Title

Nuclear Physics Network Requirements Review Final Report

Permalink

https://escholarship.org/uc/item/4qx1b4x8

Authors

Zurawski, Jason
Brown, Ben
Rai, Gulshan
et al.

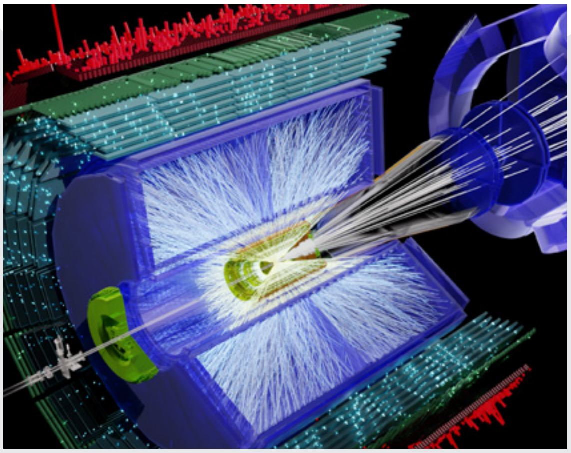
Publication Date

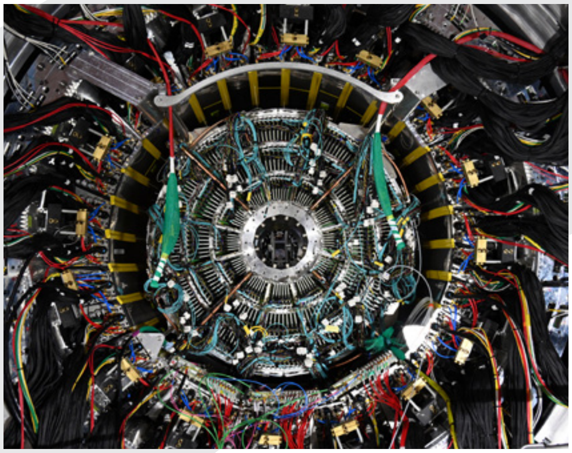
2024-07-26

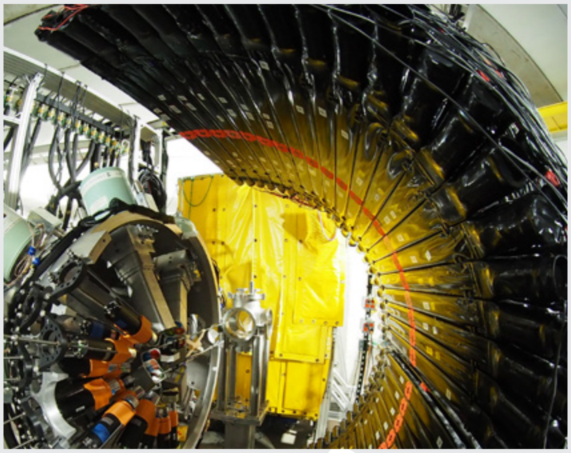
ESnet
ENERGY SCIENCES NETWORK

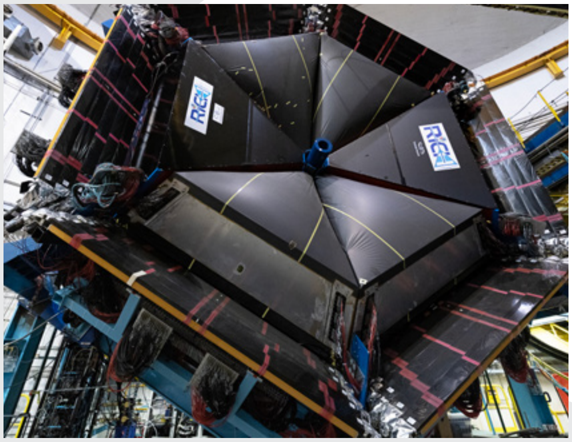
Nuclear Physics Network Requirements Review Final Report

July 2023 – October 2023









BERKELEY LAB

U.S. DEPARTMENT OF ENERGY
Office of Science

ESnet
ENERGY SCIENCES NETWORK

Nuclear Physics

Network Requirements Review Final Report

July 2023 – October 2023

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

ESnet is funded by the US DOE, Office of Science, Office of Advanced Scientific Computing Research. Carol Hawk is the ESnet Program Manager.

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

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

This is LBNL report number LBNL-2001602.

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

Cover Images:
(**Top left**) LHC-ALICE (Credit: Tapan Nayak and Simone Ragoni, CERN)
(**Top right**) BNL-sPHENIX (Credit: Brookhaven National Laboratory)
(**Bottom left**) FRIB-FDSi (Credit: Robert Grzywacz, Oak Ridge National Laboratory, U.S. Department of Energy)
(**Bottom right**) JLab-Hall B (Credit: Aileen Devlin, Jefferson Lab)

[1] https://escholarship.org/uc/item/4qx1b4x8

Participants and Contributors

John Arrington, *Lawrence Berkeley National Laboratory*

Latchezar Betev, *European Organization for Nuclear Research*

Amber Boehnlein, *Thomas Jefferson National Accelerator Facility*

Vincent Bonafede, *Brookhaven National Laboratory*

Ben Brown, *Department of Energy, Office of Science*

Giordano Cerizza, *Facility for Rare Isotope Beams*

Irakli Chakaberia, *Lawrence Berkeley National Laboratory*

Heather Crawford, *Lawrence Berkeley National Laboratory*

Jody Crisp, *Oak Ridge Institute for Science and Education*

Mario Cromaz, *Lawrence Berkeley National Laboratory*

Eli Dart, *Energy Sciences Network*

Cian Dawson, *Energy Sciences Network*

Markus Diefenthaler, *Thomas Jefferson National Accelerator Facility*

Pete Eby, *Oak Ridge National Laboratory*

Robert Edwards, *Thomas Jefferson National Accelerator Facility*

Thomas Evans, *Oak Ridge National Laboratory*

Paul Fallon, *Lawrence Berkeley National Laboratory*

Costin Grigoras, *European Organization for Nuclear Research*

Carol Hawk, *Department of Energy, Office of Science*

Bryan Hess, *Thomas Jefferson National Accelerator Facility*

Graham Heyes, *Thomas Jefferson National Accelerator Facility*

Susan Hicks, *Oak Ridge National Laboratory*

Clinton Jones, *Facility for Rare Isotope Beams*

Reiner Kruecken, *Lawrence Berkeley National Laboratory*

Eric Lancon, *Brookhaven National Laboratory*

Jerome Lauret, *Brookhaven National Laboratory*

Sean Liddick, *Facility for Rare Isotope Beams*

Mark Lukasczyk, *Brookhaven National Laboratory*

Paul Mantica, *Department of Energy, Office of Science*

Spyridon Margetis, *Department of Energy, Office of Science*

Andrew Melo, *Vanderbilt University*

Ken Miller, *Energy Sciences Network*

Nathan Miller, *Energy Sciences Network*

Shigeki Misawa, *Brookhaven National Laboratory*

Brent Morris, *Thomas Jefferson National Accelerator Facility*

Christopher Pinkenburg, *Brookhaven National Laboratory*

Mateusz Ploskon, *Lawrence Berkeley National Laboratory*

Jeff Porter, *Lawrence Berkeley National Laboratory*

Martin Purschke, *Brookhaven National Laboratory*

Gulshan Rai, *Department of Energy, Office of Science*

Kenneth Read, *Oak Ridge National Laboratory*

Thomas Rockwell, *Facility for Rare Isotope Beams*

Brad Sawatzky, *Thomas Jefferson National Accelerator Facility*

Paul Sheldon, *Vanderbilt University*

Richard Simon, *Lawrence Berkeley National Laboratory*

Rune Stromsness, *Lawrence Berkeley National Laboratory*

John White, *Lawrence Berkeley National Laboratory*

Andrew Wiedlea, *Energy Sciences Network*

Alexandr Zaytsev, *Brookhaven National Laboratory*

Jason Zurawski, *Energy Sciences Network*

Report Editors

Jason Zurawski, *Energy Sciences Network,*
zurawski@es.net

Ben Brown, *Department of Energy, Office of Science,*
Benjamin.Brown@science.doe.gov

Gulshan Rai, *Department of Energy, Office of Science,*
gulshan.rai@science.doe.gov

Eli Dart, *Energy Sciences Network,*
dart@es.net

Cian Dawson, *Energy Sciences Network,*
cbdawson@es.net

Carol Hawk, *Department of Energy, Office of Science,*
carol.hawk@science.doe.gov

Paul Mantica, *Department of Energy, Office of Science,*
paul.mantica@science.doe.gov

Spyridon Margetis, *Department of Energy, Office of Science,*spyridon.margetis@science.doe.gov

Ken Miller, *Energy Sciences Network,*
ken@es.net

Nathan Miller, *Energy Sciences Network,*
nmiller@es.net

Andrew Wiedlea, *Energy Sciences Network*
awiedlea@es.net

Table of Contents

1 Executive Summary

The US Department of Energy (DOE) Office of Science (SC) world-class research infrastructure provides the research community with premier observational, experimental, computational, and network capabilities. Each user facility is designed to provide unique capabilities to advance core DOE mission science for its sponsor SC program and to stimulate a rich discovery and innovation ecosystem. Research communities gather and flourish around each user facility, bringing together diverse perspectives. The continual reinvention of the practice of science — as users and staff forge novel approaches expressed in research workflows — unlocks new discoveries and propels scientific progress.

Within this research ecosystem, the high-performance computing (HPC) and networking user facilities stewarded by the SC's Advanced Scientific Computing Research (ASCR) program play a dynamic cross-cutting role, enabling complex workflows demanding high-performance data, networking, and computing solutions. The ASCR facilities enterprise seeks to understand and meet the needs and requirements across SC and DOE domain science programs and priority efforts, highlighted by the formal requirements review methodology.

Between July 2023 and October 2023, the Energy Sciences Network (ESnet) and the Nuclear Physics program (NP) of the DOE SC organized an ESnet requirements review of NP-supported activities. Preparation for these events included identification of key stakeholders: program and facility management, research groups, and technology providers. Each stakeholder group was asked to prepare formal case study documents about its 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.

This review includes case studies from the following NP user facilities, experiments, and joint collaborative efforts:

- Thomas Jefferson National Accelerator Facility (JLab) facilities and experiments.
- Center for Theoretical and Computational Physics and Lattice Quantum Chromodynamics (LQCD) at JLab.
- Brookhaven National Laboratory (BNL): Scientific Data and Computing Center (SDCC).
- Pioneering High Energy Nuclear Interaction eXperiment (sPHENIX) at the Relativistic Heavy Ion Collider (RHIC).
- The Facility for Rare Isotope Beams (FRIB).
- The Gamma-Ray Energy Tracking Array (GRETA).
- ALICE (A Large Ion Collider Experiment) project and ALICE-USA computing.
- The Compact Muon Solenoid (CMS) heavy ion experimentation and The Advanced Computing Center for Research and Education (ACCRE).
- The Electron-Ion Collider (EIC).

The review participants spanned the following roles:

- Subject-matter experts from the NP activities listed previously.
- ESnet Site Coordinators Committee (ESCC) members from NP activity host institutions, including the following DOE labs and facilities: BNL, FRIB, Lawrence Berkeley National Laboratory (LBNL), Oak Ridge National Laboratory (ORNL), JLab, and Vanderbilt University.
- DOE SC staff spanning ASCR and NP.
- ESnet staff.

In recent years, the research communities around the SC user facilities have begun experimenting with and demanding solutions directly integrated with HPC and data infrastructure. This rise of integrated-science approaches is well documented, and there is a broad need for integrated computational, data, and networking solutions. In response to these drivers, DOE has developed a vision for an Integrated Research Infrastructure (IRI)[1] to empower researchers to meld DOE's world-class research tools, infrastructure, and user facilities seamlessly and securely in novel ways to radically accelerate discovery and innovation.

The IRI vision is fundamentally about establishing new data-management and computational paradigms. Within these, DOE SC user facilities and their research communities build bridges across traditional silos to improve existing capabilities and create new possibilities. Implementation of IRI solutions will give researchers simple and powerful tools with which to implement multi-facility research data workflows. This work will also extend analysis done on IRI patterns[2] and discuss ways future NP workflows can benefit from the approach.

1.1 Summary of Review Findings

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

1.1.1 Facilities and Experiments

- ***JLab*** [Section 6.1]
  - JLab compute and storage requirements are estimated by the groups performing experiments.
    - JLab has planned for a roughly two-fold increase in data volumes over the next three years.
    - Each experimental hall currently has local storage close to the detector that can hold at least 48 hours of data. Data flow from the four experimental halls is expected to peak at 24–32 Gbps for the most demanding combination of experiments over the next five to seven years.
    - JLab's data archive consists of an IBM tape library with a current capacity of ~200 PB and 13,000 tapes.
    - The increase in data rates during the 12 GeV era has precipitated an increased use of off-site compute resources.
  - At any time, several experiments are operating at different points in their simulation, calibration, reduction, and analysis phases at JLab. This overlap and interplay of usage patterns leads to a base level of compute load with frequent, "bursty" times of peak demand. Off-site compute allocations offered by other institutions are part of the data processing strategy that "smooths" peak usage.
- ***LQCD*** [Section 6.2]
  - The global analysis effort for LQCD involves collaboration between experimentalists and theorists at approximately 10 universities and research laboratories across the United States. This results in data movement to and from JLab that can total several petabytes (PBs) per year, but in smaller terabyte (TB)-sized data sets. Overall, about 2 PB are generated off-site and then transferred to JLab each year.

[1] https://www.osti.gov/biblio/1984466

[2] https://www.osti.gov/biblio/2205078

- LQCD project teams seek allocations of computing time at numerous HPC facilities. While certain data may be retained for an extended period on leadership systems like the National Energy Research Scientific Computing Center (NERSC), the primary responsibility for long-term data storage lies with the member laboratories. In the case of LQCD projects related to the JLab science program, JLab will serve as the host for the extended data storage.
- The leadership computing facilities (LCFs) do not provide long-term storage for LQCD projects. Data are transferred to the US Lattice Quantum Chromodynamics (USQCD) computing facilities, JLab, Fermilab, and Brookhaven, which assume ownership.
- Data transfers to JLab to support LQCD will increase and scale with the size of new LCFs' systems. JLab will continue to serve as the repository for long-term storage. A portion of the analysis work will be conducted at the LCFs. However, the final stage of the analysis workflow is ideally suited for execution on JLab's local systems, effectively mitigating the disparity in LCF to local computing capability.
- On occasion, HPC systems in Europe are heavily used in collaborations with LQCD European researchers; these initial datasets are then used for secondary calculations also carried out at DOE HPC facilities, resulting in datasets of about 1 PB distributed over about 1,000,000 files.

- ***BNL: The SDCC and sPHENIX at RHIC*** [Section 6.3, 6.4]
  - Currently, the SDCC (Scientific Data and Computing Center) is among the 10 largest High Performance Storage System (HPSS) sites worldwide and hosts in its tape libraries 170 PB of RHIC data. About 90k CPU cores are available in the high throughput computing (HTC) farm for RHIC data processing and analysis, together with about 90 PB of Lustre disk storage. The amount of CPU is expected to double by 2025, while the data volume on tape will be over 500 PB.
  - At present, 52 Data Transfer Nodes (DTNs) are in operation at the SDCC, with the majority being utilized by programs outside of NP.
  - For sPHENIX and the Solenoidal Tracker at RHIC (STAR), datasets are disk-resident at the SDCC, and the vast majority of dataset processing will take place at the SDCC itself. Local computational resources at BNL (located within the SDCC) can be a limiting factor, and the ability to tap nonlocal resources like Open Science Grid (OSG) and other unaffiliated resources have the potential to augment capacity. The ability to utilize these resources, particularly for experiment workflows, will be limited by OSG resource availability, and the ability to transfer data between the host data center and the remote resources.
- ***FRIB*** [Section 6.5]
  - The average FRIB data set sizes ranged from a few GB to ~70 TB with an average size just over 4 TB.
  - Researchers at FRIB are increasingly interested in using off-site HPC and data infrastructure to accomplish specific goals during the execution of an experiment. One experiment group has already employed local Michigan State University (MSU) High-Performance Computing Center (HPCC) resources to expediently analyze incoming data in near real-time to direct decisions during an experiment. The FDSi[3] is exploring the use of NERSC for data analysis during ongoing experiments.

[3] https://www.es.net/news-and-publications/esnet-news/2024/frib-esnet-nersc-collaboration-fuels-high-speed-data-intensive-research

- ***GRETA*** [Section 6.6]
  - The GRETA local computing infrastructure is designed to deliver a full set of science goals. However, the project and scientific user community recognize that advances in the process of data analysis could be improved by using large scale computing (HPC) facilities. In the two-to-five-year timeframe, advances in signal processing algorithms might make the use of remote computing attractive for processing the data for some GRETA experimental scenarios. While support for these potential future activities is outside the scope of GRETA, the modular network architecture provides the flexibility to support the use of external signal processing resources. When located at either FRIB or Argonne National Laboratory (ANL), both facilities must be aware of, and support, GRETA's networking requirements.
  - The GRETA signal decomposition procedure is carried out in real time (within seconds) by a dedicated, co-located computing cluster. The resulting decomposition output, along with any data provided by auxiliary detectors, is provided to the experimental team. These data are cached locally for a period of weeks to allow sufficient time for GRETA experimental teams to transfer the data back to their local institution/computing resource for analysis.
  - GRETA's capabilities will likely represent the most significant performance challenge to the network infrastructure of FRIB or ANL. GRETA data transfer volumes can be between 50 GB and 100 TB, and performed on an ad-hoc basis when the detector is operating. Generally, analysis of GRETA data is carried out by experimental teams at their home institutions. Analysis and data interpretation is a time-consuming process (many months) but not a very computationally intensive process (can be done on local computing resources). The nature of this analysis is very much experiment dependent.
- ***ALICE Project and ALICE-USA Computing*** [Section 6.7]
  - The three-year Large Hadron Collider (LHC) Run 2 period ended in 2018. LHC is currently in Run 3, which began in July 2022. For the ALICE and LHCb experiments, Run 4 marks the beginning of the high-luminosity (HL) LHC era with data rates from the detectors up to 100 times larger than those of Run 2. ALICE data consists of ~50 PB of Compressed Time Frame (CTF)/raw data and ~5 PB of AO2D (Analysis Object). These data are stored at CERN, but also distributed among seven Tier-1 sites. Additional Monte Carlo (MC) data are stored at T1 and T2 sites (around 5PB per data taking period, 60PB overall), and 5 PB of analysis products may be found at analysis facilities.
  - The key data-management feature on the ALICE Grid is during the process of data creation. Data are sent into the computational grid after data collection and synchronous reconstruction, and are attempted to be accessed via "local" (e.g., topographically close) storage during processing.
    - When accessing local data, ALICE may read across wide area networks (WANs), requiring 1 GB/s of aggregate bandwidth. This can result in PB of data read and written across networks such as ESnet.
    - While it is true that ALICE computing is fully distributed, data processing is done locally, and all jobs are executed where the data reside. During the past year, ALICE jobs have read over 2.3 EB and written over 400 PB of data from/to the local storage, averaging about 13 GB/s and 80 GB/s for write and read traffic respectively averaged over the entire grid.
  - ALICE has yearly episodes that require more significant WAN capacities. These occur when storage is added and/or decommissioned, or when data must otherwise be redistributed between different sites. 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.

  - The current and developing ALICE computing models do not have any specific plans for use of cloud resources. However, user analysis within commercial clouds is a possibility if it is cost effective. ALICE research teams found that the use of commercial clouds was fully functional and efficient for running MC simulations, from which the produced simulated data were distributed to remote sites. 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.

- ***The CMS Heavy Ion Experimentation*** [Section 6.8]
  - CMS-HI data-taking periods occur in the final four weeks of the LHC's running year (which typically ends in October or November). Three primary data flows and formats are used for CMS-HI data: RAW, Analysis Object Data (AOD), and Mini-AOD. The volumes of these tiers, in an optimistic scenario where the accelerator performs well for the five-week run, are 27 PB RAW, 17 PB AOD, and 3 PB Mini-AOD.
  - All CMS-HI data are transferred to Fermilab for archival storage, with the overall speed being limited by available bandwidth of the tape archive. The Mini-AOD storage tier is additionally transmitted to ACCRE, located at Vanderbilt University, for access by physicists. The annual data volumes are approximately five times what they were in 2018 (e.g., 40 PB/yr at the T0 and T1, and 3 to 10 PB/yr at T2s). CMS-HI expects an additional increase by a factor of two in High Luminosity-LHC (tentatively scheduled for 2029).
  - CMS and US CMS support a few different ways to access data. Batch processing on local and grid-enabled resources remains popular, but recent years have seen an increase in interactive jobs that may leverage some command-line tools, or the use of Jupyter Notebook. In both cases, there are two main data access interfaces: Portable Operating System Interface (POSIX) mounts for local file access or remote XRootD access via the AAA (Any data, Any time, Anywhere) data federation, which federates CMS's global XRootD access points into a single access point.
  - For CMS-HI, the primary limitation is tape bandwidth available to store and recall large multi-PB datasets. While the capacity of tape cartridges greatly increases when new generations of technology are released, the bandwidth per tape has not kept pace. If trends continue, in 2030 it could take nearly a day of continuous access to read/write a tape from beginning to end. Writing to and retrieving from custodial tape is by far the most difficult issue for data movement CMS-HI will face in the coming years.
  - The largest sources/sinks of CMS-HI data are the XRootD endpoints at CERN, Fermi National Accelerator Laboratory (Fermilab), Vanderbilt, and MIT. After their initial production, very little data movement happens on the large centrally produced datasets. There is approximately TB-scale data movement of user-produced datasets to other facilities. This trend is expected to be stable until 2029, when the HL-LHC program begins. With the increased detector granularity and data acquisition rate, these numbers can scale between 5 and 10 times today's numbers.
  - CERN has, and is expected to continue to have, sufficient resources to produce smaller derived datasets for CMS-HI, and no need for DOE SC user facilities is foreseen (this is a notable difference between CMS-HI and CMS-PP).
  - An identified issue with the CMS-HI data reduction pattern is that some popular portions of

datasets are reproduced many times. At the predicted scales of HL-LHC (predicted to begin in 2029), this data duplication becomes financially burdensome to support. By storing data in object stores, these common slices of datasets could be stored and referenced by multiple end-user datasets, providing a better space efficiency for analysis. These benefits become more pronounced if there is a single copy globally of the relevant objects; this implies some level of additional WAN traffic needed to satisfy these workflows. These techniques are the target of active R&D, and initial results are expected in the 2025–2026 timescale.

  - CMS-HI participates in a number of R&D activities that may influence operations prior to the HL-LHC era. This research is ongoing, and a decision on the deployment would be made in the 2026–2027 time frame.
    - CMS-HI predicts an increase in remote science usage, namely accessing graphics processing unit (GPU) resources, via the Services for Optimized Network Inference on Coprocessors (SONIC) inference-as-a-service activity.
    - CMS is participating in Software-Defined Network for End-to-end Networked Science at the Exascale (SENSE)-Rucio R&D, which will use ESnet's ability to provide guaranteed point-to-point bandwidth to more effectively schedule data transfers. Rucio can decide to signal to the SENSE dataplane that it would like guaranteed bandwidth between two sites, and if the request is accepted, configure transfers for specific dataset(s) to transit exclusively over that guaranteed bandwidth.
    - The Network Optimized for Transfer of Experimental Data (NOTED) project uses packet marking to provide Quality of Service (QoS) guarantees, thus ensuring that bandwidth to support data transfers over tools like Rucio can remain high to support distributed job execution.

- ***The EIC*** [Section 6.9]
  - The EIC (Electron Ion Collider) is a new community-driven facility that will be constructed in the 2020s and is projected to begin data collection in 2033. The effort to estimate data volumes from the EIC's ePIC detector is in progress. Raw data rates could be hundreds of Pbps, with reduction to hundreds of Gbps possible. BNL and JLab have formed the EIC Computing and Software Joint Institute to serve as a single point of contact and organizational entity for support of the ePIC collaboration and other software and computing needs for the EIC.
  - EIC networking and computation decisions are not imminent, and will consider lessons learned from LHC computing and analysis. It is expected that major collaboration sites will contribute resources and leverage ESnet connectivity as a critical networking link to manage traffic flows. The ability to handle PB volumes of data over time and bursts of hundreds of Gbps is expected.
  - For the ePIC detector at the EIC, the ability to process data remotely is an integral part of the proposed computing model. Access to storage resources and sufficient network bandwidth to move or access data to and from remote sites is a prerequisite.

1.1.2 Cross-Cutting Data Management, Workflow, Computing, Storage, and Networking

- Current NP experiments at colliders collect large quantities of data over the course of their multi-year lifespan, typically at a rate of over hundreds of PBs per year. [Section 7]
- The use of DOE HPC facilities by NP facilities and experiments, as well as those provided by distributed grid resources like the OSG (Open Science Grid), will continue to grow. It is

expected that the volume of data generated at LCFs will continue to increase by a factor of 5 to 10 for NP workflows that use these resources. [Section 7]

- Rucio[4], a software package that manages large volumes of data spread across facilities at multiple institutions and organizations, is allowing NP facilities to re-imagine the data pipeline from the experiments, creating a single source of truth (i.e., all data are edited in only one location) for policy-based file movement of raw data around campus environments. This incorporates the use of OSG resources that can be leveraged both locally and external to the facility to deliver on computational tasks. [Section 7]
- Experimental and MC (Monte Carlo) data workflows demand substantial computational resources. The MC workflow is generally executed across multiple sites. This includes not only those sites directly affiliated with the experiment but also independent ones such as those connected to the OSG and potentially even the DOE leadership class HPC facilities (such as NERSC). The primary reason for this distribution is the computational intensity of MC workflows, which demand significant processing power but use relatively minimal data. [Section 7]
- For many NP experiments, utilization of unaffiliated computing resources not at the host data centers is a given. OSG and the supercomputers at DOE facilities (mainly NERSC for the STAR experiment at BNL, and GlueX and CLAS12 from JLab) are commonly used for simulations and are two examples of remote, unaffiliated computing resources. [Section 7]
- With the advent of new computing facilities such as Perlmutter at NERSC and Frontier at Oak Ridge Leadership Computing Facility (OLCF), the volume of data generated at LCFs is increasing by a factor of 5 to 10 for NP workflows that use these resources. Existing workflows will remain in use, leading to a corresponding increase in the amount of data that need to be transferred back to home institutions. The Aurora system at Argonne Leadership Computing Facility (ALCF) will likely mark a significant milestone in data production. An estimated three times more data are expected to be generated on exascale systems compared with the current generation of systems. [Section 7]
- Since 2020, an increasing number of NP facility users participate remotely via login to computing resources and remote conferencing facilities, and by transferring data to and from facilities. Remote science activities routinely leverage computational resources provided at partner sites. These could be located at DOE HPC facilities or distributed computing resources such as those provided by the OSG. Providing capabilities for remote users to observe the products of ongoing data analysis continues to be beneficial to increase engagement with the user community. [Section 7]
- NP simulation workflows are capable of running off-site, since the input data required are small versus those of an analysis workflow. In some cases, an experiment may require that the output of simulation runs be returned to a source institution for storage or future analysis, and this will contribute to incoming WAN traffic. This may result in traffic demands that are similar in scale to reconstruction workflow's contribution to the outgoing WAN traffic. [Section 7]
- NP facilities, experiments, and researchers have noted that integrating with other labs and data facilities presents ongoing challenges in authentication and federated identity. While frameworks like x.509, OAuth, and tokens have broad consensus, issues arise from nonstandard or incomplete implementations. Federated interactive logins that leverage approaches such as SAML, OpenAuth2, and Shibboleth are standardized and interoperable, but "automated" authentication for tasks like remote job control and large file transfers is fragmented and

[4] https://rucio.cern.ch/

challenging to debug. [Section 7]

- NP facilities, experiments, and researchers have noted that facility-to-facility trust implementations are also difficult from both policy and technical perspectives. Cross-facility systems often rely on long-lived "robot" tokens or certificates that are an extension of a user's identity. These are not always a good fit for long-term campaigns. It is often the case that actions need to be taken on behalf of or as a proxy for users, which can stretch the abilities of federated identities. [Section 7]
- NP experiments and facilities leverage several data mobility tools for sharing experimental data. CVMFS, XRootD, File Transfer Service (FTS), Rucio, and Globus are all used when exchanging data with collaboration sites (e.g., DOE HPC Facilities, OSG participants) and with end users. [Section 7]
- The majority of NP facilities and experiments (located at BNL, JLab, FRIB, ORNL, LBNL, and Vanderbilt) are connected to ESnet with a capacity of 100 Gbps, and several will upgrade to multiple 100 Gbps or 400 Gbps in the coming years to support increases in data volumes. [Section 6]
- NP facilities and experiments continue to investigate use of commercial cloud services for data processing, in particular for the case where a rapid turnaround is required that exceeds resources available locally or via dedicated (e.g., DOE HPC) or distributed computing (e.g., OSG) facilities. Due to the current cost associated with cloud computing, use is expected to be rare, and WAN requirements will not significantly increase. [Section 7]

1.1.3 IRI (Integrated Research Infrastructure) Responsiveness

- The availability of uniform interfaces at experimental facilities, HPC facilities, and network facilities will strengthen NP workflows, as well lead to an increase in external resource uses such as what is seen today between JLab and NERSC. IRI (Integrated Research Infrastructure) will allow for better EIC integration at BNL and JLab for the long-term campaign pattern, as well as other possibilities such as time-dependent workflows being managed across sites. [Section 6.1, 6.9]
- In light of the crucial role of authentication in the IRI vision, a central repository with reference designs and implementation examples for common workflows would offer significant value. [Section 7]
- A number of limitations to addressing future needs exist at FRIB, many of which can be mitigated using emerging approaches to multi-facility workflows and the IRI activity: [Section 6.5]
  - Not enough computing capability to address new instrumentation or on-site analysis.
  - Limited staff availability to adapt or convert HPC workflows to operate within FRIB.
  - No staff expertise that is capable of leveraging capabilities at other DOE facilities (e.g., DOE HPC centers, ESnet).
- The GRETA data pipeline was designed with IRI workflows in mind, and this option is actively being developed. GRETA's forward buffers can send their data over WANs to remote HPC facilities where the main data processing tasks could be carried out on interactive timescales (time-sensitive pattern). These workflows are currently being evaluated using the ESnet testbed and OLCF/ORNL IRI testbed and are expected to be ready for production use in a two-year time frame. [Section 6.6]

1.2 Summary of Review Actions

Lastly, ESnet will follow up with review participants on a number of high-level actions identified. These items are listed as guidance for future collaboration, and do not reflect formal project timelines. ESnet will review these with NP participations on a yearly basis, until the next requirements review process begins.

1.2.1 Facilities and Experiments

- ESnet will continue to monitor the network speed needs of NP facilities, and schedule upgrades to capacity and services as required. The major NP facilities and experiments (located at BNL, JLab, FRIB, ORNL, LBNL, and Vanderbilt) are all connected to ESnet with a capacity of 100 Gbps, and several will upgrade to multiple 100 Gbps or 400 Gbps in the coming years to support increases in data volumes.
- ESnet and LQCD research will discuss mechanisms to expose and transfer data from the LQCD effort at major DOE HPC facilities, as the data volumes of the data products increases to the multiple PB volume in the coming years. Deployment of high-performance portal software (e.g., based on the Globus Modern Research Data Portal) could simplify the delivery of data to collaborators.
- ESnet, LBNL, and NERSC will work with ALICE-USA to adapt their primary workflow to utilize streaming versus bulk data movement. These changes will have an impact on network use, particularly within the LBL campus, and the interconnection to ESnet.
- ESnet and ALICE-USA will work to leverage the OSG network group's perfSONAR dashboards. Implementation of this sensible design for systematic, automated WAN monitoring is very important for efficient use of ALICE-USA resources.
- ESnet will continue to work on the design and implementation of DTNs to support data-mobility needs. JLab and ALICE-USA have requested help in this area.
- ESnet, CMS-HI, and ALICE-USA will continue to coordinate on supporting operations during Run 3 (which began in July 2022), through the planning and execution of Run 4 (the HL-LHC era). This support will include coordination on data challenge (DC) activities, increasing capabilities through the development and deployment of R&D projects (SENSE, NOTED), and delivering measurement and monitoring frameworks (e.g., Stardust, perfSONAR).
- BNL, JLab, and ESnet will coordinate on the needs of the EIC during the design, construction, and implementation process. Due to the potential design implications that will result in raw data rates of hundreds of Pbps, with reduction to hundreds of Gbps, networking between the sites to support computation and storage will be critical.

1.2.2 Cross-Cutting Data Management, Workflow, Computing, Storage, and Networking

- ESnet and Rucio developers will continue to coordinate on software package development and how it relates to R&D efforts such as SENSE and NOTED. Rucio is allowing NP facilities to re-imagine the data pipeline from the experiments, creating a single source of truth (i.e., all data are edited in only one location) for policy-based file movement of raw data around campus environments.
- ESnet will continue to investigate the use of commercial cloud services for data processing. Due to the current cost of cloud computing, use cases are expected to be rare, and when used it will replace one or more of the other resources so the WAN requirements will not significantly increase. This work will benefit NP facilities and experiments, along with others in DOE SC.

1.2.3 IRI Responsiveness

- The IRI effort, led by members across DOE SC programs, must address the availability of uniform interfaces at experimental facilities, HPC facilities, and network facilities. Doing so will strengthen facility and experimental workflows, as well as lead to an increase in external resource use. The NP community will be a beneficiary of this work. In light of the crucial role of authentication in the IRI vision, a central repository with reference designs and implementation examples for common workflows would offer significant value.
- The IRI effort, led by members across DOE SC programs, must address some of the ongoing challenges in authentication and federated identity. While frameworks like x.509, OAuth, and tokens have broad consensus, issues arise from nonstandard or incomplete implementations. Federated interactive logins that leverage approaches such as SAML, OpenAuth2, and Shibboleth are standardized and interoperable, but "automated" authentication for tasks like remote job control and large file transfers is fragmented and challenging to debug. Addressing these challenges will benefit the NP community as well as other DOE SC programs.
- ESnet will continue to consult with FRIB on its workflow and data mobility needs. A number of limitations to addressing future needs exist at FRIB, many of which can be mitigated using emerging approaches to multi-facility workflows and the IRI activity.
- The IRI effort, led by members across DOE SC programs, must address the long-standing policy issue surrounding facility-to-facility trust implementations. Cross-facility systems often rely on long-lived "robot" tokens or certificates that are an extension of a user's identity. These are not always a good fit for long-term campaigns. Actions often need to be taken on behalf of or as a proxy for users, which can stretch the abilities of federated identities. Addressing these challenges will benefit the NP community, as well as other DOE SC programs.

2 Requirement Review Overview

ESnet and ASCR use 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.

2.1 Purpose and Process

The requirements review process, when performed regularly and comprehensively, surveys major science stakeholders' plans and processes to investigate data-management requirements over the next 5–10 years. Questions crafted to explore this space include the following:

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

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 leads the effort and relies on collaboration from other ESnet teams: Software Engineering, Network Engineering, and Network Security. This team meets with each individual program office within the DOE SC every three years, with an intermediate virtual update scheduled between the full review. ESnet collaborates with the relevant program managers to identify the appropriate principal investigators, and their information technology partners, to participate in the review process. ESnet organizes, convenes, executes, and shares the outcomes of the review with all stakeholders.

Requirements reviews are a critical part of a process to understand and analyze current and planned science use cases across the DOE SC. 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.

We achieve these goals by reviewing the case study documents, discussions with authors, and general analysis of the materials. The resulting output is a set of review findings and actions that will guide future interactions between NP, ASCR, and ESnet. These terms are defined as follows:

- ***Findings:*** key facts or observations gleaned from the entire review process that highlight specific challenges, particularly those shared among multiple case studies.
- ***Actions:*** potential strategic or tactical activities, investments, or opportunities that can be evaluated and potentially pursued to address the challenges laid out in the findings.

2.2 Structure

The requirements review process is hybrid, and relies on a combination of asynchronous and synchronous activities to understand specific facility and experimental use cases. The review is 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 implemented in the coming years.

2.2.1 Background

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

- 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 and 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 produced.
- Current and anticipated data output on near- and long-term timescales.
- Timeline(s) for building, operating, and decommissioning of experiments, to the degree these are known.
- Existing and planned network resources, usage, and "pain points" or bottlenecks in transferring or productively using the data produced by the science.

2.2.2 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 studies address the following sections with review participants:

Science Background: a brief description of the scientific research performed or supported, the high-level context, goals, stakeholders, and outcomes. The section includes a brief overview of the data life cycle and how scientific components from the target use case are involved.

Collaborators: aims to capture the breadth of the science collaborations involved in an experiment or facility focusing on geographic locations and how datasets are created, shared, computed, and stored.

Instruments and Facilities: 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. An overview of the composition of the datasets produced by the instrument or facility (e.g., file size, number of files, number of directories, total dataset size) is also included.

Process of Science: 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, and the integration of experimental data with simulation data or other use cases.

Remote Science Activities: use of any remote instruments or resources for 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: discussion of 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: the network architecture and bandwidth for the facility and/or laboratory and/or campus. The section includes detailed descriptions of the various network layers (local area network (LAN), MAN, and WAN) capabilities that connect the science experiment/facility/data source to external resources and collaborators.

IRI Readiness: Research communities that utilize DOE SC user facilities are experimenting with and demanding solutions integrated with HPC and data infrastructure. The Integrated Research Infrastructure Architecture Blueprint Activity (IRI-ABA) brought together domain experts from all DOE SC Programs to look for common patterns within diverse workflows across a range of scientific disciplines. Participants discovered three common patterns:

- Time-sensitive pattern.
- Data integration-intensive pattern.
- Long-term campaign pattern.

Participants are asked to discuss if experimental or facility workflows exhibit any of the IRI patterns:

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

Data-Related Resource Constraints: 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.

Data Mobility Endpoints: If a facility or experiment has dedicated infrastructure to facilitate data sharing, ESnet is interested in learning more about how it is constructed and maintained. ESnet maintains a set of well-tuned test endpoints and recommends regular testing to evaluate data transfer capabilities.

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

2.3 ESnet

ESnet is the high-performance network user facility for the US DOE SC and delivers highly reliable data transport capabilities optimized for the requirements of data-intensive science. In essence, ESnet is the circulatory system that enables the DOE science mission by connecting all its laboratories and facilities in the US and abroad. ESnet is funded and stewarded by the ASCR program and managed and operated by the Scientific Networking Division at LBNL. ESnet is widely regarded as a global leader in the research and education networking community.

ESnet interconnects DOE national laboratories, user facilities, and major experiments so that scientists can use remote instruments and computing resources as well as share data with collaborators, transfer large datasets, and access distributed data repositories. ESnet is specifically built to provide a range of network services tailored to meet the unique requirements of the 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.

2.4 About ASCR

The mission of the ASCR program is to discover, develop, and deploy computational and networking capabilities to analyze, model, simulate, and predict complex phenomena important to the DOE. A particular challenge of this program is fulfilling the science potential of emerging computing systems and other novel computing architectures, which will require numerous significant modifications to today's tools and techniques to deliver on the promise of exascale science.

To accomplish its mission and address the challenges described previously, the ASCR program is organized into two subprograms:

- The Mathematical, Computational, and Computer Sciences Research subprogram develops mathematical descriptions, models, methods, and algorithms to describe and understand complex systems, often involving processes that span a wide range of time and/or length scales.
- The HPC and Network Facilities subprogram delivers forefront computational and networking capabilities and contributes to the development of next-generation capabilities through support of prototypes and test beds.

2.5 About the NP Program

Nuclear physics seeks to understand matter in all of its manifestations. This applies to not just the familiar forms of matter seen around us, but also such exotic forms as the matter that existed in the first moments after the creation of the universe. Nuclear science is the investigation of how protons and neutrons are formed from elementary particles and how the forces between those particles produce both nuclei and the vast variety of nuclear phenomena that occur in the universe.

The mission of the NP program is to discover, explore, and understand all forms of nuclear matter. The fundamental particles that compose nuclear matter, quarks and gluons, are relatively well understood, but exactly how they fit together and interact to create different types of matter in the universe is still not fully explained. To solve this mystery, NP supports experimental and theoretical research — along with the development and operation of particle accelerators and advanced technologies — to create, detect, and describe the different forms and complexities of nuclear matter that can exist in the universe, including those that are no longer found naturally.

Nuclear physics has come to focus on four broad yet tightly interrelated areas of inquiry. These areas are described in A New Era of Discovery, a long-range plan for nuclear science released in 2023 by the Nuclear Science Advisory Committee (NSAC)[5]. The plan represents a consensus within the nuclear science community about compelling scientific thrusts. The following are the four science questions the long-range plan identified:

- How do quarks and gluons make up protons, neutrons, and ultimately, atomic nuclei?
- How do the rich patterns observed in the structure and reactions of nuclei emerge from the interactions of neutrons and protons?

[5] https://science.osti.gov/-/media/np/nsac/pdf/reports/2024/2024-NSAC-LRP-Report_Final.pdf

- What are the nuclear processes that drive the birth, life, and death of stars?
- How do we use atomic nuclei to uncover physics beyond the Standard Model?

NP has embraced the use of HPC, artificial intelligence (AI), machine learning (ML) and Quantum Information Science (QIS) technologies, all of which have led to remarkable scientific progress for nuclear physics, enabled in part by collaboration with computational scientists and applied mathematicians. As the DOE enters the era of exascale computing, with increasing numbers of communities within nuclear physics poised to take advantage of HPC, enhanced support will maximize scientific progress. Support for a coordinated effort to integrate AI/ML technologies into the nuclear physics research programs will accelerate discoveries.

3 Assessment of 2019 NP Requirement Review

ESnet recorded a set of action items from the *2019 NP Network Requirements Review*, with the goal of continuing support of collaborations funded by the NP program beyond the publication of the report. ESnet performed regular check-ins with NP between the reviews (holding webinars with the community in 2021 and 2022), to ensure progress was made on these identified areas of mutual interest. These sections outline the previous actions and discuss progress made since the publication of the report.

3.1 Prior Action Items

ESnet will take the following steps:

- Start a discussion between ESnet engineering and experimental representatives interested in sharing ESnet 6 telemetry data.
- Consider the creation of LHC 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 / Eastern Lightwave Internetworking Technology Enterprise (E-LITE) / 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 Modern Research Data Portal design considerations.
- Facilitate peering with commercial clouds, as needed, for experiments looking into pilot efforts.

The NP community will take the following steps:

- Continue discussions with ASCR facilities about 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 the Exascale Computing Project (ECP) 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 take the following steps:

- 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).

3.2 Assessment

The *2019 NP Network Requirements Review* action items were assessed and reviewed with the NP community. This section does not offer a line-by-line account for each, but instead provides an update on several major initiatives undertaken by the DOE SC programs as a whole, along with individual actions by ASCR, NP, and ESnet.

3.2.1 ESnet Actions

Since the 2019 review, ESnet completed the design, procurement, and installation of ESnet6, the next-generation networking platform that is currently running to deliver on DOE SC mission requirements. This was a multi-year project that involved complete replacement of the underlying optical and routed network that ESnet operates and resulted in upgraded capacities across the backbone and to connected laboratories and facilities. Along with the upgraded network capacities, a rich set of new measurement, monitoring, telemetry, and data-handling services were installed and integrated into a number of scientific workflows. ESnet has also initiated the following efforts and programs that address some of the actions from the *2019 NP Network Requirements Review*:

- Deployment of the Stardust measurement and monitoring platform, which allows ESnet connectors a way to visualize and categorize their scientific network patterns.
- Upgrades to all perfSONAR and DTN hardware to at least 100 Gbps capable speeds
- Full adoption and support of the LHCONE networking overlay, and the ability of other affiliated scientific efforts (e.g., Belle II, ALICE, etc.) to utilize this for network traffic that traverses similar computational sites and network paths.
- Upgrades to network capacities, including many to multiple 400 G connections, at connected sites. With upgrades, a renewed focus was made on ways that ESnet can upgrade and support critical peering with national and international research and education networks, as well as commercial partners that operate cloud computing infrastructures.
- Discussions with regional networking partners to ensure that ESnet connected sites are meeting minimum connectivity requirements of at least 100 Gbps, with options to add secondary and backup connections at similar capacities.
- An organizational commitment to participate in “co-design” efforts. These joint projects between ESnet engineering staff, along with scientific experts at other facilities and projects, afford projects such as GRETA and EJ-FAT access to technical expertise and build strong relationships throughout the scientific community.
- The creation of a program (i.e., the Fasterdata DTN Framework) as a mechanism to support data mobility needs across the backbone and regional networking infrastructure, as well as touchpoints in campus environments.

3.2.2 DOE SC Actions

A major focus for ASCR, along with other DOE SC programs like NP, in the time between the *2019 NP Network Requirements Review* and *2023 NP Network Requirements Review* was the convening of groups to discuss ways that the future generation of technology provided on the network, and within computational facilities, will grow to meet the demands of scientific experimentation.

The complexity of scientific pursuits is increasing rapidly with aspects that require dynamic integration of experiment, observation, theory, modeling, simulation, visualization, AI/ML, and analysis. Research projects across the DOE are increasingly data and compute intensive. Innovative research teams are accelerating the pace of discovery by using high-performance computational and data tools in their research workflows and leveraging multiple research infrastructures. DOE SC has embraced what can be possible with this integration and has adopted a vision for IRI:[6] to empower researchers to meld DOE's world-class research tools, infrastructure, and user facilities seamlessly and securely in novel ways to radically accelerate discovery and innovation.

IRI defined three core patterns during its early work:

- ***Time-sensitive pattern:*** requires real-time, or near real-time, response across more than one facility or resource for timely decision-making and experiment steering.
- ***Data integration–intensive pattern:*** requires combining and analyzing data from multiple sources, e.g., sites, experiments, and/or computational runs, to deepen and expand context.
- ***Long-term campaign pattern:*** requires sustained access to more than one facility or resource, often centered around rare or unique scientific instruments, over years or decades to accomplish a well-defined objective.

These activities address some of the actions from the *2019 NP Network Requirements Review*, namely in trying to define a sensible set of requirements that DOE HPC facilities can deliver to assist the future needs of experimental workflows and better scale service offerings for emerging the Electron Ion Collider. These observations come from the broader work to classify ESnet Requirements Reviews through the IRI lens[7].

IRI acknowledges that data movement, be it in the form of bulk or streaming requirements, is a core goal and must be supported fully by all experimental facilities that connect to ESnet. ESnet is committed to working with NP, and other DOE SC programs, to execute on the vision of IRI starting with the items identified in the *2023 NP Network Requirements Review.*

[6] https://www.osti.gov/biblio/1984466
[7] https://www.osti.gov/biblio/2008205

4 Review Findings

The requirements review process helps to identify important facts and opportunities from the programs and user facilities that are profiled. The following sections outline a set of findings from the NP and ESnet requirements review. These points summarize important information gathered during the review discussions surrounding case studies and the NP-managed user program in general. These findings are organized by topic area for simplicity and by common themes:

- Facility management and readiness.
- Scientific data management.
- Scientific workflow.
- Computational and storage requirements.
- Remote collaboration and operational requirements.
- Multifacility computational workflows.
- Domestic networking for local and wide-area data mobility.
- Emerging needs.
- IRI responsiveness.

