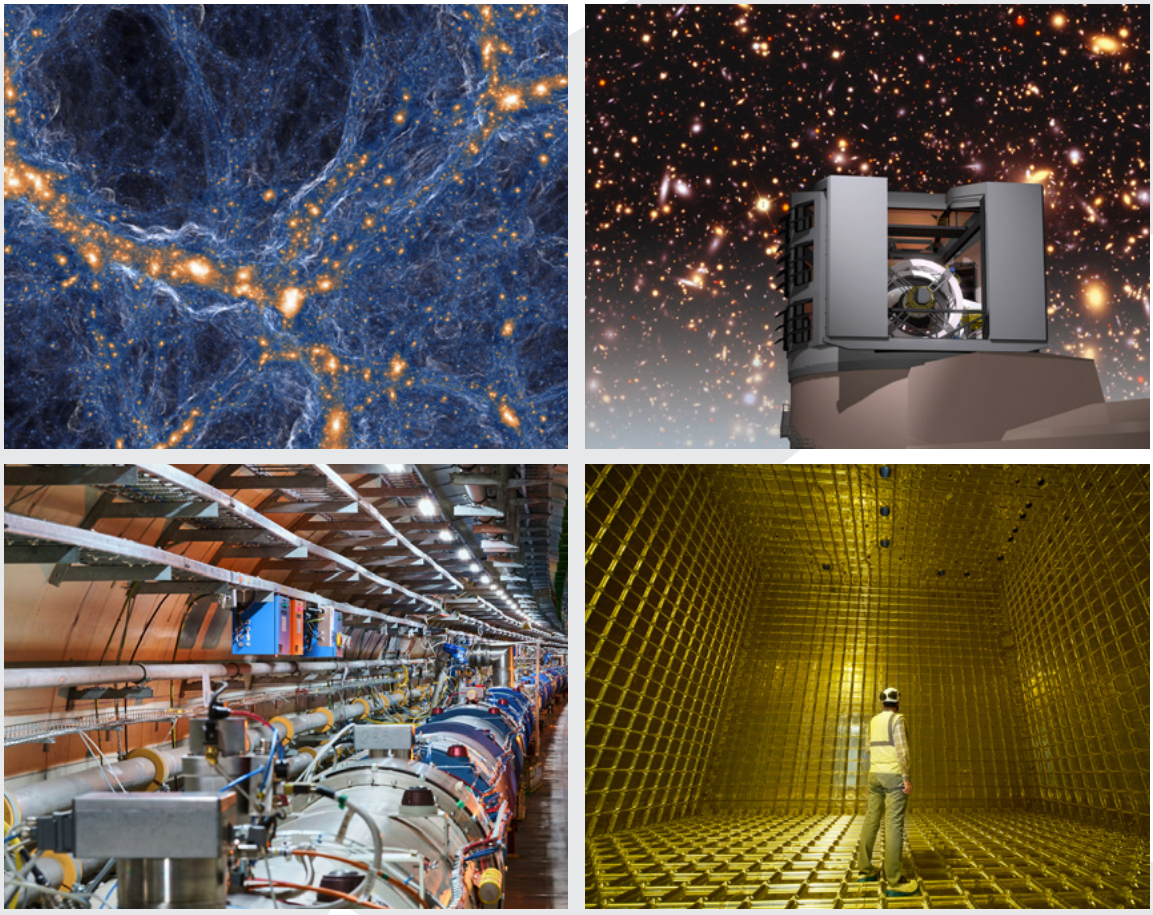
ESnet
ENERGY SCIENCES NETWORK

High Energy Physics Network Requirements Review Final Report

July – October, 2020



BERKELEY LAB

U.S. DEPARTMENT OF ENERGY
Office of Science

ESnet
ENERGY SCIENCES NETWORK

High Energy Physics
Network Requirements Review Final Report

July – October, 2020

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

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

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

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

This is LBNL report number LBNL-2001398.

Cover Images:
(**Top left**) the new universe simulation model, dubbed Illustris, courtesy of Simons Foundation
(**Top right**) Facilities Building with Simulated Night Sky, courtesy of Vera C. Rubin Observatory and Todd Mason, Mason Productions Inc. / LSST Corporation
(**Bottom left**) LHC tunnel image, courtesy of CERN
(**Bottom right**) protoDUNE detectors at CERN, courtesy of Max Brice/CERN) PHENIX image

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*

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*

Damian Hazen, *Lawrence Berkeley National Laboratory and The National Energy Research Scientific Computing Center*

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*

David Jaffe, *Brookhaven National Laboratory*

Jeff Kantor, *Rubin Observatory*

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*

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

Bill Miller, *National Science Foundation*

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*

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*

Lauren Rotman, *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, *Indiana University*

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

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

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

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

Lauren Rotman, *ESnet:* lauren@es.net

Andrew Wiedlea, *ESnet:* awiedlea@es.net

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

Table of Contents

1 Executive Summary

About ESnet

The Energy Sciences Network (ESnet) is the high-performance network user facility for the US Department of Energy (DOE) Office of Science (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 of its laboratories and facilities in the United States and abroad. ESnet is funded and stewarded by the Advanced Scientific Computing Research (ASCR) program and managed and operated by the Scientific Networking Division at Lawrence Berkeley National Laboratory (LBNL). ESnet is widely regarded as a global leader in the research and education networking community.

ESnet 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 data sets, 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.

Requirements Review Purpose and Process

ESnet and ASCR use requirements reviews to discuss and analyze current and planned science use cases and anticipated data output of a particular program, user facility, or project to inform ESnet's strategic planning, including network operations, capacity upgrades, and other service investments. A requirements review comprehensively surveys major science stakeholders' plans and processes in order 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 with collaboration from departments across the organization, including Software Engineering, Network Engineering, Infrastructure, and others. This team meets with each individual program office within the DOE SC every three years, with intermediate updates scheduled every off year. 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.

This Review

Throughout 2020, ESnet and the Office of High Energy Physics (HEP) of the DOE SC organized an ESnet requirements review of HEP-supported activities. Preparation for this event included identification of key stakeholders: program and facility management, research groups, technology providers, and a number of external observers. These individuals were asked to prepare formal case study documents about their relationship to the HEP program to build a complete understanding of the current, near-term, and long-term status, expectations, and processes that will support the science going forward. A series of pre-planning meetings better prepared case study authors for this task, along with guidance on how the review would proceed in a virtual fashion.

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 research and development (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 High Energy Physics Advisory Panel (HEPAP) are detailed in its recent long-range strategic plan, developed by the Particle Physics Project Prioritization Panel (P5)[1]. 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.

This review includes case studies from the following HEP stakeholder groups:

- Cosmological Simulation Research
- Dark Energy Science Collaboration (DESC)
- Dark Energy Spectroscopic Instrument (DESI)
- The Vera C. Rubin Observatory (Rubin Observatory) and the Legacy Survey of Space and Time (LSST)
- Cosmic Microwave Background — Stage 4 (CMB-S4)
- LZ (LUX-ZEPLIN) Dark Matter Experiment
- Muon experimentation at Fermilab
  - Muon G minus two (g-2)
  - Muon-to-electron-conversion experiment (Mu2e)
- Belle II experiment
- Neutrino experiments at Fermilab
  - Short-Baseline Neutrino Program (SBN)
  - Deep Underground Neutrino Experiment (DUNE)
- Large Hadron Collider (LHC) experimentation and operation
  - ATLAS (A Toroidal LHC ApparatuS) experiment

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

- Compact Muon Solenoid (CMS) experiment
- LHC operations
- High-luminosity (HL) era of the LHC

The review participants spanned the following roles:

- Subject-matter experts from the HEP activities listed previously.
- ESnet Site Coordinators Committee (ESCC) 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), Fermi National Accelerator Laboratory (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 High Performance Computing (HPC) facilities.
- DOE SC staff spanning both ASCR and HEP.
- Observers from other DOE SC programs and facilities.
- Observers from the National Science Foundation (NSF).
- ESnet staff supporting positions related to facility leadership, scientific engagement, networking, and R&D.

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

- Data volumes continue to increase for the profiled experiments, in most cases by orders of magnitude in the coming years (e.g., petabyte [PB] data sets are now envisioned on yearly or more frequent scales). With this significant increase in data produced by detectors, simulations, and analysis, there are concerns about the ability of supporting technology to keep pace (e.g., networks, storage, computation), particularly for highly distributed collaborations and those that involve remote components.
- International network capacity will become a significant concern for a number of collaborations in the five-year and beyond time frame, namely to support the HL-LHC era of operation for the transatlantic (TA) use case. Significant upgrades (to support Tbps [terabits per second] requirements) will be required, along with maintaining and increasing bilateral peering with research and education (R&E) partners to other worldwide locations (e.g., South America, the Asia-Pacific region, and potentially others).
- Network capacity on the domestic backbone is keeping pace with current and near-future scientific requirements. Capacity increases to site users will be an area of focus in future years. Connectivity to DOE national labs and facilities as well as to specified universities is delivered via ESnet and the DOE. Other connectivity may require external funding (e.g., universities, remote instrumentation locations, other federal agencies). Major facilities are currently connected using 100 Gbps (or better) and are expected to grow to Tbps in the coming years.
- Project portability (e.g., workflow and allocated resources) between ASCR supercomputing facilities is not easily accomplished. When experiments utilize a single facility for computation, either by design or due to resource constraints, computing resource availability and occasional downtime (e.g., due to power loss, machine maintenance, network issues, etc.) pose a considerable challenge for the progress of science. The ability to migrate analysis pipelines between computational resources is desirable, and a unified way to develop, deploy, and manage these resources could facilitate expanded usability of HPC resources.

- Previously established scientific workflows make deliberate technology choices early in their design. Those that are designed to primarily use distributed computational resources (e.g., computational grids, clouds) cannot be easily ported to adapt to architectural considerations of different approaches (e.g., HPC resources) and vice versa. Equally, optimizing these resources is heavily dependent on the underlying (and often evolving) hardware. Thus the solutions adopted have to be robust against technology and market change. The availability of alternate computing paradigms to augment existing capacity is a powerful motivator, and some R&D efforts are underway to incorporate all forms of computational resources into a workflow. This conversion of software to run across different computing architectures requires significant time and effort, and some architectural differences (e.g., data locality) still pose barriers when attempting to utilize additional computational power.
- Cyberinfrastructure (CI) innovations such as the Science DMZ, Data Transfer Nodes (DTNs), perfSONAR, and the Modern Research Data Portal continue to be important tools in supporting distributed scientific workflows. Adoption of these approaches remains high, with many collaborations requiring their implementation. Exercises such as the Petascale DTN effort[2] and Data Mobility Exhibition[3] are helping to verify and improve data movement activities across the country and globally.
- To address the growing data volumes and potential constraints in shared network bandwidth environments, considerable efforts are being put into R&D to adapt data formats, developing intelligent software for data movement and network management, and creating advanced services that can integrate workflow requirements. This collaborative effort must involve experiments, network operators, and facilities. Optimal solutions will require strategic approaches that extend beyond HEP, and must involve other SC stakeholders working to understand requirements and create synergistic technology.
- Distributed collaborations that are aligned in mission, but differ in operations and funding source (e.g., cosmological simulations that produce data products that are used by unaffiliated collaborations), typically lack a unified way to store, search, and serve critical research products. To prevent older, yet still useful, data sets (e.g., observations, simulations, etc.) from going out of circulation, creating a facility for storage, long-term management, sharing, and search is strongly encouraged.
- Partnerships that span agencies (e.g., the DOE, the NSF) are required for HEP success. The NSF's Campus Cyberinfrastructure (CC*)[4] and International R&E Network Connections (IRNC)[5] programs have enabled essential campus and international connectivity upgrades to support a wide range of science, and are critical lifelines for enabling distributed workflows and collaborations. The expanding data and capacity requirements in future years should continue to leverage agency-level cooperation to ensure that all collaborating entities have capabilities to meet scientific challenges.
- Polar network connectivity to enable remote scientific experiments is an emerging area for DOE science. Existing solutions are not directly operated by the DOE, and rely on other government agencies (e.g., the Department of Defense, the NSF) or nascent efforts by commercial entities. Investigating possible mechanisms is suggested to support research requirements. These may include ample bandwidth, low latency, and network-aware services to deliver on Service Level Agreements.

[2] Rao, Nageswara S., Liu, Qiang, Liu, Zhengchun, Kettimuthu, Rajkumar, and Foster, Ian. *Throughput Analytics of Data Transfer Infrastructures.* United States: N. p., 2019. Web. doi:10.1007/978-3-030-12971-2_2.
[3] https://fasterdata.es.net/performance-testing/2019-2020-data-mobility-workshop-and-exhibition/2019-2020-data-mobility-exhibition
[4] https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504748
[5] https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503382

- Special purpose networking overlays (e.g., LHC Open Network Environment [LHCONE]) have been incredibly useful for collaborations but are being stretched beyond original scope as the number of collaborating site users and use cases increase. Experimental stakeholders must review operational requirements and acceptable usage policy (AUP) on a regular basis to ensure alignment and resource limitations are understood and sustainable.
- The COVID-19 pandemic has influenced the scientific world in a number of ways, most notably by forcing an increase in the adoption of remote collaboration tools, but also causing scheduling changes to experimental start and run times. Progress has not been completely halted, as a number of projects have shifted to virtual mechanisms to collaborate and control experimental progress. The availability of high-performance networks remains critical to allowing continued operation.

