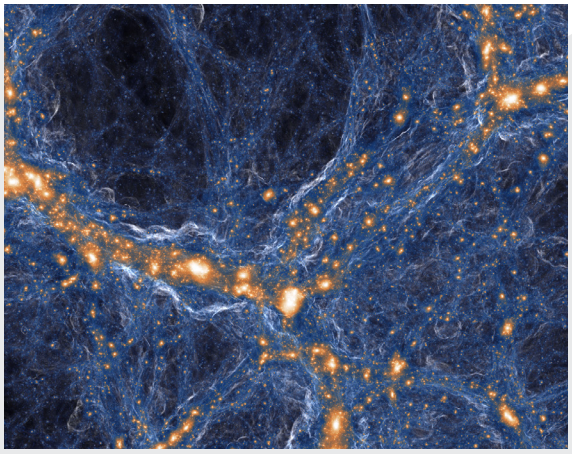
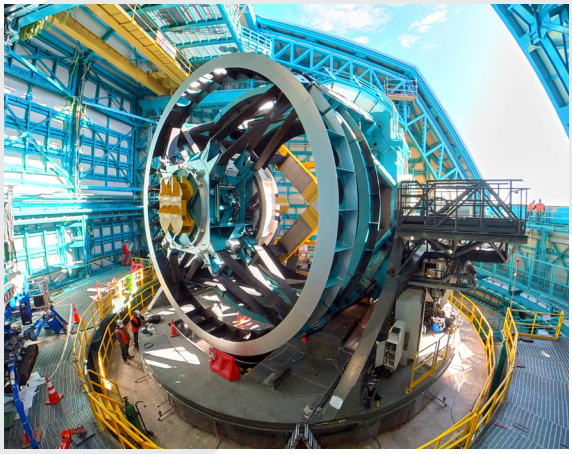
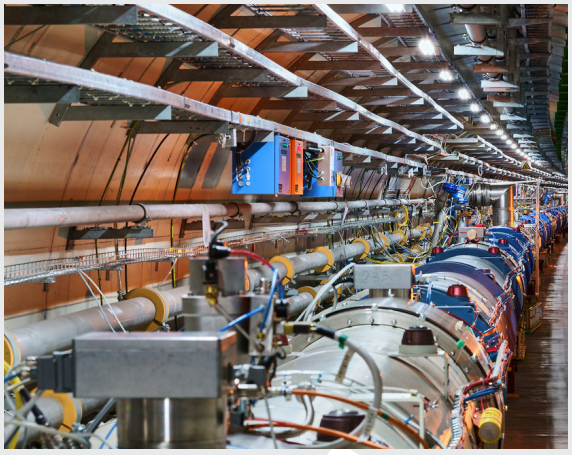
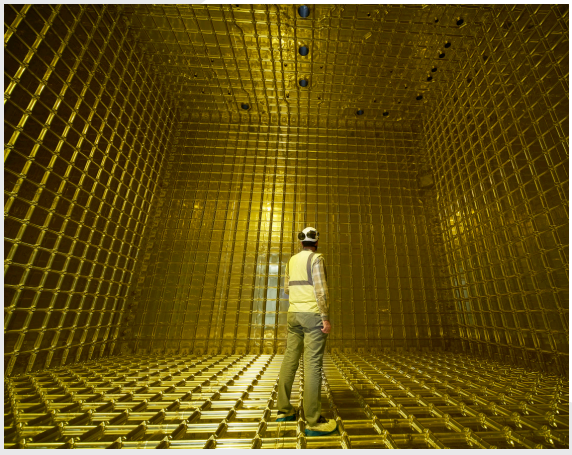
ESnet
ENERGY SCIENCES NETWORK

High Energy Physics Network Requirements Review: Two-Year Update

July 2023

BERKELEY LAB

U.S. DEPARTMENT OF ENERGY
Office of Science

ESnet
ENERGY SCIENCES NETWORK

High Energy Physics Network Requirements Review: Two-Year Update

July 2023

Office of High Energy 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 Directors of the Office of Science, Office of Advanced Scientific Computing Research, Facilities Division, and the Office of High Energy Physics.

This is LBNL report number LBNL-2001605.

Cover Images:
(**Top left**) the new universe simulation model, dubbed Illustris, courtesy of Simons Foundation
(**Top right**) Wide view of the telescope mount inside the dome at the Rubin Observatory, courtesy of H. Stockebrand/Rubin Obs/NSF/AURA
(**Bottom left**) LHC tunnel image, courtesy of CERN
(**Bottom right**) protoDUNE detectors at CERN, courtesy of Max Brice/CERN) PHENIX image

[1] https://escholarship.org/uc/item/00w301f1

Participants and Contributors

Garhan Attebury, *University of Nebraska-Lincoln*

Nicole Avila, *University of Chicago*

Stephen Bailey, *Lawrence Berkeley National Laboratory*

Justas Balcas, *California Institute of Technology*

Amanda Bauer, *Rubin Observatory*

Lothar Bauerdick, *Fermi National Accelerator Laboratory*

Chris Bee, *Stony Brook University*

Doug Benjamin, *Argonne National Laboratory*

Kurt Biery, *Fermi National Accelerator Laboratory*

Kenneth Bloom, *University of Nebraska-Lincoln*

Bob Blum, *Rubin Observatory*

Andrey Bobyshev, *Fermi National Accelerator Laboratory*

Brian Bockelman, *University of Wisconsin-Madison*

Tim Bolton, *Kansas State University*

Vincent Bonafede, *Brookhaven National Laboratory*

Julian Borrill, *Lawrence Berkeley National Laboratory and University of California, Berkeley*

Tulika Bose, *University of Wisconsin-Madison*

Joseph Boudreau, *University of Pittsburgh*

Steve Brice, *Fermi National Accelerator Laboratory*

Benjamin Brown, *Department of Energy Office of Science*

Paolo Calafiura, *Lawrence Berkeley National Laboratory*

Simone Campana, *European Organization for Nuclear Research*

Dale Carder, *Lawrence Berkeley National Laboratory and Energy Sciences Network*

John Carlstrom, *University of Chicago and Argonne National Laboratory*

Eric Colby, *Department of Energy Office of Science*

John Corlett, *Lawrence Berkeley National Laboratory*

Eli Dart, *Lawrence Berkeley National Laboratory and Energy Sciences Network*

Kaushik De, *University of Texas at Arlington*

Phil DeMar, *Fermi National Accelerator Laboratory*

Seth Digel, *SLAC National Accelerator Laboratory*

Richard Dubois, *Stanford University*

Daniel Eisenstein, *Harvard University*

Johannes Elmsheuser, *Brookhaven National Laboratory*

Simon Fiorucci, *Lawrence Berkeley National Laboratory*

Mark Foster, *SLAC National Accelerator Laboratory*

Stuart Fuess, *Fermi National Accelerator Laboratory*

Robert Gardner, *University of Chicago*

Gil Gilchriese, *Lawrence Berkeley National Laboratory*

Heather Gray, *University of California Berkeley*

Chin Guok, *Lawrence Berkeley National Laboratory and Energy Sciences Network*

Oliver Gutsche, *Fermi National Accelerator Laboratory*

Julien Guy, *Lawrence Berkeley National Laboratory*

Salman Habib, *Argonne National Laboratory*

Carol Hawk, *Department of Energy Office of Science*

Damian Hazen, *Lawrence Berkeley National Laboratory and Energy Sciences Network*

Katrin Heitmann, *Argonne National Laboratory*

Ken Herner, *Fermi National Accelerator Laboratory*

Saswata Hier-Majumder, *Department of Energy Office of Science*

Michael Hildreth, *University of Notre Dame*

Julio Ibarra, *Florida International University*

David Jaffe, *Brookhaven National Laboratory*

Jeff Kantor, *Rubin Observatory*

Heather Kelly, *SLAC National Accelerator Laboratory*

Wesley Ketchum, *Fermi National Accelerator Laboratory*

Mike Kirby, *Fermi National Accelerator Laboratory*

Alexei Klimentov, *Brookhaven National Laboratory*

Markus Klute, *Massachusetts Institute of Technology*

Robert Kutschke, *Fermi National Accelerator Laboratory*

Eric Lancon, *Brookhaven National Laboratory*

David Lange, *Princeton University*

Kevin Lannon, *University of Notre Dame*

Paul Laycock, *Brookhaven National Laboratory*

Tom Lehman, *Lawrence Berkeley National Laboratory and Energy Sciences Network*

James Letts, *University of California San Diego*

Michael Levi (Dark Energy Spectroscopic Instrument (DESI) Director), *Lawrence Berkeley National Laboratory*

Mark Lukasczyk, *Brookhaven National Laboratory*

Adam Lyon, *Fermi National Accelerator Laboratory*

Krista Majewski, *Fermi National Accelerator Laboratory*

Dan Marlow, *Princeton University*

Phil Marshall, *SLAC National Accelerator Laboratory*

Edoardo Martelli, *European Organization for Nuclear Research*

David Mason, *Fermi National Accelerator Laboratory*

Shawn Mckee, *University of Michigan*

Andrew Melo, *Vanderbilt University*

Bogdan Mihaila, *National Science Foundation*

Bill Miller, *Department of Energy Office of Science*

Ken Miller, *Lawrence Berkeley National Laboratory and Energy Sciences Network*

Inder Monga, *Lawrence Berkeley National Laboratory and Energy Sciences Network*

Maria Elena Monzani, *SLAC National Accelerator Laboratory*

Harvey Newman, *California Institute of Technology*

Will O'Mullane, *Rubin Observatory*

Verena Martinez Outschoorn, *University of Massachusetts Amherst*

Nathalie Palanque-Delabrouie, *the French Alternative Energies and Atomic Energy Commission*

Ramon Pasetes, *Fermi National Accelerator Laboratory*

Abid Patwa, *Department of Energy Office of Science*

Christoph Paus, *Massachusetts Institute of Technology*

Srini Rajagopalan, *Brookhaven National Laboratory*

Quentin Riffard, *Lawrence Berkeley National Laboratory*

Kate Robinson, *Lawrence Berkeley National Laboratory and Energy Sciences Network*

David Schelgel, *Lawrence Berkeley National Laboratory*

Heidi Schellman, *Oregon State University*

Kate Scholberg, *Duke University*

Jennifer Schopf, *Texas Advanced Computing Center (TACC)*

Uros Seljak, *University of California, Berkeley*

Elizabeth Sexton-Kennedy, *Fermi National Accelerator Laboratory*

Richard Simon, *Lawrence Berkeley National Laboratory*

Eric Smith, *Lawrence Berkeley National Laboratory and Energy Sciences Network*

Maria Spiropulu, *California Institute of Technology*

Tavia Stone Gibbins, *Lawrence Berkeley National Laboratory and the National Energy Research Scientific Computing Center*

Rune Stromsness, *Lawrence Berkeley National Laboratory*

Matevz Tadel, *University of California San Diego*

Kevin Thompson, *National Science Foundation*

Steve Timm, *Fermi National Accelerator Laboratory*

Margaret Votava, *Fermi National Accelerator Laboratory*

Paul Wefel, *Lawrence Berkeley National Laboratory and Energy Sciences Network*

Torre Wenaus, *Brookhaven National Laboratory*

Andrew Wiedlea, *Lawrence Berkeley National Laboratory and Energy Sciences Network*

Linda Winkler, *Argonne National Laboratory*

Frank Wuerthwein, *University of California San Diego*

Wei Yang, *SLAC National Accelerator Laboratory*

Xi Yang, *Lawrence Berkeley National Laboratory and Energy Sciences Network*

Jim Yeck, *University of Wisconsin, Madison*

Alexandr Zaytsev, *Brookhaven National Laboratory*

Jason Zurawski, *Lawrence Berkeley National Laboratory and Energy Sciences Network*

Report Editors

Jason Zurawski, *ESnet:* zurawski@es.net

Dale Carder, *ESnet:* dwcarder@es.net

Eric Colby, *Department of Energy Office of Science:* Eric.Colby@science.doe.gov

Eli Dart, *ESnet:* dart@es.net

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

Ken Miller, *ESnet:* ken@es.net

Abid Patwa, *Department of Energy Office of Science:* Abid.Patwa@science.doe.gov

Kate Robinson, *ESnet:* katerobinson@es.net

Andrew Wiedlea, *ESnet:* 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.

