use-rate of existing instruments in this experimental program is 5%.
15. The Low-Energy Neutron Detector Array (LENDA) is comprised of 24 plastic scintillator bars that can detect neutrons down to energies of 150 keV. LENDA is used mainly to detect neutrons from the (p,n) charge-exchange reaction in inverse kinematics. Typical use-rate of existing instruments in this experimental program is 5%.

16. The Cesium Iodide Array (CAESAR) is a high-efficiency photon counting system that contains 192 CsI(Na) scintillator detectors with a photopeak efficiency of 35% at 1 MeV. CAESAR has been used at the S800 target position for particle-gamma coincidence measurements in front of the Sweeper Magnet for three-fold particle-neutron-gamma detection, and behind the Radio-Frequency Fragment Separator for spectroscopy of proton-rich beams. Typical use-rate of existing instruments in this experimental program is 10%.
17. The Summing NaI(Tl) (SuN detector) [10%] is a total absorption spectrometer that is used for a variety of decay studies and is crucial for a new technique for predicting (neutron, gamma) reaction rates. Typical use-rate of existing instruments in this experimental program is 10%.
18. The High-Resolution Charged-Particle Array (HiRA) is an array of 20 segmented Si-Si-CsI(Tl) telescopes providing an angular resolution of 0.15° at the nominal distance of 35 cm from the target. At this distance, the telescopes cover 70% of the solid angle between scattering angles of 5° and 30°. The telescopes are designed such that they can be independently placed, which allows optimizing the geometry for each experiment. Typical use-rate of existing instruments in this experimental program is 5%.
19. The Proton Detector is a gas volume detector used in beta-delayed proton experiments. In its current iteration, the detector operates in a calorimetric mode with 13 pick-up pads, and the signals are processed using digital electronics. In the near future, the Proton Detector will be equipped with approximately 2,000 pads, and support read-out of data using high-density GET electronics. This will enable the Proton Detector to operate as a time-projection chamber (TPC), capable of distinguishing multi-particle emission events. The Proton Detector is typically surrounded by SeGA in its barrel configuration for the simultaneous detection of gamma rays. Typical use-rate of existing instruments in this experimental program is 5%.
20. GRETINA is a national resource that moves from laboratory to laboratory. A collaboration of scientists from LBNL, ANL, NSCL, ORNL, and Washington University has designed and constructed a new type of gamma-ray detector to study the structure and properties of atomic nuclei. GRETINA consists of 28 highly segmented coaxial germanium crystals. Each crystal is segmented into 36 electrically isolated elements and four crystals are combined in a single cryostat to form a quad-crystal module. The modules are designed to fit a close-packed spherical geometry that will cover approximately one-quarter of a sphere. GRETINA is the first stage of the full GRETA. Typical use-rate of existing instruments in this experimental program is 25%.
21. Gammasphere consists of up to 110 Compton-Suppressed Ge detectors. It was built by a collaboration of physicists from LBNL, ANL, ORNL, and a number of U.S. universities. The device offers excellent gamma-ray energy resolution (2.3 keV at 1 MeV) and a photopeak efficiency of ~10% at 1 MeV. Gammasphere uses the same digitization modules (14 bit, 100 MHz) as the GRETINA detector, and is able to process singles rates up to 500k/s and triple-gamma-coincidence rates up to 120k/s. Typical use-rate of existing instruments in this experimental program is 0%.
22. SuperCHICO is a 4π position-sensitive parallel-plate avalanche counter. This instrument serves as a heavy ion recoil detector and has an angular resolution of $1^{o} \times 1^{o}$ for θ and ϕ, respectively. The instrument is used in conjunction with high-resolution gamma-ray arrays (GRETINA, GRETA, etc.) for the kinematic reconstruction of transfer, Coulomb excitation, and fission reactions. Typical use-rate of existing instruments in this experimental program is 2%.

23. The SuperORRUBA detector consists of two rings of silicon detectors. The detectors cover a geometrical area, 7.5 cm×4 cm, with the front sides divided into 64 1.2 mm×4 cm strips, and the back sides segmented into 4 7.5 cm×1 cm strips. The individual elements were assembled into two dodecagonal rings, one forward of 90° in the laboratory and the other backward. The radius of the forward (backward) angle ring was 11.2(12.5) cm, respectively, as measured from the beam axis to the center of the detector, and the angular range 55–125° is covered. When fully instrumented, the SuperORRUBA detector has 70% azimuthal coverage at forward angles and 60% azimuthal coverage at backward angles. Typical use-rate of existing instruments in this experimental program is 2%.
24. The Array for Nuclear Astrophysics Studies with Exotic Nuclei (ANASEN) is an active-target detector array developed specifically for experiments with radioactive ion beams. ANASEN is a collaborative project between LSU and FSU. The array consists of 40 Si-strip detectors backed with CsI scintillators. The detectors cover an area of about 1300 cm2 providing essentially complete solid angle coverage for the reactions of interest with good energy and position resolution. ANASEN also includes a position-sensitive annular gas proportional counter that allows it to be used as an active gas target/detector. Typical use-rate of existing instruments in this experimental program is 5%.
25. The Oak Ridge Isomer Spectrometer and Separator (ORISS) is a multi-pass time-of-flight spectrometer used for isomer and decay spectroscopy experiments. The device is designed to provide pure beams of rare isotopes with a resolving power of 400,000. It consists of an injection radio-frequency quadrupole, a multi-pass beam chamber housing electrostatic mirrors, and a time-of-flight detector at the ejection point. Typical use-rate of existing instruments in this experimental program is 0%.
26. Jet Experiments in Nuclear Structure and Astrophysics (JENSA) gas jet target provides a target of light gas that is localized, dense, and pure. The JENSA system involves nearly two dozen pumps, a custom-built industrial compressor, and vacuum chambers designed to incorporate large arrays of both charged-particle and gamma-ray detectors. JENSA is used at the target position of SECAR for studying astrophysically relevant reactions in inverse-kinematics. Typical use-rate of existing instruments in this experimental program is 30%.
27. VANDLE is a highly efficient plastic scintillator array constructed for decay and transfer reaction experimental setups that require neutron detection. The array consists of 48 plastic scintillators outfitted with digital electronics and has an energy resolution of 120 keV for 1 MeV neutrons and an energy threshold of 100 keV. This instrument has been used in conjunction with LENDA to provide large solid angle coverage for neutron detection following transfer and charge-exchange reactions. Typical use-rate of existing instruments in this experimental program is 5%.

Planned Instruments (Beyond Five Years)

GRETA (will fill the role of current GRETINA) will use highly segmented hyper-pure germanium crystals together with advanced signal processing techniques to determine the location and energy of individual gamma-ray interactions, which are then combined to reconstruct the incident gamma-ray in a process called tracking. GRETA will consist of a total of 120 highly segmented large-volume, coaxial germanium crystals, with four crystals combined to form a total of 30 Quad Detector Modules, designed to cover the total solid angle with a close-packed spherical geometry. Each crystal will be electrically segmented into 36 individual elements and a core contact, and read out over custom designed, digital electronics. The detector signals will be analyzed to

reconstruct gamma-ray energies and interaction points in a dedicated HPC cluster of commercially available CPUs. Typical use-rate of existing instruments in this experimental program is 30%.

28. HRS (will fill the role of current S800) for a wide range of nuclear reaction and structure studies that is matched to FRIB beam rigidities. The HRS will have a magnetic bending power up to 8 Telsa, large momentum (10% dp/p) and angular acceptances (80x80 mrad), and momentum resolution 1 in 5,000. A high-acceptance beam transport line will deliver rare isotope beams from the A1900 focal plane to the HRS target, and this beam line can operate in either achromatic or dispersive mode. The spectrometer will have three operating modes: one for high-resolution spectroscopy, a second for invariant mass spectroscopy, and the third for mass measurements using magnetic rigidity and time of flight. The focal-plane detector system will be similar to the S800 spectrograph and will include tracking detectors, an ion chamber, plastic scintillators for timing and energy loss, and a 32-segment CsI(Na) hodoscope for particle identification. Typical use-rate of existing instruments in this experimental program is 50%.
29. The FRIB Decay Station (FDS) (will fill the role of current BCS and ancillary detectors) is a state-of-the art instrument for nuclear structure and astrophysics studies of most exotic nuclei. The FDS will contain multiple, modular detector subsystems for the observation of charged particles, photons, and neutrons. The FDS will be deployed with the ability to modify the combination of detector subsystems according to the needs of specific experimental programs. All detectors will be read out using digital electrons and some experimental programs will take advantage of the waveform acquisition capabilities of the data acquisition electronics. Multiple workshops on the FDS have been held and a white paper has been developed. Typical use-rate of existing instruments in this experimental program is 20%.
30. The Isochronous Large Acceptance Spectrometer (ISLA) will make use of reaccelerated beams following ReA energy upgrades. ISLA will support the efficient detection of rare isotope beam-induced reactions by tagging reactions at the target by mass and product atomic number, permitting a clear correspondence to be inferred between radiation products observed around the target, and the final nuclei generated. These observations by ISLA will allow for recoil-decay studies of reaction products at focal-plane implementation stations. Typical use-rate of existing instruments in this experimental program is 15%.

    This planned instrument will allow the efficient use of reactions with rare isotope beams by tagging reactions observed at the target by mass and atomic number of the products to make a clear correspondence between the radiations observed around the target in powerful detector arrays and the corresponding final nuclei of interest and allowing for recoil-decay studies of reaction products in focal-plane implantation stations. ISLA will have a high acceptance (64 msr in angle, 20% in momentum) and a high mass-to-charge resolving power (of order 1,000), unique in the world, to carry out these studies.

Compute, Storage, and Network Capabilities

Compute

FRIB/NSCL "Fireside" Slurm batch cluster consists of 56 nodes of two quad core Xeon E5620 2.4GHz processors with 24GB RAM and 250GB SATA hard drive. These nodes are connected via 1 GE and Quad Data Rate InfiniBand standard. Five newer nodes are connected with 10 GE only and total 188 Xeon cores. These nodes have 96 or 192GB each. Retirement of the Xeon E5620 hosts and the addition of more new nodes is anticipated.

For interactive use, FRIB/NSCL provides users several Linux hosts for interactive analysis with a total of approximately 100 cores. The hosts have storage access and software environments matching the Fireside batch cluster.

Storage

FRIB/NSCL utilizes a NetApp storage system to provide reliable Enterprise-class storage. Snapshots, off-site replication, and tape backups are maintained for data security. This storage is used to provide user home areas, shared project storage space, shared departments storage space, and additional storage areas requiring a high level of data protection.

Linux/ZFS on commodity hardware is used to provide high capacity research data storage. Approximately 1.8 PB of storage is spread across three Linux/ZFS servers. These have 2x10GE network links. These comprise the "off-line" storage and are accessible from Linux compute systems. This supports off-line simulation, data reduction, and analysis workflows.

A separate Linux/ZFS system provides online events storage (output of DAQ systems) and is replicated to the off-line storage (Evtdata). The system is connected at 2x10GE. Normally, only one experiment is writing data to the online storage system. Archival copies of experiment raw data are made at the end of experiment running.

Network

FRIB/NSCL operates several internal networks. A WAN connection is provided by MSU Information Technology Services (ITS) with a 2x10GE connection between FRIB and MSU. The WAN connection is subject to ITS firewall restrictions. MSU ITS also provides Wi-Fi coverage within the FRIB/NSCL office buildings.

FRIB/NSCL manages a border firewall between the internal network and MSU campus. The internal wired networks are generally configured for one and 10 Gbs Ethernet. Additional information on the FRIB network infrastructure is included in Section 7.

MSU Campus HPC

The MSU Institute for Cyber-Enabled Research (iCER) operates HPC clusters that include more than 600 compute nodes with more than 20,000 Xeon cores. Nodes supporting NVIDIA TESLA V100, other NVIDIA GPUs, and Intel Phi are available. The clusters are linked together by high-throughput, low-latency InfiniBand. The Slurm batch system is used. For storage, GPFS is used to provide a 4 PB replicated, backed-up file system and a 1 PB high-performance scratch file system. Through a buy-in process and MSU support, iCER has sustained operations and system upgrades for more than a decade. Systems are now housed at the MSU Data Center that was completed in 2018.

Resources for Managing Instruments

At FRIB, device physicists maintain facility experimental equipment and help users set up their experiments. FRIB also provides technical support for making the interface between FRIB and users' equipment.

Composition of Data for Instruments

This information is included in the instrument descriptions provided previously.

5.2.5 Process of Science

Present to Five Years

NSCLDAQ is the main data acquisition application for producing and handling data flow at FRIB/NSCL. In NSCLDAQ, the process that manages the data pipeline is the ReadoutGUI. The ReadoutGUI constructs the data pipeline and controls the run state of the system (whether the data sources are producing data or not).

Source data are managed through the use of ring buffers. Data sinks, or consumers, access the data pipeline via the DAQ-net. Data sinks can include, e.g., storage, scalers, and online analysis.

Online analysis is usually accomplished using a home-built data unpacker and histogramming program SpecTcl. This application has inherent hooks to ROOT.

Data are stored in experimental event directories transferred using a data sink to dedicated disk space via DAQ-net. This data storage is accessed by experimenters using Linux workstations in the data-taking areas for the duration of an experiment.

Spokespersons are responsible for complying with respective data policies, including long-term storage of the research data and records of the data analysis, and for responding to data access requests. The Laboratory's Business Information Technology department facilitates the recording of research data during the running of the experiment (writing data to experimental event directories), the transfer of the data to long-term storage media by the spokesperson or his/her designee, and keeps, as a courtesy, a duplicate of the recorded research data for a period of two years after completion of an experiment. As a scientific user facility, FRIB/NSCL provides resources to assist spokespersons with recording and preservation of experiment data.

At the conclusion of an experiment, experimental account access is disabled. Off-line analysis is typically performed at the spokesperson's home institution. For workflows associated with FRIB/NSCL, the raw data are transferred to Evtdata directories accessible via Office-net. A compute cluster is available for off-line data analysis. Spokespersons from other institutions typically port the raw data via tape or high-volume USB drive. Spokespersons who have accumulated data sets of more than several TB have made use of the FRIB/NSCL Globus endpoint to ship data to remote storage for off-line data analysis.

Detector simulations contribute significantly to the compute demands during off-line analysis. Most of the FRIB/NSCL instruments are using GEANT4 (or a similar application) to simulate instrument performance, including detector acceptances, efficiencies, and detection thresholds. Analyses completed in-house are making use of the compute cluster or the High-Performance Computing Center (HPCC) that is part of MSU iCER.

FRIB/NSCL has recently embarked on leveraging ML algorithms for off-line data analysis. Multiple collaborations are using standard ML tools such as Tensorflow and Keras to analyze experimental data. To date, simulation data is used to train a ML model to extract features from a simulated analysis. The simulation data can be augmented with labeled experimental data if feasible. The trained model is then transferred to perform predictions. The model training makes use of the compute cluster and the HPCC at MSU iCER. The goal for the ML efforts is to incorporate ML models into the online data analysis and potentially data reduction.

GRETINA has been hosted at FRIB/NSCL for two previous campaigns, and a third campaign began in June of 2019. The data handling process for GRETINA differs from the description provided previously, since this detector has dedicated DAQ, compute cluster, and storage. The raw data pipeline from the GRETINA detectors to the compute cluster and storage is via dedicated fiber (2 x 10Gb bond).

Beyond Five Years

The capabilities of GRETA will likely represent the most significant performance challenge to network infrastructure of FRIB. GRETA will have two primary workflows: the first being a real-time workflow where the positions and energies of gamma-ray interaction points are determined from the digitized detector signals, and the second an experiment-specific workflow carried out by the experimental team (generally at their home institution) to perform Compton tracking on the interaction point set and infer physics observables. Further details can be found in the GRETA case study.

5.2.6 Remote Science Activities

SπRIT-TPC

The SAMURAI Pion-Reconstruction and Ion-Tracker (SπRIT) is a TPC constructed at MSU as part of an international effort to constrain the symmetry-energy term in the nuclear EoS. The SπRIT TPC is used in conjunction with the SAMURAI spectrometer at the Radioactive Isotope Beam Factory at RIKEN to measure yield ratios for pions and other light isospin multiplets produced in central collisions of neutron-rich heavy ions.

Data from a recent SπRIT TPC experiment totaled nearly 250 TB. Using Globus, it took nearly three months to transfer these data from RIKEN to MSU. These data are being analyzed at MSU using the HPC in iCER. iCER has CPU power sufficient to handle the analysis, but the lack of readily available and cost-effective storage space is limiting. No direct, high-speed network connections exist between iCER and FRIB, where high-volume storage is available and affordable. Some of the large-scale analysis is presently being completed using IT resources at RIKEN.

Another set of SπRIT-TPC experiments is planned in two years' time, and an improved approach to the "process of science" for this remote resource is needed.

Tools for Nuclear Theory

Part of the "process of science" that was not discussed in Section 4 is the comparison of experimental results with theory. FRIB is home of the FRIB Theory Alliance, a coalition of scientists from universities and national laboratories who seek to foster advancements in theory related to diverse areas of FRIB science.

At present, our remote access is mainly to both leadership class and capacity computing resources in the United States (ORNL, NERSC, OSC, iCER, etc.) and abroad. Access is mostly to launch jobs by remote login. Large data transfers are typically between different HPC centers, and do not involve so much the local FRIB/NSCL networks. Access to both remote capacity and leadership-class computing resources is critical for nuclear theory efforts. For example, to support many-body theory development, there will be a growing need for leadership and capacity resources.

To advance specific calculations for increasingly heavy nuclei with proper treatment of deformation, clustering, continuum degrees of freedom, etc. implies a need for several orders of magnitude larger computational effort. That will necessarily be leadership-class applications. The leadership-class calculations will then serve as anchors and constraints for ensemble applications of computationally cheaper models that are derived either from theory or through ML techniques (emulators). This will be key for day-to-day use in conjunction with, e.g., experimental analysis, large-scale parameter exploration, uncertainty quantification, and the treatment of dynamics (interfaces between structure and reactions). While such applications may not require leadership-class facilities for individual runs, they rely on the ready availability of capacity systems to meet the (growing) need for computing time.

5.2.7 Software Infrastructure

Manage Data Resources

Data resources are managed on an ad-hoc basis. No specific tools are used for management.

Data Transfer

FRIB's Business Information Technology department facilitates the transfer of data to long-term storage and to remote collaborators at the conclusion of an approved experiment. Data transfers to tape drives and hard drives

are currently performed at NSCL using standard Linux utilities. Network data transfers to remote collaborators have been accomplished on an ad-hoc basis at the request of the remote collaborator using a variety of tools. Tools used to accomplish data transfers currently include:

- Globus (subscription based): a secure, reliable research data management service.
- scp (open source): secure copy program to copy files between hosts on a network.
- rsync (open source): a file copy tool used for mirroring data files.

There is a desire to improve the performance of data transfers between the experimental facility and the HPCC at MSU. FRIB is in discussions with MSU ITS in this regard. No immediate plans exist to change the tools used to transfer data to long-term storage or remote collaborators.

Process Raw Data

Collaborations use a variety of tools to process raw experimental data into intermediate formats and data products. Programs that are globally installed and available to users at NSCL are as follows:

- Mathematica (commercial): a platform for technical computing across a range of fields.
- MatLab (commercial): a programming platform designed specifically for engineers and scientists with its own MATLAB language, a matrix-based language allowing the most natural expression of computational mathematics.
- GEANT4 (open source): a toolkit for the simulation of the passage of particles through matter. Its areas of application include high-energy, nuclear and accelerator physics as well as studies in medical and space science.
- Radware (open source): software package for interactive graphical analysis of gamma-ray coincidence data.
- ROOT (open source): a modular scientific software toolkit that provides all the functionalities needed to deal with big data processing, statistical analysis, visualization, and storage.
- TV (open source): a graphical plotting program for gamma-ray spectra.
- SuperMongo (open source): a plotting program.
- Origin (commercial): data analysis and graphing software.
- SpecTcl (open source): a nuclear event data analysis tool with an object-oriented C++ framework for histogramming and other data analysis operations. The Tcl/TK scripting language is embedded as the program's command language.

A small number of experimental collaborations are exploring the use of ML models to augment their data analysis pipelines on the two- to five-year time horizon. Some programs used in these applications include:

- Scikit-learn (open source): a Python-based ML library.
- Tensorflow (open source): an end-to-end open source platform for ML.
- Keras (open source): a high-level deep learning library.

5.2.8 Network and Data Architecture

Present to Two Years

Network Description

The FRIB/NSCL network (Office-net) consists of 10 and 1 Gbs Ethernet supporting general business IT functions, office LAN, Linux research compute systems, including interactive and batch systems, and Linux ZFS/NFS and Ceph File System (CephFS) storage systems. Infrastructure is shared with NSCL DAQ systems (DAQ-net) supporting DAQ experiment running, DAQ systems, online analysis, and online storage. During experiment data taking, data files are replicated from online storage to Linux research storage for larger scale "nearline" and off-line analysis. Separate networks exist for NSCL Cyclotron operations (NSCLControls-net), FRIB Linac operations (FRIBControl-net), etc.

WAN connectivity is provided via the MSU campus network (see accompanying figure). FRIB/NSCL has 2x10GE connection to campus with a FRIB managed border firewall. MSU R+E traffic primarily traverses MERIT/MiLR links to R+E pops in Chicago. Michigan Lambda Rail (MiLR) is a DWDM ring owned by MSU, Wayne State University, and the University of Michigan. MiLR is operated by MERIT.

Data Transfers

External data transfers are performed with Secure SHell (SSH) or Globus Online. A Globus endpoint on a 10GE connected virtual machine provides access to specified research storage file systems. Compute resources at MSU iCER are utilized by some local researchers. Data transfers between FRIB/NSCL and MSU iCER HPCC traverse multiple firewalls. In support of experimental data analysis, a 350 MB/s transfer rate using Globus Online has been demonstrated for FRIB storage to MSU iCER storage.

Infrastructure Work

FRIB is currently engaged in network connection discussions with ESnet.

Work is ongoing to increase network segmentation of the general FRIB/NSCL network. Segmentation goals are primarily security related with needs in the business IT portion of the network addressed first. Development of additional DAQ network isolation/independence and related capabilities addressing the relationship of DAQ, local research computing, and external data transfer, are anticipated.

Design work is ongoing concerning the connectivity of DAQ and off-line Linux compute and storage. An upcoming NSCL experiment plans to write data at up to 200 MB/s, requiring changes to existing systems.

Beyond Two Years

An ESnet-managed 100 GE connection utilizing MiLR is anticipated.

Dependent on WAN changes, a Science DMZ architecture is potentially beneficial to WAN and MSU HPCC transfers. Science DMZ is expected to include Globus transfer hosts, the ability to provision DAQ/experiment transfer links (virtual circuits), performance monitoring (perfSONAR), and security monitoring (packet capture based traffic inspection).

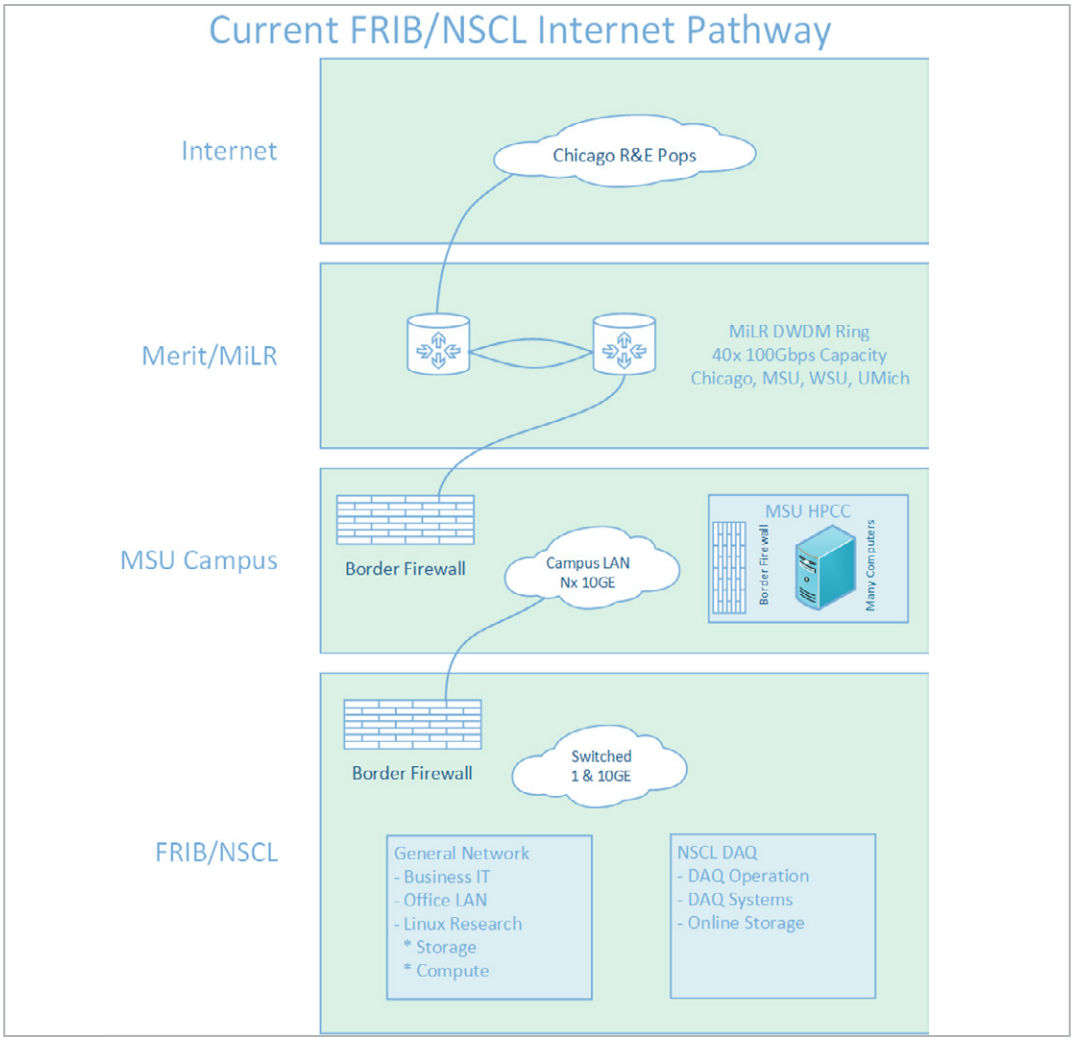

Figure 14. Diagram of the FRIB/NSCL pathway.

5.2.9 Cloud Services

MSU offers cloud services and cloud storage options. FRIB is currently not making use of these services for research data activities.

Cloud services include:

- Core Apps within Google Apps/G Street for Education Edition
- Microsoft Office Live within Spartan 365

MSU has established guidelines for the appropriate use of cloud services:

- https://tech.msu.edu/about/guidelines-policies/cloud-services-appropriate-use/

Cloud storage options:

- Kaltura Media Space
- Google Drive within Google Apps/G Street for Education Edition
- Spartan 365 (MSU implementation of Office 365)
- Remote storage within MSU's Desire-2-Learn (D2L) Learning Management System (LMS)

5.2.10 Data-Related Resource Constraints

- Network connectivity to the outside world is a single point of failure.
- FRIB has limited compute capability to address new instrumentation.
- FRIB lacks local expertise related to high-performance and exascale computing.

5.2.11 Outstanding Issues

FRIB follows a requirements-based approach for personnel and environmental safety, and property protection and information security. FRIB maintains management systems registered by National Sanitation Foundation – International Strategic Relations (NSF-ISR) to the ISO 9001 (Quality Management), ISO 14001 (Environmental Management), ISO 45001 (Integrated Safety and Health Management), and ISO 27001 (Information Security Management) standards.

IT infrastructure at FRIB needs to meet the requirements of our existing, integrated management systems. The two management systems most relevant to networking and scientific computing are the ISO 9001 and 27001 programs.

Within the ISO 9001 Quality Management System, FRIB is committed to delivering world-class beams of rare isotopes and to deliver the FRIB Project scope on schedule, within budget, with safety, and with high quality, to enable its users to achieve their scientific objectives.

FRIB was registered to the ISO 27001 Information Security Management System in 2018, with the general objective of delivering secure and reliable IT services. FRIB is committed to provide IT services that preserve the confidentiality, integrity, and availability of information in support of the laboratory's mission. Measurable objectives with the information security management system are associated with ensuring infrastructure availability and network integrity, while at the same time educating staff to recognize and react appropriately to potential threats to information and information systems.