4.1 Facility Management and Readiness

- JLab has an international user community of over 1,800 active users (324 institutions in 39 countries.). One-third of all PhDs granted in nuclear physics in the US are based on JLab research. The primary scientific instrument is the Continuous Electron Beam Accelerator Facility (CEBAF). CEBAF is a high-intensity electron accelerator with unique capabilities to probe the nuclear structure of matter at the quark level. [Section 6.1]
- JLab is an integral part of the USQCD collaboration, a consortium consisting of approximately 160 individuals from around 50 institutions, including universities and research laboratories. [Section 6.2]
- The RHIC collider at BNL is a world-class particle accelerator exploring the most fundamental forces and properties of matter and the early universe. RHIC accelerates beams of particles (e.g., the nuclei of heavy atoms such as gold) to nearly the speed of light, and smashes them together to recreate a state of matter thought to have existed immediately after the Big Bang. The RHIC facility has approximately 1,000 unique users. [Section 6.3]
- The sPHENIX and STAR detectors are the two operating detectors at RHIC. The RHIC Computing Facility (RCF) at the SDCC at BNL hosts the storage and computing resources used by the RHIC experiments. These detectors collect large quantities of data over the course of their multi-year lifespan, typically at a rate of over hundreds of PBs per year. [Section 6.3]
- sPHENIX is the first major upgrade to a nuclear physics heavy-ion experiment in the US in two decades, and along with STAR will be the final experiment taking data at the RHIC before the construction of the EIC. The sPHENIX Collaboration consists of 81 member institutions and about 400 collaborators from 14 countries. [Section 6.4]
- The operation and data analysis of the sPHENIX experiment is centered at BNL. BNL's SDCC currently provides most of the data storage and computing resources for the experiment and the collaboration. [Section 6.4]

- FRIB is a DOE-SC scientific user facility serving users organized in the FRIB Users Organization (FRIBUO). The FRIBUO has over 1,800 members representing 124 US colleges and universities, 13 national laboratories, and 52 countries. FRIB was designed to be the world's most powerful rare isotope research facility and commenced user operation in the latter half of FY 2022. [Section 6.5]
- GRETA is an advanced gamma-ray spectrometer for low-energy nuclear physics measurements and is currently in the final stages of fabrication. The array is a primary instrument for FRIB. 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. [Section 6.6]
- GRETA is nearing the end of its fabrication phase. 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. GRETA is currently planning for start of operations in 2024 and project completion (all detectors, full rate) in 2027. [Section 6.6]
- GRETA will be initially sited at FRIB with the possibility of later operations at Argonne Tandem Linac Accelerator System (ATLAS)/Argonne. It is expected that once GRETA has completed construction and been sited at FRIB, it will be operated by a local operations team with technical support provided by LBNL staff. [Section 6.6]
- The ALICE collaboration has constructed and operates a heavy-ion detector to exploit the unique physics potential of proton-proton and nucleus-nucleus interactions at collision energies of the LHC at CERN. The ALICE Collaboration consists of over 2,000 scientists, engineers, and students spread over 170 institutions in 41 countries. [Section 6.7]
- The ALICE-USA computing project was established to meet US obligations by operating ALICE Grid facilities in the US. The initial ALICE-USA obligations corresponded to about 6% of all ALICE computing resource needs and are currently about 8% of those requirements. [Section 6.7]
- The heavy ion program using the CMS detector at the LHC probes the conditions of the early universe by studying collisions of heavy nuclei at relativistic speeds (usually between two heavy nuclei, though some data were recorded of protons striking lead nuclei). [Section 6.8]
- The EIC is a new community-driven facility that targets the exploration of quantum chromodynamics (QCD) to high precision, with a particular focus on unraveling the quark-gluon substructure of the nucleon and of nuclei. The EIC will investigate the structure of nucleons and nuclei by performing precise measurements of deep-inelastic scattering (DIS) and other processes over the complete relevant kinematic range including the transition region from perturbative to nonperturbative QCD. The EIC users' group has roughly 1,400 members from nearly 300 institutions and 40 countries. [Section 6.9]
- The EIC will be constructed in the 2020s, with an extensive science case as detailed in the EIC white paper, the 2023 NSAC Long Range Plan for Nuclear Science, and the EIC yellow report. The yellow report has been an important input to the successful DOE CD-1 review and decision. It describes the physics case, the resulting detector requirements, and the evolving detector concepts for the experimental program at the EIC. The first scientific collaboration for the EIC, ePIC, was formed in 2023 to support the realization of the EIC project detector. [Section 6.9]

- The ePIC Collaboration, consisting of almost 500 members from 171 institutions, is working jointly with the EIC project team to design and establish the ePIC detector, which is poised to be primed and ready for data collection once the EIC springs into action in the early 2030s. Work on a second detector at the EIC is also in progress by a yet unnamed collaboration. [Section 6.9]
- The EIC facility is projected to begin data collection in 2033. While the guiding details of the ePIC detector, data acquisition systems, and analysis workflows are mostly understood, the final design is still in progress. The next major milestone is a technical design report (TDR) which will include studies of the detector components. No major networking requirements are anticipated before the next requirements review in 2027, when scale tests of the computing model take place. EIC/ePIC-focused R&D on streaming readout and reconstruction frameworks is expected to ramp up in the latter half of this decade to be ready for full production in the 2030s. [Section 6.9]
- EIC networking and computation decisions are not imminent, and will consider lessons learned from LHC computing and analysis. It is expected that major collaboration sites will contribute resources and leverage ESnet connectivity as a critical networking link to manage traffic flows. The ability to handle PB volumes of data over time and bursts of hundreds of Gbps is expected. [Section 6.9]

4.2 Scientific Data Management

- The raw data from JLab CEBAF experiments are shared with experimenters, and a backup is stored locally on tape. 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 the raw data taken in the lifetime of the laboratory are still stored in the tape library at the lab. Approximately 100 PB of data a year are produced through a mixture of raw scientific observation, analysis data, simulation, and backups of affiliated data sets. [Section 6.1]
- LQCD data retention at HPC facilities is typically short-term. The data are subsequently transferred back to local computing resources for long-term storage. [Section 6.2]
- The data from the sPHENIX data acquisition (DAQ) system are written to a Lustre file system and are immediately processed by the HTC farm. A copy of the data is also sent to the HPSS tape system for archiving at a sustained rate of 10 GB/sec. For STAR, data are sent by the DAQ system to HPSS, with the HTC farm retrieving data from the cache (FastOffline) or tape for first-pass processing. [Section 6.3]
- Currently, the SDCC is among the 10 largest HPSS sites worldwide and hosts in its tape libraries 170 PB of RHIC data. About 90k CPU cores are available in the HTC farm for RHIC data processing and analysis, together with about 90 PB of Lustre disk storage. The amount of CPU is expected to double by 2025, while the data volume on tape will be over 500 PB. [Section 6.3]
- No large-scale export of data or the routine use of grid services is foreseen for sPHENIX, although the possibility of hosting additional second-tier copies in particular of DSTs in other regions has been discussed. [Section 6.4]
- The average FRIB data set sizes ranged from a few GB to ~70 TB with an average size just over 4 TB. For the first year of FRIB operation, the data sizes ranged from a few GB to 24 TB with an average size just over 4 TB. [Section 6.5]
- The raw data collected from the GRETA detectors are the energies, times, and associated waveforms as captured in the ADCs/ field programmable gate arrays (FPGAs) which instrument

GRETA. From these data the location of gamma-ray interaction points can be inferred 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, are provided to the experimental team of a given measurement. These data are cached locally for a period of weeks to allow sufficient time for GRETA experimental teams to transfer the data back to their local institution/computing resource for analysis. [Section 6.6]

- The three-year LHC Run 2 period ended in 2018. LHC is currently in the Run 3 which began in July 2022. For the ALICE and LHCb experiments, Run 3 marks the beginning of the high-luminosity (HL) LHC era with data rates from the detectors reaching up to 100 times larger than those of Run 2. [Section 6.7]
- ALICE data consist of ~50 PB of CTF/raw data, and ~5 PB of AO2D data. These data are stored at CERN, but also distributed among seven Tier-1 sites (Germany, France, Italy, UK, Netherlands, Korea, and Russia). Additional MC data are stored at T1 and T2 sites (around 5PB per data taking period, 60PB overall), and 5 PB of analysis products may be found at analysis facilities. [Section 6.7]
- CMS-HI data are recorded when the LHC performs a heavy-ion run, where instead of colliding protons, the LHC collides heavier nuclei such as lead, xenon, or oxygen. In theory, these data-taking periods occur in the final four weeks of the LHC's running year (which typically ends Oct or Nov), though this is subject to change due to external factors. [Section 6.8]
- CMS-HI has three primary data flows. The data are created at CERN in three different data tiers: RAW, AOD, and Mini-AOD. The volumes of these tiers, in an optimistic scenario where the accelerator performs well for the five-week run, are 27 PB RAW, 17 PB AOD, and 3 PB Mini-AOD. [Section 6.8]
- The effort to estimate data volumes from the EIC's ePIC detector is in progress. Collision parameters, synchrotron radiation, and beam gas backgrounds from both the electron and hadron beams have been studied, but there are continued efforts to ensure that all detectors are included using proper energy thresholds and digitization schemes. Raw data volumes could be hundreds of PBps, with reduction to hundreds of Gbps possible after various forms of filtering are applied. [Section 6.9]

4.3 Scientific Workflow

- The JLab experimental workflow is designed to accumulate events based on an average rate and size. Production data account for about 60 to 70% of the wall-clock time for an experimental run. The total data volume can be estimated by multiplying the total number of events in the dataset by the average event size. Once these values are established, the compute and storage requirements for an experiment can be estimated, since they scale linearly with events in the data set. [Section 6.1]
- The Rucio project at JLab has begun to re-imagine the data pipeline from the experimental halls, creating a single source of truth (i.e., all data are edited in only one location) for policy-based file movement of raw data around the campus. This will incorporate the use of OSG resources that can be leveraged both locally and external to the facility to deliver on computational tasks. [Section 6.1]
- The global analysis effort for LQCD involves collaboration between experimentalists and theorists at approximately 10 universities and research laboratories across the United States.

The computational requirements for these analyses are comparatively lower than those for the LQCD projects. This results in data movement to and from JLab that can total several PB per year, but in smaller TB-sized data sets. [Section 6.2]

- The workflow for JLab LQCD computing involves around 30 allocated projects, with a collective allocation of 2 PB of disk storage and 15 PB of tape storage. These projects typically generate data in the range of hundreds of terabytes during their yearly allocation period, with a portion of these data being transferred off-site. [Section 6.2]
- BNL NP experiments follow certain shared workflows during their operations. Primarily, these workflows revolve around two types of data: experimental data, which come directly from the detectors, and MC data, which represent simulations of both the detector's operations and the specific physics processes under study. [Section 6.3]
- Experiment and MC data workflows at RHIC demand substantial computational resources, with the former taking more than 75% of the total compute resources and the latter taking the rest. The MC workflow is generally executed across multiple sites. This includes not only those sites directly affiliated with the experiment but also independent ones such as those connected to the OSG and potentially even the DOE Leadership class HPC facilities (such as NERSC used by the STAR collaboration). The primary reason for this distribution is the computational intensity of MC workflows, which demand significant processing power but use relatively minimal data. [Section 6.3]
- RHIC experiment workflows are more data-centric, necessitating immense data handling and processing capabilities. Due to this, they are primarily executed at extensive high-throughput batch farms located within computing facilities explicitly dedicated to the experiment. [Section 6.3]
- The primary dataset consists of the raw data files written by the sPHENIX DAQ system. A data acquisition "run" consists of about one hour of data taking. Each run dataset consists of dozens of files, each written in parallel by the DAQ process that reads a given detector component. A full run with all components produced file sets of 52 files. This number is expected to grow to more than 60 to increase the available bandwidth for some components. The output files automatically roll over when reaching an adjustable size limit, typically 20 GB, resulting in several hundred files being written during a run. [Section 6.4]
- The sPHENIX raw data files are processed into summary files, traditionally still called "DSTs" (data summary tapes). The DSTs represent a state of the data reconstruction after a number of time-consuming steps have been performed. DSTs are expected to be about 50% of the size of the original raw data, conservatively accounting for the information omitted from the DSTs for them to support all the envisioned analyses. It is possible to create further generations of output, called micro-DSTs (or nano-, or pico-...), by applying a number of filters to the DST data to further reduce the size. Past experience indicates that those files will have a size on the order of 10% of the original raw data. [Section 6.4]
- FRIB categorizes data processing into three levels in relation to experiment operations: [Section 6.5]
  - "Online" includes the DAQ systems used for recording experiment data to permanent (disk) storage. Varying degrees of software-based event building, data reconstruction and filtering are used during online data taking. Online processing is required for recording data.
  - "Nearline" indicates processing required during experiment runtime that is not directly in the data recording path. For example, event analysis may be required to verify detector output and data quality and to inform operational decisions.

  - "Offline" is processing not directly tied to experiment operations. This includes MC and other simulations, data reduction, and analysis.

- GRETA's capabilities will likely represent the most significant performance challenge to the 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. [Section 6.6]

- Generally, analysis of GRETA data is carried out by experimental teams at their home institutions. Analysis and data interpretation is a time-consuming process (many months) but not a very computationally intensive process (can be done on local computing resources). The nature of this analysis is very much experiment dependent. [Section 6.6]

- GRETA performs a real-time analysis of the digitized waveform data. This analysis 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. [Section 6.6]

- Two primary workflows are associated with each GRETA experiment. The first is the real-time signal processing workflow that occurs internal to the GRETA instrument. This is common to all experiments. The second workflow is a data analysis step carried out by the experimenter and the experimenter's group. [Section 6.6]

  - The workflow for real-time signal processing consists of 120 User Datagram Protocol (UDP) data streams with an aggregate (maximum) rate of 32 Gbps. Signal processing is carried out on a GPU-cluster co-located with the experiment. All pipeline components in the GRETA pipeline and control plane are container based. This simplifies deployment and allows for the complex orchestration needed for future IRI implementations.

  - The data analysis workflow involves clustering and ordering interaction points into likely gamma-ray tracks and rejecting partial energy deposition events. This analysis step is not considered computationally (CPU or network) intensive and requires off-the-shelf computing resources.

- The key feature for data management and access on the ALICE Grid is the distribution of the data onto the grid at the data creation. The data are then subsequently accessed only from local site storage . 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 an about 200 PB distributed file system that is primarily used as disk storage on the local cluster. [Section 6.7]

- During the past year, ALICE jobs have read over 2.3 XB and written over 400 PB of data from/to the local storage, averaging about 13 GB/s and 80 GB/s for write and read traffic respectively averaged over the entire grid. Over the same period, conversely, ALICE jobs have read just 30 PB and written 3.4 PB of data over the WAN, averaging just 1 GB/s of aggregate bandwidth. [Section 6.7]

- The wide-area data distribution mode for the ALICE-USA sites is (1) receive a fraction of ALICE AO2D data files produced at T0/T1 sites in Europe (and Korea), (2) receive MC simulation files produced at T1/T2 sites, and (3) send copies of MC simulation files and analysis-reduced data produced locally to other sites, including between the US sites. [Section 6.7]

- All CMS-HI data are transferred to Fermilab for archival storage, primarily limited by available tape bandwidth. The Mini-AOD storage tier is additionally transmitted to Vanderbilt for access by physicists. Some users then generate even smaller n-tuples from this Mini-AOD , which are primarily transmitted to MIT for interactive analysis. [Section 6.8]

- The annual data volumes for CMS-HI are approximately five times what they were in 2018 (e.g., 40 PB/yr at the T0 and T1, and 3 to 10 PB/yr at ACCRE). An additional increase by a factor of two in HL-LHC is expected (tentatively scheduled for 2029). [Section 6.8]
- CMS (and US CMS) support a few different ways to access data. Users will typically begin an analysis workflow using CMS Remote Analysis Builder (CRAB), where they provide an executable and a desired input dataset. CRAB then will split the dataset into job-sized subdivisions and run a batch job for each piece. Generally, users use CRAB for data reduction and skimming by processing the Mini-AOD as the input and producing an analysis-specific n-tuple. CMS also supports direct batch submission either to the universities themselves or to the global CMS batch queue (which spans all CMS sites). In the final stages with smaller data needs, users will use interactive command-line tools and (increasingly) Jupyter Notebook. [Section 6.8]
- CMS supports two main data access interfaces: POSIX mounts for local file access (e.g., if a job at Vanderbilt needs to access data stored at Vanderbilt), or remote XRootD access via the AAA data federation, which federates CMS' global XRootD access points into a single access point. [Section 6.8]
- BNL and JLab have formed the EIC Computing and Software Joint Institute to serve as a single point of contact and organizational entity for support of the ePIC Collaboration and other software and computing needs for the EIC. Within this structure, theoretical calculations and accelerating modeling will be conducted, also within the current estimates established by the individual BNL and JLab facility use cases. [Section 6.9]
- EIC will follow a generalized approach to the process of science that is rooted in current NP-based workflows [Section 6.9]:
  - DAQ system affiliated with the ePIC detector.
  - Streaming readout to produce raw datasets.
  - Immediate computation to support calibration, alignment.
  - Long-term storage and sharing.
  - Reconstruction and reprocessing.
  - Simulation and possible integration with digital twins.
  - Local and distributed analysis.
- The EIC's workflow and resource utilization are still being modeled. Reconstructing the simulated data within the same workflow is preferable, as doing so will avoid a storage-consuming output stage after the simulation. [Section 6.9]

4.4 Computational and Storage Requirements

- JLab compute and storage requirements for the experimental program are estimated by the groups performing experiments in the halls with the assistance of a per-hall computing coordinator. JLab has planned for a roughly two-fold increase in data volumes over the next three years. The first increase in computing requirements will begin in 2024 with the Super BigBite (SBS) /GEp-V program, and the next in 2025 when a significant upgrade to the GlueX detector is completed. Beyond five years, the next increase in requirements will be affiliated with Solenoidal Large Intensity Device (SoLID). [Section 6.1]

- Each of the JLab halls currently has local storage close to the detector that can hold at least 48 hours of data, and is also connected to the data center with redundant 100 Gbps links. Data flow from the four experimental halls is expected to peak at 24–32 Gbps for the most demanding combination of experiments over the next five to seven years. [Section 6.1]
- JLab's data archive consists of an IBM tape library with a current capacity of ~200 PB and 13,000 tapes. [Section 6.1]
- At any time, several experiments are operating at different points in their simulation, calibration, reduction, and analysis phases at JLab. This overlap and interplay of usage patterns leads to a base level of compute load with frequent, bursty times of peak demand. It is not cost effective for the laboratory to provision for peak workloads, nor is it reasonable for wait times to vary too widely. To balance these concerns, off-site compute allocations offered by other institutions are part of the data processing strategy that smooths peak usage. [Section 6.1]
- Prior to the upgrade of the JLab accelerator from 6 to 12 GeV, event and data rates from experiments were an order of magnitude lower than they are anticipated to be over the next five years. This has facilitated a move from predominant use of local computing, to a mode where an increasing percentage of the work will be performed off-site. [Section 6.1]
- To support OSG operations, JLab has expanded its OSG infrastructure to eliminate single points of failure. There is an ongoing project to create job submission and data-transfer services per collaboration that can be adjusted to make efficient use of OSG resources. [Section 6.1]
- Lattice QCD calculations rely on a formulation of the theory in Euclidean space, and as such, are not real-time calculations. MC sampling of the theory provides snapshots of configurations of quarks and gluons. These snapshots, typically generated on LCFs, are used in subsequent calculations of observables that are carried out in ensemble calculations. From these observables, theoretical calculations guide, and confront, experiment measurements. [Section 6.2]
- LQCD projects at JLab exhibit variability based on their allocations. Some may heavily rely on JLab's computing resources; these flagship projects initiate their datasets on the LCFs with initial datasets typically amounting to around 300 TB spread across approximately 10,000 files. [Section 6.2]
- With the advent of new computing facilities such as Perlmutter at NERSC and Frontier at OLCF, the volume of data generated by the LQCD community at LCFs is increasing by a factor of 5 to 10. Existing workflows will remain in use, leading to a corresponding increase in the amount of data that need to be transferred back to JLab. [Section 6.2]
- The Aurora system at ALCF will likely mark a significant milestone in data production. An estimated three times more data are expected to be generated on Exascale systems compared with the current generation of systems (Summit, Perlmutter). For LQCD projects in 2025, this could result in approximately 10 to 15 PB/year of data generated on LCFs and transferred back to the lab, with most of the data being utilized locally. [Section 6.2]
- For sPHENIX and STAR, compute resources are a limiting factor. The ability to tap nonlocal resources like OSG and other unaffiliated resources can provide additional capacity. The ability to utilize these resources, particularly for experiment workflows, will be limited by the ability to transfer data between the host data center and the remote resources. [Section 6.3]
- The RHIC aims to provide 28 "cryo weeks" per year (operation of the superconducting magnets). The plan is to have a three-year running period (2023 to 2025) with the following

breakdown for the three years of sPHENIX operation: [Section 6.4]

  - Run 1 (2023): 70 PB (expected), 11.5 PB (actual).
  - Run 2 (2024): 50 PB (expected), 78 PB (potential).
  - Run 3 (2025): 180 PB.

- sPHENIX collected only 11.5 PB of data instead of the expected 70 PB in FY 2023 due to reduced accelerator availability. [Section 6.4]
- sPHENIX datasets are kept disk-resident at the SDCC. Therefore, the vast majority of dataset processing will take place at the SDCC itself. It is not economical to transfer the DSTs that make up a dataset off-site for remote processing because the CPU usage/disk space ratio is not large enough. [Section 6.4]
- FRIB experiments use a mix of compute systems as required for nearline and online data processing. A few dozen "SPDAQ" systems are deployed as required for interfacing to detectors and electronics. These are the start of the online FRIBDaq pipelines. [Section 6.5]
  - FRIB nearline processing occurs on either individually allocated compute nodes or fixed batch processing allocations on the offline Slurm cluster.
  - FRIB offline processing is performed on Linux compute clusters using the Slurm batch system.
- Currently, FRIB operates several types of storage systems for research support. [Section 6.5]
  - A NetApp storage system provides reliable Enterprise-class storage. Snapshots, off-site replication, and tape backups are maintained for data security.
  - Higher capacity research storage currently uses either Linux/ZFS or Ceph File System (CephFS) on commodity hardware. Approximately 2 PB of storage is spread across three Linux/ZFS servers. These have 2x10GE network links. These comprise the "offline" 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. The system is connected at 2x10 GE.
  - To support experiments requiring higher disk IO (>100 MB/s continuous) and to provide increased capacity, a Ceph storage cluster with CephFS is deployed. The raw capacity (before data redundancy) of the cluster is 2 PB. The cluster nodes use dual 25 GE networking. The CephFS storage has been used to back Globus DTN transfers across ESnet.
- GRETA data transfer volumes can be between 50 GB and 100 TB, and performed on an ad-hoc basis when the detector is operating. [Section 6.6]
- GRETA can write 1 GB/s to its local disk cache, although typical rates are expected to be less than this. Data set sizes are highly dependent on the physics case being studied. They depend on the triggered gamma-ray rate, the auxiliary detectors employed, and the beam time allocated. An experiment's aggregate data size is expected to range from a minimum of 50 GB to a maximum 100 TB. Individual file sizes should be $<$ 2 TB. [Section 6.6]
- 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). [Section 6.7]

- 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. [Section 6.7]
- For CMS-HI, the primary limitation on storage is the tape bandwidth available to store and recall large multi-PB datasets. While the capacity of tape cartridges greatly increases when new generations of technology are released, the bandwidth per tape has not kept pace. If trends continue, in 2030 it could take nearly a day of continuous access to read/write a tape from beginning to end. This presents significant issues operationally. Writing to and retrieving from custodial tape is by far the most difficult issue for data movement CMS-HI will face in the coming years. [Section 6.8]
- One limiting factor for data volumes in CMS-HI Run3 is the cost of tape archival and disk storage. Additionally, the rate at which tape archival can occur is limited by the number of available tape drives at Fermilab. CMS currently records at the maximum rate possible from the detector. When the detector is upgraded in 2029, the increase in detector channels will increase the maximum rate possible, which exacerbates these resource constraints. [Section 6.8]
- EIC networking and computation decisions are not imminent, and will consider lessons learned from LHC computing and analysis. It is expected that major collaboration sites will contribute resources, and leverage ESnet connectivity as a critical networking link to manage traffic flows. The ability to handle PB volumes of data over time, and bursts of hundreds of Gbps, is expected. [Section 6.9]
- Assessment of the computation and storage requirements for the EIC's ePIC detector will consider data collection, calibration, alignment, reconstruction, and analysis, with MC-simulated events used for studies and for analysis. The ePIC data acquisition is planned to be implemented as a flexible, scalable, and efficient streaming DAQ system, as outlined by the EIC yellow report. [Section 6.9]
- MC simulation in ePIC will encompass physics simulation (event and background modeling) and (with physics simulation as input) detector simulation, both fully detailed (Geant4) and fast (parameterized, ML based). [Section 6.9]

4.5 Remote Collaboration and Operational Requirements

- A majority of JLab users participate remotely via login to computing resources and remote conferencing facilities or by transferring data to or from the facility. Remote science activities routinely leverage computational resources provided at partner sites (e.g., NERSC), and by using the OSG. [Section 6.1]
- There has been an increase in the use of networking resources to enable remote users to participate in experimental runs at FRIB. Providing a capability for remote users to observe the products of ongoing data analysis would be beneficial to increase engagement with the user community. [Section 6.5]
- Given GRETA is an instrument that is movable, its network address space is necessarily abstracted from that of the host laboratory. External network-facing components include a bastion host for remote logins, an internet service host that abstracts standard services for the instrument, a data-transfer node for moving processed data to experimenter home facilities, and the forward buffers to admit the possibility to send full waveform data to remote computing facilities. [Section 6.6]

4.6 Multifacility Computational Workflows

- JLab simulation workflows are capable of running off-site, since the input data required is small versus that of an analysis workflow. In some cases, an experiment may require that the output of simulation runs be returned to JLab for storage or future analysis, and this will contribute to incoming WAN traffic. This may result in traffic demands that are similar in scale to reconstruction's contribution to the outgoing WAN traffic. [Section 6.1]
- In addition to the use of OSG for simulation at JLab, GlueX has performed the processing of a raw data subset off site, most notably at NERSC. This is a more data-intensive workflow, with large input and output. JLab's Scientific Workflow Indefatigable Factotum (SWIF) workflow manager is used to orchestrate the movement of the raw data from tape, over the network to NERSC, and the results back again. The total raw data processed at NERSC in 2022 was 1.5 PB. The processed data output, roughly 0.5 PB, is then transferred back to JLab for storage. [Section 6.1]
- The computing requirements for CLAS12, GlueX, and Hall A experiments at JLab exceed the local computing resources. The current plan is to do as much locally as possible, but to make extensive use of remote resources, OSG, NERSC, collaborators, and in some cases commercial cloud. The future 9 PB/yr GlueX raw dataset plus the 8 PB/yr from CLAS12 and 5 PB/year from hall A will rely on current and future ESnet upgrades. [Section 6.1]
- Members of the LQCD project teams seek allocations of computing time at numerous HPC facilities. While certain data may be retained for an extended period on leadership systems like NERSC, the primary responsibility for long-term data storage lies with the member laboratories. In the case of LQCD projects related to the JLab science program, JLab will serve as the host for the extended data storage. [Section 6.2]
- Researchers at FRIB are increasingly interested in using off-site HPC and data infrastructure to accomplish specific goals during the execution of an experiment. One experiment group has already employed local MSU HPCC resources to expediently analyze incoming data in near real-time to direct decisions during an experiment. Another group is exploring the use of NERSC for data analysis during ongoing experiments. A demonstration using an existing data set occurred in fall 2023 and a planned production test with FRIB will occur in spring 2024. [Section 6.5]
- 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 real time processing itself. For example, a coupled signal decomposition and tracking algorithm would require such an infrastructure. [Section 6.6]
- In the late two- to-five-year timeframe, 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. [Section 6.6]
- The ALICE-USA Computing project officially launched in 2010 and underwent changes to participants and computing architecture in 2015 and 2018. ALICE Grid sites today are based at the ORNL CADES facility and the LBNL / High-Performance Computing Services (HPCS)

facility. In addition to the core facilities, the ALICE-USA project has a history of working to integrate HPC systems into the ALICE Grid facility. Integrations at NERSC started with Cori in 2020, and have expanded into Perlmutter. [Section 6.7]

- ALICE does not operate a Grid-enabled storage element inside NERSC, which poses an issue running analysis jobs that are I/O heavy. With the switch to Perlmutter, the initial estimations show that this bottleneck is less constraining, allowing ALICE to leverage the proximity of the LBNL T2 SE as an input source to jobs that require data. This does increase network traffic between LBNL and NERSC, which has a small latency, but is still limited. Understanding the network limits now (zero to two years) and future (two- to five-year) capacities may help guide the development of ALICE use of NERSC HPC in the era beyond Perlmutter. [Section 6.7]
- The largest sources/sinks of CMS-HI data are the XRootD endpoints at CERN, Fermilab, Vanderbilt, and MIT. After their initial production, very little data movement happens on the large centrally produced datasets. There is approximately TB-scale data movement of user-produced datasets to other facilities. This trend is expected to be stable until 2029 when the HL-LHC program begins. With the increased detector granularity and data acquisition rate, these numbers can scale between 5 and 10 times today's numbers. [Section 6.8]
- CERN has, and is expected to continue to have, sufficient resources to produce smaller derived datasets for CMS-HI, and no need for DOE SC user facilities are foreseen (this is a notable difference between CMS-HI and CMS-PP). [Section 6.8]
- For the ePIC detector at the EIC, the ability to process data remotely is an integral part of the proposed computing model. Access to storage resources and sufficient network bandwidth to move or access data to and from remote sites is a prerequisite. [Section 6.9]
- For many NP experiments, utilization of unaffiliated computing resources not at the host data centers is a given. OSG and the supercomputers at DOE facilities (mainly NERSC for the STAR experiment) are commonly used for simulations and are two examples of remote, unaffiliated computing resources. [Section 7]
- NP facilities, experiments, and researchers have noted that integrating with other labs and data facilities presents ongoing challenges in authentication and federated identity. While frameworks like x.509, OAuth, and tokens have broad consensus, issues arise from nonstandard or incomplete implementations. Federated interactive logins that leverage approaches such as SAML, OpenAuth2, and Shibboleth are standardized and interoperable, but “automated” authentication for tasks like remote job control and large file transfers is fragmented and challenging to debug. [Section 7]
- NP facilities, experiments, and researchers have noted that facility-to-facility trust implementations are also difficult from both policy and technical perspectives. Cross-facility systems often rely on long-lived “robot” tokens or certificates that are an extension of a user's identity. These are not always a good fit for long-term campaigns. It is often the case that actions need to be taken on behalf of or as a proxy for users, which can stretch the abilities of federated identities. [Section 7]

4.7 Domestic Networking for Local and Wide-Area Data Mobility

- JLab leverages several data mobility tools for sharing experimental data. CVMFS, XRootD, FTS, Rucio, and Globus are all used when exchanging data with collaboration sites (e.g., OSG, NERSC) and with end users. [Section 6.1]

- GlueX has pioneered the use of NERSC to process JLab NP data. NERSC had a preference to process raw data rather than simulate, so GlueX reconstruction was executed on Perlmutter. In 2022 GlueX used 92K node hours at NERSC, and in 2023, GlueX was awarded 85K node hours. This data-intensive workflow can saturate one of the lab's two 10 Gbps circuits during periods of file transfer. This bottleneck is being addressed by the 100 Gbps WAN upgrade currently in progress. [Section 6.1]
- Several JLab software infrastructure projects for data movement are under development and will expand; the upgrade to 100 Gbps WAN connectivity will incentivize this work. Software projects under development include XRootD-based DTNs, Rucio for policy-driven data movement to remote sites, and Globus for file movement to leadership compute facilities and users home institutions. [Section 6.1]
- JLab's Science DMZ is leveraged to support data-intensive workflows to collaboration sites. Identified flow patterns have provisions to bypass the firewalls using policy-based routing (PBR) for well-known source/destination pairs; this adheres to two primary use cases: [Section 6.1]
  - Movement of data from the experimental halls to the tape library for storage.
  - Movement of data from the tape library and Lustre disk pool for off-site processing.
- Due to the complexity of the network path from JLab to both the ESnet Atlanta and Washington point-of-presence (PoP) locations, circuit outages had been relatively common. Significant effort has been put into automatic failover at every layer, to avoid making these outages impact users. As a result of this design, 100% uptime has been achieved over the past year despite a host of fiber cuts and equipment problems off-site between JLab and ESnet. [Section 6.1]
- In the ESnet5 era, JLab was connected over a complex Layer 2 fabric that spanned the E-LITE and MARIA, with the first upstream ESnet-managed routers at Ashburn and Atlanta. This left a blind spot in traffic monitoring between JLab and the ESnet cross-connect points that made debugging difficult at times. With ESnet6 and the installation of routers at JLab, additional insight into network operations was gained. [Section 6.1]
- The LCFs do not provide long-term storage for LQCD projects. Data is transferred to the USQCD computing facilities (JLab, Fermilab, and Brookhaven), which assume ownership. Data transfers to JLab LQCD will increase and scale with the size of new LCF systems. JLab will continue to serve as the repository for long-term storage. A portion of the analysis work will be conducted at the LCFs. However, the final stage of the analysis workflow is ideally suited for execution on JLab's local systems, effectively mitigating the disparity in LCF to local computing capability. [Section 6.2]
- Over the next five years, computing allocations of a few million node-hours per year for projects will be available for LQCD projects on Frontier at OLCF. It is anticipated that similar amounts will be available in the future for Aurora at ALCF. With these computing allocations, a few PB of data will be generated for each project. These data will be transferred to JLab, BNL, and Fermilab for further analysis. In total, about 10 PB is anticipated for transfer over that period. [Section 6.2]
- RHIC data are transferred through the BNL High Throughput Science Network (HTSN) from the experiments by fiber optic cable to the SDCC for storage and further processing. The HTSN is the network fabric that ties the components of the experiments together. It provides direct, high-throughput connectivity between the RHIC experimental halls and the compute and storage resources at the SDCC. It also enables high-bandwidth transfers between the SDCC and sites on the WAN. Finally, it couples the compute and storage resources within the SDCC data center together at high bandwidth. [Section 6.3]

- At present, 52 DTNs are in operation at the BNL SDCC, with the majority being utilized by programs outside of NP. [Section 6.3]
- sPHENIX data processing is to a large extent centered around the BNL SDCC. The primary required ESnet services are therefore interactive logins and small-scale data transfers. The amount of data transferred out of the SDCC is not expected to exceed 100 TB/month. [Section 6.4]
- The FRIB network is evolving with the introduction of ESnet connection in the last year. FRIB operates several internal networks. A WAN connection is provided by MSU Information Technology (MSU IT) with a 2 x 10 Gbps connection between FRIB and MSU (some links in this path have been upgraded to 100 Gbps). The WAN connection is subject to MSU IT firewall restrictions. [Section 6.5]
- The FRIB ESnet connection consists of ESnet routers, with two 100 Gbps WAN links to the ESnet network. This connection currently supports the Science DMZ, including Globus DTN. [Section 6.5]
- At the conclusion of an FRIB experiment, experimental account access is disabled. Off-line analysis is typically performed at the spokesperson's home institution. Spokespersons who have accumulated data sets of more than several TB have made use of the FRIB Globus endpoint to ship data to remote storage for off-line data analysis. [Section 6.5]
- 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 FRIB 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 the following: [Section 6.5]
  - 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.
- Current infrastructure work at FRIB includes adding redundancy to the Science DMZ network and standardizing support of user usage of the new DTN (provisioning new users and storage access is currently a manual process). Work in planning includes further separation (both physical and logical) of business/office and research/science networks. The addition of additional network security (stateful firewall, etc.) will allow additional services to utilize the ESnet network (in addition to Science DMZ-based services). [Section 6.5]
- GRETA will interact with the WANs in three primary ways: [Section 6.6]
  - System-level access by staff with appropriate access permission for system maintenance.
  - Download of experiment data sets from the GRETA DTN to remote analysis resources.
  - Possible future signal processing modalities which require resources beyond the capabilities of the production GRETA signal decomposition cluster.
- 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. Globus will serve as our data-transfer tool. [Section 6.6]
- ALICE has yearly episodes that require more significant WAN capacities. These occur when storage is added and/or decommissioned, or when data must otherwise be redistributed

between different sites. 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. [Section 6.7]

- The ALICE T2 site at ORNL currently 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. Plans are underway this year to transition the ALICE environment to connect directly to the ORNL border routers at a peer level with the Science DMZ. The ALICE network environment is in the process of being upgraded to switches connected at 40 G. The upgraded ORNL ALICE environment is expected to be completed early in 2024. Beyond 2024, ORNL connections to ESnet are expected to migrate proportionate with ESnet backbone speeds. [Section 6.7]
- The current network topology of the ALICE T2 facility at the LBNL HPCS center features an internal connection between the worker nodes and the storage using 56 Gb InfiniBand (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 LBLnet connection to ESnet. The SEs were recently added to the LHCONE via a 10 Gb firewall. An item for the longer (two- to five-year) term is to understand and if necessary, optimize the network connectivity between the storage at the HPCS facility and Perlmutter HPC at NERSC. [Section 6.7]
- An identified issue with the CMS-HI data reduction pattern is that some popular portions of datasets are reproduced many times. For example, a user might directly copy the Mini-AOD dataset's electron collections to their private n-tuples. At the predicted scales of HL-LHC (predicted to begin in 2029), this data duplication becomes financially burdensome to support. By storing data in object stores, these common slices of datasets could be stored and referenced by multiple end-user datasets, providing a better space efficiency for analysis. Of course, these benefits become more pronounced if there is a single copy globally of the relevant objects, so this implies some level of additional WAN traffic needed to satisfy these workflows. These techniques are the target of active R&D and initial results are expected in the 2025 to 2026 timescale. [Section 6.8]

4.8 Emerging Needs

- JLab relies on cloud services for business services (e.g., email, calendar service, cloud storage). Although bandwidth to and from these services is much lower than that required for scientific data, the reliance of the laboratory on cloud services for day-to-day business has made the resilience of Internet connectivity critical. During periods of peak scientific data movement, cloud data for day-to-day work has been affected, and QoS and other traffic engineering controls were put in place to protect the business traffic from being affected by the scientific data flows. [Section 6.1]
- 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 OSG and NERSC. Experimental workflows require expedient access to computation on occasion to perform a recalibration of experimental data or reconstruction of an analysis. These use cases are portable enough to run in a cloud environment, and can be run faster than waiting in local computational queues for service. Due to the current cost of cloud computing, use cases are expected to be rare, and when used it will replace one or more of the other resources so the WAN requirements will not significantly increase. [Section 6.1]

- On occasion, HPC systems in Europe are heavily used in collaborations with LQCD European researchers; these initial datasets are then used for secondary calculations also carried out at DOE HPC facilities, resulting in datasets of about 1 PB distributed over about 1,000,000 files. [Section 6.2]
- At the present time, the use of cloud resources at BNL to meet the computational needs of RHIC experiments is limited to specialized, low-intensity (minimal compute and storage requirements) applications. Product evaluation and services in the cloud (repository, collaborative tools, and groupware applications) are use cases that utilize cloud resources, and this trend will likely increase over time. While cloud computing offers certain advantages, the bulk of computing and storage resources will continue to be located on-premises at BNL. This approach presents significant cost benefits, especially when considering egress issues associated with cloud solutions. [Section 6.3]
- The Superconducting Analyzer for Multi-particles from Radioisotope beams (SAMURAI) Pion-Reconstruction and Ion-Tracker (SπRIT) is a time-projection chamber (TPC) constructed at MSU as part of an international effort to constrain the symmetry-energy term in the nuclear equation of state (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. [Section 6.5]
  - 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 spring 2024, and an improved approach to the "process of science" for this remote resource is needed. A discussion at FRIB has started involving several departments for support.
- In the next five-year timeframe, when GRETA begins collecting data initially at FRIB and potentially subsequently at ATLAS/Argonne, real-time component of signal processing (workflow 1) will be carried out using GRETA's local computing cluster or potentially at HPC facilities. 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, the demand for use of cloud services in the final analysis should be limited. [Section 6.6]
- ALICE can benefit from additional supercomputing resources especially with GPU capabilities. [Section 6.7]
- The current and developing ALICE computing models do not have any specific plans for use of cloud resources. However, user analysis within commercial clouds is a possibility if it is cost effective. ALICE research teams in Europe found that commercial clouds were fully functional and efficient for running MC simulations, from which the produced simulated data were distributed 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. [Section 6.7]
- Systematic, automated WAN monitoring is very important for efficient use of ALICE-USA resources. This goes beyond any one specific site-to-site monitor, as it should seamlessly include

all paths between any two ALICE Grid sites. ALICE has a simple, yet effective, monitor using traceroute between every VOBox that can be used in near-real time to adjust automated data placement, but which does not provide any real diagnostic capabilities. The plan is to leverage the OSG network group's perfSONAR dashboards. That plan has not been implemented, partially due to the reduced active collaboration between ALICE Grid sites during COVID, but remains our current best model for tackling this need. The project, however, is interested in any feedback regarding future monitoring capabilities from ESnet or other service providers. [Section 6.7]

- CMS-HI expects an increase in remote science usage, namely accessing GPU resources, via the SONIC inference-as-a-service activity. This R&D is still in its infancy, and there is some uncertainty on how the service would be deployed (if SONIC is deployed). This research is ongoing and a decision on the deployment would be made in the 2026 to 2027 time frame. [Section 6.8]
- CMS is also participating in SENSE-Rucio R&D, which will use ESnet's ability to provide guaranteed point-to-point bandwidth to more effectively schedule Rucio data transfers. With SENSE-Rucio, the Rucio data movement tool can decide to signal to the SENSE dataplane that it would like guaranteed bandwidth between two sites, and if the request is accepted, configure transfers for specific dataset(s) to transit exclusively over that guaranteed bandwidth. There is a similar effort in the NOTED project, using a technology called "packet marking" to provide QoS guarantees. All these projects are under active development, and CMS will decide on the timescale of 2026 to2027 whether to move forward with these technologies at scale. [Section 6.8]

4.9 IRI Responsiveness

- JLab is implementing a model of policy-based data movement. There are two major efforts underway to achieve this goal: organize experimental metadata into a multi-experiment catalog and deploy Rucio as a data mobility platform for internal and external facility use. JLab anticipates that once uniform interfaces are available to create workflows that are portable between experimental facilities, HPC facilities, and network facilities, NP workflows can more easily use external resources and will be substantially strengthened. [Section 6.1]
- The RHIC experiments rely mostly on services located at BNL. As the SDCC is supporting users and collaborators from multiple organizations, it is increasingly deploying Federated Identity mechanisms for accessing web services and collaborative tools supported by the facility. The SDCC supports the three IRI common patterns: [Section 6.3]
  - Time-sensitive: Data produced by RHIC are reconstructed as soon as they are produced for rapid feedback on the quality of the data taking. The result may inform the machine for performance tuning and adjustments.
  - Data integration-intensive: Data from a given year are usually reprocessed and filtered several times as software and calibration evolve. Data from simulations may be embedded with real data to assess detector performance and signal reconstruction efficiencies.
  - Long-term campaign: Data from different years, taken under different conditions of the accelerator, may be produced or reproduced depending on the advancement of knowledge (calibrations, physics understanding, etc.).
- A number of limitations to addressing future needs exist at FRIB, many of which can be mitigated by the use of emerging approaches to multi-facility workflows and the IRI activity: [Section 6.5]

  - Not enough computing capability to address new instrumentation or on-site analysis.
  - Limited staff availability to adapt or convert HPC-workflows to operate within FRIB.
  - No staff expertise capable of leveraging capabilities at other DOE facilities (e.g., DOE HPC centers, ESnet capabilities).
- Two of the three IRI patterns will be predominant at FRIB: [Section 6.5]
  - Time sensitive: Rapid data transmission, analysis, and inspection are critical to enable data-informed decisions during experiment execution. The appropriate time scale is on the order of one hour.
  - Long-term campaign: Researchers need sustained access to computational resources over multiple years to refine an analysis of a particular experiment leading to publication.
- The GRETA data pipeline was designed with IRI workflows in mind, and this option is actively being developed. GRETA's forward buffers can send their data over WANs to remote HPC facilities where the main data processing tasks could be carried out on interactive timescales (time-sensitive pattern). These workflows are currently being evaluated using the ESnet testbed and OLCF/ORNL IRI testbed and are expected to be ready for production use in a two-year timeframe. [Section 6.6]
- NP facilities, experiments, and researchers have observed that integration with other labs and data facilities presents ongoing challenges in authentication and federated identity. While frameworks like x.509, OAuth, and tokens have broad consensus, issues arise from nonstandard or incomplete implementations. Federated interactive logins, that leverage approaches such as SAML, OpenAuth2, and Shibboleth, are standardized and interoperable, but "automated" authentication for tasks like remote job control and large file transfers is fragmented and challenging to debug. [Section 7]
- NP facilities, experiments, and researchers have observed that facility-to-facility trust implementations are difficult from both policy and technical perspectives. Cross-facility systems rely on long-lived "robot" tokens or certificates that are an extension of a user's identity. This approach is not a good fit for the IRI defined long-term campaign pattern. Actions often need to be taken on behalf of or as a proxy for users, which can stretch the abilities of federated identities. [Section 7]
- In light of the crucial role of authentication in the IRI vision, a central repository with reference designs and implementation examples for common workflows would offer significant value. [Section 7]

5 Review Actions

ESnet recorded a set of high-level actions from the NP user facilities and ESnet requirements review that extend ESnet's ongoing support of NP-funded collaborations. Based on the key findings, the review identified several actions for NP, ASCR, and ESnet to jointly pursue. These items are listed as guidance for future collaboration, and do not reflect formal project timelines. ESnet will review these with NP participations on a yearly basis, until the next requirements review process begins.

These are also organized by topic area for simplicity and follow common themes:

- Facility management and readiness.
- Scientific data management.
- Scientific workflow.
- Computational and storage requirements.
- Remote collaboration and operational requirements.
- Multifacility computational workflows.
- Domestic networking for local and wide-area data mobility.
- Emerging needs.
- IRI responsiveness.

5.1 Facility Management and Readiness

- BNL, JLab, and ESnet will coordinate on the needs of the EIC during the design, construction, and implementation process. Due to the potential design implications that will result in raw data rates of hundreds of Pbps, with reduction to hundreds of Gbps, networking between the sites to support computation and storage will be critical.
- ESnet, CMS-HI, and ALICE-USA will continue to coordinate on supporting operations during Run 3 (which began in July 2022), through the planning and execution of Run 4 (the HL-LHC era). This support will include coordination on data challenge (DC) activities, increasing capabilities through the development and deployment of R&D projects (SENSE, NOTED), and delivering measurement and monitoring frameworks (e.g., Stardust, perfSONAR).

5.2 Scientific Data Management

- ESnet will continue to work on the design and implementation of DTNs to support data mobility needs. JLab and ALICE-USA have requested help in this area.
- ESnet and LQCD research will discuss mechanisms to expose and transfer data from the LQCD effort at major DOE HPC facilities, as the data volumes of the data products increase to multiple PBs in the coming years. Deployment of high-performance portal software (e.g., based on the Globus Modern Research Data Portal) could simplify the delivery of data to collaborators.

5.3 Scientific Workflow

- ESnet will continue to consult with FRIB on workflow design, and data mobility needs. A number of limitations to addressing future needs exist at FRIB , many of which can be mitigated by the use of emerging approaches to multi-facility workflows and the IRI activity.

- The FRIB network is evolving with the introduction of ESnet connection in the last year. FRIB operates several internal networks. A WAN connection is provided by MSU Information Technology (MSU IT) with a 2 x 10 Gbps connection between FRIB and MSU (some links in this path have been upgraded to 100 Gbps). The WAN connection is subject to MSU IT firewall restrictions.
- FRIB staff are interested in consulting with HPC facilities on ways to increase access to computational capabilities. Doing so will allow FRIB to address new instrumentation at the facility and explore and new analysis needs.

5.4 Computational and Storage Requirements

- ESnet, ASCR, the newly awarded High Performance Data Facility (HPDF), and NP facilities will continue to ensure that network bandwidth between facilities remains high to support high-performance data transfer. One example of demand is data movement for LQCD projects. The LCFs do not provide long-term storage for the LQCD projects. Data are transferred to the USQCD computing facilities (JLab, Fermilab, and Brookhaven), which assume ownership.

5.5 Remote Collaboration and Operational Requirements

- ESnet will assist the SπRIT TPC collaboration to validate network connectivity between the MSU and RIKEN facilities. 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 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 spring 2024, and an improved approach to the "process of science" for this remote resource is needed. A discussion at FRIB has started involving several departments for support.

5.6 Multifacility Computational Workflows

- ESnet, LBL, NERSC, and ALICE should work to address the implications of adapting the ALICE workflow model to support streaming versus bulk data movement.
  - The current network topology of the ALICE T2 facility at the LBNL HPCS center features an internal connection between the worker nodes and the storage using 56 Gb 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 LBLnet connection to ESnet. The SEs were recently added

to the LHCONE via a 10 Gb firewall. An item for the longer (two- to five-year) term is to understand and if necessary, optimize the network connectivity between the storage at the HPCS facility and Perlmutter HPC at NERSC.

  - ALICE does not operate a Grid-enabled storage element inside NERSC, which poses an issue running analysis jobs that are I/O heavy. With the switch to Perlmutter, the initial estimations show that this bottleneck is less constraining, allowing ALICE to leverage the proximity of the LBNL T2 SE to use it as an input source to jobs that require data. This does increase network traffic between LBNL and NERSC, which has a small latency, but is still limited. Understanding the network limits now (zero to two years) and future (two- to five-year) capacities may help guide the development of ALICE use of NERSC HPC in the era beyond Perlmutter.

- ESnet should hold regular discussions with NP facilities to ensure that upgrade paths to deliver more services and bandwidth capacity remain available. As computing requirements for certain NP experiments are expected to exceed local resources in the future, a number of facilities are planning to make extensive use of remote resources.