In July 2023, the Energy Sciences Network (ESnet) and the High Energy Physics program (HEP) of the DOE SC organized an interim ESnet requirements review of HEP-supported activities, to follow up on the work started during the *2020 HEP Network Requirements Review*. Preparation for these events included checking back with the key stakeholders: program and facility management, research groups, and technology providers. Each stakeholder group was asked to prepare updates to their previously submitted case study documents, so that ESnet could update the understanding of any changes to the current, near-term, and long-term status, expectations, and processes that will support the science activities of the program.

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

- Large Hadron Collider (LHC) experimentation and operation:
  - ATLAS (A Toroidal LHC ApparatuS) experiment.
  - Compact Muon Solenoid (CMS) experiment.
  - LHC operations.
  - High-luminosity (HL) era of the LHC.
- Neutrino experiments at Fermi National Accelerator Laboratory (Fermilab):
  - Short-Baseline Neutrino Program (SBN).
  - Deep Underground Neutrino Experiment (DUNE).
- Dark Energy Spectroscopic Instrument (DESI).
- Belle II experiment.
- Muon experimentation at Fermilab:
  - Muon G minus two (g-2).
  - Muon-to-electron-conversion experiment (Mu2e).
- Dark Energy Science Collaboration (DESC).
- The Vera C. Rubin Observatory (Rubin Observatory) and the Legacy Survey of Space and Time (LSST).
- Cosmic Microwave Background — Stage 4 (CMB-S4).
- Cosmological simulation research.

- LZ (LUX-ZEPLIN) Dark Matter Experiment.

The review participants spanned the following roles:

- Subject-matter experts from the HEP activities listed previously.
- ESnet Site Coordinators Committee members from HEP activity host institutions, including the following DOE labs: Argonne National Laboratory (ANL), Brookhaven National Laboratory (BNL), the European Organization for Nuclear Research (CERN), Fermilab, LBNL, the National Energy Research Scientific Computing Center (NERSC), and SLAC National Accelerator Laboratory (SLAC).
- Networking and/or science engagement leads from the ASCR HPC facilities.
- DOE SC staff spanning both ASCR and HEP.
- ESnet staff supporting positions related to facility leadership, scientific engagement, networking, security, software development, and research and development (R&D).

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 multifacility research data workflows. This work will also extend analysis done on IRI patterns[2] and discuss ways future HEP 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:

- Pandemic impacts continue to persist, and are causing design, testing, construction, and experimental runtime delays.
- Increased data volumes were reported by several experiments, but not by significant margins beyond the predictions that were made during the *2020 HEP Network Requirements Review*. These increases are mostly related to minor changes in the process of analysis, as well as increased runtimes.
- Leveraging "data challenges," synthetic exercises to evaluate the performance of hardware and software using previously collected, or simulated, data sets, continues to be an effect way to plan for experimental success.
- Investigation into the use of machine learning (ML) and artificial intelligence (AI) approaches continues, with more HEP facilities and experiments looking to gain access to graphics processing unit (GPU) resources.
- The use of commercial cloud resources to support components of experimental workflows continues to be an area of active investigation, but not an area of active deployment. Data volumes, computational responsiveness, and network performance are all important factors.

[1] https://www.osti.gov/biblio/1984466
[2] https://www.osti.gov/biblio/2205078

- ESnet continues to work with sites, facilities, and R&E partners on planning to upgrade network connectivity to meet the data volumes required for projected experimental loads.
- ESnet has executed on planned expansion of transatlantic network capacity to support LHC, DUNE, and other use cases.

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 HEP participants on a yearly basis, until the next requirements review process begins:

- ESnet will be an active participant in DC24 with the LHC experiments, and will be monitoring the network at CERN, BNL, Fermilab, and LHC Open Network Environment (LHCONE)-connected Tier 2s.
- ESnet will continue to work with R&D efforts in the LHC community.
- ESnet will continue to participate with community members in discussions surrounding best common practices for data mobility and management.
- ESnet will continue to work with Research Education and Economic Development Network (REED) and Sanford Underground Research Facility (SURF) on network upgrades to support 100 Gbps connectivity to support DUNE and LZ.
- ESnet transatlantic capabilities have been upgraded, but will advance further in the coming years. ESnet will remain in contact with key HEP stakeholders on the status of this.
- ESnet will continue to work with AMPATH in supporting the international connection to Rubin Observatory.
- ESnet will continue to work with the LHCONE and Large Hadron Collider Optical Private Network (LHCOPN) planning groups.
- ESnet will participate in DOE efforts such as the Integrated Research Infrastructure Architecture Blueprint Activity (IRI-ABA).
- ESnet will continue to maintain and augment cloud connectivity options.

2 Requirement Review Overview

ESnet and ASCR use the requirements review process 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 HEP, 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], metropolitan-area network [MAN], and wide-area network [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 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. This section asks if their workflows can be categorized into the three common patterns:

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

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 HEP Program

The HEP program's mission is to understand how the universe works at its most fundamental level by discovering the elementary constituents of matter and energy, probing the interactions between them, and exploring the basic nature of space and time. This R&D inspires young minds, trains an expert workforce, and drives innovation that improves the nation's health, wealth, and security.

The scientific objectives and priorities for the field recommended by the HEP Advisory Panel are detailed in its recent long-range strategic plan, developed by the Particle Physics Project Prioritization Panel (P5).[3] HEP research is inspired by some of the most fundamental questions about our universe. What is it made of? What forces govern it? How did it evolve to the way it is today? Finding these answers requires the combined efforts of some of the largest scientific collaborations in the world, using large arrays of the most sensitive detectors in the world, at some of the largest and most complex scientific machines in the world.

HEP supports US researchers who play leading roles in these international efforts and world-leading facilities at our national laboratories that make this science possible. HEP also develops new accelerator, detector, and computational tools to open new doors to discovery science, and through the Accelerator Stewardship program, works to make transformational accelerator technology widely available to science and industry.

[3] https://science.osti.gov/~/media/hep/hepap/pdf/May-2014/FINAL_P5_Report_053014.pdf

3 Review Findings

The requirements review process helps to identify important facts and opportunities from the programs and facilities that are profiled. These points summarize important information gathered during the review discussions surrounding case studies and the HEP program in general.

- The overall schedule for LHC and HL-LHC remains similar to what was reported in the previous update. Run 3 will proceed in 2024 and 2025, the long shutdown will last between 2026 and 2028, and Run 4 is scheduled to begin in 2029.
  - Due to energy costs in Europe, the yearly LHC run schedule was altered for the end of the run in 2023 and for early 2024: this resulted in one fewer month of operation.
  - The LHC did not operate in late 2023 due to a failure, and this impacted operations for the end of 2023. The yearly Heavy-Ion operation was not completed in 2023 as a result.
- The LHC collaboration is actively preparing for DC24, a set of network, storage, and computational challenges that are meant to prepare the Tier 0, Tier 1s, and Tier 2s for the increases in data expected during HL-LHC operations.
  - DC24 was scheduled for late February and early March 2024 and targeted a 25% increase over Run 3 data volumes.
- LHC research efforts continue and are now working with several persistent testbeds to test some of the major R&D efforts. These efforts are working to integrate production tools (e.g., Rucio, File Transfer Service [FTS]) with automated ways of reserving network paths and bandwidth between facilities. The work will also allow more fine-grained management of network flows and more efficient use of limited network resources.
- SBN's new detector (Short-Baseline Near Detector [SBND]) will start taking data in 2024, joining Imaging Cosmic And Rare Underground Signals (ICARUS), which has been running since 2021.
  - SBN continues to use Open Science Grid (OSG) resources for computing, with the majority being contributed by Fermilab.
  - SBN is working to port some aspects of a ML workflow to DOE HPC facilities, specifically Argonne Leadership Computing Facility (ALCF) and NERSC. This will result in less than 1 petabyte (PB) being sent across the network for processing.
- protoDUNE will enter operation in 2024. A data challenge from CERN to Fermilab is also planned in early 2024, and will attempt to reach 40 Gbps as a load test.
- DUNE will begin operation at SURF in 2028. Currently, DUNE continues to work on overall simulation and analysis workflow performance. DUNE expects to generate 30 PB/year of data.
- DUNE is exploring options for increasing access to GPUs. For now, DUNE is trying to identify resources at DOE HPC centers and discussing with NERSC resources on Perlmutter.
- DUNE networking (between major collaboration sites) will target using LHCONE where applicable.
- The main SURF site is still limited to 10 Gbps, but looking for upgrades in 2024 or 2025. DUNE data acquisition (DAQ) commissioning is scheduled for 2027 to 2028.
  - DUNE requires 100 Gbps between SURF and Fermilab to support operations.
  - LZ requires 10 Gbps between SURF and LBNL and is currently sending a fractional amount (4 Gbps) to preserve resources.

- DESI had an initial public data release in July 2023, which resulted in over 7,000 users downloading data from NERSC.
- DESI's use of Perlmutter has worked without issue, but the project acknowledges that this is now a single point of failure for analysis.
- Belle II was in shutdown mode for maintenance and restarted in late 2023.
- Belle II will conduct a data challenge in 2024, with the expectation of moving 40 TB in a day (which equates to a sustained network performance of 4 Gbps) using LHCONE connectivity.
- Muon g-2 shut down in July 2023, and will perform data analysis on historic data sets in the coming years. This will be performed primarily at Fermilab, with the potential for some use of OSG resources at other facilities.
- Mu2e is under construction and expects to go into service for an initial run in 2026. No changes are expected to the data analysis or computing models. Mu2e's second run is scheduled for 2029 and will last four years.
- DESC has made progress on defining an experimental workflow and analysis framework for when Rubin data become available in 2026. The project will perform primary storage and computation activities at NERSC, with backup sites at IN2P3 and eInfrastructure for Research and Innovation (IRIS) at the Science and Technology Facilities Council (STFC).
- The construction of the Rubin Observatory continues and the facility expects to start collecting data in 2025, with data analysis starting in 2026. Data rates and expectations on data-movement deadlines remain the same.
- AMPATH continues to work with the Rubin Observatory and ESnet to characterize and support the networking path between the facilities.
- CMB-S4 must re-evaluate some implementation and scientific goals based on resources constraints at the South Pole. This includes sharing of limited network and power resources, as well as potentially changing the expectations for scientific analysis.
- CMB-S4 will embark on a set of data challenges in the coming years as it gets closer to implementing analysis workflows and data mobility expectations.
- The long-term viability of cosmological simulation research will rely on access to increasing amounts of computational and storage resources that persist between funding cycles and campaign runs.
- Network bandwidth and resiliency out of SURF remains a significant concern for the LZ project. This will become critical in coming years as more experiments use networking resources.