5.2.12 Case Study Contributors

- **Steve Beher,** *MSU/FRIB,* beher@frib.msu.edu
- **Georg Bollen,** *MSU/FRIB,* bollen@frib.msu.edu
- **Scott Bogner,** *MSU/NSCL,* bogner@nscl.msu.edu
- **Mario Cromaz,** *LBNL,* mcromaz@lbl.gov
- **Thomas Glasmacher,** *MSU/FRIB,* glasmacher@frib.msu.edu
- **Clinton Jones,** *MSU/FRIB,* jonesc@frib.msu.edu
- **Dean Lee,** *MSU/FRIB,* leed@frib.msu.edu
- **Sean Liddick,** *MSU/NSCL,* liddick@nscl.msu.edu
- **Paul Mantica,** *MSU/FRIB,* mantica@frib.msu.edu
- **Tom Rockwell,** *MSU/FRIB,* rockwell@frib.msu.edu
- **Brad Sherrill,** *MSU/FRIB,* sherrill@frib.msu.edu
- **Jie Wei,** *MSU/FRIB,* wei@frib.msu.edu
- **Betty Tsang,** *MSU/NSCL,* tsang@nscl.msu.edu

5.3 Gamma-Ray Energy Tracking Array (GRETA)

GRETA is being planned as a key instrument for FRIB but can also operate at ATLAS located at ANL. Based on early work in the GRETINA project, GRETA is expected to be ready for operation in approximately six years. Early estimates have this instrument capable of producing approximately 200TB of data during an experimental run, using local technology to support storage, processing, and networking.

5.3.1 Discussion Summary

- GRETA is an experiment being designed for FRIB at MSU. This work is five to seven years away from operation and is still in the design phase.
- GRETA is the next generation of GRETINA, a gamma-ray spectrometer, used at ANL and FRIB.
- GRETA makes some design assumptions that can be adapted to future use cases that include separation of the control and data channels, ability to use local or remote computation, and ability to function with or without external networking capabilities (for storage and computation). As a result, the experiment is fully mobile.
- GRETA data sets will vary in size between experiment runs and are expected to range from 14GB to 189TB. Individual file sizes should be < 2 TB. The GRETA detector will write data to local storage at 500 MB/s.

5.3.2 Science Background

GRETA is an advanced gamma-ray spectrometer for low-energy NP measurements, funded by the DOE Office of Science, Office of NP, and is currently under construction. The array is a primary instrument for the upcoming FRIB, currently under construction at MSU. Within the context of the FRIB scientific mission, GRETA will be used for measurements of nuclear structure and reactions and nuclear astrophysics, with both fast and reaccelerated rare isotope beams. Following the excitation of an atomic nucleus, typically during a reaction with a fixed target, de-excitation gamma rays are emitted as the nucleus returns to its ground state. The high-purity germanium (HPGe) detectors of GRETA, which will cover approximately 80% of the solid angle surrounding a target, detect these gamma rays with excellent energy resolution and efficiency. In addition, the segmentation of the individual GRETA detectors allows, through the signal decomposition process (explained in a subsequent section), localization of the gamma-ray interactions within several mm3, and thus the ability to reduce background through gamma-ray tracking and make the best possible Doppler correction for radiation emitted in flight. The scientific case for GRETA is centered on understanding the structure and excitation modes of the atomic nucleus across the nuclear landscape, in order to provide the input and constraints for nuclear theory to move towards a predictive description.

As an instrument at FRIB, the GRETA scientific program will consist of individual experimenter-led measurements approved by the FRIB PAC. Typically, an individual measurement requires 2–10 days of beam time, with GRETA operated in conjunction with particle detection systems or spectrometers.

The raw data collected from the GRETA detectors are the energies, times, and associated waveforms as captured in the ADCs/FPGAs which instrument GRETA's roughly 4,000 electronic channels (120 x 36-fold segmented HPGe detectors). Specifically, for a single detector (single HPGe crystal) registering a gamma-ray event, the raw data of 36 segment waveforms, times, and energies, in addition to the full-volume waveform, energy, and time, are captured. From this data we can infer the location of gamma-ray interaction points, through a procedure known as signal decomposition, which effectively fits the observed signals against a library (basis) of calculated signals on a grid of known positions. The signal decomposition procedure is carried out in real time (within seconds)

by a dedicated, co-located computing cluster. The resulting energy/interaction point data, along with any data provided by auxiliary detectors, is provided to the experimental team of a given measurement. These data are cached locally for a period of weeks to allow sufficient time to transfer the data to their local institution/computing resource for analysis.

Generally, analysis of the previously provided data is carried out by the experimental team at their home institutions. Analysis and data interpretation are a time-consuming process (many months) but not a very computational-intensive process (can be done on local computing resources) The nature of this analysis is very much experiment dependent.

5.3.3 Collaborators

GRETA will be sited primarily at FRIB. As a major item of equipment (MIE), GRETA use is overseen by a user proposal process overseen by a PAC. Beam time and use of GRETA at FRIB are granted based on PAC evaluation of proposal scientific merit. It is envisioned that users will employ GRETA for 2- to 10-day experiments.

Proposals are generated by PIs at U.S. universities, U.S. national laboratories, and international institutions. Data produced during the approved beam time for a given experiment must be transferred from the host facility to the analysis point, which is generally their home facility (be it a university or laboratory).

User/ collaborator and location	Is a primary or secondary copy of the data stored?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of data set? (report in bytes, e.g. 125GB)	Frequency of data transfer or download? (e.g. ad-hoc, daily, weekly, monthly)	Are data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g., difficult tools, slow network)?
U.S. university-based PIs	Primary	Data transfer	14GB–189TB	Ad-hoc (< 4 weeks)	No	N/A
U.S. national lab-based PIs	Primary	Data transfer	14GB–189TB	Ad-hoc (< 4 weeks)	No	N/A
International PIs	Primary	Data transfer	14GB–189TB	Ad-hoc (< 4 weeks)	No	N/A

Table 5. GRETA Data Summary.

5.3.4 Instruments and Facilities

This case study describes the 4π γ-ray tracking array GRETA which will be a powerful new instrument needed to accomplish a broad range of experiments in low-energy nuclear science. GRETA marks a major advance in the development of γ-ray detector systems and can provide order-of-magnitude gains in sensitivity compared to existing arrays. It uses highly segmented hyper-pure germanium (HPGe) crystals together with advanced signal processing techniques to determine the location and energy of individual γ-ray interactions, which are then combined to reconstruct the incident γ-ray in a process called tracking. Full GRETA comprises 30 Quad Detector Modules (120 HPGe crystals) packed in a spherical geometry to surround a fixed target, a digital electronics system, a local computing cluster, and a support frame to precisely locate the detectors.

GRETA is currently in its design phase with an expectation that CD2/3 will be achieved in the summer of 2020. The project scope will be delivered in two phases (CD-4A and CD-4) to enable the possibility of early science with the Phase-1 delivery of electronics, computing, and mechanical subsystems and initial detector modules, followed by Phase-2, procurement of the balance of detector modules over several years. We are currently planning for CD4A in 2024 and CD4 (all detectors, full rate) in 2027. GRETA will be initially sited at FRIB with

the possibility of later operations at ATLAS/Argonne (n.b., the device is designed to be moved either within a facility or between facilities to take advantage of different beamlines/accelerators). Once GRETA has been built, it will most likely be operated at FRIB by a local operations team, supported by LBNL technical staff. Given the GRETA construction schedule, comments in this report apply to the five-plus-year timescale (strategic planning).

GRETA performs a real-time analysis of the digitized waveform data from the 120 individual HPGe detector crystals. This analysis (primarily) consists of determining the positions and energies of gamma-ray interaction points. This is a computationally intensive process and requires the use of a dedicated GRETA computing cluster co-located with the experiment. To keep pace with the required event rate, a 5,000-core cluster would be currently required (part of GRETA design is refactoring the signal decomposition code, potentially porting part of it to GPUs so the exact configuration of the cluster is not yet known). Processed data will be stored to a 1 PB storage array, which acts as a cache until it is moved to the experimenter's home facility for later analysis. A diagram showing the main components of the planned GRETA computing system is given in **Figure 15.**

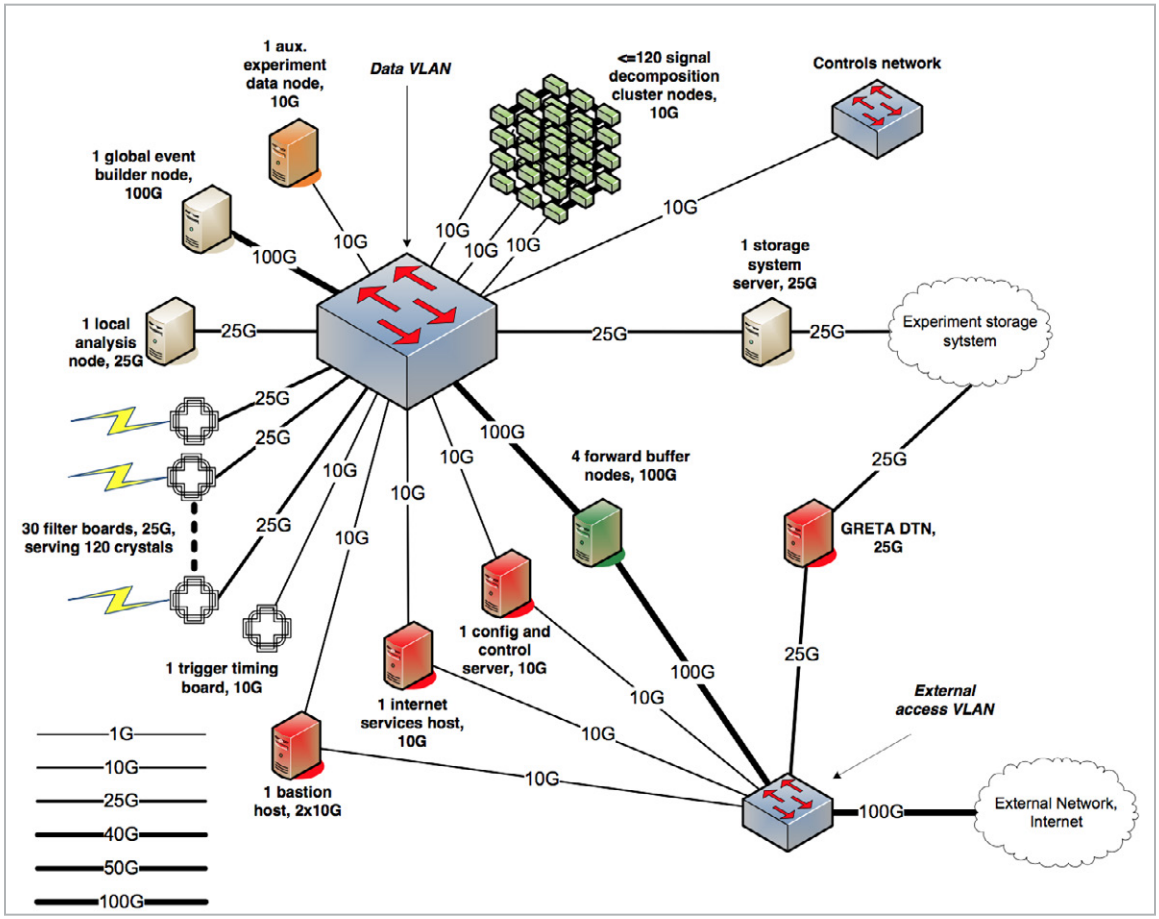

Figure 15. Major components of the GRETA computing system.

The GRETA network is a self-contained instrument that is movable, thus its network address space is abstracted from that of the host laboratory. External network facing components (in red) include a bastion host for remote logins, an internet service host that abstracts standard services for the instrument, a DTN for moving processed data to experimenter home facilities, and the forward buffers (in green) to admit the possibility to send full waveform data to remote computing facilities.

The data sets produced contain gamma-ray interaction points, energies, event times, and any data collected by auxiliary detector systems. Also included are metadata concerning the experimental configuration. This data format itself is custom and will likely be encoded by a standard object serialization scheme.

Running at the maximum design rate, GRETA can write 500 MB/s to its local disk cache, although we expect typical rates to be less than this. Data set sizes are highly dependent on the physics case being studied. The size of the data set is dependent on the triggered gamma-ray rate, the auxiliary detectors employed, and the beam time allocated. Assuming an average five-day beam time allocation, we expect a per experiment data generation amount to vary between 14GB and 189TB. Individual file sizes should be < 2 TB.

5.3.5 Process of Science

Associated with each GRETA experiment are two primary workflows. The first is the real-time signal processing workflow that occurs internal to the GRETA instrument. This is common to all experiments. The second is a data analysis step carried out by the experimenter and the experimenter's group. A brief description of both of these workflows is given in a subsequent section.

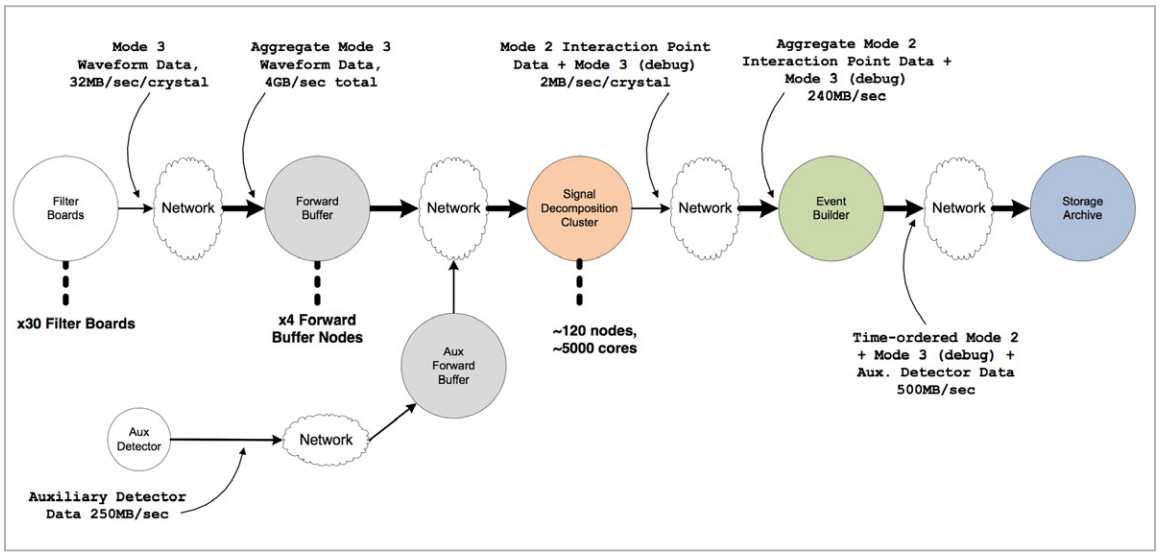

Figure 16. Signal processing workflow for the GRETA spectrometer.

The workflow for real-time signal processing is shown in **Figure 16.** The primary data producers are the FPGA-based filter boards of the GRETA electronics subsystem. These boards output windowed waveforms and energy/timing filter values as User Datagram Protocol (UDP) packets. There are 120 such UDP streams, each corresponding to a detector crystal, and their aggregate (maximum) rate is 4GB/s with the possibility of load asymmetry of up to 7:1. These data are captured by up to four forward buffer nodes which serve this data to a computing cluster for signal analysis and provide global flow control.

Signal processing is carried out in signal decomposition algorithms using a container-based architecture. One container is allocated for each GPU-equipped cluster node and is responsible for processing the data from a subset of detectors. This subset is determined by a run control system which balances workload across the cluster. Processing the data from each detector requires a large in-memory detector simulation which is tailored to the characteristics of a specific detector crystal. Multiple nodes/containers can be assigned to process data from a given detector type to accommodate the asymmetric workloads presented by some types of experiments. All signal decomposition containers forward their processed events to a global event builder. The event builder is

an aggregation component that orders events according to their timestamps to create global events. At this point auxiliary data are also time correlated. These global events are then written to the disk cache to be eventually transferred to the user home facility.

The second workflow involves the analysis of data by the experimenter(s). The first step in this process involves clustering and ordering interaction points into likely gamma-ray tracks and rejecting partial energy deposition events. This is followed by a number of experiment-dependent steps, which may include gating on auxiliary detectors, coincidence analysis, and comparison with simulation. This analysis step is not considered computationally (CPU or network) intensive and requires off-the-shelf computing resources.

5.3.6 Remote Science Activities

Experimenters are present at the facility during data taking, and all data processing prior to data storage is carried out on the local GRETA computing cluster. Remote access to the internal GRETA network is required by LBNL-based staff for technical support of the instrument.

The GRETA local computing infrastructure and signal processing algorithms are designed to deliver the full GRETA science goals. However, the project and scientific user community recognize that advances in algorithms could enhance the experimental sensitivity of GRETA and that then could benefit from using large-scale computing (HPC) facilities. In this case, waveform data from the forward buffers would be forwarded to local storage at the HPC facilities for signal basis optimization or potentially the real time processing itself (workflow 1). For example, a coupled signal decomposition and tracking algorithm would require this infrastructure.

5.3.7 Software Infrastructure

Application-specific software developed by the GRETA project and user community is used to do real-time signal processing (workflow 1) and analyze processed data (workflow 2). We expect to use Globus for file transfers between GRETA and the users' home institutions, but this decision has not been finalized.

5.3.8 Network and Data Architecture

The local network architecture for GRETA is centered on the data acquisition and signal processing components as described in section 5.3.5. The network diagram, given by **Figure 15,** shows several key components, including:

- The filter boards, which send the data captured by the detector electronics to the computing infrastructure for signal processing;
- The forward buffer nodes, which provide queuing and flow control for the data streams coming from the filter boards so that the data can be sent to the cluster for signal processing without packet loss;
- The signal decomposition cluster, which provides the computing necessary for signal processing;
- The global event builder, which aggregates the data from the cluster and produces event data files which will be analyzed by the experiment team; and
- A storage system, which provides non-volatile cache storage for the data files produced by the global event builder.

Another view of this workflow is shown in **Figure 16,** which gives the workflow from a dataflow perspective rather than from a network architecture perspective.

Three primary ways in which GRETA will interact with the WAN are:

1. System-level access by staff with appropriate access permission for system maintenance;
2. Download of experiment data sets from the GRETA DTN to remote analysis resources; and

3. Possible future signal processing modalities which require resources beyond the capabilities of the production GRETA signal decomposition cluster.

Each of these is more completely described in a subsequent section.

System-level access for maintenance and troubleshooting is not expected to occur routinely, as there will be an operational team at FRIB able to perform normal and expected system administration and maintenance tasks. However, it is possible that remote access by system experts from other sites (e.g., LBNL, ORNL) will be necessary. This is expected to be normal SSH-based access as one would expect for any remote system login and is not expected to be a significant driver of network requirements. The GRETA network design includes a bastion host for secure remote access, in accordance with network security best practices.

Download of experiment data sets will occur using the GRETA DTN in accordance with the Science DMZ design pattern, consistent with best practice for remote access to large-scale scientific data sets. The data-transfer tools to be used on the DTN have not yet been decided. However, it is expected that Globus, or a toolset with capabilities similar to Globus (automation of large transfers, automated fault recovery, parallel stream capability, etc.) will be used.

It is possible that future advances in signal processing algorithms might make the use of a remote HPC facility attractive for processing the data for some GRETA experimental scenarios. While support for these potential future activities is outside the scope of the GRETA project (the GRETA signal processing cluster is fully capable of supporting all currently-envisioned GRETA experiment scenarios), the GRETA network architecture provides the flexibility to support the use of external signal processing resources if that should prove advantageous in the future. This mode of operation would be an example of the superfacility model of integrating an experimental facility with an HPC facility by means of a high-performance network. In this scenario, the forward buffer hosts would serve data to a remote signal processing cluster or HPC resource. This would require up to 4GB/sec (36Gbps) of real-time data-transfer performance between the GRETA site at FRIB and the remote resource used for signal processing.

5.3.9 Cloud Services

For the next four years, during the design and production stages of the GRETA project, we do not plan to use cloud services.

Following this, when the instrument begins to collect data at FRIB, we expect the real-time component of signal processing (signal decomposition) to be carried out on its local computing cluster. Subsequent data analysis carried out by experimenters (workflow 2) may or may not use cloud resources. Given that the computational needs of this analysis are currently modest, we expect the demand for use of cloud services in the final analysis to be limited.

5.3.10 Data-Related Resource Constraints

The GRETA project does not anticipate or foresee future resource constraints to meet the project scientific goals.

5.3.11 Outstanding Issues

N/A

5.3.12 Case Study Contributors

- **Mario Cromaz,** *LBNL, NSD,* mcromaz@lbl.gov
- **Eli Dart,** *LBNL, ESnet,* dart@es.net
- **Heather Crawford,** *LBNL, NSD,* hlcrawford@lbl.gov
- **Paul Fallon,** *LBNL, NSD,* pfallon@lbl.gov

5.4 Argonne National Laboratory — Gammasphere / Argonne Tandem Linear Accelerator System (ATLAS)

ATLAS and the Gammasphere detector are located at ANL. This infrastructure is capable of hosting GRETINA and will also be capable of hosting GRETA in the future. Experimental data volumes are manageable, but there is a need to increase sophistication around the analysis pipeline, in particular migration to higher-order computation and improving software tools.

5.4.1 Discussion Summary

- The Low-Energy NP Research Group (LER) operates ATLAS and the Gammasphere detector at ANL.
- The internal network infrastructure of LER is 1Gbps, and the ANL Laboratory Computing Resource Center (LCRC) provides computational resources.
- The amount of raw data produced in an experiment is variable and dependent on experimental conditions. At most it is possible to produce 1.5TB of data on a daily basis, with experiments running typically between three to seven days (4.5 to 11TB).
- Globus is available for use, but not used by the majority of users. Experimental groups are typically used to traveling with external storage resources.
- ATLAS also hosts GRETINA and will host GRETA when the time comes. The experiment is designed to be mobile.
- The analysis of Gammashphere is labor intensive. Typically, there is a two-year analysis window, and the process is not automated in an extensive way. More computing being available would help but not fix the problem of stitching together the resources, which is a manual process.

5.4.2 Science Background

The LER in the physics division at ANL strives to understand the structure and stability of the nuclei around us and in the cosmos, to explore their astrophysical origin, and to use nuclei as sensitive probes in searches for new physics. The scientific questions guiding this research are:

- What are the limits of nuclear stability?
- What are the mechanisms responsible for shell evolution in nuclei?
- What are the astrophysical processes of nucleosynthesis and which nuclear properties are key to their understanding?
- What is the physics beyond the Standard Model?

To address these questions, LER leads the nuclear research program at ATLAS, the flagship DOE user facility for stable isotope beams. Our group maintains experimental devices, which are utilized by the facility to make measurements and store data. These devices include the helical spectrometer, HELIOS, several ion traps, a vacuum mode and gas-filled recoil spectrometer, FMA and the Argonne Gas-Filled Fragment Analyzer (AGFA), and several gamma-ray spectrometers, including Gammasphere and GRETINA. Since GRETINA is only here for targeted campaigns, Gammasphere is clearly the most prolific data-producing device and will form the basis for this exercise.

Since ATLAS is a user facility, experimental groups from all over the world can make proposals to use the facility. These groups typically collaborate with members of the LER group. The Office of NP provides funding for the facility, ATLAS, and the LER GROUP.

The Gammasphere device is composed of ~100 Compton-Suppressed Ge detectors. One Gammasphere module consists of 1 HPGe detectors and 1 Bismuth Germanate (BGO) scintillator side shield. Data are collected utilizing a digital data acquisition system, which stores information regarding the detectors, which have registered hits for each triggered event. Each Gammasphere module can record up to four channels of data (Ge central contact, Ge side contact, BGO sum signal, and BGO hit pattern). The DAQ trigger selects which data are read out and stored on a hard disk. A file on the hard disk is generated for each digitizer module. The data are divided into "Runs" which typically last for one hour. Once a run is finished, the data files are merged into one file, ordered in time. The merged, time-ordered data set is used to build events and allow for analysis of the data.

The PI of any particular experiment takes ownership of the data and works with the collaborators on a plan for analysis with the goal publication. The merged data files are copied and shared with members of the collaboration utilizing the local computer facility at ATLAS. Analysis of the data can take anywhere from several months to several years depending on the complexity of the data set and the individuals performing the analysis.

5.4.3 Collaborators

The ATLAS facility works on an individual investigator model. A researcher submits a proposal in response to a call from the facility. These calls are issued every 9 to 12 months. Proposals are then reviewed by a PAC, which makes recommendations on which proposals provide the highest-impact science. Each proposal is led by one or two PIs. It is the PI's responsibility to assemble the team that will perform and analyze the data. In addition, the PI while in residence at the facility copies the data and distributes to the collaboration. Consequently, no data are transmitted to investigators over the network; rather, copies of the hard drives are made available to those who need the data to perform analysis.

For informational purposes, investigators from ORNL, LBNL, BNL, LANL, and Lawrence Livermore National Laboratory (LLNL) perform experiments at the ATLAS facility as either PIs or collaborators. Other users include groups from Massachusetts Lowell, Washington University St. Louis, the University of Connecticut, Florida State University, MSU, Louisiana State University, Notre Dame University, and the University of Tennessee.

5.4.4 Instruments and Facilities

As indicated previously, ATLAS is a national user facility funded by the Office of NP in the DOE's Office of Science. The LER group takes responsibility for the experimental equipment which is used to perform measurements of interest to nuclear structure/reactions, nuclear astrophysics, and fundamental symmetries. The large detectors, HELIOS, Gammasphere, FMA, AGFA and Multi-Sampling Ionization Chamber detector (MUSIC), and GRETINA (when in-house) utilize the same digital data acquisition hardware. This acquisition system has been designed to allow high-throughput data acquisition and utilizes parallelism to accomplish this. One major upgrade planned for the facility is to add multiuser capability. This would allow for two beams to be supplied to two different end stations simultaneously. This upgrade could allow for 30–50% more experiments to be performed on a yearly basis and should be completed within the next five years.

Our experimental devices are operated on an internal network. The network operates at 1GB/s. The Gammasphere detector, which is the basis of this case study, provides users access to the network and desktop computers. The computers themselves allow hard drives to be plugged into the system, allowing data to be written directly to these hard drives. There are approximately 10 workstations available for Gammasphere users.

The amount of data produced in an experiment is variable and dependent on experimental conditions. Here we give numbers for a standard experiment using the maximum DAQ throughput. The data consist of information associated with each digitizer channel, such as timestamp, energy of detector gamma-ray, constant fraction timing, etc. Such an experiment would generate ~1500 Gbyte/day. These experiments operate typically for three to seven days and require 4.5 to 11Tbytes of hard drive storage for the raw data.

5.4.5 Process of Science

Once the data are taken, the investigators perform an analysis using standard techniques and software developed by the community over the years. Visualization of the data is provided by the use of Root. Once the data have been taken, networking capabilities are not utilized.

The most time-consuming aspect of post-data processing is the merging of our individual data files into a single merged and time-ordered data set. The merging is applied on a run-to-run basis.

5.4.6 Remote Science Activities

Currently, no remote resource is utilized or planned.

5.4.7 Software Infrastructure

Software utilized for post-processing has been written by the collaboration. The only open-source software used for analysis is Root from CERN.

5.4.8 Network and Data Architecture

ATLAS operates on a 1Gb/sec private network. There are plans to upgrade the network to 10Gb/sec in the next few years. While increasing the capability to move the data to other lab resources has been discussed, a decision has not yet been made. It is critical for ESnet to understand the network resources used to move data from the data source location to the wider facility/campus network, and to external collaborators or other data resources. Implementation of Globus into the network has been identified as a useful tool for the future.

5.4.9 Cloud Services

No cloud services are currently used or planned.

5.4.10 Data-Related Resource Constraints

None.

5.4.11 Outstanding Issues

None

5.4.12 Case Study Contributors

- **Michael Carpenter,** *ANL,* carpenter@anl.gov
- **Torben Lauritsen,** *ANL,* Torben@anl.gov

5.5 Argonne National Laboratory — CLAS12 / Electron-Ion Collider (EIC)

The ANL Physics group is a user of the JLab CLAS12 experiment. This use involves ESnet as a middle entity between experimentation components.

5.5.1 Discussion Summary

- The ANL LCRC provides computational resources, and there is a desire to use the ALCF.
- The JLab CLAS12 writeup provides information on data production and curation.
- JLab on-site computingresources are oversubscribed for user analysis which challenges availability ofresources. Through the use of ANL computing, additional user analysis can be performed.
- The tools that exist for CLAS12 do not lend themselves to the HPC environment and analyses may require multiple runs. ANL is developing new tools to leverage ML. The ANL group is attempting to transfer data (via Globus) for analysis and then transfer back results to JLab for storage to tape. The ANL group hopes to create a similar Globus-based transfer mechanism for raw data in future.
- The EIC is far out, so less is known about this. The ANL group expects that a similar workflow and capabilities will be needed to support CLAS12.