Lastly, ESnet will be following up with participants in the coming years on a number of actions that were identified:

- ESnet will work with laboratories, user facilities, experiments, domestic networking peers, universities, partners, and international peers in the coming years to evaluate and upgrade network connections to support essential science missions.
- ESnet will continue to provide leadership in the development, implementation, and operation of cyberinfrastructure components (e.g., the Science DMZ, DTNs, perfSONAR, and the Modern Research Data Portal) that are used by the scientific user community. Emerging testbeds and technologies, including AutoGOLE, FABRIC, and SDN for End-to-End Networked Science at the Exascale (SENSE), will see increased adoption as they are deployed more widely in the coming years.
- ESnet will continue to evaluate and expand international connectivity options in support of international scientific requirements.
- ESnet will continue to work closely with LHC R&D efforts to develop and deploy new capabilities and services that address scientific use cases. These efforts include methods to better measure and track network traffic, manage and provision networks, and improve in-network caching capabilities.
- ESnet will remain active in the design, implementation, and operation of the LHCONE effort as it expands scope and usage to support the LHC and affiliated scientific experiments.
- ESnet will collaborate with ASCR computing facilities to ensure connectivity matches scientific data requirements at all stages of operation (intake, dissemination, etc.).

2 Review Findings

The requirements review process helps to identify important facts and opportunities from the programs and facilities that are profiled. The following sections outline a set of findings from the HEP and ESnet requirements review starting in July 2020 and running through October 2020. These points summarize important information gathered during the review discussions surrounding case studies and the HEP program in general. These findings are organized by topic area for simplicity and are organized by common themes:

- **Experimental timelines, collaboration, and COVID-19:** collaboration and management of projects and special topics related to addressing the global pandemic and the impacts this collaboration and management will or may have to progress and productivity of the science.
- **Domestic networking for local and wide-area uses cases:** predominantly involves issues related to provisioning of domestic network resources (local to either the experiment or to distributed sites around the country) to support the science.
- **International and transoceanic networking:** predominantly involves issues related to provisioning of international network resources and in many cases involves transoceanic connections with multiple collaborators and stakeholders to support the science.
- **Scientific data management: storage, dissemination, and volume:** topics related to the management of scientific data. This includes but is not limited to how and where data are stored, how data can be shared in structured and unstructured ways, and the increase in data volume in the near and long term.
- **Data mobility:** observations and challenges involving the transmission of scientific data. Data management activities (e.g., the long-term storage and curation of scientific data sets) overlap with sharing and transfer, as do the demands of providing networking services for this core use case.
- **Computational resources:** ways in which collaborations compute their data (e.g., HPC, high-throughput computing (HTC) / grids, cloud computing).
- **Software infrastructure:** topics related to software infrastructure of scientific experiments.

2.1 Experimental Timelines, Collaboration, and COVID-19

DESC will consume data produced by an optical telescope (the Rubin Observatory and the affiliated LSST) over a 10-year period. [Section 5.2, Section 5.4]

- DESI will use instrumentation located at Kitt Peak National Observatory (KPNO) to create a 3D map of the universe over a five-year runtime starting in 2021. DESI will capture observational data and then transfer results to NERSC in Berkeley, California, for processing. Mirrors of the data products will be housed at the NSF's National Optical Infrared Astronomy Research Laboratory (NOIRLab), operated by the Association of Universities for Research in Astronomy, Inc. (AURA). [Section 5.3]
- The Rubin Observatory will carry out the LSST using the Simonyi Survey Telescope and the Rubin Observatory LSST Camera. It is expected that 5 PB of data per year will be generated and grow to 500 PB (factoring in all project data) by the end of the project in 2035. [Section 5.4]
- CMB-S4 is ground based (with instruments located at the South Pole and the Chilean Atacama Desert) and will be jointly funded by the DOE and NSF. Its goals include detecting primordial gravitational waves and species of light relic particles, mapping the matter in the universe as it

relates to galaxy clusters, and finally detecting mm-wave transients. The project is planned to be constructed by 2028. [Section 5.5]

- LZ will explore dark matter through the use of a detector that is located at the Sanford Underground Research Facility (SURF) in Lead, South Dakota. Detector events will be analyzed by computational infrastructure located at NERSC in Berkeley, California, and backed up to a UK-based secondary facility provided by GridPP. [Section 5.6]
- The currently operating muon g-2 experiment and the planned Mu2e are located at Fermilab. The g-2 experiment will run through at least 2022 with additional runs possible if Mu2e is delayed for any reason. Mu2e first beam is scheduled for 2023, and the entire experiment will run five to seven years. [Section 5.7]
- Belle II utilizes the SuperKEKB asymmetric electron-positron collider located at the Japanese High Energy Accelerator Research Organization (KEK) in Tsukuba, Japan. It is expected to operate through 2030 and is a worldwide collaboration, with BNL as a major supplier of computation and storage to the overall collaboration. [Section 5.8]
- Belle II has used an extension of the Distributed Infrastructure with Remote Agent Control (DIRAC) framework (BelleDIRAC) to manage its distributed computing needs. As of this writing, the collaboration is migrating to a more modern data management framework, Rucio. During the transition, and during operation, the experimental operations staff will be watching latency-based interactions between the United States and Japan to ensure performance remains consistent. [Section 5.8]
- The SBN Program at Fermilab will rely on a chain of three particle detectors: ICARUS, MicroBooNE, and the SBN near detector (SBND). The program is under construction and will begin commissioning in 2021. The work of the SBN Program will prepare for DUNE, which is scheduled to start in several years' time. [Section 5.9]
- DUNE is an international neutrino experiment that will be conducted with the international Long-Baseline Neutrino Facility (LBNF) at Fermilab and SURF in Lead, South Dakota. Operations will proceed in three major phases. ProtoDUNE at CERN will stage two runs between 2021 and 2022, and will continue data reconstruction and analysis through 2025. Installation and commissioning of the far and near detectors in South Dakota and at Fermilab will occur over the period between 2025 and 2029. Physics operations running with both the near and far detectors will occur between 2028 and 2040. [Section 5.9]
- The LHC experiments are a global collaboration:
  - The ATLAS collaboration has approximately 6,000 members spread among nearly 200 institutions in 38 countries. [Section 5.10.5]
  - The CMS collaboration is made up of more than 3,000 members from more than 50 countries. Researchers at US institutions comprise about 30% of the collaboration. [Section 5.10.6]
- As a result of delays incurred due to the COVID-19 pandemic, in June 2020 the CERN Directorate issued a revised plan for the start of LHC Run 3. This plan foresees the re-start of LHC operations in February 2022. Run 3 will last until the end of 2024. All of the equipment needed for the HL-LHC, the LHC's successor, and its experiments will be installed during a long shutdown between 2025 and mid-2027. The HL-LHC is scheduled to come into operation at the end of 2027 or early 2028. [Section 5.10]
- The LHCOPN (the LHC Optical Private Network) was initially designed to offer LHC national data facility (DF) (Tier 1) sites a dedicated primary path with the central data store (Tier 0 at CERN) to ensure experiment success for data exchange. As other global science projects share

facilities with LHC science (e.g., DUNE, Belle II), it was decided that affiliated science traffic could (and now does) utilize the same infrastructure. [Section 5.10.7]

- The LHCONE is an overlay network to provide connectivity between LHC sites, especially those not allowed to use the LHCOPN, to provide a delineated science data path with a target usage policy. It is constructed through collaboration with a global set of R&E partners (computing facilities, regional and national networks, international link providers, etc.) from a variety of funding sources. [Section 5.10.7]
- The LHC and its associated experiments will undergo a major upgrade in the next six years, leading to HL-LHC operations around 2027. 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. [Section 5.10.8]
- COVID-19 related delays have permeated the HEP community and have resulted in delays in most timelines (e.g., design of new experiments, builds for others, and operations for the remainder). These delays will result in schedule changes for run times and reduced expectations on scientific output in the near term. [5.10]

2.2 Domestic Networking for Local and Wide-Area Uses Cases

- All DESC use cases will involve transmission of data from Rubin, which will store all of its data at the US Data Facility (USDF) at SLAC. The data go from SLAC to the NERSC facility for computational analysis, storage, and sharing with collaborators. After 10 years, the image data are expected to occupy about 500 PB of space, while the object catalogs would occupy about 5 PB. Use of an interim DF (hosted in the Google Cloud Platform [GCP]) will begin in 2021/2022. The goal of this interim activity is to evaluate planned operational procedures for full Rubin operation. [Section 5.2, Section 5.4]
- DESI requires reliable network connectivity between the KPNO in Arizona and NERSC to ensure stable operations. Limited buffering space is available in the event of network events that may prevent transmission to NERSC. Existing capacity is limited to 1 Gbps for the entire shared facility (which supports multiple projects funded by multiple federal research agencies). The 1 Gbps link is contracted through a commercial provider, which then connects KPNO to the University of Arizona in Tucson. From there, ample (e.g., 10 Gbps and 100 Gbps) capacity exists through the Sun Corridor Network, which connects to Internet2, and then ESnet to transmit the data directly to NERSC. [Section 5.3]
- Muon experimentation computation relies heavily on grid computational resources provided directly by Fermilab, and thus does not require extensive wide area networking (WAN). In the case of currently operating g-2, 90% of the production and analysis jobs run at Fermilab, and others may run off-site. Data transfers in these off-site cases 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%) will use opportunistic computing resources provided by other Open Science Grid (OSG) sites. [Section 5.7]
- BNL has implemented a vendor agnostic, resilient, scalable, and modular Tbps High-Throughput Science Network (HTSN), which serves as the primary network transport for all data-intensive collaborations at BNL, such as ATLAS and Belle II. It provides high-throughput connectivity to all HPC and HTC collaborations and supports the timely transfer of large amounts of scientific data via diverse 100 Gbps paths across ESnet, and averages multiple PBs of data transferred monthly. Scalability to beyond 100 Gbps is expected in future years. [Section 5.8, Section 5.10.5]

- BNL participates in LHCONE for multiple HEP experiments (LHC Tier 1 for ATLAS and Belle II). Participating in overlay networks like LHCONE brings challenges both technical and policy based, namely in adherence to AUPs. This infrastructure can be complex to support at a multipurpose lab utilizing a unified network perimeter, particularly if individual experiments want exclusivity over a Virtual Private LAN Service (VPLS) or Layer 3 Virtual Private Network (L3VPN) circuits while utilizing Border Gateway Protocol (BGP). Scaling users for these networks remains difficult, forcing choices that make BNL operate more as a service provider versus an end-user environment. [Section 5.8, Section 5.10.5]
- The DUNE far detector relies on a wide-area network that originates at SURF in South Dakota. DUNE must transfer all of its experimental data back to Fermilab. This emphasis on near-constant network connectivity is shaping the choices made for buffering, storage, and analysis at both locations. The DUNE near detector located at Fermilab will not require WAN. [Section 5.9]
- A typical LHC regional (Tier 2) facility, for its ATLAS or CMS experiments, must provide a 10 Gbps base level of network capacity. In practice, many site users are able to provide a higher level of service (between 20 Gbps and 100 Gbps). Each Tier 2 has unique LAN/WAN architecture developed in coordination with local and regional network managers. Each participant is connected through LHCONE, which requires coordination with ESnet. Given the shared nature of the connectivity, it is possible to see an average of a 15 Gbps network throughput over the course of a year, with peaks of 70 Gbps (or more). Projections for the HL-LHC, with a planned start in 2027, are a 100 Gbps average over the year, with 400 G bursts lasting hours. Not all Tier 2s may be ready for this. Preparation is possible if awareness is raised during Run 3 (2022–2024). [Section 5.10]
- Fermilab's WAN architecture is based on separating science data traffic from general internet traffic (e.g., the Science DMZ architecture[1]). Most US-CMS Tier 1 traffic is via the science data path(s), specifically across the LHCOPN and LHCONE overlays. [Section 5.10.6]
  - LHCOPN supports movement of raw data from Tier 0 (CERN) and production data movement to other Tier 1s. The LHCOPN configuration consists of three OSCARS (On-demand Secure Circuits and Reservation System) circuits (primary, secondary, and tertiary) to CERN, which provide levels of redundancy with differing bandwidth guarantees for that traffic.
  - LHCONE supports production data movement use cases (e.g., to other Tier 1s not on the LHCOPN, Tier 2s, and some Tier 3s). Connectivity to the LHCONE is via geographically redundant (primary/fail-over) paths.
  - US-CMS Tier 1 WAN traffic that does not utilize either the LHCOPN or LHCONE paths traverses the laboratory's general internet path instead.
- Fermilab currently has three 100 Gbps links to ESnet via a geographically redundant metro ring. Two 100 Gbps links are used to support the science data network paths, including LHCOPN and LHCONE. The third 100 Gbps link supports the laboratory's general internet traffic, which includes non-categorized CMS traffic to and from locations that are not participating in LHCONE. [Section 5.10.6]
- The data volume that can be handled by the networks within, and coming out of, the CMS detector facility far exceeds what can be handled offline within current central processing unit (CPU), storage, and networking infrastructures. Therefore, data are reduced by a multistage compute facility (e.g., TriDAS, the combined Trigger and Data Acquisition System) close to the detector, in an effort to retain data of highest interest to the CMS physics program as possible. [Section 5.10.6]

[1] E. Dart, L. Rotman, B. Tierney, M. Hester, and J. Zurawski, "The Science DMZ: A network design pattern for data-intensive science," *SC '13: Proceedings of the International Conference on High Performance Computing, Networking, Storage, and Analysis*, Denver, CO, 2013, pp. 1–10, doi: 10.1145/2503210.2503245.