4 Review Actions

ESnet recorded a set of recommendations from the HEP and ESnet requirements review that extend ESnet's ongoing support of HEP-funded collaborations. Based on the key findings, the review identified several recommendations for HEP, ASCR, ESnet, and ASCR HPC facilities to jointly pursue.

- ESnet will be an active participant in DC24 with the LHC experiments, and will be monitoring the network at CERN, BNL, Fermilab, and LHCONE-connected Tier 2s.
- ESnet will continue to work with R&D efforts in the LHC community.
- ESnet will continue to participate with community members in discussions surrounding best common practices for data mobility and management.
- ESnet will continue to work with REED and SURF on network upgrades to support 100 Gbps connectivity for DUNE and LZ.
- ESnet transatlantic capabilities have been upgraded, but will advance further in the coming years. ESnet will remain in contact with key HEP stakeholders on status.
- ESnet will continue to work with AMPATH in supporting the international connection to Rubin Observatory.
- ESnet will continue to work with the LHCONE and LHCOPN planning groups.
- ESnet will participate in DOE efforts including IRI-ABA.
- ESnet will continue to maintain and augment cloud connectivity options.

5 HEP Case Study Updates

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 funded by the HEP program of the DOE SC. These updated case studies were reviewed virtually in July 2023. The case studies profiled, and featured in the *2020 HEP Network Requirements Review*, include the following:

- LHC experimentation and operation:
  - ATLAS experiment.
  - CMS experiment.
  - LHC operations.
  - HL era of the LHC.
- Neutrino experiments at Fermilab:
  - SBN.
  - DUNE.
- DESI.
- Belle II experiment.
- Muon experimentation at Fermilab:
  - Muon G minus two (g-2).
  - Muon-to-electron-conversion experiment (Mu2e).
- DESC.
- The Rubin Observatory and the LSST.
- CMB-S4.
- Cosmological simulation research.
- LZ Dark Matter Experiment.

5.1 LHC Experimentation and Operation (e.g., ATLAS, CMS, and Shared Operations)

The LHC at the European Laboratory for Particle Physics (CERN) is the most powerful particle accelerator in the world. Highly energetic protons, traveling almost at the speed of light around a 27-kilometer-long ring in both directions, are steered to collide head-on, creating new particles and new interactions to probe fundamental natural laws.

The two main general-purpose collaborations at the LHC are ATLAS and CMS, and both collaborations have thousands of collaborators distributed around the globe who require access to the data generated by their detectors at CERN and to simulated data generated at sites around the world.

All LHC experiments follow a general pattern of operation: capture of the raw data from the instrument at CERN, storage and dissemination of this raw data along with creation of a variety of formats that can be used for further analysis, creation and dissemination of "simulated" data sets used to assist in understanding and planning for analysis and calibration, and sharing of the data with the user community that analyzes and publishes the results.

The LHC collides protons more than a billion times every second. Experiments can select interesting collisions from this activity at a rate of 10,000 times every second. Starting in 2029, a major upgrade to the LHC (e.g., HL) will change the rates and sizes of collected physics data. With over 2,000 hours of data collection every year when the HL-LHC starts running in 2029, both collaborations will have a huge data sample for physicists to analyze worldwide. The HL-LHC program is expected to last for a decade. Large improvements in networking will be required to enable the ambitious physics goals of the HL-LHC.

The HL-LHC will accumulate roughly the same amount of integrated luminosity of data in three years as the entire LHC running period has produced. This implies that the science capabilities are expected to be roughly equivalent to the data taking from runs 1, 2, and 3 combined. The entire HL-LHC era will last for 10 years of data taking, with 12- to 24-month maintenance periods interspersed roughly every three years.

5.1.1 ATLAS Case Study Summary

- ATLAS is a global collaboration, with approximately 6,000 members spread among nearly 200 institutions in 38 countries. Data management from the single source of experimentation (CERN) to the highly distributed scientific population is an ongoing challenge.
- The ATLAS grid infrastructure consists of the Tier 0 computing site at CERN and 11 Tier 1, 70 Tier 2, and about 30 Tier 3 sites distributed worldwide. Basically, all workflows are executed at all tiers: the Tier 0, Tier 1, and Tier 2 sites. Tape storage to store raw and Analysis Object Data (AOD) files is available at the Tier 0 and Tier 1 sites.
- The Worldwide LHC Computing Grid (WLCG) collects computing resources worldwide and enables their usage by the LHC experiments. The mission of the WLCG is to provide global computing resources to store, distribute, and analyze the ~50–70 PBs of data expected every year of operations from the LHC.
- The US ATLAS Tier 1 is hosted at BNL's Scientific Data and Computing Center (SDCC). The ATLAS connection to ESnet is shared with other programs hosted at the SDCC. The US Tier 1 is the largest of the ATLAS experiment; it represents about 25% of the Tier 1 computing resources of ATLAS.
- Four ATLAS Tier 2 centers are in the US: Northeast Tier 2 (NET2), Great Lakes Tier 2 (AGLT2), Midwest Tier 2 (MWT2), and Southwest Tier 2 (SWT2). These centers are used for all distributed production and user analysis workloads. Each Tier 2 center consists of multiple university-based clusters. All Tier 2 sites are required to provide a minimum of 10 Gbps connectivity. However, all US Tier 2 sites provide 20–100 Gbps. The US goal is to achieve 40 Gbps links at all Tier 2 sites at the start of Run 3.
- ATLAS has a long history of successfully using HPC resources during Run 2 at the LHC. From 2016 to 2020, US-based HPC resources supported 10–25% of ATLAS simulation production. ATLAS plans to run all forms of workloads at HPCs. This will put much higher demands on networking.
- ATLAS computing is fully distributed. All computing activities are free to occur at any site, irrespective of their tier, based on intelligent brokering of tasks and jobs. Distributed analysis jobs are also brokered by site capability: users are discouraged from choosing a specific site. The distributed nature of ATLAS computing drives the network performance requirements between ATLAS sites. All ATLAS workloads and workflows may be run on demand at any time.
- The open-source software framework Rucio is used to organize, manage, and access the ATLAS data. Rucio consists of a central database at CERN that contains a dataset catalog (for all data the experiment produces). Rucio stages data between facilities based on processing requests.

Around 1–2PB per day are migrated worldwide in this manner. Rucio leverages other tools (FTS, etc.) to physically transfer the datasets.

- The PanDA Production and Distributed Analysis System(PanDA) ecosystem manages all workflows and workloads in ATLAS. It is designed to handle complex multistep workflows, running over thousands of files, using many different application workloads.
- BNL has implemented a vendor-agnostic, resilient, scalable, and modular terabit per second (Tbps) High Throughput Science Network (HTSN), which serves as the primary network transport for all data-intensive collaborations at BNL.
- Each Tier 2 site has unique LAN/WAN architecture developed in coordination with local and regional network managers.
- PanDA+Rucio can use commercial cloud resources interchangeably with grid-based WLCG resources, though such resources are currently not available in HEP. ATLAS expects a few PB of data transfers are possible to cloud sites.
- Capabilities to monitor and manage data transfers automatically are a high priority. Given the size, complexity, and fully distributed nature of ATLAS computing, all workflow and data distribution need to be optimized and managed with AI.
- As ATLAS begins Run 3, network needs will grow gradually. Increasing network capacity and performance will be needed at US ATLAS Tier 1, Tier 3, and Tier 3 analysis facilities.

5.1.2 CMS Case Study Summary

- CMS is divided into tiers of operation. CERN is considered "Tier 0" and is the home of a complete backup of the raw data set, along with partial copies of other formats used for calibration, reconstruction, and simulation. The globally distributed Tier 1 and Tier 2 facilities are responsible for data archiving, simulated data generation, analysis data storage, and physics analysis activities.
- The US operates one Tier 1 facility (Fermilab), which is responsible for 40% of CMS Tier1 capacity. The majority of the traffic flows affiliated with Fermilab are related to raw data from CERN during operations, but may also be related to reprocessing the raw data, producing/sharing simulations, and producing/sharing user analysis. Fermilab has 27 PB of active disk storage available for use.
- The US has seven Tier 2 facilities. Data typically move from these facilities (and the Tier 1 at Fermilab) to other universities as analysis data sets are reduced and refined during the analysis process. These facilities each contribute around 3 PB (or more) of active discussion storage.
- Tier 3 facilities are loosely organized (and nonfunded) resources that perform user-level analysis. Access patterns here are usually in the form of downloading analysis formats for local processing, and the potential to upload results to group storage at other locations.
- HPC facilities (National Science Foundation [NSF] and DOE funded) are not a primary use case for US CMS, but can be used for certain aspects of the overall workflow, typically for simulation production.
- Today, and in the future, the global CMS collaboration together with WLCG and OSG will define services that sites perform.
- In the future, centers of a given tier may no longer provide all the services that today would be expected from that tier. In addition, it is likely that the HL-LHC data and processing infrastructure will no longer support the full global mesh of data flows among all tiers.
- Run 2 produced around 45 PB of total data during the four years of operation, and a roughly

similar set is expected for Run 3, as no major technology upgrades occurred beyond changes to file formats on the analysis side. Run 4 will usher in a new era of scientific technology and will produce 350 PB of data per year starting in 2029.

- CMS supports streaming data access to any data on disk across its grid facilities from any location with an internet connection at any time. This is called AAA, for "Any Data, Anytime, Anywhere."
- FTS is used to manage scheduling and file transfer. For bulk transfers, CMS has historically used PhEDEx to handle transfers at the dataset (i.e., groups-of-files) level. In November of 2020, CMS will switch to using Rucio instead of PhEDEx and Dynamo to manage dataset storage and dataset transfers (while still relying on FTS underneath).
- US CMS is in the process of retiring the use of gridFTP, and replacing it with HTTP-TPC, implemented via XRootD servers. Sites typically have multiple such servers, each providing 10 Gbps, all having access to the same file system. Large bandwidth transfers are thus accomplished by orchestrating many flows across many servers.
- Fermilab's WAN architecture is based on separating its high-impact science data traffic from its general internet traffic. Conceptually, this design is analogous to a Science DMZ architecture.
- CMS does not currently use cloud resources to any significant extent. Previous studies have shown them to be a more costly model than the owned-resource model that CMS currently relies on. At this point, CMS tools are generally able to use cloud services, typically via infrastructure at one of the tiered sites, but we do not have plans to use cloud services extensively in the near or longer term.
- CMS is currently a major user of the transatlantic network links. The raw data transferred to Fermilab alone are expected to average more than 10 GB/second during HL-LHC operations. CMS tools do not prioritize site proximity (in the networking sense) when scheduling data transfers. Streaming data across the transatlantic link is allowed (even if discouraged). If the current growth rate in transatlantic link use by CMS continues, the size of the transatlantic link becomes a major limitation by Run 3.
- As is the case with HPCs, reliable networking can be used to reduce disk replica requirements either by the use of tape recall or caching. By the end of Run 4, a copy of the entire CMS MiniAOD will be around 100 PB. If 10% of this is used during any given month in a caching system, one can estimate the need for 10 PB/month of transfers to keep the cache up to date with the most recently used data. Understanding caching use cases and needs is part of ongoing R&D.