5.7 Domestic Networking for Local and Wide-Area Data Mobility

- ESnet and ALICE-USA will work to leverage the OSG network group's perfSONAR dashboards. Implementation of this sensible design for systematic, automated WAN monitoring is very important for efficient use of ALICE-USA resources.
- ESnet will continue to support JLab resiliency, particularly when upgrades are required to support EIC and HPDF. The upgrades to JLab's network path to both the ESnet Atlanta and Washington PoP locations have resulted in 100% uptime over the past year despite a host of fiber cuts and equipment problems off-site.
- ESnet and JLab will continue to work to support the metro and wide-area networking requirements. In the ESnet5 era, JLab was connected over a complex Layer 2 fabric that spanned the E-LITE and MARIA, with the first upstream ESnet-managed routers at Ashburn and Atlanta. This left a blind spot in traffic monitoring between JLab and the ESnet cross-connect points that made debugging difficult at times. With ESnet6 and the installation of routers at JLab, additional insight into network operations was gained.
- ESnet will continue to monitor the network speed needs of NP facilities and schedule upgrades to capacity and services as required. The major NP facilities and experiments (located at BNL, JLab, FRIB, ORNL, LBNL, and Vanderbilt) are all connected to ESnet with a capacity of 100 Gbps and several will upgrade to multiple 100 Gbps or 400 Gbps in the coming years to support increases in data volumes.

5.8 Emerging Needs

- ESnet and Rucio developers will continue to coordinate on software package development and how it relates to R&D efforts such as SENSE and NOTED. Rucio is allowing NP facilities to re-imagine the data pipeline from the experiments, creating a single source of truth (i.e., all data are edited in only one location) for policy-based file movement of raw data around campus environments.
- ESnet and JLab should continue to ensure that their connectivity to cloud services for business services remains stable. Bandwidth to and from these services is much lower than that required for scientific data, but the reliance of the laboratory on cloud services for day-to-day business

has made the resilience of Internet connectivity critical. During periods of peak scientific data movement, cloud data for day-to-day work has been affected, and QoS and other traffic engineering controls were put in place to protect the business traffic from being affected by the scientific data flows.

- ESnet will continue to investigate the use of commercial cloud services for data processing. Due to the current cost of cloud computing, use cases are expected to be rare, and when used it will replace one or more of the other resources so the WAN requirements will not significantly increase. This work will benefit NP facilities and experiments, along with others in DOE SC.
- ESnet will continue to partner with CMS-HI in a number of R&D activities that may influence operations prior to the HL-LHC era. These include the following three:
  - The SONIC inference-as-a-service activity.
  - SENSE-Rucio R&D.
  - The NOTED project.

5.9 IRI Responsiveness

- The IRI effort, led by members across DOE SC programs, must address the availability of uniform interfaces at experimental facilities, HPC facilities, and network facilities. Doing so will strengthen facility and experimental workflows, as well as lead to an increase in external resource use. The NP community will be a beneficiary of this work.
- In light of the crucial role of authentication in the IRI vision, it is recommended that a central repository with reference designs and implementation examples for common workflows would offer significant value.
- The IRI effort, led by members across DOE SC programs, must address some of the ongoing challenges in authentication and federated identity. While frameworks like x.509, OAuth, and tokens have broad consensus, issues arise from nonstandard or incomplete implementations. Federated interactive logins that leverage approaches such as SAML, OpenAuth2, and Shibboleth are standardized and interoperable, but “automated” authentication for tasks like remote job control and large file transfers is fragmented and challenging to debug. Addressing, these challenges will benefit the NP community as well as other DOE SC programs.
- ESnet will continue to consult with FRIB on workflow design and data mobility needs. A number of limitations to addressing future needs exist at FRIB , many of which can be mitigated by the use of emerging approaches to multi-facility workflows and the IRI activity.
- The IRI effort, led by members across DOE SC programs, must address the long-standing policy issue surrounding facility-to-facility trust implementations. Cross-facility systems often rely on long-lived “robot” tokens or certificates that are an extension of a user’s identity. These are not always a good fit for long-term campaigns. It is often the case that actions need to be taken on behalf of or as a proxy for users, which can stretch the abilities of federated identities. Addressing, these challenges will benefit the NP community, as well as other DOE SC programs.

6 NP 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 principal investigators (PIs) funded by the Office of NP of the DOE SC. These case studies were discussed virtually between July 2023 and October 2023.

The case studies were presented, and are organized in this report, in a deliberate format to present an overview based on individual experiments, larger facilities, and in some cases the encompassing laboratory environments that provide critical resources for operation. The case studies profiled include:

- JLab Facilities and Experiments
- Theory Group and LQCD at JLab
- BNL: The SDCC
- sPHENIX at the RHIC
- FRIB
- GRETA
- ALICE Project and ALICE-USA Computing
- The CMS Heavy Ion Experimentation and ACCRE
- The EIC

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 will be 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 will provide a snapshot of the discussion and focus during the in-person review.

6.1 JLab Facilities and Experiments

JLab is a US DOE SC national laboratory. Scientists worldwide utilize the lab's unique particle accelerator, known as the CEBAF, to probe the most basic building blocks of matter - helping us to better understand these particles and the forces that bind them - and ultimately our world.

In addition, the lab capitalizes on its unique technologies and expertise to perform advanced computing and applied research with industry and university partners, and provides programs designed to help educate the next generation in science and technology.

6.1.1 Discussion 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 biggest experimental contributors to the JLab computation workload and WAN bandwidth requirements will be GlueX in hall-D and CLAS12 in hall-B. Egress (ingress) is estimated to peak at around 15 PB/year (3 PB/year) by 2025/26. While internal data rates from the other halls (hall-A in particular) will rise to similar scales, their analysis program is expected to have minimal impact on the WAN until SoLID is operational (~2030).

The JLab LQCD Theory group is planning to increase engagement with the Aurora system at Argonne (coming online in 2024) and OLCF Frontier within the next few years. This is expected to result in 10–15 PB/yr to be transferred from those sites back to JLab beginning in 2025. Please refer to the separate JLab LQCD/Theory case study for additional details.

Both GlueX and CLAS12 continue to make use of off-site computing resources. Historical constraints driven by a 2x10Gbit WAN link have been addressed through an upgrade to 2x100Gbit connections supported through ESnet.

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 QoS. This would also provide a reliable connection over which, potentially, experiments could be controlled remotely.

Research Projects related to IRI make use of ESnet and SC capabilities to examine facility-to-facility operations including data-intensive, long-term, and near-real-time workflows. In the future, the ability to light dark fiber at JLab may be a natural next step for dedicating bandwidth to these efforts.

The following bullets were extracted from the case study, and in-person discussion. Several appear in Section 1 and Section 3.

- JLab has an international user community of over 1800 active users (324 institutions in 39 countries.). One-third of all PhDs granted in nuclear physics in the US are based on JLab research. The primary scientific instrument is the CEBAF. CEBAF is a high intensity electron accelerator with unique capabilities to probe the nuclear structure of matter at the quark level.
- The raw data from JLab CEBAF experiments are shared with experimenters, and a backup is stored locally on tape. 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. Approximately 100 PB of data a year are produced through a mixture of raw scientific observation, analysis data, simulation, and backups of affiliated data sets.
- The JLab experimental workflow is designed to accumulate events based on an average rate and size. Production data account for about 60-70% of the wall-clock time for an experimental run. The total data volume can be estimated by multiplying the total number events in the dataset by the average event size. Once these values are established the compute and storage requirements for an experiment can be estimated since they scale linearly with events in the data set.
- The Rucio project at JLab has begun to re-imagine the data pipeline from the experimental halls, creating a single source of truth (i.e., all data are edited in only one location) for policy-based file movement of raw data around the campus. This will incorporate the use of OSG resources that can be leveraged both locally, and external to the facility to deliver on computational tasks.
- JLab compute and storage requirements for the experimental program are estimated by the groups performing experiments in the halls with the assistance of a per-hall computing coordinator. JLab has planned for a roughly two-fold increase in data volumes over the next three years. The first increase in computing requirements will begin in 2024 with the SBS/GEp-V program, and again in 2025 when a significant upgrade to the GlueX detector is completed. Beyond five years, the next increase in requirements will be affiliated with SoLID.
- Each of the JLab halls currently has local storage close to the detector that can hold at least 48 hours of data, and is also connected to the data center with redundant 100 Gbps links. Data flow

from the four experimental halls is expected to peak at 24–32 Gbps for the most demanding combination of experiments over the next 5–7 years.

- JLab's data archive consists of an IBM tape library with a current capacity of ~200 PB and 13,000 tapes.
- At any time, there are several experiments operating at different points in their simulation, calibration, reduction, and analysis phases at JLab. This overlap and interplay of usage patterns leads to a base level of compute load with frequent, bursty times of peak demand. It is not cost effective for the laboratory to provision for peak workloads, nor is it reasonable for wait times to vary too widely. To balance these concerns, off-site compute allocations offered by other institutions are part of the data processing strategy that smooths peak usage.
- Prior to the upgrade of the JLab accelerator from 6 to 12 GeV, event and data rates from experiments were an order of magnitude lower than they are anticipated to be over the next five years. This has facilitated a move from predominant use of local computing, to a mode where an increasing percentage of the work will be performed off-site.
- To support OSG operations, JLab has expanded its OSG infrastructure to eliminate single points of failure. There is an ongoing project to create job submission and data-transfer services per collaboration that can be adjusted to make efficient use of OSG resources.
- A majority of JLab users participate remotely via login to computing resources and remote conferencing facilities or by transferring data to or from the facility. Remote science activities routinely leverage computational resources provided at partner sites (e.g., NERSC), and through the use of the OSG.
- JLab simulation workflows are capable of running off-site, since the input data required are small versus that of an analysis workflow. In some cases, an experiment may require that the output of simulation runs be returned to JLab for storage or future analysis, and this will contribute to incoming WAN traffic. This may result in traffic demands that are similar in scale to reconstruction's contribution to the outgoing WAN traffic.
- In addition to the use of OSG for Simulation at JLab, GlueX has performed the processing of a raw data subset off site, most notably at NERSC. This is a more data-intensive workflow, with large input and output. JLab's SWIF workflow manager is used to orchestrate the movement of the raw data from tape, over the network to NERSC, and the results back again. The total raw data processed at NERSC in 2022 was 1.5 PB. The processed data output, roughly 0.5PB, is then transferred back to JLab for storage.
- The computing requirements for CLAS12, GlueX, and hall A experiments at JLab exceeds the local computing resources. The current plan is to do as much locally as possible, but to make extensive use of remote resources, OSG, NERSC, collaborators and in some cases commercial cloud. The future 9 PB/yr GlueX raw dataset plus the 8 PB/yr from CLAS12 and 5 PB/year from hall A will rely on current and future ESnet upgrades.
- JLab leverages several data mobility tools for sharing of experimental data. CVMFS, XRootD, FTS, Rucio, and Globus are all used when exchanging data with collaboration sites (e.g., OSG, NERSC) and with end users.
- GlueX has pioneered the use of NERSC to process JLab NP data. NERSC had a preference to process raw data rather than simulate so GlueX reconstruction was executed on Perlmutter. In 2022 GlueX used 92K node hours at NERSC, and in 2023, GlueX was awarded 85K node hours. This data-intensive workflow can saturate one of the lab's two 10 Gbps circuits during periods of file transfer. This bottleneck is being addressed by the 100 Gbps WAN upgrade currently in progress.

- Several JLab software infrastructure projects for data movement are under development and will expand; the upgrade to 100 Gbps WAN connectivity will incentivize this work. Software projects under development include XRootD based DTNs, Rucio for policy-driven data movement to remote sites, and Globus for file movement to leadership compute facilities and users home institutions.
- JLab's Science DMZ is leveraged to support data-intensive workflows to collaboration sites. Identified flow patterns have provisions to bypass the firewalls using PBR for well-known source/destination pairs; this adheres to two primary use cases:
  - movement of data from the experimental halls to the tape library for storage
  - movement data from the tape library and Lustre disk pool for off-site processing
- Due to the complexity of the network path from JLab to both the ESnet Atlanta and Washington PoP locations, circuit outages had been relatively common. Significant effort has been put into automatic failover at every layer, to avoid making these outages user impacting. As a result of this design, 100% uptime has been achieved over the past year despite a host of fiber cuts and equipment problems off-site between JLab and ESnet.
- In the ESnet5 era, JLab was connected over a complex Layer 2 fabric that spanned the E-LITE and MARIA, with the first upstream ESnet-managed routers at Ashburn and Atlanta. This left a blind spot in traffic monitoring between JLab and the ESnet cross-connect points that made debugging difficult at times. With ESnet6 and the installation of routers at JLab, additional insight into network operations was gained.
- JLab relies on cloud services for business services (e.g., email, calendar service, cloud storage). Although bandwidth to and from these services is much lower than that required for scientific data, the reliance of the laboratory on cloud services for day-to-day business has made the resilience of Internet connectivity critical. During periods of peak scientific data movement, cloud data for day-to-day work has been affected, and QoS and other traffic engineering controls were put in place to protect the business traffic from being affected by the scientific data flows.
- 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 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 expediently a computationally expensive task. Due to the current cost of cloud computing, use cases are expected to be rare and when used it will replace one or more of the other resources so the WAN requirements will not significantly increase.
- JLab is implementing a model of policy-based data movement. Two major efforts are underway to achieve this goal: organize experimental metadata into a multi-experiment catalog and deploy Rucio as a data mobility platform for internal and external facility use. JLab anticipates that once there are uniform interfaces available to create workflows that are portable between experimental facilities, HPC facilities, and network facilities, NP workflows can more easily use external resources and will be substantially strengthened.

6.1.2 JLab Facility Profile

JLab is funded by the SC for the US 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 user community of over 1800 active users. One-third of all PhDs granted in nuclear physics in the US are based on JLab research with 735 PhDs granted to-date and 181 in progress.

6.1.2.1 Science Background

Complementary to the Nuclear Physics 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. Additional details on those programs can be found in [Section 6.2].

The raw data from the 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.

6.1.2.2 Collaborators

The more than 1800 JLab users are from 324 institutions in 39 countries. Figure 6.1.1 shows the geographic distribution of collaborators. The number of users varies depending on the institution. At any time only a fraction of the user population is physically present at the JLab site. Most participate remotely via login to JLab computing resources, remote conferencing facilities or by transferring data to or from JLab.

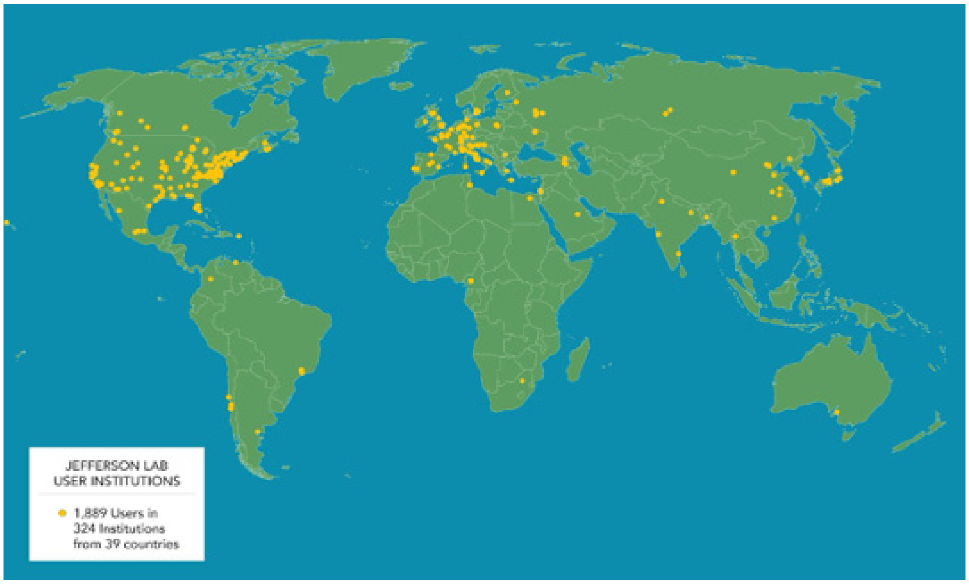


Figure 6.1.1: Geographic Distribution of JLab Collaborators

User/collaborator and location	Do they store a primary or secondary copy of the data?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of dataset? (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)?
OSG SITES - SIMULATION	Secondary	CVMFS, XRootD	MC Output 1 PB/year	daily	no	Active project to scale up XRootD storage for data to JLab
USQCD COLLABORATORS (FERMILAB, BNL, UNIVERSITIES)	Both	Globus	20-50 PB/year	ad hoc	no	no
NERSC - HALL PRODUCTION	Secondary	Globus	3 PB/year	monthly	yes, processed and returned with Globus	file pre-staging/ post job retrieval automation can have logistical challenges (credential lifetime, file lifetime)
INDIVIDUAL USER ANALYSIS/ UNIVERSITY OR LAB	Secondary	Globus	1 PB/year	ad hoc	no	instruction to users, expectations for performance
COLLABORATING INSTITUTION DATA EXPORT	Secondary	Rucio (FTS+XRootD)	10 PB/year	weekly	no	scaling this up now.
(EXASCALE LCFS (AURORA, OLCF TITAN) FOR LQCD DATA SETS (EXPECTED TO RAMP UP IN 2025)	Secondary	Lustre file systems and tape storage at JLab. Access via tape and Globus	10–15 PB/yr in 2025	monthly	no	no

Table 6.1.1: Collaborative Data Mobility

6.1.2.3 Instruments and Facilities

JLab's primary scientific instrument is the 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, halls A, B, C and D, and CEBAF is able to deliver beam to up to 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 the electron beam itself or a photon beam derived from the electron beam is directed onto a target. Targets can be solid, liquid or high-pressure gas. In certain materials 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. 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. 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. The 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.

The properties of the detectors in the four halls are summarized in the table below, and will be described in detail in the subsequent sections.

Hall D	Hall B	Hall C	Hall A
Polarized photons	electron luminosity 10^{35}	electron luminosity 10^{39}	
excellent hermeticity	good hermeticity	precision spectrometers	
photon energies of ~8.5-9 GeV		11 GeV beamline	
hoton luminosity 10^{8} photons/s		target flexibility	
good momentum/angle resolution		excellent momentum resolution	
high multiplicity reconstruction		energy reach	custom installations

Table 6.1.2: Scientific Properties of CEBAF Detectors

Hall-A

Hall-A is physically the largest of the four halls. Historically, it contained two precision high momentum spectrometers that are movable to various angles. They were used to study interactions of the electron beam with a target at the pivot point of the spectrometers Typical hall-A experiments were short lived, weeks or months on the floor and raw data rates were of order 1–10 MB/sec, or a fraction of a PB/yr. However, this has changed with the advent of several large-installation experiments that employ new detector hardware and multi-year run periods. Hall-A is presently running the SBS program, which will be followed by Measurement of a Lepton-Lepton Electroweak Reaction (MOLLER) and the proposed SoLID detector. Each will be summarized below.

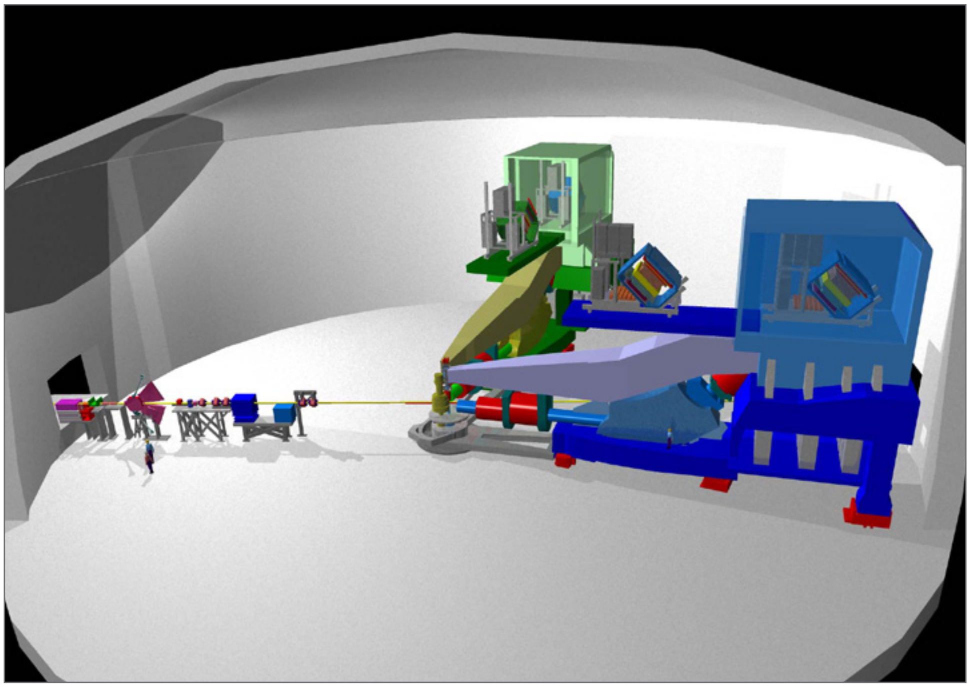


Figure 6.1.2 : Hall A High Resolution Spectrometer configuration.

Super BigBite

The Super BigBite apparatus is a relatively new modular detector system focused on a series of polarized form factor measurements on the proton and neutron over a broad, currently unmeasured, Q2 range. This program began with a measurement of GMn in 2022 and continues with a measurement of GEn using a polarized He3 target that will run through 2023. The program is expected to close with a year-long 'flagship' measurement on GEp running from 2024–25. The core of this system is a large acceptance, medium-precision, magnetic spectrometer. This consists of a ~2 Tesla-meter dipole for momentum separation of charged particles and a configurable detector stack consisting of gas-electron-multiplier (GEM) tracking layers, scintillators, gas cherenkov, and calorimeter devices as required. The apparatus is designed to operate at beam-target luminosities up to 1039 cm-2s-1. The SBS experimental group is expected to acquire roughly 2 PB of data in 2023 and 6 PB/yr during the 2024–25 GEp measurement (comparable to hall-B and D data volumes). The SBS collaboration plans to do the bulk of the analysis and simulation on-site using the JLab Farm.

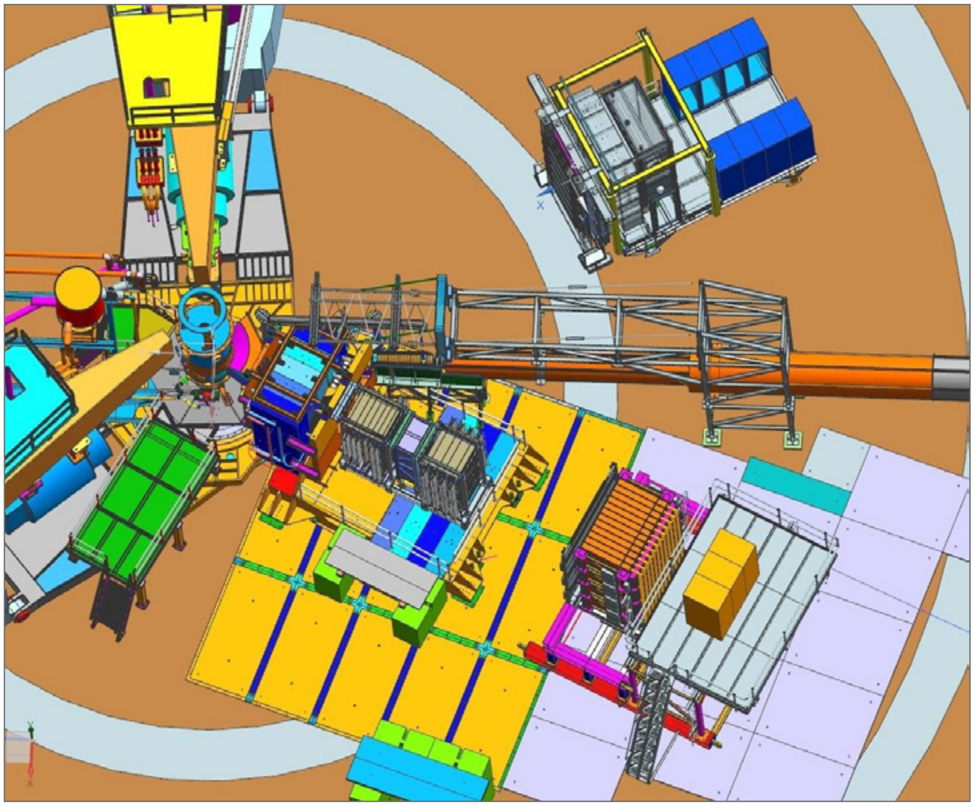
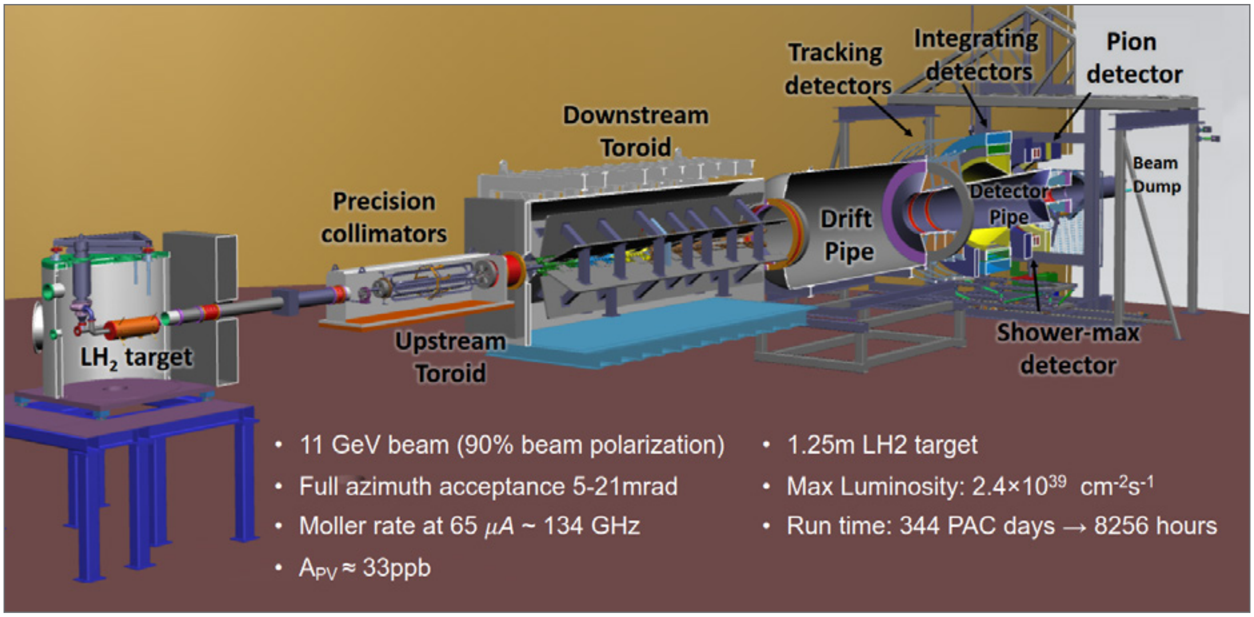


Figure 6.1.3: Hall A / Super BigBite configuration.

MOLLER

The MOLLER facility is an approved 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. 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.



- 11 GeV beam (90% beam polarization)
- Full azimuth acceptance 5-21mrad
- Moller rate at 65 μA ~ 134 GHz
- $A_{PV} \approx$ 33ppb
- 1.25m LH2 target
- Max Luminosity: 2.4×10^{39} $cm^{-2}s^{-1}$
- Run time: 344 PAC days → 8256 hours

Figure 6.1.4: Hall A / MOLLER target and detector layout.

The detector and experiment funding were formally approved in 2020 and construction is now underway. Installation into Hall A is expected to begin in 2025, with initial beam tests starting in 2026. The full program is scheduled to run through 2029.

The MOLLER Experiment plans to acquire about 4 PB/year of raw data over a three-year run period. To put this into context GlueX acquires ~8 PB/year of raw data over the same period. Raw data analysis will be carried out at JLab. An intermediate step in the analysis will generate about 20 PB of data over a ~six-year production run and heavy analysis period. These intermediate data will again be analyzed with compute resources at JLab. Over the course of the experiment and analysis (four+ years) about 1 PB of data will be transported off site for analyses at collaborating institutions. 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 are generated.

SoLID

SoLID is a proposed large acceptance detector intended for use over multiple run groups 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.

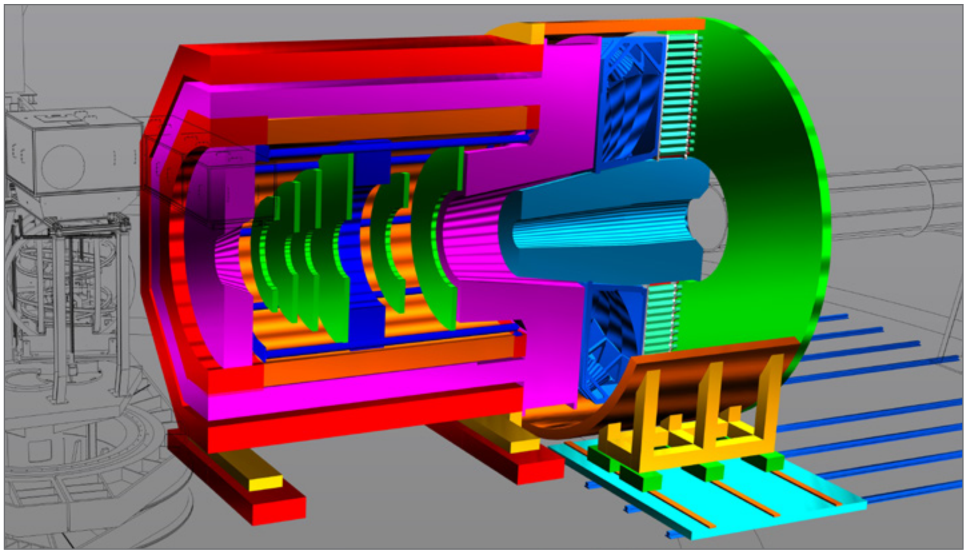


Figure 6.1.5: Hall A / SoLID detector.

In one configuration, SoLID will be used for Semi-Inclusive Deep Inelastic Scattering (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 Parity Violating Deep Inelastic Scattering (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.

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 and would follow the MOLLER program. That would mean installation of SoLID could not happen before 2030.

In the SIDIS J/Psi mode the SoLID detector will produce data at a rate of 3–4 GB/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 6 GB/s which would add 175 PB of raw data over the same three-year period. These numbers should be treated with caution because they do not take into account reductions in dataset size due to filtering, online processing before storage, compression algorithms etc. Also, these rates assume a traditional triggered data acquisition (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, this mod also envisions an online compute farm that would implement a software “trigger” and other techniques to reduce the size of the stored dataset. 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 models employed by halls B and D, using 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 fully developed. It is likely that highly reduced and pre-processed data sets would likely be exported off site at a (roughly estimated) rate of 2–4 PB/year over a six-year production and analysis period.

Hall-B

Hall-B is currently instrumented with the CLAS12 detector. 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.

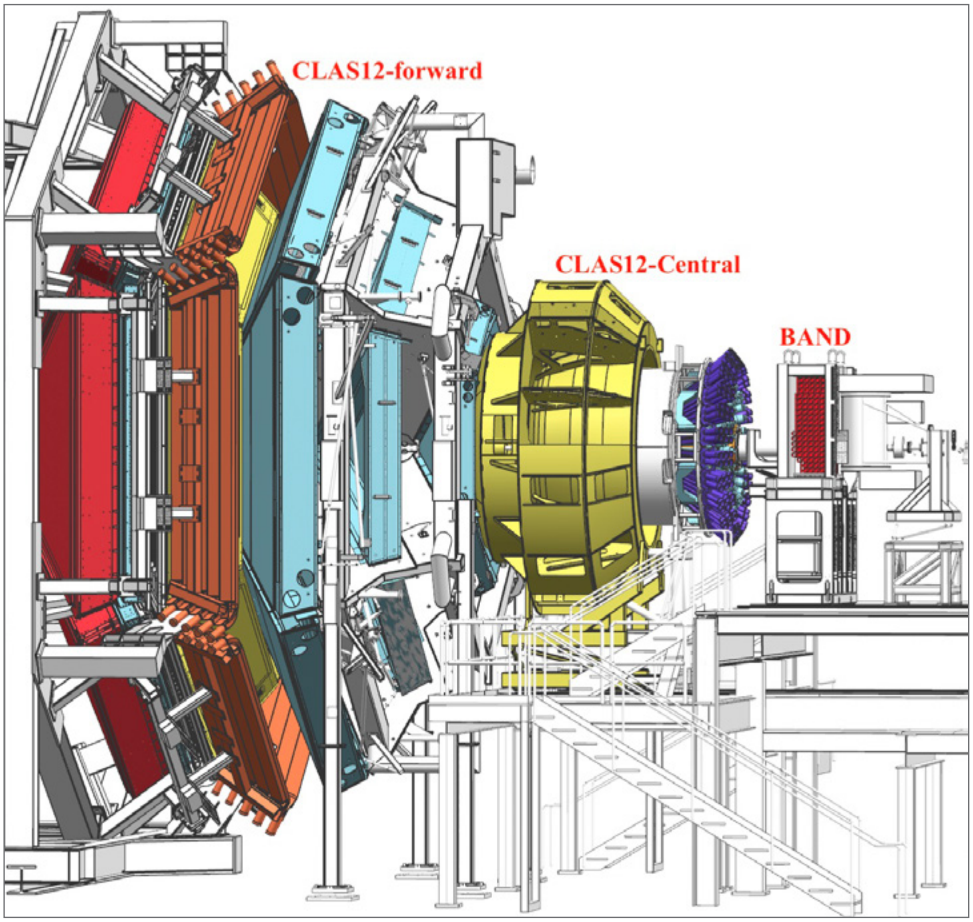


Figure 6.1.6: Hall B / CLAS12 detector.

The CLAS12 detector is designed to simultaneously run multiple experiments that have the same target and beam condition requirements. For example, Run Group C that took data over 2022—3 produced a dataset that will be shared by eight different studies. In its highest rate mode CLAS12 takes data at ~1 GB/s, accumulating ~4–8 PB/yr.

Hall-C

The core detector systems in hall-C are two high-momentum spectrometers. Similar to hall-A, the experimental program in hall-C generally consists of a series of relatively short-lived experiments. Although hall-C can perform experiments requiring high beam energies and luminosities, the historical data rates and event sizes generate a data volume at about 1/10 the scale of halls-B and D. However, newer technologies and experimental approaches requiring digitized waveform analysis and pipelined readout are expected to increase data rates in hall-C by at least an order of magnitude in the next few years. The current experimental program involves installation of a new highly-segmented calorimeter called NPS (Neutral Particle Spectrometer). It is designed to sit at forward angles and withstand high luminosity neutral and charge particle flux off the target. Particle ID and background suppression requires digitized waveform data and will impose peak data rates on the order of 300 MB/sec for extended periods with total volumes of roughly 2–3 PB per year from 2023–24. NPS comes off the floor in 2025 and the hall program is expected to revert to more historical data rates, collecting <0.5 PB/year. Analysis of hall-C data will use the JLab Farm with minimal off-site data transfers.

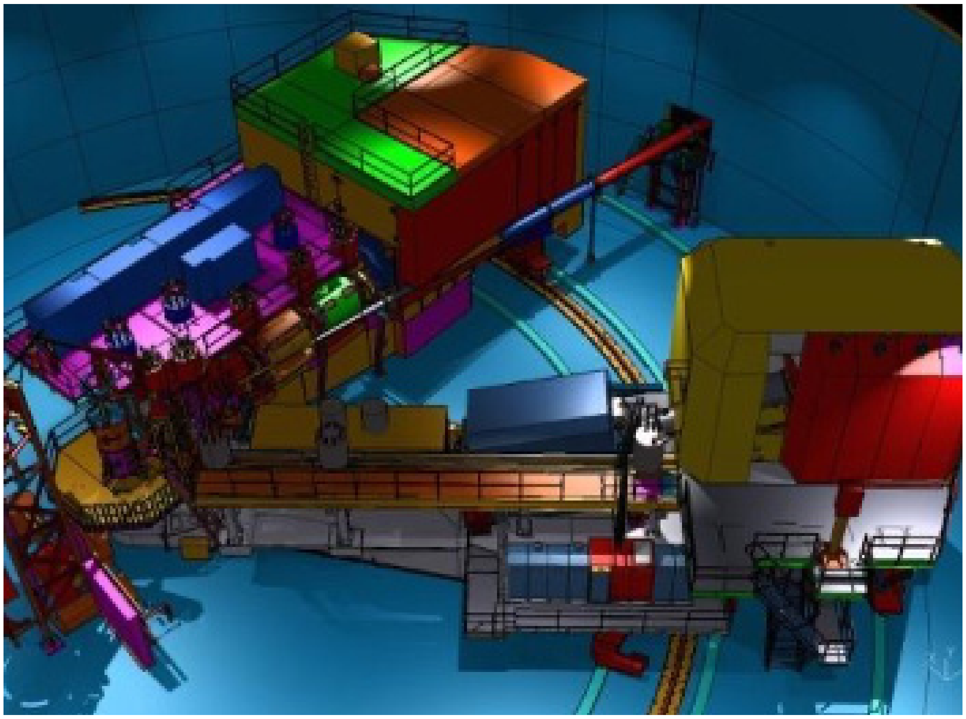


Figure 6.1.7: Hall C spectrometer configuration.

Hall-D

Hall-D is the newest of the four halls and was constructed as part of the 12GeV accelerator upgrade. The main instrument in hall-D is the GlueX detector shown in Figure 6.1.8. The GlueX Collaboration consists of about 150 members from more than 30 international institutions. GlueX is designed to operate with a high-intensity polarized photon beam generated by the CEBAF electron beam. The GlueX target is surrounded by the detector. which is itself inside a solenoid magnet. The detector is asymmetric along the beam axis since most of the particles coming from interactions in the target travel in that direction. that direction. GlueX completed its first phase of operations in 2018, where it generated about 500 MB/s, taking data at an event rate of 40kHz. The second GlueX phase including an upgraded detector is presently running at twice that event and data rate, roughly 1 GB/s. Phase II concludes in 2025, collecting an estimated 10 PB of raw data spread over 2023–5. Subsequent experiments are expected to run at similar data rates (4–8 PB/year).

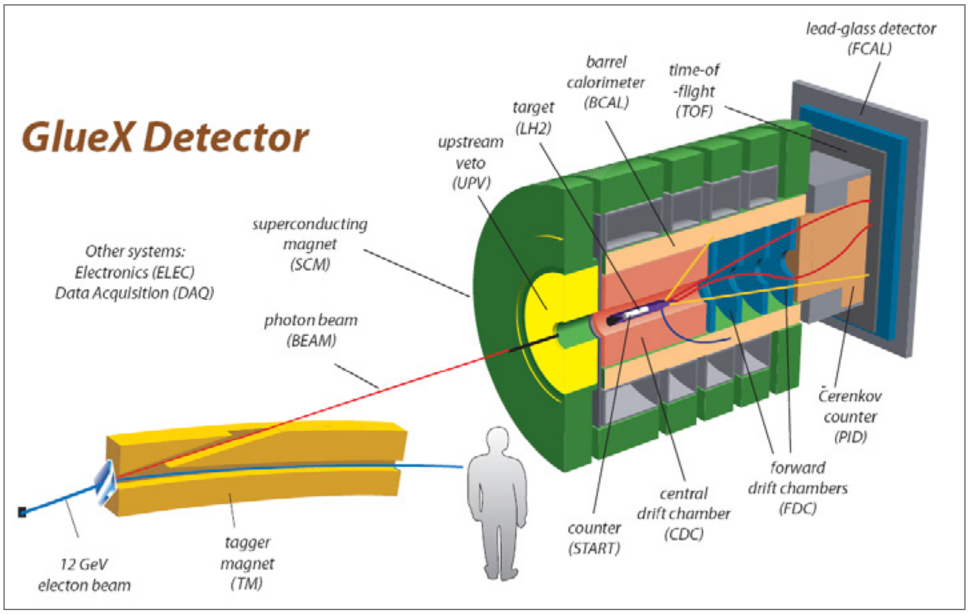

GlueX Detector
lead-glass detector (FCAL)
barrel calorimeter (BCAL)
time-of -flight (TOF)
target (LH2)
upstream veto (UPV)
superconducting magnet (SCM)
Other systems: Electronics (ELEC) Data Acquisition (DAQ)
photon beam (BEAM)
Čerenkov counter (PID)
forward drift chambers (FDC)
central drift chamber (CDC)
counter (START)
tagger magnet (TM)
12 GeV electon beam

Figure 6.1.8: Hall D / GlueX detector.

Compute and Storage

The compute and storage requirements for the experimental program are estimated 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 every other year by an external review panel. The ENP computing coordinator works with IT to develop a strategy that will meet the requirements in the current to two-year timescale and plan strategically for experiments and equipment upgrades lying further in the future. The most recent external review was completed in early-2022, and the next is scheduled for Feb 2024. JLab has planned for a roughly two-fold increase in data volumes is planned over the next three years as SBS/GEp-V begins production in hall-A and a detector upgrade is completed in hall-D. The next large jump is expected to be from SoLID toward the end of this decade.

The science workflow will be described in more detail in the [Section 6.1.2.4] 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 only taken for a fraction of the wall clock time, maybe 60-70%. The total number of events in the dataset can be estimated in advance by multiplying the average event rate by availability and the total run time of the experiment. Similarly, the total data volume can be estimated by multiplying the total number events in the dataset by the average event size. Once these values are established the compute and storage requirements for an experiment can be estimated since they scale linearly with events in the data set. Each hall currently has local storage close to the detector that can hold at least 48 hours of data. Each Hall is connected to the data center with redundant 100 Gbit/sec links. Data flow from the four experimental halls is expected to peak at 3–4GB/sec for the most demanding combination of experiments over the next 5–7 years. Note that those rates are driven by beam-on data production. Daily/weekly average rates are significantly less. Once the data are staged on the local storage, control is passed to systems operated by Scientific Computing Operations that copy the data over the LAN to staging area in the data center. From there the data are archived in an IBM tape library with a current capacity of ~200 PB and 13,000 tapes.

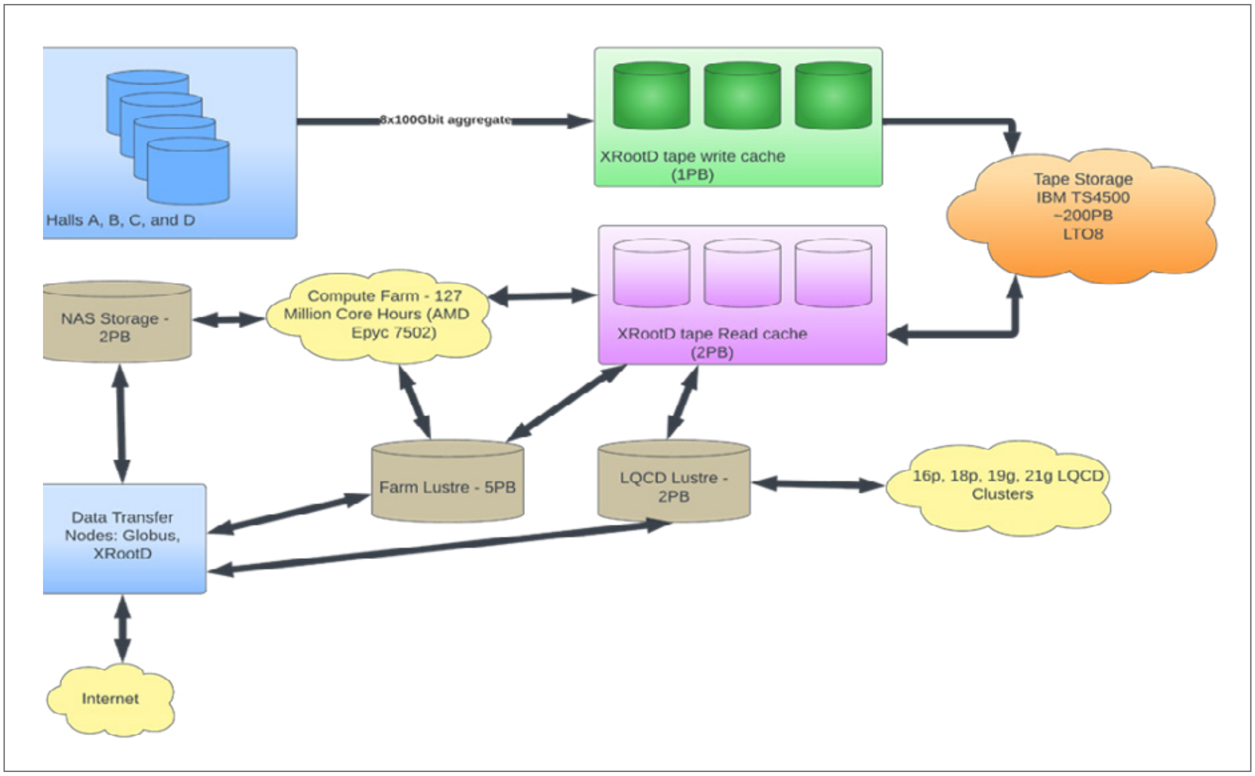


Figure 6.1.9: Block diagram of JLab local compute resources.

The JLab local computing resources are outlined Figure 6.1.9, as well as the systems described in the previous paragraph there are clusters that are used for Theory/LQCD (covered in a separate case study). An HTC cluster is used by ENP to process data from the experiments, as described in [Section 6.1.2.4]. There is a path from the mass storage system to the WAN that will be described in more detail in [Section 6.1.2.7].

6.1.2.4 Generalized Process of Science

The detectors forming the baseline equipment in the four halls were described in [Section 6.1.2.3]. The process of science begins with individual or groups of PIs presenting proposals for experiments to an advisory committee, the Program Advisory Committee (PAC), whose tasks are to prioritize experiments based on scientific merit, recommend experiments that should be scheduled, and for how many days. The scheduling unit is the so called "PAC day" which is equivalent to two calendar days. As of Spring 2023 there are 2,940 PAC days of experiment approved spread across more than 51 individual experiments. Since several halls can receive beam simultaneously, one calendar year provides 300–370 PAC days of experiment time and the current list of experiments represents a backlog of between 7–10 years. Each year the PAC may approve more experiments and 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 digitizes, formats, and stores 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 that are described below.

Calibration

A sample of the raw dataset, often runs taken under specific calibration 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 5–10% of the total dataset 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 dataset 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. Consequently, JLab provisions local resources to cover the calibration workflow.

Reconstruction

In the reconstruction phase the calibration is applied to the bulk of the dataset to convert raw data into measured physical parameters. Reconstruction requires access to the entire dataset. Expected data volumes from halls A, B, and D are expected to be 6–10 PB/year/hall over the next 5–10 years. The output of reconstruction jobs varies by hall but are 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 roughly 20% of the size of the input, and so is about 1.2 PB/yr. Usage of off-site resources supporting reconstruction (particularly NERSC and OSG) has expanded over the last several years and imposes an increasing load on the WAN and ESnet bandwidth. This is included in the provided projections.

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 dataset the reconstructed data are first sorted into physics channels of interest to individual experiments. The size of this dataset relative to the raw data varies by experiment. For example, GlueX is a single experiment but 80% of the events in the dataset do not represent physics of interest to GlueX so only 20% are included in final stage analysis. Even so, the GlueX raw dataset, at ~6–10 PB/yr, is large enough that the input for analysis is still ~1 PB/yr. For some experiments taking part in the run groups using CLAS12 data only 1% of the total CLAS12 dataset is of interest and the analysis input is only tens of terabytes. Unlike prior years, where hall-A data sets were small compared to CLAS12 and GlueX, the SBS program in hall-A is expected to generate 5–6 TB/year. Current planning is for that analysis to be completed using internal Farm resources. The next major challenge is expected to be SoLID in hall-A (2030 timeframe). It is expected that that program will learn from, and follow analysis models similar to CLAS12/GlueX, with a large fraction of their data moving off-site for distributed analysis at non-JLab datacenters. The program planned for hall-C has proportionately small data volumes compared to the other three halls.

Simulation

A fourth computational process is simulation. Here a model of the detector along with, detector calibration and numerical models of physical processes is 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) 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 will simulate a number of events proportional to the number in the experimental dataset with a ratio that depends upon the required statistical accuracy. Using an earlier example, the experiment that uses 1% of the CLAS12 data, of O(100 TB), as the input to its analysis may require ten times that number of simulated events, which is of O(1 PB) simulated dataset. These datasets 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 ten 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 that the output of simulation be returned to JLab for storage or analysis so it has a contribution to the incoming WAN traffic that may be similar in scale to reconstruction's contribution to the outgoing WAN traffic.

The JLab scientific computing requirements will jump in 2024 when the SBS/GEp-V program ramps up in hall-A, and again in 2025 when a significant upgrade to the GlueX detector in hall-D comes online. As noted previously, the next increase in requirements will be towards the end of the decade associated with SoLID in hall-A.

6.1.2.5 Remote Science Activities

At any time, there are several experiments at different points in their simulation, calibration, reduction, and analysis phases. This overlap and interplay of usage patterns leads to a base level of compute load with frequent, bursty times of peak demand. It is not cost effective for the laboratory to provision for peak workloads, nor is it reasonable for wait times to vary too widely. To balance these concerns, off-site compute allocations offered by other institutions are part of the data processing strategy that smooths peak usage.