2.3 International and Transoceanic Networking

- The global use of R&E networks will have a different character on the timescale of the HL-LHC; e.g., other science domains (astronomy, biology, engineering, and others) are emerging as powerful and prolific network use cases, and the availability of networks is based on the fact that they are finite resources that must be shared. Early adopters of networks built technology footprints (e.g., computation and workflow software) that did not take into account costs and quantities. Understanding capacity trajectory, in addition to how new intelligent network services developed through prototypes, is critical for the future success of LHC and other science collaborations. [Section 5.10.8, Section 6]
- The Rubin Observatory will produce approximately 20 TB of data per night, which will accrue at a rate of around 5 PB per year, and 500 PB (factoring in all project data) by project end in 2035. These data must be transferred from the Rubin Observatory location (Chile) to a USDF at SLAC for primary processing and storage. A secondary DF located in France (CC-IN2P3) will also receive a copy of the data sets for additional storage and processing work. WAN requirements are focused on availability to transfer, within seconds of observation, latency to the USDF at SLAC and bandwidth capacity to accommodate the scientific data volume. To ensure stable and continuous operations, there will be a primary and secondary path to ensure continuous operation from the experiment in Chile to the USDF at SLAC. Connectivity will be provided through a mixture of 10 Gbps, 40 Gbps, and 100 Gbps connections to ensure adequate bandwidth both domestically and internationally. [Section 5.4]
- CMB-S4 has initial estimates of data volumes from Chile approaching 14 TB and up to 8 TB at the South Pole, for nightly observations. Due to the remote nature of both sites, network connectivity is challenging to arrange and to keep stable over time. The South Pole data volume in particular vastly exceeds current available capacity of existing satellite networks (currently less than 100 Mbps, which would require a transfer time of a week or more to handle a single day's data volume back to the United States). Without a significant increase in bandwidth, bulk-data transfer will be limited to annual shipments of disks, which could have significant impacts on overall data quality and systematic errors due to intermittent detector monitoring as well as the potential risk of total loss without available backup capability. The data will be transferred to a main USDF at NERSC, where processing and analysis of the data will occur and long-term storage will be conducted. [Section 5.5]
- The success of Belle II relies on transpacific networking capacity provided by a number of R&E partners (and often funded by the NSF) that participate in the LHCONE overlay network. [Section 5.8]
  - Belle II connects directly to the Science Information Network (SINET) and Asia Pacific Advanced Network (APAN) networks in Tokyo, Japan, as the first leg of a multiple provider path.
  - SINET then connects to the United States via multiple direct and indirect paths:
    - 100 Gbps via PacWave to Los Angeles, California.
    - 100 Gbps via TransPac to Seattle, Washington.
    - 100 Gbps via the Japan Gigabit Network (JGN), which connects through Hong Kong and Guam in collaboration with the Pacific Islands Research and Education Network (PIREN) and TransPac.
    - 100 Gbps via the JGN via Singapore, where peering to other R&E partners is possible.
  - PacWave facilitates connections to ESnet, Internet2, CENIC, and others.

- The SBN Program's raw data backups will require ~75 GB/hr (~20 MB/sec) to CNAF (the national center of INFN [Italian Institute for Nuclear Physics] in Italy) for the ICARUS detector, and 130 GB/hr (~35 MB/sec) for the SBND detector (location not yet identified). [Section 5.9]
- The DUNE project currently relies on ESnet TA connectivity to provide access between Fermilab and ProtoDUNE, which is operating at CERN. To date, the network requirements for this experiment have not run into congestion or network capacity problems, and are not expected to through the end of the ProtoDUNE run through 2022. [Section 5.9]
- LHCONE is designed to provide a friction-free network by offering a succinct usage policy and technology profile to fit science use cases different from those of traditional networks that have performance impediments. Connecting to LHCONE requires compliance with an AUP that may not be readily adoptable by non-LHC facilities that are nonetheless capable of providing resources (e.g., commercial clouds, HPCs). To utilize LHCONE for these resources, there would need to be an effort to explore how to dynamically identify LHC activities and how to then connect to specific aspects; this is an area in which research into WAN network orchestration may yield solutions. [Section 5.10.7]
- A critical component to the success of HEP-funded experiments and facilities (e.g., the LHC, Belle II, DUNE, the Rubin Observatory, and others) is the use of shared R&E resources that facilitate international networking. A number of connectivity options are in place today provided by different funding sources. ESnet (funded by the DOE) has multiple dedicated TA 100 Gbps paths provisioned to the LHC. Other science use cases between the European Union and United States now also take advantage of these connections. The NSF also funds several links for general-purpose use by the R&E community. Other consortia of R&E operators have also collaborated to ensure traffic sharing and peering arrangements that make capacity available in the event of link maintenance and failure. Additional capacity to address these cases must be on the roadmap, and the costs related to implementing new connectivity options are continuing to drop, thus making acquisition and implementation of new resilient capacity achievable. [Section 6]
- To ensure efficient use of TA bandwidth, changes will be required to the operational approach to data mobility software and practices. The "any data, anywhere, anytime" model adopted by the LHC community can lead to situations where nonoptimal choices are made during data migration (e.g., selecting paths that have a long latency, may feature lower bandwidth, or are congested due to other use cases). As data sets' sizes grow, and more users compete for available resources, three main areas of growth should be considered in this space: (a) introducing more compact analysis formats to reduce the required transfer sizes; (b) securing more bandwidth to meet growing experimental data set size and volume increases; and lastly (c) altering the data staging approaches to leverage more intelligent methods (caching, staging into "data lakes," facilitating fetching from more geographically relevant locations). [Section 5.10]
- Connectivity between the LHC at CERN and the distributed computing centers that make up the tiers is complex and relies on a tapestry of partnerships that provide networking. Traditionally the concern has been international R&E connectivity between continents, which has been historically limited and a critical bottleneck to overcome. There are concerns that, in the future, domestic networking capacity (e.g., between the national R&E backbones and site users, almost always provided by a regional network operator), may become a limiting factor as the number of use cases increases and capacity upgrades do not keep pace with demand. R&D efforts (some ongoing, some planned) have resulted in new ways to manage data mobility challenges without putting increased strain on networks, but this will not supplant the need for network capacity upgrades in the coming years to ensure that the distributed scientific mission

is successful. Programs like the NSF CC*[2] efforts to upgrade university campus infrastructure are critical to ensuring success of science programs. Participation is encouraged as a means for upgrading campus infrastructure where many DOE collaborators are housed. ESnet will also continue to work with DOE-connected labs and computing facilities to ensure that capacity is not a factor on the wide-area network, but has minimal control over local area handling of networking to computation and storage infrastructure. Emulating a program like CC* within the DOE could facilitate upgrades to meet the needs of the science, providing true end-to-end network performance. [Section 5.10]

- The success of Belle II relies on transpacific networking capacity provided by a number of R&E partners (and often funded by the NSF) that participate in the LHCONE overlay network. ESnet does not currently control a transpacific network link, and relies on the circuits provided by collaborating entities to ensure BNL/Belle II success. Belle II has a distributed software environment, with BNL running services critical for the operation of the Belle II facility in Japan. ESnet should investigate the risk associated with this operational reality. [Section 5.8]

2.4 Scientific Data Management: Storage, Dissemination, and Volume

- Cosmological simulation output will continue to increase resolution and scope, which implies an overall increase in data volume on a per-product and per-catalog basis. Current data set sizes are in the terabyte (TB) range; emerging and future data sets will reach petabyte size. These datasets often exceed the size and usable lifespan of their optical and microwave observations, particularly as they are created and used for a number of years before observations may begin during the facilities' life cycle. Longevity of tens of years is common, and usage is variable to the number of ongoing projects and users under design or operation. [Section 5.1]
- Cosmological simulations can greatly benefit from a unified solution for storage and sharing of data products. Currently they are housed at a creation facility (e.g., DOE HPC facilities, a university, or other location that provides computation). At the current time, no unified way exists to search or retrieve data due to the nature of the funding sources and resource allocation. To ensure long-term availability of data products over time, as well as improve the usability for a number of communities, implementation of mechanisms to store, locate, and share these simulations are recommended. This can be centralized or distributed with a unification scheme. [Section 5.1]
- DESI will capture observational data during a 15-minute exposure (resulting in 715 megabytes [MB] of raw data), and then transfer the raw results to NERSC for processing into 10 GB data products. 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). [Section 5.3]
- The Rubin Observatory does not expect extensive off-site data use, and will provide a data access platform designed for on-site analysis by the user community. Off-site use for affiliated projects (e.g., DESC) will be organized in a structured manner to allow for bulk-data movement (potentially yearly, to coincide with data product releases). The major data streams will thus be Chile to the USDF at SLAC, and the USDF at SLAC to France. [Section 5.2, Section 5.4]
- CMB-S4 will rely on two observational locations with multiple telescopes: the South Pole and the Chilean Atacama Desert. In aggregate, 22 TB (~8 TB at the South Pole and ~14 TB in Chile) of data will be generated daily, leading to an accrual of 3 PB of data annually. The total data set by project completion could reach 100 PB. [Section 5.5]

[2] https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504748

- CMB-S4 is planning for the capability to perform on-site analysis, and the ability to buffer data in times of network blackout that may be caused when network connectivity at either remote facility is not available. The experiment is creating mechanisms that can be used to reduce data set size in these circumstances that would prioritize higher value observations during blackout periods to save storage space and optimize the available computational power. It is fully expected that the South Pole instruments will physically send higher resolution data over mechanical means on a yearly basis and utilize limited network capabilities for critical events only. [Section 5.5]
- LZ data taking at SURF, and analysis at NERSC, is expected to begin in the autumn of 2020, and will operate in stable and continuous condition for five years. Experiment data flow, hardware, and software infrastructure will remain unchanged during this time (at both SURF and NERSC), and LZ will produce around 1 PB of data per year (all varieties) with an expectation of 5 PB by project completion. [Section 5.6]
- The g-2 experiment at Fermilab will produce at least 10 PB in overall data volume (simulation, production, analysis, raw) through 2022, with an upper window of 20 PB by experimental completion. [Section 5.7]
- The Mu2e at Fermilab is estimated to produce around 15 PB of data a year when running (simulation, production, analysis, raw) during the five to seven years of run time starting in 2022 or 2023. [Section 5.7]
- 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 (e.g., 2 PB per year growth pattern). Belle II expects upgrades to the instrument in 2021, 2022, and 2026. Data volume could increase by more than five times as a result of these upgrades, and data challenges indicate as much as a 42 TB/day rate could occur by 2027. [Section 5.8]
- SBN Program event data will consist of beam events, cosmic rays, and detector measurements, all of which will be processed and written to storage at Fermilab. The data lifetime (derived and bulk) is two years and is expected to be 6–7 PB per year. [Section 5.9]
  - Raw data will arrive from ICARUS at a steady rate of about 370 GB/hr (~100MB/sec).
  - Raw data will arrive from SBND at about 320 GB/hr (~90 MB/sec).
  - All data will be delivered to Fermilab and be made available for immediate processing, which will run on-site at Fermigrid, but can be accessed and run at OSG sites.
- The DUNE near detector at Fermilab will generate around 250 TB of local used/stored data per year for the tasks of beam and cosmic events, as well as calibration. The near detector will not leverage WAN connections for data movement. [Section 5.9]
- 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. 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. [Section 5.9]
  - 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: between 100–200 TB per event, and 1.4 PB per year that must run in parallel with beam and cosmic-ray trigger operation (some caching may be permitted of the

latter two). Instantaneous processing will be required during these windows, resulting in an extreme need for reliable and predictable networking between SURF and Fermilab.