5.5.2 Science Background

The medium energy physics group from the Physics Division at Argonne conducts and participates in many JLab experiments and has active membership in the EIC users' group. The JLab experiments, conducted mostly in halls A, B, and C, use JLab's 12 GeV electron beam, with large particle detectors and magnetic spectrometers, to study QCD and nuclear matter. In halls A and C, due to the small angular acceptance of the magnetic spectrometers, the amount of data collected are small relative to the data produced by the CLAS12 detector found in Hall B. For reference, from January 2018 through March 2019, the CLAS12 detector collected 3.4 PB of raw data while halls A and C only produced a combined 0.13 PB. For this reason, we focus primarily on the data produced from the CLAS12 detector. However, it should be noted that the SoLID project (Hall A) could potentially produce similar types and quantities of raw data.

In addition to raw data, intermediate steps of data processing have comparable sizes but are typically transient from the point of view of a final analysis of the fully reconstructed data. However, these intermediate forms of data are where detector calibrations, run conditions, and new algorithms live and work. Therefore, having the ability to store and analyze these steps is highly beneficial.

Experiment MC and detector simulations also play a critically important role in data analysis and generate data sets similar in size to the raw data. The simulation data are important for understanding efficiencies, resolutions, and acceptances.

The raw data are archived on JLab's tape library system immediately after being recorded by the data acquisition. At this point, a typical analysis would use JLab's computing farm to process the data. Furthermore, sophisticated analysis tools often require multiple iterations through different data sets, dramatically increasing the compute resources needed to effectively deploy. To mitigate this, the ANL group will transfer the data (via Globus) to Argonne and process and analyze the data using the unique computing capabilities found at the Argonne site. The processed data will then be transferred to JLab and archived.

5.5.3 Collaborators

User/ collaborator and location	Is a primary or secondary copy of the data stored?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of data set? (report in bytes, e.g., 125GB)	Frequency of data transfer or download? (e.g., ad-hoc, daily, weekly, monthly)	Are data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g., difficult tools, slow network)?
CLAS Collaboration (JLab)	Primary raw data	JLab Tape Library, Globus data transfer	1 TB per run	Ad-hoc or daily depending on if the JLab is running	Yes, reduced in size via Globus	Working with and sharing large data sets with collaboration
EIC User Group	Both	Data portal / transfer	> 1 TB	Ad-hoc	Maybe	Managing distributed data sets and their metadata

Table 6. ANL CLAS12/EIC Data Summary.

The CLAS Collaboration is international and has many locations of data analysis. The primary location is JLab but with limited compute resources there is a strong desire to leverage external capabilities.

5.5.4 Instruments and Facilities

The CLAS12 science program will continue to take data at roughly 300 MB/s for the next two years (only running ~20 to 22 weeks a year). The CLAS12 reconstruction software is expected to improve over this time, but the amount of data accumulated and processing times will put significant strain on the JLab computing farm, which is used by all users to analyze past and present experiments and prepare for future experiments. Use of new data analysis tools and techniques, such as ML, will require moving the data to ANL. Therefore the goal in this time frame is to explore new workflows using the ANL LCRC infrastructure and map out the data processing and movement workflows that work for the Argonne group and the CLAS Collaboration. Currently, raw data are saved in 2GB chunks, with each experimental run (approximately an hour of beam time) yielding about 1 to 2 TB. This chunk size will be increased soon (producing fewer files per run). MC simulation data will also be produced at ANL and will be comparable in size with the raw data.

On the technology horizon, the ANLwould like to explore the possibility to move the CLAS12 raw data processing towards HPC, requiring significant CLAS12 software development. In addition to the LCRC computing facilities, we plan to leverage the resources found at the ALCF user facility and develop a proposal to bring exascale computing to CLAS12.

A possible strategy to consider over the next five years includes running the CLAS12 detector at higher luminosity, which will increase the data sizes proportionally. A factor of 5- to 10 luminosity upgrade will require significant computing resources. We would like to provide a Tier-2 computing center at Argonne that would require data transfers from JLab to ANL in a streaming fashion, where the raw data (already archived at JLab) is not saved after processing.

Present to two years: (300 MB/s) Use ANL's LCRC(Bebop) to process data and run simulations.

Next two to five years: (>500 MB/s) Process significant fraction of CLAS12 using LCRC and ALCF resources.

Beyond five years: (>5GB/s) Process luminosity upgraded CLAS12 data.

The EIC efforts will certainly be focused on physics and detector simulations for optimizing the machine and detector designs. The data sizes will become increasingly larger as the granularity of the detector grows smaller and simulation methods become sophisticated.

5.5.5 Process of Science

Today, raw CLAS12 event data are processed at JLab on the computing farm and outputs reconstructed Data Summary Tape (DST) files, which reduce the data raw data by a factor of 10 Significant additional experiment control and transient data streams are also produced during experiments. These additional transient data streams are roughly the same size as raw detector data generated, but do not need to be saved following the experiment.

5.5.6 Remote Science Activities

Currently we are using Argonne's LCRC resources to explore data processing for CLAS12 and the EIC. We will also pursue projects to use the ALCF, an Office of Science user facility.

5.5.7 Cloud Services

The ANL group is not using any commercial cloud services and does not plan to in the immediate future. However, the CLAS Collaboration has been exploring the use of the OSG. Recently the collaboration has been discussing the possibility of utilizing commercial compute resources but has not made use of these yet.

Within the EIC users' group there could possibly be, over the next few years, a strong case to make use of cloud resources. This avoids biasing the site selection and detector design decisions towards one group or laboratory.

5.5.8 Data-Related Resource Constraints

- **Present to two years:** We do not see significant network constraints in the next two years. This time frame will focus on the computing infrastructure and workflow development.
- **Next two to five years:** Increased data processing rate. This will place high demands on the network resources.
- **Beyond five years:** We expect an order-of-magnitude increase in the data throughput from JLab, other laboratories, and universities.

5.5.9 Case Study Contributors

- **David Potterveld,** *ANL,* potterveld@anl.gov
- **Linda Winkler,** *ANL, CELS,* winkler@mcs.anl.gov
- **Sylvester Joosten,** *ANL,* sjoosten@anl.gov

5.6 Brookhaven National Laboratory — Relativistic Heavy Ion Collider (RHIC) and ATLAS Computing Facility (RACF)

The RHIC and RACF are located at BNL. RHIC computation needs rely upon a mixture of local BNL, and other DOE facility, resources. The flow of data is managed by efficient and well-supported software tools. The facility is preparing for the EIC era by upgrading network, storage, and computational facilities.

5.6.1 Discussion Summary

- The RHIC and RACF are part of the BNL Scientific Data and Computing Center (SDCC).
- The RHIC takes advantage of computational capabilities elsewhere, including DOE HPC facilities (ORNL/ANL/NERSC) and NSF centers (National Center for Supercomputing Applications [NCSA]), Texas Advanced Computing Center [TACC]) for simulation workflows.
- Computation and storage are plentiful, and advanced tools are utilized for data movement and management.
- Network and data architecture are well supported. Science DMZ available, internal architecture in various stages of 10Gbps, 25Gbps, 40Gbps, and 100Gbps flavors. Effort to upgrade capacity as the experimental needs increase.
- Planning for the EIC, although not expecting for 10 years.
- The RACF also provides compute resources to Belle II/KEK out of Japan and ATLAS/LHC.
- The RHIC could greatly benefit from the ability to orchestrate data movement in a "third-party" fashion, e.g., staging information between participating sites. This functionality is being explored, but not readily available.
- The top collaboration sites for RACF are those of LHC facilities: CERN, University of Michigan/MSU (AGLT2), University of Chicago/University of Illinois Urbana Champaign (MWT2), JANET (UK), IN2P3 (France), ASGARR (Italy), TACC, etc.

5.6.2 Science Background

The RHIC and RACF are part of the BNL SDCC. As part of the host laboratory, RACF serves the computing and storage needs of the RHIC science program. Data from the RHIC experiments are stored, archived, and curated by SDCC, which provides the means for data processing and analysis for about 1,200 users distributed worldwide. Collaborative tools required by the large and lively supported scientific user community are also provided by the SDCC.

Over the years, the expertise of the RACF in the operation of reliable services led to the extension to the RHIC and ATLAS Computing Facility, and finally the SDCC, which recently started to provide central computing resources for the Belle II experiment at KEK in Japan. In this process the SDCC went through a transformation of a geographical decoupling of experimental facilities from storage and computing facilities. For the ATLAS experiment at CERN (Switzerland), the SDCC is one of the biggest external resource providers worldwide. About 25% of ATLAS data are transferred over the Atlantic. BNL distributes data to five secondary U.S. sites connected to ESnet and universities participating in the U.S. ATLAS program. Simulations produced at DOE (and NSF) HPC centers are shipped back for storage at BNL. For Belle II, the SDCC is the experiment data center outside of Japan. It is responsible for handling all of its raw data and supports central services (databases and data distribution system), traditionally hosted at the host lab. BNL's potential for hosting further medium and large new projects in the future will build upon this experience.

The SDCC is one of the largest HPSS sites in the United States and hosts in its tape libraries over 100 PB of RHIC data. About 30k CPU cores are available for RHIC data processing and analysis together with about 20PB of disk storage.

5.6.3 Collaborators

Through the NP- and HEP-supported scientific programs, the SDCC provides access to local resources to about 2,000 users worldwide.

Interfaced to the OSG, computing and storage resources are also accessible to collaborators of BNL-supported science programs (and a small fraction for opportunistic usage by the OSG community).

5.6.4 Remote Science Activities

Supercomputers at DOE facilities will be used mainly for simulation workloads by future NP programs. The bandwidth requirements for these applications are small in comparison to the needs from other programs supported at BNL.

5.6.5 Software Infrastructure

Data management (distribution, cataloging, retention policies) for HEP experiments is orchestrated by Rucio (open software), which has proven to be reliable and scalable. It allows centrally managed data distribution and also data discovery and subscription for individual users. Rucio is under evaluation for future NP and other scientific programs at BNL.

Data transfers are scheduled and organized with the GRID Data Transfer Service used at CERN, File Transfer Service (FTS). Several protocols are used (GridFTP, HTTPS, Xroot,). Third-party copy is a requirement for data transfers. Globus is also used for transferring data to/from BNL for various use cases.

Each experiment develops its own software for data transformation from raw to user data products. These transformations require detailed knowledge and integration of experiment specifics and cannot be shared between projects. However, future experiments are evaluating common software frameworks to develop these tasks.

Local development is underway to provide users an analysis batch processing capability based on Jupyter.

5.6.6 Network and Data Architecture

5.6.6.1 BNL Network Architecture

BNL has implemented a vendor agnostic, resilient, scalable, and modular terabit per second (Tbps) High-Throughput Science Network (HTSN) which serves as the primary network transport for all data-intensive collaborations at BNL. It provides high-throughput connectivity to all HPC and high-throughput computing collaborations and supports the timely transfer of large amounts of scientific data via the internet.

The HTSN has five key components:

1. **Network Perimeter**
   - Two (soon to be three) diverse 100Gbps circuits that peer with ESnet in New York City. These circuits are utilized by all scientific and administrative communities at BNL. All traffic to and from BNL flows through these ESnet circuits.
   - The BNL Network Perimeter transfers on average 7 to 8 Petabytes of data monthly, with spikes up to ~12 Petabytes.
2. **Science DMZ**
   - Supports open, high-speed WAN/internet access for all scientific collaborations throughout the BNL campus.

3. **Science Core**
   - A Tbps Science and Data Center Interconnect for data-intensive collaborations at BNL. This Science Interconnect enables high-speed connectivity between collaborations such as ATLAS, STAR, PHENIX, CAD, the Center for Functional Nanomaterials (CFN), NSLS-II, HPC clusters, RHIC, and RACF.
   - Intelligence and routing policies are applied within the HTSN Science Core to restrict or grant access to specific resources within the RACF.
4. **Spine**
   - A Tbps network spine that interconnects all leaf switches. Leaf switches consist of top of rack or chassis-based switches that connect compute, storage, or general infrastructure service servers.
   - The responsibility of the spine is fast packet forwarding and flexibility, not policy insertion or server termination.
   - External Border Gateway Protocol (eBGP) is utilized throughout the HTSN. EBGP was chosen for its ability to immensely scale and to create modularity and fault domain isolation down to the rack level. Each spine group shares the same Autonomous System Number (ASN) but does not have Internal BGP (iBGP) peerings between them. Each leaf or pairs of leaves will require its own ASN.
5. **Storage Core**
   - A redundant terabit per second switching block that aggregates high-performance storage services.

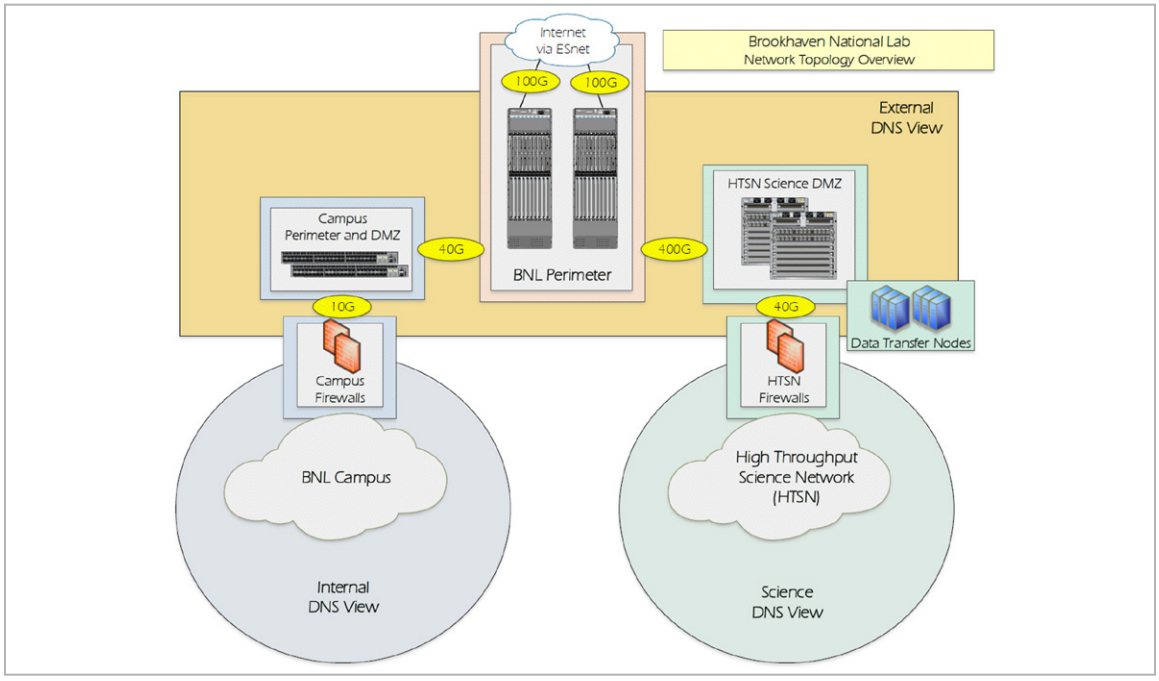

Figure 17. High-level overview of the BNL network perimeter and domain name system (DNS) architecture.

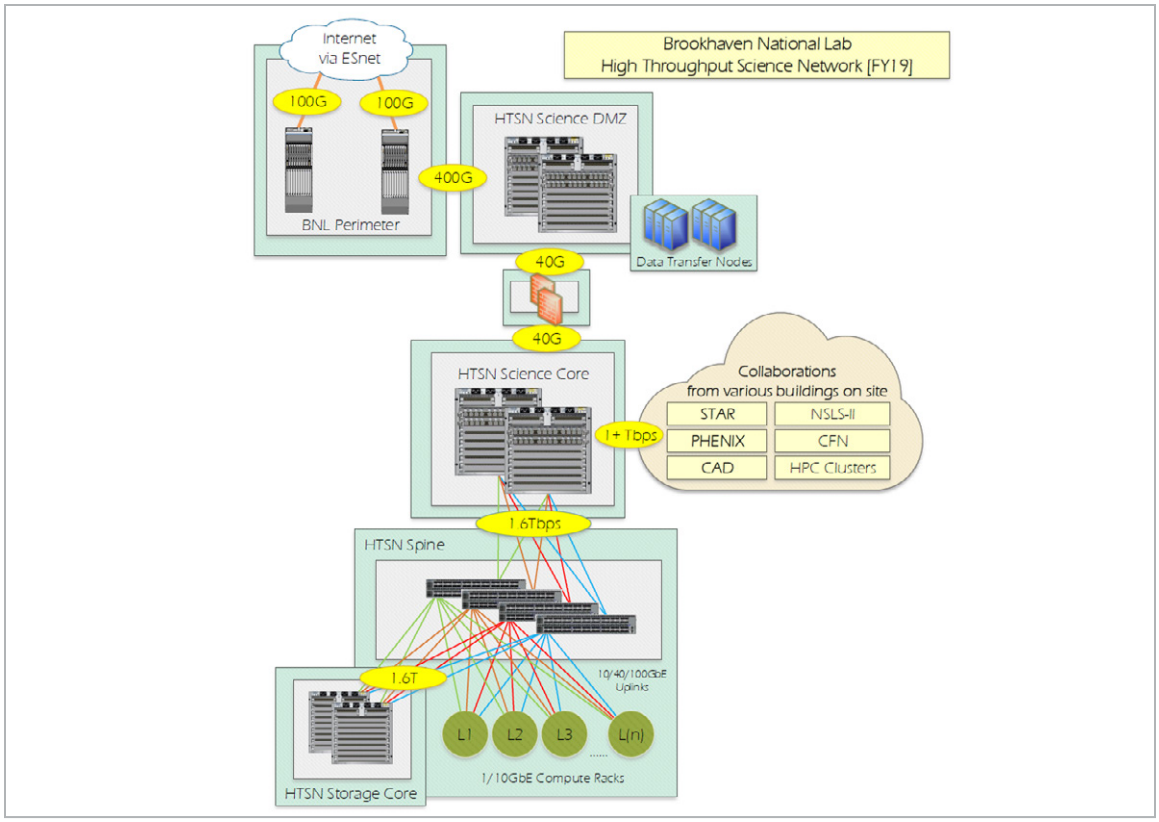

Figure 18. High-level overview of the BNL HTSN in FY19.

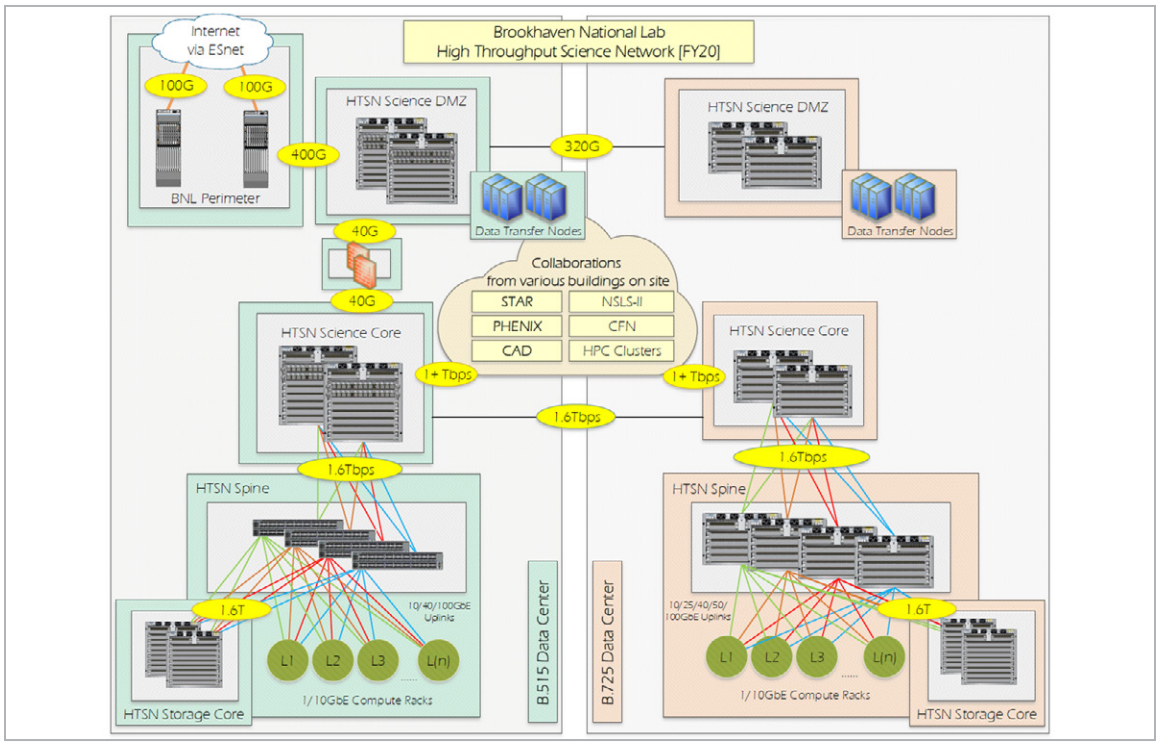

Figure 19. High-level overview of the BNL HTSN in FY20. (Includes HTSN expansion into new Data Center, right hand side).

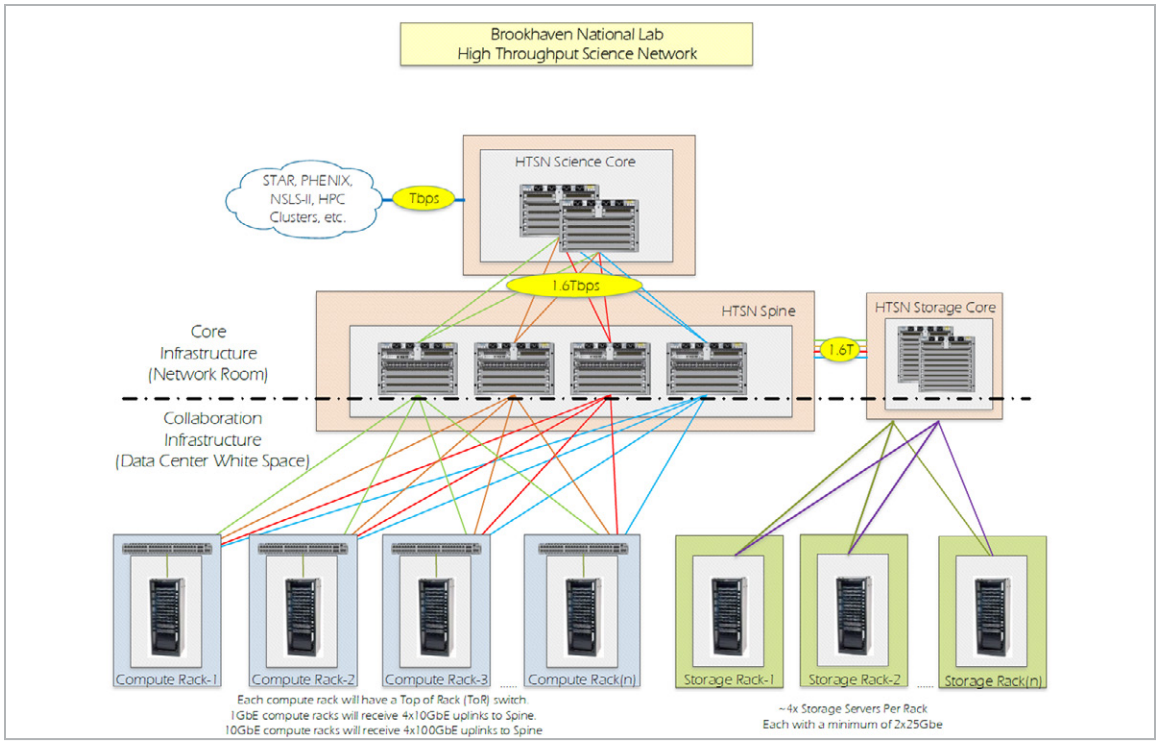

Figure 20. High-level overview of the BNL HTSN. Demarcation points between core and collaboration infrastructures.

The current set of WAN-connected systems includes:

- **DTNs**
  - 20 GridFTP/XRootD/WebDAV proxy servers with 2x10Gbps (WAN) + 2x10Gbps (LAN) each for ATLAS dCache storage. They are the primary DTNs for ATLAS storage at BNL.
  - Four GridFTP/SRM/XRootD/WebDAV servers with 2x10Gbps(WAN) + 2x10Gbps (LAN) each for Belle II dCache storage. They are the primary DTNs for Belle II storage at BNL.
  - One Globus connect server with 2x40Gbps (WAN) + 2x40Gbps (LAN) each. Mainly used by NSLS-II and the CFN.
  - Four BNLBox (BNL cloud storage solution) gateway nodes with 2x10Gbps (WAN) + 2x10Gbps (LAN) each.
  - Four STAR Grid proxy servers with 2x10Gbps (WAN) + 2x10Gbps (LAN) each.
  - Three BNL FTS servers with 2x10Gbps (WAN) each.
- **Caching**
  - Two XCache servers with 2x40Gbps (WAN) + 2x40Gbps (LAN) each. They are data cache service for ATLAS.
  - Four CVMFS (the CERN VM File System used for software distribution) proxy servers with 2x10Gbps (WAN) + 2x10Gbps (LAN) each.
- **PerfSONAR**
  - One perfSONAR server with 2x40Gbps (WAN) each.
  - One perfSONAR server with 2x10Gbps (WAN) each.

Technology Refresh Cycle

- **Present to two years (current budget horizon)**
  - One high-bandwidth perfSONAR service with 2x100Gbps to LHCOPN/LHCONE to be added in calendar 2020 to 2021.
  - Possible increase in the number of XCache servers with 2x40Gbps (WAN) + 2x40Gbps (LAN) each from two to four hosts in 2020 and from four to eight in 2021.
  - Possible increase in the number of Globus connect server with 2x40Gbps (WAN) + 2x40Gbps (LAN) each from one to two in 2020 and from two to four in 2021.
- **Next two to five years (current technology horizon)**
  - Planned conversion of 20 ATLAS dCache GridFTP/XRootD/WebDAV proxy servers from 2x10Gbps (WAN) + 2x10Gbps (LAN) connectivity to 2x25Gbps (WAN) + 2x25Gbps (LAN).
  - Planned conversion of four Belle dCache GridFTP/XRootD/WebDAV proxy servers from 2x10Gbps (WAN) + 2x10Gbps (LAN) connectivity to 2x25Gbps (WAN) + 2x25Gbps (LAN).
  - Possible conversion of XCache servers from 2x40Gbps (WAN) + 2x40Gbps (LAN) connectivity to 2x100Gbps (WAN) + 2x100Gbps (LAN).
  - Planned conversion of three BNL FTS servers with 2x10Gbps (WAN) connectivity to 2x25Gbps (WAN).
  - Planned conversion of three BNLBox gateway nodes with 2x10Gbps (WAN) + 2x10Gbps (LAN) connectivity to 2x25Gbps (WAN) + 2x25Gbps (LAN).
  - Planned conversion of four CVMFS proxy servers with 2x10Gbps (WAN) + 2x10Gbps (LAN) connectivity to 2x25Gbps (WAN) + 2x25Gbps (LAN).
  - Possible need for deploying up to four sPHENIX DTN nodes with 2x25Gbps (WAN) + 2x25Gbps (LAN) each.
  - Possible need for deploying up to 16 DTN nodes with 2x25Gbps (WAN) + 2x25Gbps (LAN) each for BNL's participation in DUNE (Deep Underground Neutrino Experiment, LBNF/DUNE). The DUNE requirements are expected to be at the level of ATLAS experiment requirements with respect to WAN connectivity for BNL from the same timeframe (2023).
- **Beyond five years (strategic planning)**
  - The ATLAS experiment is expected to start the data talking with high-luminosity LHC in 2026, resulting in an increase of the data rate by a factor of five. This could result in the increase of WAN network traffic by the factor of two to five for LHCOPN/LHCONE at BNL ATLAS Tier-1.
  - The previous upgrade of 20 ATLAS dCache GridFTP/XRootD/WebDAV proxy servers should be able to handle this increase in the network traffic.
  - The EIC and its detector experiments are expected to come online around 2030. The WAN data rate for the EIC is expected to be similar to the ATLAS experiment of the same time period (2030).

5.6.6.2 Performance and Statistics

The WAN bandwidth utilization over the past year is shown on the following graph. After the dip related to the winter stop of the LHC accelerator at CERN, network utilization has been steady.