5.1.3 LHC Operations Case Study Summary

- The experiments are utilizing HPC environments, provided by both DOE- and NSF-funded facilities, for event simulation workloads. This is expected to continue into the future, particularly as new resources come online.
- Large sources of computation that exist "outside" of the experimental control (e.g., commercial clouds, HPC facilities) can be problematic to access via the LHC network infrastructure, which was designed to prioritize and facilitate intraresource communication above all. Thus, the use of "off collaboration" resources is subject to external factors (R&E peering points, commercial exchanges, etc.).
- The LHCOPN and LHCONE networks have expanded their original scope and use cases beyond design to include other facilities (e.g., HPC centers) and science use cases (e.g., DUNE, Belle II). This was simplified for a practical reason: at large DOE labs, separating traffic from one experiment becomes challenging when it is accessing other large DOE labs.

- LHCONE is currently lacking good monitoring for traffic details by experiment and traffic purpose. In addition, a single source of truth suitable for automated consumption for management and configuration is needed. Both of these are critical topics to address in the short term.
- The experiments are performing R&D for situations with constrained network resources and potentially intelligent network services. The Software-Defined Network for End-to-end Networked Science at the Exascale (SENSE) architecture, models, and demonstrated prototype define the mechanisms needed to dynamically build end-to-end virtual guaranteed networks across administrative domains with no manual intervention.
- The annual growth in network bandwidth used ranges from about 40% to 60%; 40% annual growth means doubling every two years, and x15 growth in eight years (2020 to 2028, the nominal beginning of the HL-LHC era). A 60% annual growth rate implies a x43 increase by 2028. Thus, the annual data volume for a single reconstruction version of data and simulations increases at this step function from about 22 PB to 634 PB.
- Considering transfers, remote reads for analysis, and pileup mixing, it is likely that HL-LHC computing requires 1 Tbps links for network backbones and larger sites to support ATLAS and CMS needs together with those of the other experiments. For example, CMS transfers from CERN to Tier 1s during 2018 were already peaking above the 16 Gbps level, with similar peaks generated by ATLAS. Part of this data flow is raw data: the event rate and event size will increase by factors of 7.5 and 7, respectively, in Run 4.

5.1.4 HL Era of the LHC Case Study Summary

- Each data-taking year, the LHC experiments are expected to accumulate roughly one exabyte of new data.
- The impact of HL-LHC on storage and compute resources is significant. The increase in luminosity not only generates significantly more data, but also significantly more complex events, which require more processing to resolve.
- The expanded use of resources at HPC centers will have an impact on the availability of compute resources (storage and networking) for the LHC experiments. These HPC centers are increasing in computing power, and several exa-flop scale machines will be operational during the start of the HL-LHC. These machines will be capable of producing a large volume of simulated data. The data produced will need to be quickly transferred to data centers for subsequent processing.
- The HL-LHC (i.e., Run 4) will start in 2029. The physics events that drive the experiment will be collected at a rate 10 times more than during previous runs. Challenges will be involved in collecting, storing, reconstructing, and analyzing the data volume; it is expected that Monte Carlo (MC) simulation events will need to be produced in similar numbers in the preceding years.
- Economizing storage is an important goal for HL-LHC computing. Opportunistic storage does not exist; optimizing storage by breaking out of the disk/tape paradigm to a finer grained spectrum of storage cost-reliability-latency is being pursued. This includes mechanisms to stage data from tape to a sliding window disk buffer when they are required for processing, reducing by 50% or more the input sample volume resident on disk.
- For the HL-LHC era, the predictions show a mismatch between the computing and storage resources the experiments can afford versus the resources needed to reach science goals. In response to this gap, the experiments are exploring alternatives in how to utilize storage, computing, and network infrastructure. The network baselines are currently being planned to be terabit-scale (1–2 Tbps) backbone networks with the largest resource sites connected at the

multiple-gigabit scale (200–800 Gbps). Network use will be at least a factor of 10 larger than Run 2.

- Joint ATLAS and CMS use of Rucio for distributed data management prior to HL-LHC will be an appropriate mechanism to interact with ESnet (and other R&E networks), communicating near-term data-movement intents and perhaps negotiating for any required quality of service (QoS) or deadline requirements.
- This vast quantity of data must be distributed around the globe for processing and physics analysis. The data distribution model for the HL-LHC is commonly referred to as the "data lake" model. A data lake is defined as a cluster of computing facilities that have a single entry point and multiple storage endpoints that are geographically distributed.
- Data transfer between two lakes is a top-down-controlled activity governed by Rucio and executed via FTS using third party copy HTTPS or XRoot transfer protocols with capability token authentication.
- We expect the bulk (more than 90%) of the compute resources to be used by central production workflows, while the bulk of the storage resources will be used to support end-user analysis workflows. Both types of workflows have significant data flows, and thus an impact on the networks.
- We expect that US-based processing facilities will be part of US data lakes only. Data lakes do not span the Atlantic or the Pacific. However, it seems likely that processing facilities in South America, in fact all of Latin America, will be part of the US-based data lake infrastructure.
- Networking has been fundamental to the success of LHC computing to date, enabling the exploitation of globally distributed resources for computationally limited science. This will remain the case to meet the budget-constrained computing challenges of HL-LHC. Strategies for HL-LHC computing are based on extensive use of powerful networks to reduce data replication by streaming over the net, and consolidating distributed resources into cohesive virtual federations, such as data lakes.
- For HL-LHC, four main requirements have been identified:
  - Capacity: Run 3 is moving to multiple 100 G links for large sites, while Run 4 (HL-LHC) is targeting Tbps links.
  - Capability: It is necessary to understand the impact of new features in networking (SDN/NFV) by testing, prototyping, and evaluating impact. The experiments will need to evolve applications, facilities, and computing models to meet the HL-LHC challenges.
  - Visibility: As the ESnet Blueprinting meetings have shown, the ability to understand WAN network flows is limited. New methods to mark and monitor network use are needed.
  - Testing: Developing, prototyping, and testing network features at suitable scale will be needed.
- A typical LAN configuration today aggregates worker node connections into 10 Gbps switches with multiple 40–100 Gbps uplinks to the WAN. The WAN connection is typically a (set of) 100 Gbps link(s), shared with the entire institution. It is common for the LHC program to dominate the WAN link use at Tier 2 institutions. This may change in the future.
- The LHC experiments are planning for 100 Gbps sustained use for all Tier 2s, with occasional bursts to 400 Gbps, throughout the first run of the HL-LHC.
- Tier 1s will require Tbps burst capabilities. Steady state network bandwidth consumption is expected to be between 200 and 300 Gbps, at a minimum.

- Tier 2s will require 400 Gbps burst capabilities. Steady state network bandwidth consumption is expected to be around 100 Gbps.
- The large exascale HPC centers funded by the DOE will require Tbps burst capabilities to pursue the workflows described.
- If the NSF were to fund exascale systems in the future, then those would require the same Tbps burst capabilities as the DOE systems.
- To optimally use the Exascale HPC systems of the HL-LHC era, each must be connected to ESnet at Tbps.
- It is expected that there will be some diversity in WAN connectivity for the Tier 2s.

5.1.5 Discussion

The overall schedule for LHC and HL-LHC remains similar to what was reported in the previous update. Run 3 will proceed in 2024 and 2025, the long shutdown will last between 2026 and 2028, and Run 4 is scheduled to begin in 2029. Due to energy costs in Europe, the yearly run schedule was altered in 2023 and 2024: this may result in one fewer month of operation. The LHC was not operating due to a failure at the end of 2023; an electrical perturbation from a compressor impacted several magnets. Due to this, the yearly Heavy-Ion run was not completed in 2023.

Recent R&D has produced minor changes to the data-rate estimates. The transatlantic traffic patterns may be increased to 10-20 Gbps (average) due to data-taking rates from the detector being increased. These rates will be slightly higher than in previous reports, but still within manageable levels for the networking capabilities that are planned for intrasite traffic.

The LHC collaboration is actively preparing for DC24, a set of network, storage, and computational challenges that are meant to prepare the Tier 0, Tier 1s, and Tier 2s for the increases in data expected during HL-LHC operations. The WLCG's Data Organization, Management, and Access (DOMA) group met in November 2023 to finalize one of the plans. DC24 is scheduled for late February and early March 2024, and will target a 25% increase over Run 3 data volumes. A notable change from DC21 will be the inclusion of Tier 2s, and increased instrumentation of the storage, computation, and network to better understand the macro and micro behaviors of the systems. This will address some gaps from DC21 (see Figure 5.1.2.1). The GNA-G Data Intensive Sciences Working Group and AutoGOLE/SENSE Working Group are providing input into this process.