  - 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.

- Datasets in ATLAS are collections of files organized by category/workflow. Individual datasets vary largely in size. [Section 5.10.5]
  - Raw datasets are in the range of 1 to about 50 TB.
  - Analysis object data (AOD) datasets are in the range of 1 GB to about 50 TB.
  - Derived AOD (DAOD) datasets are in the range of 1 GB to about several TB.
  - HITS datasets are in the order of several TB.
- In Run 3 of the LHC, starting in 2022, a major change will be implemented with regards to file formats in ATLAS and CMS; the smaller DAOD_PHYS and NanoAOD formats will be preferred to facilitate less network use and faster computation. This will reduce the number of the larger AOD formatted files in active circulation for both experiments. [Section 5.10.5, Section 5.10.6]
- CMS data formats range from the most versatile and complete (raw and AOD) to the easiest, smallest, and fastest to use (MiniAOD and NanoAOD). Data formats differ in the level of detail stored per collision. [Section 5.10.6]
  - Raw data size is approximately 1 MB currently and will grow to 6.5 MB during Run 4.
  - AOD format data are reduced to approximately 400 KB but will be approximately 2 MB during Run 4.
  - MiniAOD is approximately 60 KB currently and will grow to 250 KB during Run 4.
  - NanoAOD is approximately 1 KB in size and will grow to 2 KB during Run 4.
- The CMS NanoAOD format, designed for interactive end-user analysis, will be more widely adopted in its Run 3. It is anticipated that this format will be utilized in 50% (or more) of data transfers. Given that the NanoAOD format is more than an order of magnitude smaller than MiniAOD, it will reduce network bandwidth and increase processing speeds. While other formats can be used for analysis, the goal is to keep these potential use cases to a minimum. At HL-LHC scales, CMS may not be able to afford to keep larger formats (e.g., AOD) on disk. In that scenario, access to AOD would require retrieval from archival storage, which would increase the complexity and resources required. [Section 5.10.6]
- CMS produced approximately 45 PB of raw data during the four years of operation for Run 2, and a roughly similar set is expected for Run 3. There were no major technology upgrades beyond changes to file formats on the analysis side. Run 4 will usher in a new era of scientific technology, and 350 PB per year is expected starting in 2028. [Section 5.10.6]
- CMS has produced simulations of roughly two to three times as many collisions and plans to continue this practice during Run 3. CMS has about 140 PB of Run 2 Monte Carlo (MC) simulation datasets; this is representative of the total four-year production of simulation. HL-LHC currently envisions roughly the same number of events from simulation as from the detector (e.g., 300+ PB). [Section 5.10.6]
- Near the end of a typical year, ATLAS and CMS perform a "reprocessing" phase where raw data are repeated and run through the most recent software and analysis infrastructure to recreate experimental results. This is also performed at the end of the run cycle, coinciding with experimental shutdown. [Section 5.10.5, Section 5.10.6]

- Both ATLAS and CMS have a "steady state" annual growth in network bandwidth consumed ranging from approximately 40% to 60%. [Section 5.10.5, Section 5.10.6]
  - A 40% annual growth means x2 every two years, and a total of x15 growth in eight years (e.g., through 2028).
    - A 60% annual growth rate implies a total of x43 increase by 2028.
    - Step-function increases between 2024 and 2028 (to coincide with HL-LHC) may be required for the data volume changes.
    - Investment into network infrastructure at all layers (facilities, regional providers, national backbones, and international connectivity) must keep pace, along with efforts in R&D to reduce data set expectations in the absence of capacity improvements.
- The HL-LHC, commencing with Run 4 in 2027, will deliver unprecedentedly complex events. These events will be collected at a rate ten times more than during previous runs. Each data-taking year during the HL-LHC, the experiments ATLAS and CMS combined are expected to accumulate roughly 1 exabyte (EB) of new raw data, which will require upgraded network capabilities across the world to ensure a smooth and efficient pipeline to link computational and storage resources. [Section 5.10.8]
- In the expected three-year operational time of the first HL-LHC run, it is expected that the experiment will accumulate roughly the same amount of integrated luminosity of data that has been collected during the entirety of the LHC experiment. This implies that the science capabilities are expected to be roughly equivalent to the data taken from 2010–2024, or runs 1, 2, and 3 combined. The entire HL-LHC era will last for 10 years, with 12–24-month maintenance periods interspersed roughly every three years. [Section 5.10.8]
- The adoption of new, and more compact, data formats (e.g., DAOD_PHYSLITE for ATLAS and NanoAOD for CMS) will have a large impact on resource use: smaller files can be transferred faster, will take up less storage space, and can be processed more quickly. By increasing adoption of these technologies, and reducing use of legacy formats, network bandwidth on critical paths can be conserved. This will affect several parts of the workflow (analysis, simulation, and production). [Section 5.10.5, Section 5.10.6]
- Through the review it has become evident that across the entire DOE SC landscape, a succinct definition for a "data set" remains elusive between facilities, experiments, and projects. This makes intra-group communication challenging as the ability to accurately depict this unit of measurement to those that provide computation, storage, or networks highly differs. [Section 6]

2.5 Data Mobility

- The cosmological simulations community is adept at data movement between major facilities (e.g., DOE HPC facilities) due to assistance from ESnet and the Petascale DTN effort. It is now routine to translate TB- and PB-sized data sets between well-connected locations. Network mobility to other collaborators can still be problematic due to end-to-end network performance issues at the destination. This applies to both streaming and bulk-download use cases. [Section 5.1]
- DESI data transfer from KPNO to NERSC is done on an approximate 10-minute cadence initiated by NERSC. 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). These data sets are expected to be on the order of GB size, or in rare instances, TB size for yearly data releases. [Section 5.3]

- It is expected that the Rubin Observatory will generate 20 TB of raw data each evening that will flow from the Rubin Observatory location (Chile) to both a USDF at SLAC and one in France (CC-IN2P) for processing and storage. There are also time-sensitive requirements on certain nightly captured data streams. In addition to the regular mode of operation, there exists the possibility of capturing "transient" events. These are defined as short-time window bursts (approximately 13 GB) and will require near-real time processing at SLAC (e.g., within one minute of capture). [Section 5.4]
- All Belle II raw DFs are affiliated with the larger Worldwide LHC Computing Grid (WLCG) effort, and are connected to the LHCONE overlay network to facilitate movement. [Section 5.8]
- The nominal event size from the ProtoDUNE Single-Phase (SP) detectors was 180 MB via the data acquisition system (DAQ), resulting in a final compressed event size of approximately 75 MB. Early data challenges over a period of six weeks produced over 570 TB in raw beam data, and 1 PB in cosmic-ray data. These were divided into 8 GB files, consisting of 100–130 events each. A sustained rate of 2 GB/sec was achieved between Fermilab and CERN for network and disk performance. [Section 5.9]
- Organizations including the US ATLAS/CMS operations programs, IRIS-HEP, and the WLCG Data Organization, Management, and Access (DOMA) working groups are organizing data challenges to build up to the scale projected for HL-LHC. One challenge under development is meant to test the ability to transfer and process 10 PB of data at an HPC center, and then return the output back to the sending source. To do this under realistic HL-LHC scenarios will require data transfer speeds of 1 Tbps over the course of a single day. This particular use case corresponds with the scale of data movement required to process an entire year's worth of HL-LHC data, but processed in a window of 100 days. It brings a number of technical challenges, including tape recall, managing disk buffers, and managing network usage. [Section 5.10.8]
- All CMS researchers can access storage via streaming or by grid analysis jobs. This is called "any data, anytime, anywhere." This model is an evolution from the original tiered model where connectivity was extremely limited. When the tiers were hierarchical, data flowed between adjacent levels only. The global CMS collaboration, together with WLCG and OSG, is working to better define services that participating members perform. In the future, services may return to a more centralized model to conserve resources as data sizes increase. [Section 5.10.6]
- CMS uses "top-down data placement" at Tier 1 and Tier 2 centers, combined with intelligent applications that are able to specify specific datasets that are required. When this occurs, the executable jobs are then automatically routed and executed at the location that is currently housing the data. This mode of operation prevents the need to have to rely on a wide-area network to transfer data, and causes all data access to be local, which saves networking resources. [Section 5.10.6]
- 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 tradeoff is reasonable when there is little input/output (IO) per CPU, but does not work in some architectural paradigms (e.g., HPC facilities) where computational units may allow only access to local data. [Section 5.10.6]
- CMS can perform "bottom-up data placement," as is implicit in caching. Here the applications are routed to localities with caches, applications access the cache locally, and cache misses are handled by the CMS XROOTD Data Federation. CMS allows access in this fashion for all of the MiniAOD and NanoAOD formats, but nothing that may not be disk resident (e.g., AOD and raw). Caching is expected to become the dominant data access for end users of MiniAOD and NanoAOD formats. [Section 5.10.6]

- CMS central production workflows use top-down data access and streaming access for workflows that have a very small data to CPU ratio. Caching access is not (yet) used in central production. [Section 5.10.6]
- The LHC experiments delegate wide-area network performance measurement collection, via the perfSONAR framework, to the OSG and will continue to do so for the foreseeable future. All Tier 2s and the Tier 1 are expected to keep up this perfSONAR instrumentation with the bandwidth requirements. [Section 5.10.7]
- In the HL-LHC era, exploiting networking resources must be balanced against the availability of storage and computational resources. Given the distributed nature of the ATLAS and CMS collaborations, networks are critical for operational success, and usage can be improved via novel approaches that measure, monitor, and better utilize available resources: [Section 5.10.8]
  - Reducing data replication by streaming over the network and consolidating distributed resources into cohesive virtual federations, such as data lakes.
  - Marking network packets to identify sources of traffic.
  - Economizing storage resources by keeping valuable data available, and archiving and compressing older, less valuable data sets.
  - Developing new services, improving existing tools, and tuning workflows to deliver more targeted fine-grained data sets as needed.
- During the HL-LHC era, it is expected 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 and Latin America will be part of the US-based data lake infrastructure. [Section 5.10.8]
  - A data lake is a collection of computing facilities that have a single entry point for interactions outside the data lake. The data lake model is under conceptual development, and its technologies (e.g., enhancements to Rucio and File Transfer Service [FTS]) are part of the HL-LHC R&D program.
  - Both caching and streaming are foreseen as use cases for data movement within a lake for lakes that are attached to processing resources.
  - The concept of data streaming could only be supported for applications and between locations where round-trip time (RTT) is small enough to guarantee good performance.
- Many distributed experiments that utilize a grid-computing paradigm (e.g., LHC experiments, muon and neutrino experiments, Belle II, etc.) have adopted the use case of any data, anywhere, anytime, which is shorthand notation for being able to locate, download, and perform analysis on experimental data wherever capabilities exist and using approaches to "stream" information across networks on demand. This behavior treats the wide-area network as an appliance with almost infinite capacity; thus network capabilities have become more and more an integral part of the computing model. This approach cannot scale infinitely, as data set sizes and network capacities do not grow at the same rate. To make better use of limited network resources, R&D efforts have focused on: [Section 6]
  - Adoption of smaller data formats for sharing and computation.
  - Utilization of caching approaches (simple and intelligent) to better position data closer to where they may be needed.
  - Changes to the underlying tools to allow better understanding of network capabilities, and data locality.