The largest example of this are the long-term simulation campaigns run by GlueX and CLAS12 on the OSG. Figure 6.1.10 shows the variation in use of OSG resources over the last year. In the 2019 case study GlueX dominated OSG usage. CLAS12 has now taken the lead in consumed core hours, and the two together represent one of the largest single uses of the OSG. Figure 6.1.11 shows the breakdown of OSG usage by institution and collaboration.

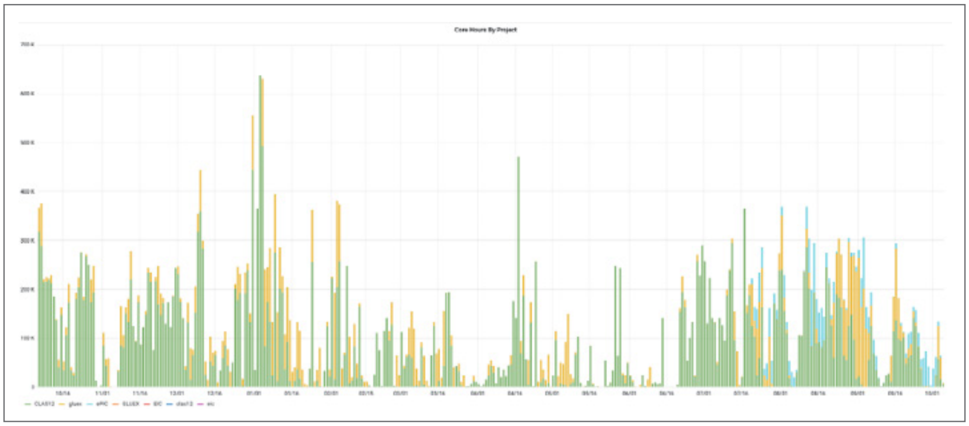


Figure 6.1.10: OSG resource usage variation over the last 12 months (2023).

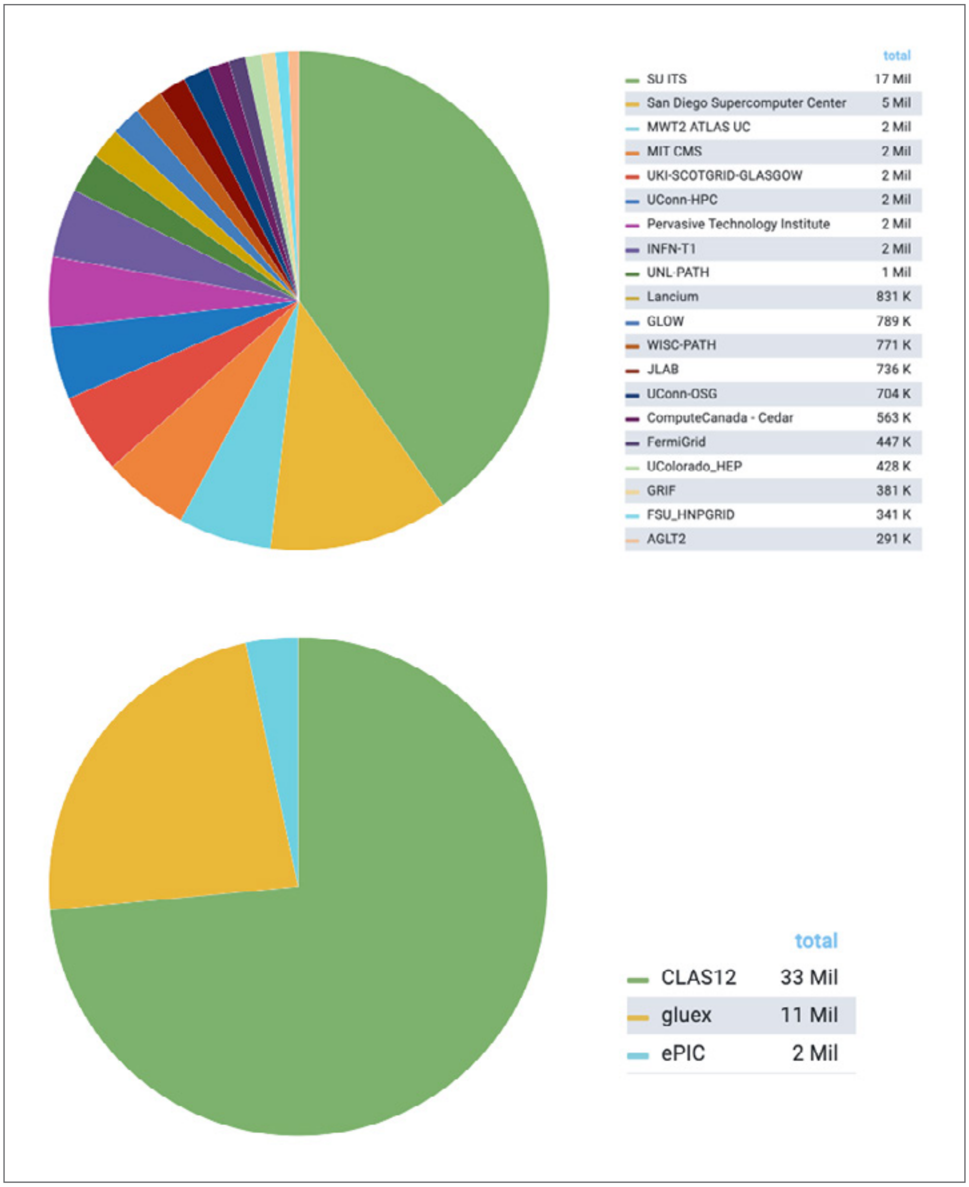


Figure 6.1.11: OSG usage breakdown by institution and Collaboration (top). Example of OSG core hour use for GlueX, CLAS12, and ePIC over the 2023 (bottom).

To support OSG operations, the laboratory has expanded its OSG infrastructure to eliminate single points of failure. There is an ongoing project to create job submission and data-transfer services per collaboration that can be adjusted to make efficient use of OSG resources. For example, CLAS12 has institutions that have dedicated cycles to CLAS12 simulation. This tailoring of OSG resources is reflected in the design of JLab's data transfer and submission nodes.

In addition to OSG for Simulation, GlueX has performed the processing of a raw data subset off site, most notably at NERSC. This is a more data-intensive workflow, with large input and output. JLab's SWIF workflow manager is used to orchestrate the movement of the raw data from tape, over the network to NERSC, and the results back again. The total raw data processed at NERSC in 2022 were 1.5 PB, and .5 PB (to date) in 2023. The processed data output transferred back to JLab for storage were roughly 1/3 of those figures.

Although NERSC is the single largest contributor to GlueX raw data processing, Pittsburgh Supercomputer Center (PSC) Bridges-II and at Indiana University Big Red are also used. At the Pittsburgh Supercomputing Center, GlueX used 2.8M core hours between 2021-07-01 and 2022-06-30 on Bridges-II. This corresponds to roughly 8k node hours at NERSC Perlmutter. Figure 6.1.11 shows the previous year of contributions by OSG sites to processing for CLAS12, GlueX, and ePIC, in excess of 44 million core hours.

GlueX have also pioneered the use of NERSC to process JLab NP data. NERSC had a preference that we process raw data rather than simulate so we have run GlueX reconstruction on Perlmutter. In 2022 GlueX used 92K node hours at NERSC, and in 2023, GlueX was awarded 85K node hours. This data-intensive workflow can saturate one of the lab's two 10 GBit/sec circuits during periods of file transfer. This bottleneck is being addressed by the 100 GBit WAN upgrade currently in progress.

6.1.2.6 Software Infrastructure

There are two categories of software used within JLab: those used within the facility for scientific workflow management, and those used to face the internet for data mobility.

Facility Software Infrastructure

Prior to the upgrade of the accelerator from 6 to 12 GeV, which included the addition of GlueX and the upgrade of the CEBAF Large Acceptance Spectrometer (CLAS) to CLAS12, the event and data rates from the experiments were an order of magnitude lower than they are anticipated to be 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.

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.

Internet-Facing Software Infrastructure

Several software infrastructure projects for data movement are under development and will expand. We anticipate that the completed build out of the 100 Gbit circuits will incentivize this work. Software projects under development include XRootD based DTNs; Rucio for policy-driven data movement to remote sites; and Globus for file movement to leadership compute facilities and users home institutions.

The Rucio project at JLab is in its early stages, but multiple experiments and use cases have been identified. Starting in summer 2023, a project was begun to incorporate Rucio into the data pipeline from the experimental halls, creating a source of truth for policy-based file movement of raw data. Rucio naturally builds upon the XRootD storage that is already part of the lab's tape storage system

A project to use XRootD storage with SciToken based authorization for writing from OSG jobs is also underway. This mechanism, still in early user testing, makes use of the OSG's transition to bearer tokens, and writes files back to JLab through DTNs on the Science DMZ.

Globus endpoints, also on the Science DMZ network, are used most extensively by USQCD members, particularly with OLCF.

SWIF is a JLab development workflow management tool that coordinates data movement from tape with Slurm job submission. SWIF orchestrates complex workflows for data analysis, and has the ability to retry jobs and report on overall workflow capabilities. SWIF has extensions to allow users to run jobs at NERSC, moving data from tape to NERSC using Globus. Integration of Rucio and XRootD transport into SWIF is in development.

As described in [Section 6.1.2.5], JLab has a significant presence on the OSG. This now includes the capability to run OSG jobs at JLab for projects that are part of the facility mission. This presently includes GlueX, CLAS12, MOLLER, and EIC. The ability to run OSG jobs at JLab, while not the primary driver, gives the ability for some OSG jobs to “flow back” to JLab and improve cluster utilization by backfilling production.

Containerization of jobs is an area of growing emphasis. In the case of both OSG and NERSC the software is containerized using Apptainer and Docker. GlueX and CLAS12 use the CernVM distributed file system (CVMFS) as part of their workflow. Due to its caching architecture, this is not a high bandwidth application, but the availability of a distributed file system is important to the ability to run code off-site. Together, containers, CVMFS, and streaming storage with Rucio/XRootD provide a powerful set of tools for making jobs more site-agnostic.

6.1.2.7 Network and Data Architecture

JLab’s network is anchored by a pair of redundant 100Gbit core routers that act as a pure layer 3 transport for the site’s various network collections, “network pods”, 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. Figure 6.1.12 shows a high-level overview of the lab’s layer 3 design as it exists today.

Each network pod has redundant head-end routers and firewalls. Those firewall/router pairs are housed in separate physical locations and meshed using campus fiber that follows diverse paths. 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, the lab’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.

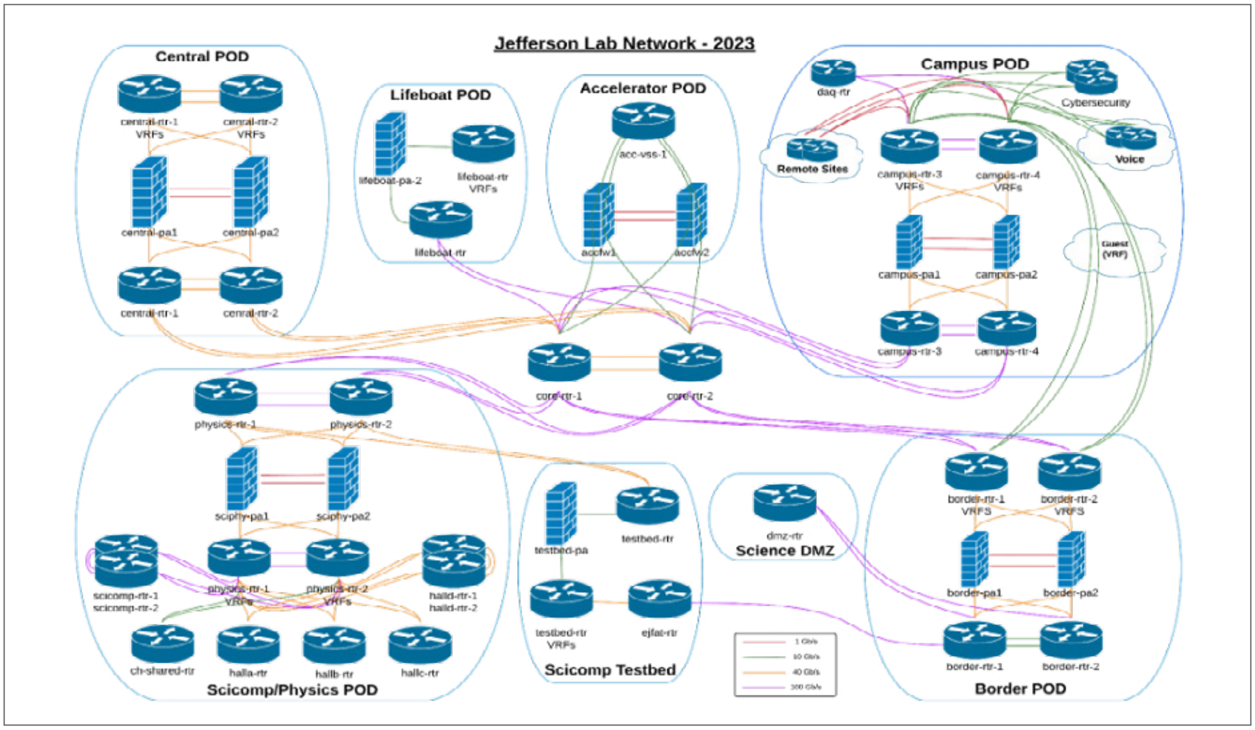


Figure 6.1.12: JLab layer 3 design.

Although the routers in the core and border pods are currently 100Gbit capable Arista equipment, the diagram shows them at their currently connected speeds as of September 2023. The top left pod, for Scientific Computing and Experimental Physics ("SciPhy") is where networking for each of the four experimental halls originates along with the batch farm, LQCD clusters, and mass storage. The Border pod, bottom right, is where connectivity to ESnet is managed, including the Science DMZ which sits outside the border firewall.

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

Both of these cases are illustrated in Figure 6.1.13 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 scientific computing and experimental physics (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.

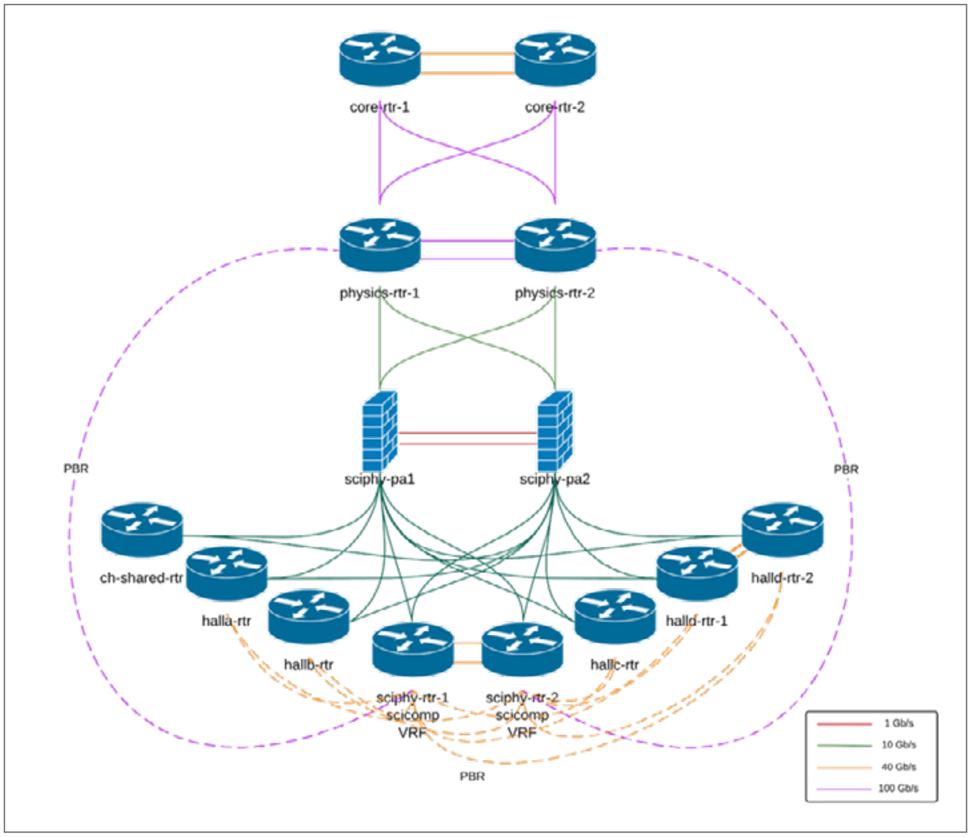


Figure 6.1.13: Block diagram of the scientific computing network layout.

Figure 6.1.14 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 E-LITE, the Metro area ring, which is operated by Old Dominion University. E-LITE Member institutions include Old Dominion University The College of William and Mary, and the National Oceanic and Atmospheric Administration (NOAA). As of this writing, E-LITE is a 10Gbit/sec fully protected dense wavelength-division multiplexing (DWDM) ring. Two physically diverse E-LITE routers provide layer 2 exit from the ring to the wide area. There are two E-LITE DWDM optical nodes located at JLab, collocated with the JLab border routers.

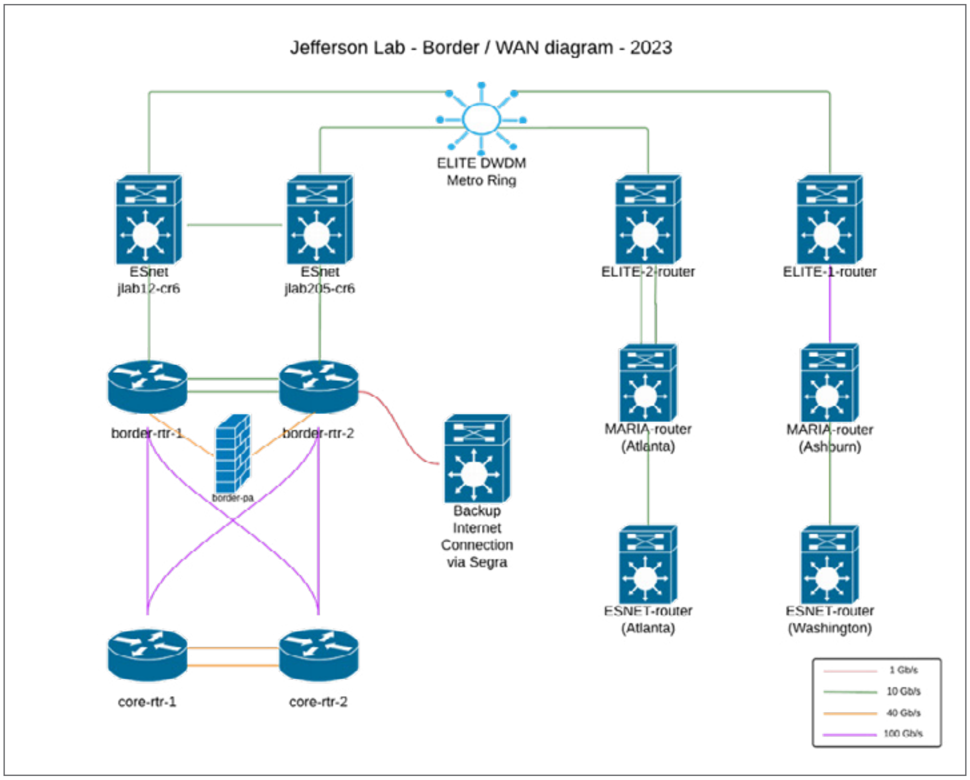


Figure 6.1.12: Layer 2 path to ESnet

The layer 2 path from E-LITE to Atlanta and from E-LITE 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 E-LITE-1 router to Ashburn is a protected 100Gbit/sec circuit. The path from the E-LITE-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 E-LITE 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 in summer of 2023. The Lumos router is a 1Gbit/sec path of last resort.

In addition to the necessary router and switch upgrades, lifecycle refresh of the DTNs and storage capability has been completed. The aim of this project was to build out a robust set of 100Gbit DTNs for data movement to other facilities using Globus and XRootD most commonly.

The Networking and Scientific Computing groups have built out a redundant 100Gbit network, from the data center to the internet border. As soon as the ESnet 100Gbit circuits are in operation, they can be fully utilized by the scientific computing systems, including DTNs.

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

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.

6.1.2.8 IRI Readiness

Within the physics division, a project to organize experimental metadata into a multi-experiment catalog is underway. This work tracks the Rucio development, so that policy-based data movement is made based on discoverable metadata.

The lab network infrastructure has been designed with IRI patterns in mind. This is shown in the network design of the external Science DMZ network for data transfer and a firewalled science portals network for services.

IRI benefits NP by providing a uniform interface to disparate resources in a way that is easier to integrate with, better fit to NP workflows. The JLab NP workflows that benefit most from an IRI approach are those currently running at NERSC and on the OSG. These most accurately fit the long-term campaign pattern.

Most time-sensitive patterns, those at the sub-second scale, are processed locally because of the data volumes involved.

IRI would allow for better EIC integration at BNL, JLab, and other partners as those workflows are developed.

6.1.2.9 Cloud Services

In 2019 JLab moved to using Microsoft Office 365 for email, calendar service, Sharepoint, and OneDrive cloud storage. Although, bandwidth to and from these services is much lower than that required for scientific data, the reliance of the laboratory on cloud services for day-to-day business has made the resilience of Internet connectivity critical. During periods of peak scientific data movement, cloud data for day-to-day work has been affected, and QoS and other traffic engineering controls were put in place to protect the business traffic from being affected by the scientific data flows.

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 OSG and NERSC. Experimental workflows require expedient access to computation on occasion, either to perform a recalibration of experimental data or reconstruct an analysis. These use cases are portable enough to run in a cloud environment, and can be run faster than waiting in local computational queues for service. Due to the current cost of cloud computing we expect that use cases will be rare and when used it will replace one or more of the other resources so the WAN requirements will not significantly increase.

6.1.2.10 Data-Related Resource Constraints

As discussed in [Section 6.1.2.4], the computing requirements for CLAS12, GlueX, and now hall A experiments exceed the local computing resources. Our current plan is to do as much locally as we can but to make extensive use of remote resources, OSG, NERSC, collaborators and maybe cloud. The future 9 PB/yr GlueX raw dataset plus the 8 PB/yr from CLAS12 and 5 PB/year from hall A will rely on the 100Gbit/sec ESnet upgrade that is in progress.

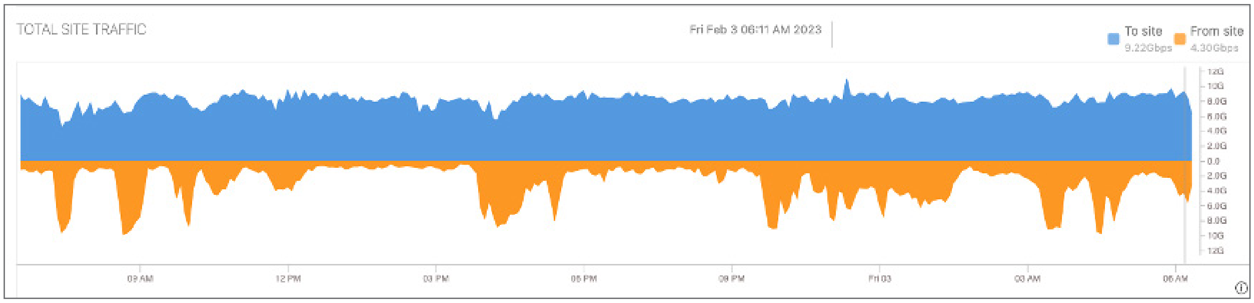


Figure 6.1.15: ESnet traffic associated with GlueX.

6.1.2.11 Data Mobility Endpoints

JLab operated two Globus Endpoints, jlab#gw1 and jlab#gw2. These endpoints both have access to the two Lustre parallel filesystems for experimental physics and LQCD. These endpoints have been tested and evaluated with ESnet to baseline their performance at 10Gbit. Once the 100Gbit ESnet upgrade is completed, these systems will need to be baselined again. The endpoint has 100Gbit network interface controllers (NICs) on the Science DMZ and the Lustre systems can deliver 100Gbit/sec. The present bottleneck is the 10Gbit Internet connection that is in the process of being upgraded.

6.1.2.12 Outstanding Issues

In the ESnet5 era, JLab has been connected to ESnet over a complex Layer 2 fabric that spanned the E-LITE and MARIA, with the first upstream ESnet-managed routers at Ashburn and Atlanta. This left a blind spot in traffic monitoring between JLab and the ESnet cross-connect points that made debugging difficult at times. With ESnet6 and the installation of routers at JLab, additional insight into network operations was gained.

The primary outstanding issues are bandwidth limitations and the complexity of the layer 2 fabric. The bandwidth limit is the most common user-visible problem. Both of these concerns are being addressed by the upgrade from 2x10Gbit to 2x100Gbit internet connections. The 100Gbit circuits are ESnet provided lit fiber, with no shared bandwidth at layer 2 between JLab and ESnet.

Despite the complexity of the layer 2 transport over E-LITE and MARIA, The level of resilience provided by ESnet for JLab networking has been excellent. The upgrade to ESnet6 routers (done) and 100Gbit (in progress) will sustain that level of resilience.

Integrating with other labs and data facilities presents ongoing challenges in authentication and federated identity. While frameworks like x.509, OAuth, and tokens have broad consensus, issues arise from nonstandard or incomplete implementations.

Federated interactive logins that leverage approaches such as SAML, OpenAuth2, and Shibboleth are standardized and interoperable, but “automated” authentication for tasks like remote job control and large file transfers is fragmented and challenging to debug.

In addition to federated user identities, facility-to-facility trust implementations are also difficult from both policy and technical perspectives. Cross-facility systems often rely on long-lived “robot” tokens or certificates that are an extension of a user’s identity. These are not always a good fit for long-term campaigns. It is often the case that actions need to be taken on-behalf-of or as-a-proxy for users which can stretch the abilities of federated identities.

In light of the crucial role of authentication in the IRI vision, a central repository with reference designs and implementation examples for common workflows would offer significant value.

6.1.2.13 Facility Profile Contributors

JLab Representation

- Amber Boehnlein, JLab, amber@jlab.org
- Bryan Hess, JLab, bhess@jlab.org
- Graham Heyes, JLab, heyes@jlab.org
- Brad Sawatzky, JLab, brads@jlab.org

ESCC Representation

- Brent Morris, JLab, bmorris@jlab.org

6.2 Theory Group and LQCD at JLab

JLab operates large clusters of computers for LQCD, as part of the Nuclear and Particle Physics LQCD Computing Initiative (NPPLCI) established by the DOE SC. Their mission is to extend the fundamental understanding of nucleons and their quark constituents and to provide essential dedicated computing capability for critical nuclear theory calculations that are complementary to its experimental program. JLab will continue ongoing R&D in mixed architectures incorporating general-purpose graphics processor units (GPGPU), as well as potential future architectures such as those proposed for the ECP. Time on these clusters is scheduled by the USQCD collaboration, complementing the multi-petaflops resources deployed at the DOE and National Science Foundation (NSF) supercomputing centers.

6.2.1 Discussion Summary

The following bullets were extracted from the case study, and in-person discussion. Several appear in Section 1 and Section 3.

- JLab is an integral part of the USQCD collaboration, a consortium consisting of approximately 160 individuals from around 50 institutions, including universities and research laboratories.
- LQCD data retention at HPC facilities is typically short-term. The data are subsequently transferred back to local computing resources for long-term storage.
- The global analysis effort for LQCD involves collaboration between experimentalists and theorists at approximately 10 universities and research laboratories across the United States. The computational requirements for these analyses are comparatively lower than those for the LQCD projects. This results in data movement to and from JLab that can total several PB per year, but in smaller TB-sized data sets.
- The workflow for JLab LQCD computing involves around 30 allocated projects, with a collective allocation of 2 PB of disk storage and 15 PB of tape storage. These projects typically generate data in the range of hundreds of terabytes during their yearly allocation period, with a portion of these data being transferred off-site.
- LQCD calculations rely on a formulation of the theory in Euclidean space, and as such, are not real time calculations. MC sampling of the theory provides snapshots of configurations of quarks and gluons. These snapshots, typically generated on LCFs, are used in subsequent calculations of observables that are carried out in ensemble calculations. From these observables, theoretical calculations guide, and confront, experiment measurements.
- LQCD projects at JLab exhibit variability based on their allocations. Some may heavily rely on JLab's computing resources; these flagship projects initiate their datasets on the LCFs with initial datasets typically amounting to around 300 TB spread across approximately 10,000 files.
- With the advent of new computing facilities such as Perlmutter at NERSC and Frontier at OLCF, the volume of data generated by the LQCD community at LCFs is increasing by a factor of 5 to 10. Existing workflows will remain in use, leading to a corresponding increase in the amount of data that needs to be transferred back to JLab.
- The Aurora system at ALCF will likely mark a significant milestone in data production. An estimated 3x more data are expected to be generated on Exascale systems compared to the current generation of systems (Summit, Perlmutter). For LQCD projects in 2025, this could result in approximately 10 to 15 PB/year of data generated on LCFs and transferred back to the lab, with most of the data being utilized locally.
- Members of the LQCD project teams seek allocations of computing time at numerous HPC facilities. While certain data may be retained for an extended period on leadership systems

like NERSC, the primary responsibility for long-term data storage lies with the member laboratories. In the case of LQCD projects related to the JLab science program, JLab will serve as the host for the extended data storage.

- The LCFs do not provide long term storage for LQCD projects. Data are transferred to the USQCD computing facilities (JLab, Fermilab, and Brookhaven) which assume ownership. Data transfers to JLab LQCD will increase and scale with the size of new LCF systems. JLab will continue to serve as the repository for long-term storage. A portion of the analysis work will be conducted at the LCFs. However, the final stage of the analysis workflow is ideally suited for execution on JLab's local systems, effectively mitigating the disparity in LCF to local computing capability.
- Over the next five years computing allocations of a few million node-hours per year for projects will be available for LQCD projects on Frontier at OLCF. It is anticipated that similar amounts will be available in the future for Aurora at ALCF. With these computing allocations, a few PB of data will be generated for each project. These data will be transferred to JLab, BNL and Fermilab for further analysis. In total, about 10 PB is anticipated for transfer over that period.
- On occasion, HPC systems in Europe are heavily used in collaborations with LQCD European researchers; these initial datasets are then used for secondary calculations also carried out at DOE HPC facilities, resulting in datasets of about 1 PB distributed over about 1,000,000 files.

6.2.2 LQCD Experimental Case Study

The origin of the proton and neutron's structure and the interactions between them can be traced back to a fundamental quantum field theory known as quantum chromodynamics (QCD). This theory governs the behavior of quarks and gluons, which are the elementary components of the observable matter in our surroundings. Experiments conducted at high energies have extensively validated QCD, providing valuable insights into the workings of nature at scales smaller than nucleons (the collective term for protons and neutrons). However, when dealing with low energies or larger distances, QCD becomes a challenging framework, and attempts to derive fundamental nuclear physics phenomena directly from it have seen limited success. The US DOE NP program has long been dedicated to unraveling how QCD in this low-energy regime translates into the observed spectrum of hadrons and nuclear phenomena. Moreover, it aims to employ QCD to make reliable predictions for processes that are beyond the reach of experimental investigations. These theoretical endeavors play a crucial role in supporting the DOE's nuclear experimental initiatives, particularly those taking place at JLab, Brookhaven's RHIC, Michigan State's FRIB, and the planned EIC.

6.2.2.1 Science Background

In JLab's Theory Group, various approaches are being pursued to unveil this intricate structure. One major undertaking involves the use of Lattice QCD to generate predictions that not only guide experiments but also serve as a basis for comparison. These calculations rely on HPC resources available at DOE and NSF centers, as well as local computing facilities. Since data retention at these facilities is typically short-term, the data are subsequently transferred back to local computing resources for long-term storage.

Lattice QCD calculations rely on a formulation of the theory in Euclidean space, and as such, are not real time calculations. MC sampling of the theory provides snapshots of configurations of quarks and gluons. These snapshots, typically generated on LCFs, are used in subsequent calculations of observables that are carried out in ensemble calculations. From these observables, theoretical calculations guide, and confront, experiment measurements.

There are several stakeholders. Domain scientists in nuclear and high energy physics design the calculational campaigns. In cooperation, with applied and mathematical scientists, the algorithmic and software infrastructure needed for the campaigns are assembled.

The LCFs do not provide long term storage for the projects. Data are transferred to the USQCD computing facilities (JLab, Fermilab, and Brookhaven), which assume ownership.

Researchers in the Theory Group at JLab also collaborate with the Global Analysis efforts to combine data produced from the lattice QCD campaigns as well as experimental efforts. These projects utilize both sets of data to constrain the parameterizations of observables.

6.2.2.2 Collaborators

The Theory Group at JLab comprises laboratory staff, joint staff members affiliated with neighboring universities, and bridge faculty members. This diverse team is engaged in a multifaceted endeavor aimed at comprehending the fundamental origin of matter.

JLab is an integral part of the USQCD collaboration, a consortium consisting of approximately 160 individuals from around 50 institutions, including universities and research laboratories. These researchers are affiliated with lattice QCD groups that typically span multiple institutions. Within this collaboration, there are local computing facilities hosted at JLab, Fermilab, and BNL, facilitating coordinated calculations. Additionally, parts of the calculations are conducted using DOE and NSF LCFs at Argonne, Oak Ridge, and NERSC as well as European facilities. Subsequently, the results of these calculations are returned to the laboratories for further analysis, utilizing local computing resources supported by Brookhaven, Fermilab and JLab.

Members of the lattice QCD project teams seek allocations of computing time on all these facilities. While certain data may be retained for an extended period on leadership systems like NERSC, the primary responsibility for long-term data storage lies with the member laboratories. In the case of LQCD projects related to the JLab science program, JLab will serve as the host for the extended data storage.

The global analysis effort involves collaboration between experimentalists and theorists at approximately 10 universities and research laboratories across the United States. The computational requirements for these analyses are comparatively lower than those for the LQCD projects.

User/collaborator and location	Do they store a primary or secondary copy of the data?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of dataset? (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)?
J. DUDEK, R. EDWARDS, K. ORGINOS, D. RICHARDS, A. RODAS, J. CHEN, E. ROMERO, F. WINTER (JLAB) R. BRICENO (LBNL) B. JOO (ORNL)	Primary	Multiple different data sets of 500 TB, totaling ~2 PB generated off-site and moved to JLab per year. On-site, generating about 1 PB/yr Total of 3 PB/yr	Lustre file systems and tape storage at JLab. Access via tape and Globus	Ad hoc	About 100 TB/year sent to NERSC, OLCF, ALCF	Globus is sufficient
W. MELNITCHOUK, N. SATO, (JLAB)	Secondary	10 TB datasets < 0.1 PB/year of external data transfers	Data portals	Over next 5 years, evolving to support real-time analysis	No	In 5-10 years, data analysis rate requirements significantly increasing
OTHER USQCD MEMBERS, INCLUDING ANL, BNL, LBNL, ORNL, MICHIGAN STATE, KENTUCKY, U. WASH., U. CONN.	Primary	Multiple datasets of 100 TB each. < 1PB/year of external data transfers Total storage at JLab from LQCD, 4 PB/yr	Lustre file systems and tape storage at JLab. Access via tape and Globus	Ad hoc	Data sent to universities	
ALL USQCD MEMBERS	Primary	New LCF systems on line by 2025. Expect ~3x increase in data set sizes generated by USQCD member. Size estimates from 10 PB to 15 PB.	Data portals, transferring to Lustre/tape storage at JLab	Ad hoc	About 500 TB/year send to NERSC, OLCF, ALCF	Globus should be sufficient for transfers.

Table 6.2.1: Collaborative Data Mobility

6.2.2.3 Use of Instruments and Facilities

The HPC and LQCD Computing Systems at JLab encompass several clusters: an 8-node octal Advanced Micro Devices, Inc. (AMD) CPU and AMD MI-100 GPU system with Infiniband; a 440-node Xeon Phi (Knights Landing) cluster with Omnipath, and a 32-node eight-way RTX-2080 cluster equipped with 256 GPUs. These systems are complemented by a 3 PB Lustre parallel distributed file system, as well as an IBM TS3500 Tape Library shared between the Experimental Physics and the HPC programs. Over the next two years, upgrades are planned for the Lustre file system, along with an increase in its storage capacity. Furthermore, the computational capabilities of the LQCD GPU system are set to increase with the installation of a new 100-node Intel CPU Saphire Rapids system in the fall of 2023, and a new computing system in 2025.

LQCD projects at JLab exhibit variability based on their allocations within the JLab system. Flagship projects, such as those focused on hadron spectroscopy and hadron structure, heavily rely on JLab's computing resources. These projects initiate their datasets on the LCFs, including the ALCF, the OLCF and NERSC systems, with initial datasets typically amounting to around 300 TB spread across approximately 10000 files. In addition, HPC systems in Europe, including France, Italy, and Finland are heavily used collaborations with European researchers. These initial datasets are then used for secondary calculations also carried out at ALCF, OLCF, NERSC, resulting in datasets of about 1 PB distributed over about 1,000,000 files. All these datasets are subsequently transferred to JLab for further analysis within the scope of a yearly allocation. Overall, about 2 PB are generated off-site and then transferred to JLab each year.

These externally generated datasets from LCFs are combined with other datasets, resulting in approximately 1 PB of data generated at JLab, which is primarily utilized there and does not leave the lab.

With the advent of new computing facilities such as Perlmutter at NERSC, Frontier at OLCF (currently operational), the volume of data generated at LCFs is increasing by a factor of 5 to 10. Existing workflows will remain in use, leading to a corresponding increase in the amount of data that needs to be transferred back to JLab.

Looking ahead two to five years, the Aurora system at ALCF (expected to be operational in 2024) will likely mark a significant milestone in data production. Members of the USQCD collaboration have been actively involved in the ECP, collaborating directly with Intel and ALCF to deploy efficient codes on Aurora. The project finished in 2023 the development of codes and optimizations for the Frontier computing system in collaboration with AMD. The LQCD group has met, in 2023, the project requirement of 50x improvement on benchmarks, that serve as progress markers for the ECP, moving from the OLCF Titan to the new Frontier system. An estimated three times more data are expected to be generated on these Exascale systems compared to the current generation of systems (Summit, Perlmutter). For LQCD projects in 2025, this could result in approximately 10 to 15 PB/year of data generated on LCFs and transferred back to the lab, with most of the data being utilized locally.

6.2.2.4 Process of Science

As previously detailed, the workflow for JLab LQCD computing involves around 30 allocated projects, with a collective allocation of 2 PB of disk storage and 15 PB of tape storage. These projects typically generate data in the range of hundreds of terabytes during their yearly allocation period, with a portion of these data being transferred off-site.

Planning for the next two to five years is aligned with the availability of resources from the LCFs. However, the growth in the capabilities of LCF systems occurs in intermittent bursts. Specifically, when large systems are deployed, there is a surge in data generation, which stabilizes as these systems remain in operation for more than five years. In contrast, the expansion of JLab facilities happens more gradually. The Intel Knights Landing facilities at JLab, nearing the end of operational life, increased the computational power necessary to analyze the existing 15 PB of data. The deployment of the JLab NVIDIA and AMD GPU systems, and soon the Intel Sapphire Rapids system, have enhanced the capabilities, and consequently, the data analysis rates to process the new generation of data produced on the LCF-s.

The USQCD collaboration has maintained records of the effective flops (floating-point operations per second) for LQCD applications over the past years. It has been observed that the computational power delivered by LQCD facility projects is on par with the flops acquired from LCFs. The deployment of the Exascale computing systems disrupts this trend. LQCD applications currently receive approximately 15% of the available computational cycles from LCFs across the nation, a fraction comparable to USQCD resources. However, the DOE's investments in Exascale systems have accelerated compared to historical levels. Appearing now are Exascale computing nodes that are commercially available. Consequently, with the incremental growth of local computing facilities, a closer alignment with the proportion of Exascale computing resources available for LQCD is expected in the long-term, exceeding five years.

In light of this, for planning purposes in the two- to five-year timeframe, it is envisaged that data transfers to JLab LQCD will increase and scale with the size of these new LCF systems. JLab will continue to serve as the repository for long-term storage. A portion of the analysis work will be conducted at the LCFs. However, the final stage of the analysis workflow is ideally suited for execution on JLab's local systems, effectively mitigating the disparity in LCF to local computing capability.

Currently long-term planning for the LQCD computing facilities is through FY29. The deployment of computational systems at JLab, Fermilab, and BNL will follow a two-year, staggered, time cycle of deployments.

6.2.2.5 Remote Science Activities

For the LQCD computing workflow, the first components, namely configurations of gluons fields, are generated on the LCFs. Those data are transferred to the local computing facilities and analyzed there. We anticipate this workflow to continue.

Over the next five years, computing allocations of a few million node-hours per year for projects will be available for LQCD projects on Frontier at OLCF. It is anticipated that similar amounts will be available in the future for Aurora at ALCF. With these computing allocations, a few PB of data will be generated for each project. These data will be transferred to JLab, BNL and Fermilab for further analysis. In total, about 10 PB is anticipated for transfer over that period.

6.2.2.6 Software Requirements

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.

6.2.2.7 Network and Data Architecture Requirements

JLab Network and Data Architecture Requirements can be found in [Section 6.1.2.7].

6.2.2.8 IRI Readiness

JLab IRI Readiness can be found in [Section 6.1.2.8].

6.2.2.9 Use of Cloud Services

JLab Use of Cloud Services can be found in [Section 6.1.2.9].

6.2.2.10 Data-Related Resource Constraints

JLab Data-Related Resource Constraints can be found in [Section 6.1.2.10].

6.2.2.11 Data Mobility Endpoints

JLab Data Mobility Endpoints can be found in [Section 6.1.2.11].

6.2.2.12 Outstanding Issues

JLab Outstanding Issues can be found in [Section 6.1.2.12].

6.2.2.13 Facility Profile Contributors

JLab Representation

- Robert Edwards, JLab, edwards@jlab.org

ESCC Representation

- Brent Morris, JLab, bmorris@jlab.org

6.3 BNL: The Scientific Data and Computing Center (SDCC)

The SDCC at Brookhaven Lab began in 1997 when the RHIC, pronounced "Rick") and ATLAS Computing Facility was established. Then called Resource Access Control Facility, the full-service scientific computing facility has since supported some of the most notable physics experiments, including Broad RAnge Hadron Magnetic Spectrometers (BRAHMS), Pioneering High Energy Nuclear Interaction eXperiment (PHENIX), PHOBOS Collaboration, and STAR, by providing dedicated data processing, storage, and analysis resources for these diverse, expansive experiments with general computing capabilities and support for users.

Today, this history of providing useful, resilient, and large-scale computational science, data management, and analysis infrastructure has grown, and the SDCC has evolved to support additional facilities and experiments. These include the National Synchrotron Light Source II (NSLS-II) and Center for Functional Nanomaterials (CFN) at Brookhaven Lab, as well as other DOE SC user facilities; the ATLAS experiment at CERN's LHC in Europe; and Belle II at the Japanese High-Energy Accelerator Research Organization (KEK)and Deutsches Elektronen-Synchrotron (DESY) in Germany. SDCC also is planning for its role in future experiments with the sPHENIX, Deep Underground Neutrino Experiment, and EIC.

6.3.1 Discussion Summary

The following bullets were extracted from the case study, and in-person discussion. Several appear in Section 1 and Section 3.

- The RHIC collider at BNL is a world-class particle accelerator exploring the most fundamental forces and properties of matter and the early universe. RHIC accelerates beams of particles (e.g., the nuclei of heavy atoms such as gold) to nearly the speed of light, and smashes them together to recreate a state of matter thought to have existed immediately after the Big Bang. The RHIC facility has approximately 1,000 unique users.
- The data from the sPHENIX DAQ system are written to a Lustre file system and is immediately processed by the HTC farm. A copy of the data is also sent to the HPSS tape system for archiving at a sustained rate of 10 GB/sec. For STAR, data are sent by the DAQ system to HPSS, with the HTC farm retrieving data from the cache (FastOffline) or tape for first-pass processing.
- Currently, the SDCC is among the ten largest HPSS sites worldwide and hosts in its tape libraries 170 PB of RHIC data. About 90k CPU cores are available in the HTC farm for RHIC data processing and analysis, together with about 90 PB of Lustre disk storage. The amount of CPU is expected to double by 2025, while the data volume on tape will be over 500 PB.
- BNL NP experiments follow certain shared workflows during their operations. Primarily, these workflows revolve around two types of data: experimental data, which come directly from the detectors, and MC data, which represent simulations of both the detector's operations and the specific physics processes under study.
- Experiment and MC data workflows at RHIC demand substantial computational resources, with the former taking more than 75% of the total compute resources and the latter taking the rest. The MC workflow is generally executed across multiple sites. This includes not only those sites directly affiliated with the experiment but also independent ones such as those connected to the OSG and potentially even the DOE Leadership class HPC facilities (such as NERSC used by the STAR collaboration). The primary reason for this distribution is the computational intensity of MC workflows, which demand significant processing power but use relatively minimal data.
- RHIC experiment workflows are more data-centric, necessitating immense data handling and processing capabilities. Due to this, they are primarily executed at extensive high-throughput batch farms located within computing facilities explicitly dedicated to the experiment.
- For sPHENIX and STAR, compute resources are a limiting factor. The ability to tap nonlocal resources like OSG and other unaffiliated resources can provide additional capacity. The ability to utilize these resources, particularly for experiment workflows, will be limited by the ability to transfer data between the host data center and the remote resources.
- RHIC data are transferred through the BNL HTSN from the experiments by fiber optic cable to the SDCC for storage and further processing. The HTSN is the network fabric that ties the components of the experiments together. It provides direct, high throughput connectivity

between the RHIC experimental halls and the compute and storage resources at the SDCC. It also enables high bandwidth transfers between the SDCC and sites on the WAN.
Finally, it couples the compute and storage resources within the SDCC data center together at high bandwidth.

- At present, 52 DTNs are in operation at the BNL SDCC, with the majority being utilized by programs outside of NP.
- At the present time, the use of "cloud" resources at BNL to meet the computational needs of RHIC experiments is limited to specialized, low-intensity (minimal compute and storage requirements) applications. Product evaluation and services in the cloud (repository, collaborative tools, and groupware applications) are use cases that utilize cloud resources, and this trend will likely increase over time. While cloud offers certain advantages, the bulk of computing and storage resources will continue to be located on-premises at BNL. This approach presents significant cost benefits, especially when considering egress issues associated with cloud solutions.
- The RHIC experiments rely mostly on services located at BNL. As the SDCC is supporting users and collaborators from multiple organizations, it is increasingly deploying Federated Identity mechanisms for accessing Web-services and collaborative tools supported by the facility. The SDCC supports the three IRI common patterns:
  - Time-Sensitive: Data produced by RHIC are reconstructed as soon as they are produced for rapid feedback on the quality of the data taking. The result may inform the machine for performance tuning and adjustments.
  - Data Integration-Intensive: Data from a given year are usually reprocessed and filtered several times as software and calibration evolve. Data from simulations may be embedded with real data to assess detector performance and signal reconstruction efficiencies.
  - The Long-Term Campaign: Data from different years, taken under different conditions of the accelerator, may be produced or reproduced depending on the advancement of knowledge (calibrations, physics understanding, etc.).

6.3.2 SDCC Facility Profile

The RHIC and plans for its successor, the EIC, a future facility poised to further this investigative journey, are two collider facilities at BNL. Particle collisions created by these facilities are used to probe the interaction between fundamental forces and particles.

6.3.2.1 Science Background

Current NP experiments at colliders collect large quantities of data over the course of their multi-year lifespan, typically at a rate of over hundreds of PBs per year. The sPHENIX and STAR detectors are the two operating detectors at RHIC. The RCF at the SDCC at BNL hosts the storage and computing resources used by the RHIC experiments.

The byproducts of this NP research are the technologies developed and the expertise that is gained from developing the apparatus to create and observe the collisions and the computational tools and techniques to simulate collisions and collect, store, and analyze the data obtained from the experiments. The apparatus includes superconducting magnets, particle accelerators, particle detectors, and high-speed analog and digital circuits. On the computing side, techniques and technologies include the application of AI/ML techniques and algorithms, the use of data analysis and analytics, data-management systems, data analysis tools (like Jupyter), high-performance and high-throughput computing, and high-performance networking.

6.3.2.2 Collaborators

At RHIC, sPHENIX and STAR are actively engaged in data collection using their respective detectors. This modus operandi is anticipated to continue through 2025. Subsequently, the cessation of RHIC operations will pave the way for the construction of the anticipated EIC. In parallel, the ePIC Collaboration, part of the EIC experimental program, is diligently working on a detector, which is poised to be primed and ready for data collection once the EIC springs into action in the early 2030s. Work on a second detector at the EIC is also in progress by a yet unnamed collaboration.