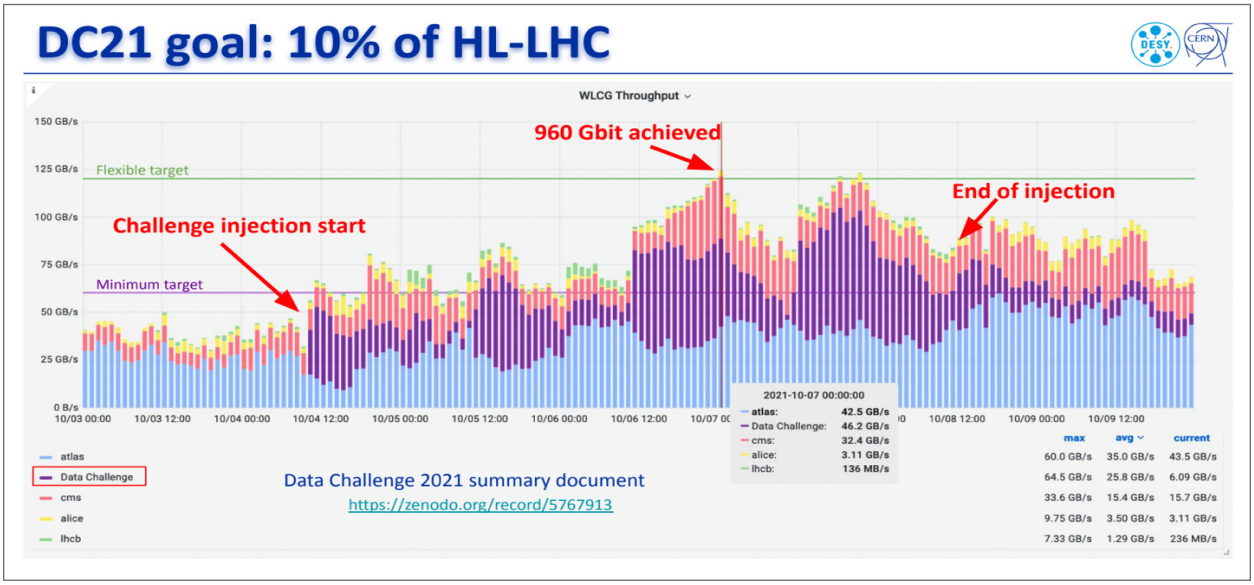


Figure 5.1.2.1: LHC DC21 Outcomes

The LHC experiments are now working with several persistent testbeds to test some of the major R&D efforts: AutoGOLE/SENSE, Global P4 Lab, Global Research Platform (GRP), National Research Platform (NRP), Atlantic Research Platform (ARP), and FABRIC. These efforts are working to integrate production tools (e.g., Rucio, FTS) with automated ways of reserving network paths and bandwidth between facilities. The work will allow more fine-grained management of network flows, and more efficient use of limited network resources. High-throughput data-transfer milestones of being able to achieve 400 Gbps disk-to-disk transfers between Caltech and the University of California, San Diego, were reached in mid-2023. Plans exist to scale this out to other sites in the near future starting with modest goals of 100 Gps for longer latency paths initially, and then stretching to 400 Gbps. The research groups are also actively talking with ESnet about the inclusion of High-Touch capabilities and how they may benefit LHC operations.

The International Committee on Future Accelerators (ICFA), composed of the laboratory directors and national representatives of HEP, is considering the formation of an Open Science Panel that will coordinate international efforts on computing, storage, networking, and workflow. A draft mandate was finalized at the ICFA Seminar in November 2023. Its mission will be related to the system mentioned previously as well as data preservation.

5.2 Neutrino Experiments at Fermilab

Fermilab features a number of experiments focused on neutrinos, an often hard to detect phenomena that will help to understand the origin of matter as well as the unification of forces. This detection can be done over both short and long distances, and the two experiments profiled, SBN and DUNE, are examples of this.

Both focus on the study of neutrino oscillations, and use a similar set of scientific technology for observation. The work of SBN and the ProtoDUNE experiments at CERN will prepare for DUNE, which is scheduled to start in several years' time. DUNE experimentation will occur in South Dakota at the SURF facility as well as Fermilab, while the SBN detectors and beamline are contained within Fermilab.

Both experiments will utilize grid-computing approaches provided by OSG software for data movement, cataloging, simulation, and analysis. The majority of cycles will be provided by Fermilab, with some use allocated to other participating sites. DUNE has the added challenge of relying on a WAN that originates at SURF in South Dakota, and must transfer all data back to Fermilab. This emphasis on near-constant network connectivity is shaping the choices made for buffering, storage, and analysis at both locations.

5.2.1 SBN Case Study Summary

- SBN will rely on a chain of three particle detectors — ICARUS, MicroBooNE, and the SBND — that probe a beam of neutrinos created by Fermilab's particle accelerators. Portions of the experiment are still under construction, with MicroBooNE (the middle detector) being currently operational. ICARUS will begin its physics run in 2021, with SBND coming into service shortly afterward.
- SBN will measure and inform behaviors that will influence the future long-baseline neutrino experiment DUNE.
- SBN event data are dominated by the data from the detector instrumentation: each instrument can sample and record behaviors during events. All these data (beam events, cosmic rays, measurements) are processed and written to storage at Fermilab.
- Raw data must be processed for analysis (signal processing, reconstruction, neutrino interaction analysis). Cosmic-ray backgrounds are a constant presence, and must be accounted for in the data.
- SBN experiments plan for a roughly yearly data production cycle. Raw data are collected from the detectors throughout the year and stored permanently, but derived data will be reproduced each year. Volumes of 6–7 PB per year (all data types) are expected.

- Most computing will be done by Fermilab, with some being handled by collaborators at OSG-affiliated sites or HPC facilities. These volumes could reach or exceed 2–4 PB per year. Domestic and foreign sites could be involved in these use cases.
- Simulation workloads (e.g., small input, large output jobs designed to create data sets used in analysis and calibration) can be done at HPC facilities such as ALCF and NERSC, along with the expected use of computational grids. The HPC use case will remain constant throughout the experiments (SBN and DUNE).

5.2.2 DUNE Case Study Summary

- DUNE is an international neutrino experiment that will be conducted with the international Long-Baseline Neutrino Facility (LBNF) at Fermilab and SURF. DUNE is under active design and construction, and will come online in the late 2020s (estimated to be 2026–2027).
- DUNE will span several US states: the beam will originate at Fermilab using a pulse rate of approximately once per second, 24 hours each day, during running periods with approximately 15 million pulses per year. The far end in South Dakota will house the detection equipment.
- DUNE far detector data generation from SURF will come in four major forms for each of the four modules: beam events, cosmic rays, supernova triggers, and calibration activities.
  - Beam events will be the smallest data volume, and will occur on the order of 41 per day, producing 6 GB per event (47 TB over the course of a year).
  - Cosmic rays will be the largest data volume, and will be seen the most frequently (4,500 per day). Each of these events will also be 6 GB in size, but could approach 10 PB per year in data volume.
  - Supernova triggers will be rare (e.g., one per month), but when observed will produce a large data volume: 115 TB per event, and 1.4 PB per year.
  - Calibration data to better understand and adapt the detector and beam will be captured twice per year, resulting in a total of 1.5 PB of data volume.
- Overall, DUNE will generate around 13 PB of data per year per module, with the project expecting to retain 30 PB of this per year on Fermilab storage.
- Supernova candidates pose a unique problem for data acquisition and reconstruction. The supernova triggers involve short, very large, bursts of data, which are collected in parallel with normal beam and cosmic-ray trigger operations. If a supernova trigger occurs, normal data may be cached locally while the supernova data are transferred. A compressed supernova from DUNE will approximately 200 terabytes (TB) in size and take a minimum of four hours to transfer over a 100 Gbps network. Instantaneous processing will be required during these windows.
- Fermilab will be the single largest provider of computation and storage resources for DUNE: current estimates are between 25% and 50% of the total that is required. The remaining resources will come from distributed OSG- and WLCG-affiliated sites (domestic and foreign), placing a heavy emphasis on networking for the overall success of the workflow. Data volumes could reach or exceed 30 PB per year.
- DUNE reconstruction and analysis will be a constant process during operation, with most computation happening at Fermilab, with other use cases leveraging the computational grid of other contributors.
- The OSG tools that DUNE will use facilitate a streaming data model, where externally operated reconstructions jobs may stream multiple GB files over the network. DUNE simulation

workflows will also be grid based and look like analysis or reconstruction, but may also occur at HPC facilities like NERSC. The data requirements for the wide area will increase as the data volumes from operation ramp up (2028 and beyond).

- The current DUNE data-transfer rate between remote sites and Fermilab is limited by the capacity of the Fermilab public dCache to sink the data. We anticipate that by the late 2020s, there will be a need to sink 100 GBs from the SURF site to Fermilab and redistribute those data simultaneously to sites worldwide.

5.2.3 Discussion

SBN has some minor updates to the *2020 HEP Network Requirements Review*. ICARUS has been taking data for two years and has no changes in operation or data rate. A new detector (SBND) will start taking data in 2024; this has the highest neutrino rate of any detector so far and will produce slightly more data than the report indicated. SBN will still use OSG resources for computing, with the majority being contributed by Fermilab.

SBN is working to port some aspects of an ML workflow to DOE HPC facilities, specifically ALCF and NERSC. This will result in less than 1 PB being sent across the network for processing. An additional preprocessing data step is being tested at SLAC, which will result in a new flow pattern, but the data volume is not expected to be high or require special support. SBN also continues to work with CNAF (Italy) on backups of data; the rate is tarted to be less than 1 Gbps.

protoDUNE work on two detectors continues at CERN. These are still under construction and not taking data, but are expected to start in 2024. A data challenge from CERN to Fermilab is planned in early 2024, and will attempt to reach 40 Gbps as a load test. The typical protoDUNE experience will be 10 Gbps expected during regular data operations.

DUNE continues to work on overall simulation and analysis workflow performance. These have resulted in a set of smaller data challenges, where a simulated data set has been shared to remote sites, analysis was performed, and results were returned. These were fractional, compared to the 30 PB/year expected when DUNE begins operation at SURF in 2028.

DUNE is exploring options for increasing access to GPUs. For now, DUNE is trying to identify resources at DOE HPC centers and talking with NERSC about resources on Perlmutter.

DUNE networking (between major collaboration sites) will target using LHCONE where applicable. This will be limited to DUNE collaboration sites that have existing LHCONE connections. The main SURF site is still limited to 10 Gbps, but looking for upgrades in 2024 or 2025. DUNE DAQ commissioning is scheduled for 2027–2028, where 100 Gbps will be needed between SURF and Fermilab.

5.3 DESI

DESI is a scientific research instrument for conducting spectrographic astronomical surveys of distant galaxies. It will utilize the Nicholas U. Mayall Telescope[4] (a four-meter telescope) located at Kitt Peak National Observatory (KPNO)[5] near Tucson, Arizona.

The overall process of science is focused on creating a 3-D map of the universe. To do this, spectral exposure of approximately 5,000 objects will be performed every 15 minutes every night over a five-year period that will aim to map 35 million galaxies. The data volumes are expected to be around 700 MB for an image, which will be combined into data sets that approach 10 GB after processing. The workflow involves use of local networking to transit the observational data periodically from KPNO to NERSC for all data processing. The resulting data products will be stored at NERSC, as well as mirrored back to Arizona, for sharing with collaborators.

4 https://www.desi.lbl.gov/telescope
5 https://www.noao.edu/kpno

Reprocessing is expected on a yearly basis, and an estimated 10 TB of data will be produced over the five-year experimental run.

Given the highly automated nature of the work, a stable and performant network is expected. Today 10 Gbps exists as provided by KPNO, although upgrades and redundancy are stretch goals. The experiment can buffer data when connectivity is lacking through the use of some local computation and storage, and a workflow manager controlled at NERSC.

DESI expects a model like other astronomical experiments, where most (if not all) user analysis will be done at data location (e.g., NERSC). A portal system, with available storage and compute, will be made available. External downloads are possible, but will not be the common use case. For instances where that is required, DESI will leverage existing NERSC infrastructure (Data Transfer Nodes and software) to facilitate transfers off-site. Use of traditional HTTP-based portals may also be required (with modern modifications), as some collaborators are more comfortable with that approach.