  - Creation of regionally deployed storage and computation resources to absorb some of the requests more centrally and better scale scientific use cases.
  - Creation and adoption of smarter infrastructure that can be managed more efficiently and dynamically.
- Some projects/experiments (such as LZ, DUNE, and DESI) feature a "two-sided" workflow: data are migrated from an instrument to a first level of computation/storage, and then a second that involves a network to more plentiful/high-performance storage (the latter could be fully distributed on a grid or to HPC/DFs). From there, additional "fan out" of data to users may be possible. Automation of these two different sides has increased significantly in recent years, which has facilitated increased network usage, and sometimes limited available resources. Mechanisms to monitor bandwidth and reduce usage during times of resource exhaustion, or during high-profile use cases such as a time-sensitive event, are needed. [Section 6]

2.6 Computational Resources

- Cosmological simulations are performed at a number of computational facilities, including Argonne Leadership Computing Facility (ALCF), NERSC, and Oak Ridge Leadership Computing Facility (OLCF) as well as NSF computing facilities, along with institutional compute resources available at the collaborating institutions. [Section 5.1]
- DESC user-level analysis will be performed almost exclusively at NERSC and CC-IN2P3 (France) through the use of dedicated computational and storage allocations, analysis toolchains, and other aspects of project curation. The collaboration will encourage use of these resources, but will also make data sets available for off-site use. [Section 5.2]
- DESI will perform a yearly "data assembly" effort at NERSC to reprocess all of the data taken to date. The product of this will be publicly released after the initial set of cosmology papers is published by the DESI collaboration. DESI projects that the total data volume will grow to 10 PB by the end of the survey lifespan in 2025. [Section 5.3]
- DESI leverages cycles, workflow, and data movement tools supported by NERSC for nightly observations. As a result, uptime is critical. Thus there are limited options when planned downtime or systemic problems arise. Utilizing other facilities within the DOE HPC facility complex is not immediately possible, but could be considered if the ability to migrate resources was more readily available. Commercial cloud resources are a consideration for extreme downtime scenarios, but this mode of operation would not be viewed as a primary or secondary. [Section 5.3]
- The Rubin Observatory will provide an analysis platform 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. This will run at resources at the USDF at SLAC and in France at CC-IN2P. An interim data facility (IDF) will be utilized via the GCP between 2021 and 2023 and limited to testing operational readiness. The IDF is active as of October 2020. [Section 5.4]
- The primary CMB-S4 DF will be NESRC, with secondary facilities established at the ALCF, OSG, the Extreme Science and Engineering Discovery Environment (XSEDE), and potentially via the FABRIC project. [Section 5.5]
- LZ computation will be performed by NERSC. SURF has 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, and then from NERSC to the UK for backup purposes. [Section 5.6]

- The scientific workflow for g-2 and Mu2e involves the use of computational grids (the majority of resources provided by Fermilab) to perform simulation, reconstruction, analysis, and long-term storage of results that are collected on-site. HPC resources (NERSC and ALCF) are being used by g-2, and may be used by Mu2e, for the simulation aspect of the workflow. [Section 5.7]
- Belle II employs a grid paradigm, utilizing a three-level hierarchical structure of computing: raw data centers (all connected to the LHCONE overlay network), regional data centers, and MC (i.e., simulated data) production centers. Belle II analysis is fully distributed around the world and relies on data movement to migrate raw output to facilities that can convert the data into more usable analysis formats. The Belle II raw data consist of two copies: one copy at KEK and the second copy at BNL. Beginning 2021, the second copy will be distributed between BNL (30%), Canada (15%), France (15%), Germany (20%), and Italy (20%). [Section 5.8]
- Both SBN and DUNE will utilize grid-computing approaches provided by the OSG software for data movement, cataloging, simulation, and analysis. In the case of SBN, the majority of computation cycles will be provided by Fermilab, with some use allocated to other participants around the world. DUNE will use a more distributed approach, where Fermilab will still provide the largest number of computational and storage resources (between 25% and 50%). The remainder will be provided by domestic and internal partners through OSG and WLCG contributions. [Section 5.9]
- SBN computing will be done by Fermilab, with some handled by collaborators within the OSG or at the DOE HPC facilities. These volumes could reach or exceed 2–4PB per year, depending on the needs for external processing and available network bandwidths. Domestic and foreign computing locations could be involved in these use cases. Raw data are collected from the detectors throughout the year and stored permanently. Derived data will be reproduced each year. Simulation and signal-processing outputs will be produced roughly once per year, while 3D reconstruction outputs will be produced roughly twice per year (using the previous stages as inputs). [Section 5.9]
- Fermilab will be the largest provider of computation and storage resources for DUNE, providing 25% and 50% of the total computation and storage resources, respectively, required. The remaining resources will come from any number of distributed OSG and WLCG participants (domestic and foreign), placing a heavy emphasis on networking to the overall success of the workflow. Data volumes could reach or exceed 30 PB per year. [Section 5.9]
- SBN and DUNE simulation workloads can be performed using 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. [Section 5.9]
- DUNE reconstruction and analysis will be constant during operation; the primary copy of the raw data will be stored at Fermilab. Secondary storage and reconstruction and simulation CPU will be spread among the international partners; the specific locations are still to be determined. OSG tools facilitate a streaming data model, which would imply that reconstruction across wide-area networks may increase during operational ramp-up (2028 and beyond). [Section 5.9]
- The WLCG serves as the LHC's distributed computing facility. The mission of the WLCG is to provide global computing resources to store, distribute, and analyze the ~50–70 PB of data expected every year of operations from the LHC. WLCG management and operation requires participation by the major LHC experiments, major computing centers, software development efforts, and network providers. [Section 5.10.7]
- The ATLAS grid infrastructure consists of the Tier 0 computing facility at CERN, 11 Tier 1s, 70 Tier 2s, and about 30 Tier 3 participants distributed worldwide. All workflows are executed at all tiers. Tape storage for raw and AOD files is available at the Tier 0 and Tier 1 locations. [Section 5.10.5]

- The US ATLAS Tier-1 is hosted at BNL's Scientific Data and Computing Center (SDDC). The ATLAS connection to ESnet is shared with other programs hosted at the SDDC. The US Tier 1 is the largest participant of the ATLAS experiment, representing about 25% of the Tier 1 computing resources of ATLAS.
- ATLAS Tier 2 centers consist of multiple university-based clusters. There are four collaborative Tier 2 centers in the United States: NorthEast Tier 2 (Boston University and Harvard University), Great Lakes Tier 2 (University of Michigan and Michigan State University), MidWest Tier 2 (University of Chicago, Indiana University, and the University of Illinois at Urbana-Champaign), and SouthWest Tier 2 (University of Texas at Arlington and Oklahoma University).

- CMS is divided into several tiers. CERN is considered the 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 (at the Tier 1 only), simulated data generation, analysis data storage, and physics analysis activities (primarily at the Tier 2s). [Section 5.10.6]
  - The United States operates one Tier 1 facility (Fermilab), which is responsible for 40% of CMS Tier 1 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 available for use.
  - The United States has seven Tier 2 facilities: the University of Florida; the University of California, San Diego (UCSD); California Institute of Technology (Caltech); the Massachusetts Institute of Technology; the University of Wisconsin; the University of Nebraska; and Purdue University. Data move from these facilities to other universities as analysis data sets are reduced and refined during the analysis process. The US CMS Tier 2 facilities each contribute approximately 3 PB (or more) of active discussion storage.
  - Tier 3 facilities are loosely organized (and nonfunded) resources that perform user-level analysis using donated local computation and storage. 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.
- ATLAS and CMS have both experimented with simulation workflows at DOE and NSF HPC facilities including the DOE's ALCF, NERSC, and the NSF-funded Texas Advanced Computing Center (TACC) with plans to also utilize the DOE's OLCF. In this use case, it is possible to create simulation files and transfer the results back to any major experimental facility. The workflow is expected to continue to be used on the current generation of machines into Run 3 (2022 through 2024). Additional work to experiment with machine learning (ML) on CPUs and graphics processing units (GPUs) is a part of R&D efforts. In the future, more workflows can be converted to use HPC resources, putting much higher demands on networking. [Section 5.10.5, Section 5.10.6]
- ATLAS and CMS can use commercial cloud resources interchangeably with grid-based WLCG resources, but large-scale adoption is not expected. A notable exception in the use of commercial clouds would be the case of needed resource bursts for either CPU or network resources. Both experiments have evaluated the use of commercial cloud for processing and for TA transfer. In general, it was found that both are not cost-effective, at present. The experiments have shown that they have the ability to make large-scale use of cloud resources if the cost structure were to change. However, if cloud services are used, routing of traffic to and from those services must be done carefully because varying ingress and egress points can have significant cost implications. ESnet maintains peering points to major cloud providers and can

assist the LHC community to find solutions that are required to support the network workflow. [Section 5.10.5, Section 5.10.6]

- Budgetary gaps show a mismatch between the computing and storage that are affordable versus what is required to meet science goals within the LHC community. R&D efforts to better use computational, storage, and networking resources are critical. Network baselines are currently being planned to be terabit-scale (1–2 Tbps) backbone networks, with the largest resource site users connected at 100 G scale (200–800 Gbps). Network use will be at least a factor of 10 larger than Run 2 to overcome the gap in available computational and storage resources. [Section 5.10.7]
- HL-LHC R&D is being performed to understand the flexibility of the tiered computing model. It is not anticipated there will be a fixed hierarchy of computing for data processing use cases, although to improve usage efficiency, participants will be primarily categorized by size, service level, and capability to better utilize computing resources. [Section 5.10.8]
- During the HL-LHC era, the ATLAS experiment expects distribution to seven large facilities in the United States, which will all be required to have a full range of distributed computing capabilities. They will store both primary and secondary data, will provide access to hundreds of users, and will participate in continuous data transfers. These will consist of the current US ATLAS Tier 1 (BNL), Tier 2s (Great Lakes, Midwest, Northeast, and Southwest), as well as SLAC, and a few HPCs. [Section 5.10.8]
- During the HL-LHC era, the expanded use of HPC facilities will have an impact on HPC center storage and networking resources. These HPC centers are increasing in computing power, and several exascale 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 ATLAS data centers for subsequent processing. [Section 5.10.8]
- The concept of a dedicated analysis facility (AF), a specific location with enhanced capabilities, may emerge in the HL-LHC era as certain Tier 2s are able to keep or exceed the required storage, computational, and networking requirements. Intelligent networks and software components (e.g., workflow tools) can leverage these destinations to better allocate available resources. [Section 5.10.8]
- Experiments, such as ATLAS and CMS, that have traditionally utilized a grid paradigm (due to their highly distributed collaboration space) have experimented with and found success for certain aspects of their workflows by using HPC resources. Typically, these are limited to small input/big output tasks, such as creation of simulation data. To better support this use case, HPC centers will: [Section 5.10.5, Section 5.10.6]
  - Experience increased network usage, and will require more capacity to ESnet and to the connected data-ingest systems.
  - Require consideration of integrating common transfer tools used by the grid community to facilitate data ingest. This could be in the form of bulk-data movement or facilitating streaming use cases.
  - Investigate community-wide approaches to develop/adopt common Application Program Interfaces (APIs) that would facilitate software portability.
- Commercial cloud use is growing but is not wide across the HEP program area. Costs remain a major barrier, along with defining use cases that are better met versus other available resources provided by HPC and HTC facilities. Ensuring network performance to the commercial cloud is a secondary concern, particularly if the path taken is not able to leverage high-speed network

resources (such as ESnet). Adopting a cloud approach also comes with significant technical debt and risk that some projects and experiments are not willing to take on, namely creating a workflow and environment that can be ported, versus something that is finely tuned to use existing DOE resources such as an HPC facility or distributed grid environments. With wide availability of HPC and HTC resources in the DOE SC environment, commercial cloud adoption may remain low unless the previous factors are mitigated. [Section 5.10.5, Section 5.10.6]

- A fundamental choice for emerging experiments is the adoption of locality regarding computational and storage resources. For many collaborations, this means building their infrastructure (data management, computation, collaboration space) to use a centralized set of resources typically at a few core locations (e.g., ASCR HPC facilities), or a fully distributed grid-based model that can be built over time by contributing computation and storage resources, and linking them via high-speed network connections. In either case, it is paramount to ensure that a data sharing mechanism be thought out, tested, and verified regularly to ensure proper operation. [Section 6]
  - Collaborations that adopt a use case of keeping the computation and data centralized must:
    - Accept they need to establish a workflow from the instrument or facility (potentially remote) that involves knowing the full network path, ensuring high performance, and building appropriate toolchains to ensure a steady and predictable data flow to the allocated storage and computational resources.
    - Build a set of tools that is potentially portable but often tailored to the operating environment, for end-user analysis in the most efficient way possible.
  - For those collaborations that adopt a use case of allocating computational and storage resources anywhere, data must be migrated. This implies:
    - Network paths between collaborators are not always known, and often may change over time, which could affect productivity.
    - Shared resources imply that fate is shared more evenly between participants.
    - Tooling is more general purpose, since it must be developed and supported to run in a variety of environments.