Data Transfer - A permanent direct conduit between data source and destination that is designed for continuous delivery of data at high bandwidth.

Data Portal - An access point where data can be uploaded or downloaded at moderate bandwidth. Batch and interactive data transfers are supported. Various platforms and servers are employed to manage data, each coming with its distinct authentication method to ensure security and controlled access: S3 Servers allow access key access-based authentication as well as authentication through an identity provider; Globus endpoint uses OAuth2 for authentication while FTP/GridFTP leverages the OSG approach. Access through the S3 protocol, however, allows access to the data from anywhere (a change in the data access approach).

Collaboration	User/ Collaborator and Location	Do they store a primary or secondary copy of the data?	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)?
SPHENIX @ RHIC	sPHENIX Experimental Hall/BNL Upton, New York	Source (No permanent storage)	Data Transfer		Continuous	N/A	
	SDCC/BNL Upton, New York	Primary	Data Portal		Ad hoc	No	
	sPHENIX Collaborators/ Worldwide	N/A			Ad hoc	Yes ad-hoc upstream data portal	
STAR @ RHIC	STAR Experimental Hall/BNL Upton, New York	Source (No permanent storage)	Data Transfer		Continuous	N/A	
	SDCC/BNL Upton, New York	Primary	Data Portal		Ad hoc	No	
	STAR Collaborators/ Worldwide	Subset (< 10%)			Ad hoc	Yes ad-hoc upstream data portal	

Table 6.3.1: Collaborative Data Mobility

6.3.2.3 Instruments and Facilities

The RHIC is the currently operational particle collider at BNL. It is expected to operate through 2025. The EIC will replace RHIC at BNL and is expected to begin colliding particles sometime in the 2030s.

Collider/Accelerator	Years of Operation	Location	Experiment(s)	Particle beam or colliding particles
RHIC	2001 - ~2025	BNL New York, USA	sPHENIX STAR	proton - proton proton - heavy ion gold ion - gold ion
EIC	~2030 +	BNL New York, USA	ePIC	electron - ion

Table 6.3.2: Collaborative Data Mobility

The major "instruments" at BNL are the STAR and sPHENIX detectors, both at RHIC. For the EIC, the first detector is called ePIC, and the eventual second is as yet unnamed. The STAR detector, the last of the four originally installed at RHIC's inception, remains in operation, having undergone numerous incremental upgrades that have replaced many of its parts over the years. The sPHENIX detector is one of the first next-generation particle detectors as it is expected to generate hundreds of PBs of data per year when fully commissioned, roughly an order of magnitude more data than previous NP detectors. A leading edge "streaming" DAQ system is used to continuously read out data from the detector. The DAQ system sends the data to the computing facility at an average rate of 10 gigabytes per second to disk for reconstruction and to the HPSS tape system.

Collider/Accelerator	Years of Operation	Detector Form Factor	Major Subdetectors
RHIC	2001 - ~2025	BNL New York, USA	sPHENIX STAR
EIC	~2030 +	BNL New York, USA	ePIC

Table 6.3.3: RHIC and EIC Capabilities

The major "instruments" at BNL are the STAR and sPHENIX detectors, both at RHIC. For the EIC, the first detector is called ePIC, and the eventual second is as yet unnamed. The STAR detector, the last of the four originally installed at RHIC's inception, remains in operation, having undergone numerous incremental upgrades that have replaced many of its parts over the years. The sPHENIX detector is one of the first next-generation particle detectors as it is expected to generate hundreds of PBs of data per year when fully commissioned, roughly an order of magnitude more data than previous NP detectors. A leading edge "streaming" DAQ system is used to continuously read out data from the detector. The DAQ system sends the data to the computing facility at an average rate of 10 gigabytes per second to disk for reconstruction and to the HPSS[8] tape system.

Experiment	Years of Operation	Detector Form Factor	Major Subdetectors
SPHENIX	2023 - 2025	Toroidal	TPC silicon strip/pixel calorimeters
STAR	2001 - 2025	Toroidal	TPC silicon strip/pixel calorimeters
EPIC	2030s - TBD	Toroidal	TPC silicon strip/pixel calorimeters

Table 6.3.4: BNL Instrumentation

Each of these experiments is supported by their respective "host" computing facilities. They archive data from the detectors and provide a significant portion of the computational resources required by the experiments. The majority of these resources are in the form of an HTC farm, which is composed of thousands of rack-mounted servers located within the host computing center.

[8] https://hpss-collaboration.org

For sPHENIX, the data from the DAQ system are written to a Lustre file system and are immediately processed by the HTC farm. A copy of the data is also sent to the HPSS tape system for archiving at a sustained rate of 10 GB/sec. For STAR, data are sent by the DAQ system to HPSS, with the HTC farm retrieving data from the cache (FastOffline) or tape for first-pass processing.

Currently, the SDCC is among the ten largest HPSS sites worldwide and hosts in its tape libraries 170 PB of RHIC data. About 90k CPU cores are available in the HTC farm for RHIC data processing and analysis, together with about 90 PB of Lustre disk storage. The amount of CPU is expected to double by 2025, while the data volume on tape will be over 500 PB.

The SDCC also provides critical experiment support services, such as databases, authentication and authorization services, websites, software development infrastructure, and collaborative tools (groupware).

6.3.2.4 Generalized Process of Science

All nuclear physics experiments follow certain shared workflows during their operations. Primarily, these workflows revolve around two types of data: experimental data, which come directly from the detectors, and MC data, which represent simulations of both the detector's operations and the specific physics processes under study. These two pivotal workflows converge in the advanced stages of data analysis—often referred to as the end-stage physics analysis—where definitive physics results are derived from the experimental data.

For the experiment workflow, the process is roughly as follows:

1. Large volumes of data generated by the detector (raw data) at terabits per second under various experiment conditions are collected by the data acquisition system (DAQ) at the experiment hall.
2. The data are noise-reduced, zero suppressed, and possibly compressed by the DAQ and online computing systems, reducing the data bandwidth to ~ hundreds of gigabits per second.
3. The reduced "raw" data are sent to the host computing facility for permanent storage (HPSS).
4. The data may be sent to other computing facilities so that a second copy of the data is available.
5. At the computing facility, a first pass processing is made over all the raw data. This step reconstructs the events from the millions of channels of detector data. The reconstruction process itself consists of many steps
    a. Reconstruction of particle trajectories produced in collisions (events) from coarse two and three-dimensional space points generated by particles as they interact with the elements of the sub-detectors in the detector;
    b. Particle identities are assigned to detected particles based on reconstructed trajectories and energy deposition in various "calorimeters" in the detector.
    c. Initial collision points or interaction vertices are determined as secondary vertices from decays of particles produced from the initial collision.
    d. Event information is distilled to quantities needed for specific physical processes.
    e. The distilled event information, also known as DSTs from first-pass processing, is written for bulk online storage and archival storage. They are self-described formats and represent the persistent data model for the experiments.
    f. Other formats may be produced during this data processing phase (micro, pico, or nano DSTs, which we will refer to as xDST). When present, they are the primary source for physics harvesting.
6. Raw data may be reprocessed multiple times as reconstruction algorithms are improved, and the detector performance is better understood. DSTs may be converted to new sets of xDST as applied.

7. The xDSTs generated by reconstruction are then taken by individuals or groups of researchers for further analysis and distillation.
8. The product of the Experiment workflow is combined with the output of the MC workflow in the end-stage physics analysis discussed below.

For the MC data, the workflow can be described as follows:

1. Billions of simulated particle collisions are created with "physics generators" such as Pythia[9].
2. The products of these collisions are "propagated" through a digital model of the detector, resulting in a simulation of an event as it would be recorded by the detector. Geant4[10] is one of the toolkits used to build these models.
3. Specific detector simulators merge the raw simulation output and convert it into signals similar to those coming from the experimental devices and DAQ.
4. The output of this propagation step is fed into the first pass reconstruction pipeline outlined in the Experiment workflow, yielding MC DSTs. Provenance information and simulation conditions are also propagated in this case.
5. These MC DSTs are then consumed by individuals and researchers for further analysis and distillation
6. The product of the MC workflow is combined with the output of the analysis workflow in the end-stage physics analysis discussed below.

Both the Experiment and MC data workflows demand substantial computational resources, with the former taking more than 75% of the total compute resources and the latter taking the rest. The MC workflow is generally executed across multiple sites. This includes not only those sites directly affiliated with the experiment but also independent ones such as those connected to the OSG and potentially even the DOE Leadership class HPC facilities (such as NERSC used by the STAR collaboration). The primary reason for this distribution is the computational intensity of MC workflows, which demand significant processing power but use relatively minimal data.

On the other hand, the Experiment workflows are more data-centric, necessitating immense data handling and processing capabilities. Due to this, they are primarily executed at extensive high-throughput batch farms located within computing facilities explicitly dedicated to the experiment.

End-stage physics analysis consists of the following workflow

1. MC and experiment DSTs are analyzed and compared to quantify the performance of the detector and reconstruction algorithms.
2. Algorithms and detector calibrations are adjusted to improve performance and reflect a better understanding of experimental conditions
3. Experiment and MC DSTs are further filtered and reduced to isolate physics processes of interest
4. Steps 1 through 3 are repeated as the analysis turns to the extraction of physics results from the data.
5. End stages of analysis typically involve interactive analysis of data utilizing interactive analysis and data visualization systems like ROOT[11] and Jupyter[12] Notebook.
6. Results are then presented in papers that are submitted for publication.

[9] https://pythia.org
[10] https://geant4.org
[11] https://root.cern
[12] https://jupyter.org

End-stage physics typically involves fewer computing resources and requires access to online data stores due to the interactive nature of the analysis process.

The integrity of the entire analysis chain is paramount. It must undergo rigorous checks and validations at multiple levels and stages to ensure the accuracy of the final results. In addition, comprehensive documentation of this chain is imperative. This documentation serves not only as a testament to reproducibility but also supports independent verification, both of which are foundational to the scientific method. To exemplify this commitment, the PHENIX experiment, which collected data at RHIC until 2016, has harnessed the Reana[13] toolkit—developed by CERN and hosted at the SDCC—to preserve its knowledge. Meanwhile, the STAR collaboration is actively refining its approach to data and knowledge preservation.

For STAR and sPHENIX, steps 1 (data collection) through 5 (first pass reconstruction) of the experiment workflows will be ongoing as the experiments will be actively collecting data. In 2025, both experiments aim at acquiring massive datasets hence, after the RHIC shutdown, the experiment will finalize their data production workflow (this will extend beyond one year) and will move towards those workflows mostly related to the analysis of DSTs data and end-stage analysis. Reprocessing of raw data may happen as warranted. The MC workflows will continue, although likely at a reduced volume, as will end-stage analysis. Analysis of data from the RHIC experiments is expected to taper off slowly beyond at least five years from the shutdown of RHIC.

The EIC experimental program and EIC projects stand apart from traditional NP workflows. Until the EIC is operational, they won't handle actual experimental data; instead, their main emphasis is on simulation.

6.3.2.5 Remote Science Activities

The NP collaborations at BNL are international; remote access to computing resources, from the perspective of the researchers, is paramount, as is access to data. Collaborative tools, e.g., communication tools, code, and document repositories, are needed to allow researchers to work together effectively.

For all NP experiments, utilization of unaffiliated computing resources not at the host data centers is a given. OSG and the supercomputers at DOE facilities (mainly NERSC for the STAR experiment) are commonly used for simulations and are two examples of remote, unaffiliated computing resources.

6.3.2.6 Software Infrastructure

Data management (distribution, cataloging, retention policies, etc.) is crucial for large NP experiments. STAR has developed its own data-management system. With STAR, several data-transfer mechanisms are used, including Globus and third-party transfers via Condor-G[14], as well as simpler methods like sftp/scp orchestrated by a home-developed data-management system. sPHENIX, as a new NP experiment, has adopted Rucio[15], an open-source data management system created by the ATLAS[16] collaboration at CERN. This system has proven to be reliable and scalable and has gained widespread adoption in the NP and High Energy Physics (HEP) communities. The Rucio service for sPHENIX is operated at BNL. With Rucio, data transfers are scheduled and organized with FTS[17]. FTS can use different protocols (GridFTP, HTTPS, XRoot[18], etc.) to transfer the data. Globus is also used for transferring data to/from BNL for various use cases.

Each experiment develops its own software frameworks for data transformation from raw to user data products. These transformations require detailed knowledge of the composition and performance of their detectors, as well as an understanding of the physical processes that are being investigated. Future experiments are strategizing to maximize the utilization of shared software frameworks and toolkits. These common elements will streamline the development of specialized code necessary for each phase of the data analysis process.

[13] https://reana.io
[14] https://htcondor.org
[15] https://rucio.cern.ch
[16] https://atlas.cern
[17] https://fts.web.cern.ch/fts
[18] https://xrootd.slac.stanford.edu

6.3.2.7 Network and Data Architecture

BNL Network Architecture

RHIC data are transferred through the BNL HTSN from the experiments by fiber optic cable to the SDCC for storage and further processing.

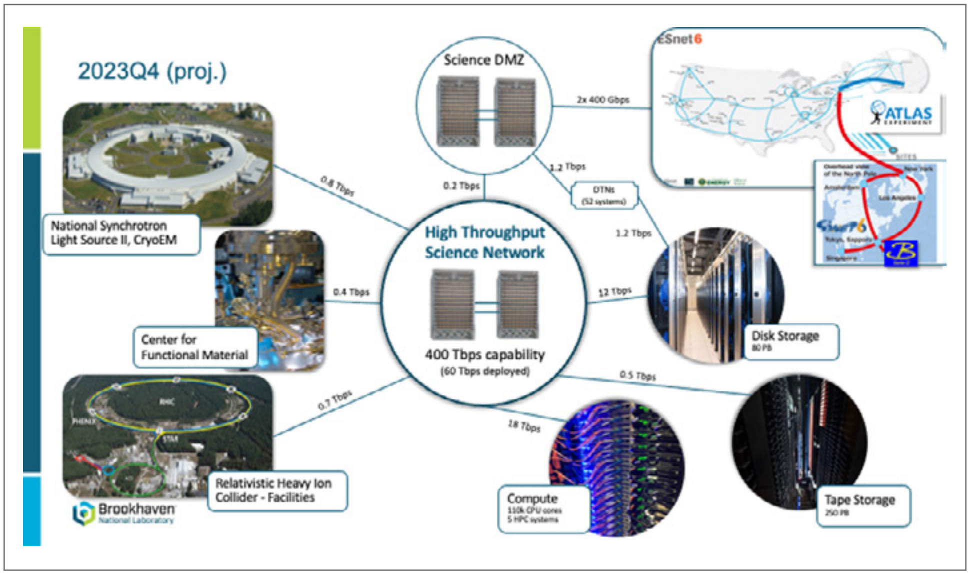


Figure 6.3.1: BNL Internal Connectivity

The HTSN is the network fabric that ties the components of the experiments together. It provides direct, high throughput connectivity between the RHIC experimental halls and the compute and storage resources at the SDCC. It also enables high bandwidth transfers between the SDCC and sites on the WAN. Finally, it couples the compute and storage resources within the SDCC data center together at high bandwidth.

The HTSN has five key components:

1. *Network Perimeter*
   a. Currently consists of two redundant Nokia 7750 SR-2s with building diversity.
   b. Two circuits that peer with ESnet, one 200 Gbps, and the other 100 Gbps. These circuits are utilized by all scientific and administrative communities at BNL. All traffic to and from BNL flows through either of these circuits.
   c. The BNL Network Perimeter transfers, on average, 20 PB of data monthly, with spikes up to ~25 PB.
   d. In December 2023, connectivity to ESnet will be upgraded to two 800 Gbps circuits with two routers, each 2x400 GE.
2. *Science DMZ*
   a. Supports open, high-speed WAN/Internet) access for all scientific collaborations throughout the BNL campus.
3. *Science Core*
   a. A Tbps network dedicated to transporting scientific data. It connects instruments in the BNL campus, outside of the data center, to the SDCC data center and resources in the data center together. These include

    b. STAR, sPHENIX, Collider Accelerator Division, CFN Electron Microscopy, NSLS-II, and CryoEM

    c. SDCC-based HPC Clusters and HTC farms

    d. SDCC-based high-performance and high-capacity storage and archive storage.

    e. Intelligence and routing policies are supported by the Science core routers, enabling selective connectivity among endpoints and resources.

4. *Spine*

    a. A Tbps network Spine that interconnects all Leaf switches. Leaf switches can consist of Top of Rack (ToR) or chassis-based switches that connect compute, storage, or general infrastructure service servers.

    b. External Border Gateway Protocol (eBGP) is utilized throughout the HTSN.

5. *Storage Core*

    a. A redundant terabit per second switching block that aggregates high-performance storage services.

The BNL HTSN has been designed to accommodate the needs of the entire BNL campus, including the NP experiments, for the foreseeable future. As bandwidth requirements increase, link speeds and link count can be increased, but the basic architecture remains intact.

At present, 52 DTNs are in operation, with the majority being utilized by programs outside of NP.

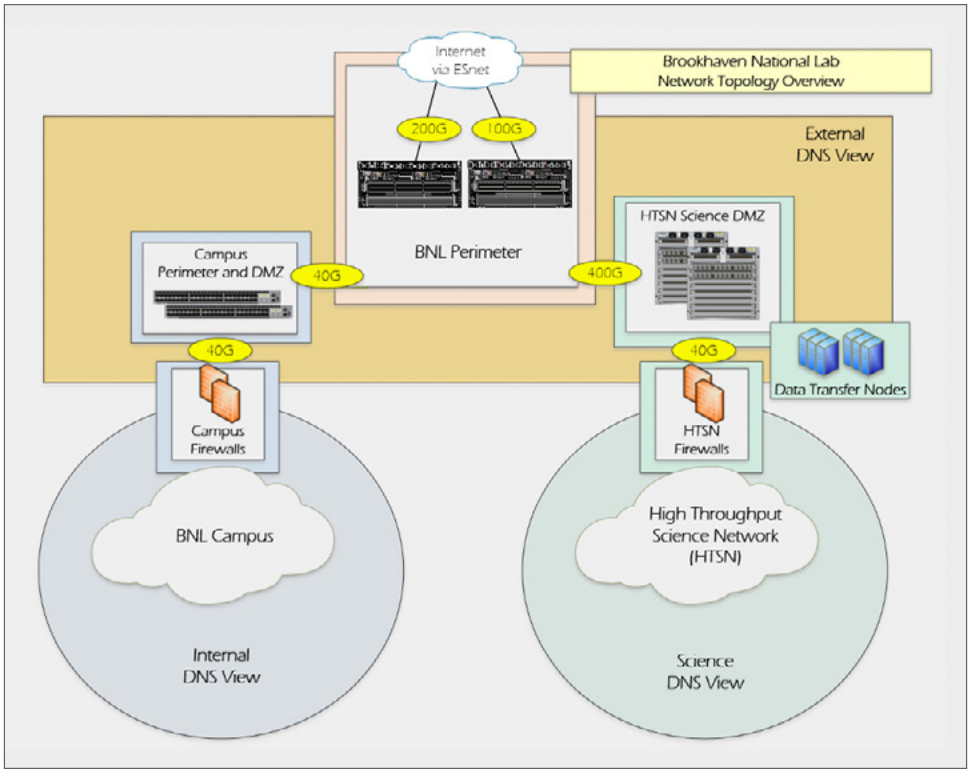


Figure 6.3.2: High Level Overview of the BNL Network Perimeter and Domain Name System (DNS) Architecture

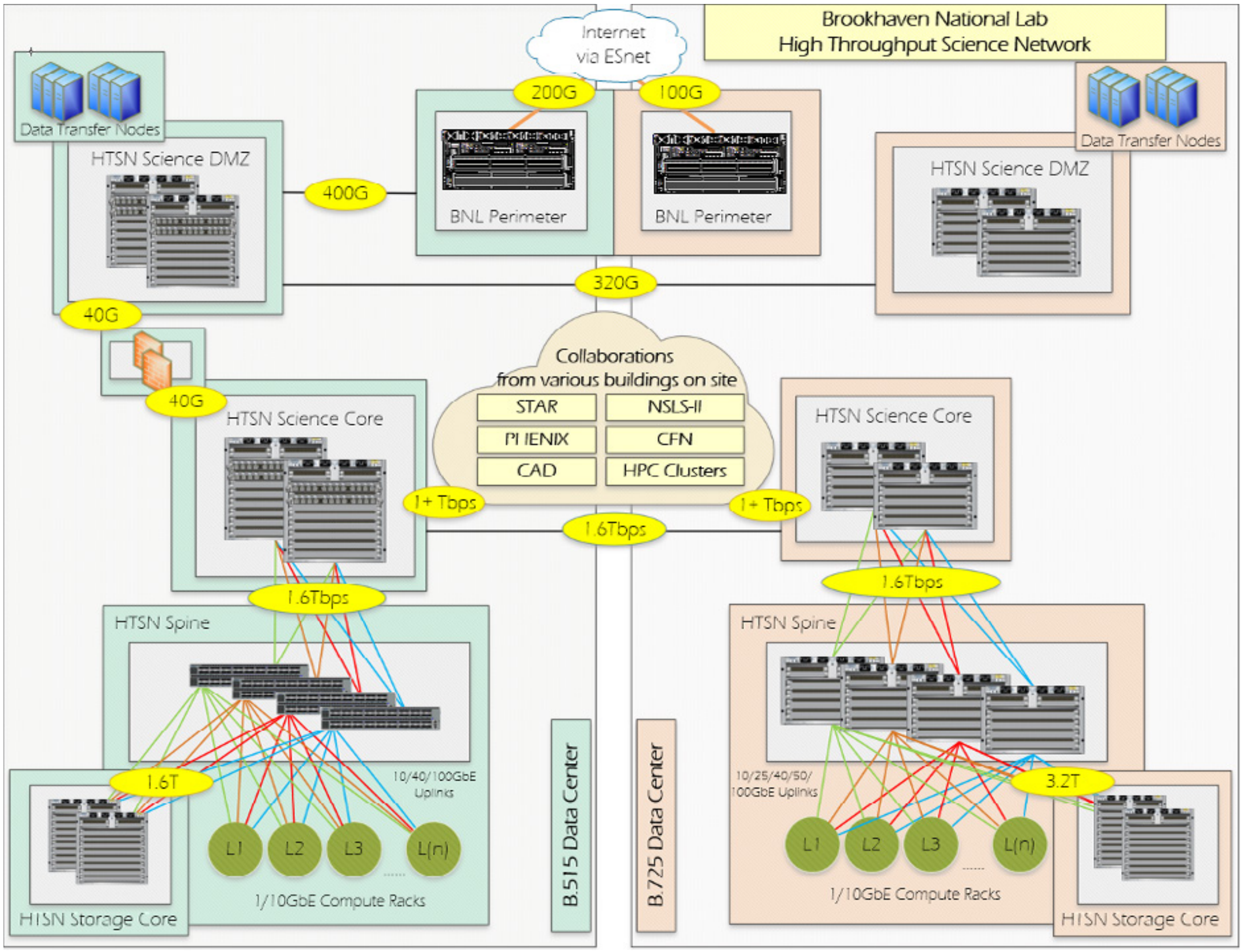


Figure 6.3.3: High Level Overview of BNL HTSN in FY23

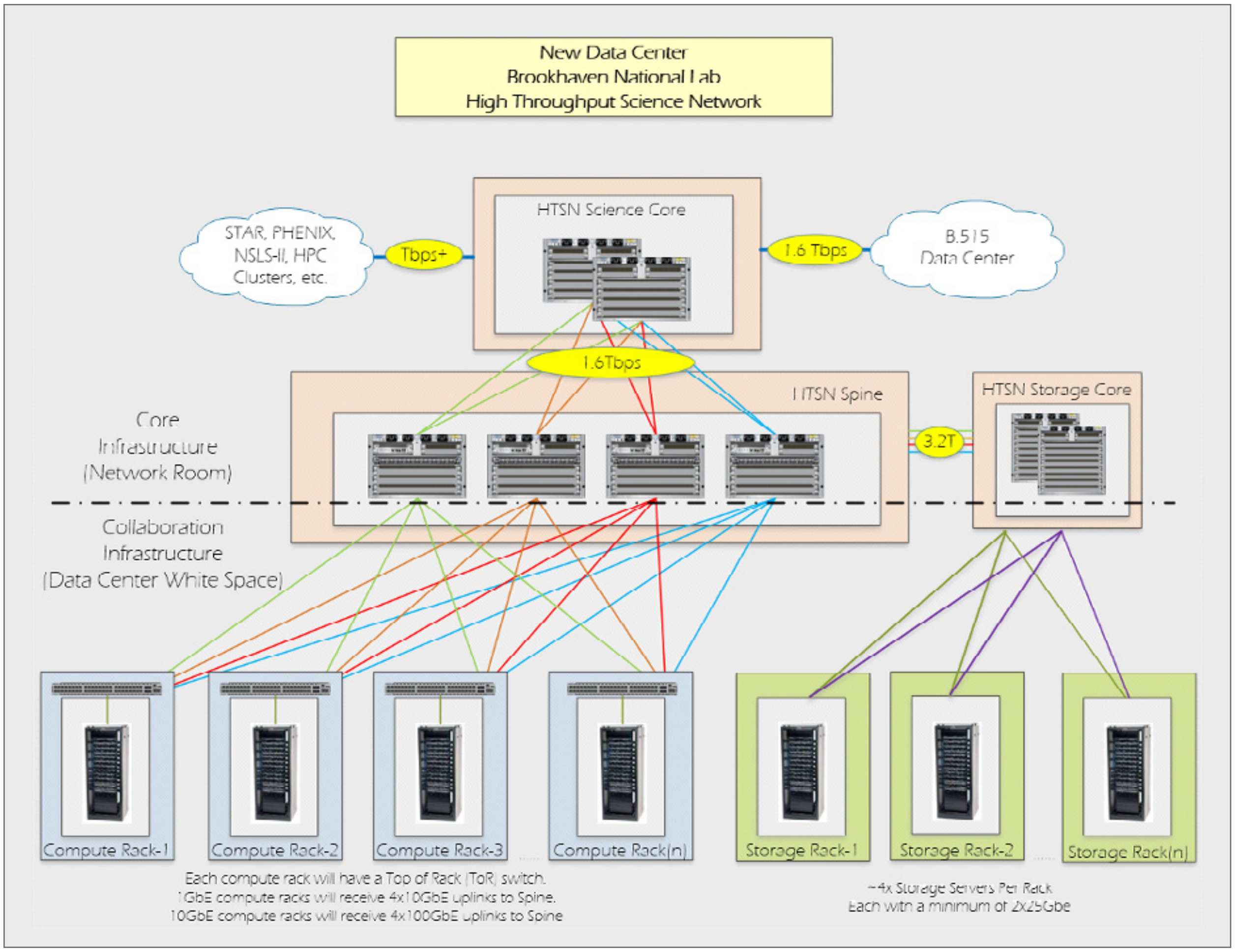

New Data Center
Brookhaven National Lab
High Throughput Science Network
HTSN Science Core
STAR, PHENIX, NSLS-II, HPC Clusters, etc.
Tbps+
1.6 Tbps
B.515 Data Center
1.6Tbps
HTSN Spine
HTSN Storage Core
3.2T
Core Infrastructure (Network Room)
Collaboration Infrastructure (Data Center White Space)
Compute Rack-1
Compute Rack-2
Compute Rack-3
Compute Rack(n)
Storage Rack-1
Storage Rack-2
Storage Rack(n)
Each compute rack will have a Top of Rack (ToR) switch.
1GbE compute racks will receive 4x10GbE uplinks to Spine.
10GbE compute racks will receive 4x100GbE uplinks to Spine
~4x Storage Servers Per Rack
Each with a minimum of 2x25Gbe

Figure 6.3.4: High Level Overview of BNL HTSN and Demarcation Points Between Core and Collaboration Infrastructure

6.3.2.8 IRI Readiness

The RHIC experiments rely mostly on services located at BNL. As the SDCC is supporting users and collaborators from multiple organizations, it is increasingly deploying Federated Identity mechanisms for accessing Web-services and collaborative tools supported by the facility.

The SDCC supports the three IRI common patterns:

- ***Time-Sensitive:*** Data produced by RHIC are reconstructed as soon as they are produced for rapid feedback on the quality of the data taking. The result may inform the machine for performance tuning and adjustments.
- ***Data Integration-Intensive:*** Data from a given year are usually reprocessed and filtered several times as software and calibration evolve. Data from simulations may be embedded with real data to assess detector performance and signal reconstruction efficiencies.
- ***The Long-Term Campaign:*** Data from different years, taken under different conditions of the accelerator, may be produced or reproduced depending on the advancement of knowledge (calibrations, physics understanding, etc.).

6.3.2.9 Cloud Services

At the present time, the use of "cloud" resources is limited to specialized, low-intensity (minimal compute and storage requirements) applications. Product evaluation and services in the cloud (repository, collaborative tools, and groupware applications) are use cases that utilize cloud resources, and this trend will likely increase over time. While cloud offers certain advantages, the bulk of computing and storage resources will continue to be located on-premises at BNL. This approach presents significant cost benefits, especially when considering egress issues associated with cloud solutions.

6.3.2.10 Data-Related Resource Constraints

For sPHENIX and STAR, compute resources will always be a limiting factor. The ability to tap nonlocal resources like OSG and other unaffiliated resources can provide additional capacity. The ability to utilize these resources, particularly for experiment workflows, will be limited by the ability to transfer data between the host data center and the remote resources.

For ePIC, the ability to process data remotely is an integral part of their proposed computing model. Access to storage resources and sufficient network bandwidth to move or access data to and from remote sites is a prerequisite.

6.3.2.11 Data Mobility Endpoints

The NP communities supported by BNL utilize an array of methods for sharing research data. The "compute model" adopted by the community is the primary determinant of how research data is shared. For those experiments with data center-centric compute models (sPHENIX and STAR), the vast majority of data sharing occurs within the host data center. In this case, network file systems (Lustre, GPFS, NFS) are the dominant mode of data sharing through dCache and XRootD have been used for final stage analysis data access. For these experiments, data sharing with remote sites is predominantly ad hoc, through sftp gateways or Globus endpoints and similar data portals. These transfers are mostly driven by individual researchers or small research groups at the tail end of the data analysis chain.

The ePIC Collaboration is planning to adopt a distributed computing model, which means that large, automated data flows are expected to be used for sharing data among the collaborating computing centers, including the two host data centers. Rucio is the prime candidate for this task. Other mechanisms are likely to be used for ad hoc sharing and access of data, but these mechanisms have yet to be enumerated by ePIC.

6.3.2.12 Outstanding Issues

Nothing to report.

6.3.2.13 Facility Profile Contributors

BNL Representation

- Eric Lancon, BNL, elancon@bnl.gov
- Jerome Lauret, BNL, jlauret@bnl.gov
- Shigeki Misawa, BNL, misawa@bnl.gov

ESCC Representation

- Vincent Bonafede, BNL, bonafede@bnl.gov
- Mark Lukasczyk, BNL, mlukasczyk@bnl.gov

6.4 sPHENIX at the RHIC

sPHENIX is a radical makeover of the PHENIX experiment, one of the original detectors designed to collect data at Brookhaven Lab's RHIC. It includes many new components that significantly enhance scientists' ability to learn about quark-gluon plasma (QGP), an exotic form of nuclear matter created in RHIC's energetic particle smashups.

The sPHENIX detector is about the size of a two-story house and weighs 1,000 tons. Like a giant, 3D digital camera, the detector will capture snapshots of 15,000 particle collisions per second, more than three times faster than PHENIX. With these capabilities, sPHENIX will take advantage of the many accelerator improvements that have been made to increase collision rates at RHIC.

6.4.1 Discussion Summary

The following bullets were extracted from the case study, and in-person discussion. Several appear in Section 1 and Section 3.

- The sPHENIX and STAR detectors are the two operating detectors at RHIC. The RCF at the SDCC at BNL hosts the storage and computing resources used by the RHIC experiments. These detectors collect large quantities of data over the course of their multi-year lifespan, typically at a rate of over hundreds of PBs per year.
- sPHENIX is the first major upgrade to a nuclear physics heavy-ion experiment in the US in two decades, and along with STAR will be the final experiments taking data at the RHIC)before the construction of the EIC. The sPHENIX Collaboration consists of 81 member institutions and about 400 collaborators from 14 countries.
- The operation and data analysis of the sPHENIX experiment is centered at the BNL. BNL's SDCC currently provides the vast majority of data storage and computing resources for the experiment and the collaboration.
- No large-scale export of data or the routine use of grid services is foreseen for sPHENIX, although the possibility of hosting additional 2nd-tier copies in particular of DSTs in other regions has been discussed.
- The primary dataset consists of the raw data files written by the sPHENIX data acquisition (DAQ) system. A data acquisition "run" consists of about one hour of data taking. Each run dataset consists of dozens of files, each written in parallel by the DAQ process that reads a given detector component. A full run with all components produced file sets of 52 files. This number is expected to grow to more than 60 in order to increase the available bandwidth for some components. The output files automatically roll over when reaching an adjustable size limit, typically 20GB, resulting in several hundred files being written during a run.
- The sPHENIX raw data files are processed into summary files, traditionally still called DSTs. The DSTs represent a state of the data reconstruction after a number of time-consuming steps have been performed. DSTs are expected to be about 50% of the size of the original raw data, conservatively accounting for the information omitted from the DSTs in order for them to support all of the envisioned analyses. Further generations of output, called micro-DSTs (or nano-, pico-...) that apply a number of by-then standard cuts and filters to the DST data to further reduce the size. Past experience indicates that those files will have a size in the order of 10% of the original raw data.
- The RHIC aims to provide 28 "cryo weeks" per year (operation of the superconducting magnets). The plan is to have a three-year running period (2023-2025) with the following breakdown for the three years of sPHENIX operation:
  - Run 1 (2023): 70 PB (expected), 11.5 PB (actual)
  - Run 2 (2024): 50 PB (expected), 78 PB (potential)
  - Run 3 (2025): 180 PB
- sPHENIX only collected 11.5 PB of data instead of the expected 70 PB in FY 2023 due to reduced accelerator availability.

- sPHENIX datasets are kept disk-resident at the SDCC. Therefore, the vast majority of dataset processing will take place at the SDCC itself. It is not economical to transfer the DSTs that make up a dataset off-site for remote processing because the CPU usage/disk space ratio is not large enough.
- sPHENIX data processing is to a large extent centered around the BNL SDCC. The primary required ESnet services are therefore interactive logins and small-scale data transfers. The amount of data transferred out of the SDCC is not expected to exceed 100 TB/month.

6.4.2 sPHENIX Experimental Case Study

The sPHENIX experiment began taking data in 2023 with a 12-week commissioning period, followed by an envisioned 8-week physics/production period.

Due to a fault with the accelerator, beam operations had to be prematurely terminated for the year on August 1, 2023, in week 10 of the sPHENIX commissioning. sPHENIX never entered the production period in 2023. Instead of the expected 70 PB of data, sPHENIX only collected 11.5 PB of data.

sPHENIX data processing is to a large extent centered around the BNL SDCC. The primary required ESnet services are therefore interactive logins and small-scale data transfers. The amount of data transferred out of the SDCC is not expected to exceed 100 TB/month.

6.4.2.1 Science Background

sPHENIX will be the final experiment taking data at the RHIC before the construction of the EIC. The experiment is capable of measuring jets, jet correlations and upsilons to determine the temperature dependence of transport coefficients of the QGP. Fig. 6.4.1 shows an artist's view of the sPHENIX experiment.

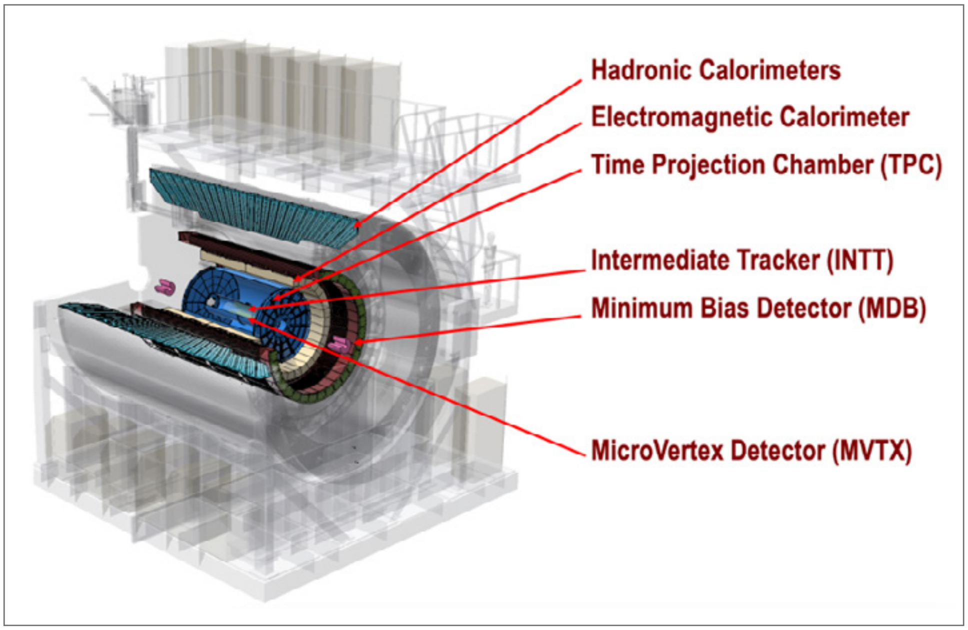

Hadronic Calorimeters
Electromagnetic Calorimeter
Time Projection Chamber (TPC)
Intermediate Tracker (INTT)
Minimum Bias Detector (MDB)
MicroVertex Detector (MVTX)

Figure 6.4.1: An Artist's view of the sPHENIX experiment

6.4.2.2 Collaborators

At the time of this writing, the sPHENIX Collaboration consists of 81 member institutions and about 400 collaborators from 14 countries. The collaboration has been growing steadily and will accept new institutions in the future. Fig. 6.4.2 shows a breakdown of sPHENIX institutions by Country.

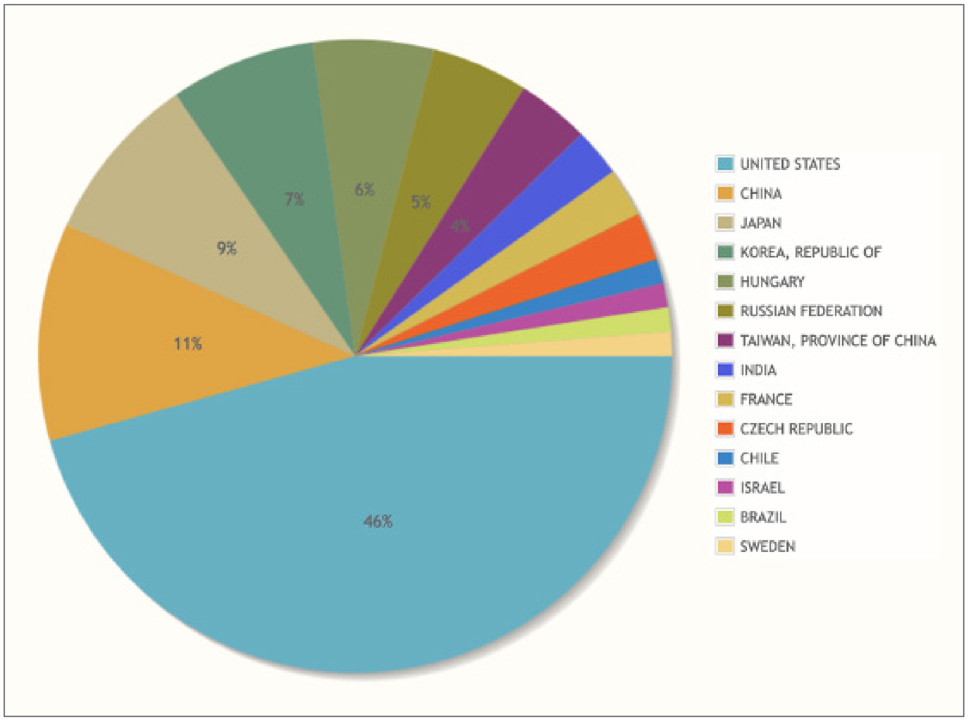


Figure 6.4.2: Breakdown of sPHENIX institutions by Country

The operation and data analysis of the sPHENIX experiment is centered at the BNL. BNL's SDCC currently provides the vast majority of data storage and computing resources for the experiment and the collaboration. This includes

- permanent storage of the sPHENIX raw data in the HPSS tape storage system;
- temporary storage of raw data on disk for processing and during code development;
- storage of reconstructed data summary files (historically called DSTs) on tape and disk
- interactive logins for collaboration members
- batch processing support for reconstruction and analysis.
- General support such as email, mailing lists, messaging, document storage, and similar services.

At this time, no large-scale export of data or the routine use of grid services is foreseen, although the possibility of hosting additional 2nd-tier copies in particular of DSTs in other regions has been discussed.

6.4.2.3 Use of Instruments and Facilities

The RHIC aims to provide 28 "cryo weeks" per year (operation of the superconducting magnets). The plan is to have a three-year running period (2023-2025). The first year has now ended.

The original "Beam Use Proposal" foresaw the following breakdown for the three years of sPHENIX operation, with the "Run 1" (2023) later reduced to eight weeks of physics running:

Run/year	Collision System	weeks	events	estimated dataset size
1 - 2023	Au+Au 200GeV/n	13 (later reduced to 8)	43 billion	70 PB
2 - 2024	p+p, p+A	21 weeks	70 billion	78 PB
3 - 2025	Au+Au	24.5 weeks	107 billion	180 PB

Table 6.4.1: the original Run plan for the sPHENIX experiment

The plan for the first year had since been amended to include a 12-week commissioning period (which does not preclude acquiring some physics-grade data), followed by a reduced 8-week “physics” period to acquire the main 2023 physics dataset.

Unfortunately, on August 1, in week 10 of the 12-week commissioning schedule, RHIC developed a fault in a valve box that led to the premature end of the beam operations for the year. sPHENIX never entered the envisioned physics/production running period. Therefore, the entire 2023 dataset fell far short of expectations, consisting of 11.5 PB total only.

As a result, the plan for 2024 has been revised to include a make-up period for the Au+Au component from Run 1. This will likely lead to the elimination of the planned p+A running in its entirety. The exact schedule is still in flux and is driven by the scope and progress of the required accelerator repairs.

At the same time, the Collider-Accelerator Department revised the earlier projected luminosity figures downwards due to a number of new technical obstacles, such as the difficulties running the accelerator in the increasingly high temperatures during the summer months. We maintain the expected numbers of 50 PB in 2024 and 180 PB in 2025 but consider them a best-case scenario now.

Datasets

The primary dataset consists of the raw data files written by the sPHENIX data acquisition (DAQ) system. A data acquisition “run” (not to be confused with the yearly “Run” of RHIC in 2023, 2024…) consists of about one hour of data taking. That time is determined by the size of the resulting dataset that still needs to fit into the typically available disk space available on modern machines, as well as detector-performance considerations, chiefly the stability of detector gains that do not typically change significantly over the time span of one hour.

Each run dataset consists of dozens of files, each written in parallel by the DAQ process that reads a given detector component. A full run with all components produced file sets of 52 files from the same number of DAQ processes in 2023. This number is expected to grow to more than 60 in order to increase the available bandwidth for some components. The output files automatically roll over when reaching an adjustable size limit, typically 20GB, resulting in several hundred files being written during a run.

The raw data files are then processed into summary files, traditionally still called DSTs. The DSTs represent a state of the data reconstruction after a number of time-consuming steps have been performed, such as the application of calibration constants, clustering of adjacent detector elements into “hits”, and an early version of the tracking output. Another important aspect of DSTs is that they will already contain the combined information of the aforementioned individual streams, so that the analysis of DSTs compared to the actual raw data is dramatically simplified.

The first generations of DSTs still contain enough information that the tracking can be re-run from the data contained on the DSTs. This will aid the improvement and performance tuning of the tracking algorithms while at the same time already providing a reproducible tracking performance benchmark for physics analyses.

We foresee a number of “primary” DST generations to be produced over time, as we gain a better understanding of the data. The first generations of DSTs are expected to be about 50% of the size of the original raw data, due to the fact that we need to be conservative with the information omitted from the DSTs in order for them to support all of the envisioned analyses. Over time, we will learn what additional information can be dropped from the DSTs and the eventual size is expected to shrink.

Traditionally, we have produced further generations of output, called micro-DSTs (or nano-, pico-...) that apply a number of by-then standard cuts and filters to the DST data to further reduce the size. Past experience indicates that those files will have a size in the order of 10% of the original raw data. At the time of this writing, we have not arrived at a stage where such micro-DST output is generated.

6.4.2.4 Process of Science

The DSTs are later vetted for correctness and grouped into certain datasets. Those datasets are usually (except for code development and similar purposes) analyzed in their entirety by individual scientists or analysis groups. The fact that the composition of those datasets is static over time allows us to more easily compare results from different analyses, because the input to each analysis is the same, and predictable.

The analysis of those datasets is the main way that the sPHENIX data are analyzed for physics results.

The goal is to keep those datasets disk-resident at the SDCC. Therefore, the vast majority of this processing will take place at the SDCC itself. It is not economical to transfer the DSTs that make up a dataset off-site for remote processing because the CPU usage/disk space ratio is not large enough. sPHENIX has about 60,000 processing slots (roughly equivalent to processor cores) at its disposal.

The earliest output that is a viable candidate for a transfer to a local (e.g., University) cluster is the output of those dataset analyses. It is still expected that this is not a very common occurrence.

In FY2024 we expect to perform most of the reconstruction/DST production for the 2023 data, while already acquiring the new 2024 data, and getting a head-start with the 2024 data reconstruction/DST production. If sPHENIX ends taking data after 2025 as currently planned, this will lead to a five-year-long process of concurrent reconstruction of new data, and analysis of earlier data.

Taking the experience of the PHENIX experiment as guide, where data are still actively analyzed 22 years after the first data-taking run, the analysis of sPHENIX data will continue well into 2040.

6.4.2.5 Remote Science Activities

Nothing to report.

6.4.2.6 Software Requirements

Most of the ingredients for our data processing needs are in place. The vast majority of the components are open-source, with minor exceptions such as licensed virtual desktop management software (NX), and a current third-party service contract for the (open-source) Lustre file system management.

Some of the ingredients for the data analysis are

- The root package (open source)
- databases (Postgres and others) (open source)
- User job management (condor) (open source)
- Meta-management packages for jobs (e.g., PanDA) (open source)

The root package provides the common analysis framework that standardizes common tasks such as data storage/IO, histogramming, and data visualization.

The databases are used to select runs by various criteria, and to retrieve per-run information such as calibration constants, and file locations, etc.

Condor is the de-facto job submission standard technology in widespread use.

A Meta-package is usually needed to manage a huge number of individual user jobs.

6.4.2.7 Network and Data Architecture Requirements

BNL Network and Data Architecture Requirements can be found in [Section 6.3.2.7].

6.4.2.8 IRI Readiness

BNL IRI Readiness can be found in [Section 6.3.2.8].

6.4.2.9 Use of Cloud Services

BNL Use of Cloud Services can be found in [Section 6.3.2.9].

6.4.2.10 Data-Related Resource Constraints

BNL Data-Related Resource Constraints can be found in [Section 6.3.2.10].

6.4.2.11 Data Mobility Endpoints

Based on past experience, it is estimated that a rather small amount of data will be transferred by sPHENIX out of SDCC; approximately 50 TB/month.

6.4.2.12 Outstanding Issues

BNL Outstanding Issues can be found in [Section 6.3.2.12].

6.4.2.13 Facility Profile Contributors

sPHENIX Representation

- Christopher Pinkenburg, BNL, pinkenburg@bnl.gov
- Martin Purschke, BNL, purschke@bnl.gov

ESCC Representation

- Vincent Bonafede, BNL, bonafede@bnl.gov
- Mark Lukasczyk, BNL, mlukasczyk@bnl.gov

6.5 FRIB

FRIB is a scientific user facility for the DOE-SC, supporting the mission of the DOE-SC Office of NP. User facility operation is supported by the DOE-SC NP as one of the DOE-SC user facilities.

6.5.1 Discussion Summary

The following bullets were extracted from the case study, and in-person discussion. Several appear in Section 1 and Section 3.

- FRIB is a DOE-SC scientific user facility serving users organized in the FRIB Users Organization (FRIBUO). The FRIBUO has over 1,800 members, representing 124 US colleges and universities, 13 national laboratories, and 52 countries. FRIB was designed to be the world's most powerful rare isotope research facility and commenced user operation in the latter half of FY 2022.
- The average FRIB data set sizes ranged from a few GB to ∼70 TB with an average size just over 4 TB. For the first year of FRIB operation the data sizes ranged from a few GB to 24 TB with an average size just over 4 TB
- FRIB categorizes data processing into three levels in relation to experiment operations.
  - "Online" includes the DAQ systems used for recording experiment data to permanent (disk) storage. Varying degrees of software-based event building, data reconstruction and filtering are used during online data taking. Online processing is required for recording data.