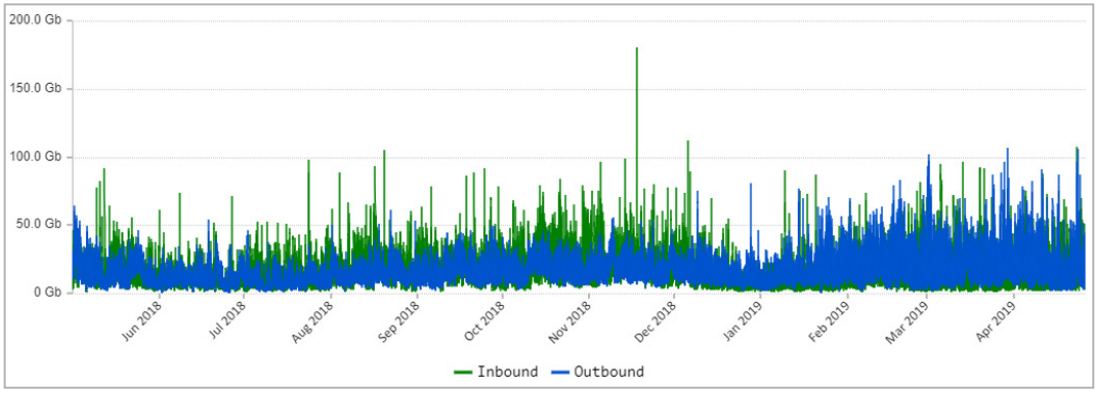

Figure 21. BNL WAN bandwidth utilization (May 2018 to April 2019).

The data volume transferred per month over the WAN since January 2017 is shown on the following graph. An increase of data-transfer volume of +30% compared to the previous year occurred in 2019. The average data volume transferred by month is close to 10PB.

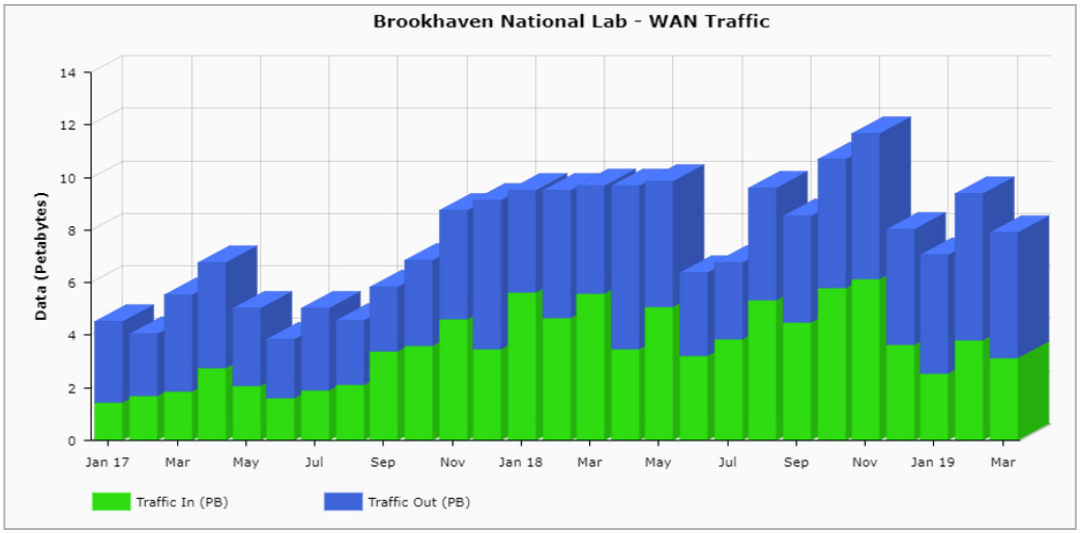

Figure 22. Monthly BNL WAN data-transfer volume (PB) since Jan. 2017.

The origins and destinations of the WAN traffic are shown by country on the following two pie charts. The United States is the primary origin and destination of WAN traffic.

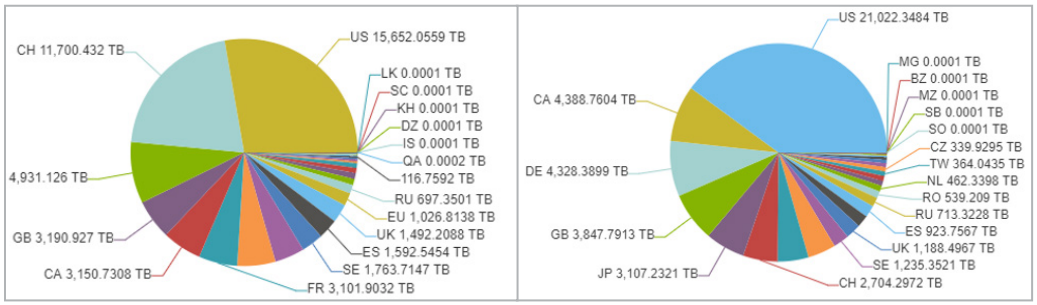

Figure 23. BNL WAN sources of incoming (left) and outgoing (right) data (by country) (May 2018 to April 2019).

The origins and destinations of BNL WAN traffic by ASN are shown on the following two graphs. Data volumes over 1PB/y are exchanged between 15 and 20 different ASNs.

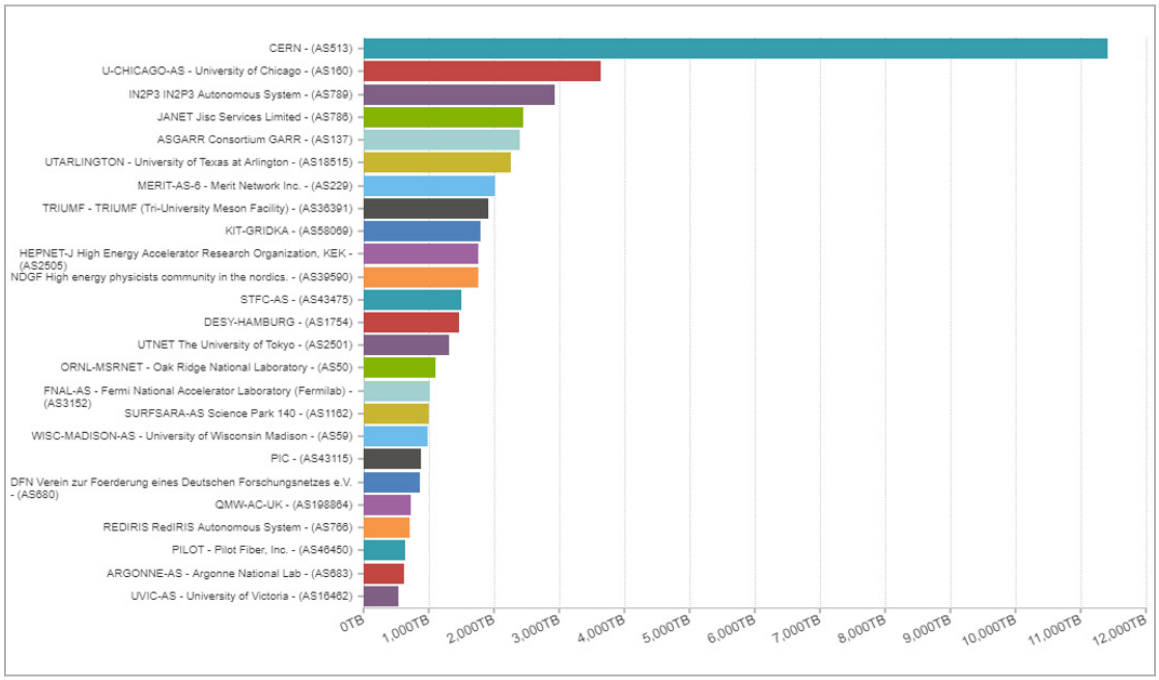

Figure 24. BNL WAN sources of incoming data (by ASN) (May 2018 to April 2019).

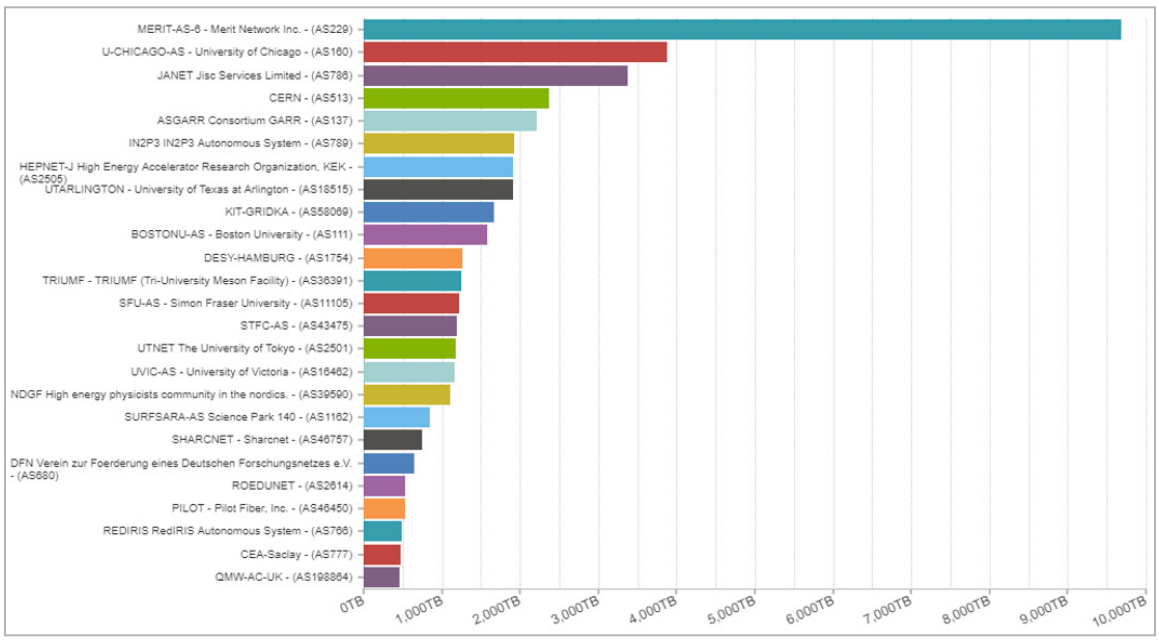

Figure 25. BNL WAN destinations of outgoing data (by ASN) (May 2018 to April 2019).

The data volume exchanged through the DMZ is around 2.5PB per week, as can be seen on the following graph. The largest user is the ATLAS experiment data flow.

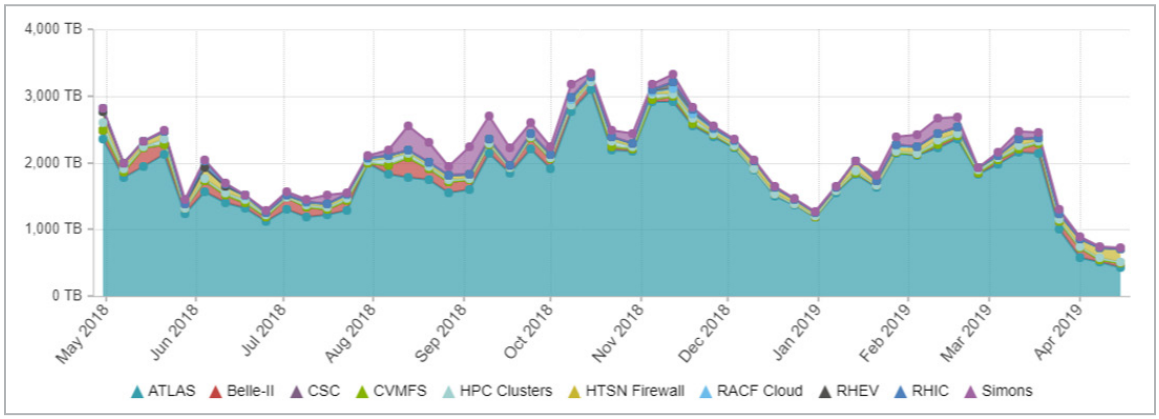

Figure 26. Weekly BNL Science DMZ bandwidth utilization per week (April 2018 to April 2019).

The internal network (LAN) traffic between the compute nodes in the STAR CAS and STAR CRS Linux clusters is shown in **Figure 27;** several GB/s of transfer rates are achieved. Each computing cluster is equipped with a several PB capacity distributed storage system.

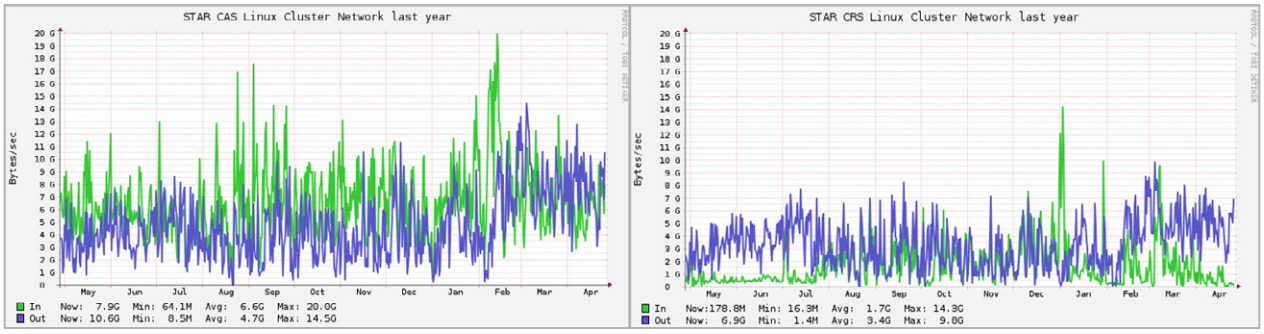

Figure 27. STAR CAS and CRS Linux clusters aggregate LAN traffic (April 2018 to April 2019).

The LAN network traffic between all RHIC Linux Farms (STAR and PHENIX combined) and the recently deployed RHIC central storage is shown in **Figure 28.**

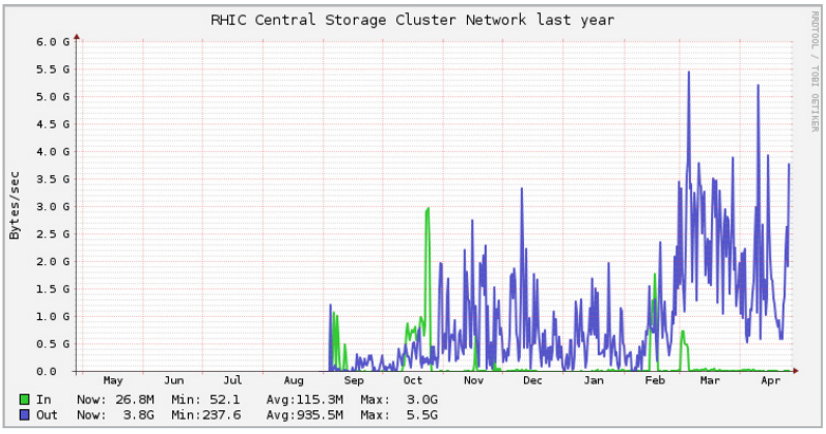

Figure 28. Aggregate LAN traffic between all Linux clusters (STAR and PHENIX combined) and the RHIC central storage (April 2018 to April 2019).

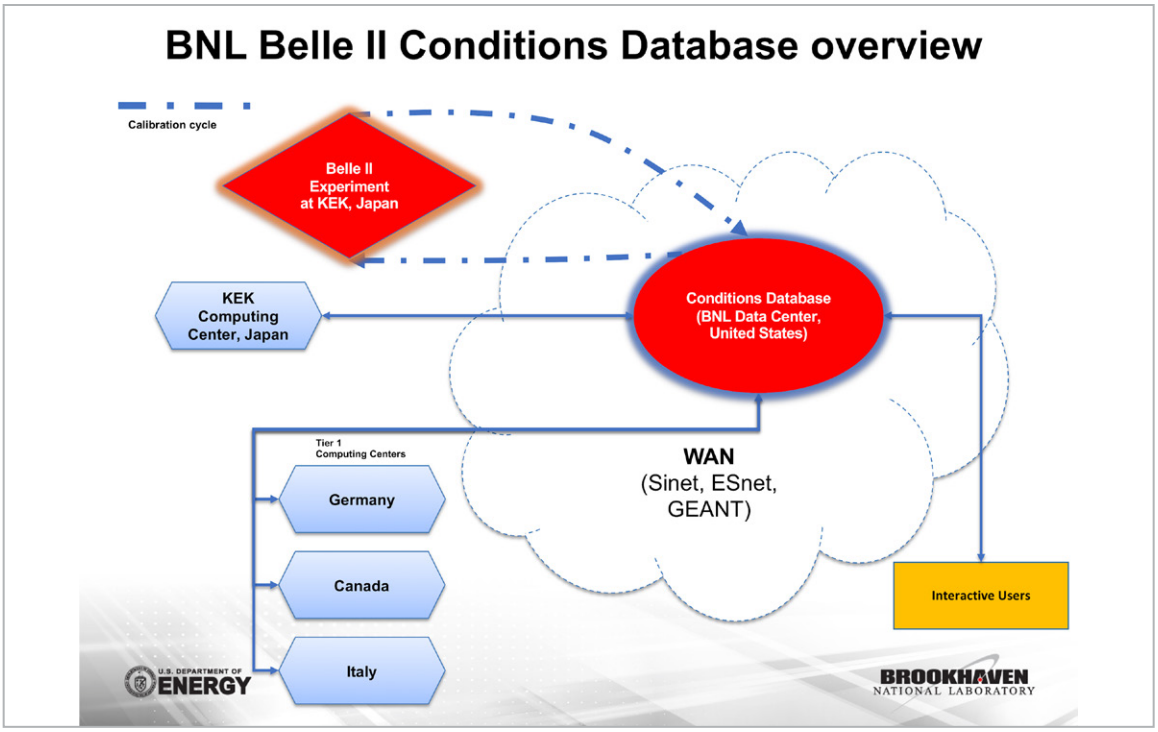

Figure 29. The Belle II experiment condition database infrastructure operated at BNL is used for data calibration and quality check at the experiment site in KEK. Conditions are accessed by various large centers distributed throughout the world.

5.6.7 Cloud Services

Services from cloud providers are not anticipated to be used in the near or medium term.

5.6.8 Case Study Contributors

- **Vincent Bonafede,** *BNL,* bonafede@bnl.gov
- **Mark Lukasczyk,** *BNL,* mlukasczyk@bnl.gov
- **Hironori Ito,** *BNL,* hito@bnl.gov
- **Alexandr Zaytsev,** *BNL,* alezayt@bnl.gov
- **Eric Lancon,** *BNL,* elancon@bnl.gov

5.7 Brookhaven National Laboratory — The Solenoidal Tracker at RHIC (STAR)

STAR at BNL is currently the last running experiment at the RHIC due to the Pioneering High-Energy Nuclear Interaction eXperiment (PHENIX) shutdown and sPHENIX upgrade. STAR has plans to run in parallel of sPHENIX, with the period of 2023-2025 run plans driven by the sPHENIX collaboration. No upgrades are envisioned ruing this period.

As a result of the experimental stability, no major upgrades are planned to workflow, computation, storage, or networking needs. Computation is handled with local and remote resources, and data mobility has functioned well. The lack of a network can become critical after several days, when the backlog of experimental data requiring analysis will exceed onsite storage resources.

5.7.1 Discussion Summary

- Data from STAR are analyzed during collection in order to perform real time experimental calibration. Processing/reprocessing is done local to BNL. The remote workflow aspects @ NERSC are in the process of getting formal allocation time. There is a larger project to look at OSG computational cycles, but NERSC is currently the largest external focus due to the capability of compute, storage, and network.
- Current raw data out of STAR can reach the .5 PB size. Creating the processed data sets can require between 3PB and 14PB of space (multiple format options). These rates will grow over the next several years (towards STAR shutdown) to around 20PB of raw data, and between 7PB and 40PB of processed data.
- The introduction of a new compact format for analysis (aka "picoDST") may very well mean that data sets can be transferred on the WAN to remote institutions. However, it is too early to tell where and when this will happen.
- Movement of data for processing at NERSC was accomplished at a rate of 700MB/sec. This involves a mixture of caching and streaming. Lots of effort is occurring to get things working well there, with optimization between storage, network, compute (CVMFS).
- STAR may survive in network isolation for a limited time, as many as one to two days. For quality assurance reasons during operation, computation is needed and often that occurs off-site using the WAN.
- STAR still uses a primary copy of data with no back-up at BNL.
- The EIC will involve BNL, JLab, and ANL, most likely.
- Lack of support for old GridFTP is causing evaluation of other protocols.

5.7.2 Science Background

The STAR experiment is one of the two large NP U.S.-based experiments at the RHIC and the only one running at the moment. Located at BNL in Long Island, New York, the facility has been one of the greatest successes of the U.S. NP research program and the first to observe convincing evidence of a new state of quark-gluon matter. In addition, it is the world's only polarized proton collider. RHIC has been extremely productive in delivering and accomplishing its scientific mission, and the first decade of physics deliverables produced in STAR alone has formed and produced 280 PhD students, 264 scientific peer reviewed papers, and 639 published papers with 29,000 citations.

Since its inception, the STAR detector's hallmark has been full acceptance mid-rapidity measurements with excellent particle identification capabilities. Recently installed upgrades to maximize the physics output from

Beam Energy Scan (BES) Phase II substantially enhance STAR's already excellent capabilities. These upgrades also enable STAR to continue its unique, ground-breaking, mid-rapidity science program in the period following BES-II. STAR additionally proposes measurements of forward photons, electrons from J/y and Drell-Yan processes, and inclusive jet, dijet, and hadron/jet correlation probes at both 200 and 500 GeV center of mass energies and demonstrates measurement capability and sensitivity through simulations. These measurements allow STAR to probe the fundamental structure of nucleons in new kinematic regimes and where existing data still provide rather poor constraints. One aspect is the composition of nucleon spin in terms of quark and gluon degrees of freedom; the other is to go beyond the one-dimensional picture of nucleons in momentum-space by correlating the information on the individual parton contribution to the spin of the nucleon with its transverse momentum and spatial distribution inside the nucleon. In p+A collisions, these measurements will enable STAR to study cold nuclear matter. The proposed forward tracking and calorimetry provide kinematic access to very small momentum fractions x in nuclei, facilitating investigations into the dynamics and nonlinear evolution effects in the regime of high gluon-density.

In addition to QCD studies with polarized p+p and p+A collisions, the proposed upgrade will facilitate the determination of quark-gluon plasma (QGP) properties in A+A collisions through improved measurements of the initial density fluctuations as well as the collective flow seeded by these fluctuations. The proposed forward upgrade will further quantify η/s through improved understanding of initial conditions via measurements of longitudinal flow de-correlation, multiple harmonics, and event-shape engineering in A+A collisions, and will allow studies of the possible existence of and limits on hydrodynamics and jet-medium interaction in small systems at RHIC energies. In A+A collision measurements with unprecedented precision, using deep penetrating leptons and photons as probes, scientists will be able to observe entire collision evolutions using A+A measurements with unprecedented precision.

The proposed upgrades and RHIC's extraordinary versatility are integral to all these measurements. Our program is designed to capitalize on STAR's existing resources, which include a proven multipurpose detector, established calibration techniques, and highly developed software infrastructure, and are a natural extension of the current STAR physics program. The programs outlined previously are essential steps towards the completion of the RHIC mission and will provide a natural transition to the highly anticipated EIC program.

Our run plan is shown in **Table 7.**

	2019	2020	2021	2022	2023	2024	2025
Dectector Configuration	iTPC/eOF	Partial Forward	Forward Calorimeter sTGC	FST	NO Detector Upgrade ONLY Operation Driven by sPHENIX		
Operation	19.6 GeV 14.5 GeV FXT ?	11.5 0.1 FXT	9.1 7.7	pp500	Au+Au200	pp pA	Au+Au200
Weeks	24	24	20	16	24	24	24

Table 7. STAR Data Summary.

Predictions are necessarily tentative beyond 2023 when sPHENIX experiments will predominate data production.

5.7.3 Collaborators

STAR remains a strong collaboration composed of 67 institutions from 14 countries, with a total of 668 individual collaborators total, to date. Noticeably (comparing to the last review), in the era of STAR being the sole running RHIC experiment with sPHENIX under construction, a newly agreed-upon membership model for the RHIC

collaborations has made possible the belonging of individuals and institutions to multiple RHIC experiments (the previous model implied exclusive membership). As a result, the flux of institutions has changed. New institutions to STAR since our last requirement document include Stony-Brook, Darmstadt, UC Riverside, Southern Connecticut State University, NCKU (Taiwan), Lehigh, Abilene Christian University, Rutgers, Cairo/Egypt, Germany/Heidelberg (synergistic collaboration with FAIR/CBM), Hungary/Eotvos, Japan/Tsukuba, China/Fudan, China/Huzhu, and several Indian-based institutes (Berhampur, Patna, Tirupati). This dynamic shows a strong interest in the BES programs and the potentials of our proposed forward upgrade. The geographical distribution is shown in **Figure 30**, accompanied by a detailed evolution over the past workshops.

Figure 30. STAR Collaborators.

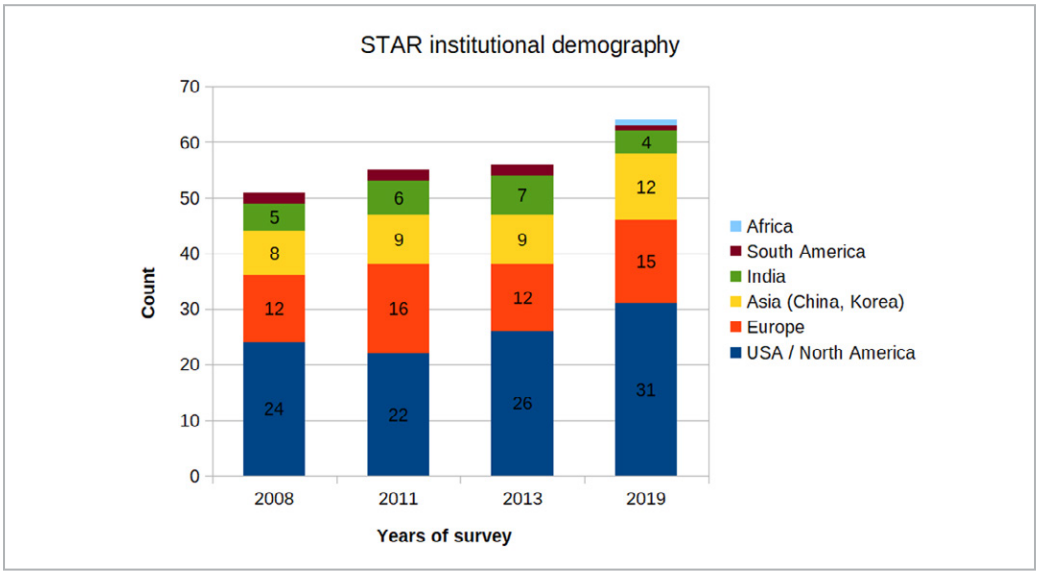

Figure 31. STAR Institutional Demography over experimental lifetime.

Our collaborators are generally encouraged to remotely log into the core facility at BNL RHIC and ATLAS Computing Facility (or RACF) located at BNL as STAR's Tier-0 center. A few collaborators, constituting the core of the efficiency correction production (aka "embedding"), routinely log into the NERSC/Cori HPC center that currently serves as a Tier-1 center for STAR. Emergency data production at NERSC/Cori has also been demonstrated as feasible and could very well occur on a case-by-case basis. The implied model while running at NERSC is "produce and bring the data back" (to the Tier-0 center, that is back to the RACF/BNL facility). This approach is a dramatic paradigm shift for the collaboration as previously, the NERSC / Parallel Distributed Systems Facility (PDSF) center also served as secondary analysis center (but had no real-data production capabilities) and hence, many data sets were brought to/from NERSC/PDSF for analysis support. While this has not yet crystallized, STAR is looking into shifting its infrastructure support from the now obsolete PDSF facility to a nascent LBNL Institutional Cluster. If successful, our analysis requirements at this center would not change compared to previous (PDSF) years. Modest workflows are also run at Dubna, usually constrained to data reduction processing that requires little network bandwidth. Any data transferred to facilities supporting data production workflows such as the one at NERSC or Dubna would use GridFTP as a base tool. STAR is at this moment equipped with four Grid gatekeepers fully instrumented with the OSG software stack. At least two of those are reserved for data transfers and provide a sustainable data-transfer rate for the foreseeable future.

Additionally, the introduction of a new compact format for analysis (aka "picoDST") may very well mean that data sets can be transferred on the WAN to remote institutions. However, it is too early to tell where and when this will happen (it will highly depend on the institution's own strategy and resources).

5.7.4 Instruments and Facilities

The BNL RHIC, and RACF hosts all RHIC experiments. The core operation and role of the facility is to provide the core CPU computing cycles for most of our user analysis needs, the whole of data reconstruction, support for data calibration, data reduction, databases, and some local need for sparse simulations.

The STAR experiment and apparatus computational needs are satisfied by resources co-located at the experimental facility. The STAR group within BNL's Physics Department is charged with the proper operation and ongoing improvement of the STAR apparatus and ensuring the reliable collection of the massive amount of data the detector generates, transformation of that raw data into information useful to researchers, and the transfer of that information to long-term storage systems for subsequent analysis by the scientific community. This requires a complex network of special-purpose devices and computers collectively known as the STAR Data Acquisition System (SDAS) Subsystem. SDAS is operated by STAR operations personnel but remote access to SDAS complies with BNL's Cyber Security Program Plan. In effect, authentication relies on a common infrastructure (but SDAS runs its own set of network services and information service data caching). During the data-taking campaign, it is not unlikely to have detector experts connecting remotely to SDAS. "SSH" is used in this case and its path goes through well maintained and controlled gateways. Remote monitoring of the shift operations is also accessible via web interfaces (https being reverse proxies).