2.7 Software Infrastructure

- Policy-level differences between operational facilities have an impact on tool adoption within DESI. As KPNO and NOIRLab, which KPNO operates, are NSF-funded facilities, their policies differ from those of NERSC and DESI, which are DOE-funded experiments. This results in the use of a software toolchain that is not unified. As data volumes increase, this lack of unification can further exacerbate negative impacts of a nonunified software environment, making shared analytics and research needlessly difficult. [Section 5.3]
- LZ user-level analysis will use computational and storage allocations at NERSC along with a common toolset. All software tools will be deployed via containers, which allow for portability to supported systems at NERSC (a decision that was made to ensure operation during maintenance windows due to the continuous nature of the experiment). Off-site data access is permitted, but not expected to be significant. [Section 5.6]
- Both g-2 and Mu2e use a similar set of software to manage workflow, analysis, and transfer (provided by the OSG). Data movement is typically handled as streaming, and coordinated through tools like XROOTD and Rucio. [Section 5.7]

- US CMS is in the process of retiring the use of GridFTP, and replacing it with TPC (Third-Party Copy) HTTPS, implemented via XROOTD servers. Participants typically have multiple such servers that each provide 10 Gbps, and all have access to the same filesystem. Large bandwidth transfers are thus accomplished by orchestrating very many flows across many servers. [Section 5.10.6]
- ATLAS and CMS share common software components for data mobility and management. At the lowest layers, the FTS tool is used to manage scheduling and file transfer as well as the XROOT protocol for data streaming use cases. ATLAS uses the PanDA workflow tool, and both experiments will use Rucio to assist with distributed data management at the start of the next run in 2022. The use of other tools (e.g., CMS's use of PhEDEx, Dynamo, and GridFTP) may be reduced after migration. [Section 5.10.5, Section 5.10.6]
- The data collected from perfSONAR at participating LHC AFS, as well as additional network-related data, are being gathered by OSG/WLCG and sent to an analytics platform at the University of Chicago. The data are stored in Elasticsearch and are publicly accessible via Kibana dashboards. [Section 5.10.7]
- Conversion of software from one major computing paradigm (HTC/grid versus HPC) to utilize another is problematic. Not only is rewriting software for a different use case time consuming, it is often done as a last resort and not funded. In many cases, software is not included in projects' budgets; thus they choose whatever is available, whether or not that is ideal for their needs. In rare cases, new software may be created to fill gaps. In the latter case, creation of hard to support/non-battle tested tools results, which can make the overall success of the research suffer as a result. [Section 6]
- Software development for scientific use cases remains a challenge. There are two main approaches: using software that is developed/supported by others for the same or similar use cases or attempting to write one's own (either funded by a project or unfunded). The former is encouraged; the latter is not recommended, but sometimes must occur. When applicable, leverage existing software, or ensure that future projects/experiments are budgeting for a software development lifecycle. Centralizing software development and support to knowledgeable teams, who are better equipped to implement best practices and ensure long-term usability, versus having experimentation fund dedicated resources, may be a more sustainable approach for future projects. [Section 6]

3 Review Action Items

ESnet recorded a set of action items from the HEP-ESnet Requirements Review that extend ESnet's ongoing support of HEP-funded collaborations. Based on the key findings, the review identified several actions for HEP, ASCR, ESnet, and ASCR HPC facilities to jointly pursue. These actions are also organized by topic area for simplicity and follow common themes:

- **Domestic networking for local and wide-area uses cases:** predominantly involves issues related to provisioning of domestic network resources (local to the experiment or distributed around the country) to support the science.
- **International and transoceanic networking:** predominantly involves issues related to provisioning of international network resources, often crossing oceans and involving multiple collaborators, to support the science.
- **Scientific data management: storage, dissemination, and volume:** topics related to the handling and management of scientific data. This includes but is not limited to how and where data are stored, how data can be shared in structured and unstructured ways, and the increases in volume in the coming years.
- **Data mobility:** observations and challenges involving the transmission of scientific data, and how these overlap with issues related to networking as a service and data management as an activity.
- **Computational resources:** computational topics (e.g., HPC, HTC/grids, cloud computing).
- **Software infrastructure:** topics related to software infrastructure of scientific experiments.

3.1 Domestic Networking for Local and Wide-Area Uses Cases

- ESnet will work with SLAC and NERSC to understand and contribute to the success of the WAN requirements for the Rubin Observatory and DESC experiment as they approach operational state in 2024. This includes but is not limited to providing dedicated connectivity between the ESnet POP (Point of Presence) in Atlanta and SLAC and peering arrangements with domestic sources (e.g., GCP) and international networks (AMPATH, GEANT). [Section 5.2, Section 5.4]
- ESnet will work with DESI and the Sun Corridor Network to understand the capabilities and constraints on networking at KPNO in Arizona. Network capacity is not viewed as an immediate concern, as the raw data sizes are not expected to grow on the constrained pathway between the instrument and NERSC over the course of the project. Network availability is viewed as a concern given the last mile challenges between KPNO and NOIRLab (single shared 1 Gbps connection). [Section 5.3]
- ESnet will continue to ascertain the networking needs between SURF and the DOE facilities that support and house research data of operating experiments (e.g., Fermilab, NERSC). Currently, SURF is served by a regional R&E network (Research Education and Economic Development Network [REED]) with a 10 Gbps connection and a 1 Gbps commercial-provider backup link available. Given the strategic importance of SURF for several DOE projects, understanding options that will lead to increased capacity, redundancy, and high-performance operation is recommended.
  - LZ maintains storage at SURF, which enables approximately 90 days of storage when a network disruption prevents transit to NERSC. [Section 5.6]
  - Due to the scale and relatively remote locality of DUNE, the networking technology that links the facilities (Fermilab and SURF) is critical for experimental success. By late in the

2020s, DUNE anticipates needs of 100 Gbps from SURF to Fermilab, and redistribution to other participating computing grid partners worldwide. [Section 5.9]

- ESnet will coordinate with Fermilab, BNL, and the LHCONE effort to understand and plan for the growing needs of LHC network capacities in preparation for Run 4 (2027 and beyond). Increasing network capacity and performance is anticipated at US-based Tier 1, Tier 2, and Tier 3 AFs (the first relies on DOE funding, and the latter two depend on external funding from the participants, experimental members such as universities, and agencies such as the NSF). [Section 5.10]
  - ESnet will coordinate with BNL, as BNL expects to scale beyond 100 Gbps in future years to facilitate LHC, Belle II, and other science use cases based at the lab. This should coincide with activities related to Run 4 (2027 and beyond), which will begin as Run 3 (2022–2024) is operating.
  - ESnet will coordinate with Fermilab to understand the growth trajectory of the network perimeter infrastructure in the coming years. It is expected that 400 Gbps technology could be deployed by the lab by the FY21–FY22 time frame. Additional WAN capacity from ESnet, either in the form of additional 100 GE WAN links or a 400 GE WAN link, should be investigated during Run 3 (2022–2024).
- ESnet will coordinate with the LHC experiments to fully understand and prepare for the networking needs of the emerging HL-LHC era (planned start in 2027). Observations from 2018 show utilization peaks of 16 Gbps between CERN and Tier 1s (raw data, as well as other use cases). The projected increases in event size (up to 8x) and event rate (up to 10x) for Run 4 will require matching network speeds. [Section 5.10.8]
  - It is anticipated that network capacity will be provisioned to match the scale of the available resources for each site user, which will exceed 1 Tbps for Tier 1s and hundreds of Gbps for Tier 2s. For Tier 1s, this will be supported by the DOE. Tier 2s have external funding sources and rely on upgrades to national and regional networks, as well as the AFs.
  - Network backbones will need to support multiple Tbps capacities.
- ESnet will partner with LHC R&D efforts to better understand the role of caching in the experimental workflow. 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 (2027 and beyond), a complete copy of the most succinct data formats will exceed 100 PB. By caching some percentage, a significant amount can be spared from local storage requirements, but in turn will be pushed into the network. [Section 5.10.8]

3.2 International and Transoceanic Networking

- ESnet will work with the Rubin Observatory, SLAC, and IN2P3 to understand and prepare for the WAN requirements of the Rubin Observatory and affiliated experiments (e.g., DESC) in the present term and operational period starting in 2024. The path from the Rubin Observatory to SLAC is long and complex, with multiple network providers. The international collaboration team has arranged relationships with carriers along the path (Chile, Brazil, the United States, and France) to guarantee operational stability. Performance monitoring and engineering coordination between organizations will be critical to ensure that the path is performant and stable. [Section 5.2, Section 5.4]
- ESnet will work with CMB-S4 to utilize and understand available international networking resources in extremely remote locations. In particular, South Pole networking is primarily accomplished via limited bandwidth and long-latency satellite connectivity. Current

expectations for performance are ~ 60 Mbps via these means, where instrumentation may produce requirements closer to 750 MB/sec. Approaches to minimize the immediate use of networks will reduce the requirement but mean a regular cadence of sending information via physical mechanisms will be needed. [Section 5.5]

- ESnet will continue to participate in, and advance the capabilities of, the LHCONE overlay network. [Section 5.10]
  - This collaboration supports a number of scientific projects, including ATLAS, CMS, and Belle II. The access to TA and transpacific networking connectivity is critical to experimental success, particularly as the resources available come from diverse funding sources.
  - LHCONE participation requires support for the technical considerations as well as any policy-based challenges. LHCONE is discussing ways to better understand and create AUP language that fits DOE science use cases. This effort will include mechanisms to enable multipurpose laboratory environments (e.g., LHC Tier 1 centers such as BNL and Fermilab) and computational facilities, such as the OLCF, ALCF, and NERSC, or commercial cloud. Solutions could leverage emerging R&D in WAN network orchestration (SENSE). LHCONE is currently lacking a unified set of 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.
- ESnet will continue to work in collaboration with Fermilab during operation of the ProtoDUNE experiment, which requires TA networking capabilities due to the location at CERN. Data rates of between 2–3 GB/sec are possible during runs. A total of 2–10 PB of data will be generated during the 2021–2022 operational phase. [Section 5.9]
- ESnet will coordinate with LHC R&D efforts that are working to support HL-LHC use cases. These include: [Section 5.10.8]
  - Understanding network capacity trajectory.
  - Adapting to growing use cases from other science domains (astronomy, biology, and engineering) that are emerging as global network users.
  - Researching and deploying new intelligent network services developed through prototypes.
  - Arriving at a system that provides transparency of use, management, and planning.
  - Connectivity between distributed storage and AFs remains critical, with capacity required to grow by orders of magnitude in the coming years. With overall experimental membership for ATLAS and CMS not expected to change significantly, the volume of data sent for computing and storage will increase to the existing members.
- ESnet will continue to engage with and understand the requirements surrounding the HEP program's usage of TA networking capability. [Section 5.10]
  - The LHC experiments are currently a major use case for the available resources. Raw data streams transferred between CERN and the US Tier 1 centers are expected to average more than 10 GB/sec during HL-LHC operations.
  - Current LHC software tools do not prioritize topological proximity when scheduling data transfers. As a result, streaming data across the TA links is allowed, yet inefficient. While efforts to reduce usage are ongoing, the projected growth rate will still require bandwidth augmentation. It is expected that the available capacity to support TA operations will need to be increased significantly near the later stages of Run 3 (2022–2024) and in preparation for Run 4 (2027 and beyond).

- ESnet will continue to provide the global R&E networking community relevant information on the best common practices for network architecture supporting efficient scientific data movement. This will better inform scientific facilities as they upgrade to advanced capacities in the coming years. [Section 6]

3.3 Scientific Data Management

- ESnet, ASCR, and HEP will explore ways to facilitate the long-term storage and sharing of research data sets (e.g., simulations, observations, derived data sets) required by a number of HEP experiments and facilities. Along with the general requirement of storage space that lasts longer than the run of any given project/experiment, the affiliated functions of cataloging, searching, and efficient download would be required. This need spans all focus areas of HEP, including Cosmic Frontier, Energy Frontier, and Intensity Frontier, and is complex to solve due to disjointed factors such as diverse funding sources, the need for collaboration space, the need for persistent and "off-project" resources, etc. [Section 5.1, Section 6]
- ESnet will work with DESC, NERSC, SLAC, and the Rubin Observatory to understand the available bindings to the data, software, hardware, and networks to ensure stable and reliable operational patterns in the future. DESC is heavily reliant on the Rubin Observatory as the source of scientific data for this project over the estimated 10-year run time (2024 through 2033). [Section 5.2, Section 5.4]
- ESnet will work with DESI and NERSC to understand and prepare for long-term solutions for data dissemination. Long-term curation of DESI data sets is expected to remain at NERSC beyond project lifetime. This implies that a solution that fits the entire cosmological landscape would be desirable to unify the tasks of storing, searching, and transferring data sets produced by a number of different projects over time. [Section 5.3]
- ESnet will collaborate with CMB-S4 on a series of data challenges in the coming years as the project moves from design to implementation. These will grow larger and test the ability of NERSC, ALCF, OSG, and XSEDE to handle the expected data volumes (network, storage, processing). The systems that will ultimately process the data sets may be years away and could take the form of both HPC- and HTC-enabled workflows. [Section 5.5]

3.4 Data Mobility

- ESnet will work with DESI to investigate other methods of data availability that better fit the use cases of the user community. The ability to easily replicate the entire catalog to other computational/storage resources is desired by the collaboration both for backup purposes and "fate sharing," in the event of extended downtimes at the primary computing facility. [Section 5.3]
- ESnet will work with DUNE to develop a comprehensive model of how many input/output operations per second (IOPS) are needed to handle incoming data between SURF and Fermilab, copy the data elsewhere, and serve distributed production and analysis users. This must take into account the locality of users to conserve network bandwidth on critical long-distance network paths. [Section 5.9]
- ESnet will collaborate with the LHC R&D efforts on ways to improve efficient data staging activities. To support the coming HL-LHC era, the experiments will move away from the any data, anywhere, anytime model, which allows for streaming data as needed. These R&D projects include: [Section 5.10.8]
  - Efforts to adopt intelligent data management systems (for all types of experimental data) that attempt to keep useful data in circulation based on availability, resources, and recent usage patterns.