  - "Nearline" indicates processing required during experiment runtime that is not directly in the data recording path. For example, event analysis may be required to verify detector output and data quality and to inform operational decisions.
  - "Offline" is processing not directly tied to experiment operations. This includes MC and other simulations, data reduction and analysis.

- FRIB experiments use a mix of compute systems as required for nearline and online data processing. A few dozen "SPDAQ" systems are deployed as required for interfacing to detectors and electronics. These are the start of the online FRIBDaq pipelines.
  - FRIB nearline processing occurs on either individually allocated compute nodes or fixed batch processing allocations on the offline Slurm cluster.
  - FRIB offline processing is performed on Linux compute clusters using the Slurm batch system.
- Currently, FRIB operates several types of storage systems for research support.
  - A NetApp storage system provides reliable Enterprise-class storage. Snapshots, off-site replication, and tape backups are maintained for data security.
  - Higher capacity research storage currently uses either Linux/ZFS or CephFS on commodity hardware. Approximately 2 PB of storage is spread across three Linux/ZFS servers. These have 2x10GE network links. These comprise the "offline" 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. The system is connected at 2x10GE.
  - To support experiments requiring higher disk IO (>100 MB/s continuous) and to provide increased capacity, a Ceph storage cluster with CephFS is deployed. The raw capacity (before data redundancy) of the cluster is 2 PB. The cluster nodes use dual 25GE networking. The CephFS storage has been used to back Globus DTN transfers across ESnet.
- There has been an increase in the use of networking resources to enable remote users to participate in experimental runs at FRIB. Providing a capability for remote users to observe the products of ongoing data analysis would be beneficial to increase engagement with the user community.
- Researchers at FRIB are increasingly interested in using off-site HPC and data infrastructure to accomplish specific goals during the execution of an experiment. One experiment group has already employed local MSU HPCC resources to expediently analyze incoming data in near real-time to direct decisions during an experiment. Another group is exploring the use of NERSC for data analysis during ongoing experiments. A demonstration using an existing data set occurred in Fall 2023 and a planned production test with FRIB will occur in Spring 2024.
- The FRIB network is evolving with the introduction of ESnet connection in the last year. FRIB operates several internal networks. A WAN connection is provided by MSU Information Technology (MSU IT) with a 2 x 10 Gbps connection between FRIB and MSU (some links in this path have been upgraded to 100 Gbps). The WAN connection is subject to MSU IT firewall restrictions.
- The FRIB ESnet connection consists of ESnet routers, with two 100 Gbps WAN links to the ESnet network. This connection currently supports the Science DMZ including Globus DTN.

- At the conclusion of an FRIB experiment, experimental account access is disabled. Off-line analysis is typically performed at the spokesperson's home institution. Spokespersons who have accumulated data sets of more than several TB have made use of the FRIB Globus endpoint to ship data to remote storage for off-line data analysis.
- 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 FRIB 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.
- Current infrastructure work at FRIB includes adding redundancy to the Science DMZ network and standardizing support of user usage of the new DTN (provisioning new users and storage access is currently a manual process). Work in planning includes further separation (both physical and logical) of business/office and research/science networks. Addition of additional network security (stateful firewall, etc.) will allow additional services to utilize the ESnet network (in addition to Science DMZ based services).
- 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 Spring 2024, and an improved approach to the "process of science" for this remote resource is needed. A discussion at FRIB has started involving several departments for support.
- A number of limitations to addressing future needs exist at FRIB, many of which can be mitigated by the use of emerging approaches to multi-facility workflows and the IRI activity:
  - Not enough computing capability to address new instrumentation or on-site analysis.
  - Limited staff availability to adapt or convert HPC-workflows to operate within FRIB.
  - No staff expertise that is capable of leveraging capabilities at other DOE facilities (e.g., DOE HPC centers, ESnet).
- Two of the three IRI patterns will be predominant at FRIB.
  - Time-sensitive: Rapid data transmission, analysis, and inspection is critical to enable data-informed decisions during experiment execution. The appropriate time scale is on the order of one hour.

- Long-term campaign: Researchers need sustained access to computational resources over multiple years to refine an analysis of a particular experiment leading to publication.

6.5.2 FRIB Facility Profile

FRIB was designed to be the world's most powerful rare isotope research facility and commenced user operation in the latter half of FY 2022, and in FY22 delivered 3677 scheduled hours for scientific users.

6.5.2.1 Science Background

FRIB enables 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.

Goals of the Science

FRIB capabilities provide unprecedented opportunities to study the origin and stability of nuclear matter. It is 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 at final power 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 US 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 Laboratories Involved

FRIB is 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 August 2023, the FRIBUO has 1,874 members, representing 124 US colleges and universities, 13 national laboratories, and 52 countries. A listing of the membership is available at fribusers.org.

Stakeholders

DOE, FRIB Laboratory, FRIBUO, NSF, MSU, and the state of Michigan

Data Life Cycle

The Experimental Co-spokesperson (lead investigator on an approved experiment) and collaborators are expected to promptly prepare and submit for publication, with authorship that accurately reflects the contributions of those involved and all significant findings of an experiment. The Experimental Co-spokesperson shall ensure the availability of processed data used to generate charts, figures, illustrations, etc., in published works, and the published work shall indicate how these data can be accessed. The Experimental Co-spokespersons are encouraged to share research data upon request from other researchers, unless 1) the sharing would incur an unreasonable burden of cost or time or 2) the requested data are protected data. Data are considered protected data when the sharing of data would usurp the scientific results of the experiment participants or their students, or cause significant negative impact to intellectual property rights, innovation, program and operational improvements, and US competitiveness. Protected data are protected from immediate public disclosure per the Special Terms and Conditions of the DOE Contractual Agreement. Protected data are not released, and sharing of data requires permission from Experimental Co-spokespersons.

The FRIB Business Information Technology Department will provide standard FRIB tools to support the recording of research data during the running of the experiment and may support its processing. If experiments use the standard FRIB tools, then at the end of the experiment, a copy of the directory designated to contain the research data will be provided by FRIB to the Experiment Co-spokespersons or Contact Spokesperson (the point-of-contact for the facility for an experiment) upon request to the Manager for User Relations. For experiments using the standard FRIB tools, the Laboratory will also, as a courtesy, keep a duplicate of the recorded research data for a period of two years after completion of an experiment accessible from the internal network as provided by the Contact Spokesperson. After that period, data will be available for at least five years upon request.

6.5.2.2 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. Experimental research performed at 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 [Section 6.5.2.1].

The working groups include:

- Astrophysics and Separator for Capture Reactions (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 SECAR optimized for measurements of radiative capture reactions with low-energy FRIB radioactive beams.
- Advanced Targets: The Advanced Targets - Center for Accelerator Target Science (AT-CATS) Working Group is interested in creating and maintaining advanced targets for use in astrophysics, structure, and reactions studies.
- 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 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.
- Heavy Elements: The primary purpose of the US Heavy Element Working Group is to regularly bring together the US heavy element community for discussions of US-based heavy element science and capabilities.
- 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 Large multi-Institution Scintillator 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: ReAccelerator (ReA) 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.
- Neutron Source: This working group is focused on establishing an intense neutron source at FRIB to measure neutron-induced reactions on FRIB harvested radioisotopes.
- 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 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 FRIB 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 convener's 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.
- Short-lived Atoms and Molecules: The Short-Lived Atoms and Molecules (SLAM) working group aims to bring together a community of scientists and engineers working across a wide range of fields, including atomic, molecular, and optical physics, nuclear physics, radiochemistry, physical chemistry, radioactive beams, astrophysics, precision measurement, and more, to realize the opportunities offered by atoms and molecules containing short-lived, radioactive nuclei.
- 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	Do they store a primary or secondary copy of the data?	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 (FRIB, MSU, ANL, ORNL, UT, MIT, NCSU)	Primary	Data transfer, tapes, hard drive	*	Ad hoc	No	No
ADVANCED TARGETS (ORNL, ANL)	Primary and secondary	Data transfer, tapes, hard drive	*	Ad hoc	No	No
NUCLEAR DATA (NNDC, BNL, ANL, TUNL)	Primary and secondary	Data transfer, tapes, hard drive	*	Ad hoc	No	No
DATA ACQUISITION (ORNL, ANL, LBNL, MSU, FRIB, UROCHESTER)	Primary and secondary	Data transfer, tapes, hard drive	*	Ad hoc	No	No
DETECTORS FOR EOS PHYSICS (MSU, NOTRE DAME, OHIO STATE, WU, TAMU)	Primary and secondary	Data transfer, tapes, hard drive	*	Ad hoc	No	No
HEAVY ELEMENTS (LBNL, OSU, ORNL, ANL, LLNL)	Primary and secondary	Data transfer, tapes, hard drive	*	Ad hoc	No	No
HIGH-RESOLUTION IN-FLIGHT GAMMA-RAY SPECTROSCOPY (FSU, ANL LBNL, FRIB, MSU, ORNL)	Primary and secondary	Data transfer, tapes, hard drivel	*	Ad hoc	No	No
HRS: HIGH RIGIDITY SPECTROMETER (FRIB, MSU)	Primary and secondary	Data transfer, tapes, hard drive	*	Ad hoc	No	No
ION TRAPS (MSU, FRIB, UMICH, ANL)	Primary and secondary	Data transfer, tapes, hard drive	*	Ad hoc	No	No
ISLA:REA12 RECOIL SEPARATOR (FRIB, MSU, LBNL, ANL, UBUCKNELL, ANL)	Primary and secondary	Data transfer, tapes, hard drive	*	Ad hoc	No	No
ISOTOPES AND APPLICATIONS (LLNL, LANL, UALABAMA, FRIB, MSU, HOPEC)	Primary and secondary	Data transfer, tapes, hard drive	*	Ad hoc	No	No
LASER SPECTROSCOPY AND NEUTRAL ATOM TRAPS (FRIB, ANL)	Primary and secondary	Data transfer, tapes, hard drive	*	Ad hoc	No	No
NEUTRON DETECTION (ORNL, UTK, LSU, FRIB, MSU)	Primary and secondary	Data transfer, tapes, hard drive	*	Ad hoc	No	No

Table 6.5.1: Collaborative Data Mobility

**Average Data Set Sizes* - For the experiments performed at the end of the operation of the National Superconducting Cyclotron Laboratory (NSCL) with primary data stored on site data sizes ranged from a few GB to ~ 70 TB with an average size just over 4 TB. For the first year of FRIB operation the data sizes ranged from a few GB to 24 TB with an average size just over 4 TB. This applies to all the experiments.

Affiliation Abbreviations:

- ANL – Argonne National Laboratory,
- BNL – Brookhaven National Laboratory
- Caltech – California Institute of Technology
- FSU – Florida State University
- HopeC – Hope College
- IndianaU – Indiana University
- LANL – Los Alamos National Laboratory
- LBNL – Lawrence Berkeley National Laboratory
- LLNL – Lawrence Livermore National Laboratory
- LSU – Louisiana State University
- MIT – Massachusetts Institute of Technology
- MSU – Michigan State University
- NCSU – North Carolina State University
- NNDC – National Nuclear Data Center
- Notre Dame – University of Notre Dame
- NSCL – National Superconducting Cyclotron Laboratory
- Ohio State – Ohio State University
- Oregon – Oregon State University
- ORNL – Oak Ridge National Laboratory
- OSU – Oregon State University
- RutgersU – Rutgers University
- TAMU – Texas A&M University
- TUNL – Triangle Universities Nuclear Laboratory
- UAlabama – University of Alabama
- UConn – University of Connecticut
- UCSB – University of California Santa Barbara
- UMass – University of Massachusetts, Lowell
- UMich – University of Michigan
- URochester – University of Rochester
- UT – University of Tennessee, Knoxville
- WU – Washington University, St. Louis

6.5.2.3 Instruments and Facilities

Instrument Descriptions

The FRIBUO has 22 working groups for experimental instrumentation that develop plans for using existing devices at FRIB as well as proposals for new equipment.

Existing/Near Future Instruments (Present to Five Years)

1. The Advanced Rare Isotope Separator (ARIS) 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 ARIS 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. The ARIS 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 ARIS 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) 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 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, FRIB, 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 US 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° × 1° for ▯ and ▯, 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. 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%.

26. The FRIB Decay Station Initiator (FDSi) is an efficient, granular, and modular multi-detector system capable of performing spectroscopy with multiple radiation types over a range of beam production rates spanning ten orders of magnitude. Typical use-rate of existing instruments in this experimental program is 20%.

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%.

28. SOLARIS is a large bore solenoid spectrometer that can apply a magnetic field up to 4 Tesla parallel to the beam direction for experiments with the AT-TPC or with a silicon detector array. Typical use-rate of existing instruments in this experimental program is 5%.

29. Resonance-ionization Spectroscopy Experiment (RiSE) is a laser spectroscopy instrument that will be integrated in the BECOLA facility for high-sensitive measurements with rare isotopes available as stopped beams at FRIB. Typical use-rate of existing instruments in this experimental program is 2%.

30. The Positron Polarimeter consists of a pair of identical superconducting solenoids, with each solenoid capable of producing a maximum field of 2 T, used for fundamental interactions studies. Typical use-rate of existing instruments in this experimental program is 2%.

Planned Instruments (Beyond Five Years)

1. 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%.

2. 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 Tesla, 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%.

3. The FRIB Decay Station (FDS) will be built on the FDSi and be 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 electronics 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%.

4. 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. 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. Typical use-rate of existing instruments in this experimental program is 15%.

Compute, Storage, and Network Capabilities

FRIB categorizes data processing into three levels in relation to experiment operations.

- **"Online"** includes the DAQ systems used for recording experiment data to permanent (disk) storage. Varying degrees of software-based event building, data reconstruction and filtering are used during online data taking. Online processing is required for recording data.
- **"Nearline"** indicates processing required during experiment runtime that is not directly in the data recording path. For example, event analysis may be required to verify detector output and data quality and to inform operational decisions.
- **"Offline"** is processing not directly tied to experiment operations. This includes MC and other simulations, data reduction and analysis.

Compute

FRIB experiments use a mix of compute systems as required for online data processing. A few dozen "SPDAQ" systems are deployed as required for interfacing to detectors and electronics. These are the start of the online FRIBDaq pipelines. Generally, these are smaller x86 based modules in electronics crates or PC systems. Event building and online reconstruction may require larger commodity server systems. Currently, typical specifications are 1 or 2 x86 processors with 16 to 32 cores and 128 GB RAM with 10GE networking.

Nearline processing occurs on either individually allocated compute nodes or fixed batch processing allocations on the offline Slurm cluster. Typical individually allocated nodes have 1 x86 processor, 16 to 32 cores and 128 to 512 GB RAM with 10 GE networking.

Offline processing is performed on Linux compute clusters using the Slurm batch system. Newer compute nodes have 32 or 64 core AMD Epyc CPUs with 256 to 512 GB of RAM (6+ GB per core) with 25GE networking. Older compute nodes include 12 to 40 core Intel Xeonsystems with 4+GB RAM per CPU core and 10GE networking. Approximately 1200 total cores are available.

A handful of Linux login hosts with similar specifications provide access to the software environment including batch job management and enable interactive data processing with tools such as ROOT.

Storage

Currently, FRIB operates three types of storage systems for research support. A NetApp storage system provides 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, and additional storage areas requiring a high level of data protection.

Higher capacity research storage currently uses either Linux/ZFS or CephFS on commodity hardware. Approximately 2 PB of storage is spread across three Linux/ZFS servers. These have 2x10GE network links. These comprise the "offline" 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.

To support experiments requiring higher disk IO (>100 MB/s continuous) and to provide increased capacity, a Ceph storage cluster with CephFS is deployed. The raw capacity (before data redundancy) of the cluster is 2 PB.

The cluster nodes use dual 25GE networking. The CephFS storage has been used to back Globus DTN transfers across ESnet.

Network

The FRIB network is evolving with the introduction of ESnet connection in the last year. FRIB operates several internal networks. A WAN connection is provided by MSU Information Technology (MSU IT) with a 2x10GE connection between FRIB and MSU (some links in this path have been upgraded to 100GE). The WAN connection is subject to MSU IT firewall restrictions. MSU IT also provides Wi-Fi coverage within the FRIB office buildings.

FRIB manages a border firewall between the internal network and MSU campus. The internal wired networks are generally configured for one and 10 Gbps Ethernet. The switched network core uses 100GE switches.

The ESnet connection consists of ESnet routers located at FRIB with two 100GE WAN links to the ESnet network. This connection currently supports the ScienceDMZ including Globus DTN.

Additional information on the FRIB network infrastructure is included in [Section 6.5.2.7].

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.

6.5.2.4 Generalized Process of Science

Present to Five Years

FRIBDAQ is the main data acquisition application for producing and handling data flow at FRIB. In FRIBDAQ, 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). Nuclear Physics 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 DAQnet. This data storage is accessed by experimenters using Linux workstations in the data-taking areas for the duration of an experiment.

Co-Spokespersons are responsible for complying with respective data policies, and adhere to the FRIB Data-Management Plan, as described in [Section 6.5.2.1], paragraph "Data Life Cycle"

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, 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 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 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 HPCC that is part of MSU iCER.

FRIB 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 are 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 for three previous campaigns, and a third campaign began in June of 2023. 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 10 Gb 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.

6.5.2.5 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.

Another set of SπRIT-TPC experiments is planned in Spring 2024, and an improved approach to the "process of science" for this remote resource is needed. A discussion at FRIB has started involving several departments for support, although there is no ongoing discussion regarding setting up a direct, high-speed network connection between RIKEN, iCER, and FRIB, where high-volume storage is available and affordable. The bulk of the large-scale analysis is planned by using IT resources at RIKEN.

Tools for Nuclear Theory

Part of the "process of science" that was not discussed in [Section 6.5.2.4] 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 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.

Remote Experiment Participation

There has been an increase in the use of networking resources to enable remote users to participate in experimental runs at FRIB. Providing a capability for remote users to observe the products of ongoing data analysis would be beneficial to increase engagement with the user community.

6.5.2.6 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 FRIB 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 FRIB 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.

An increasing 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.

6.5.2.7 Network and Data Architecture

Present to Two Years

Network Description

The FRIB network (Office-net) consists of 100, 25, 10 and 1 Gbps Ethernet supporting general business IT functions, office LAN, Linux research compute systems, including interactive and batch systems, and Linux ZFS/NFS and CephFS storage systems. Infrastructure is shared with FRIB DAQ systems (DAQnet) 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 FRIB Linac operations (FRIBControl-net), other Experimental Physics and Industrial Control System (EPICS) based controls (NSCLControls-net), etc.

WAN connectivity is provided via the MSU campus network (see accompanying figure). FRIB has 2x10GE connection to campus with a FRIB managed border firewall. MSU R+E traffic primarily traverses Michigan Educational Research Information Triad (MERIT) managed links to R+E pops in Chicago. MERIT operates multiple DWDM links in lower Michigan providing redundant paths from MSU campus to Chicago. This redundancy includes multiple DWDM connection points on the MSU campus.

ESnet and FRIB have established a connection point at FRIB. A pair of ESnet routers are deployed at FRIB. These are uplinked via MSU campus fiber and MERIT DWDM links to ESnet locations in Chicago. Two independent paths disperse 100GE links are active. MERIT and MSU IT manage the physical and lower layer (layer 1) aspects of these links while ESnet manages all routing and traffic control. This provides ESnet managed connections end-to-end between FRIB and other ESnet connection points.

The ESnet connection currently serves a Science DMZ including Globus DTN and perfSONAR monitoring node. See diagram. The Science DMZ has minimal connection to the internal network (specific storage traffic only). Throughputs greater than 10Gbs for disk-disk transfers for both short (Midwest, East Coast US) and longer distance (West Coast US, NERSC) have been demonstrated for disk-disk transfers.

Data Transfers

External data transfers currently use the ESnet/Science DMZ based DTN or Secure SHell (SSH) or Globus Online to the Office network. Work is ongoing to migrate usage from the legacy DTN in Office network to the new Science DMZ DTN. The new DTN will become the preferred data-transfer system for FRIB science users.

Compute resources at MSU iCER are utilized by some local researchers. Data transfers between FRIB and MSU iCER HPCC across the campus network 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. Currently, transfers from the FRIB Science DMZ to MSU iCER HPCC would go via Chicago.

As the FRIB internal network evolves with additional segmentation between science and office functions, a DTN may be leveraged for internal transfers as well.

Infrastructure Work

Current infrastructure work includes adding redundancy to the Science DMZ network and standardizing support of user usage of the new DTN (provisioning new users and storage access is currently a manual process). Work in planning includes further separation (both physical and logical) of business/office and research/science networks. Addition of additional network security (stateful firewall, etc.) will allow additional services to utilize the ESnet network (in addition to Science DMZ based services).

Planning is beginning to define the GRETA external network utilizing ESnet. This will allow a DTN with access to GRETA storage to transfer data to external sites.

Beyond Two Years

In this time frame the focus is on utilizing ESnet to provide reliable production services to FRIB science users and exploration of the IRI model for FRIB users. These include remote access, and user managed data transfers via Globus and DTNs, and remote processing activities.

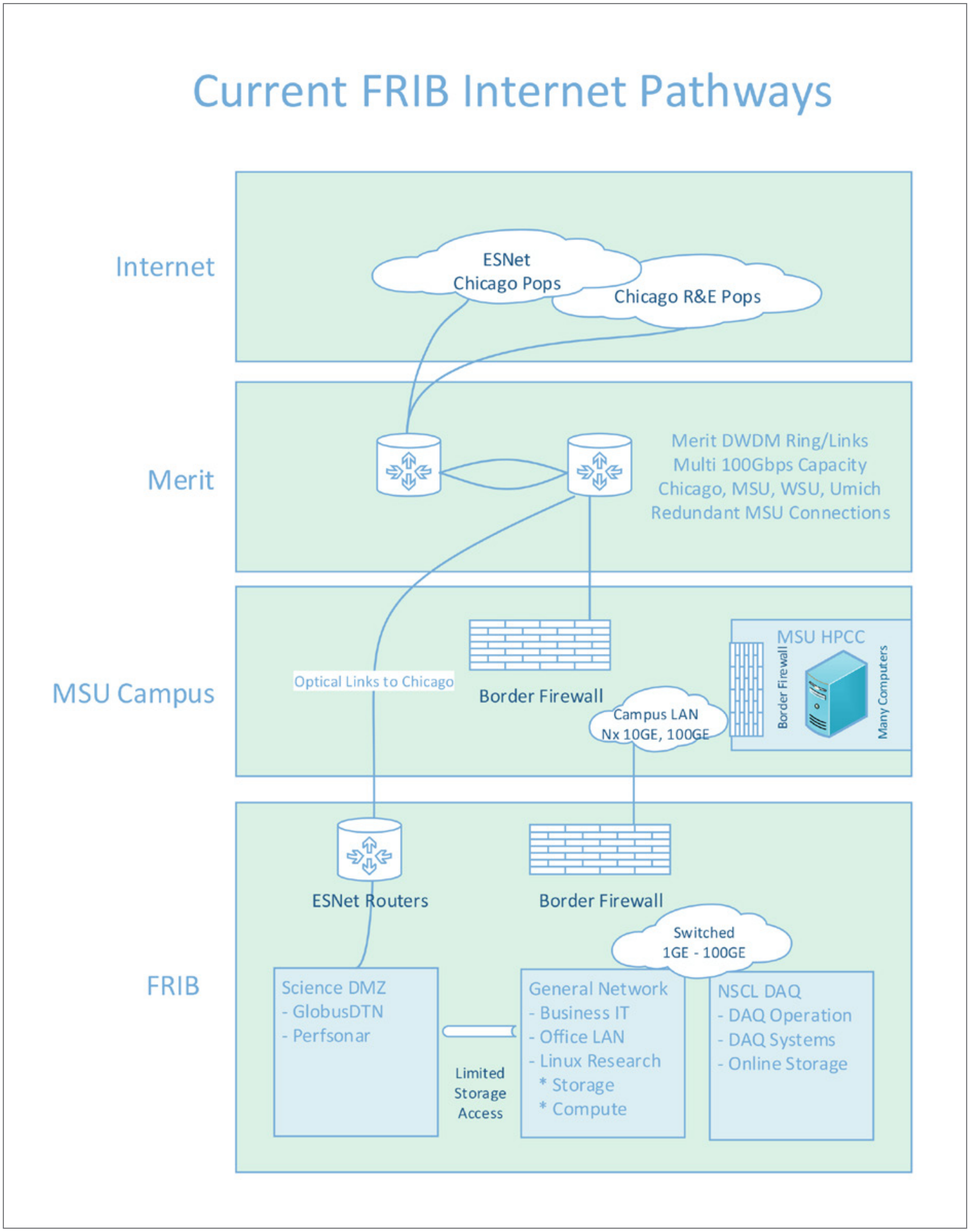


Figure 6.5.1: Diagram of the FRIB internet pathways

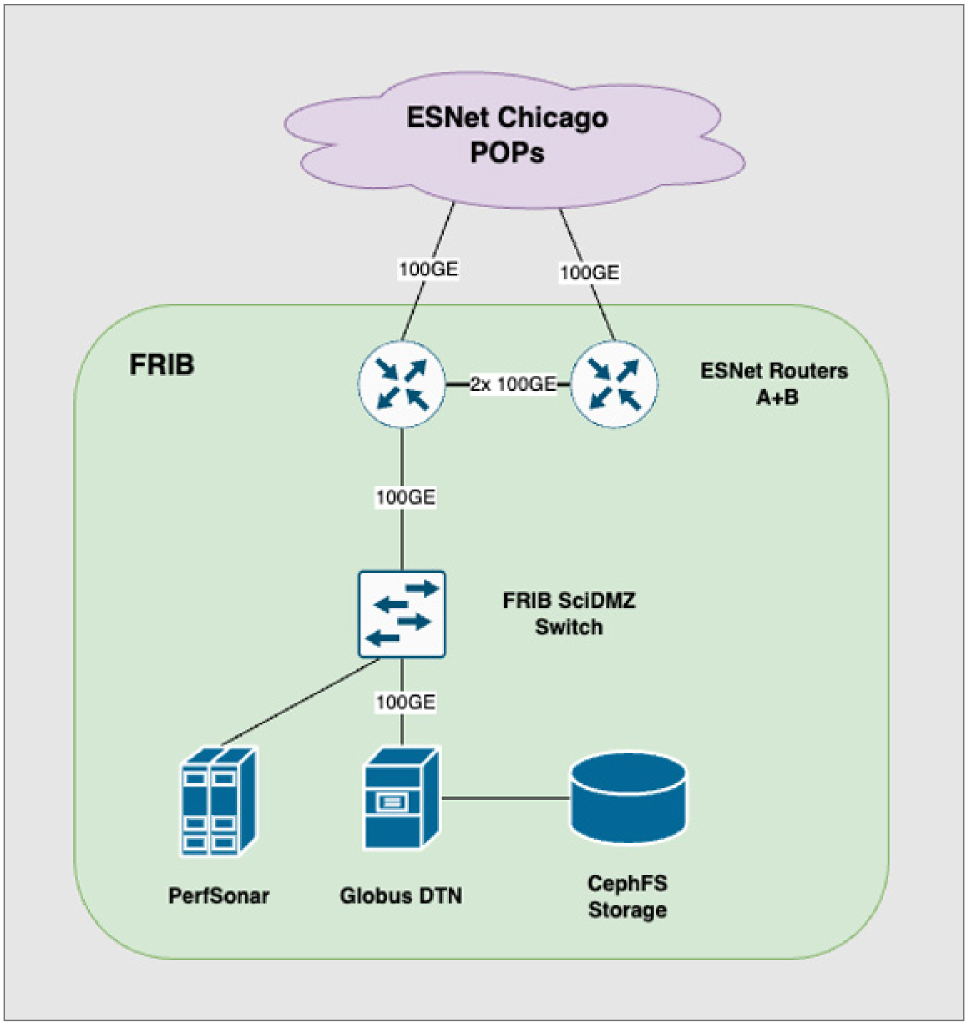


Figure 6.5.2: Diagram of initial Science DMZ

6.5.2.8 IRI Readiness

Researchers at FRIB are increasingly interested in using off-site HPC and data infrastructure to accomplish specific goals during the execution of an experiment. One experiment group has already employed local MSU HPCC resources to expediently analyze incoming data in near real-time to direct decisions during an experiment. Another group is exploring the use of NERSC for data analysis during ongoing experiments. A demonstration using an existing data set occurred in Fall 2023 and a planned production test with FRIB will occur in Spring 2024. The GRETA instrument also has long term plans to leverage multiple facilities for experimental analysis which are detailed in that case study. With the upcoming HPDF, the ability to store data and take advantage of the DOE HPCC facilities throughout the country to off-load data processing in an efficient and balanced way becomes a key goal. Furthermore, the partnership between JLab and ESnet to define an edge to data center traffic shaping/steering transport capability to increase the science rate for high data rates/volumes with its keystone ESnet JLab FPGA Accelerated Transport (EJFAT) project will be an asset for the NP community.

Two of the three IRI patterns will be predominant at FRIB.

- Time-Sensitive: Rapid data transmission, analysis, and inspection is critical to enable data-informed decisions during experiment execution. The appropriate time scale is on the order of one hour.

- The Long-Term Campaign: Researchers need sustained access to computational resources over multiple years to refine an analysis of a particular experiment leading to publication.

6.5.2.9 Cloud Services

MSU offers cloud services and cloud storage options, and has established guidelines for the appropriate use of cloud services[19]. 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

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)

6.5.2.10 Data-Related Resource Constraints

FRIB has limited ways to address the growth of data in the coming years:

- Local compute capability to address new instrumentation and new analysis needs will be limited, and off-site options (e.g., MSU, DOE HPC, etc.) will be explored as new instruments and experiments come online.
- Local expertise related to high-performance and exascale computing is limited, and specialized to the resources directly operated by MSU and FRIB.
- local expertise to implement the IRI vision is limited, and FRIB is interested in partnering with entities like the DOE HPC centers, other experimental facilities, and ESnet

6.5.2.11 Data Mobility Endpoints

Nothing to report.

6.5.2.12 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

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

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.

6.5.2.13 Facility Profile Contributors

FRIB Representation

- Giordano Cerizza, FRIB, cerizza@frib.msu.edu
- Sean Liddick, FRIB, liddick@nscl.msu.edu

ESCC Representation

- Clinton Jones, FRIB, jonesc@frib.msu.edu
- Thomas Rockwell, FRIB, rockwell@frib.msu.edu

6.6 GRETA

The study of atomic nuclei is central to our understanding of the world around us. Comprising 99.9% of the visible matter in the universe nuclei are, in multiple aspects, central to fundamental questions in physics, such as our understanding of the origin of the elements and how complex many-body quantum systems organize. FRIB [Section 6.5] was constructed at MSU, and produces thousands of new short-lived radioactive (rare) isotopes and will greatly expand our reach to evermore exotic nuclei heavier in mass and closer to the limits of existence.

GRETA will be a key instrument at FRIB, capable of reconstructing the energy and three-dimensional position of γ-ray interactions. Its design provides the unprecedented combination of full solid-angle coverage and high efficiency, excellent energy and position resolution, and good background rejection (Peak-to-Total (P/T)) needed to carry out a large fraction of the nuclear structure and nuclear astrophysics science programs at FRIB. GRETA will be movable between the various beam-lines at FRIB and will also be used at other facilities such as the ATLAS stable-beam facility at ANL.

6.6.1 Discussion Summary

The following bullets were extracted from the case study, and in-person discussion. Several appear in Section 1 and Section 3.

- GRETA is an advanced gamma-ray spectrometer for low-energy nuclear physics measurements, and is currently in the final stages of fabrication. The array is a primary instrument for the FRIB. 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.
- GRETA is nearing the end of its fabrication phase. 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. GRETA is currently planning for start of operations in 2024 and project completion (all detectors, full rate) in 2027.
- GRETA will be initially sited at FRIB with the possibility of later operations at ATLAS/ Argonne. It is expected that once GRETA has completed construction and sited at FRIB it will be operated by a local operations team with technical support provided by LBNL staff.
- The raw data collected from the GRETA detectors are the energies, times and associated waveforms as captured in the ADCs/FPGAs which instrument GRETA. From these data the

location of gamma-ray interaction points can be inferred, 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, are provided to the experimental team of a given measurement. These data are cached locally for a period of weeks to allow sufficient time for GRETA experimental teams to transfer the data back to their local institution/computing resource for analysis.

- 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.
- Generally, analysis of GRETA data is carried out by the experimental team at their home institutions. Analysis and data interpretation is a time-consuming process (many months) but not a very computationally intensive process (can be done on local computing resources) The nature of this analysis is very much experiment dependent.
- GRETA performs a real-time analysis of the digitized waveform data. This analysis 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.
- Associated with each GRETA experiment there 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 workflow is a data analysis step carried out by the experimenter and their group
  - The workflow for real time signal processing consists of 120 UDP data streams with an aggregate (maximum) rate of 32 Gbps. Signal processing is carried out on a GPU-cluster co-located with the experiment. All pipeline components in the GRETA pipeline and control plane are container based. This simplifies deployment and allows for the complex orchestration needed for future IRI implementations.
  - The data analysis workflow involves clustering and ordering interaction points into likely gamma-ray tracks and rejecting partial energy deposition events. This analysis step is not considered computationally (CPU or network) intensive and requires off-the-shelf computing resources.
- GRETA data transfer volumes can be between 50 GB and 100 TB, and performed on an ad-hoc basis when the detector is operating.
- GRETA can write 1 GB/s to its local disk cache although typical rates are expected to be less than this. Data set sizes are highly dependent on the physics case being studied. It depends on the triggered gamma-ray rate, the auxiliary detectors employed, and the beam time allocated. An experiment's aggregate data size is expected to range from a minimum of 50 GB to a maximum 100 TB. Individual file sizes should be < 2 TB.
- Given GRETA is an instrument that is movable, its network address space is necessarily abstracted from that of the host laboratory. External network facing components 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 to admit the possibility to send full waveform data to remote computing facilities

- 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 real time processing itself. For example, a coupled signal decomposition and tracking algorithm would require such an infrastructure.
- In the late 2–5-year timeframe, 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.
- There are three primary ways in which GRETA will interact with the wide area network:
  - System-level access by staff with appropriate access permission for system maintenance;
  - Download of experiment data sets from the GRETA DTN to remote analysis resources; and
  - Possible future signal processing modalities which require resources beyond the capabilities of the production GRETA signal decomposition cluster.
- 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. Globus will serve as our data-transfer tool.
- In the next five-year timeframe, when GRETA begins collecting data initially at FRIB and potentially subsequently at ATLAS/Argonne, real-time component of signal processing (workflow 1) will be carried out using GRETA's local computing cluster or potentially at HPC facilities. 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, the demand for use of cloud services in the final analysis should be limited.
- The GRETA data pipeline was designed with IRI workflows in mind and this option is actively being developed. GRETA's forward buffers can send their data over WANs to remote HPC facilities where the main data processing tasks could be carried out on interactive timescales (time sensitive pattern). These workflows are currently being evaluated using the ESnet testbed and OLCF/ORNL IRI testbed and are expected to be it ready for production use in a two-year timeframe.

6.6.2 GRETA Experimental Case Study

GRETA is an advanced gamma-ray spectrometer for low-energy nuclear physics measurements, funded by the DOE SC, Office of NP, and is currently in the final stages of fabrication. 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 toward a predictive description.

The array is a primary instrument for the upcoming FRIB which recently started operation 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.

6.6.2.1 Science Background

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 will correspond to of order 2 to 10 days of beam time, with GRETA operated in conjunction with particle detection systems or spectrometers.

The function of a gamma-ray spectrometer such as GRETA is to measure the energy and tracks of gamma-rays. 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 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 below), 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 raw data collected from the GRETA detectors are the energies, times and associated waveforms as captured in the ADCs/FPGAs which instrument GRETA's 4800 electronic channels. 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 these 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, are provided to the experimental team of a given measurement. These data are cached locally for a period of weeks to allow sufficient time for GRETA experimental teams to transfer the data back to their local institution/computing resource for analysis.

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

6.6.2.2 Collaborators

GRETA is a scientific instrument (MIE) which will be sited primarily at FRIB. GRETA is not a collaboration. Use of the instrument is proposal driven and governed by a PAC at the host institution as outlined in the response to question 1 above. Beam time / use of GRETA at the facility is granted on the basis of scientific merit for individual experimenter-driven proposals. Beam time requests are typically 2 to 10 days in duration.

Proposals are generated by experimenter-PIs at US universities, US 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	Do they store a primary or secondary copy of the data?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of dataset? (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)?
US UNIVERSITY-BASED PIS	Primary	Data transfer	50 GB-100 TB	ad hoc (< 4 weeks)	No	N/A
US NATIONAL LAB-BASED PIS	Primary	Data transfer	50 GB-100 TB	ad hoc (< 4 weeks)	No	N/A
INTERNATIONAL PIS	Primary	Data transfer	50 GB-100 TB	ad hoc (< 4 weeks)	No	N/A

Table 6.5.1: Collaborative Data Mobility

6.6.2.3 Use of 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 track of the incident γ-ray.

GRETA is currently nearing the end of its fabrication phase. 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 beam-lines/accelerators). It is expected that once GRETA has completed construction and sited at FRIB it will be operated by a local operations team with technical support provided by LBNL staff. Given the GRETA construction schedule, comments in this report apply to the five-year timescale (current budget horizon, current technology horizon).

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 modern GPU-equipped mid-scale computing cluster is required. Processed data will be stored to a 300 TB high performance SSD-based storage array, which acts as a cache until it is moved to the experimenter's home facility through a DTN for later analysis. A diagram showing the main components of the GRETA computing system is given in Figure 6.6.1.

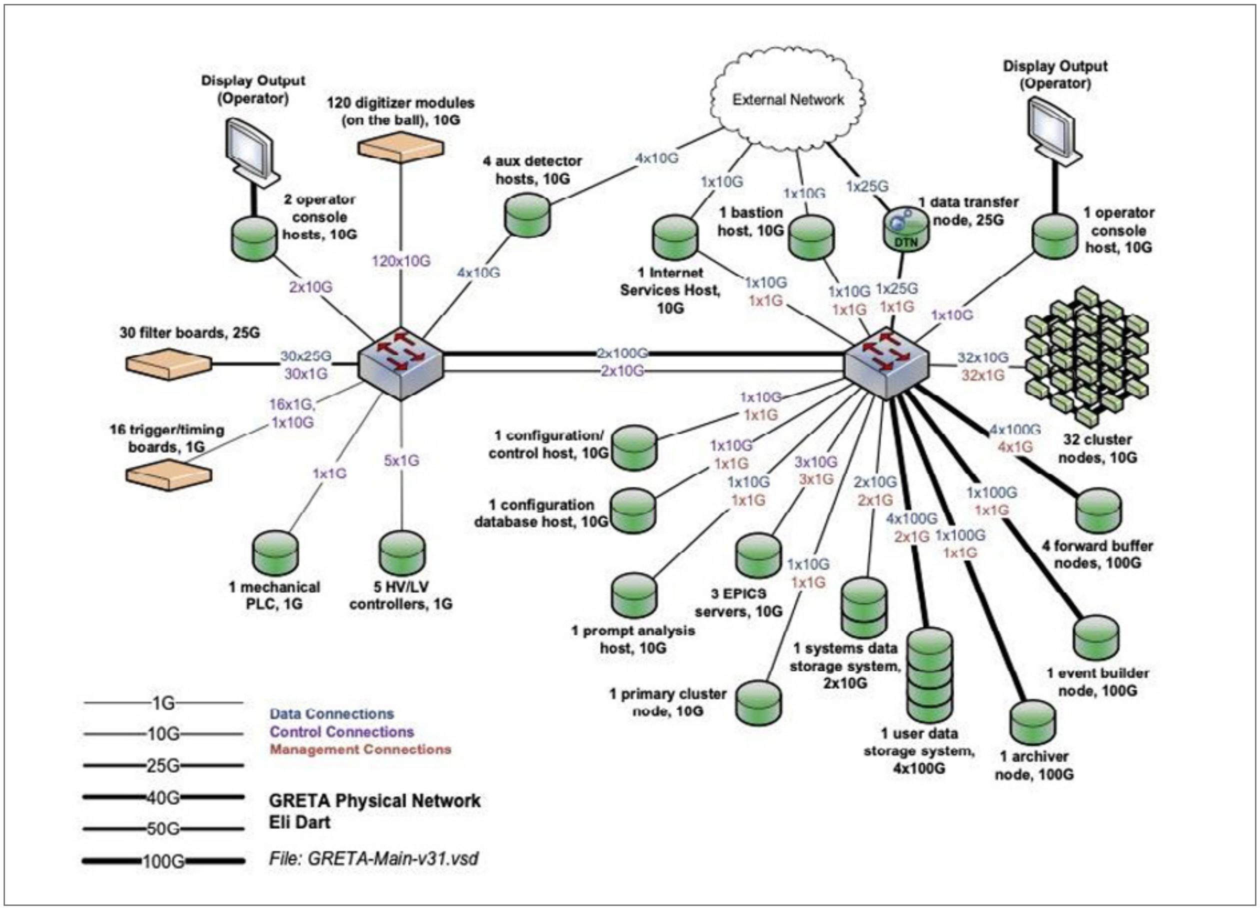


Figure 6.6.1: Network diagram of GRETA computing infrastructure co-located with experiment.

Given GRETA is an instrument that is movable, its network address space is necessarily abstracted from that of the host laboratory. External network facing components 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 to admit the possibility to send full waveform data to remote computing facilities (see [Section 6.6.2.8])

The datasets 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 a custom binary format consisting of routing headers which define transport through the pipeline and a payload of event-level experiment data whose processing depends on a type/subtype pair.

GRETA can write 1 GB/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. It depends on the triggered gamma-ray rate, the auxiliary detectors employed, and the beam time allocated. We expect an experiment's aggregate data size to range from a minimum of 50 GB to a maximum 100 TB. Individual file sizes should be < 2 TB.

6.6.2.4 Process of Science

Associated with each GRETA experiment there 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 workflow is a data analysis step carried out by the experimenter and their group. A brief description of both of these workflows is given below.

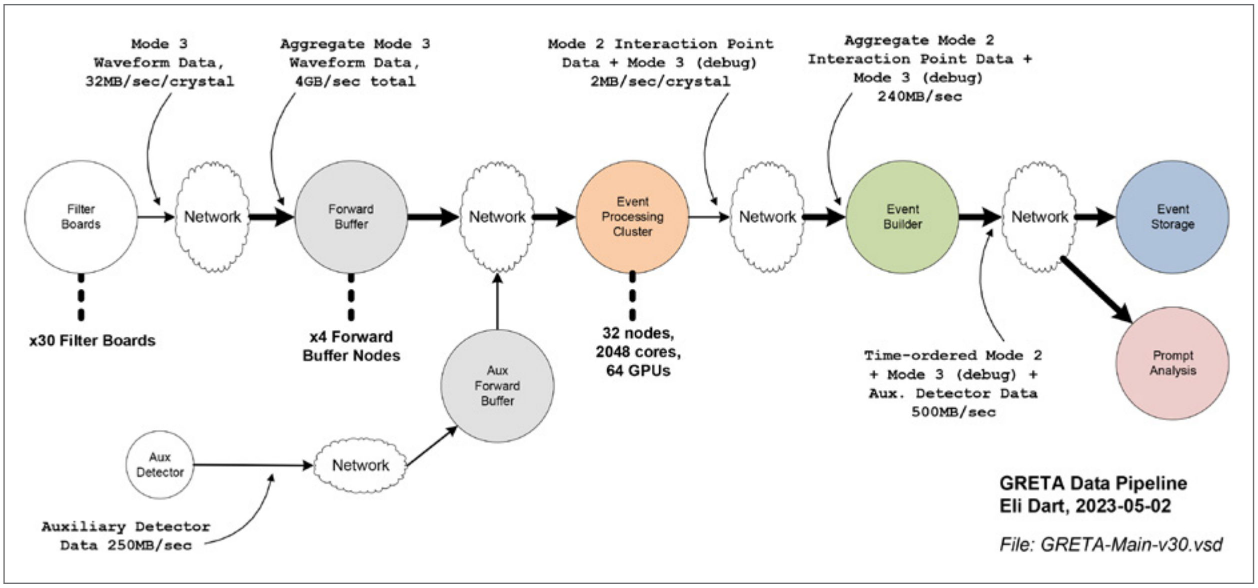


Figure 6.6.2: Signal processing workflow for the GRETA spectrometer.

The workflow for real time signal processing is shown in Figure 6.6.2. 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 UDP packets. There are 120 such UDP streams, each corresponding to a detector crystal, and their aggregate (maximum) rate is 4 GB/s with the possibility of load asymmetry of up to 7:1. These data are captured by up to 4 forward buffer nodes which serve these data to a computing cluster for signal analysis and provide global flow control.

Signal processing is carried out on a GPU-cluster co-located with the experiment running GRETA's signal decomposition algorithms to locate interaction points. Each node provides several processing 'slots', roughly equaling the number of cores, which can be allocated to a given detector crystal. This is necessary as each detector has different physical characteristics and a large in-memory detector simulation, tailored to the characteristics of a specific detector, is required. Load balancing within the cluster is accomplished by altering the allocation of the number of slots to detector crystals on the cluster.

All signal decomposition nodes 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.

All pipeline components in the GRETA pipeline and control plane are container based. This simplifies deployment and allows for the complex orchestration needed for future IRI implementations.

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.

6.6.2.5 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 real time processing itself (workflow 1) as described in [Section 6.6.2.8]. For example, a coupled signal decomposition and tracking algorithm would require such an infrastructure.

6.6.2.6 Software Requirements

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.

6.6.2.7 Network and Data Architecture Requirements

The local network architecture for GRETA is centered around the data acquisition and signal processing components as described in [Section 6.6.2.3]. The network diagram, given by Figure 6.6.1, shows several key components including:

1. The FPGA-based filter boards, which send the data captured by the detector electronics to the computing infrastructure for signal processing;
2. 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;
3. The signal decomposition cluster, which provides the computing necessary for signal processing;
4. 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
5. A storage system which provides nonvolatile cache storage for the data files produced by the global event builder.

Another view of this workflow is given in [Section 6.6.2.4], which describes the workflow from a dataflow perspective rather than from a network architecture perspective.

There are three primary ways in which GRETA will interact with the wide area network:

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 below.

System-level access for maintenance and troubleshooting is not expected to occur routinely, as there will be an operational team at FRIB which is 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. Globus will serve as our data-transfer tool.

In the late two- to five-year timeframe, 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. Further discussion of this mode of operation is given in [Section 6.6.2.8].

6.6.2.8 IRI Readiness

The GRETA data pipeline was designed with IRI workflows in mind and we are actively developing this option. GRETA's forward buffers can send their data over WANs to remote HPC facilities where the main data processing tasks could be carried out on interactive timescales (time sensitive pattern). This would allow for more complex tracking algorithms to be applied to the data then could be accomplished on GRETA's local computing cluster. These workflows are currently being evaluated using the ESnet testbed and OLCF/ORNL IRI testbed and we expect it ready for production use in a two-year timeframe.

Given the transport and scientific processing aspects of the GRETA pipeline are well abstracted, the pipeline can be applied to a wide variety of experiments which follow a time sensitive pattern.

6.6.2.9 Use of Cloud Services

In the next five-year timeframe, when the instrument begins to collect data initially at FRIB and potentially subsequently at ATLAS/Argonne, we expect the real-time component of signal processing (workflow 1) to be carried out using GRETA's local computing cluster or potentially at HPC facilities. 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.

6.6.2.10 Data-Related Resource Constraints

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

6.6.2.11 Data Mobility Endpoints

GRETA will have a Globus endpoint for data distribution at the experimental facility where it is sited (FRIB initially). Endpoints where the data will be received are dependent on the analysis sites of PAC approved experiments.

6.6.2.12 Outstanding Issues

Nothing to report.

6.6.2.13 Facility Profile Contributors

GRETA Representation

- Heather Crawford, LBNL, hlcrawford@lbl.gov
- Mario Cromaz, LBNL, mcromaz@lbl.gov
- Paul Fallon, LBNL, pfallon@lbl.gov

ESCC Representation

- Richard Simon, LBNL, rsimon@lbl.gov
- Rune Stromsness, LBNL, rstrom@lbl.gov

6.7 ALICE Project and ALICE-USA Computing

ALICE is a detector dedicated to heavy-ion physics at the LHC. It is designed to study the physics of strongly interacting matter at extreme energy densities, where a phase of matter called QGP forms.

The ALICE collaboration uses the 10 000-tonne ALICE detector – 26 m long, 16 m high, and 16 m wide – to study QGP. The detector sits in a vast cavern 56 m below ground close to the village of St Genis-Pouilly in France, receiving beams from the LHC.