5.3.1 Case Study Summary

- DESI is a cosmology collaboration with the goal of creating a 3D map of the universe over a five-year runtime.
- DESI has adopted the use of a primary HPC facility located at NERSC in Berkeley, California. This will serve as the critical home for nearly all aspects of initial processing, user-level analysis, and long-term storage for the length of the project.
- DESI will observe using a 15-minute exposure that generates 715 MB of data. After observation, data must be sent to NERSC in semi-real time. Processing will result in a 10 GB data product per exposure. The data products will be returned to KPNO to make adjustments to a night's observations, or influence the targets for future nights.
- The DESI data volume at NERSC will grow at a rate of 1 PB/year, and will reach 10 PB for the lifecycle of the project (raw and processed).
- Data transfer from KPNO to NERSC is done on an approximate 10-minute cadence initiated by NERSC. Network or computational events may influence if this operation succeeds on the prescribed schedule, but ensures that all data (even in the event of pileup) will be synchronized for processing.
- DESI requires network connectivity between KPNO and NERSC to ensure stable operations. Limited buffering space is available, which helps mitigate network events that may prevent transmission to NERSC (e.g., storage of nightly results and forgoing the use of prior calibrations to influence observational behavior).
- A computational facility is affiliated with KPNO, and it is provided by the NSF's National Optical-Infrared Astronomy Research Laboratory (NOIRLab) and operated by the Association of Universities for Research in Astronomy (AURA). Mirrors of the DESI data products, after creation at NERSC, will be housed at NOIRLab. The facility can also be used for limited processing, but this is not considered a primary use case.
- User-level analysis will be conducted at NERSC through dedicated compute and storage allocations, along with yearly tasks to reprocess datasets in preparation for public releases. Data access for user-level analysis is expected to be utilized at NERSC for most use cases (using allocations and tools such as Jupyter), but can be taken off-site using tools supported by NERSC (scp, http, or Globus).
- Existing connectivity is limited to 1 Gbps for the entire shared facility (several other funded astronomical projects from different agencies), and this connection is contracted through a commercial provider to link to the University of Arizona in Tucson. From there, ample (e.g.,

10 Gbps and 100 Gbps) connectivity is available through the Sun Corridor regional network, Internet2, and ESnet.

- Network capacity is not viewed as an immediate concern, as the raw data sizes are not expected to grow for DESI on the constrained pathway between the instrument and NERSC over the course of the project. Network availability is viewed as a concern given the fragile nature of the connection between KPNO and NOIRLab (single shared 1 Gbps connection). Adding network redundancy or working toward a network capacity increase would help to add stability to the trajectory of the experiment.
- Policy level differences between operational sites have an impact on tool adoption. Policies at KPNO and NOIRLab, as NSF-funded facilities, differ from those of NERSC and DESI, which are DOE-funded experiments. This results in the use of a software toolchain that is not as efficient as it can be.
- NERSC uptime is critical. Thus, there are limited options when systemic problems (site downtime, upgrades, maintenance windows) may reduce capacity for storage or processing. Utilizing other facilities within the DOE LCF complex is not immediately possible, but could be considered if the ability to migrate resources were more readily available.
- Commercial cloud resources are also a consideration for extreme downtime scenarios, provided the workflow tools and environment could be transitioned, but this mode of operation would not be viewed as a primary or secondary.
- Long-term curation of data sets is expected to remain at NERSC beyond project lifetime, but does face a problem seen with similar collaborators: does the location, relative usability, and importance of the data indicate that a more permanent location for astronomical/cosmological data (observed or simulated) is required? The use of NSF- and DOE-funded resources for most of these projects complicates the answer, but does indicate more emphasis must be placed on creating a solution that can be applied to multiple disciplines.

5.3.2 Discussion

DESI had an initial public data release in July 2023, which resulted in over 7,000 users downloading data from NERSC. This was a major milestone, and the project is happy with the scalability of the portal for delivering this to the user community.

The DESI collaboration is satisfied with the computational and storage infrastructure, and is not making any changes since the approach documented in the 2020 HEP Network Requirements Review. The use of Perlmutter does have a downside, however; it is now a single point of interruption, and thus if the infrastructure is not available science analysis cannot occur.

DESI can report on a change to the schedule; the collaboration aspires to continue to run the instrument beyond 2026, with the associated linear increase in data volume. Discussions are ongoing with DOE HEP regarding this.

5.4 Belle II Experiment

Belle II is a third generation "B meson" experiment located at the Japanese High Energy Accelerator Research Organization (KEK). It is expected to operate through 2030, and is a worldwide collaboration (of which BNL is a major supplier of computation and storage). Analysis functions using a grid paradigm, where analysis is fully distributed around the world, and relies on data movement to migrate raw output to centers that can convert into more usable analysis formats. A set of advanced software is used to curate and control the data movement and analysis activities.

Belle II shares many similarities with the operational approaches of the LHC community, including use of some common software components modified to fit the use case. Due to the distributed nature of the collaboration

space, the use of high-speed networks (particularly those that link continents) is of high concern to ensure sound operational approaches.

5.4.1 Case Study Summary

- BNL is a major supplier of computation and storage to the overall collaboration. The Belle II raw data consist of two copies; one copy at KEK and the second copy distributed between BNL (30%), Canada (15%), France (15%), Germany (20%) and Italy (20%).
- Belle II data storage at BNL (simulated, raw, processed, and user analysis) will scale from approximately 5 PB in 2020 to more than 30 PB by experiment end (i.e., a 2 PB per year growth pattern).
- Belle II is expected to operate through 2030. Upgrades are expected in 2021, 2022, and 2026, implying some change to the underlying data volumes. Data challenges indicate as much as 42 TB/day rate could occur by 2027.
- The Belle II computing operations use a grid paradigm, where analysis is fully distributed around the world. The grid-computing model uses a three-level hierarchical structure of computing sites: raw data centers (all connected to the LHCONE overlay network), regional data centers, and MC (simulated data) production centers.
- The collaboration uses Rucio for data orchestration, which is operated by BNL. During operation, the experiment monitors latency-based interactions between the US and Japan to ensure performance remains consistent.
- BNL network and data access to support Belle II is delivered via the HTSN, the primary mechanism for all HPC and high-throughput computing (HTC) functions. This infrastructure features diverse 100 Gbps paths to ESnet, and averages multiple PB of data transferred monthly.
- Belle II's success relies on networking capacity via the R&E community provided by NSF-funded links (e.g., TransPAC[6], Pacific Wave[7]) as well as those provided by the Japanese science collaborations (Science Information NETwork [SINET]). These links provide sufficient capacity and failover for a number of projects that collaborate between the Asia pacific region and the US.
- BNL participates in LHCONE for use in both LHC experiments (LHC Tier1 for ATLAS), and Belle II. One of the most complex areas in operating this type of network infrastructure is the adherence to the LHCOPN and LHCONE Acceptable Usage Policies (AUPs). In a multipurpose lab utilizing a unified Network Perimeter, this becomes exponentially complex as scientific programs want exclusivity over a Virtual Private LAN Service (VPLS) or L3VPN circuits while utilizing BGP (e.g., LHCONE, LHCOPN or the possibility of a MultiOne deployment).

5.4.2 Discussion

Belle II was in the middle of a planned long shutdown to upgrade components of the project and restarted in late 2023. All aspects of the case study presented at the *2020 HEP Network Requirements Review* are still valid. The one recent update (that was discussed as something being planned) is the full adoption of Rucio as a data management and transfer platform. The transition has been smooth, with no issues in data mobility. The use of LHCONE continues and has not impacted science operations.

Belle II is conducting a data challenge in 2024. Effort is still needed to define the parameters, but early discussion is aiming for 40 TB/day (e.g., approximately 4 Gbps). These values should be easily achievable for Tier 0 to Tier 1s, as well as for Tier 2s.

6 https://internationalnetworks.iu.edu/projects/TransPAC/index.html
7 https://cenic.org/initiatives/pacific-wave

5.5 Muon Experimentation at Fermilab

The case study profiles two aspects of the muon research program at Fermilab: Mu2e and Muon g-2. Both focus on using particles called muons to search for rare and hidden phenomena in the quantum realm. Simply stated, muons are heavy, ephemeral cousins of the electron, living for two millionths of a second before decaying. By producing and examining the interactions, it is possible to make measurements that will help to understand other aspects of physics beyond the standard model.

Muon g-2 operated at Fermilab until the summer of 2023. Additional reprocessing is expected, and the potential for more runs exists depending on the commissioning schedule of Mu2e. All computation and storage use Fermilab connected grid-computing resources. Recent R&D efforts are looking into incorporation of ML and AI, both of which may influence future operations for Mu2e.

Mu2e is under construction, and will go into operation in 2026 with a five-year run cycle. It is expected that Mu2e will use a similar set of software and hardware to Muon g-2, with upgrades to support more storage and processing capabilities.

Both experiments utilize grid-computing approaches provided by OSG software for data movement, cataloging, simulation, and analysis. Most cycles will be provided by Fermilab, with some use allocated to other participating sites (a minority of the expected computation and storage power).

5.5.1 Muon g-2 and Mu2e Case Study Summaries

- Muon g-2 and Mu2e are two experiments at Fermilab involving the study of muon particles. These are distinct but share some common components.
- The Muon g-2 experiment started in 2018 and operated until summer 2023. Mu2e is under construction (with the first beam scheduled for 2025/2026, and five to seven years of runtime).
- The primary workflow for both experiments is to perform on-site observational science, and utilize computational grids to perform simulation, reconstruction, analysis, and long-term storage of results.
- The Muon g-2 and Mu2e experiments rely on a detector and data acquisition systems (DAQs) that capture events. Data are captured and then assembled on site. Analysis can be performed by the Fermilab team, or individuals who are collaborating via the distributed grid/software infrastructure.
- The Muon g-2 experiment will produce at least 10 PB in overall data volume (simulation, production, analysis, raw), with an upper window of 20 PB by experimental completion.
- The experiments utilize grid resources at Fermilab for analysis, as well as other distributed resources among collaborators.
- The Mu2e experiment is estimated to produce around 15 PB of data a year when running (simulation, production, analysis, raw).
- When required, data transfers to off-site resources can be on the order of small GB files to multiple TB data sets. When Mu2e enters production, a larger number of jobs (as high as 50%) may use opportunistic resources outside of Fermilab.
- Simulation for both Muon g-2 and Mu2e experiments can use grid-affiliated resources (Fermilab or opportunistic) or emerging use cases at specific HPC facilities (ALCF and NERSC).
- Both the Muon g-2 and Mu2e experiments utilize OSG[8] software (with modifications) when applicable, which facilitates a majority grid-computing use case. Migration to some forms of HPC is not feasible due to project timelines and available funding.

[8] https://opensciencegrid.org

- Data movement is typically handled as streaming, and coordinated through tools like XRootD[9] and Rucio[10].

5.5.2 Discussion

Muon g-2 ended experimental data taking after six years in July 2023. The experiment was considered a success, and met all of the defined design and experimental goals. Scientific analysis will still be performed on data; it is expected that systematic studies will continue for a number of years. These reprocessing workflows have the potential to be network intensive with between 1–2 PB per run needing to be processed, and began in October of 2023.