During data taking, the STAR DAQ system streams data to a cache space spread over 14 nodes known as a "buffer box" (nodes collecting and aggregating the data into streams and files) for a total of 210 TB of disk space. In this configuration, and depending on the DAQ rate, assuming a maximum of 2,000 MB/sec rate (theoretical limit for the iTPC system), STAR would be able to hold its ground for ~24 hours without network connectivity. At the standard data rate (1,500 or less during the BES data acquisition time), 48 hours is possible. Currently, the data are pushed from the experimental facility to the High-Performance Storage System (HPSS) tape archiving system using 4x10 GbE lines spread over two redundant switches. In the most data-intensive high-luminosity years, we expect up to 3x10 Gbs Ethernet connections to be needed to handle streaming data movement. As the data are acquired, local (to SDAS) processing handles the High-Level Triggering (HLT running on multicore nodes) and makes event characterization-like processing and triage. The events are then sent to "streams" that may uniquely identify a physics of interest (for example, events containing an Uspillon candidate may be directed to an "upsilon stream"). A large fraction of the data are taken as "minimum bias" Through our data acquisition

process, event counters are kept for later reference (and cross-section, trigger-efficiency evaluation). A fraction of the data are analyzed online, but a second-pass processing (described in the next section) takes care of the bulk of the quality assurance process.

Our biggest data set to date has been in 2016 (a large Au+Au data set in support of the Heavy Flavor Tracker program) and was of the order of 7 PBytes in a single year and 1.5-million files (the file size is highly dependent on the event selection at trigger level).

Our projected data set sizes for years 2019 to 2025 are shown in **Table 8.**

Year	Species	Number of events [B]	Additional [PB]	Additional [PB]	Additional [PB]	Total Space Needed [PB]	Needed [PB]
2019–2020	Various Au+Au BES-II	2	0.45	0.32	0.05	14.14	3.03
2021	Various Au+Au BES-II (sTGC)	1	0.24	0.17	0.03	14.31	3.06
2022	500 GeV p+p	7	6.79	3.71	0.57	18.02	3.63
2023	200 GeV Au+Au	9	15.3	10.71	1.67	28.73	5.30
2024	200 GeV p+Au (or Al)	7	5.12	3.45	0.54	21.47	4.17
	200 GeV p+p	7	4.76	3.33	0.52	24.81	4.69
2025	200 GeV Au+Au	12	20.4	14.28	2.23	39.09	6.92

Table 8. STAR Data Set Size Projection.

The last columns show the total space needed for holding the data sets in a micro-DST and picoDST scenario, respectively. The picoDST space is indicative of the storage needed on disks (and to be possibly partially transferred to other facilities for analysis). Currently, STAR has ~10 PBytes of storage in XRootD space, 2.5 PBytes of central storage, and 500 TB of user analysis space. Within our projections and our current space allocation, it is clear that only the picoDST scenario is viable (the HPSS storage would need to accommodate for the added sizes from columns four to five).

5.7.5 Process of Science

The STAR computing resources available for data processing currently consist primarily of the resources available at the RACF, where at most 14,000 CPU slots are allocated to the data production campaign and the handling of data calibration tasks. The rest of the RACF slots are reserved for user analysis jobs, with the overflow directed to opportunistically utilize unused resources from other experiments by sharing resources. This new model (aka "shared pool") vastly increases the number of slots reserved for user analysis and as a result, other production workflows are often run expanding beyond the 14,000 slots target (for example, data transformation workflows are run routinely to regenerated picoDST from micro-DST). As the shared pool has been adjusted to allow for runtimes of three days, in fact, it is also possible to extend some of our real-data production workflows to the pool.

While the number of slots used by the data production and calibration workflows is ~14,000 on average across the year, this number fluctuates and is lower during RHIC run time (when additional resources are taken for near real-time incremental calibrations) and increases just after major conferences in the field (when user's jobs slightly deplete). During the run, as data are moved from SDAS to HPSS, a process known as "FastOffline" immediately recovers the data (most likely from HPSS cache and not tape) and submits them for local processing. The results of this workflow are used for quality assurance and feedback to the experiment and complement the real-time online monitoring process. Full event reconstruction using the latest code and calibration constants is used for this process, producing a "best" subsystem-independent known calibration state by the end of the run. Special data productions are requested by subsystem experts throughout the run to verify the quality of their subsystem in regards to others. Finally, at the end of the run, additional calibrations are made that consider multiple subsystems (fine tuning of relative alignments, distortions, and calibration such as the energy-loss

procedure known as "dEdx" are part of this processing pass). Once calibration activities are completed, the experimental data collection campaign may begin.

In the early years of RHIC, STAR had the resources to produce a given run's data 2.2 to 2.4 times within a year of an experiment. Severe resource constraints for the past several years have forced the STAR Software and Computing team to first reduce this number to a conservative 1.2 to 1.5 passes, and then remove the constraint that the 1.5 passes fit with the one-year schedule. Processing time needed for the calibration passes described previously varies from one and a half to three months depending on whether the collision system is well known and all calibrations follow expectations, or if the studied system is novel (i.e., isobars in Run 18 and beyond which required a "blind procedure" and extreme vetting to be defined) and/or presents a unique new challenge, such as unusually high luminosity, additional distortions, or inclusion of a new detector subsystem. Typically, the scheduling of the production of quality high-priority physics data sets may occur as soon as two months after a given run ends. Then entire (prioritized) data set production would be launched: the data produced would be saved to BNL's HPSS. A copy of the picoDST could be left (as data are produced) available on disk for the analyzer's access.

While our workflow efficiencies are very high, for planning purposes, we typically use a 93% compounded efficiency where the compounded efficiency is defined as the product of the resource utilization efficiency (our ability to saturate the available slots) times the application's run-time efficiency. As per previous estimates, the facilities can support 14,000 processing power slots per year. Accordingly, using these assumptions, estimated processing times are shown in **Table 9.**

Year	Species	Total Reco Time, 1 Pass 2017 Farm [months]	1.5 Passes Equivalent (needed)	Required (calib and R&D added [months]
2019–2020	Various Au+Au BES-II	0.77	1.16	2.66
2021	Various Au+Au BES-II (sTGC)	0.39	0.58	2.58
2022	500 GeV p+p	10.64	15.95	17.45
2023	200 GeV Au+Au	11.34	17.01	
2024	200 GeV p+Au (or Al)	2.81	4.22	10.36
	200 GeV p+p	3.1	4.65	
2025	200 GeV Au+Au	15.12	22.68	24.18

Table 9. STAR Computation Time Projection.

In 2022 and 2023, experimental data generated are predicted to exceed computing resources allocated to STAR. STAR will be making use of additional external resources to fill this shortfall in capability, since adding STAR-internal resources for this gap-period is not economical. Our considered remote resources are primarily used to provide a boost to the data production workflows: for example, lower priority data sets could be moved to sparse resources (remote sites) with predictable estimated time to delivery while the resources at major facilities (including leadership facilities) may be used to provide additional production passes (emergency production on the approach of conferences).

Possible external resources that may be employed during this gap-period include either the computational cluster available at the Dubna institution or the HPC resources available at NERSC.

In 2016, STAR demonstrated the very first high-efficiency use of HPC resources for real-data production in the nuclear and particle physics communities. This work was presented at the CHEP 2016 conference and showed an end-to-end workflow efficiency nearing 95%. The STAR collaboration was subsequently provided with an allocation of 25 million hours. Further attempts to push production workflow showed an improved efficiency in excess of 99% (well matching the RACF local efficiency as reported in doi :10.1088/1742-6596/898/8/082023). Our projected throughput at the time required the transfer of ~100 TB/week and ~15 TB/week of reconstructed data

back to BNL or approximately a continuous transfer rate of 200 MB/s in steady state production (though we made intense use of data caching on both ends to be able to "ride" network performance fluctuations).

NERSC/Cori has also shown to be suitable for our complex simulation workflow known as "embedding," where simulation events are embedded into real events and the efficiency corrections thereafter determined. All embedding workflows continue to be carried out at NERSC (on Cori) but a significant change has been that the result of embedding is brought back immediately to BNL for quality assurance (the quality assurance process, done by analyzers, was previously done at NERC/PDSF: the HPC cluster is not practical for user processing). From 200 to 300 TB of embedding data production is brought back to BNL per year. The transfer of the data was shown to reach 300 MB/sec using standard GridFTP tools and more typically, 1 TB of data takes about 30 minutes to one hour to be brought back to the RACF (on a standard shared bandwidth day).

For the data processing workflows we are running at NERSC, a network speed of ~ 700 MB/sec we demonstrated we currently have is enough to sustain the full cycle of data production (300 MB/sec + 200 MB/sec + margin). The margin is present in case of large production needs where a balance of caching and "streaming" mode may be needed. With the challenging outer years of our run plan (2022 and 2025 especially) and in order to handle "boost" production needs and accelerate science delivery, it is likely STAR would request another allocation similar to the one made in 2017 towards the NERSC/Cori facility.

Present to two years, the STAR collaboration will continue and complete the BES program. This program will not generate a major amount of data; the collaboration will make full use of the recently added new detector system and harvest its science to the fullest. We expect little changes in data rates and needs. However, the flow of picoDST to remote facilities will become clearer.

Next two to five years, the forward spin program will be enhanced by the forward detector systems. Proton on proton collision will be part of our accumulated data sets and likely be the largest accumulated to date. The process of science (production, local needs) will not change but a boost of production need in 2022 (via a temporary allocation, for example) is likely. The two DTNs provide a 2x10 GbE link, sufficient to sustain a 700 MB/sec data production workflow and encompass the needs for our data and embedding production.

Beyond five years (beyond 2025), we do not expect STAR to continue to run. The last data set taken by the STAR collaboration will be in 2025, and we expect the collaboration to make the best of it and accumulate as much data as possible (and the best heavy ion reference sample to date). To absorb the overflow of production needs, another allocation may be requested at this time. No network requirement changes are expected.

5.7.6 Remote Science Activities

In the previous section, we described our use of the NERSC/Cori leadership facility. All workflows carried to/from Cori are done using the OSG software stack. Additionally, simulation and event generation may be run opportunistically on the OSG. Today's workflows are highly simplified by the wide use of containers (Docker) and CVMFS for software distribution. CVMFS service is hosted at BNL (as its stratum 0 or main copy or "truth"). Using containers and CVMFS allows for seamless software provisioning across all facilities while validation can happen at BNL using the same environment seen at remote sites. We do not foresee a change to the landscape and remote usage in the next five years. Due to diverse factors, we will not be able to maintain our Dubna site, but it is likely that a resource of similar size would appear.

5.7.7 Software Infrastructure

The data management software of STAR is composed of a file, location, and metadata catalog, a data aggregation tool for local analysis and standard data-transfer tools for remote processing. Our catalog has been in-house developed and shown to scale to a million files with no sign of scalability problems, heavily relying on highly distributed daemon processes keeping data set consistency (each "node" in XRootD for example taking care of its own data, each storage element as well. The aggregation is visible by all users at any point in time). It is the core of many tools in STAR and has allowed "virtualized" access to our data sets (users no longer need to "know"

where the files are but refer to data sets by metadata queries). The data aggregation itself is taken care of by XRootD services (nowadays maintained by the RACF personnel). Data transfers are handled by GridFTP. Users would typically use "scp" or other transfer protocols using transfer gateways provided by the RACF.

Data processing, simulations, and analysis can be performed using the STAR standard framework known as "root4star". This single framework relies on the ROOT package and XRootD plugin is a de-facto component installed along the STAR software. The software stack runs seamlessly on both 32- and 64-bits architectures. Parallelism is for now considered at the event level, but the framework is fully capable of using vectorization (eigen or Vc, the most dramatic gain being in 64 bits mode where 15 to 20% gains were observed).

STAR may collaborate with the newly created Nuclear and Particle Physics Software Group and the RACF to evaluate Rucio data management in the next two to five years. Since STAR already has a working solution, the evaluation of additional capabilities as they become available may have large benefits for all parties (no loss for STAR, chances to consolidate and learn). We will also need to evaluate new transfer protocols (GridFTP may not be maintained much longer) and integrate in our distributed computing workflows (and data placement tools). Beyond this timeframe, it is unlikely to see STAR making dramatic architecture changes.

5.7.8 Network and Data Architecture

We refer to the RACF/SDCC document for details and overview of the campus network topologies. We would like to add a few bullets specific to STAR:

- As previously noted, STAR makes use of GridFTP services for transfer to its main facilities, employing dedicated DTNs for this purpose.
- Currently, STAR is not leveraging the capabilities of perfSONAR. Its facility having converged to a few, large-scale sites, an integrated monitoring has not seen a high level of priority.
- The STAR Counting House (where the "Instrument" is located) belongs to the "Campus Network" while the core of the RACF resources rests on the ScienceDMZ. This brings uptime requirements for both the local network and WAN.

5.7.9 Cloud Services

While an early adopter and tester of cloud processing workflows, the STAR collaboration is currently not making use of cloud services. However, as STAR has a long-standing participation within the OSG and one of the OSG core proposals relates to outreaching to cloud providers, it is not inconceivable to imagine that modest workflow (piggy-backing on research allocations) could happen within the next two years.

5.7.10 Data-Related Resource Constraints

We believe that the network capabilities in place on the WAN at the moment will suffice to satisfy STAR's established data processing needs. STAR has already shown that HPC resources can be exploited for real-data production workflows. The implication is that as usable network connections to leadership facilities need to be as good (if not better) as the capabilities currently in place between BNL and NERSC. Additionally, a possible move from NERSC/PDSF to the LBNL Institutional Cluster resources may force us to readdress and understand better the connectivity and path between BNL and those resources (as any shift in the HPC landscape would).

For the LAN, we would like to note that our previous XRootD model relied on a scaling of storage, network connectivity, and compute power: as dedicated STAR resources and compute node would be added to the facility, their storage would be added to the XRootD data aggregation and the connectivity (in terms of GB/sec) scale accordingly and along the growth of CPU power. The shift to a "shared pool" and separation of compute and storage raise a concern as per (a) the concentration on a few servers (losing one server with a low cardinality has high impact) and (b) their network connectivity need to scale in the several 10GB /sec today and proportionally to the CPU growth. Access to central storage needs to be equally scaling with analysis growth.

5.7.11 Outstanding Issues

5.7.11.1 Understanding LAN and WAN Uptime

We have previously noted (in Section 3) that STAR may survive from one to two days of network isolation. We need to emphasize at this stage that for operational reasons, during data acquisition campaigns, such downtime would not be desirable and may severely impede our science. During run time, quality assurance needs to be carried out on a continuous level. Assuming six hours beam "fill," at least one hour of network access is required to be able to transfer "some" data to HPSS and hence, retrieve them for "FastOffline" processing. This implies a LAN downtime of three to four hours maximum. We are pleased to note that our local networking team has achieved this level of service. WAN access has different requirements. We also noted in previous sections that many detector experts connect to SDAS using SSH. Their interaction may be crucial to the experiment as they may be needed to diagnose an issue, change detector settings via controlled access, or require access for expert data flow monitoring. Beyond the experts, on-call personnel access web-based interfaces providing crucial information on the run sanity. A WAN access interruption of 20 minutes would start to weigh heavily on the operation team but would be manageable (the on-site support team could handle expert instructions via phone call with difficulty and impact upon other operations, while the remote support team would be unable to further support for that duration). A WAN downtime beyond three hours would, however, have a significant negative impact on the experiment as a whole. It is also important to note that STAR has considered "Remote Quality Assurance shifts," off-loading all quality assurance shifts to institutions that may not be able to travel or access BNL in person. Long WAN network downtimes would be detrimental to our plans.

5.7.11.2 Electron-Ion Collider (EIC) Era

While outside the scope of this report, it is worth mentioning that at the dawn of the EIC and with commitments from both BNL and JLab in its vision and program, good networking capabilities and interconnects between those two labs seems essential. No matter the outcome of the site selection, it is likely that the "other lab" would play a significant role in processing the data collected at the EIC, following a model similar to a Tier model. This matter will become clearer at the next ESnet workshop.

5.7.11.3 Data Preservation

The possibility of a full copy of raw data to a secondary facility for the long-term preservation and safety of STAR data has been long discussed, considered but never resolved (funding to save a secondary copy was never allocated). For this process and the current bandwidth and considering an accumulated raw total data set of 72 PBytes (and counting), would require 3.5 years to transfer. This plan seems rather impractical at this stage and likely unfeasible by the end of STAR's lifetime (unless serious commitment is made to fund storage at a secondary facility).

5.7.11.4 Other

The STAR collaboration would like to renew our thanks to ESnet's and BNL local networking teams' assistance with many past efforts. Without the dedication of teams from both sides of the BNL to/from NERSC/Cori exercise, we would have not been able to tune the network to appropriate requirements and to transform an idea to a transformative success story. This is one of the many examples where networking was part of the panoply of "resources" needed for an efficient end-to-end workflow. We would like to further thank ESnet for considering our past feedback (seeing it integrated in the form and format of this workshop especially).

5.7.12 Case Study Contributors

- PI for this document: **Jerome Lauret,** *BNL,* jlauret@bnl.gov
- PI for the experiments, Spokespersons: **Zhangbu Xu,** *BNL,* xzb@bnl.gov and **Helen Caines,** *Yale University,* helen.caines@yale.edu

5.8 Brookhaven National Laboratory — Pioneering High-Energy Nuclear Interaction eXperiment (PHENIX) / sPHENIX

PHENIX at BNL was shut down after 16 years, and BNL is preparing for sPHENIX upgrades to complete by 2023. Increases in data volume are expected, but the general workflow and use of local and remote resources is expected to remain the same. Current work has not stopped, and there are development and design activities to leverage more communal computation resources such as the OSG.

5.8.1 Discussion Summary

- sPHENIX is the planned next generation of PHENIX.
- Despite shutdown, activity on reconstruction and analysis of PHENIX data remains. PHENIX data exist as a collection of 10GB files. Processed data can be 1k to 40k files and raw data can be up to 300k files. Data sets size depends on the collision system and the energy. Au+Au (Gold) collisions result in the largest data volumes since they involve the most energy. The largest data set from PHENIX was 3PB taken in 2014.
- The raw data is reconstructed using a production-pass algorithm which involves CPU-intensive tasks. The output of this production pass is suitable for analysis (DST) and contains enough information for final calibrations, which are applied on the fly during user analysis. DSTs are typically 25% of raw size.
- sPHENIX start is expected to start in 2023 and will run for three years during 22-week RHIC operation periods.
- The currently expected data set sizes are 75PB in year 1, 143 PB in year 2, and 205 PB in year 3. If the running time gets extended to five years, year 2 and 3 would be repeated, resulting in similar data set sizes.
- If sPHENIX is used as a basis for a day one EIC detector, an average signal rate of up to 100Gbps is estimated. This is less than the anticipated rate for Au+Au running in 2023, which is around 175Gbps.
- Some of the sPHENIX simulations are foreseen to run opportunistically, using, e.g., the OSG computing resources. Given that sPHENIX simulations are purely cpu bound, the network bandwidth needed for this effort is expected to be small.

5.8.2 Science Background

PHENIX was a first-generation experiment, and its results were essential for the discovery of a new state of matter QGP. sPHENIX is a second-generation experiment, designed to understand the inner workings of the QGP.

PHENIX (and in the future sPHENIX) are housed in the 1008 complex at one of the interaction regions of the RHIC ring. The data from the collisions are recorded locally in the counting house and then transferred to the RHIC RACF for long-term storage in the HPSS storage system and processing. The local buffering capacity of three to four days allows the experiment to send data at an average rate rather than peak rates, which reduces the requirements on the RACF in terms of bandwidth and i/o capabilities. It also insulates PHENIX and sPHENIX from the possibility of service interruptions. The size of the data sets depends on the collision system and the energy. Au+Au runs at top energy, resulting in the largest data volumes. The largest data set from PHENIX was 3PB taken in 2014. For sPHENIX it is expected to reach 200PB for the Au+Au run planned in its third year of operation.

The raw data are reconstructed in a production pass which involves CPU-intensive tasks like tracking. The output of this production pass is suitable for analysis (DST) and contains enough information for final calibrations, which

are applied on the fly during user analysis. In addition to the raw data, all DSTs are stored in the HPSS storage system. In the past, PHENIX has sent whole raw data sets over the network to remote facilities (Japan, Vanderbilt and France) to run production. However, this was not deemed to be efficient and, given the increased capabilities of RACF, data processing is now done exclusively at RACF. But large-scale simulations which do not require the transfer of large amounts of data have been run on the OSG, and we will preserve this capability. sPHENIX is looking at the same computing model where the storage and the processing are done in RACF.

The versatility of RHIC in terms of beam energies and species leads to a large number of different data sets, most of which are actively analyzed. PHENIX stores all DSTs for analysis going back to 2003 on distributed storage (for a total of 8 PB stored in a dCache instance) for immediate access. PHENIX employs a coordinated analysis scheme — the Analysis Taxi — where users submit their analysis code, which is then checked for errors and run centrally over the requested data set(s). This approach enables collaborators to spend more time on their analysis rather than dealing with computing issues when running tens of thousands of jobs. The current sPHENIX run plan foresees only three different but very large data sets: p+p, p+Au, and Au+Au at 200GeV. Nevertheless, those data sets need to be accessible simultaneously, and we will employ an Analysis Taxi approach here as well.

There are no restrictions on collaborators who want to export/import data to/from their home institutions, and collaborators can archive data in HPSS in their own dedicated area. The volume of this data tends to not exceed the few GB range.

5.8.3 Collaborators

PHENIX and sPHENIX are separate collaborations (the sPHENIX collaboration was formed in December 2016). PHENIX has more than 500 members from 75 institutions from 14 countries. sPHENIX currently has 250 members from 77 institutions from 13 countries. This is expected to grow. Substantial overlap exists between member institutions for PHENIX and sPHENIX, but sPHENIX also includes members of the STAR collaboration and the LHC experiments. Every PHENIX or sPHENIX collaborator can have an account in the RACF computing facility, which enables access to the collaboration resources. Data analysis is typically done via our Analysis Taxi but collaborators are entitled to run their own processing.

Teleconferencing tools and low-latency networking to support them are essential for our daily communication.

5.8.4 Instruments and Facilities

The PHENIX experiment has been decommissioned, but the collaboration is actively analyzing the data, and data reconstruction is still ongoing. The raw data of a large Au+Au data set are in the 2 to 3PB range. The resulting DSTs, which are the input of the user analysis, are about 25% of the raw data size. We keep the file sizes at around 10GB, which is optimal for our tape access and network transfers patterns. With the previously listed data set sizes, our DST-type data sets for analysis contain between 1,000 and 40,000 files, and the raw data up to 300,000 files. The Analysis Taxi running at RACF is optimized for latency. We strive for a turnaround time of less than 24 hours, so users have the results of their analysis by the next day. It is expected that the current usage will continue for the next two years and then slowly taper off over the following years as ongoing analyses are finalized and sPHENIX starts to take data.

The sPHENIX experiment is scheduled to start taking data in 2023. The current run plan covers three years of running. The expected data set sizes are 75PB in year 1, 143 PB in year 2, and 205 PB in year 3. If the running time gets extended to five years, year 2 and 3 would be repeated, resulting in similar data set sizes (143PB and 205PB). The peak data rates that need to be handled by the HPSS system, accounting for the expected averaging by the local three- to four-day buffering — "the best week" — would be 110 Gb/sec in year 1/2 and 175Gb/s in year 3. The expectation is that the archiving and processing of the raw data is done by RACF. Simulations, on the other hand, will use external resources if not enough capacity exists in RACF.

In the far future, if sPHENIX turns into a day one EIC detector, the expected signal rate is 100Gb/s where detector noise would have to be added. But even if detector noise is added, it is expected to stay well below the peak rate in Au+Au running during year 3.

5.8.5 Process of Science

Analysis work in PHENIX is required to be associated with a Physics Working Group (PWG). The PWG has local resources at the RACF for its members, which it manages largely autonomously within the different analysis projects. Most local and remote collaborators draw on the resources of the PWG in question. sPHENIX has "topical groups" for this purpose. Their task is to prepare the tools needed for a given high-level analysis. Within the next two years it is foreseen that they will have dedicated resources (mainly disk space) which are available to their members. In the long term these topical groups will turn into PWGs which will manage "their" resources autonomously. But just like in PHENIX, sPHENIX collaborators will always have the ability to pursue their own analysis on collaboration wide resources.

5.8.6 Remote Science Activities

Some of the sPHENIX simulations are foreseen to run opportunistically, using, for example, the OSG. The job submission system for sPHENIX will be agnostic where those jobs are run. With the advent of containers and cvmfs basically available on most sites, running on remote sites became a lot simpler. Given that sPHENIX simulations are purely CPU computed (currently running a central heavy ion event through a full GEANT4 simulation of sPHENIX takes 45 minutes) the network bandwidth needed for this effort is expected to be small. The simulation activities are expected to ramp up over the coming years as more collaborators become involved in the preparations for the analysis of sPHENIX data.

5.8.7 Software Infrastructure

Facility data management tools are written by the collaboration and RACF. For sPHENIX, workflows will employ open-source data management tools mainly from the LHC community. Globus is being used to transfer simulation output from LLNL to RACF. There are no plans to use commercial software.

5.8.8 Network and Data Architecture

Please see the RACF document

5.8.9 Cloud Services

The sPHENIX facility does not use cloud services and does not plan to use them in the future due to cost.

5.8.10 Data-Related Resource Constraints

The expectation is that for sPHENIX the data-intensive processing will be done in RACF. Needs from remotely running simulations are met by the currently available network bandwidth.

5.8.11 Case Study Contributors

- **Martin Purschke,** *BNL*, purschke@bnl.gov
- **Jin Huang,** *BNL*, jhuang@bnl.gov

5.9 Compact Muon Solenoid (CMS) Heavy Ion Experimentation

The CMS Heavy Ion (CMS-HI) group is closely tied to the LHC and the HEP efforts. Many needs and requirements at Vanderbilt University are shared between these programs as a result. Given the current shutdown period for LHC, activities are being limited to reprocessing and simulation using existing computation and storage distributed globally in the Worldwide LHC Computing Grid (WLCG). After the upgrades, data volumes will increase, forcing network, storage, and computational upgrades. Work is also being done on software that manages data movement to take more advantage of streaming workflows and new computational resources provided by ASCR computing facilities.

5.9.1 Discussion Summary

- CMS-HI is closely tied to the LHC and the HEP effort. Many needs and requirements at Vanderbilt are shared between these programs as a result.
- The LHC is currently in shutdown for upgrades, but work is being performed. Simulation and reconstruction are common and still occurring around the WLCG. The LHC will resume operation in 2021, and heavy ion collisions should resume their schedule of ~24 days of recording data in November/December in the winters of 2021,2022, and 2023. The LHC will shut down in early 2024 for upgrades and will be in service again from 2027 to 2029.
- Experimental data come in two varieties: raw and Analysis Object Data (AOD).
- LHC/CMS produces raw data sets and performs reconstruction on WLCG computing resources The output of this process is AOD. Both are stored at CERN on tape.
- Data are then transferred via ESnet to Fermilab (FNAL), which houses the second copy of raw and AOD, and to Vanderbilt, where only the AOD files are stored.
- The AOD format is one-third the size of the raw data.
- Reconstruction of experimental results (e.g., reading of raw, and creating a new AOD set) can occur at a later date, and typically happens when:
  - Different calibrations are being experimented with.
  - New algorithms are being tested.
  - Sites are re-populating (or commissioning) storage.
- The pipeline from the detector to Vanderbilt's storage is extremely CPU limited. The 2018 data-taking period is demonstrative. Even using 50,000 CPUs at CERN to perform the reconstruction, it took nearly a month to reconstruct the 6PB of raw data to produce 2.7PB of AOD outputs. These AOD outputs were produced and immediately transmitted to Vanderbilt/ Fermilab at an average rate of 1.5GByte/sec.
- CMS and other LHC experiments are migrating away from a data-tethering model where compute and storage resources are linked. The use of streaming tools (XRootD, etc.) relies more on bandwidth as a resource that can be used to decouple the storage and processing requirements. This results in more efficient use of available CPU resources, at the cost of larger data transfer band width requirements. As a result of this, CMS is exploring a similar model where two large storage sites (and a deployment of caches) can be utilized for storage needs, with computing occurring at different analysis locations.
- CMS is working with NERSC on these models, in particular getting XRootD integrated into worker nodes within their infrastructure. Typically, worker nodes within an HPC do not have WAN access for security and performance reasons. Use of local caches to manage data

flow between compute resources and the outside world is helping to make non-data tethered workflows possible.