6.7.1 Discussion Summary

The following bullets were extracted from the case study, and in-person discussion. Several appear in Section 1 and Section 3.

- The ALICE collaboration has constructed and operates a heavy ion detector to exploit the unique physics potential of proton-proton and nucleus-nucleus interactions at collision energies of the LHC at CERN. The ALICE Collaboration consists of over 2000 scientists, engineers and students spread over 170 institutions in 41 countries.
- The ALICE-USA computing project was established to meet US obligations by operating ALICE Grid facilities in the US. The initial ALICE-USA obligations corresponded to about 6% of all ALICE computing resource needs and are currently about 8% of those requirements.
- The three-year LHC Run 2 period ended in 2018. LHC is currently in the Run 3 which began in July 2022. For the ALICE and LHCb experiments, Run 4 marks the beginning of the high luminosity (HL) LHC era with data rates from the detectors reaching a up to 100 times larger than those of Run 2.
- ALICE data consist of ~50 PB of CTF/raw data, and ~5 PB of AO2D data collected and produced annually (since Run 3). These data are stored at CERN, but also distributed among seven Tier-1 sites (Germany, France, Italy, UK, Netherlands, Korea, and Russia). Additional MC data are stored at T1 and T2 sites (around 5PB per data taking period, 60PB overall), and 5 PB of analysis products may be found at analysis facilities.
- The key feature for data management and access on the ALICE Grid is the distribution of the data onto the grid at the data-creation. The data are then subsequently accessed only from local site storage . 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 about 200 PB distributed file system that is primarily used as disk storage on the local cluster.
- During the past year ALICE jobs have read over 2.3 XB and written over 400 PB of data from/to the local storage, averaging about 13 GB/s and 80 GB/s for write and read traffic respectively averaged over the entire grid. Over the same period, conversely, ALICE jobs have read just 30 PB and written 3.4 PB of data over the WAN, averaging just 1 GB/s of aggregate bandwidth.
- The wide area data distribution mode for the ALICE-USA sites is: 1) receive a fraction of ALICE AO2D 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 US sites.
- 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 SE.
- 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.

- The ALICE-USA Computing project officially launched in 2010, and underwent changes to participants and computing architecture in 2015 and 2018. ALICE Grid sites today are based at the ORNL CADES facility and the LBNL/HPCS facility. In addition to the core facilities, the ALICE-USA project has a history of working to integrate HPC systems into the ALICE Grid facility. Integrations at NERSC started with Cori in 2020, and have expanded into Perlmutter.
- ALICE does not operate a Grid-enabled storage element inside NERSC, which poses an issue running analysis jobs that are I/O heavy. With the switch to Perlmutter the initial estimations show that this bottleneck is less constraining, allowing ALICE to leverage the proximity of the LBNL T2 SE to use it as an input source to jobs that require data. This does increase network traffic between LBNL and NERSC, which has a small latency, but is still limited. Understanding the network limits now (zero to two years) and future (two to five year) capacities may help guide the development of ALICE use of NERSC HPC in the era beyond Perlmutter.
- ALICE has yearly episodes that require more significant WAN capacities. These occur when storage is added and/or decommissioned, or when data must otherwise be redistributed between different sites. 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.
- The ALICE T2 site at ORNL currently 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. Plans are underway this year to transition the ALICE environment to connect directly to the ORNL border routers at a peer level with the Science DMZ. The ALICE network environment is in the process of being upgraded to switches connected at 40G. The upgraded ORNL ALICE environment is expected to be completed early in 2024. Beyond 2024 ORNL connections to ESnet are expected to migrate proportionate with ESnet backbone speeds
- The current network topology of the ALICE T2 facility at the LBNL HPCS center features an internal connection between the worker nodes and the storage using 56 Gb 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 LBLnet connection to ESnet. The SEs were recently added to the LHCONE via a 10 Gb firewall. An item for the longer (two to five year) term is to understand and if necessary, optimize the network connectivity between the storage at the HPCS facility and Perlmutter HPC at NERSC.
- ALICE can benefit from additional supercomputing resources especially with GPU capabilities.
- The current and developing ALICE computing models do not have any specific plans for use of cloud resources. However, user analysis within commercial clouds is a possibility if it is cost effective. ALICE research teams in Europe found that the use of commercial clouds was fully functional and efficient for running MC simulations, from which the produced simulated data were distributed 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 is significantly less than a MB/sec, which would allow thousands of such tasks (jobs) to be run concurrently on a Cloud service.
- Systematic, automated WAN monitoring is very important for efficient use of ALICE-USA resources. This goes beyond any one specific site-to-site monitor as it should seamlessly include all paths between any two ALICE Grid sites. ALICE has a simple, yet effective, monitor using

traceroute between every VOBox, that can be used in near-real time to adjust automated data placement, but which does not provide any real diagnostic capabilities. The plan is to leverage the OSG network group's perfSONAR dashboards. That plan has not been implemented, partially due to the reduced active collaboration between ALICE Grid sites during COVID, but remains our current best model for tackling this need. The project, however, is interested in any feedback regarding future monitoring capabilities from ESnet or other service providers.

6.7.2 ALICE Experimental Case Study

The ALICE collaboration has constructed and operates a heavy ion detector to exploit the unique physics potential of proton-proton 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 interacting 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 archival storage within the sponsoring country in proportion to its Ph.D. participation. The ALICE-USA computing project was established to meet US obligations by operating ALICE Grid facilities in the US. The initial ALICE-USA obligations corresponded to about 6% of all ALICE computing resource needs and are currently about 8% of those requirements. A brief history of the computing project is provided here.

6.7.2.1 Science Background

The ALICE-USA Computing project officially launched in 2010 and originally deployed and operated US ALICE Grid sites located within two computing centers, Livermore Computing (LC) at LNLL and NERSC Parallel Distributed Systems Facility (PDSF) at LBNL. In 2015, the ALICE group at LLNL withdrew from ALICE, which led the US 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, NERSC planned to decommission PDSF facility to make room for its future generations of supercomputers and storage systems. As a result, the US computing project developed a transition plan to build a new ALICE Grid site in the LBNL/HPCS 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-US obligations coming from LBNL. The LBNL/HPCS officially took over as the second US ALICE Grid site in April of 2019 with the decommissioning of the NERSC PDSF cluster. An overview of this history is illustrated in Figure 6.7.1, which shows the number of concurrent ALICE jobs in each center during the thirteen years of the ALICE-USA computing project.

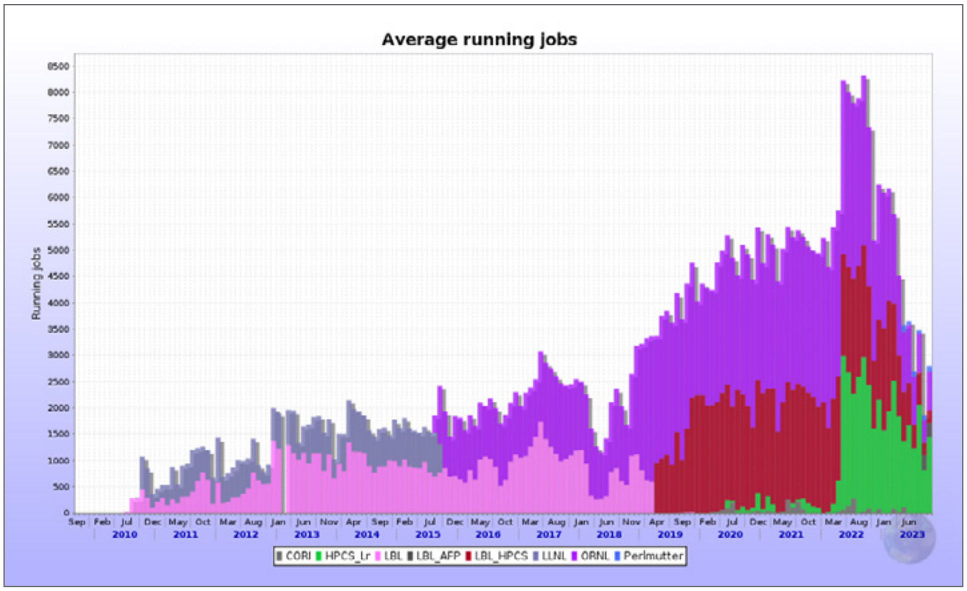


Figure 6.7.1: Number of concurrent ALICE jobs on each US Grid site over the lifetime of the ALICE-USA Computing Project.

The apparent decline, seen in Figure 6.7.1, in the number of jobs on US sites around yearly 2023 is a result of switching the site compute node operations to the whole-node scheduling mode. In this mode of operation entire compute node is running a single pilot job, called JobRunner, which then matches the node resource (available CPU and RAM) to the jobs submitted to the central services based on multiple parameters (such as proximity to data, recent reliability of the compute element [CE], etc.) and keeps launching payload within the pilot job, while optimizing the node use.

In addition to the use of conventional clusters at LLNL/LC, NERSC/PDSF, ORNL/CADES and LBNL/HPCS, the ALICE-USA project has a history of working to integrate HPC systems into the ALICE Grid facility, such as Titan at ORNL and Edison, Cori, Perlmutter at NERSC. Such efforts at ORNL were not 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 the NERSC Cori system was added to the ALICE grid with modest use beginning in 2020. This use has been further expanded onto the current NERSC machine, Perlmutter.

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 multi-year periods of operation, separated by two(three)-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. LHC is currently in the Run 3 which began in July 2022. For the ALICE and LHCb experiments Run 4 marks the beginning of the high luminosity LHC era with data rates from the detectors reaching a up to 100 times larger than those of Run 2.

To deal with such a drastic increase in data readout ALICE reinvisioned its software model into the new Online-Offline (O2) software. With Run 3 ALICE switched to the so-called triggerless/streaming data readout mode. To handle data readout in this mode O2 software is running to produce highly filtered CTF data in the online/synchronous regime during data taking. This synchronous processing runs on powerful Event Processing Nodes (EPN) that are fed data from the First Level Processing (FLP) farm using ~ 1 TB/s bandwidth. Near-online calibration reconstruction produces data that are consequently used during the offline/asynchronous processing of the collected data.

In addition to the raw/CTF data, a comparable volume of MC simulation data, used to evaluate measurement efficiencies and systematic uncertainties required with each data set, are produced and stored on the ALICE Grid facility. These quantities are translated into the global CPU and storage requirements from ALICE from which the US 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 asynchronous processing produces a refined set of quantities known as Analysis Object Data (AO2D), used in most end-user analyses. For the proton-proton collision data taken in Run 3, event skimming of the datasets is performed on the AO2D datasets based on the physics analysis goals. Based on this additional selection, only CTF files associated with the skimmed pp datasets are stored for archival. Conversely, the entire CTF data are stored for the HI runs.

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 SE. Raw event data are 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 over 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 2nd 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 AO2D 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) & 2) above, long-term custodial storage of data and additional reconstruction passes. An overview of the ALICE Grid is shown in Figure 6.7.2: a world map of the ALICE Grid on which each dot is a grid sites and each line represents data transfer at the moment the image was made.

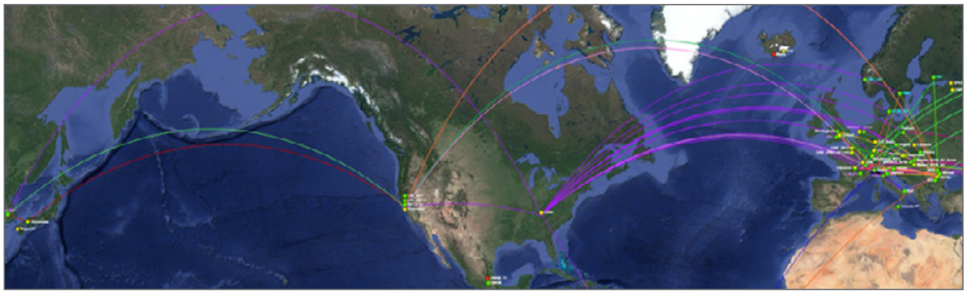


Figure 6.7.2: The ALICE Grid facility. Green dots represent sites while the lines represent data transfers occurring at the time the snapshot was taken. US sites are LBL_HPCS, HPCS_Lr, LBL_AFP and Perlmutter at LBNL and ORNL at ORNL.

The key feature for data management and access on the ALICE Grid is the distribution of the data onto the grid at the data-creation. The data are then subsequently accessed only from local site storage[20]. 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 about 200 PB 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 2.3 XB and written over 400 PB of data from/to the local storage, averaging about 13 GB/s and 80 GB/s for write and read traffic respectively averaged over the entire grid. Over the same period, conversely, ALICE jobs have read just 30 PB and written 3.4 PB of data over the WAN, averaging just 1 GB/s of aggregate bandwidth.

6.7.2.2 Collaborators

The ALICE Collaboration consists of over 2000 scientists, engineers and students spread over 170 institutions in 41 countries. The ALICE VO is a worldwide virtual organization for use by ALICE scientists interacting with Grid organizations such as the WLCG and the OSG in the US. The registry of members including information about roles with respect to computing activities is maintained in the Virtual Organization Management and

[20] The model can fall back to pull a copy from a nonlocal resource, but this is done at the 5% level.

Registration Service (VOMRS) by the WLCG at CERN. WLCG also maintains accounting and reporting for ALICE and other LHC experiments' computing resources in Computing Resource Information Catalog (CRIC). The Virtual Organization (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 with Irakli Chakaberia (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 (Chakaberia) and ALICE representatives from each of the two sites: K. Read, P. Eby, and S. Moulton at ORNL and J. White, K. Fernsler, J. Porter and I.Chakaberia at LBNL. The group holds biannual in-person meetings that include participation from the ALICE Offline 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. A new system, referred to internally as the Hyperloop system, was developed for the Run 3 data analysis, which can leverage multi-core processing, and is being phased in to replace the older train system. In either case, refined analysis results are stored on the Grid but logically linked to the scientist's personal allocations.

LBNL also coordinates the US share [eight hours in convenient to US time zone five days a week] of the Hyperloop shifts dedicated to support of the job submission and management for the individual scientists to the new job management system.

Those results are typically small and can be copied to personal compute 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 reflected in Table 6.7.1.

User/ Collaborator and Location	Do they store a primary or secondary copy of the data?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of dataset? (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)?
TIER-0 @ CERN	Primary	ALICE Grid	~50 PB CTF/ raw ~5 PB AO2D	Immediately after creation, unless occupied by other data, in which case data kept at O2 buffers until buffers are needed for data taking.	N	N
7 TIER-1 SITES GERMANY, FRANCE, ITALY, UK, NETHERLANDS, KOREA, NORDIC SITES	Primary & secondary	ALICE Grid	~50 PB CTF ~5 PB AO2D ~1 PB MC	At creation. Data is routinely redistributed as needed.	Y - during redistribution of data on the ALICE grid	N
OVER 70 TIER-2 SITES	Secondary	ALICE Grid	~5 PB AO2D ~5 PB MC	At creation. Data is routinely redistributed as needed.	Y - during redistribution of data on the ALICE grid	N
ANALYSIS FACILITIES	Secondary	ALICE Grid	~5 PB AO2D	At creation Redistributed as requested by the PWGs.	Y - during redistribution of data on the ALICE grid	N

Table 6.7.1: Collaborative Data Mobility

6.7.2.3 Use of 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 Korea Institute of Science and Technology Information (KISTI) in Korea. That cluster of resources in Europe is illustrated in the map shown in Figure 6.7.2. The two ALICE-USA sites, ORNL and LBNL/HPCS, are operated as T2 centers and as such, do not participate in processing of raw data, however this approach could be revisited due to HPC availability at US sites that are part of the ALICE grid. The raw data reconstruction passes at the T0 and T1 sites noted in Table 6.7.2 produce AO2D 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 mode for the ALICE-USA sites is: 1) receive a fraction of ALICE AO2D 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 US sites.

Task	Activity	Location	Input & source	Output & destination
SYNCHRONOUS RECONSTRUCTION	During data-taking	EPN nodes near experiment	Detector readout	CTF (raw) data
RAW DATA RECONSTRUCTION	Organized & managed	T0 & T1	CTF/raw data Experiment or Tape	AO2D 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	AOD/AO2D from local SE	User ROOT files to ALICE Grid SE + Copied off by hand
USER ANALYSIS ON THE GRID	Chaotic	T1 & T2	AOD/AO2D from local SE	User ROOT files to ALICE Grid SE + Copied off by hand

Table 6.7.2: Types of processing carried out by ALICE scientists on the ALICE Grid and EPNs.

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 6.7.2. The fraction of the ALICE grid CPU resources used by each type of processing over the past year for the entire grid and for the US T2s is shown in Figure 6.7.3. With the exception of raw data reconstruction, all resources are available for use by all types of processing tasks allowing relatively high grid utilization even while the mix of job types fluctuates. (Note: The drop seen in utilization during 2003 is due to ongoing work tuning Run 3 MC simulations. This type of elastic resource usage would be lost if facilities are restricted to host only a single type of task.

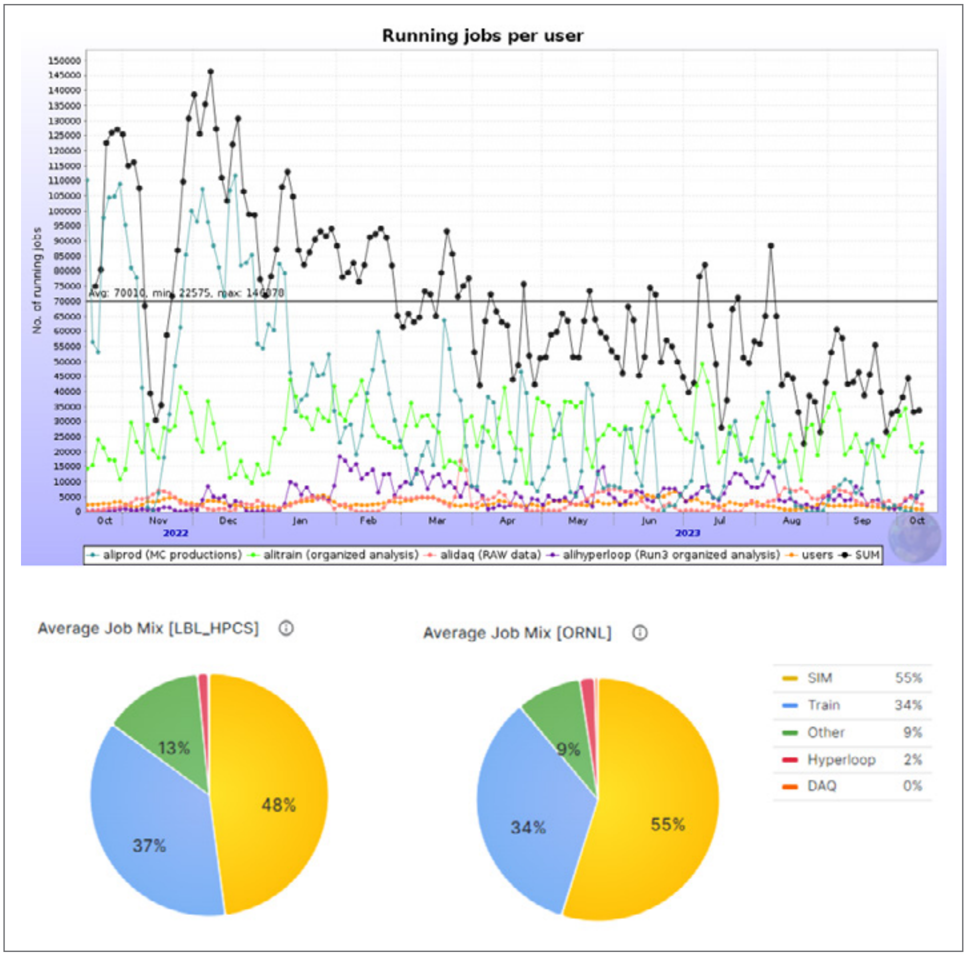


Figure 6.7.3: The overall job-mix on the ALICE Grid facility (top) and job-mix on the ALICE-USA T2 sites(bottom)

To demonstrate the stability of the total number of jobs across the ALICE grid, the total number of CPU cores in use is plotted on Figure 6.7.4. This plot factors out the fact that ALICE has started increasingly running multi-core jobs on the grid (so the number of jobs in this case is not the correct measure of the CPU capacity utilization) and the three-year range in Figure 6.7.4 clearly demonstrated the effect of the lack of MC jobs in 2023.

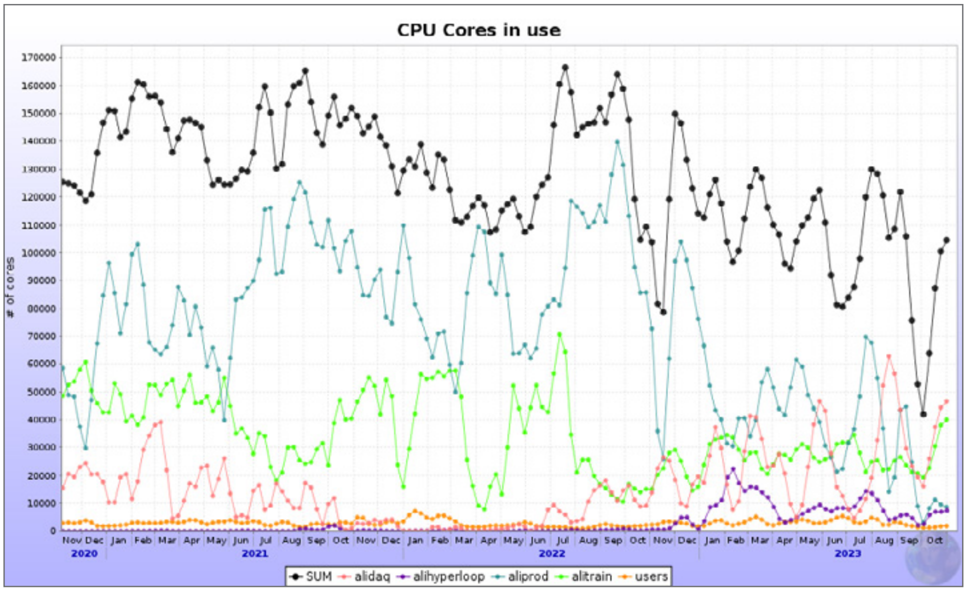


Figure 6.7.4: The overall CPU utilization on the entire ALICE the ALICE Grid facility over the last three years.

The CPU and disk capacity of the ALICE-USA facility are currently about 4600 CPU cores and 8 PB of disk storage, split between the two sites. The network activity produced by the computing model is shown in Figure 6.7.5, in which the rates data are written to (top plot) and read from (bottom plot) the LBNL/HPCS 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 2400 ALICE jobs during the period when the plots were generated. The peaks correspond to periods with large fractions of analysis jobs which require ~3 MB/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.

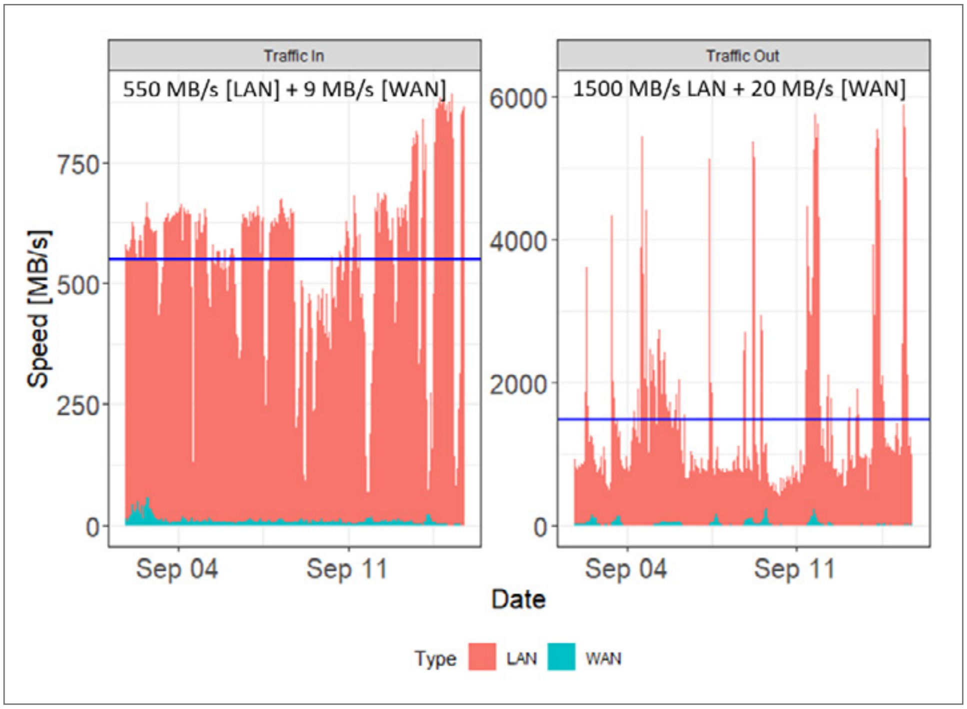


Figure 6.7.5: Traffic written to (left) and read from (right) the LBNL/HPCS storage servers, in September of 2023 separately for WAN and LAN traffic.

In addition to the bandwidth characteristics of normal operations as shown in Figure 6.7.5, 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 6.7.6 when newly added storage at LBNL/HPCS 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.

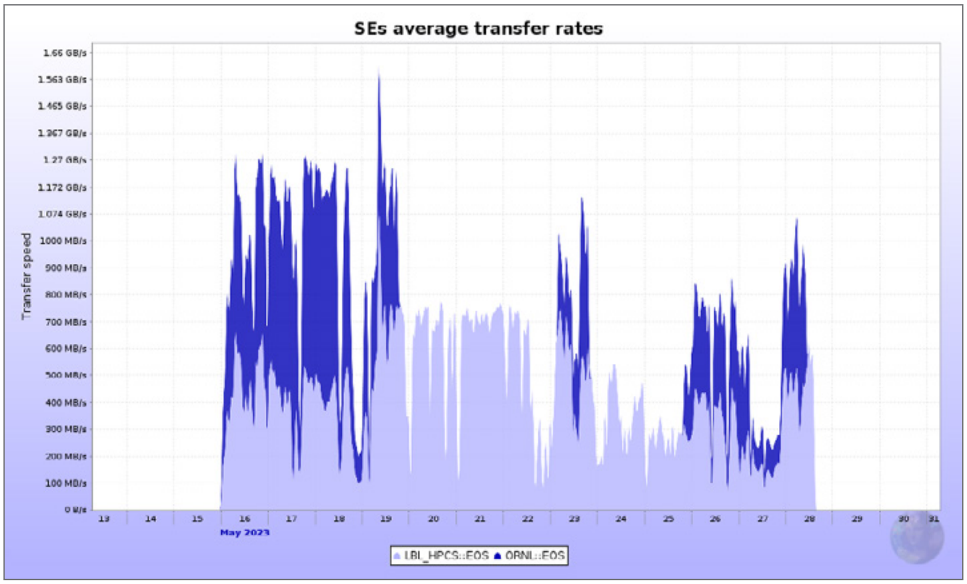


Figure 6.7.6: Network (WAN) utilization during the period of data redistribution on the ALICE Grid.

Under the scenario described, the expected CPU, disk, network and computing model for the periods requested in the case study are shown in Table 6.7.3. With Run 3 in addition to the new ALICE O2 software model the use of GPUs for the asynchronous data reconstruction has been implemented. This gives about a 2.5 increase in the reconstruction speed. We have recently tested offloading CTF reconstruction to the US HPCs either streaming data directly from the source (usually at CERN) or reading from the local T2 storage. For the two- to five-year period we plan to offload some of the GPU-based reconstruction onto the US HPCs as well, which will demand more of the network due to the faster turnaround times from leveraging the HPC GPUs for data processing. ALICE also foresees more aggressive use of HPC facilities for ML/AI technique-based payloads. Even with such significant changes to the overall ALICE computing model and capabilities, the resource obligation to ALICE from the US should remain at a steady growth in CPU, disk and network capacities.

Period	CPU (cores)	Disk (PBs)	Aggregate LAN BW (GB/s)	Aggregate WAN BD (GB/s)	Grid-enabled resource types and conf.
0-2 YEARS	10k - 15k	14-16	10-20	4	Conventional clusters, disk storage, and HPC, Analysis Facility
2-5 YEARS	15-22k	18-25	40	4-10	GPU/Accelerators, ML technique
5+ YEARS	22k+	25+	40+	10+	...

Table 6.7.2: Expected ALICE-USA computing resources for the periods requested by this case study

6.7.2.4 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 6.7.2 and these will continue over the next zero to two years. The detector streams the readout data to the EPN which in an online regime (synchronous processing) process and package it into CTF output. Almost concurrently (with a delay on the order of hours) this “raw” CTF data are used for detector calibrations. The calibration data are used to run offline reconstruction (asynchronous) over the CTF data. Two such asynchronous passes are envisioned, but more may be required. All of the above is managed by the central team to produce data files (AO2D) 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 are the priority given to each task. In LHC Run 3 ALICE transitioned to the new software model. The goal of that new model, referred to as O2 (Online/Offline combined), is to more efficiently process and reduce the large amounts of data coming off the detector. In addition, after this transition the additional processing of Run 1 and 2 data was done to convert old AOD into the new AO2D datasets. With this done ALICE will entirely move into the unified job submission system - hyperloop. Hyperloop job management infrastructure was developed to cope with a drastic increase of Pb-Pb collision data in Run 3 compared to the Run 1 and 2 data-taking periods. This infrastructure replaces the old Lightweight Environment for Grid Operators (LEGO) train submission system. But until then there is a mix of old and new analysis jobs on the grid.

The two- to five-year period covers the LHC Run 3 when the new ALICE computing model is being further tuned for event size and data compression from each detector. Additional dedicated efforts will be made to extend GPU based asynchronous reconstruction to the grid computing sites, this is currently possible only at CERN T0 and the online EPN farm. The growing use of ML and AI techniques for scientific research makes available GPU systems on HPC resources more valuable and may necessitate delivery of the data closer to such resources. With the increase of the accumulated data the fraction of the data analysis jobs on the grid compared to other types of jobs (e.g., MC production) is increasing. In this scenario the expected use of dedicated ALICE Analysis Facilities (AFs), which are optimized for the I/O intensive analysis jobs, is increasing. The ALICE AFs are expected to support fast turn-around analysis iterations on significant subsets of AO2D data using the new Hyperloop processing system.

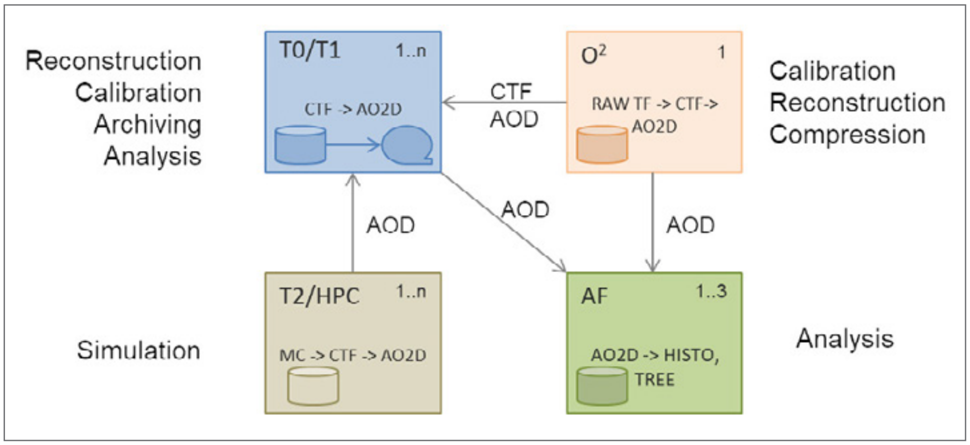


Figure 6.7.7: ALICE O2 Computing model concept consisting of a single O2 facility at the experiment site, many Grid sites (T0/T1/T2s) around the world and a handful of AFs.

An illustration of the new ALICE O2 computing model is shown in Figure 6.7.7 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, with a summed capacity expected for Run 3 of 10s of thousands of CPU cores and 10s 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 are three AFs operating at GSI, Germany, Wigner Computing Center, Hungary and a recently installed one at LBL, US.

6.7.2.5 Remote Science Activities

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 throughout this case study.

6.7.2.6 Software Requirements

There are two distinct types of ALICE software infrastructure: 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 will be described first followed by the scientific analysis framework.

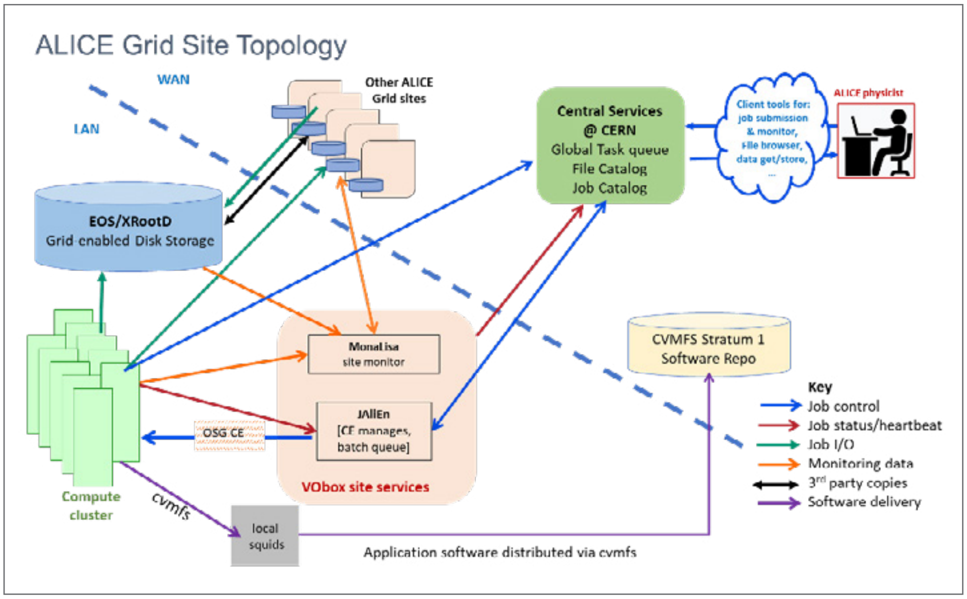


Figure 6.7.8: ALICE Grid site topology and connection to the larger ALICE Grid.

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 6.7.8. Each site runs an ALICE VO box, 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 6.7.8 are listed in Table 6.7.4 with a brief description of their operation. The ALICE Grid facility will continue operations during the next zero to two years and beyond. In recent years most of the ALICE grid services, called AliEn, were moved from a mix of Perl, Java and C to primarily a Java based system, JAliEn. The overall functionality of the ALICE Grid and the types of site services depicted in Figure 6.7.8 and listed in Table 6.7.4 are expected to be supported in both the two- to five-year timeframe and beyond in the five+ year time frame.

Task	Activity	Location	Input & source
SOFTWARE DEPLOYMENT	CVMFS	compute node cache + squid access to software repo	CERN solution for distributed software delivery
SITE MONITORING	MonALISA	VOBox server	Monitoring data is stored locally, and then is replicated to a central data collection at CERN
PROCESS MONITORING	MonALISA probes	SE, with jobs on VOBox	Processes probe data and sends to local ML server
JOB MANAGEMENT	JAliEn	VOBox server	Evaluated existing resources, matches them with available jobs and launches the jobs
LOCAL QUEUE	SLURM, OSG HTCondor	Compute cluster	ALICE pilot jobs are submitted to the batch system of the cluster
JOB PROCESSING	JObAgent	Worker nodes	JobAgent actually runs the ALICE job payload
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 6.7.4: ALICE Grid service types and the software used to perform the services.

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. With the recent upgrade to the O2 software the new approaches to data compression are being investigated. In the zero- to two-year time frame the optimal compression should be solidified to both storage optimization but also for the memory use for the data processing. The changes with triggerless data collection with synchronous processing and the new O2 software infrastructure in Run 3 dominate future computing landscape of ALICE, such that the five+ year time frame is expected to largely be an extension of the Run 3 O2 computing model, with optimization based on experience we are gaining with the ongoing Run.

6.7.2.7 Network and Data Architecture Requirements

ALICE ORNL

The ALICE T2 site at ORNL currently 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. Plans are underway this year to transition the ALICE environment to connect directly to the ORNL border routers at a peer level with the Science DMZ.

The ALICE network environment is in the process of being upgraded to Arista 7280SR3 core switches connected at 40G, and Arista 7010TX switches for management connectivity. The upgraded ORNL ALICE environment expected to be completed early in 2024 is depicted below.

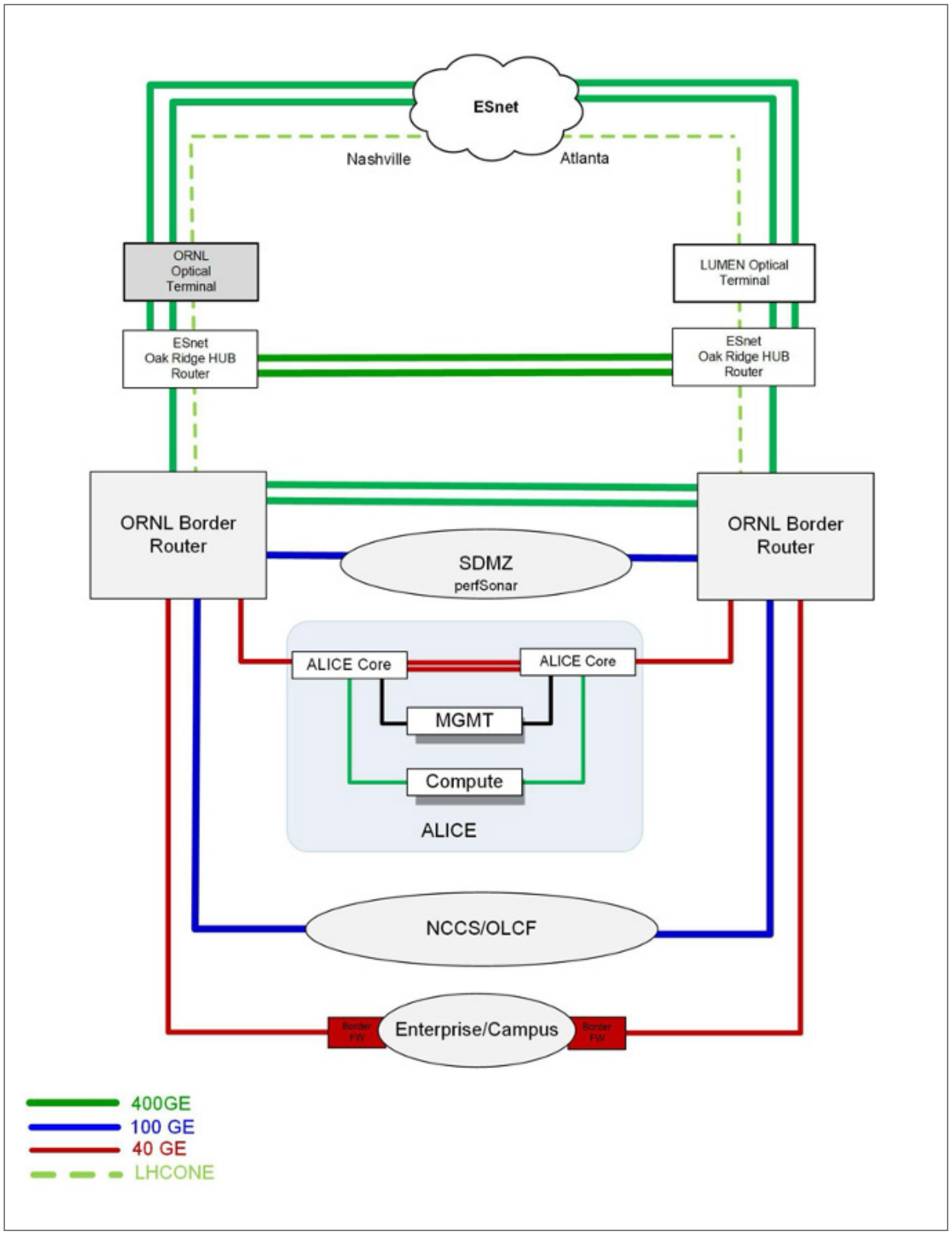


Figure 6.7.9: Network layout of the ALICE T2 facility at ORNL relative to ESnet peering

Beyond 2024 ORNL connections to ESnet are expected to migrate proportionate with ESnet backbone speeds. With direct connections to the ORNL border routers and upgrades to the ALICE environment as needed, bandwidth to collaborators should not be an issue.

ALICE LBNL

The current network topology of the ALICE T2 facility at the LBNL HPCS center is shown in Figure 6.7.10. Internal connection between the worker nodes and the storage is done over 56 Gb 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 LBLnet connection to ESnet. The SEs were recently added to the LHCONE via a 10 Gb firewall. Our ALICE-USA computing project plan calls for adding a perfSONAR service installed on the same external route at the storage. This requirement is expected to be fulfilled within next year. An item for the longer (two to five year) term is to understand and if necessary, optimize the network connectivity between the storage at the HPCS facility and Perlmutter HPC at NERSC, in order to run analysis jobs and more importantly ALICE CTF reconstruction on this site.

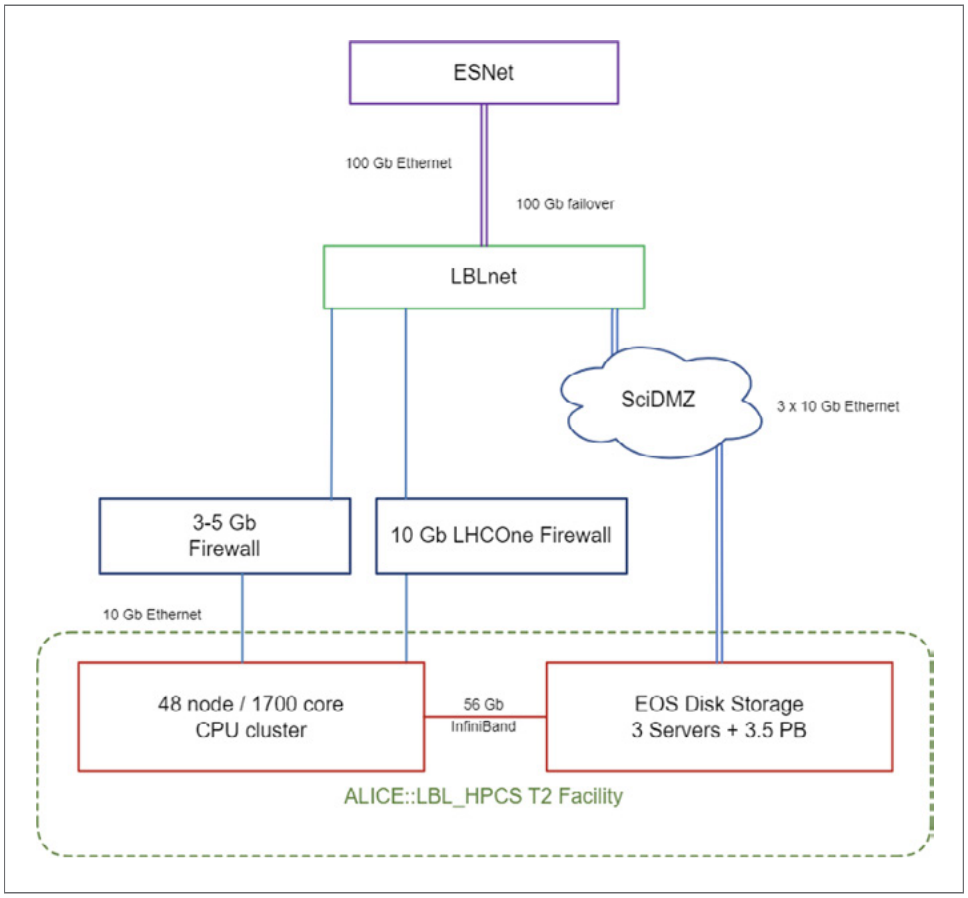


Figure 6.7.10: Network layout of the ALICE T2 facility at LBNL/HPCS relative to ESnet peering.

ALICE NERSC

We have successfully moved our workflow from the NERSC's previous Cori HPC to the new Perlmutter HPC. Job submission to Perlmutter is done using the NERSC's SuperFacility API as accessed by the ALICE Compute Element (CE) service that is running on an external VOBox to manage the job flow on the NERSC system. While Cori was mostly used as the R&D platform, the improvement in the network for the Perlmutter allowed us to use it as an additional resource for the ALICE grid. In the next few years, we will continue the stable use of this resource and may use Perlmutter as an additional pledged CPU resource, which will allow us to increase the storage capacity at each of our Tier-2 facilities.

6.7.2.8 IRI Readiness

JAliEn software is very flexible and could be adapted to different scenarios. As mentioned above we have added support in JAliEn for using the NERSC SuperFacility API to be able to leverage Perlmutter resources. If other HPC sites were to adopt this API and satisfy minimum requirements, such as outgoing network connectivity from worker nodes and availability of CVMFS, ALICE could in principle deploy those resources as an ALICE grid site. Specifically, providing API for scheduling and data management as in case of Perlmutter to other HPC facilities would greatly simplify the process of including them into the common workflow.

ALICE can certainly benefit from additional supercomputing resources especially with GPU capabilities.

6.7.2.9 Use of Cloud Services

The current and developing ALICE computing models do not have any specific plans for use of cloud resources. However, user analysis within commercial clouds is a possibility if it is cost effective. There is an experience of using a cloud resource which was 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 6.7.3, 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.

6.7.2.10 Data-Related Resource Constraints

One specific data-related constraint that we faced with Cori was how to efficiently use our NERSC HPC allocation for data analysis with minimal impact on the ALICE Grid model. ALICE does not operate a Grid-enabled storage element inside NERSC, which posed an issue running analysis jobs that are I/O heavy. With the switch to Perlmutter the initial estimations show that this bottleneck is less constraining allowing us to leverage the proximity of the LBNL T2 SE to use it as an input source to jobs that require data. That is, the network connectivity between NERSC and our LBNL site allows us to operate NERSC HPC resources as an ALICE Grid site without a local SE at NERSC. We are currently increasing the analysis job load and plan to even run data reconstruction from the CTF files where data are streamed from nearby SEs, potentially even from our ORNL T2 site. This will help us evaluate the current network capabilities. Understanding the network limits now (zero to two years) and future (two to five year) capacities may help guide the development of ALICE use of NERSC HPC in the era beyond Perlmutter.

6.7.2.11 Data Mobility Endpoints

Data placement and mobility is fully encapsulated within the ALICE Grid infrastructure and is largely automated. As such, all storage used within this infrastructure has to be provisioned specifically for ALICE, have multi-year lifespans, and be managed using one of two community-supported storage services, CERN EOS or XRootD. For this reason, and as mentioned in [Section 6.7.2.9 and Section 6.7.2.10], general use of "external" compute resources like cloud or HPC by ALICE is limited by the network access between those compute resources and ALICE permanent storage elements. Using the aggregate CPU capacities and LAN bandwidths in Table 6.7.3 gives an approximate I/O capability of 2 to 3 MB/s per CPU core between the external CPU and ALICE storages.

6.7.2.12 Outstanding Issues

As presented in the previous case study, systematic automated WAN monitoring is very important for efficient use of ALICE-USA resources. This goes beyond any one specific site-to-site monitor as it should seamlessly include all paths between any two ALICE Grid sites. ALICE has a simple, yet effective, monitor using traceroute between every VOBox, that can be used in near-real time to adjust automated data placement, but which does not provide any real diagnostic capabilities. We previously noted our plan to leverage the OSG network group's perfSONAR dashboards. That plan has not been implemented, partially due to the reduced active collaboration between ALICE Grid sites during COVID, but remains our current best model for tackling this need. The project, however, is interested in any feedback regarding future monitoring capabilities from ESnet or other service providers.

6.7.2.13 Facility Profile Contributors

ALICE Representation

- John Arrington, LBNL, jarrington@lbl.gov
- Latchezar Betev, European Organization for Nuclear Research (CERN), Latchezar.Betev@cern.ch
- Irakli Chakaberia, LBNL, iraklic@lbl.gov
- Pete Eby, ORNL, ebypi@ornl.gov
- Thomas Evans, ORNL, evanstm@ornl.gov
- Costin Grigoras, European Organization for Nuclear Research (CERN), Costin.Grigoras@cern.ch
- Reiner Kruecken, LBNL, rkruecken@lbl.gov
- Mateusz Ploskon, LBNL, mploskon@lbl.gov
- Jeff Porter, LBNL, rjporter@lbl.gov
- Kenneth Read, ORNL, readkf@ornl.gov
- John White, LBNL, JWhite@lbl.gov

ESCC Representation

- Susan Hicks, ORNL, hicksse@ornl.gov
- Richard Simon, LBNL, rsimon@lbl.gov
- Rune Stromsness, LBNL, rstrom@lbl.gov

6.8 The CMS Heavy Ion Experimentation and ACCRE

The CMS is a general-purpose detector at the LHC. It has a broad physics program ranging from studying the Standard Model (including the Higgs boson) to searching for extra dimensions and particles that could make up dark matter. The CMS experiment is one of the largest international scientific collaborations in history, involving about 5500 particle physicists, engineers, technicians, students and support staff from 241 institutes in 54 countries (as of May 2022).