  - Data lakes that will help to make clear regional distinctions by weighting network capabilities (capacity, latency, usage) differently, and thus reduce network usage through locality.
  - ESnet will collaborate with the US ATLAS/CMS operations programs, IRIS-HEP, and the WLCG DOMA working groups to organize data challenges to build up to the scale projected for HL-LHC. An example being explored is the ability to transfer and process 10 PB of data from an LHC facility to an HPC center, and back, within a day. It is very likely that the experiments will arrive at a detailed set of data challenges across multiple use cases within the next year or two. Such a program of work would then be executed and refined over the next five years. [Section 5.10]
- ASCR, HEP, ESnet, and ASCR HPC facilities will collaborate on ways to define and address advanced data-sharing capabilities that span experimental and facility boundaries. Analysis formats make this harder, as there is a desire to ensure that the unit of analysis contains enough information to be useful, but is compact enough to be shared. Once a format is created, there are the issues of the tools used, hardware required, and how it all interacts over the wide-area network. ESnet is in a unique position due to the close relationships it has with the facilities that share data and thus can encourage the use of intelligent tools and systems to simplify data sharing. Dedicated AFs are part of this solution, and are already used in some experiments and being investigated by others. This would create well connected and supported facilities with the only job of ingesting and egressing large amounts of data directly to ESnet and its connected resources and peers. [Section 6]

3.5 Computational Resources

- ASCR, HEP, ESnet, and ASCR HPC facilities will investigate and assist experiments that are "single sourced" to a specific computational facility to scale to other facilities and methods when resources become scarce or unavailable (due to scheduled or unscheduled maintenance, a natural disaster, or a systematic failure). This has potential impacts on productivity unless the following steps have been taken to build in redundancy: [Section 6]
  - Workflow portability, through the use of software containers that may facilitate deployment using other resources (within a facility, or at others within the ASCR HPC environment).
  - Cooperative computing and storage agreements (e.g., the ability to provide backup resources for a given allocation) between the ASCR HPC facilities.
  - Unified data transfer environments at ASCR HPC facilities, facilitated by ESnet.
  - Common APIs and middleware that would allow easier migration of software between ASCR HPC facilities.
  - Commercial cloud resources are a consideration for extreme downtime scenarios, but this mode of operation would not be viewed as a primary or secondary.
- ESnet will work with LHC R&D efforts to understand and support scientific use cases surrounding the use of commercial clouds. Some of the data and workflow management tools (e.g., PanDA, Rucio, etc.) are being adapted to transparently use commercial cloud resources alongside the traditional WLCG. This use case would open the possibility for more data exchange between R&E networks and commercial peering locations that support clouds, and could potentially increase to the volume of a typical Tier 2 facility. Research will investigate if this mode can be supported prior to and during Run 3 (2022–2024). [Section 5.10.8]
- ESnet will work with the HEP program and the LHC community to understand the volumes of data expected for use cases that involve streaming data. Reliable and high-capacity streaming of

input data, either raw or pileup simulation, would considerably reduce the disk requirements at HPC as well as non-dedicated computing facilities. Network capacity to these locations must keep pace with that of LHC-affiliated facilities. [Section 5.10]

- ESnet will work with ATLAS and CMS to address some of the challenges in connecting to large-scale resources that are dynamically accessed, e.g., clouds and HPC centers. For the LHC collaborator, excellent networking and connectivity to LHCONE is typical. But for "outside" computational resources that are opportunistically accessed, the networking may be challenging, since it is nondeterministic (e.g., not on LHCONE and controlled by external entities). [Section 5.10]
- ASCR, HEP, ESnet, and the ASCR HPC facilities will continue to research the implications of expanded use to support LHC-specific usage patterns prior to the start of HL-LHC. HPC centers are rapidly increasing in computing power: several exascale machines will be operational at the start of the HL-LHC era. These machines will be capable of producing larger volumes of simulated data beyond current capabilities used by the LHC community (e.g., distributed grids and current HPC facilities). Adaptation to other workflows that require more input data, such as those affiliated with experimental analysis, traditionally done using streaming tools, will require discussions on hardware and software architecture. [Section 5.10]
- ESnet will work with LHC R&D efforts to explore the efficient use of artificial intelligence/machine learning (AI/ML) in R&D efforts. The long-term adoption and impacts of these emerging research areas are not currently known, but they have shown initial promise in the areas of simulation creation and data analysis. Use of ESnet-curated data (e.g., network performance and telemetry) in AI/ML approaches will provide useful input to a number of intelligent network management prospects. [Section 5.10]
- ESnet, ASCR, and HEP will continue to explore the emerging use of commercial cloud resources. Science use is still limited, but growing, as some experiments and projects have conducted testing to evaluate usability. Costs remain high, which is the largest barrier to adoption. ESnet will continue to ensure that efficient access to clouds remains available as the use cases grow in the coming years. [Section 6]
- ASCR, HEP, ESnet, and ASCR HPC facilities will explore the concept of special purpose facilities (e.g., AFs, simulation facilities) that could create specific locations with enhanced capabilities to centralize certain aspects of highly distributed experiments that are typical of the HEP collaborations. In doing so, limited resources (e.g., computation, storage, networking) could be centralized and more efficiently managed. [Section 5.10]

3.6 Software Infrastructure

- ESnet will work with DESI to better understand and suggest mechanisms to enable "user facing" data search and import. Managing large data catalogs via a more modern graphical interface that would facilitate automatic synchronization (e.g., using fast transfer tools) between facilities would better fit workflows for user-level analysis functions. [Section 5.3]
- ESnet will collaborate with LHC R&D efforts to improve the capabilities that monitor and manage data transfers automatically. Given the size, complexity, and fully distributed nature of LHC computing, all workflow and data distribution must be optimized and managed with advanced technology. ATLAS and CMS will participate in R&D projects such as SENSE, Automated GOLE (AutoGOLE), and FABRIC to understand how to transition to managed network usage and evaluate R&D success in production operations. Collaborations with ESnet via the WLCG Network Throughput Working Group and the Global Network Advancement Group are critical. [Section 5.10.7, Section 5.10.8]

- In preparation for Run 3 (2022–2024), US ATLAS and US CMS would like to engage with ESnet and other partners on transitioning some of the SENSE functionality from R&D to production. The experiments would like to identify both appropriate links and appropriate production-ready functionality in SENSE, and integrate that into LHC tools for production use. [Section 5.10]
- ESnet will collaborate with LHC R&D efforts to research ways to acquire sufficiently fine-grained network monitoring and performance information to help debug and automate data transfers in real time. R&D into transfer accounting at the software layer will help to understand network usage patterns at a highly granular level. The extent to which this should include traffic tagging and/or flow tagging is unclear, but will be a part of the longer development efforts with networking partners. [Section 5.10]
- ESnet will collaborate with ATLAS and CMS to better account for the bulk of the usage of networking resources. This will be done using a mixture of approaches that can instrument software components and reason about network use (local and wide-area) along with measurement and monitoring performed on the network itself. Sharing network bandwidth with other science projects is expected: network management will be a core concern and area of research. Joint ATLAS and CMS use of Rucio for distributed data management prior to HL-LHC will be an appropriate mechanism, communicating near-term data movement intents and perhaps negotiating for any required quality of service (QoS) or deadline requirements. [Section 5.10]
- ESnet will collaborate with LHC R&D efforts to understand how the many deployed monitoring systems may tie together. With a more holistic view, the usage of networks will be better understood, contributing to improved data movement decision making. This includes efforts to understand traffic patterns from the applications directly, in addition to network-level reporting. Monitoring is being performed via "Packet Marking": identifying network traffic by owner and purpose, enabling identification and accounting of traffic anywhere along the network path. These efforts will require collaboration between the scientific community and network operators to manage complexity surrounding the solution space. In particular, establishing a set of rules and behaviors that can be used to guide how the traffic is handled on both traditionally operated networks, as well as those that feature advanced network infrastructures, will be necessary. Given the operational impact of implementation, care must be taken to ensure that risks and mitigations are well understood. [Section 5.10]
- ESnet will work with the SAND project, the OSG Networking team (part of IRIS-HEP), and the WLCG Network Throughput Working Group to develop and maintain an archive of measurement data. The collection and availability of network measurement data are of increasing importance as next-generation infrastructure will be able to automatically manage the network as a constrained and controllable resource. [Section 5.10]

4 Requirements Review Structure

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

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

4.1 Background

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

- Existing and planned data-intensive science experiments and/or user facilities, including the geographical locations of experimental site(s), computing resource(s), data storage, and research collaborator(s).
- For each experiment/facility project, a description of the "process of science," including the goals of the project 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 that are 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.

4.2 Case Study Methodology

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

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

The case study format seeks a network-centric narrative describing the science, instruments, and facilities currently used or anticipated for future programs; the network services needed; and how the network will be used over three timescales: the near term (immediately and up to two years in the future); the medium term (two to five years in the future); and the long term (greater than five years in the future).

The case study template has the following sections:

Science Background: 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 data sets 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 data sets produced by the instrument or facility (e.g., file size, number of files, number of directories, total data set 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 used in the process of science and how this work affects or may affect the network. This could include any connections to or between instruments, facilities, people, or data at different sites.

Software Infrastructure: 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: what is the network architecture and bandwidth for the facility and/or laboratory and/or campus? The section includes detailed descriptions of the various network layers (LAN, Metro Area Network [MAN], and WAN) capabilities that connect the science experiment/facility/data source to external resources and collaborators.

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.

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

5 High-Energy Physics Case Studies

The case studies presented in this document are a written record of the current state of scientific process, and technology integration, for a subset of the projects, facilities, and PIs funded by the Office of HEP of the DOE SC. These case studies were discussed virtually between June 2020 and November 2020.

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

- Cosmological Simulation Research
- DESC
- DESI
- The Rubin Observatory and the LSST
- CMB-S4
- LZ Dark Matter Experiment
- Muon Experimentation at Fermilab
  - Muon G minus two (g-2)
  - Muon-to-electron-conversion experiment (Mu2e)
- Belle II Experiment
- Neutrino Experiments at Fermilab
  - SBN
  - DUNE
- LHC Experimentation and Operation
  - ATLAS Experiment
  - CMS Experiment
  - LHC Operations
  - HL Era of the LHC

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

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

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

5.1 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 them, 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.

5.1.1 Discussion Summary

The following discussion points were extracted from the case study and virtual meetings with the case study authors. These are presented as a summary of the entire case study, but do not represent the entire spectrum of challenges, opportunities, or solutions.

- Cosmological simulations are typically created on HPC resources at large HPC-focused facilities (e.g., leadership computing facilities [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., the 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 users' 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 that are 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.

5.1.2 Cosmological Simulation Research Case Study

5.1.2.1 Background

Large-scale simulations provide detailed theoretical predictions to understand dark matter and dark energy, cosmological constraints on neutrino physics, and the nature of primordial fluctuations. Simulation results are essential for analyzing and interpreting results from cosmological surveys across multiple wavelengths, as well as aiding in survey design and optimization, and in the estimation and control of systematic errors. As modern surveys increase their focus on the complex relationships between galaxies and dark matter, the level of resolution and volume required of the simulations will increase, driving data volumes significantly upwards.

For the purpose of this case study, supercomputers function as data-generating instruments running flagship-scale simulations. Indeed, cosmological simulations can easily generate far more data than optical and microwave observations. Halo/galaxy catalogs from simulations are typically far larger than their on-sky counterparts: estimation of statistical errors relies on hundreds to tens of thousands of simulated examples, and cosmologists often need to simulate results obtained from dozens to hundreds of varying simulation input conditions. While the on-sky pixel-level data can in some cases be larger, the simulated catalogs are already at the level of PBs for large simulation suites. Many simulations are also needed whenever one wants to match the simulations to the actual data, by reconstruction of initial conditions. In this inverse problem application, one iterates on the initial guess, and after convergence one wants to sample from the posterior, where each operation is a separate simulation.

Typically, the dataflow consists of two stages: (1) within the facility where the simulation is run (supercomputer, storage, analysis system) and (2) from the host facility to a remote analysis/archive site. The data motion may be staged in a scheduled manner or can be highly bursty, depending on the use case. In the future, control of the dataflow may be centralized or distributed to a few "trunk" sites. As simulations become more integral to the work of individual researchers, there is a potential to see the concept of distributed analysis sites that will better adjust to the workflow demands. Further, if the data sets are too large to store locally by individual research groups, then it may be necessary to stream the data for on-the-fly analysis, which can then lead to repeated downloads and analyses coordinated with the data flow.

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 them, 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.

Data sets are typically owned by the team that generated them, unless the work was specifically part of a collaboration and/or commissioned by the collaboration. But the increasing sentiment toward open science as well as a sense that such simulations are useful for a wide range of applications has led to more interest in placing simulations in the public domain. Funding agencies can put requirements on the data sets corresponding to making subsets publicly available. The primary barrier to making simulation data sets public is the lack of available storage and/or server-side public-access compute capability.