- CMS-HI, and in the greater sense universities that participate in LHC-related experiments, typically do not use cloud resources. Power, cooling, and networking are typically subsidized heavily in a university environment. Thus using a cloud resource is often far more expensive than local resources.

5.9.2 Science Background

The CMS is one of four experiments taking data produced by the LHC accelerator located at the CERN laboratory near Geneva, Switzerland. For the most part, the LHC operates in the proton-proton (pp) collision mode designed to explore the frontiers of HEP. The announcement on July 4, 2012, that a Higgs-like boson particle had been discovered is one of the spectacular successes of this HEP research at the LHC. While several months of the year have been devoted to p-p physics, the LHC is also capable of colliding heavy ion nuclei, such as lead-on-lead (Pb-Pb), or even asymmetric collisions, such as protons-on-lead (p-Pb). In each of the past three years, the heavy ion running has taken place for a period of about 24 days after the proton running has completed.

The goal of the heavy ion research program is to create and study the properties of a novel state of matter called the QGP. This is a state of matter predicted to be created in heavy ion collisions where the produced temperatures and densities are so large that the normal nuclear matter constituents of protons and neutrons melt into their composite quarks and gluons. This phase is thought to be the state of matter persisting in the early universe until a few microseconds after the Big Bang. One of the early surprises in this field of high-energy NP was the discovery that the QGP behaves like a strongly interacting liquid with accompanying hydrodynamic behavior, such as flow, and not like a weakly interacting partonic gas, as had been expected. This discovery was made by colliding gold nuclei (AuAu collisions) at the RHIC at BNL.

The LHC heavy ion research program also analyzes p-p and p-Pb collisions as so-called reference data for the more complex Pb-Pb collisions. Originally it had been assumed that the QGP would not be formed in the simpler systems, and that QGP effects in the Pb-Pb data could be disentangled from normal nuclear matter effects by scaling up from the p-p and p-Pb measurements. However, close analyses of the p-p and the p-Pb data taken in 2012 and 2013 led to a further discovery: that there is a highly correlated pattern in some of the produced particles ("the ridge") reminiscent of the flow-like behavior seen in the RHIC AuAu and the LHC Pb-Pb collisions. These surprising results were further studied during the 2015 to 2018 period of LHC operation (Run 2).

Run 2 of the LHC recorded a significant amount of both Pb-Pb and p-Pb data. Additionally, the LHC accelerator division unexpectedly provided several hours of Xenon-Xenon (XeXe) collisions "for free" (i.e., the beam time did not subtract from the heavy ion program's budget).

In all cases, these raw data were first reconstructed at CERN's large compute farm to produce new data sets containing physics-relevant quantities collectively known as AOD. Both the raw and the AOD are first stored on archival tape at CERN, then transferred across the Atlantic to Fermilab for archival storage. Simultaneously, the AOD is transferred to Vanderbilt for access by heavy ion researchers. This reconstruction process is occasionally repeated in order to produce AOD with updated calibrations or reconstruction algorithms.

The pipeline from the detector to Vanderbilt's storage is extremely CPU limited. The 2018 data-taking period is demonstrative. Even with using 50,000 CPUs at CERN to perform the reconstruction, it took nearly a month to reconstruct the 6PB of raw data to produce 2.7PB of AOD outputs. These AOD outputs were produced and immediately transmitted to Vanderbilt/Fermilab at an average rate of 1.5GByte/sec.

5.9.3 Collaborators

Many heavy ion research groups exist in the United States at the University of Kansas, University of Maryland, MIT, Purdue, Rice, Rutgers, University of California Riverside, University of Illinois at Chicago, and Vanderbilt. There are also important overseas CMS-HI groups in France, Korea, Hungary, New Zealand, and Russia. The overwhelming majority of the analysis-level data are stored at Vanderbilt, with a small fraction of the data stored in France.

5.9.4 Instruments and Facilities

The data are recorded by the CMS detector at CERN. Two periods of upgrade of the detectors/accelerators are scheduled: (1) the present to spring 2021 and (2) spring 2024 to winter 2026 (or possibly 2027). During these shutdown periods, no new data will be recorded, and afterwards the detectors will be upgraded to record more granular data per second. The effective data rate, however, will be largely constrained by the frontend data acquisition hardware, storage constraints, and CPU hours required to reconstruct the data, so only modest increases of the annually recorded data are expected in the next decade.

Outside of the detector itself, CMS-HI primarily uses the compute resources at CERN, the tape archival system at FNAL, and the analysis facility at Vanderbilt. There will be some moderate additional use of compute facilities, mostly in the United States.

5.9.5 Process of Science

There are roughly two distinct phases of data production and analysis. These are handled by CMS's production and analysis frameworks. The production framework is concerned with producing the raw and AOD-tier data formats, while the analysis framework is used by researchers to produce science from the AOD-tier data. Each of these frameworks, on average, consumes 50% of CMS's available CPU resources (approximately 250k CPU cores at steady state).

The production framework is centrally controlled, scheduled, and managed by CMS management. When data are being recorded from the detector, the 40MHz bunch crossing rate is reduced to O(KHz) of events by the CMS triggering system, consisting of an FPGA-accelerated Level-1 Triggering (L1T) system. This selects O(100KHz) of events followed by a 15,000 core HLT, which further reduces these 100KHz of events to 1KHz. The events that pass the trigger are then "repacked" into a raw data tier (this repacking process "pivots" N independent detector streams into a format with events). These raw data are then stored on a disk buffer and tape archival storage at CERN. Additionally, these raw data are enqueued for transfer from the CERN disk buffer to the FNAL tape archive storage (all CMS raw data are stored on two independent tape archive systems to reduce the chance of a catastrophic loss of data).

To reduce these raw data into the analysis-level AOD format, an extremely CPU-intensive process known as "reconstruction" is performed. If this happens while the detector is recording data, it is known as "prompt reconstruction." Otherwise, it is referred to as "re-reconstruction." Reconstruction does several things, including applying best-known calibration and alignment parameters and "connect the dots" of particles traveling through detector elements to produce a list of particles (electrons, muons, photons, etc.) and their trajectories through the detector. It is these trajectories (known as tracks) and information from calorimetry that are used by physicists to perform their research. An additional benefit of this reconstruction process is that the AOD format is one-third the size of the raw. Since the raw is not needed day-to-day, this means the disk space requirements of the change to analysis facility disk space requirements can be greatly reduced as well.

For both prompt analysis results and reconstruction analysis of previous experiments, the overall workflow is the same: the production framework tracks the location of the raw files on disk (file locations on tape are not considered, since they are inaccessible from any CPU resources). The difference is where the raw files are coming from. In the case of prompt reconstruction, the raw files are born at CERN and moved by the data

management framework to site(s) with CPU resources. In the case of re-reconstruction, tape recall requests are issued to CERN/FNAL, and as raw files are copied to disk buffers, the data management framework moves them from disk buffers to the processing site(s). As files arrive at processing sites, the production framework injects jobs to process the raw files and produce new AOD files. These AOD are initially stored at the processing site where they are produced and are then asynchronously transferred to FNAL/CERN for archival, with an additional disk copy transmitted to Vanderbilt.

There is one important caveat about this workflow, from a network design point of view. When these data are transferred from one site to another, the data management framework treats these requests as "site X needs file Y," not "site X needs file Y from site Z." For example, if a file is produced at CERN and is requested to be moved to both FNAL and Vanderbilt, these transfer requests happen asynchronously from each other. The file could first be transferred to FNAL, then on the next processing cycle for Vanderbilt's input queue, the data management framework could choose to move the file from FNAL instead of from CERN. These "non-source optimized" transfers can contribute to a significant amount of traffic, particularly in the case where one site's storage was temporarily inaccessible.

Two different modes of operation exist for these reconstruction passes. The first method is to perform all the reconstruction at CERN. This is operationally preferred by CMS management because CERN has both enough disk space to stage all the raw data and enough CPU resources to complete the reconstruction on a reasonable timescale. The second mode is to divide the raw into subsets and scatter it to 10 to 20 sites, who each perform a portion of the reconstruction. This obviously incurs the extra delay to transmit these raw data to many input sites, as well as the operational difficulty where existing data sets on these sites have to be evicted in order to free up the space for both the raw and the temporary storage for the AOD, etc. CMS is currently performing a re-reconstruction of the 2018 data using the latter method, and **Figure 32** shows the inbound bandwidth to Vanderbilt as accounted by the CMS data management system.

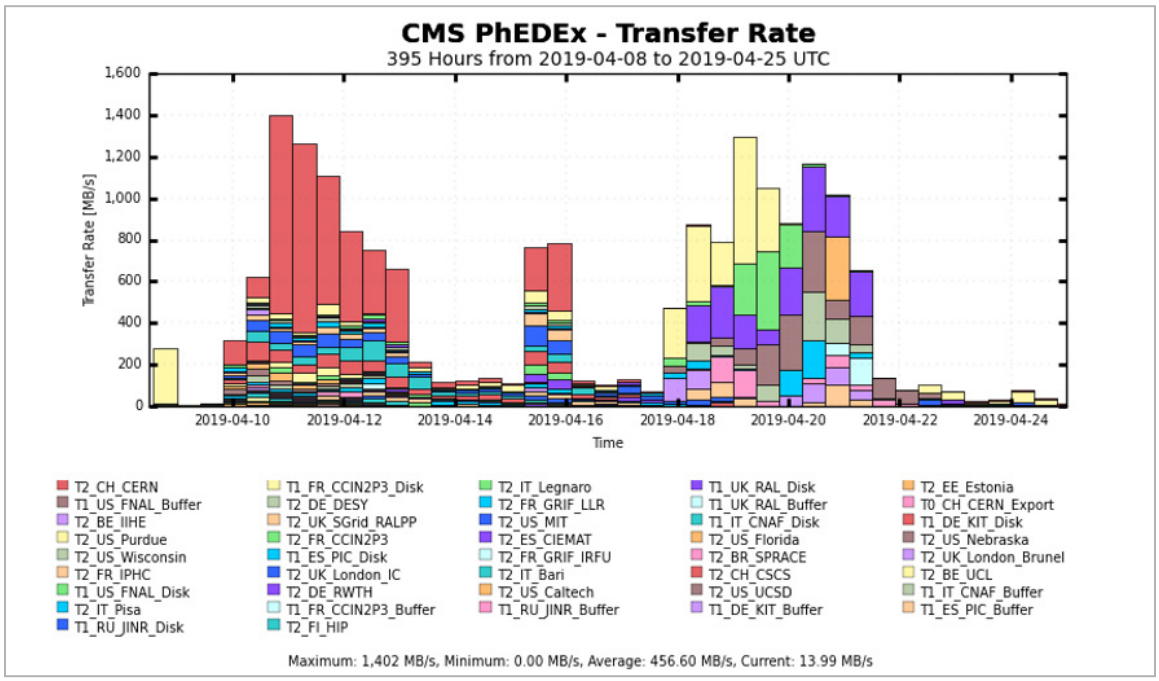

Figure 32. Vanderbilt University CMS-HI T2 Data Transfer Performance.

There is the initial burst of data from CERN of raw data to Vanderbilt (Vanderbilt was chosen as a site to perform some of the reconstruction), then the subsequent trickle of AOD data from many sites as they completed their portions of the reconstruction. Unfortunately, the data management framework does not allow us to disentangle the data flows from different processing tasks, so a similar plot for either CERN or FNAL are not representative of CMS-HI's needs (it includes a very significant amount of p-p data movement in preparation for a p-p re-reconstruction pass).

For the three time periods requested, we expect the following:

- **Present to two years:** no planned activities after the current re-reconstruction pass. The LHC is shut down.
- **Next two to five years:** the LHC will resume operation in 2021, and heavy ion collisions should resume their schedule of ~24 days of recording data in November/December in the winters of 2021,2022, and 2023.
- **Beyond five years:** the LHC will shut down again in early 2024 for another long shutdown and resume operation in 2027 and continue to 2029. The five-plus-year schedule has a chance of slipping by a year depending on the work that occurs during the previous long shutdown).

The analysis picture is much more chaotic than production and depends heavily on individual researcher's priorities/needs (e.g., conferences, hints of new science from competing experiments, etc.). In general, the CMS-HI computing model assumes that the centrally produced analysis-level data are stored at Vanderbilt. Researchers run jobs that either directly produce results from these data or, more commonly, these jobs produce smaller, more highly processed data sets which analyzers use daily. These user-produced data sets O(100TB) and are transmitted to other analysis facilities, primarily MIT and CERN's user analysis facilities.

Previously CMS used a data-tethered model where jobs were required to run where their input files were located. With the increase of bandwidth available at large university research computing resources, CMS has begun to decouple this data model and allows some subset of jobs to run at sites with non-local input data. Currently, jobs that have been idle at a site for more than six hours are re-brokered by the central job management system to specify that they can run at additional "nearby" sites (nearby, in this case, is specified by hand). So, for example, if there is sudden demand to access data at Vanderbilt, jobs that remain idle for a long time will be reconfigured to specify that they can run on any site in the United States, the assumption being that since the other large U.S. research institutions all have 100Gigabit network connections and have relatively low latency to each other, remote file access should incur a minimal performance penalty compared to local file access. **Figure 33** is a representative week of remote file accesses from Vanderbilt, of which the overwhelming majority is heavy ion data being accessed from jobs running at remote sites.

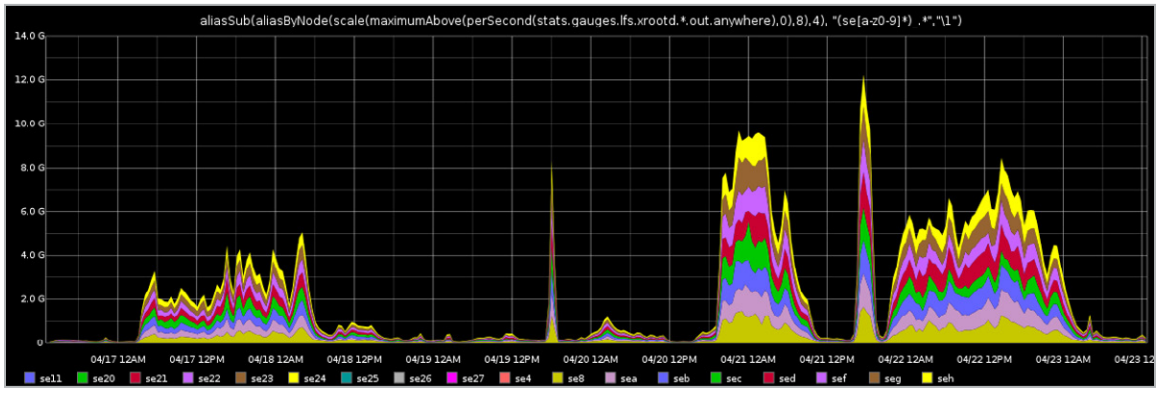

FIGURE 33. Vanderbilt University CMS-HI T2 File Access Performance.

5.9.6 Remote Science Activities

CMS is devoting significant resources to integrate nationally funded leadership-class facilities in its job submission infrastructure. One limitation, however, is the difficulty in staging our large data sets into these supercomputers. The issue is not necessarily the bandwidth to/from the site, but the difficulty in getting our data past the perimeter to the site local storage in an effective way. This is compounded for heavy ion data, since its input data sets are larger than regular p-p data. Currently, CMS is making most use of these facilities for simulation data, which require little-to-no input data, but a lot of CPU time to produce simulation data sets.

The community of various CMS stakeholders are working together to resolve these issues. If successful, these facilities would attract usage from CMS-HI and enable more freedom for when re-reconstruction passes occur.

5.9.7 Software Infrastructure

Currently, all internet-facing storage interfaces in CMS expose both GridFTP and XRootD protocols. This allows the disparate storage technologies in use at different sites to interact with common interfaces. File transfers between sites use GridFTP as the base protocol, with FTS (https://fts.web.cern.ch/) performing the actual transfers and our local data management software scheduling global movement of data (e.g., subscription of data set X to site Y). XRootD is currently used for remote file access.

During 2020 to 2021, CMS will deprecate the usage of GridFTP in favor of XRootD, meaning that in the medium term, XRootD will be the only protocol used for both remote file access and data transfers. Recent updates of XRootD support a lightly-modified variant of the WebDAV protocol instead of the proprietary XRootD protocol, which should greatly simplify both the client and server (WebDAV is roughly https, so standard http clients can be used, and unlike GridFTP, additional data ports do not have to be opened on the servers).

5.9.8 Network and Data Architecture

Vanderbilt's research computing network is segregated from the greater campus network via a firewall. This research computing network is a spine-leaf topology with four leaves and seven spines, with each leaf having 40G uplinks to all four spines and Open Shortest Path First handling the failover/redundancy. Compute nodes are connected via 1gbit connections to this network (with a few nodes using10G connections). The Maximum Transmission Unit [measured in bytes] on the internal network is 9000. CMS's data are stored across 49 machines with up to 36 spinning disks each. These storage devices are all connected via 10g links to their spine switches. Peak data rates of 180Gbit/sec have been observed from this storage system, and sustained rates in excess of 50Gbit/sec are common.

There is a physically separate Intelligent Platform Management Interface (IPMI) network. All hardware is connected to the IPMI network via dedicated ports, and this network is accessible only via hardened admin servers.

At the university border, Vanderbilt University peers with both ESnet and Southern Crossroads (SOX) for Commodity Internet. In both cases, Vanderbilt is connected to a peering point in downtown Nashville (approximately five miles away). Vanderbilt peers with SOX at this Point of Presence and connects to (currently lit) fiber towards the SOX peering point in Atlanta, where Vanderbilt University peers with ESnet. All CMS-related nodes are dual-homed with this external network in our Science DMZ, meaning there is only one hop from our DTNs to the Vanderbilt border router. We additionally have two perfSONAR nodes on our external router, providing bandwidth and latency measurements.

We currently have 16 DTNs connected with 10g connections to the external interface (we have so many because the CPU on these nodes are relatively underpowered). These nodes will be replaced by two powerful nodes connected via 40/56g in the next couple of months.

5.9.9 Cloud Services

Along with the work with integrating leadership-class supercomputers in CMS's workflow management system, effort has been expended with Google, Microsoft, and Amazon's cloud computing resources. For the same reasons, the use case for these services is predominantly for simulation. Previous proofs-of-concepts have shown CMS can effectively use up to 150k CPUs from these resources to provide simulated data.

5.9.10 Data-Related Resource Constraints

One of the limiting factors in the amount of data that can be recorded is the cost to store these data, both in terms of tape backup and for disk size. It should be noted, though, that the CMS detectors data acquisition system is the ultimate limiting factor.

5.9.11 Outstanding Issues

None.

5.9.12 Case Study Contributors

- **Andrew Melo,** *Vanderbilt University,* andrew.m.melo@vanderbilt.edu

5.10 Large Ion Collider Experiment (ALICE) Project and ALICE-USA Computing

ALICE is operated within the United States by LBNL Laboratory Research Computing (LRC) and ORNL. These two facilities provide a combined 6% of computing resources for the project as a whole. Due to ALICE's location as an LHC detector, it follows the schedule of the facility closely and has close ties to the HEP efforts. During the current ALICE shutdown, reprocessing and simulation computation efforts are underway. These will continue until the next experimental campaign in 2020. Data volumes are expected to increase, and as such ALICE is experimenting with changes to workflow that leverage new computational and storage resources, and the concept of streaming to facilitate different models for analysis.

5.10.1 Discussion Summary

- ALICE-USA represents 6% of all ALICE computing resource needs, and is currently split between resources located at LBNL LRC and ORNL Compute and Data Environment for Science (CADES). LLNL is no longer participating in providing resources for the experiment.
- The use of LBNL LRC resources was a recent migration. There is still a desire to use NERSC storage resources, but a known bandwidth constraint exists between the resources.
- A desire exists to use HPC resources at LCF (e.g., ORNL's Titan and/or NERSC's Cori). To function properly, ALICE compute requires local data access to support compute jobs. This will not scale at LCF without use of proxy tools.
- ALICE is experimenting with the use of GPUs and CPUs but will require software changes to facilitate both use cases.
- Due to ALICE's location as an LHC detector, it follows the schedule of the facility closely and has close ties to the HEP efforts. The three-year LHC Run 2 period ended in 2018 and is currently at the beginning of the second long shutdown, LS2, which will last through 2020. The start of the three-year LHC Run 3 begins in 2021, which for the ALICE and LHCb experiments marks the beginning of the high-luminosity LHC era.
- The high-luminosity LHC era projects 100-fold larger data rates from detectors than observed during Run 2.
- ALICE data volume is a function of the data on a per collision (event) basis. All technology requirements (network, storage, and computing) are reduced to per event quantities of event size and processing time, and then multiplied by the event collection rate or total number of events collected.
- ALICE data flows are related to two primary activities: experimental results and simulation data, produced via MC methods. Both data sets are produced, stored, and shared using the ALICE Grid facility.
- Experimental data come in three varieties:
  - Raw event data: stored at CERN, where detector calibrations and initial event reconstruction passes are run.
  - Event Summary Data (ESD): first-level processed data, used for some analysis. Subject to further processing, which includes pattern recognition and filtering.
  - AOD: produced from refined ESD and used in most end-user analyses.
- ALICE Grid consists of about 70 additional T1 and T2 facilities, distributed about the world. Typically, ALICE workflows operate on the assumption that compute and storage are in close proximity to each other. There is no chaotic model for computing and access. As a result, most

networking needs are "local" in nature to support local read/write needs. Data are written once and read 10 to 20 times.

- T1 facilities are relied upon for (1) long-term custodial storage of a second copy of the raw and reconstructed data, (2) additional reconstruction passes over the raw data, (3) further processing and analysis of the reconstructed data, (4) disk resident storage of and access to ESD/AOD data, (5) processing and storage of MC simulation data in quantities comparable to the real event data, and (6) running end-user analysis tasks.
- The T2 facilities provide the same functions as the T1 facilities except for (1) and (2), long-term custodial storage of data and additional reconstruction passes.
- Data sizes for all of the tiers are as follows:
  - Data sizes @ CERN: ~5 PB raw & 2 PB ESD/AOD
  - Data sizes @ all T1s: 5 PB raw, 2 PB ESD/AOD, 1 PB MC
  - Data sizes @ all T2s: 2 PB ESD/AOD, 2 PB MC
- All participation in ALICE is the same from a technology provisioning perspective, and this is done to simplify the operation model. There are regular coordinated efforts to make MC data and perform coordinated analysis. User initiated analysis is less common, but possible. Analysis facilities, for example specialized places to handle user requested jobs, are being explored.
- Storage re-balancing between facilities is performed on a yearly basis. This is a data-intensive operation that requires network bandwidth. The limiting factor for this exercise is storage performance, although network bandwidth estimates are:
  - 2021: 8-16Gb/s
  - 2024: 24-40Gb/s (LHC Run 3)
  - 2024+: 80+ Gb/s
- WAN use is heavy when new resources are commissioned or existing resources are backed up. During rare events of this type, WAN expectations can be several Gbps versus the normal hundreds of Mbps that are needed normally.
- Remote IO (e.g., the XRootD tool) has been experimented with to reduce the need for local storage, but the results were mixed. Placing extra pressure on the network can occasionally result in performance problems that do not show up when using local compute and storage.
- Wide-area data movement typically works well without incident. In 2019 a problem was reported that affected normal operations between some facilities. Due to the timing of the incident, it was easier to work around the problem (e.g., deliver data from multiple other copies at different locations). ALICE is exploring more use of perfSONAR, similar to other LHC experiments, to understand and improve network performance.

5.10.2 Science Background

The ALICE collaboration has constructed and operates a heavy ion detector to exploit the unique physics potential of p-p and nucleus-nucleus interactions at collision energies of the LHC at CERN. The principal goal of the experiment is to study the physics of strongly QGP, a novel phase of matter produced at extreme energy densities. These studies are carried out with measurements from Pb-Pb, p-Pb and p-p collisions at the LHC. In order to extract the most physics information from the measurements, ALICE, like all of the LHC experiments, requires collecting and processing an unprecedented amount of experiment data. The LHC experiments have adopted

a distributed computing model for the processing, analysis, and archival of data organized within the Worldwide LHC Computing Grid (WLCG) collaboration. For ALICE, all participating countries are expected to contribute CPU, disk, and mass storage within the sponsoring country in proportion to its PhD participation. The ALICE-USA Computing Project was established to meet U.S. obligations by operating ALICE Grid facilities in the United States. The initial ALICE-USA obligations corresponded to about 6% of all ALICE computing resource needs and are currently about 7% of those requirements. A brief history of the computing project is provided here.

The ALICE-USA Computing Project officially launched in 2010 and originally deployed and operated U.S. ALICE Grid sites located within two computing centers, Livermore Computing (LLNL/LC) and NERSC PDSF at LBNL. In 2015, the ALICE group at LLNL withdrew from ALICE, which led the U.S. computing project to establish a new ALICE Grid site at the ORNL CADES facility and to decommission its LLNL/LC site that year. During the next couple of years, it became clear that the NERSC PDSF facility would eventually need to be decommissioned to make room at NERSC for its future generations of supercomputers and storage systems. As a result, the U.S. computing project developed a transition plan to build a new ALICE Grid site in the Lawrencium facility, a cluster operated by the LBNL scientific computing group in the Lab's Information and Technology Division. After a successful prototype deployment in 2018, the project expanded the LBNL/HPCS facility so that it had the capacity to fulfill the portion of the ALICE-USA obligations coming from LBNL. Lawrencium officially took over as the second U.S. ALICE Grid site in April of 2019 with the decommissioning of the NERSC PDSF cluster. An overview of this history is illustrated in **Figure 34,** which shows the number of concurrent ALICE jobs on each center during the nine years of the ALICE-USA Computing Project.

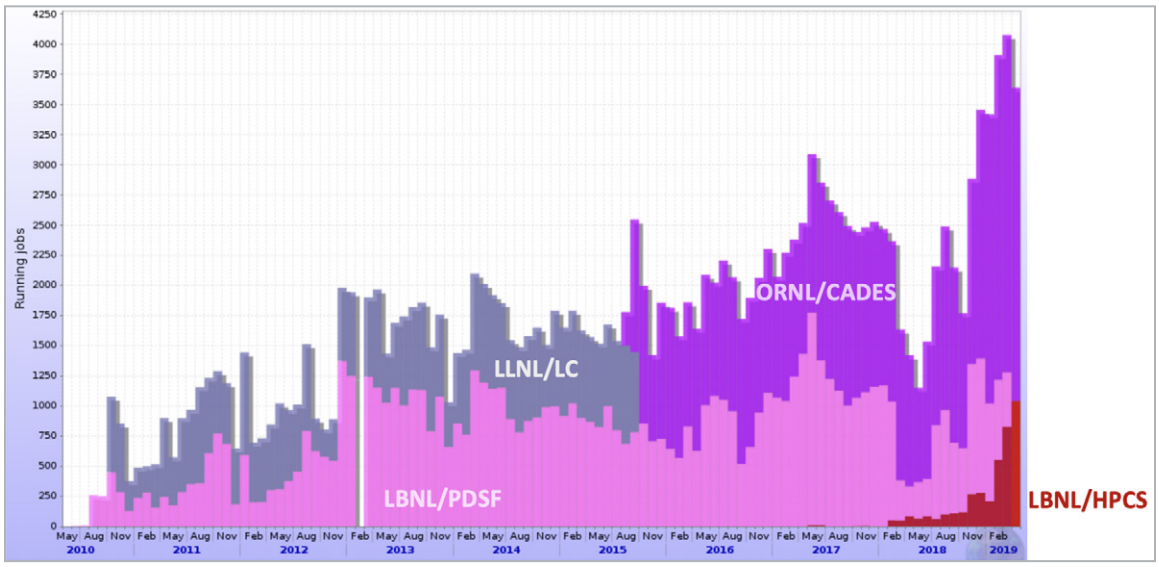

Figure 34. Number of concurrent ALICE jobs on each U.S. Grid site over the lifetime of the ALICE-USA Computing Project.

In addition to the use of conventional clusters at LLNL/LC, NERSC/PDSF, ORNL/CADES, and Lawrencium, the ALICE-USA project has worked on integrating HPC systems at ORNL (Titan) and NERSC (Edison and Cori) into the ALICE Grid facility. Such integrations at ORNL have not been successful due to a mismatch of specific requirements, such as outgoing network connectivity from the worker nodes to the WAN. Integrations at NERSC have been successful, and plans are being developed for modest use of the NERSC Cori system, which can be ramped up with the deployment of the next generation NERSC system, Perlmutter, to coincide with the upcoming LHC Run 3 in 2021 as described later. Expected use of NERSC HPC systems within the ALICE Grid and the challenges that such use imposes on the project will be included in the case study narrative presented here.