For a few millionths of a second, shortly after the Big Bang, the universe was filled with an astonishingly hot, dense soup made of all kinds of particles moving at near light speed. This mixture was dominated by quarks – fundamental bits of matter – and by gluons, carriers of the strong force that normally “glue” quarks together into familiar protons and neutrons and other species. In those first evanescent moments of extreme temperature, however, quarks and gluons were bound only weakly, free to move on their own in what’s called a QGP.

To recreate conditions similar to those of the very early universe, powerful accelerators make head-on collisions between massive ions, such as gold or lead nuclei. In these heavy-ion collisions the hundreds of protons and neutrons in two such nuclei smash into one another at energies of upwards of a few trillion electronvolts each. This forms a miniscule fireball in which everything “melts” into a QGP. The LHC performs a once-a-year run with heavy ions, and the CMS detector is used to observe the results of the scientific process.

6.8.1 Discussion Summary

The following bullets were extracted from the case study, and in-person discussion. Several appear in Section 1 and Section 3.

- The heavy ion program using the CMS detector at the LHC probes the conditions of the early universe by studying collisions of heavy nuclei at relativistic speeds (usually between two heavy nuclei, though some data were recorded of protons striking lead nuclei).
- CMS-HI data are recorded when the LHC performs a Heavy-Ion run, where instead of colliding protons, the LHC collides heavier nuclei such as lead, xenon or oxygen. In theory, these data-taking periods occur in the final four weeks of the LHC's running year (which typically ends Oct or Nov), though this is subject to change due to external factors.
- CMS-HI has three primary data flows. The data are created at CERN in three different data tiers: RAW, AOD, and Mini-AOD. The volumes of these tiers, in an optimistic scenario where the accelerator performs well for the five-week run, is 27 PB RAW, 17 PB AOD, and 3 PB Mini-AOD.
- All CMS-HI data are transferred to Fermilab for archival storage, primarily limited by available tape bandwidth. The Mini-AOD storage tier is additionally transmitted to Vanderbilt for access by physicists. Some users then generate even smaller n-tuples from this Mini-AOD , which are primarily transmitted to MIT for interactive analysis.
- The annual data volumes for CMS-HI are approximately 5x what they were in 2018 (e.g., 40PB/yr at the T0 and T1, and 3 to 10PB/yr at ACCRE). An additional increase by a factor of two in HL-LHC is expected (tentatively scheduled for 2029).
- CMS (and US CMS) support a few different ways to access data. Users will typically begin an analysis workflow using CRAB, where they provide an executable and a desired input dataset, CRAB then will split the dataset into job-sized subdivisions and run a batch job for each piece. Generally, users use CRAB for data reduction and skimming by processing the Mini-AOD as the input and producing an analysis-specific n-tuple. CMS also supports direct batch submission either to the universities themselves or to the global CMS batch queue (which spans all CMS sites). In the final stages with smaller data needs, users will use interactive command-line tools and (increasingly) Jupyter Notebook.
- CMS supports two main data access interfaces - POSIX mounts for local file access (e.g., if a job at Vanderbilt needs to access data stored at Vanderbilt), or remote XRootD access via the AAA data federation, which federates CMS' global XRootD access points into a single access point.
- For CMS-HI, the primary limitation on storage is the tape bandwidth available to store and recall large multi-PByte datasets. While the capacity of tape cartridges greatly increases when new generations of technology are released, the bandwidth per-tape has not kept pace. If trends continue, in 2030 it could take nearly a day of continuous access to read/write a tape from beginning to end. This presents significant issues operationally. Writing to and retrieving from custodial tape is by far the most difficult issue for data movement CMS-HI will face in the coming years.
- One limiting factor for data volumes in CMS-HI Run3 is the cost of tape archival and disk storage. Additionally, the rate at which tape archival can occur is limited by the number of available tape drives at Fermilab. CMS currently records at the maximum rate possible from the detector. When the detector is upgraded in '29, the increase in detector channels will increase the maximum rate possible, which exacerbates these resource constraints.

- The largest sources/sinks of CMS-HI data are the XRootD endpoints at CERN, Fermilab, Vanderbilt, and MIT. After their initial production, very little data movement happens on the large centrally-produced datasets. There is approximately TB-scale data movement of user-produced datasets to other facilities. This trend is expected to be stable until 2029 when the HL-LHC program begins. With the increased detector granularity and data acquisition rate, these numbers can scale between 5x and 10x of today's numbers.
- CERN has, and is expected to continue to have, sufficient resources to produce smaller derived datasets for CMS-HI, and no need for DOE SC user facilities are foreseen. (this is a notable difference between CMS-HI and CMS-PP).
- An identified issue with the CMS-HI data reduction pattern is that some popular portions of datasets are reproduced many times. For example, a user might directly copy the Mini-AOD dataset's electron collections to their private n-tuples. At the predicted scales of HL-LHC (predicted to begin in 2029), this data duplication becomes financially burdensome to support. By storing data in object stores, these common slices of datasets could be stored and referenced by multiple end-user datasets, providing a better space efficiency for analysis. Of course, these benefits become more pronounced if there is a single copy globally of the relevant objects, so this implies some level of additional WAN traffic needed to satisfy these workflows. These techniques are the target of active R&D and initial results are expected in the 2025-2026 timescale.
- CMS-HI expects an increase in remote science usage, namely accessing GPU resources, via the SONIC inference-as-a-service activity. This R&D is still in its infancy and there is some uncertainty how the service would be deployed (if SONIC is deployed). This research is ongoing and a decision on the deployment would be made in the 2026-2027 time-frame.
- CMS is also participating in SENSE-Rucio R&D, which will use ESnet 6's ability to provide guaranteed point-to-point bandwidth to more effectively schedule Rucio data transfers. With SENSE-Rucio, Rucio can decide to signal to the SENSE dataplane that it would like guaranteed bandwidth between two sites, and if the request is accepted, configure transfers for specific dataset(s) to transit exclusively over that guaranteed bandwidth. There is a similar effort in the NOTED project, using a technology called "packet marking" to provide QoS guarantees. All of these projects are under active development, and CMS will decide on the timescale of 2026-2027 whether or not to move forward with these technologies at scale.

6.8.2 CMS-HI & ACCRE Experimental Case Study

The heavy ion program at CMS probes the conditions of the early universe by studying collisions of heavy nuclei at relativistic speeds (usually between two heavy nuclei, though some data were recorded of protons striking lead nuclei). At the extreme temperatures within these collisions, quarks and gluons which are normally tightly bound into particles like protons and neutrons "melt" together, producing a state of matter known as the QGP. After its creation, the QGP expands and cools, by measuring the behavior of the QGP as it cools, researchers are able to explore the properties of the very universe as well as verify our understanding of the fundamental behavior of quarks and gluons.

6.8.2.1 Science Background

The data are recorded when the LHC performs a Heavy-Ion run, where instead of colliding protons, they collide heavier nuclei such as lead, xenon or oxygen. In theory, these data-taking periods occur in the final four weeks of the LHC's running year (which typically ends Oct or Nov), though this is subject to change due to external factors. For instance, the sharp increase in European energy prices in '22 due to the war in Ukraine caused an early termination of the LHC run, meaning no Heavy-Ion data were recorded. To somewhat compensate, the four-week run in '23 was extended to five weeks, leading to a 25% increase in predicted data volumes (though issues with the ongoing '23 run have greatly reduced the volume recoded by the detector).

There are three primary data flows. The data are created at CERN in three different data tiers: RAW, AOD, and Mini-AOD. The volumes of these tiers, in an optimistic scenario where the accelerator performs well for the five-week run, is 27 PB RAW, 17 PB AOD, and 3 PB Mini-AOD. All of these data are transferred to Fermilab for archival storage, primarily limited by available tape bandwidth. The Mini-AOD storage tier is additionally transmitted to Vanderbilt for access by physicists. Some users then generate even smaller ntuples from this Mini-AOD, which are primarily transmitted to MIT for interactive analysis.

6.8.2.2 Collaborators

This list of collaborators is relatively stable, though the annual data volumes are approximately 5x what they were in 2018. We expect an additional increase by a factor of two in HL-LHC (tentatively scheduled for 2029)

User/Collaborator and Location	Do they store a primary or secondary copy of the data?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of dataset? (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	Primary Tape Archive	XRootD	40PB/yr	ad hoc (< 4 weeks)	N	N
FERMILAB	Secondary Tape Archive	XRootD	40PB/yr	ad hoc (< 4 weeks)	N	N
ACCRE, VANDERBILT	Secondary	XRootD	3-10PB/yr	ad hoc (< 4 weeks)	N	N
MIT	Secondary	XRootD	100s of TB/yr	ad hoc (< 4 weeks)	N	N

Table 6.8.1: Collaborative Data Mobility

6.8.2.3 Use of Instruments and Facilities

The data are produced at CERN. There are three primary data flows. The data are created at CERN in three different data tiers: RAW, AOD, and Mini-AOD. The volumes of these tiers, in an optimistic scenario where the accelerator performs well for the five-week run, is 27 PB RAW, 17 PB AOD, and 3 PB Mini-AOD. All of these data are transferred to Fermilab for archival storage, primarily limited by available tape bandwidth. The Mini-AOD storage tier is additionally transmitted to Vanderbilt for access by physicists. Some users then generate even smaller ntuples from this Mini-AOD, which are primarily transmitted to MIT for interactive analysis.

We expect the following volumes over different timescales:

- Two years, as above
- Two to five years: Beginning in 2026, there will be no new data recorded until 2029
- Five+ years: Annual data taking will resume in 2029 at a factor of 2 greater than today

6.8.2.4 Process of Science

The CMS-HI workflow is conceptually straightforward. Raw data are recorded by the detector at CERN and promptly reconstructed into the AOD format at CERN's computing facilities. This is possible because the Heavy-Ion runs are at the end of running years (for various reasons), which means that after the run completes, 10s of thousands of CPUs normally used to trigger and capture the raw data become idle. An additional processing step converts this AOD data into Mini-AOD data, which both makes the data into a more convenient-to-process format, but also performs a 5x volume reduction on the data. The expectation in 2023 is that analysis happens exclusively on this Mini-AOD data tier, which means the AOD will not need to be on disk, it will simply be stored on tape as an input to a future re-production of the Mini-AOD data. All three data tiers are transmitted asynchronously to Fermilab for archival storage via XRootD, FTS and Rucio data-management tools. Once the data are successfully written to custodial tape storage, the disk copies are removed.

An additional copy of the Mini-AOD is stored on disk at Vanderbilt, which is the primary facility for heavy ion analyses. Once the data are at Vanderbilt, there are a few different tools and access methods users can use to process the data.

CMS (and US CMS) support a few different ways to access data. Users will typically begin an analysis workflow using CRAB, where they provide an executable and a desired input dataset, CRAB then will split the dataset into job-sized subdivisions and run a batch job for each piece. Generally, users use CRAB for data reduction and skimming by processing the Mini-AOD as the input and producing an analysis-specific ntuple that has their interested features. CMS also supports direct batch submission either to the universities themselves or to the global CMS batch queue (which spans all CMS sites). In the final stages with smaller data needs, users will use interactive command-line tools and (increasingly) Jupyter Notebook.

In all cases, CMS supports two main data access interfaces - POSIX mounts for local file access (e.g., if a job at Vanderbilt needs to access data stored at Vanderbilt), or remote XRootD access via the AAA data federation, which federates CMS' global XRootD access points into a single access point.

We foresee the above access methods to remain into the far future, but one additional access method is being investigated by CMS – the use of object stores. Once issue with the data reduction pattern above is that some popular portions of datasets are reproduced many times. For example, a user might directly copy the Mini-AOD dataset's electron collections to their private ntuples. At the predicted scales of HL-LHC (predicted to begin in 2029), this data duplication becomes financially burdensome to support. By storing data in object stores, these common slices of datasets could be stored and referenced by multiple end-user datasets, providing a better space efficiency for analysis. Of course, these benefits become more pronounced if there is a single copy globally of the relevant objects, so this implies some level of additional WAN traffic needed to satisfy these workflows. These techniques are the target of active R&D and we expect initial results in the 2025-2026 timescale.

6.8.2.5 Remote Science Activities

The largest sources/sinks of data are the XRootD endpoints at CERN, Fermilab, Vanderbilt, and MIT. After their initial production, very little (if any) data movement happens on the large centrally produced datasets. There is some O(100 TB/yr) movement of user-produced datasets to other facilities, though these are small enough that we generally don't account for the movement in our modeling. We expect this trend to be stable until 2029 when the HL-LHC program begins, and with the increased detector granularity and data acquisition rate, these numbers can scale between 5x and 10x of today's numbers.

The only other predicted large remote science usage foreseen is the possibility to use large GPU farms to provide inference-as-a-service using the SONIC service. This R&D is still in its infancy and there is some uncertainty how the service would be deployed (if SONIC is deployed). One model would be large national GPU deployments, but once could also see deploying SONIC instances at each site, removing the need to transmit data over the WAN. This research is ongoing and a decision on the deployment would be made in the 2026-2027 time frame.

6.8.2.6 Software Requirements

CMS is invested heavily in the XRootD-WebDav, FTS, Rucio stack of data management software. These software are all open-source, and CMS provides some effort towards their maintenance. Barring something unforeseen, CMS expects to use this software long-term.

CMS is also participating in SENSE-Rucio R&D, which will use ESnet 6's ability to provide guaranteed point-to-point bandwidth to more effectively schedule Rucio data transfers. With SENSE-Rucio, Rucio can decide to signal to the SENSE dataplane that it would like guaranteed bandwidth between two sites, and if the request is accepted, configure transfers for specific dataset(s) to transit exclusively over that guaranteed bandwidth. There is a similar effort in the NOTED project, using a technology called "packet marking" to provide QoS guarantees. All of these projects are under active development, and CMS will decide on the timescale of 2026-2027 whether or not to move forward with these technologies at scale.

To produce (intermediate) data products, CMS will continue to develop and use CMS Offline Software (CMSSW), which is the framework used for every stage of data manipulation, from the High-Level Trigger at CERN down to end-user analysis. CMSSW is built on a number of other projects, namely ROOT and Intel's Thread-Building Blocks (TBB) (which provides underlying multithreading support). In addition to CMSSW, a lot of effort is going into newer "pythonic" and "columnar" data analysis frameworks, which leverage popular-in-industry tools like numpy, dask, and matplotlib for end user analysis. These efforts are becoming increasingly mature and several AFs are being built around these tools (though none specifically for CMS-HI, as part of the CMS collaboration, HEavy-Ion contributors also can and do use these facilities for their analysis).

6.8.2.7 Network and Data Architecture Requirements

For CMS-HI, our primary limitation is the tape bandwidth available to store and recall large multi-PByte datasets. While the capacity of tape cartridges greatly increases when new generations of technology are released, the bandwidth per-tape have not kept pace. If trends continue, in 2030 it could take nearly a day of continuous access to read/write a tape from beginning to end. This presents significant issues operationally. Writing to and retrieving from custodial tape is by far the most difficult issue for data movement CMS-HI will face in the coming years.

Something that CMS is generally exploring for data processing & analysis is the use of SONIC for GPU-accelerated workflows. This can be considered, broadly, as "GPU-inference-as-a-service", where jobs needing to perform ML inference will transmit their inputs to the SONIC service, which will pass the inputs to a load balanced pool of GPU nodes which will perform the inference and transmit results back to the client. This technology performs well over the WAN, so if it were to be adopted, there would be a need for CMS sites to be able to have high-bandwidth, low-latency links between sites to transmit these data back and forth.

The use of this SONIC service is currently in active R&D, decisions about whether to deploy this at scale will happen within the next two to five years and would be generally deployed in the 2027-time scale.

6.8.2.8 IRI Readiness

CERN has, and is expected to continue to have, sufficient resources to produce smaller derived datasets for CMS-HI, and no need for DOE SC user facilities are foreseen. (this is a notable difference between CMS-HI and CMS-PP).

6.8.2.9 Use of Cloud Services

Nothing to report.

6.8.2.10 Data-Related Resource Constraints

One limiting factor for data volumes in Run3 is the cost of tape archival and disk storage. Additionally, the rate at which tape archival can occur is limited by the number available tape drives at Fermilab. CMS currently records at the maximum rate possible from the detector. When the detector is upgraded in '29, the increase in detector channels will increase the maximum rate possible, which exacerbates these resource constraints.

6.8.2.11 Data Mobility Endpoints

The relevant CMS institutes (CERN, Fermilab, MIT, and ACCRE) exchange data using the Web Distributed Authoring and Versioning (WebDAV) protocol support provided by XRootD. The transfer of datasets occurs with hundreds of simultaneous transfers, with each file typically being 2-10 GByte in size. These transfers are managed by the FTS, https://fts.web.cern.ch/fts/), which manages tuning the number of concurrent transfers and retries in case of errors. Above FTS, the Rucio data-management framework manages the movement of datasets by taking a series of rules requesting the placement of datasets, then triggering FTS transfers for each file that needs to be moved.

6.8.2.12 Outstanding Issues

Nothing to report.

6.8.2.13 Facility Profile Contributors

Vanderbilt Representation

- Andrew Mello, Vanderbilt University, andrew.melo@gmail.com
- Paul Sheldon, Vanderbilt University, paul.sheldon@vanderbilt.edu

6.9 The EIC

Understanding the electromagnetic force between the atomic nucleus and the electrons that orbit it is a critical foundation of modern science. Little is known about the microcosm within the protons and neutrons that make up the atomic nucleus. The EIC will be constructed to look inside the nucleus and its protons and neutrons.

The EIC will be a particle accelerator that collides electrons with protons and nuclei to produce snapshots of those particles' internal structure—like a CT scanner for atoms. The electron beam will reveal the arrangement of the quarks and gluons that make up the protons and neutrons of nuclei. The force that holds quarks together, carried by the gluons, is the strongest force in Nature. The EIC will allow us to study this "strong nuclear force" and the role of gluons in the matter within and all around us. What we learn from the EIC could power the technologies of tomorrow.

6.9.1 Discussion Summary

The following bullets were extracted from the case study, and in-person discussion. Several appear in Section 1 and Section 3.

- The EIC is a new community driven facility that targets the exploration of QCD to high precision, with a particular focus on unraveling the quark-gluon substructure of the nucleon and of nuclei. The EIC will investigate the structure of nucleons and nuclei, and will be accomplished by performing precise measurements of DIS and other processes over the complete relevant kinematic range including the transition region from perturbative to nonperturbative QCD. The EIC users' group has roughly 1400 members from nearly 300 institutions and 40 countries.
- The EIC will be constructed in the 2020s, with an extensive science case as detailed in the EIC white paper, the 2023 NSAC Long Range Plan for Nuclear Science, and the EIC yellow report. The yellow report has been an important input to the successful DOE CD-1 review and decision. It describes the physics case, the resulting detector requirements, and the evolving detector concepts for the experimental program at the EIC. The first scientific collaboration for the EIC, ePIC, was formed in 2023 to support the realization of the EIC project detector.
- The ePIC Collaboration, consisting of almost 500 members from 171 institutions, is working jointly with the EIC Project team to design and establish the ePIC detector, which is poised to be primed and ready for data collection once the EIC springs into action in the early 2030s. Work on a second detector at the EIC is also in progress by a yet unnamed collaboration.
- The EIC facility is projected to begin data collection in 2033. While the guiding details of the ePIC detector, data acquisition systems, and analysis workflows are mostly understood, the final design is still in progress. The next major milestone is a TDR which will include studies of

the detector components. No major networking requirements are anticipated before the next requirements review in 2027, when scale tests of the computing model take place. EIC/ePIC focused R&D on streaming readout and reconstruction frameworks is expected to ramp up in the latter half of this decade to be ready for full production in the 2030s.

- EIC networking and computation decisions are not imminent, and will take into account lessons learned from LHC computing and analysis. It is expected that major collaboration sites will contribute resources, and leverage ESnet connectivity as a critical networking link to manage traffic flows. The ability to handle PB volumes of data over time, and bursts of hundreds of Gbps is expected.
- The effort to estimate data volumes from the EIC's ePIC detector is in progress. Collision parameters, synchrotron radiation, and beam gas backgrounds from both the electron and hadron beams have been studied, but there are continued efforts to ensure that all detectors are included using proper energy thresholds and digitization schemes. Raw data volumes could be hundreds of PBps, with reduction to hundreds of Gbps possible after various forms of filtering is applied.
- BNL and JLab have formed the EIC Computing and Software Joint Institute to serve as a single point of contact and organizational entity for support of the ePIC Collaboration and other software and computing needs for the EIC. Within this structure, theoretical calculations and accelerating modeling will be conducted, also within the current estimates established by the individual BNL and JLab Facility use cases.
- EIC will follow generalized approach to the process of science that is rooted in current NP-based workflows:
  - DAQ system affiliated with the ePIC detector
  - Streaming readout to produce raw datasets
  - Immediate computation to support calibration, alignment
  - Long-term storage and sharing
  - Reconstruction and reprocessing
  - Simulation, and possible integration with digital twins
  - Local and distributed analysis
- The EIC's workflow and resource utilization are still being modeled. It is anticipated that reconstructing the simulated data within the same workflow is preferable, e.g., avoiding a storage-consuming output stage after the simulation, and avoiding the complication of distinct MC simulation/production workflows.

- EIC networking and computation decisions are not imminent, and will take into account lessons learned from LHC computing and analysis. It is expected that major collaboration sites will contribute resources, and leverage ESnet connectivity as a critical networking link to manage traffic flows. The ability to handle PB volumes of data over time, and bursts of hundreds of Gbps is expected.

- Assessment of the computation and storage requirements for the EIC's ePIC detector will consider data collection, calibration and alignment, reconstruction and analysis, with MC simulated events used for studies and for analysis. The ePIC data acquisition is planned to be implemented as a flexible, scalable, and efficient streaming DAQ system, as outlined by the EIC yellow report.

- MC simulation in ePIC will encompass physics simulation (event and background modeling) and (with physics simulation as input) detector simulation, both fully detailed (Geant4) and fast (parameterized, ML based).
- For the ePIC detector at the EIC, the ability to process data remotely is an integral part of their proposed computing model. Access to storage resources and sufficient network bandwidth to move or access data to and from remote sites is a prerequisite.

6.9.2 EIC Facility Profile

Although the building blocks of the nucleon have been known for decades, a comprehensive theoretical and experimental understanding of how the quarks and gluons form nucleons and nuclei, and how their strong dynamics determine the properties of nucleons and nuclei, has been elusive. Most of the information about the nucleon's inner structure has emerged from the study of DIS process, an activity that has established QCD as the theory of the strong interaction.

Dual advances in perturbative QCD and computation have laid the foundation to imaging quarks and gluons and their dynamics in nucleons and nuclei. The theoretical accuracy of modern perturbative QCD calculations has recently been advanced to next-to-next-to leading order (NNLO) and beyond; these advances enable lepton-hadron scattering as a discovery tool via precision measurements.

6.9.2.1 Science Background

The EIC will investigate the structure of nucleons and nuclei at an unprecedented level. This will be accomplished by performing precise measurements of DIS and other processes over the complete relevant kinematic range including the transition region from perturbative to nonperturbative QCD. Highly polarized beams and high luminosity will allow probes of the spatial and spin structure of nucleons and nuclei, leading to high-precision determinations of PDFs and other quantum correlation functions. These investigations will advance our understanding of hadronization as well as QCD factorization and evolution.

The frontier accelerator facility in the US will be constructed in the 2020s, with an extensive science case as detailed in the EIC white paper[21], the 2015 *Nuclear Physics Long Range Plan*,[22] and the EIC yellow report.[23] The yellow report has been an important input to the successful DOE CD-1 review and decision. It describes the physics case, the resulting detector requirements, and the evolving detector concepts for the experimental program at the EIC. The first scientific collaboration for the EIC, ePIC, was formed in 2023 to support the realization of the DOE EIC Project detector. The EIC User Group is leading the efforts to establish a collaboration dedicated to the development of a second detector, responding to the EIC community's emphasized need for two detectors at the EIC.

Timeline and High-Level Milestones

The EIC facility is projected to begin data collection in 2033. While the guiding details of the ePIC detector, data acquisition systems, and analysis workflows are mostly understood, the final design is still in progress. The next major milestone is a TDR which will encompass comprehensive studies of the overall performance and science reach of the ePIC detector, leveraging full detector simulations for a broad selection of measurements.

The ePIC Detector subsystems need to be constructed by 2030. During this construction phase, the detector designs will undergo optimization based on cost and material availability. Prototypes of the detector systems will be tested, aiding in software testing and validating simulations with test-beam measurement results. The extended duration of the construction phase is ideal for developing and implementing the ePIC Streaming Computing Model. This includes setting up data acquisition and software for test beams, handling streaming challenges from data acquisition to offline reconstruction, conducting data challenges to test the scale and

[21] https://link.springer.com/article/10.1140/epja/i2016-16268-9
[22] https://nap.nationalacademies.org/catalog/25171/an-assessment-of-us-based-electron-ion-collider-science
[23] https://arxiv.org/abs/2103.05419

capability of distributed ePIC computing resources, and addressing analysis challenges related to autonomous alignment, calibrations, and end-to-end workflows from raw data to analysis.

No major networking requirements are anticipated before the next requirements review in 2027.

6.9.2.2 Collaborators

The ePIC Collaboration, currently consisting of 173 institutions from 24 countries, works jointly with the DOE EIC Project to realize the ePIC experiment. The host laboratories for the EIC, BNL and JLab, have formed the EIC Computing and Software Joint Institute to serve as a single point of contact and organizational entity for support of the ePIC Collaboration and other software and computing needs for the EIC. Within this structure, theoretical calculations and accelerating modeling will be conducted, also within the current estimates established by the individual BNL and JLab Facility use cases.

6.9.2.3 Instruments and Facilities

The EIC will be a frontier accelerator facility in the US. The versatile collider will support the exploration of nuclear matter over a wide range of center-of-mass energies, $\sqrt{s}$ = 20–140 GeV, and ion species, using highly-polarized electrons to probe highly-polarized light ions and unpolarized heavy ions. The high instantaneous luminosity of up to $L=10^{34}\,cm^{-2}s^{-1}$ and the highly-polarized beams at this frontier particle accelerator facility will allow for the precision study of the nucleon and the nucleus at the scale of the sea quarks and gluons.

The compute resources and the streaming computing model of the ePIC experiment and the resulting computing fabric for the EIC are under development. The current state can be found in the report on the "ePIC Streaming Computing Model."[24]

6.9.2.4 Generalized Process of Science

ePIC will utilize streaming readout for its data acquisition (DAQ) system. This approach will eliminate trigger bias and enable precise estimation of uncertainties during event selection, using all available detector data for a holistic reconstruction. The streaming DAQ will capture every collision signal, including background. This capability will significantly reduce background and related systematic uncertainties in an unprecedented manner.

Physics events will be reconstructed in near real time from the streaming data, modulo time varying calibrations that will require later reprocessing for a final fully calibrated reconstruction. The prompt availability of reconstructed data, and concurrent calibration cycle consuming it, is a crucial element of ePIC's objective to have a rapid, near real time turnaround of the raw data to production.

Simulations will play a crucial role in designing the ePIC experiment, prototyping analysis, and estimating systematic uncertainties. MC simulation in ePIC will encompass physics simulation (event and background modeling) and (with physics simulation as input) detector simulation, both fully detailed (Geant4) and fast (parameterized, ML based).

Analysis

The EIC has a broad science program. While some analysis prototyping and specific types of analysis can be conducted using computing resources at home institutes, numerous complex studies, such as imaging the quark-gluon structure of the nucleon, require the use of distributed computing resources. The traditional approach for these analyses is rooted around immediate data reduction of large amounts of detected particles into multi-dimensional histograms. Corrections for experimental effects, such as background effects, limited detector acceptance and resolution, and detector inefficiencies can then be deconvoluted from the observable of interest through simple arithmetic and matrix transformations. This procedure of deconvoluting experimental effects from histogrammed observables is referred to as unfolding. In contrast, there are emerging analysis techniques at the event level. The event-level approach requires a reversal of the traditional procedure of correcting and unfolding

[24] https://indico.bnl.gov/event/20481/attachments/49818/86296/ePIC-StreamingComputingModel.pdf

measured histograms: here, idealized events from theory have to be folded with the relevant experimental effects. After folding, the theoretical calculations can then be directly compared with the experimental events at the detector level. The accuracy and precision of these methods depend on intricate simulations in the unfolding scenario and detailed modeling of experimental effects in the folding scenario.

6.9.2.5 Remote Science Activities

Currently, the ePIC experiment utilizes the OSG for simulation production, including both detector and physics simulations. The ePIC experiment will follow a lineage of “big science” collaborations that leverage computing resources on a global scale, as outlined in the report on the “ePIC Streaming Computing Model”[25]. As ePIC develops its computing infrastructure in coordination with the EIC Computing and Software Joint Institute (ECSJI), opportunistic resources such as the OSG are anticipated to remain integral, especially for simulation production purposes.

6.9.2.6 Software Infrastructure

It is expected that data-management and transport frameworks such as Rucio and XRootD together with FTS and Globus will play a significant role. Object-based data storage and movement (supporting the S3 API) is increasingly common.

6.9.2.7 Network and Data Architecture

See [Sections 6.1.2.7 and 6.3.2.7] for information on the current state of JLab and BNL network architectures. As the computing model is decided, the network architecture of both sites may be adapted.

6.9.2.8 IRI Readiness

Due to the early state of this project, there are no IRI related items to report. It is expected that many of the IRI patterns will be present, and reflected, in the design of the technology to support the EIC.

6.9.2.9 Cloud Services

See [Sections 6.1.2.9 and 6.3.2.9] for information on the JLab and BNL cloud services. Due to the early nature of this project, information related to cloud computing is expected to change in the future.

6.9.2.10 Data-Related Resource Constraints

None to report at this time.

6.9.2.11 Data Mobility Endpoints

See [Sections 6.1.2.11 and 6.3.2.11] for information on the JLab and BNL data mobility services. Due to the early nature of this project, information related to data mobility is expected to change in the future.

6.9.2.12 Outstanding Issues

None to report at this time.

6.9.2.13 Facility Profile Contributors

BNL and JLab Representation

- Amber Boehnlein, JLab, amber@jlab.org
- Markus Diefenthaler, JLab, mdiefent@jlab.org
- Eric Lancon, BNL, elancon@bnl.gov
- Brad Sawatzky, JLab, brads@jlab.org

[25] https://indico.bnl.gov/event/20481/attachments/49818/86296/ePIC-StreamingComputingModel.pdf

ESCC Representation

- Vincent Bonafede, BNL, bonafede@bnl.gov
- Mark Lukasczyk, BNL, mlukasczyk@bnl.gov
- Brent Morris, JLab, bmorris@jlab.org

7 Case Study Discussion

In October 2023 an in-person review was held that invited participation from the NP community, DOE SC staff spanning ASCR and NP, and ESnet staff. NP activity host institutions, including the following DOE labs and facilities, were present for the review: BNL, FRIB, LBNL, ORNL, JLab, and Vanderbilt University. The purpose of this review was to go over the case studies presented in [Section 6] and discuss common findings and potential actions for the future support of the NP community by ESnet.

7.1 Outcomes

The following discussion points, many of which are core findings of this report and denoted in [Section 4], were discussed in this session. These observations have also helped to produce a set of actions in [Section 5] for the participants, that will drive future collaboration.

7.1.1 Facility and Experiment Discussions

- ***JLab and LQCD***
    - JLab has planned for a roughly two-fold increase in data volumes over the next three years.
    - The increase in data rates during the 12 GeV era has precipitated an increased use of off-site compute resources.
    - LQCD project teams seek allocations of computing time at numerous HPC facilities. While certain data may be retained for an extended period on leadership systems like NERSC, the primary responsibility for long-term data storage lies with the member laboratories. In the case of LQCD projects related to the JLab science program, JLab will serve as the host for the extended data storage.
    - The LCFs do not provide long-term storage for LQCD projects. Data are transferred to the USQCD computing facilities (JLab, Fermilab, and Brookhaven), which assume ownership.
    - Data transfers to JLab to support LQCD will increase and scale with the size of new LCF systems. JLab will continue to serve as the repository for long-term storage. A portion of the analysis work will be conducted at the LCFs. However, the final stage of the analysis workflow is ideally suited for execution on JLab's local systems, effectively mitigating the disparity in LCF to local computing capability.
- ***BNL: The SDCC and sPHENIX at RHIC***
    - At present, 52 DTNs are in operation at the SDCC, with the majority being utilized by programs outside of NP.
    - For sPHENIX and STAR, datasets are disk-resident at the SDCC, and the vast majority of dataset processing will take place at the SDCC itself. Local computational resources at BNL (located within the SDCC) can be a limiting factor, and the ability to tap nonlocal resources like OSG and other unaffiliated resources has the potential to augment capacity. The ability to utilize these resources, particularly for experiment workflows, will be limited by OSG resource availability, and the ability to transfer data between the host data center and the remote resources.
- ***FRIB***
    - The average FRIB data set sizes ranged from a few GB to ~70 TB with an average size just over 4 TB.

    - Researchers at FRIB are increasingly interested in using off-site HPC and data infrastructure to accomplish specific goals during the execution of an experiment. One experiment group has already employed local MSU HPCC resources to expediently analyze incoming data in near-real time to direct decisions during an experiment. The FDSi is exploring the use of NERSC for data analysis during ongoing experiments.

- ***GRETA***
    - The GRETA local computing infrastructure is designed to deliver a full set of science goals. However, the project and scientific user community recognize that advances in analysis could benefit from using large-scale computing (HPC) facilities. In the two- to five-year timeframe, advances in signal processing algorithms might make the use of remote computing attractive for processing the data for some GRETA experimental scenarios.
    - GRETA's capabilities will likely represent the most significant performance challenge to network infrastructure of FRIB or ANL. GRETA data transfer volumes can be between 50 GB and 100 TB and performed on an ad-hoc basis when the detector is operating. Generally, analysis of GRETA data is carried out by the experimental team at their home institutions. Analysis and data interpretation is a time-consuming process (many months) but not a very computationally intensive process (can be done on local computing resources). The nature of this analysis is very much experiment dependent.

- ***ALICE Project and ALICE-USA Computing***
    - The key data-management feature on the ALICE Grid is during the process of data-creation. Data are sent into the computational grid after experimentation, and are attempted to be accessed via "local" (e.g., topographically close) storage during processing.
        - When accessing local data, ALICE may read across WANs, requiring 1 GB/s of aggregate bandwidth. This can result in PB of data read and written across networks such as ESnet.
        - While it is true that ALICE computing is fully distributed, data processing is done locally, and all jobs are executed where the data reside. During the past year ALICE jobs have read over 2.3 EB and written over 400 PB of data from/to the local storage, averaging about 13 GB/s and 80 GB/s for write and read traffic respectively averaged over the entire grid.
        - ALICE has yearly episodes that require more significant WAN capacities. These occur when storage is added and/or decommissioned, or when data must otherwise be redistributed between different sites. 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.

- ***The CMS Heavy Ion Experimentation***
    - All CMS-HI data are transferred to Fermilab for archival storage, primarily limited by available tape bandwidth. The Mini-AOD storage tier is additionally transmitted to ACCRE, located at Vanderbilt University, for access by physicists. The annual data volumes are approximately five times what they were in 2018 (e.g., 40 PB/yr at the T0 and T1, and 3 to 10 PB/yr at T2s). CMS-HI expects an additional increase by a factor of two in HL-LHC (tentatively scheduled for 2029).
    - For CMS-HI, the primary limitation is tape bandwidth available to store and recall large multi-PB datasets. While the capacity of tape cartridges greatly increases when new generations of technology are released, the bandwidth per tape has not kept pace. If trends

continue, in 2030 it could take nearly a day of continuous access to read/write a tape from beginning to end. Writing to and retrieving from custodial tape is by far the most difficult issue for data movement CMS-HI will face in the coming years.

  - An identified issue with the CMS-HI data reduction pattern is that some popular portions of datasets are reproduced many times. At the predicted scales of HL-LHC (predicted to begin in 2029), this data duplication becomes financially burdensome to support. By storing data in object stores, these common slices of datasets could be stored and referenced by multiple end-user datasets, providing a better space efficiency for analysis. These benefits become more pronounced if there is a single copy globally of the relevant objects; this implies some level of additional WAN traffic needed to satisfy these workflows. These techniques are the target of active R&D and initial results are expected in the 2025 to 2026 timescale.

- ***The EIC***
  - The EIC is a community-driven facility that will be constructed in the 2020s and is projected to begin data collection in 2033. The effort to estimate data volumes from the EIC's ePIC detector is in progress. Raw data volumes could be hundreds of PBps, with reduction to hundreds of Gbps possible. BNL and JLab have formed the ECSJI to serve as a single point of contact and organizational entity for support of the ePIC Collaboration and other software and computing needs for the EIC.
  - EIC networking and computation decisions are not imminent, and will consider lessons learned from LHC computing and analysis. It is expected that major collaboration sites will contribute resources and leverage ESnet connectivity as a critical networking link to manage traffic flows. The ability to handle PB volumes of data over time and bursts of hundreds of Gbps is expected.
  - For the ePIC detector at the EIC, the ability to process data remotely is an integral part of the proposed computing model. Access to storage resources and sufficient network bandwidth to move or access data to and from remote sites is a prerequisite.

7.1.2 Cross-Cutting Data Management, Workflow, Computing, Storage, and Networking

- Current NP experiments at colliders collect large quantities of data over the course of their multi-year lifespan, typically at a rate of over hundreds of PBs per year.
- The use of DOE HPC facilities by NP facilities and experiments, as well as those provided by distributed grid resources like the OSG, will continue to grow. It is expected that the volume of data generated at LCFs will continue to increase by a factor of 5 to 10 for NP workflows that use these resources.
- Rucio, a software package that manages large volumes of data spread across facilities at multiple institutions and organizations, is allowing NP facilities to re-imagine the data pipeline from the experiments, creating a single source of truth (i.e., all data are edited in only one location) for policy-based file movement of raw data around campus environments. This incorporates the use of OSG resources that can be leveraged both locally and external to the facility to deliver on computational tasks.
- Experimental and MC data workflows demand substantial computational resources. The MC workflow is generally executed across multiple sites. This includes not only those sites directly affiliated with the experiment but also independent ones such as those connected to the OSG and potentially even the DOE Leadership class HPC facilities (such as NERSC). The primary

reason for this distribution is the computational intensity of MC workflows, which demand significant processing power but use relatively minimal data.

- With the advent of new computing facilities such as Perlmutter at NERSC and Frontier at OLCF, the volume of data generated at LCFs is increasing by a factor of 5 to 10 for NP workflows that use these resources. Existing workflows will remain in use, leading to a corresponding increase in the amount of data that needs to be transferred back to home institutions. The Aurora system at ALCF will likely mark a significant milestone in data production. An estimated three times more data are expected to be generated on exascale systems compared with the current generation of systems.
- Since 2020, an increasing number of NP facility users participate remotely via login to computing resources and remote conferencing facilities, and by transferring data to and from facilities. Remote science activities routinely leverage computational resources provided at partner sites. These could be located at DOE HPC facilities, or distributed computing resources such as those provided by the OSG. Providing capabilities for remote users to observe the products of ongoing data analysis continues to be beneficial to increase engagement with the user community.
- NP simulation workflows are capable of running off-site, since the input data required are small versus those of an analysis workflow. In some cases, an experiment may require that the output of simulation runs be returned to a source institution for storage or future analysis, and this will contribute to incoming WAN traffic. This may result in traffic demands that are similar in scale to reconstruction workflow's contribution to the outgoing WAN traffic.
- NP experiments and facilities leverage several data mobility tools for sharing of experimental data: CVMFS, XRootD, FTS, Rucio, and Globus are all used when exchanging data with collaboration sites (e.g., DOE HPC Facilities, OSG participants) and with end users.
- The majority of NP facilities and experiments (located at BNL, JLab, FRIB, ORNL, LBNL, and Vanderbilt) are all connected to ESnet with a capacity of 100 Gbps, and several will upgrade to multiple 100 Gbps or 400 Gbps in the coming years to support increases in data volumes.
- NP facilities and experiments continue to investigate use of commercial cloud services for data processing, in particular for the case where a rapid turnaround is required that exceeds resources available locally or via dedicated (e.g., DOE HPC) or distributed computing (e.g., OSG) facilities. Due to the current cost associated with cloud computing, use is expected to be rare, and WAN requirements will not significantly increase.

7.1.3 IRI Responsiveness

- JLab anticipates that once uniform interfaces are available to create workflows that are portable between experimental facilities, HPC facilities, and network facilities, NP workflows can more easily use external resources and will be substantially strengthened. IRI will allow for better EIC integration at BNL and JLab for the long-term campaign pattern, as well as other possibilities such as time-dependent workflows being managed across sites.
- In light of the crucial role of authentication in the IRI vision, a central repository with reference designs and implementation examples for common workflows would offer significant value
- For many NP experiments, utilization of unaffiliated computing resources not at the host data centers is a given. OSG and the supercomputers at DOE facilities (mainly NERSC for the STAR experiment) are commonly used for simulations and are two examples of remote, unaffiliated computing resources.

- NP facilities, experiments, and researchers have noted that integrating with other labs and data facilities presents ongoing challenges in authentication and federated identity. While frameworks like x.509, OAuth, and tokens have broad consensus, issues arise from nonstandard or incomplete implementations. Federated interactive logins that leverage approaches such as SAML, OpenAuth2, and Shibboleth are standardized and interoperable, but "automated" authentication for tasks like remote job control and large file transfers is fragmented and challenging to debug.
- NP facilities, experiments, and researchers have noted that facility-to-facility trust implementations are also difficult from both policy and technical perspectives. Cross-facility systems often rely on long-lived "robot" tokens or certificates that are an extension of a user's identity. These are not always a good fit for long-term campaigns. It is often the case that actions need to be taken on behalf of or as a proxy for users, which can stretch the abilities of federated identities.
- In light of the crucial role of authentication in the IRI vision, a central repository with reference designs and implementation examples for common workflows would offer significant value.

List of Abbreviations

AAA	Any data, Anytime, Anywhere
ACCRE	Advanced Computing Center for Research and Education
ASCR	Advanced Scientific Computing Research
AF	Analysis Facilities
AI	Artificial Intelligence
ALCF	Argonne Leadership Computing Facility
ALICE	A Large Ion Collider Experiment
AMD	Advanced Micro Devices, Inc.
ANASEN	Array for Nuclear Astrophysics Studies with Exotic Nuclei
ANL	Argonne National Laboratory
AOD	Analysis Object Data
ARIS	Advanced Rare Isotope Separator
ASCR	Advanced Scientific Computing Research
ATLAS	Argonne Tandem Linac Accelerator System
BCS	Beta Counting system
BECOLA	beam-cooler and laser spectroscopy
BNL	Brookhaven National Laboratory
BRAHMS	Broad RAnge Hadron Magnetic Spectrometers
CAESAR	Cesium Iodide Array
CE	Compute Element
CEBAF	Continuous Electron Beam Accelerator Facility
CephFS	Ceph File System
CFN	Center for Functional Nanomaterials
CLAS	CEBAF Large Acceptance Spectrometer
CMS	Compact Muon Solenoid
CMSSW	CMS Offline Software
CPU	central processing unit
CRAB	CMS Remote Analysis Builder
CRIC	Computing Resource Information Catalog
CTF	Compressed Time Frame
CVMFS	CernVM File System
DAQ	data acquisition
DESY	Deutsches Elektronen-Synchrotron
DIS	deep-inelastic scattering
DNS	Domain Name System
DOE	Department of Energy
DST	data summary tapes
DTN	Data Transfer Node
DWDM	dense wavelength-division multiplexing

ECP	Exascale Computing Project
ECSJI	EIC Computing and Software Joint Institute
EIC	Electron-Ion Collider
EJFAT	ESnet JLab FPGA Accelerated Transport
E-LITE	Eastern Lightwave Internetworking Technology Enterprise
ENP	Experimental Nuclear Physics
EoS	equation of state
EPICS	Experimental Physics and Industrial Control System
EPN	Event Processing Nodes
ESCC	ESnet Site Coordinators Committee
FDS	FRIB Decay Station
FLP	First Level Processing
Fermilab	Fermi National Accelerator Laboratory
FPGA	field programmable gate arrays
FRIB	Facility for Rare Isotope Beams
FRIBUO	FRIB Users Organization
FSU	Florida State University
FTS	File Transfer Servic
GEANT	GEometry ANd Tracking
GEM	gas-electron-multiplier
GPGPU	general-purpose graphics processor units
GPU	graphics processing unit
GRETA	Gamma-Ray Energy Tracking Array
HEP	High Energy Physics
HiRA	High-Resolution Charged-Particle Array
HPC	high-performance computing
HPCC	High-Performance Computing Center
HPCS	High-Performance Computing Services
HPDF	High Performance Data Facility
HPSS	High Performance Storage System
HRS	High-Rigidity Spectrometer
HTC	high throughput Computing
HTSN	High Throughput Science Network
IB	InfiniBand
IRI	Integrated Research Infrastructure
ISLA	Isochronous Large Acceptance
IT	Information Technology
JANUS	Joint Array for NUclear Structure
JENSA	Jet Experiments in Nuclear Structure and Astrophysics
JLab	Thomas Jefferson National Accelerator Facility
KEK	Japanese High-Energy Accelerator Research Organization

KISTI	Korea Institute of Science and Technology Information
LAN	local area network
LBNL	Lawrence Berkeley National Laboratory
LC	Livermore Computing
LCF	Leadership Computing Facilities
LEBIT	Low-Energy Beam and Ion Trap
LEGO	Lightweight Environment for Grid Operators
LENDA	Low-Energy Neutron Detector Array
LHC	Large Hadron Collider
LHCONE	LHC Open Network Environment
LISA	Large multi-Institution Scintillator Array
LLNL	Lawrence Livermore National Laboratory
LQCD	Lattice QCD
LSU	Louisiana State University
MARIA	Mid-Atlantic Research Infrastructure Alliance
MC	Monte Carlo
MERIT	Michigan Educational Research Information Triad
ML	machine learning
MOLLER	Measurement of a Lepton-Lepton Electroweak Reaction
MoNA	Modular Neutron Array
MSU	Michigan State University
NERO	Neutron Emission Ratio Observer
NERSC	National Energy Research Scientific Computing
NIC	network interface controller
NNLO	next-to-next-to leading order
NOAA	National Oceanic and Atmospheric Administration
NOTED	Network Optimized for Transfer of Experimental Data
NP	Nuclear Physics
NPPLCI	Nuclear and Particle Physics LQCD Computing Initiative
NPS	Neutral Particle Spectrometer
NSCL	National Superconducting Cyclotron Laboratory
NSF	National Science Foundation
OLCF	Oak Ridge Leadership Computing Facility
ORNL	Oak Ridge National Laboratory
OSG	Open Science Grid
PAC	Program Advisory Committee
PB	petabyte
PBR	policy-based routing
PDSF	Parallel Distributed Systems Facility
PI	principal investigator
PoP	point-of-presence

POSIX	Portable Operating System Interface
PPAC	parallel-plate avalanche counters
PSC	Pittsburgh Supercomputer Center
PVDIS	Parity Violating Deep Inelastic Scattering
QCD	quantum chromodynamics
QGP	quark-gluon plasma
QoS	Quality of Service
RCF	RHIC Computing Facility
ReA	ReAccelerator
RHIC	Relativistic Heavy Ion Collider
RiSE	Resonance-ionization Spectroscopy Experiment
SAMURAI	Superconducting Analyzer for Multi-particles from Radioisotope beams
SBS	Super BigBite Spectrometer
SDCC	Scientific Data and Computing Center
SE	storage elements
SECAR	Separator for Capture Reactions
SeGA	Segmented Germanium Array
SENSE	Software-Defined Network for End-to-end Networked Science at the Exascale
SIDIS	Semi-Inclusive Deep Inelastic Scattering
SLAM	Short-Lived Atoms and Molecules
SoLID	Solenoidal Large Intensity Device
SONIC	Services for Optimized Network Inference on Coprocessors
sPHENIX	super Pioneering High-Energy Nuclear Interaction eXperiment
SSH	Secure SHell
STAR	Solenoidal Tracker at RHIC
SWIF	Scientific Workflow Indefatigable Factotum
TB	terabyte
TBB	Thread-Building Blocks
TDR	Technical Design Report
ToR	Top of Rack
TPC	time-projection chamber
UDP	User Datagram Protocol
USQCD	US Lattice Quantum Chromodynamics
VO	Virtual Organization
VOMRS	Virtual Organization Management and Registration Service
WAN	wide-area network
WebDAV	Web Distributed Authoring and Versioning
WLCG	Worldwide LHC Computing Grid