Reprocessing of Muon g-2 data will involve the use of OSG resources, most of which are located at Fermilab. The use of external resources is possible, but not easily accomplished due to the nature of the data sets. Distributing 1–2 PB to participating sites can be a challenge, and may not provide any overall speedup to the reprocessing effort. The analysis will complete sometime in 2024 or 2025, and will depend on the quality of the results and the available resources for the project.

The Mu2e construction project was re-baselined in fall 2022, which slipped due to pandemic delays. The current schedule has commissioning occurring in summer 2026, with a brief physics run scheduled to operate for a short period before a planned shutdown in 2027–2029. A longer run will occur starting in 2029 for a full four years. Mu2e has not changed any estimates on resource needs (computing, storage, networking), and data estimates are still accurate as published in the *2020 HEP Network Requirements Review.*

5.6 The Rubin Observatory and the LSST and DESC

Rubin Observatory, previously referred to as the Large Synoptic Survey Telescope, is an astronomical observatory currently under construction in Chile and the USA. The main task is to perform an astronomical survey, the LSST, with an expected 10-year run time. The Rubin Observatory has a wide-field reflecting telescope with an 8.4-meter primary mirror that will photograph the entire available sky every few nights. The telescope will deliver images over a 3.5-degree diameter field of view using a 3.2-gigapixel charged-coupled device imaging camera. For the purposes of the DOE, several dark energy experiments (notably DESC) will utilize data produced by Rubin on a yearly basis. The COVID-19 pandemic stopped some progress, namely the physical construction at site (which was restarted November 2020). Work on the camera has proceeded, with some promising early results in a laboratory environment at the SLAC National Accelerator Laboratory (SLAC).

5.6.1 Rubin Observatory and LSST Case Study Summary

Rubin expects to capture the entire night's sky every three days, and as a result will produce approximately 20 TB of raw data per night. These data will be streamed instantaneously from the telescope site, (possibly) through local data storage facilities, to the US Data Facility (USDF) at the SLAC National Accelerator Laboratory. ESnet will serve as a critical component in the network path, and will ultimately be used to transit portions of the US network to the USDF and to collaborating sites like DESC, which will operate at NERSC.

A primary driver for science and technology will be the ability to handle "transient" events. These are deemed to be critical observations that require immediate processing and must be completely handled within 60 seconds. This time budget allows for the event (typically based on two or more observational results) to be observed on site, raw data identified and transferred from the top of the mountain and to the USDF, processed using the analysis tool chain, and then made available through a series of brokers that will distribute the data to interested parties. A robust network (e.g., 40 Gbps, preferably with path diversity) as well as ample storage and computational infrastructure, will be required to handle these frequent events.

Outside of processing transient events, the USDF, along with a facility located at IN2P3 in France, will spend

[9] https://xrootd.slac.stanford.edu
[10] https://rucio.cern.ch

most of the year processing raw data for a yearly data release. This release will then be made available to scientists in the US and Chile, and select collaborators in countries with data rights agreements with Rubin. Rubin will follow a model of "bringing people to the data," and will make an end-user analysis platform available using dedicated computation and storage resources. It is unknown at this time how well this will scale to a potential pool of thousands of users, but there are plans to stage data trials using simulated data sets ("data previews") and both the interim cloud infrastructure, and the USDF.

- It is expected that 20 TB of raw data will be captured each evening that will flow from instrument location (Chile) to both a USDF at SLAC, and one in France (CC-IN2P3), for primary and secondary processing and storage.
- Transient events are defined to be short time window bursts (that are sized around 13 GB in two images) of objects that will require special processing. Given the transient requirements (e.g., data available to the USDF at SLAC within six seconds, and a total turnaround time of one minute) the network latency must be as stable and minimal as possible.
- A yearly release of a data-product catalog for end-user analysis is expected, and will also be used by affiliated projects (e.g., DESC).
- It is expected that 5 PB of data per year can be generated, and 500 PB by the end of the project in 2035.
- An analysis platform will be provided for end-user analysis on processed data sets, with limited bulk transfers available with affiliated projects, like DESC, to support off-site scientific reprocessing and analysis.
- Some storage and computation is available on site in Chile to support the Rubin Observatory and Chilean science community; it is expected that most (if not all) processing, reprocessing, user analysis, and long-term storage will be done by the primary USDF at SLAC, and the secondary facility at CC-IN2P3.
- The major data streams will thus be Chile to the USDF at SLAC, and the USDF at SLAC to CC-IN2P3 in France.
- Wide-area networking requirements focus on availability, latency, and capacity. To ensure stable and continuous operations, there will be a primary and secondary path to ensure continuous operation from the experiment site. Connectivity will be provided through a mixture of 10 Gbps, 40 Gbps, and 100 Gbps connections to ensure adequate bandwidth. Due to lack of control for the entire path, the international collaboration team has arranged relationships with carriers along the path (Chile, Brazil, US, and France) to guarantee operational stability.

5.6.2 DESC Case Study Summary

DESC will consume data released via the Rubin Observatory's LSST. The scientific goals include releasing analyzed and transformed data related to cosmological parameters needed for research into dark energy. This will be accomplished by taking Rubin data products (released yearly), and performing analysis at NERSC. Network connectivity between the Rubin USDF at SLAC and NERSC will be critical to ensure data flows between storage and analysis. The collaboration is still in the early stages of planning, but plans to work on simulation workflows, in addition to data trials that involve domestic and international partners (e.g., IN2P3 in France) to fully understand the capabilities and limitations of the technology in the coming years.

- DESC is heavily reliant on the Rubin Observatory as the source of scientific data for this project. As a result, the scientific facility is separate from the science community that will use the data. DESC must fully understand the available bindings to the data, software, hardware, networks (etc.) to ensure stable and reliable operational patterns in the future.
- DESC will fully adopt the "bring the user to the data" approach with scientific workflow. The

allocation of storage and computational cycles at NERSC will be used to perform operational duties (e.g., reprocessing the Rubin data yearly), but will also be used for user-level analysis of the data products.

- The largest data-movement activity will relate to the yearly data release from the Rubin USDF at SLAC to the DESC allocations at NERSC. Subsequent data movement will be ad hoc to users and to secondary/tertiary sites that participate.

5.6.3 Discussion

Construction on the Rubin Observatory has made significant progress since the last update. The camera system has been integrated at SLAC in California and will be shipped to the site in 2024. The systems are being installed and integrated at the observatory in Chile, with end-to-end integration tests performed in 2023.

The Rubin Observatory is still expecting to start the sky survey in mid-2025. There will be six months of data taking performed initially, with analysis operations expected to start in early 2026. The analysis and data expectations have not changed significantly since the 2020 HEP Network Requirements Review; 10 years of processing will result in 300 PB of data to be stored and shared. It is expected that by 2035, 1 PB a day will be entering and leaving SLAC's data facilities to support the survey.

Rubin has adopted Rucio and FTS as the data orchestration and management framework. This will be used to stage data to and from the secondary site at IN2P3 in France. User analysis is still expected to be accomplished within the Google cloud-based platform that is being constructed, with all primary data storage still being housed at SLAC.

AMPATH, the high-performance internet exchange point in Miami, Florida, has been assisting with international networking requirements between the US and Chile. The long-haul network (LHN) from Chile to Atlanta is operational, but not fully at 100 G end-to-end. The work to upgrade all segments should be completed by the first quarter of 2024. Additional work to add resiliency to the under-sea portions of the network began in 2023 and lasted into 2024. A virtual network operations center is being created, and will be managed by Indiana University's GlobalNOC group.

The DESC project has no significant changes at this time beyond adding more staff in preparation for Rubin starting to take data. The expectation is still to perform analysis at NERSC, and DESC is relying on Rubin to manage data intake and transfer to the US.

DESC has recently developed computing node hour and storage needs for the survey Year 1 analyses (2026 to 2027 timeframe). Roughly speaking, DESC will copy about 10% of the LSST survey image data from the USDF to NERSC, and redistribute some of that image data to CC-IN2P3 and IRIS. A reasonable approximation may be that DESC will need 5–10% of the network bandwidth between NERSC and CC-IN2P3, and NERSC and IRIS, that Rubin needs between NERSC and those centers.

5.7 CMB-S4

The ground-based CMB-S4 is a collaboration bringing together the US ground-based CMB community to field a single next-generation ground-based CMB experiment. This will grow to be an order of magnitude bigger than all current experiments combined. Given the collaborative nature, it is a joint effort between DOE and NSF funding, with Berkeley Lab being the lead institution on the DOE side, and the University of Chicago leading for the NSF. When CMB-S4 is complete, 3 large and 18 small telescopes will be deployed between two sites: the South Pole and Chilean Atacama Desert. Each site has a specific use case:

- South Pole will specialize in drilling down on a single ~5% sky patch with large and small telescopes.
- Chilean Atacama will be used for surveying ~70% of the sky with large telescopes.

The project has elevated the role of data management early, and as such it has been fully scoped and budgeted. The project is still in the early stages of planning, so no specific choices regarding software, hardware, or computing approach are set at this stage. There is a strong commitment to the use of "superfacility" models (e.g., joining the experimental source to computational and storage resources via ESnet and intelligent workflow tools). A critical requirement for success will be network availability from the remote sites, both of which are not in the best of environments for high-speed networking. Efforts are being made to ensure that operation can proceed with limited (or severed) resources, with the goal of increasing the available connections where possible.

5.7.1 Case Study Summary

- CMB-S4 is the result of combining cross-agency-funded science into a single project. This merger combines years of work and will have some challenges in combining the science and technology views.
- Unique opportunities for the science exist through the use of the two locations; each offers a different breadth and depth of operation. The site in Chile can observe a wider range of sky, but is not as precise. The site at the South Pole is narrower, but is more precise; thus, situations where one event is observed by both locations will offer multiple windows into the data.
- Computational and storage needs for the project are still being evaluated, but some parts are known. NERSC will be a primary facility, with other HPC and HTC facilities being added over time.
- Software (data movement, analysis, workflow) will build on existing tools, extended to meet the requirements of the unprecedented data volume and the constraints of coming architectures. It is expected that common tools from HPC/HTC use cases will be adopted where applicable.
- The CMB-S4 project will be responsible for delivering maps and alerts to the collaboration; the collaboration will then be responsible for all the subsequent science analyses. How these science analyses will be supported is still to be determined, although it is expected that this will involve some combination of allocated HPC/HTC resources and individual members' own institutional resources.
- A full set of data challenges (to evaluate computation and data movement) will be started in future years and involve the various components of the scientific workflow.
- Due to the distributed, and international, aspects of this project, network connectivity is a core concern. Connectivity to Chile has been established with international partners that are already supporting large science projects, and will scale in the coming years. Connectivity to the South Pole, on the other hand, is a major concern. Due to the use of limited satellite connectivity, the scientific transfer of data will be limited daily, and additional methods to buffer and physically ship data are required. There are several years until the project starts, and during this time the R&E community will be considering ways to fix this problem to support CMB-S4 as well as other polar programs.
- Physical infrastructure at the two sites (South Pole and Chile) may be limited. Due to this, on-site storage and compute may be scoped to deal with outage situations, but not large-scale processing or analysis.
- CMB-S4 is committed to working with ESnet on the data-movement strategy.