The ALICE detector operates in conjunction with the running schedule of the LHC at CERN, taking data during p-p and Pb-Pb (or p-Pb) collision periods each year. The broad LHC schedule consists of multiyear periods of operation, separated by two-year shutdown periods for collider maintenance and upgrades to both the collider and the experiments. The three-year LHC Run 2 period ended in 2018, and we are currently at the beginning of the second long shutdown, LS2, which will last through 2020. The start of the three-year LHC Run 3 begins in 2021, which for the ALICE and LHCb experiments mark the beginning of the high-luminosity LHC era with data rates from the detectors reaching a 100-fold larger than those during Run 2.

Data from the experiment are collected per detected collision (event), such that relevant quantities for network, storage, and computing requirements reduce to per event quantities of event size and processing time, multiplied by the event collection rate or total number of events collected. In addition to the raw data, a comparable volume of MC simulation data used to evaluate measurement efficiencies and systematic uncertainties required with each data set is produced and stored on the ALICE Grid facility. These quantities are translated into the global CPU and storage requirements from ALICE from which the U.S. obligation is evaluated. The scientific workflow is a sequence of processing over the collected (or simulated) data based on detector and event characteristics. Each step in the process refines and reduces the data, which are then stored for further analysis. The first-level processed data, referred to as ESD, are used directly in some analysis tasks but also processed further using standard sets of pattern recognition and filtering algorithms to produce a refined set of quantities known as AOD, used in most end-user analyses. During Run 3, the larger ESD data sets will not be stored and ALICE physicists will need to migrate analyses to the smaller and newly redefined AOD data sets.

The vast majority of ALICE computing work is done on the ALICE Grid facility. The Grid is a set of computing sites composed of a single Tier 0 (T0) center at CERN for primary data storage and initial processing, seven Tier 1 (T1) centers providing additional processing and both tape and disk storage capacities, and many Tier 2 (T2) centers with CPU and grid-enabled disk storage capacities, referred to as Storage Elements (SE). Raw event data is stored at the single T0 computing facility at CERN, where detector calibrations and initial event reconstruction passes are run. The rest of the computing workflow is done on the ALICE Grid consisting of about 70 additional T1 and T2 facilities distributed about the world. The T1 facilities are relied upon for (1) long-term custodial storage (tape) of a second copy of the raw and reconstructed data, (2) additional reconstruction passes over the raw data, (3) further processing and analysis of the reconstructed data, (4) disk resident storage of and access to ESD/AOD data, (5) processing and storage of MC simulation data in quantities comparable to the real event data, and (6) running end-user analysis tasks. The T2 facilities provide the same functions as the T1 facilities except for (1) and (2), long-term custodial storage of data and additional reconstruction passes. An overview of the ALICE Grid is shown in **Figure 35:** a world map of the ALICE Grid on which each dot is a grid site and each line represents data transfer at the moment the image was made.

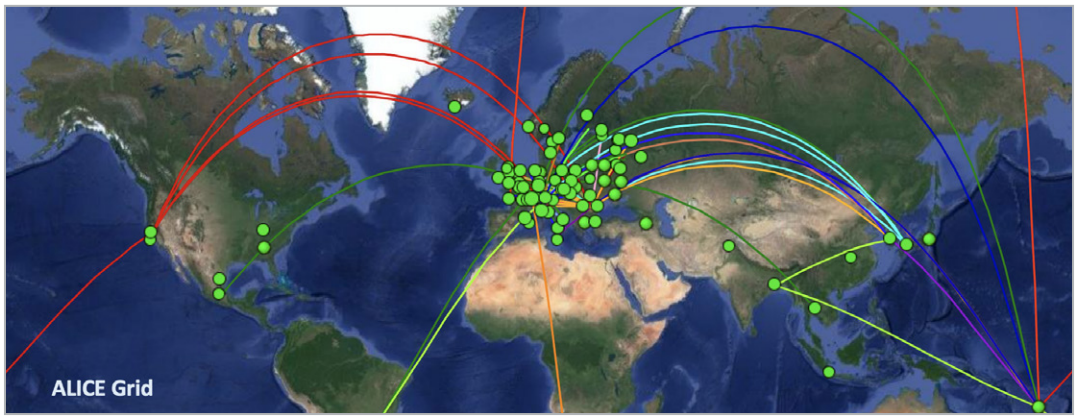

Figure 35. The ALICE Grid facility. Green dots represent sites while the lines represent data transfers occurring at the time the snapshot was taken. U.S. sites are HPCS and NERSC at LBNL and CADES and Titan at ORNL. The ORNL Titan site is not actively used.

The key feature for data management and access on the ALICE Grid is the distribution of the data onto the grid at the data's creation. The data are then subsequently accessed only from local site storage.[1] That is, while the ALICE computing is fully distributed, data processing is done locally; jobs are sent to where the data reside. In effect, the ALICE Grid operates a 90PB distributed file system that is primarily used as disk storage on the local cluster. During the past year, for example, ALICE jobs have read over 930PB and written over 75PB of data from/to the local storage, averaging about 35GB/s. Over the same period, conversely, ALICE jobs have read just 48PB and written 14PB of data over the WAN, averaging just 2GB/s of aggregate bandwidth.

5.10.3 Collaborators

The ALICE collaboration consists of almost 2,000 scientists, engineers, and students spread over 175 institutions in 40 countries. The ALICE Virtual Organization (VO) is a worldwide organization for use by ALICE scientists interacting with Grid organizations such as the WLCG and the OSG in the United States. The registry of members, including information about roles with respect to computing activities, is maintained in the Virtual Organization Management and Registration Service by the WLCG at CERN for ALICE.[2] The VO manager is Latchezar Betev (CERN), who is also in charge of grid operations for the ALICE Grid facility. There are several hundred ALICE scientists registered with the ALICE VO who directly use the ALICE Grid facility.

The ALICE-USA Computing Project is a DOE-funded project, run out of LBNL, with Jefferson Porter (LBNL) as the project lead and active member of the ALICE Computing Board. Operations at the ALICE-USA sites are coordinated by a steering committee that meets monthly and consists of the project leader (Porter) and ALICE representatives from each of the two sites: K. Read, P. Eby, and M. Galloway at ORNL and J. White, K. Fernsler, and J. Porter at LBNL. The group holds biannual in-person meetings that include participation from the ALICE off-line team from CERN and members attend an annual ALICE Tier-1 / Tier-2 workshop in which many ALICE site managers from around the world participate.

In the context of collaborators working on data management and processing, the primary driver is the ALICE Grid facility, not the individual scientists. Individual scientists submit jobs to the Grid, either as single user tasks or combined into organized analysis trains that are centrally managed and operated. In either case, refined analysis results are stored on the Grid but logically linked to the scientist's personal allocations. Those results are typically small and can be copied to personal computer resources for final analysis or preparation for presentation. Understanding the ALICE Grid structure is the most relevant aspect for understanding data generation and access patterns and is thus the context by which the access table has been completed

[1] The model can fall back to pull a copy from a non-local resource, but this is done at the 5% level.
[2] https://lcg-voms.cern.ch:8443/vo/alice/vomrs

User/ collaborator and location	Is a primary or secondary copy of the data stored?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of data set? (report in bytes, e.g., 125GB)	Frequency of data transfer or download? (e.g., ad-hoc, daily, weekly, monthly)	Are data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g., difficult tools, slow network)?
CERN T0	Primary	ALICE Grid	~5 PB raw 2 PB ESD/ AOD	At creation. Raw data pushed to T1s.	N	N
Combined 7 T1 Sites: Germany, France, Italy, UK, Netherlands, Russia, Korea	Primary and secondary	ALICE Grid	5 PB raw 2 PB ESD/ AOD 1 PB MC	At creation + routine managed data redistribution	Y, as the ALICE Grid is 'the source'	N
Combined 60+ T2 Sites	Secondary	ALICE Grid	2 PB ESD/ AOD 2 PB MC	At creation + routine managed data redistribution	Y, as the ALICE Grid is 'the source'	N

Table 10. ALICE Data Summary.

5.10.4 Instruments and Facilities

The ALICE experiment and its T0 facility are located at CERN in Geneva, Switzerland. Most of the ALICE Grid resources are located in Europe, including all current ALICE T1 sites except KISTI in Korea. That cluster of resources in Europe is illustrated in the map shown in **Figure 35**. The two ALICE-USA sites, ORNL/ CADES and Lawrencium, are operated as T2 centers and as such, do not participate in processing of raw data. The raw data reconstruction passes at the T0 and T1 sites noted in **Table 11** produce ESD/AOD files, which are automatically replicated and distributed on the ALICE Grid at the time of their creation. This same model is used for MC simulations run on both T1 and T2 sites to produce reference data for efficiency analysis and model comparisons. Thus the wide-area data distribution workflow model for the ALICE-USA sites is (1) receive a fraction of ALICE ESD/AOD data files produced at T0/T1 sites in Europe (and Korea), (2) receive MC simulation files produced at T1/T2 sites, (3) send copies of MC simulation files and analysis-reduced data produced locally to other sites, including between the U.S. sites.

Processing task	Activity	Location	Input & source	Output & dest.
Raw data reconstruction	Organized & managed	T0 & T1	Raw data Experiment or Tape	ESD/AOD onto ALICE Grid SE
MC simulation + reconstruction	Organized & managed	T1 & T2	Configuration data from remote SE	Simulated data onto ALICE Grid SE
Analysis trains = multiple analyses in single process	Organized & managed	T1 & T2	ESD/AOD from local SE	User ROOT files to ALICE Grid SE + Copied off by hand
User analysis on the Grid	Chaotic	T1 & T2	ESD/AOD from local SE	User ROOT files to ALICE Grid SE + Copied off by hand

Table 11. Types of processing carried out by ALICE scientists, all on the ALICE Grid.

About 85% of the processing on the ALICE Grid is devoted to data analysis or MC simulation. As a result, there is little distinction between T1 and T2 facilities for the general work carried out on the ALICE Grid facility. Sites with large storage, all T1 and many larger T2 sites, will accommodate more data-intensive user analysis tasks. All work, from managed production to chaotic individual user analysis, uses the same Grid submission and job management tools. Types of ALICE data processing are summarized in **Table 11.** The fraction of the ALICE Grid CPU resources used by each type of processing over the past year is shown in **Figure 36**. With the exception of raw data reconstruction, all resources are available for use by all types of processing tasks allowing the overall job utilization to remain fairly constant even while the mix of job types fluctuates. This type of elastic resource usage would be lost if facilities are restricted to host only a single type of task.

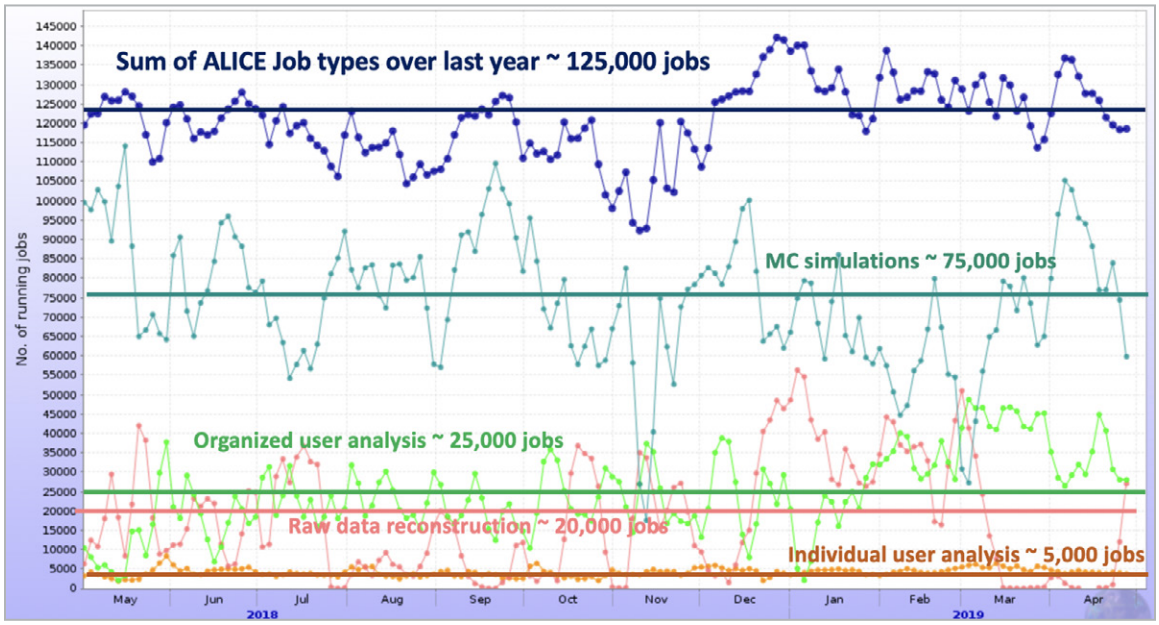

Figure 36. The job-mix on the ALICE Grid facility as split between MC simulations, organized analysis, raw data reconstruction, and individual user analysis. The sum total of jobs remains much more stable than individual job types.

The CPU and disk capacity of the ALICE-USA facility are currently about 3,500 CPU cores and 5PB of disk storage, split about evenly between the two sites. The network activity produced by the computing model is shown in **Figure 36,** in which the rates data are written to (top plot) and read from (bottom plot) the Lawrencium disk storage as monitored by the ALICE are plotted. From local monitoring by the HPCS internal systems, we have the breakdown between LAN and WAN rates annotated in the plots. The site was running about 1,000 ALICE jobs during the period when the plots were generated. The peaks correspond to periods with large fractions of analysis jobs, which require ~3MB/s per job, while the valleys occur when the cluster is running mostly MC simulations. The annotations clearly show that the majority of bandwidth used in the ALICE computing model is between the local compute resources and the local storage.

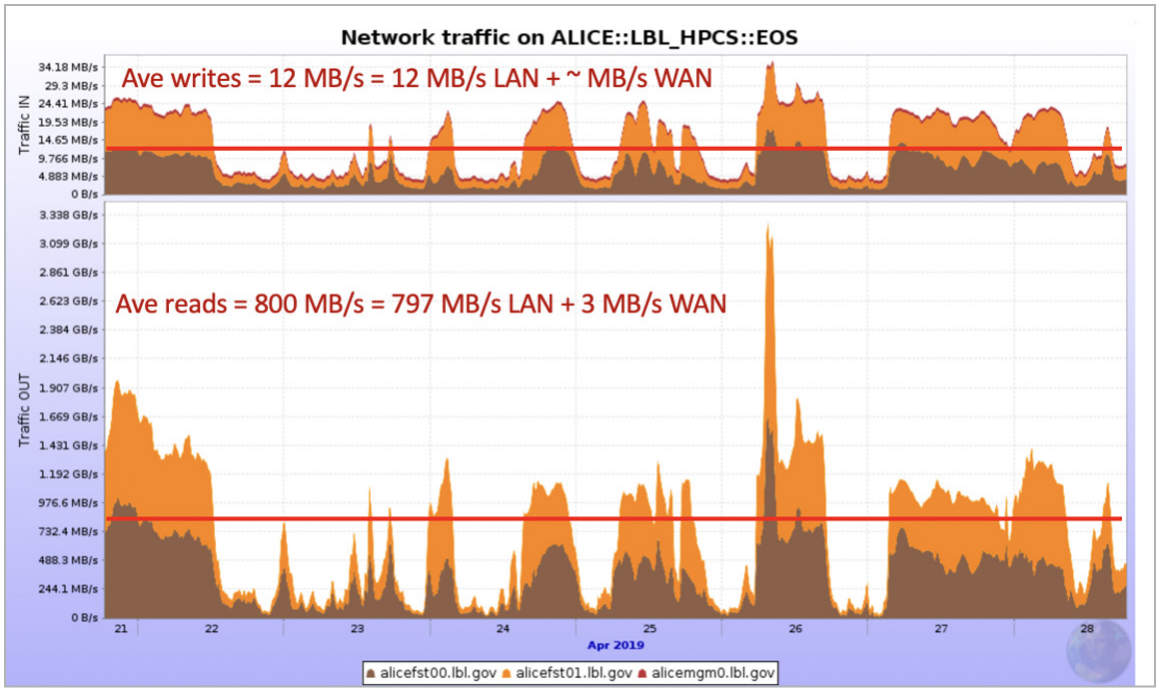

Figure 37. Traffic written to (top) and read from (bottom) the Lawrencium storage servers, alicefst00 and alicefst01, for a week in April 2019. Annotations show the average rates broken by LAN and WAN network.

In addition to the bandwidth characteristics of normal operations as shown in **Figure 37,** there are episodes that require more significant WAN capacities. These occur a few times a year when storage is added and/or decommissioned or when data must otherwise be redistributed between different sites. One such period is shown in **Figure 38** when newly added storage at Lawrencium was the primary target for data redistribution. During those periods, the WAN network requirements are on the order of GB/s instead of the 10 MB/s capacities needed during normal operations.

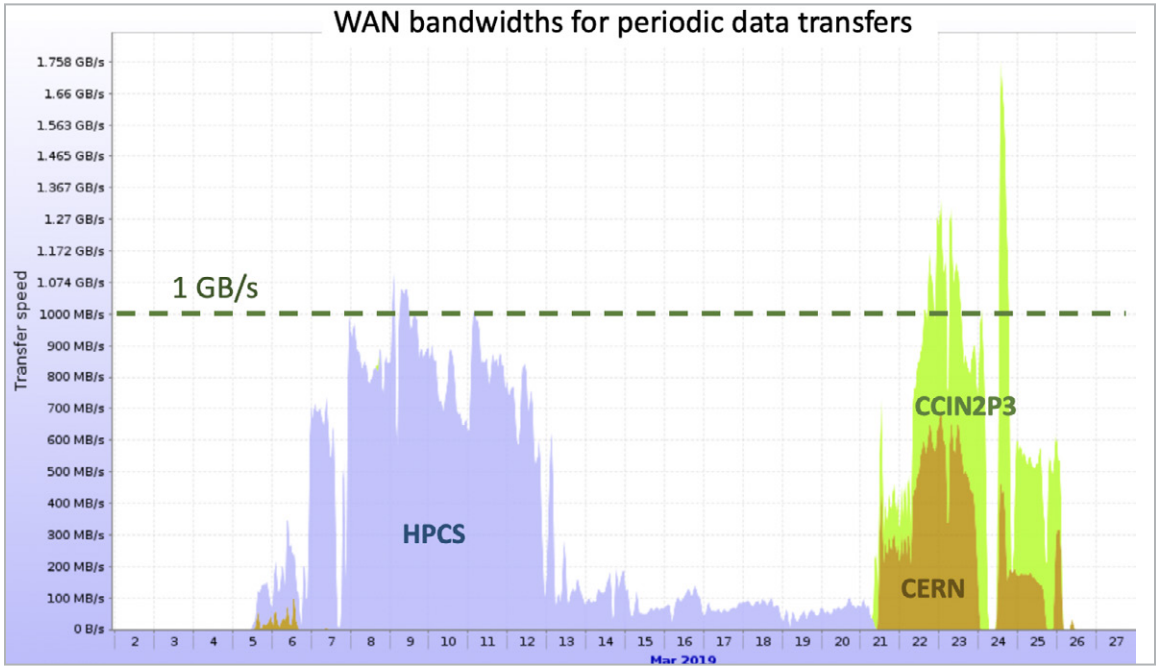

Figure 38. Network utilization during period of data redistribution on the ALICE Grid.

The previous description is applicable to the current ALICE computing model that has been in place for the past decade and will remain the model for the next two years through LS2. After that, a different model has been proposed for Run 3, which marks the high-luminosity era for ALICE with an expected 100x increased data acquisition rate at the detector.[3] In that updated computing model, data are highly filtered during the reconstruction passes done at the experiment site, eventually producing an improved AOD format that will be migrated to the ALICE Grid. That migration may include the addition of a set of ALICE Analysis Facilities (AF) with very large, performant disk storage and processing capacities. While this change in the ALICE computing model may be significant, its effect on the individual grid sites is planned to be minimal. In fact, while there may be adjustments in the balance of disk and CPU resources, the current plan is to retain an annual site-level resource growth rate of ~25% per year.

Period	CPU (cores)	Disk (PBs)	Aggregate LAN BW (GB/s)	Aggregate WAN BW (GB/s)	Computing model	Grid-enabled resource types & configurations
0–2 yrs	4k-7k	8-10	5-10	0.5	Run 1 & 2 model	Conventional clusters & disk storage
2–5 yrs	8k-12k	10-18	10-20	1-5	Run 3, 02 model	Integrate use of: HPC facilities, Accels/GPUs, ML techniques, Tiered storage, Analysis Facility
5+ yrs	20k+	20+	20+	5+	Run 3, 02 model	continued ...

Table 12. Expected ALICE-USA computing resources for the periods requested by this case study.

Under the scenario described, the expected CPU, disk, network, and computing model for the periods requested in the case study are shown in **Table 12.** The current model and framework will remain in place as the Run 2 data are processed and analyzed. During this transition period, the ALICE software stack is undergoing a significant rewrite, which will leverage use of GPUs and optimize I/O performance by reducing the work needed for deserialization of the data objects. Changes to the overall infrastructure will include more aggressive use of HPC facilities, ML techniques, and use of tiered storage systems. Even with such significant changes to the overall ALICE computing model and capabilities, the resource obligation to ALICE from the United States should remain as a steady growth in CPU, disk, and network capacities.

5.10.5 Process of Science

A significant amount of processing carried out within the scientific investigation is done within an organized production model. The types of tasks are listed in **Table 11,** and these will continue over the next **zero to two years.** Once the raw data are taken and detector calibrations determined, one or more reconstruction passes are run over the raw data, managed by the central team to produce data files that can be used by individual physicists. Similar production campaigns are carried out for MC simulations. The ALICE Grid facility is constructed to manage these productions by the central team in the same way all users can perform their analysis tasks directly on the Grid facility. Individual scientists connect directly to the ALICE Grid using client software on their personal laptop or on commonly used local clusters. They submit tasks to the grid or have them run within analysis trains as if the grid were a monolithic cluster. Those tasks analyze the data generated during the production campaigns and produce further refined data that can be accessed directly by individual scientists for final inspection and interpretation. The main difference between managed production and tasks run by individual users is the priority given to each task.

[3]The high luminosity era of the LHC, Run 4 for ATLAS and CMS, begins in Run 3 for the ALICE and LHCb experiments.

The two- to five-year period covers the LHC Run 3 when the ALICE computing model changes significantly. However, the goal of that new model, referred to as O2 (Online/Off-line combined), is to more efficiently process and reduce the large amounts of data coming off the detector. The software changes needed for Run 3 present an opportunity to revisit how analyses are done, both by the infrastructure and user codes. For example, dynamic parallel processing using shared memory within nodes and message passing services between nodes is being built into the software developed in the O2 framework. Such features should allow additional flexibility in the types of resources used by ALICE computing operations. In addition, the growing use of ML techniques for scientific research may be better leveraged by codes that take advantage of site-specific resources not usable within the current framework. If such site-specific resources are impactful to the scientific discovery process, then the ALICE Grid services will need to better support directed data movement. In that case the network bandwidth capacities between specific sites may need to be enhanced.

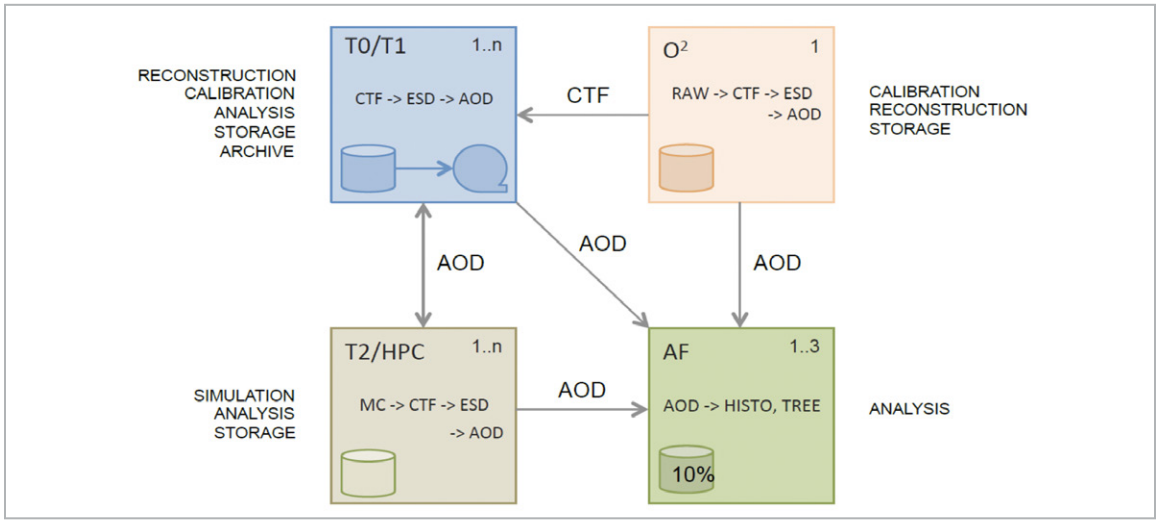

Figure 39. ALICE O2 Computing model concept, consisting of a single O2 facility at the experiment site, many Grid sites (T0/T1/T2s) around the world, and a handful of AFs. The model is attractive in that the AFs can be configured for the high I/O requirements of user analysis.

An illustration of the change just described is shown in **Figure 39** in which user analysis is redirected to run only on ALICE AFs. Such a model will require explicit high-bandwidth connections between an AF and the rest of the ALICE Grid. The advantage of this model is that an AF can be configured for the high I/O requirements of user analysis. The challenge to this model is that the ALICE Grid would need to support several AFs, each with capacities of tens of thousands of CPU cores and tens of PB of performant disk storage. It is likely that consistent use of those large amounts of resources would fluctuate, offsetting the efficiency gained by restricting user analysis to run only on these highly optimized facilities. At present, there is one prototype AF at GSI[4] in Germany. It is unclear whether deploying such AFs will be cost effective to ALICE. It may be that such a model will be realized during Run 3 or even in the **five-plus-year timeframe.** It is also possible ALICE T2 facilities will become sufficiently high performant so as to replace the conceptual need for ALICE AF centers.

5.10.6 Remote Science Activities

As noted in previous sections, the computing requirements of the ALICE experiment and its scientists are unique in that they are fully realized in the distributed computing model implemented in the ALICE Grid facility. As such, "remote science activities" are well described in other sections of this case study.

[4] https://www.gsi.de

5.10.7 Software Infrastructure

The two distinct types of ALICE software infrastructure are the software and services used to operate the ALICE Grid facility and the software framework used by ALICE physicists to analyze ALICE data. The Grid infrastructure is described first followed by the scientific analysis framework.

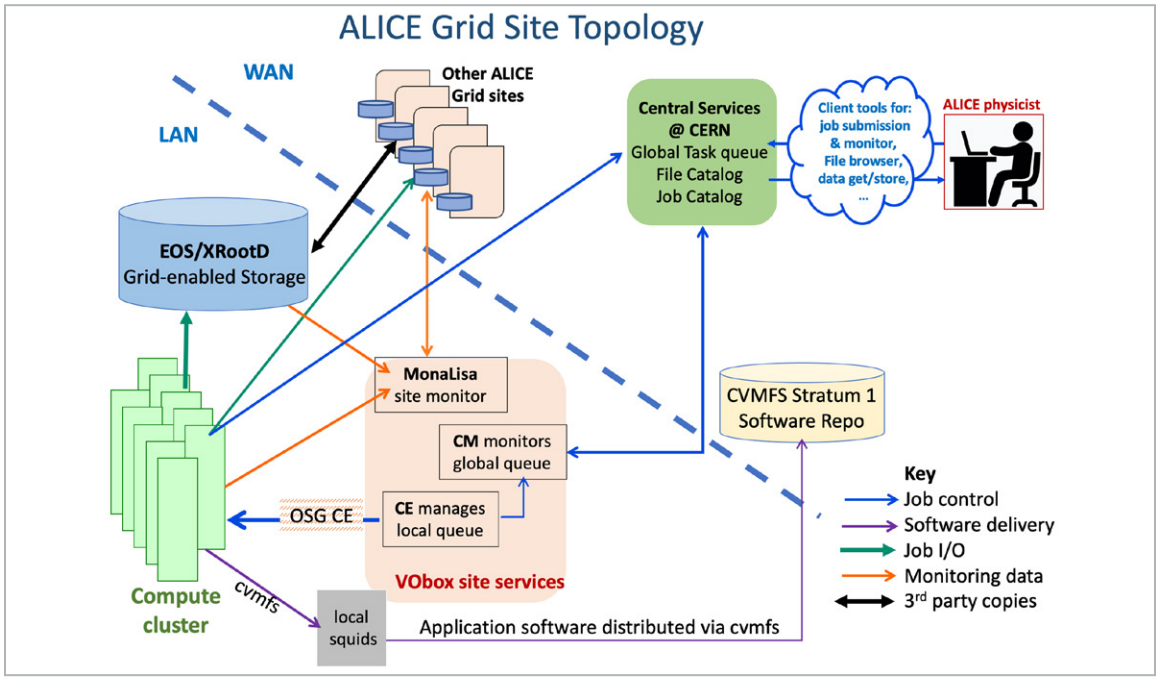

Figure 40. ALICE Grid site topology and connection to the larger ALICE Grid.

Each data center participating in the ALICE computing model is configured to run within the ALICE Grid facility via ALICE middleware services run on the site, illustrated by a site topology diagram in **Figure 40**. Each site runs an ALICE VOBox, an independent machine with network access to the ALICE central services managed at CERN and direct access to the local compute cluster and SE. Services on the VOBox receive and collect monitoring information about the cluster and local SE and are responsible for managing the work on the local cluster. The services and software to support those services shown in **Figure 40** are listed here in **Table 13** with a brief description of their operation.

The ALICE Grid facility will continue operations during the next **zero to two years** and beyond. Currently many of the site services are undergoing a full rewrite from a mix of Perl, Java, and C to primarily a Java-based system. The functionality of these services is not being changed, though some of the techniques used are different. Examples include GSI authentication being replaced by Tokens, processes being run in Linux containers, and the Global File Catalog being migrated from pure MySQL to a different set of database technologies. The overall functionality of the ALICE Grid and the types of site services depicted in **Figure 40** and listed in **Table 13** are expected to be supported in both the **two- to five-year** timeframe and beyond in the **five-plus-year time frame.**

Service	Software	Location	Description
Software (SW) deployment	CVMFS	Compute node cache + squid access to SW repo	CERN solution for distributed SW delivery
Site monitoring	MonaLisa	VOBox server	Local monitoring data store, feeds central data collection at CERN
Process monitoring	MonaLisa probes	On SE, with jobs, on VOBox	Processes probe data & send to local ML server
Job management	ClusterMonitor + CE	VOBox servers	ClusterMonitor evaluates global work. CE launches / monitors active and queued jobs.
Local queue	OSG/CE or batch systems: PBS, SGE Slurm, Condor.	Compute cluster	ALICE CE submits job to the local batch or through OSG CE
Job processing	JobAgent pilot jobs	Compute nodes	JobAgent runs payload jobs from central services
Grid-enabled storage & data transfers	EoS / XRootD	Storage servers	EoS storage system based on XRootD, data access via XRootD protocol. Dynamic data transfers done via xrdcp.

Table 13. Expected ALICE-USA computing resources for the periods requested by this case study.

The ALICE software framework used now and in the **zero- to two-year** time frame for processing data is based on the ROOT framework, C++ class libraries for reading, writing, processing, and presenting data. All data files are stored as compressed ROOT files and leverage ROOT I/O libraries for data access, including streaming from storage as supported by the XRootD protocol. The framework supports a large number of standard tools and algorithms for pattern recognition and data presentation used by HEP/NP physicists.

In the **two- to five-year** time frame, the ALICE software will undergo a significant rewrite (currently underway) in which the data formats and software frameworks will be very different from current implementations. Experience with I/O costs for storing and recovering complex objects is leading to a flatter data model that no longer requires (but may still use) the complex I/O capabilities of ROOT libraries. This change, needed by the online event processing in the O2 facility, will allow the user and software developer more flexibility in designing data processing methods that may exploit new computing architectures. As in other sections, the changes ALICE needs in Run 3 dominate its future computing landscape, such that the **five-plus-year** time frame is expected to largely be an extension of the Run 3 O2 computing model.

5.10.8 Network and Data Architecture

5.10.8.1 ALICE Oak Ridge National Laboratory (ORNL)

The ALICE T2 site at ORNL connects directly to the ORNL Science DMZ, which is positioned at the border of the ORNL network with dual peerings with both the ESnet backbone and LHCOne. The ALICE environment consists of Arista 7150S core switches connected at 40G, and three Arista 7010 switches for management connectivity. The Science DMZ network is based on Arista 7280R switches with 40/100G capability. The current ORNL ALICE environment is depicted in **Figure 41.**

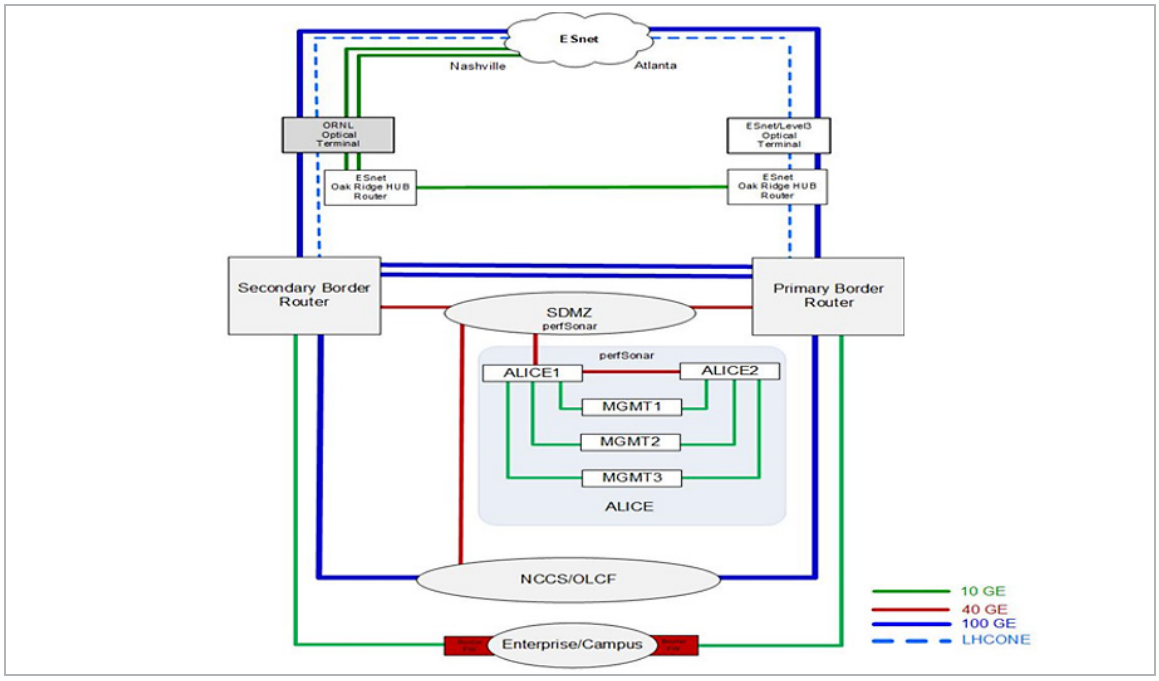

Figure 41. ORNL ALICE Networking Environment.

Increased use of the existing ALICE 40Gbs uplink to ORNL is expected to result in capacity saturation peaks in 2019. and sustained need to exceed thereafter. When LHC Run 3 begins in 2021, network demands to all WLCG sites (including the ORNL T2 site) are expected to be significant. In preparation, ALICE ORNL is looking to purchase new switches (between 2019 and 2020) with 100G uplink capabilities to the ORNL Science DMZ/ LHCOne border to meet the expected increase. The target network environment by 2021 is shown in **Figure 42.**

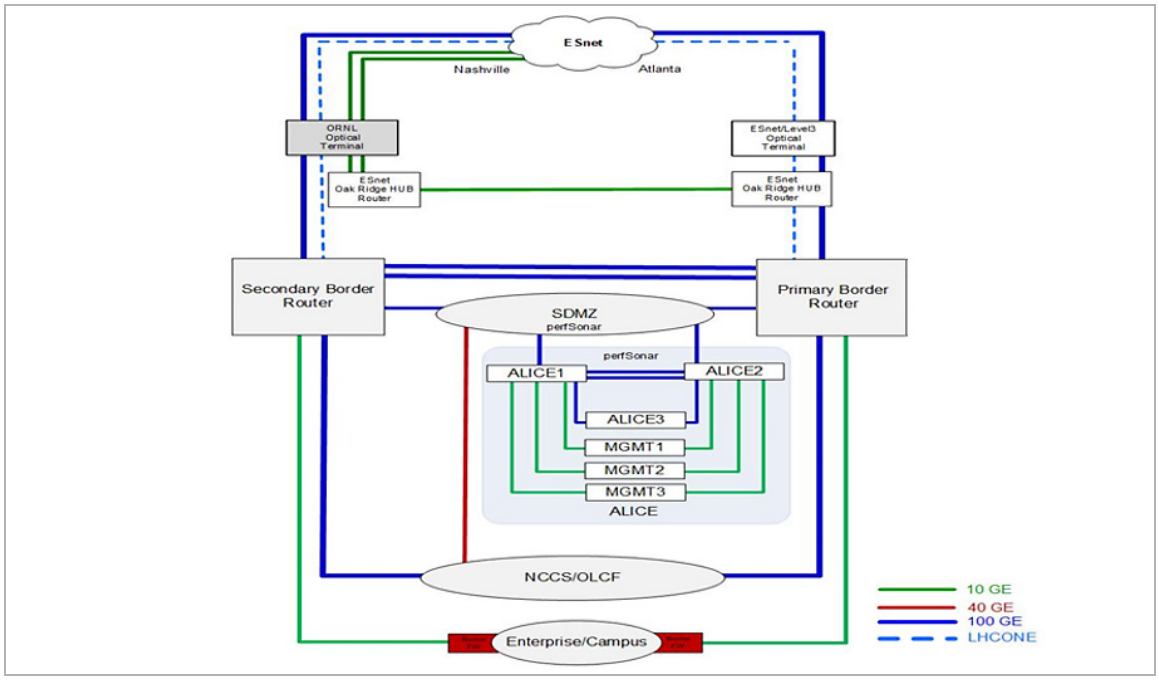

Figure 42. Planned Upgrades to ORNL ALICE Networking Environment.

Beyond 2020, ORNL connections to ESnet are expected to migrate to diverse 400G connections. This could include a period of multiple 100G connections on each border router depending on market availability, but the target environment is native 400G connections.

5.10.8.2 ALICE Lawrencium

The current network topology of the ALICE T2 facility at the LBNL HPCS center is shown in **Figure 43**. Internal connection between the worker nodes and the storage is done over 56Gb IB. Connectivity to the WAN is different between the CPU cluster and the storage with the storage connected directly via the Science DMZ and the compute cluster routed through a local firewall before reaching an LBLnet connection to ESnet. Our ALICE-USA Computing Project plan calls for adding a perfSONAR service installed on the same external route at the storage and to investigate adding LHCOne virtual routing to the EoS SE. Both of these are expected to be completed (or evaluated) this year to support the **zero- to two-year** timeframe. An item for the longer (**two- to five-year**) term is to optimize the network connectivity between the storage at the HPCS facility and HPC resources at NERSC, which will allow the conventional permanent storage at the HPCS site to support process at NERSC. This is discussed in Section 10.

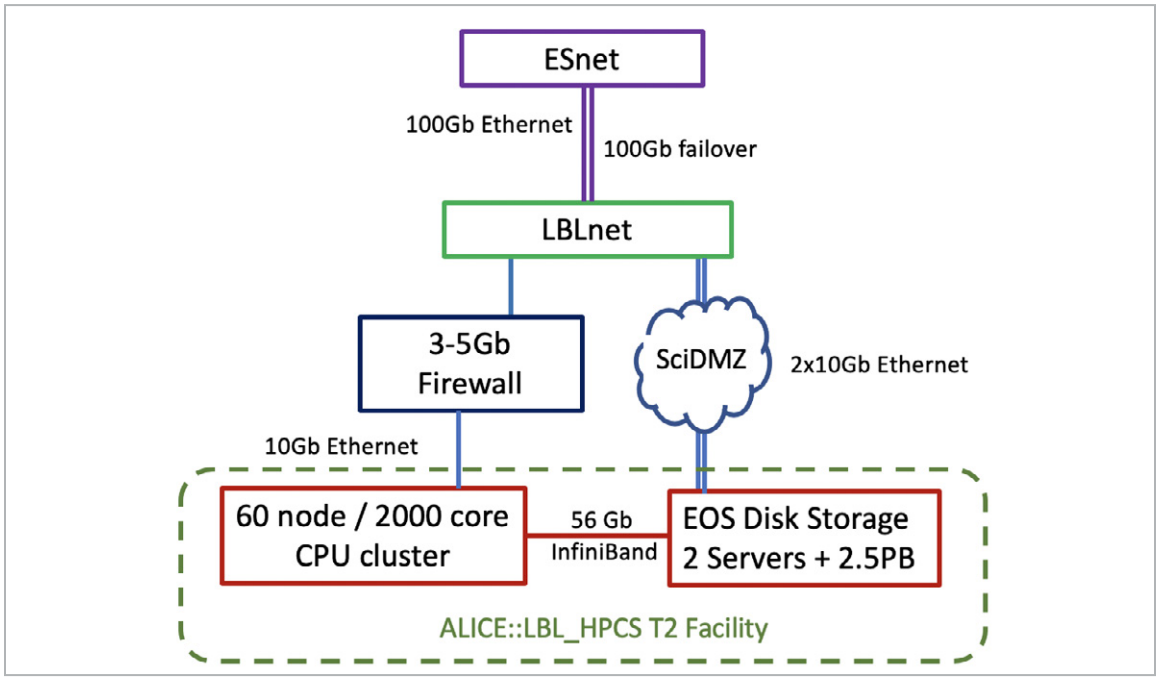

Figure 43. Network layout of the ALICE T2 facility at Lawrencium relative to ESnet peering.

5.10.8.3 ALICE Lawrence Berkeley National Laboratory (LBNL) / NERSC

Use of NERSC HPC systems in the next **zero to two years** will be primarily for simulations as there is no current plan to have a NERSC-based grid-enabled storage element. This will require modest network bandwidth capacity (much like in the "Cloud Services" discussion below). In the **two- to five-year period,** we hope to leverage use for data analysis, in which case we need to evaluate how to move data through the system from the local LBNL HPCS facility. This is discussed in item 9.

5.10.9 Cloud Services

The current and developing ALICE computing models do not have any specific plans for use of cloud resources, nor do they prohibit such resources use should they become cost effective. In fact, the "dot" over Iceland in **Figure 35** represents a cloud resource made available to ALICE by an EU research project. The ALICE group at CERN installed the ALICE middleware on the cloud service and integrated its use directly into the ALICE Grid as if it were a conventional cluster without a connected SE. The team found that the resource was fully functional and efficient for running MC simulations, from which the produced simulated data were distributed only to remote sites.

For estimating future use of cloud services, it is easiest to limit their use to MC simulations. Current bandwidth requirements for a simulation task are significantly less than a MB/sec, which would allow thousands of such tasks (jobs) to be run concurrently on a cloud service that supported network bandwidth capacity of a few hundred MB/s. As shown in **Figure 36,** large numbers of MC simulation jobs are constantly being run on the ALICE Grid. Thus, as long as a cloud service can be presented as a normal ALICE Grid site with even modest network capacity, the ALICE Grid system will be able to scale to use those resources as allowed by the cost of the services. However, it must be understood that the general ALICE resource providers, such as the ALICE-USA Computing Project, must present grid-enable disk storage directly connected to large amounts of CPU capacity to meet the full workload demand of the ALICE computing model. As a result, unloading a significant amount of the obligated CPU capacity to Cloud services for MC simulations is not a viable option for meeting the full computing needs of ALICE.

5.10.10 Data-Related Resource Constraints

One specific data-related constraint that we currently face is how to efficiently use our NERSC HPC allocation for data analysis with minimal impact on the ALICE Grid model. Since PDSF has been decommissioned, ALICE does not operate as a grid-enabled storage element inside NERSC. ALICE can operate its NERSC HPC resources as an ALICE Grid site without an SE, which means only MC simulation jobs will run on the facility, producing simulation data that is copied to remote storage systems. Integration of the NERSC HPC into the ALICE Grid facility in this way breaks the multi-task use model that other (nearly all) sites run under. Our options are to run an EoS or XRootD-based storage system within NERSC or to have NERSC HPC resources well connected to the nearby EoS system run on the LBNL HPCS facility. Both of these options are interesting and viable, but, in the context of this case study and perhaps simplicity of deployment, we would like to understand the current and future bandwidth capabilities to between the two centers located at LBNL. It may be some form of tiered storage between the facilities in which data are dynamically cached at NERSC for data analysis processing. Understanding the network limits now (**zero- to two-year**) and future (**two- to five-year**) capacities may help guide the development of ALICE use of NERSC HPC in the Perlmutter era and beyond.

5.10.11 Outstanding Issues

A recurring issue faced when managing interconnected resources used in highly distributed data processing frameworks is the sometimes subtle but inconsistent behaviors in the connectivity between the resources and the software components that tie them together. One example is illustrated in **Figure 37,** which shows the network bandwidth during a redistribution of data on the ALICE Grid in which HPCS was the prime target for that redistribution. After running for about six days at average bandwidths approaching 1.0GB/s, the rates dropped to about 100 MB/s. There were no known changes to the distribution framework or the available storage on the HPCS system that would account for that drop, and yet for over a week, the transfers could not exceed 100 MB/s to the HPCS storage. The impact of such occurrences is mitigated by the overall resilience of the ALICE Grid model which generally does not rely on such high bandwidths. Also, in this case and as shown in the figure, the data redistribution software simply retargeted the transfers to other sites to complete the task.

Another example comes from the ALICE monitoring of its SEs around the world. To do that, a process on the ALICE central services copies a small file onto each SE and another process reads the file. This process provides a simple read/write test that is run every few minutes on each SE, yielding a running availability measurement as the percentage of successful reads and writes. Last year, after we had made some significant changes to both the ORNL SE and the new LBNL HPCS SE, we noticed a drop in the availability rates on those systems. This caused us to work on debugging the "local" issues. We then noticed that our LBNL PDSF SE was having similar problems, but we had not modified that resource. Eventually that let us compare the availability rates at

all sites from which we saw a pattern emerge, which is illustrated in **Figure 44.** All the ALICE sites which were accessed from CERN via a cross-Atlantic link had much worse availability rates. But the failures were not 100%. We discovered that by upgrading the XRootD client used in the tests, the failure rate dropped significantly as that client was more resilient. After engaging LHC network engineers, the problem was eventually isolated to one of the multiple paths coming from the CERN campus network. As the figure shows, the problem lasted for several weeks before it was determined to be a network issue instead of site storage problems.

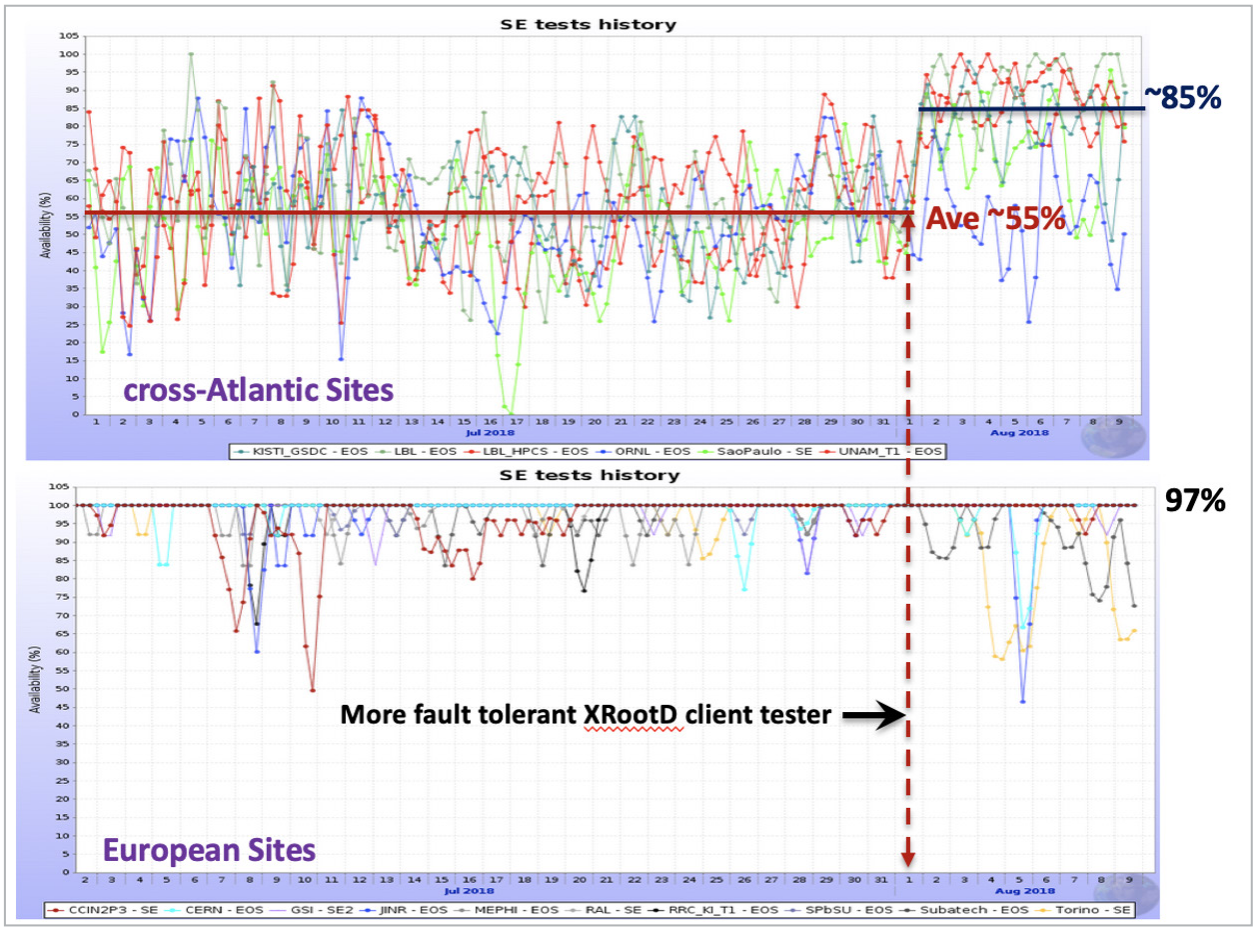

Figure 44. ALICE storage tests results for cross-Atlantic sites and those in Europe. Results indicated a network problem non-local to the grid sites instead of a local grid storage problem as originally thought.

Due to such occurrences and the effort they cost site engineers, we have begun the process of installing perfSONAR at our U.S. sites (see, for example, the ORNL network description of 7.1). Many of the ALICE Tier 1 sites, which also support ATLAS and CMS, already include perfSONAR systems. We hope to connect our monitoring systems into existing infrastructures run by OSG and WLCG. The project would appreciate guidance during this process.

5.10.12 Case Study Contributors

- **R. Jefferson Porter,** *LBNL,* rjporter@lbl.gov
- **Pete Eby,** *ORNL,* ebypi@ornl.gov
- **Susan Hicks,** *ORNL,* hicksse@ornl.gov
- **John White,** *LBNL,* JWhite@lbl.gov
- **Latchezar Betev,** *CERN,* Latchezar.Betev@cern.ch
- **Costin Grigoras,** *CERN,* Costin.Grigoras@cern.ch

6 Appendix

6.1 Appendix A: List of Abbreviations

AF	Analysis Facilities
AGFA	Argonne Gas-Filled Fragment Analyzer
ALCF	Argonne Leadership-Class Computing Facility
ALICE	A Large Ion Collider Experiment
ANASEN	Array for Nuclear Astrophysics Studies with Exotic Nuclei
ANL	Argonne National Laboratory
AOD	Analysis Object Data
ASCR	Advanced Scientific Computing Research
ASN	Autonomous System Number
ATLAS	Argonne Tandem Linear Accelerator System
AT-TPC	Active-Target Time-Projection Chamber
BCS	Beta Counting System
BECOLA	Beam-Cooler and Laser Spectroscopy Facility
BES	Beam Energy Scan
BGO	Bismuth Germanate
BNL	Brookhaven National Lab
CADES	Compute and Data Environment for Science
CAESAR	Cesium Iodide Array
CEBAF	Continuous Electron Beam Accelerator Facility
CephFS	Ceph File System
CFN	Center for Functional Nanomaterials
CMS	Compact Muon Solenoid
CPU	Central processing unit
DAQ	Data acquisition
DB	Database
DNS	Domain Name System
DOE	Department of Energy
DST	Data Summary Tape
DTN	Data-transfer node
DUNE	Deep Underground Neutrino Experiment
DWDM	Dense Wavelength Division Multiplexing
EIC	Electron-Ion Collider

ELITE	Mid-Atlantic Broadband Cooperative/COX Communications/Old Dominion University Virginia Regional Network
ENP	Experimental Nuclear Physics
EoS	Equation of State
ESD	Event Summary Data
ESnet	Energy Sciences Network
FDS	FRIB Decay Station
FNAL	Fermi National Accelerator Laboratory
FRIB	Facility for Rare Isotope Beams
FRIBUO	FRIB Users Organization
FSU	Florida State University
FTS	The GRID Data Transfer Service used at CERN
GEANT4	GEometry ANd Tracking
GlueX	Particle physics experiment located at JLab
GPU	Graphics Processing Unit
GRETA	Gamma-Ray Energy Tracking Array
GRETINA	Gamma-Ray Energy Tracking In-beam Nuclear Array
HEP	High-Energy Physics
HLT	High-Level Trigger
HPC	High-performance computing
HPCC	High-Performance Computing Center
HPCS	High-Performance Computing System
HPSS	High-Performance Storage System
HRS	High-Rigidity Spectrometer
HTSN	High-Throughput Science Network
iCER	MSU Institute for Cyber-Enabled Research
INFN	Italian National Institute for Nuclear Physics
IPMI	Intelligent Platform Management Interface
ISLA	Isochronous Large Acceptance
IT	Information technology
ITS	Information Technology Services
JANUS	Joint Array for NUclear Structure
JENSA	Jet Experiments in Nuclear Structure and Astrophysics
JLab	Thomas Jefferson National Accelerator Facility
JLEIC	JLab Electron-Ion Collider
KEK	High-Energy Accelerator Research Organization (Japan)

KISTI	Korea Institute of Science and Technology Information
LAN	Local Area Network
Lawrencium	LBNL's internal high-performance computing system
LBNL	Lawrence Berkeley National Laboratory
LCF	Leadership Computational Facility
LCRC	Laboratory Computing Resource Center
LEBIT	Low-Energy Beam and Ion Trap experiment at NSCL/MSU
LENDA	Low-Energy Neutron Detector Array
LER	Low-Energy Nuclear Physics Research
LHC	Large Hadron Collider
LHCONE	Large Hadron Collider Open Network Environment
LLNL	Lawrence Livermore National Laboratory
LQCD	Lattice Quantum Chromodynamics
LRC	Laboratory Research Computing
LSU	Louisiana State University
MAN	Metro Area Network
MARIA	Mid-Atlantic Research Infrastructure Alliance
MC	Monte Carlo
MIE	Major Item of Equipment
MiLR	Michigan Lambda Rail
ML	Machine learning
MOLLER	Measurement of a Lepton-Lepton Electroweak Reaction
MoNA-LISA	Modular Neutron Array
MSU	Michigan State University
MTU	Maximum Transmission Unit [measured in bytes]
MUSIC	Multi-Sampling Ionization Chamber detector, located at ANL
NASA	National Aeronautics and Space Administration
NCSA	National Center for Supercomputing Applications
NERO	Neutron Emission Ratio Observer
NERSC	National Energy Research Scientific Computing Center
NP	Nuclear Physics
NSCL	National Superconducting Cyclotron Laboratory
NSF	National Science Foundation
NSRL	NASA Space Radiation Laboratory
OLCF	Oak Ridge Leadership Computing Facility
ORNL	Oak Ridge National Laboratory

OSC	Ohio Supercomputer Center
OSG	Open Science Grid
PAC	Program Advisory Committee
PB	Petabyte
Pb-Pb	Lead atom collision
PBR	Policy-based routing
PHENIX	High-Energy Nuclear Interaction eXperiment
PI	Principle investigator
p-p	Proton-proton collision
p-Pb	Proton-lead atom collision
PVDIS	Parity-Violating Deep Inelastic Scattering
PWG	Physics Working Group
QGP	Quark-gluon plasma
RACF	RHIC and ATLAS Computing Facility
RHIC	Relativistic Heavy Ion Collider
SDAS	STAR Data Acquisition System
SDCC	Scientific Data and Computing Center
SE	Storage elements
SECAR	Separator for Capture Reactions
SeGA	Segmented Germanium Array
SIDIS	Semi-Inclusive Deep Inelastic Scattering
SoLID	Solenoidal Large Intensity Device
SSH	Secure SHell
STAR	Solenoidal Tracker At RHIC
STEM	Science, technology, engineering, and mathematics
SW	Software
SWIF	Scientific Workflow Indefatigable Factotum
TACC	Texas Advanced Computing Center
TB	Terabytes
Tbps	Terabit per second
TPC	Time-projection chamber
UDP	User Datagram Protocol
VO	Virtual Organization
WAN	Wide-Area Network
WLCG	Worldwide LHC Computing Grid
XeXe	Xenon-Xenon