5.7.2 Discussion

CMB-S4 has had a number of changes since the initial case study at the 2020 HEP Network Requirements Review. Due to the cross-agency nature of the project (DOE, NSF), competing requirements and resources will impact the funding and schedule. The current trajectory for the project is still to begin operations in 2033, but

this will depend on construction schedules and budgets. CMB-S4 still plans to maintain two observations sites: Chile and South Pole.

The Chilean site will feature two large telescopes with a sustained bandwidth capacity of 1 Gbps. The precursor experiment in Chile (e.g., the Simons Observatory) has doubled in size in recent years, and has increased network connectivity as a result. Connectivity is supplied to Atacama Large Millimeter/submillimeter Array (ALMA) by the Red Universitaria Nacional (REUNA) network. Efforts to characterize the path with perfSONAR are underway.

The South Pole site aims to deploy 18 small telescopes (with a sustained bandwidth expectation of 500 Mbps), but has encountered issues in availability of network bandwidth that is shared with other experiments and limited power resources for networking, storage, and computing. The smaller (nontelescope) instruments produce data that can fit into the limited network bandwidth, so other efforts are being made to see if this can be delivered in near-real time, instead of via the slower data delivery method of a yearly shipment. As a result of the challenges at the South Pole with networking and power, CMB-S4 is re-configuring what may be possible with respect to on-site computing versus remote computing. Options include trying to bring in more power-generation capabilities, changing the overall computing and storage models, or reducing the number of telescopes.

Within the US, CMB-S4 is making progress on analysis workflows. One area of focus has been trying to reduce one of the data detection expectations from hours to minutes by redesigning some aspects of data management. The solution may involve the use of data streaming from the remote sites, and will require the use of tools that can facilitate this workflow. Rucio is being considered, since it is already in use by other HEP projects, and would need to be integrated into the computing and storage resources at NERSC. CMB-S4 is also in discussion with FABRIC as a possible way to handle transient detection for the Chilean portion of the project.

Lastly, CMB-S4 is sketching out plans for data challenges similar to those conducted within the LHC. These are still years away (the target is 2030), and many changes to the infrastructure are expected between now and then. There will be a focus on ensuring that real-time results can exit the experimental sites and be made available at the primary and secondary data analysis facilities, as well as on the replication between data analysis facilities.

5.8 Cosmological Simulation Research

Cosmological simulations are used to provide detailed theoretical predictions to understand dark matter and dark energy, cosmological constraints on neutrino physics, and the nature of primordial fluctuations. These data products are essential for analyzing and interpreting results from physical cosmological surveys, as well as aiding in survey design and optimization, and in the estimation and control of statistical errors. The level of resolution and volume required of the simulations has been, and will continue, to increase in the coming years, driving data volumes significantly upwards.

Supercomputers function as data-generating instruments to create simulations, and in practice generate data volumes that can overtake a traditional optical or microwave observation program. This is due to complexity: simulations form the basis for the creation of algorithms and software testing. Thus, the more simulations that are generated, the better the subsequent products can be in production when coupled to scientific instruments. Current volumes are already PB in size, and will grow in volume and quantity as supercomputing resources become faster and more numerous.

The data generated from the simulations exist at multiple levels, from the basic representation used in the codes (particles, grid information), to science-level information (e.g., density and velocity fields, halo information), to catalog-level information (properties of simulated galaxies). The usage pattern varies from the group that generated the data analyzing the data, to working within distributed collaborations, and to making the results publicly available. Results from major simulation runs can be useful for many years, up to a decade or more in some cases. A critical problem in this space is finding long-term locations to store the results of this work over time. As volumes

increase, and locality versus the original creation point changes, a unified view of the available simulations is required for long-term usage across the community.

5.8.1 Case Study Summary

- Cosmological simulations are typically created on HPC resources at large HPC-focused facilities (e.g., LCFs). The location where they are created is also where they are typically stored/served to those who need them.
- Simulations are typically larger than observational counterparts because they are used to help create software and bound-error calculations.
- Future data sizes for a given object catalog or sky map could be in the PB range. It is expected that the intricacy of resolution, as well as the overall volume, of cosmological simulations will increase as computational resources improve.
- Funding may span agencies (e.g., DOE, NSF, etc.) and the usefulness of a particular simulation may go on beyond the specific project funding stream, in some cases decades after creation. This causes two particular conflicts with regards to long-term storage approaches:
  - No central repository to find or track the location of surveys.
  - Storage resources that are built out of a patchwork of locations.
- Data transfer between HPC facilities has not been an issue, but transfer between HPC facilities and a user community (home institution, etc.) can be problematic due to the size of data sets, as well as not knowing the capabilities of the end-user's software, hardware, and network infrastructure. Unsophisticated users may prefer to download more than is needed, which exacerbates the problems.
- Cosmological simulation remains rooted at facilitates that support HPC. Software is created for, operated on, and shared via the resources available. Portability is possible, but not something that can happen without some modification to codes and workflow process.
- Emerging distributed analysis paradigms will complicate and exacerbate the need for a set of resources that can provide for long-term curation of simulated data sets.

5.8.2 Discussion

The cosmological simulation group reports no significant changes to the case study since the *2020 HEP Network Requirements Review*. The group is actively participating in IRI requirements gathering and hopes to see some impacts to its storage and computation challenges in future years.

5.9 LZ Dark Matter Experiment

The LZ Dark Matter Experiment is located at SURF in South Dakota and is managed primarily by Berkeley Lab. The scientific focus is on dark matter direct detection through the use of DAQ systems deployed within SURF, with analysis being performed at NERSC after the data are streamed. The experiment has an expected five-year runtime (i.e., it does not operate in bursts, and will be in a constant state of acquisition), implying that network connectivity is critical to keep in place. Gaps in connectivity can be overcome through local buffering/storage mechanisms.

The group has made all decisions about computation and storage and is awaiting experimental start. Given the use of NERSC, almost all the LZ workflow has been developed and deployed using container technology (CernVM File System), which gives a layer of protection and redundancy to cope with resource constraints that may exist at NERSC due to maintenance.

5.9.1 Case Study Summary

- LZ will explore dark matter through the use of a detector located one mile underground at SURF in Lead, South Dakota. The captured events will be analyzed by computational infrastructure located primarily at NERSC in Berkeley, California. User-level analysis is expected to be done at NERSC, through a set of computational and storage allocations, along with a set of tools that can be used.
- Data taking and analysis is expected to begin in the autumn of 2020, and will operate in stable and continuous condition for five years (e.g., continuously). The data flow, hardware, and software infrastructure will remain unchanged during this time.
- The SURF facility will have limited computational and storage resources available for LZ, and these will be viewed only as a forward buffer (~90 days' worth) to be used temporarily while data transits the network connection between SURF and NERSC.
- LZ will produce around 1 PB of data per year, with an expectation of 5 PB by project completion.
- All software tools will be deployed via containers, which allow for portability to supported systems at NERSC. This decision was made to ensure operation during maintenance windows due to the continuous nature of the experiment.
- The network connectivity between SURF and LZ is a critical component, given the workflow of doing all scientific analysis and storage off-site. The buffering capability at SURF for LZ is limited to a 90-day window, meaning that there is tolerance when connectivity is severed or reduced.
- Given the strategic importance of the facility for several DOE projects (LZ, DUNE, etc.), establishing a pathway for increased capacity, redundancy, and high-performance operation is recommended.

5.9.2 Discussion

LZ construction has been completed, and LZ has been taking data since October 2021. There are no significant changes in collaborators, system architecture, process of science, network and data architecture, and resource constraints since the *2020 HEP Network Requirements Review*. Some upgrades to hardware located at SURF have occurred, but none have altered data volumes in any significant way.

As of July 2023, 1.5 PB of data had been transferred to NERSC from LZ. LZ is limiting network traffic to 4 Gbps to not overwhelm the shared 10 Gbps connection provided by REED to SURF. This is accomplished using data buffers that were implemented to deal with computational, storage, or network outages outside of the facility.

LZ is currently the primary user of network bandwidth at SURF, until other experiments (e.g., DUNE) begin operation in the coming years. SURF is still only connected at 10 Gbps via REED, and will need to be upgraded.

List of Abbreviations

AI	artificial intelligence
ALCF	Argonne Leadership Computing Facility
ALMA	Atacama Large Millimeter/submillimeter Array
ANL	Argonne National Laboratory
AOD	Analysis Object Data
ARP	Atlantic Research Platform
ASCR	Advanced Scientific Computing Research
ATLAS	A Toroidal LHC ApparatuS
AUP	Acceptable Usage Policy
AURA	Association of Universities for Research in Astronomy
BNL	Brookhaven National Laboratory
CMB	Cosmic Microwave Background
CMS	Compact Muon Solenoid
CPU	central processing unit
DAQ	data acquisition
DESC	Dark Energy Science Collaboration
DESI	Dark Energy Spectroscopic Instrument
DOE	Department of Energy
DOMA	Data Organization, Management, and Access
DUNE	Deep Underground Neutrino Experiment
FABRIC	FABRIC is Adaptive ProgrammaBle Research Infrastructure for Computer Science and Science Applications
FTS	File Transfer Service
GPU	graphics processing unit
GRP	Global Research Platform
HEP	High Energy Physics program
HL	high luminosity
HPC	high-performance computing
HTC	high-throughput computing
HTSN	High Throughput Science Network
ICARUS	Imaging Cosmic And Rare Underground Signals
ICFA	International Committee on Future Accelerators
IRI	Integrated Research Infrastructure
IRIS	eInfrastructure for Research and Inovation at STFC
KEK	Japanese High-Energy Accelerator Research Organization
KPNO	Kitt Peak National Observatory
LAN	local-area network
LBNF	Long-Baseline Neutrino Facility
LBNL	Lawrence Berkeley National Lab

LHC	Large Hadron Collide
LHCONE	LHC Open Network Environment
LHCOPN	Large Hadron Collider Optical Private Network
LHN	long-haul network
LSST	Legacy Survey of Space and Time
LZ	LUX-ZEPLIN
MAN	metropolitan-area network
MC	Monte Carlo
ML	machine learning
NERSC	National Energy Research Scientific Computing Center
NSF	National Science Foundation
NRP	National Research Platform
OSG	Open Science Grid
PanDA	PanDA Production and Distributed Analysis System
PB	petabyte
QoS	Quality of Service
R&D	research and development
R&E	research and education
REED	Research Education and Economic Development Network
REUNA	Red Universitaria Nacional
SBN	Short-Baseline Neutrino
SBND	Short-Baseline Near Detector
SC	Office of Science
SDCC	Scientific Data and Computing Center
SENSE	Software-Defined Network for End-to-end Networked Science at the Exascale
SINET	Science Information NETwork
SLAC	SLAC National Accelerator Laboratory
STFC	Science and Technology Facilities Council
SURF	Sanford Underground Research Facility
TB	terabyte
USDF	US Data Facility
VPLS	Virtual Private LAN Service
WAN	wide-area network
WLCG	Worldwide LHC Computing Grid