5.1.2.2 Collaborators

All the HEP Labs (ANL, BNL, Fermilab, LBNL, and SLAC) participate in this work, as well as other labs (Los Alamos National Laboratory, Lawrence Livermore National Laboratory) and a large number of collaborating universities, including the University of California, Berkeley, the University of Chicago, Harvard University, the University of Illinois Urbana-Champaign, the University of Pennsylvania, Stanford University, the University of Washington, Yale University, and many others. Many of the collaborations involve directly working with the

associated experiments (Dark Energy Survey[1], the DESI[2], the Extended Baryon Oscillation Spectroscopic Survey[3], the South Pole Telescope[4]). Computational facilities used include ALCF, NERSC, and OLCF as well as NSF computing facilities, along with institutional resources available at the participating institutions.

The data products generated by simulation groups are used downstream by a large number of users who are members of science working groups in cosmological surveys and experiments. A common use pattern is to build a mock survey from the simulation data, then perform a clustering analysis on that survey. In the future, the ratio of the user community to the size of the simulation groups will only increase as the role of simulations continues to become more important. Thus, the dataflow pattern will likely become more diverse and there will be increased pressure on data transfer facilities.

User/collaborator and location	Primary or secondary copy of the data	Data access method	Avg. size of data set	Frequency of data transfer or download	Are data sent back to the source and method	Any known issues with data sharing
ANL	Primary/secondary	Data transfer via Globus	Variable, TB to PB	Intermittent, week/month	Rarely, via Globus	Lack of stable long-term storage
HARVARD UNIVERSITY	Primary/secondary	Data transfer via Globus	Variable, TB to PB	Intermittent, week/month	Rarely, via Globus	Lack of stable long-term storage
LBNL	Primary/secondary	Data transfer via Globus	Variable, TB to PB	Intermittent, week/month	Rarely, via Globus	Lack of stable long-term storage
OAK RIDGE NATIONAL LABORATORY	Primary	Data transfer via Globus	Variable, TB to PB	Intermittent, week/month	Rarely, via Globus	Lack of stable long-term storage

Table 1: Cosmology simulation data projections

There are too many potential endpoints to note individually in the previous table. The cosmology collaborations can have many hundreds of collaborators from a large number of institutions (~100) who all have the right (and need) for various simulation products. Each consumer endpoint also varies significantly in how much data throughput and volume it can support (from individual desktops to institutional clusters).

5.1.2.3 Instruments and Facilities

The work performed is primarily computational — the major facilities used are primarily those at the ALCF[5] (Theta[6]/Cooley[7]), NERSC[8] (Cori[9]), and OLCF[10] (Summit[11]). The use case consists of running large simulations on the primary systems, carrying out analyses on smaller satellite machines, and making the data available for downstream work. Smaller local resources are available at a number of participating institutions. Institutional storage ranges from ~100 TB to a few PB, significantly smaller than the total storage available at the HPC centers. Internal data transfer rates range from 100s of GB/s bandwidths to as low as 10 Gb/s internal institutional links.

[1] https://www.darkenergysurvey.org
[2] https://www.desi.lbl.gov
[3] https://www.sdss.org/surveys/eboss/
[4] https://pole.uchicago.edu
[5] https://www.alcf.anl.gov
[6] https://www.alcf.anl.gov/support-center/theta
[7] https://www.alcf.anl.gov/support-center/cooley
[8] https://www.nersc.gov
[9] https://www.nersc.gov/systems/cori/
[10] https://www.olcf.ornl.gov
[11] https://www.olcf.ornl.gov/summit/

We do not foresee a change in the modus operandi over the next five years (and even beyond, unless cloud facilities become competitive). Transfer rates are expected to improve as infrastructure upgrades are made.

Data storage needs post mid-2020s are likely to be in the ~300 PB range of spinning disk equivalent, requiring transfer rates of >300 Gb/s (intermittent). These numbers are consistent with those in the ASCR/HEP Exascale Requirements Review Report[12].

5.1.2.4 Process of Science

The simulation results can be viewed as corresponding to three levels. Level 1 is the raw data from the simulations, level 2 is intermediate-level analysis data, and level 3 is level 2 data reduced to the level of catalogs/databases (when possible). Level 1 analysis can be performed both in in situ or post-processing modes, while level 2 and level 3 analyses are primarily in post-processing mode. All three levels are part of the local and remote "process of science." Level 1 and 2 data analyses involve batch processing, while level 3 analyses can have a significant level of interactivity, so the data access patterns can be quite different.

In situ analysis uses the supercomputer's own network, or co-scheduled computational resources, whereas level 2 and level 3 analyses can be conducted locally or remotely and may involve moving data from file systems back to the host supercomputer or to analysis resources. This is where the networking support is most important. Level 2/3 data may also be moved over in batches to remote sites where they can be locally analyzed. Note that the actual mass of data at the three levels is roughly similar, except that the granularity increases significantly from level 1 to level 3.

Future activity will be along the lines mentioned previously, and this should not change over the next five years.

5.1.2.5 Remote Science Activities

As already stated, the major computational facilities used are primarily those at the ALCF, NERSC, and OLCF. Storage is provided at the facilities at the level of hundreds of TB per sub-project of disk (potentially larger, in some instances) and substantially more on tape. The exceptions have included special dispensations for the ASCR Leadership Computing Challenge (ALCC) and Innovative and Novel Computing Theory and Experiment Program (INCITE) projects. The major data sources are the supercomputers and other associated compute resources (analysis and visualization clusters). The use of cloud resources may also be considered in the near future, both commercial and institutional. WAN is typically via Internet2 and ESnet.

We expect this mode of operation to remain unchanged over the next five years.

The current data stream involves typically moving ~10–100 TB of bulk data, with larger transfer carried out on occasion. The much smaller "user" downloads to local storage typically involve <1 TB chunks of data. Progress in remote visualization methods should allow (almost) real-time visualization of many large data sets. Examples of this already exist.

As to the supercomputers themselves, there will be significant architectural changes, but these will not be disruptive with respect to the networks.

Over this time span, a few major simulation data archive/analysis centers will emerge (e.g., the community file systems at the LCFs and NERSC). Most of the data will eventually be hosted there; they should also have substantial local analysis computing available (since computing must follow the data). In this case, the data stream will consist of two major types:

- "Feeders" to the data centers, moving ~PB, and
- Data center to "user" links which would typically move <100 TB chunks of the data itself.

Remote visualization methods should allow real-time visualization of many large data sets.

[12] S. Habib, R. Roser, et al, arXiv:1603.09303 [physics.comp-ph]

We expect some changes to the simulation software, but they are unlikely to be very significant (at least partly due to inertia, because of the size of the current software base). Major changes should be expected beyond the early 2020s, however, in step with the architecture changes in next-generation supercomputers. The expectation is only evolutionary changes in the tool infrastructure.

5.1.2.6 Software Infrastructure

Software infrastructure varies widely. The use of code repositories is now widespread. Higher-level workflow tools to manage simulations and simulation analysis (including remote analysis) are slowly emerging, although most of this work is still performed by hand-written scripts. Software containers are seeing more use for complex software stacks. Data transfer is primarily through Globus[13], which has proven very effective both within institutions and externally as a graphical tool or when coupled with a portal[14]. A number of tools are used to process data sets, including embedded capabilities (e.g., data compression within I/O). The use of data containers (HDF5[15], PnetCDF[16]) is sometimes limited by their reliance on message passing interface (MPI) -IO; HDF5 is also considered to be too complex by some users. Native I/O tools written for simulation codes still obtain the best performance.

Over the next two to five years, there will be an evolution in workflow tools to manage local and remote simulation data analysis (Jupyter notebooks[17]). This area is still in flux, and community desires are only now being captured in a design process. Beyond five years, software management infrastructure is expected to be significantly improved on this timescale because the software complexity will be much higher than at present.

5.1.2.7 Network and Data Architecture

These factors are highly specific to individual programs. For purposes of this review, networking and data architecture features are not foreseen to be a major source of requirements versus future HPC center/university compute requirements.

5.1.2.8 Cloud Services

As far as cloud computing and storage are concerned, usage is primarily a question of latency, bandwidth, amount of associated computing, and cost for computing, data transfer, and storage. Primarily due to cost issues, the cosmological simulation community does not see the commercial option as being viable currently, but this situation could change very quickly. There are several attractive features of the cloud model, including virtualization and resource elasticity; there is an expectation that these will become ever more important with time and to drive increased use of cloud resources. Another advantage is the ability to make data available, with server-side compute, while placing the cost of using the data onto the end user, who may be supported by a different funding agency/source or country.

More recently, cloud hosting of supercomputing and specialized AI/ML-optimized hardware, such as Tensor Processing Units (TPUs), has added a new layer of capability that has potential benefits for cosmological simulations and the associated data analysis, especially since the AI/ML capabilities can be easily exploited with productivity languages such as Python. Cost concerns do remain an issue.

5.1.2.9 Data-Related Resource Constraints

This work continues to be substantially limited by the availability of storage, particularly when considering how to distribute the simulation data beyond the group that performed the simulation. Further, the need for storage beyond the initial computing project is critical; the lifetime of the need is often set by the life cycle of the analysis, which is happening in larger collaborations.

[13] https://www.globus.org
[14] https://mrdp.globus.org
[15] https://support.hdfgroup.org/HDF5/whatishdf5.html
[16] https://parallel-netcdf.github.io
[17] https://jupyter.org

The network speeds between major centers are not a limitation, and transferring PB-sized data sets is almost routine. This is expected to increase in the coming years. User-level transfer remains smaller, as downloading entire catalogs is still not a typical use case.

5.1.2.10 Outstanding Issues

The following factors are considered to be high priority for the long-term success of this community:

- A possible disruptive element is the role of AI/ML in low-level simulation and analysis code, as well as at the catalog analysis level. Since such techniques are typically very data intensive, it is conceivable that the associated network requirements might be significant, especially if compute resources are not local to where the data are available.
- Storage for simulations will continue to be a critical issue over time. This has several dimensions to consider:
  - Size of simulations will increase.
  - Locations producing simulations are expected to increase.
  - Funding spans different agencies.
  - Usefulness over time, and to different communities, is increasing.
- All of the factors that exacerbate the issue of long-term storage may not have a single unified solution, but require a solution. If a single location cannot become the home for data sets, there must be efforts to unify a cataloging system to enable locality searches, despite the distributed nature.
- Data movement between facilities is not a critical problem, but to the user community remains challenging. Simplifying the access method to use updated portal software, or tools like Globus, has improved data mobility.

5.1.2.11 Case Study Contributors

Cosmological Simulation Representation

- Salman Habib[18], ANL
- Daniel Eisenstein[19], Harvard University
- Uros Seljak[20], University of California, Berkeley

ESnet Site Coordinator Committee Representation

- Linda Winkler[21], ANL

5.2 DESC

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.

[18] habib@anl.gov
[19] deisenstein@cfa.harvard.edu
[20] useljak@berkeley.edu
[21] winkler@mcs.anl.gov

5.2.1 Discussion Summary

The following discussion points were extracted from the case study and virtual meetings with the case study authors. These are presented as a summary of the entire case study, but do not represent the entire spectrum of challenges, opportunities, or solutions.

- DESC is heavily reliant on the Rubin Observatory as the source of scientific data for this project. As a result, the scientific facility is completely 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 has not made final decisions regarding the full chain of software as a result of this binding to Rubin, and will closely observe the setup of the Rubin DFs in the coming years. Adoption of several key software packages will be chosen after Rubin has indicated their process. This may include the same, or a complimentary, approach to data management and mobility, as well as a computational analysis framework.
- DESC will fully adopt the "bring the user to the data" approach with its 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 method to accomplish these goals is still being defined, but will use existing NERSC tools when possible.
- 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. Use of these other large facilities (e.g., grid sites in the UK or other DOE LCF facilities like ALCF or OLCF) are not known at the time of writing.
- Collaboration size and scope for DESC will change as the project approaches start in 2023/2024.
- The next-generation NERSC resource (Perlmutter[22]) is not operational at the time of this review, and will influence some of the choices for analysis (software, workflow, etc.).
- The operational platform defined by Rubin is not fully known, as the USDF at SLAC was named in October 2020. The use of HPC or HTC computing resources, and the software used to power the platform, are thus still in the planning stages. DESC's use of HPC resources at NERSC has the potential to be different than Rubin. Thus there will be discussion on the ways to best integrate and reuse components between the experiments.
- In summary, DESC has several areas of uncertainty at the current time, but will know the ways forward as the project approaches 2023/2024 startup. These areas of uncertainty include:
  - Software tools used for most aspects (data movement, computation, etc.).
  - A more succinct notion of the data release schedule and data sizes.
  - The proper bindings to interface with the Rubin USDF at SLAC.
  - The role and bindings to the EU DF at CC-IN2P3.
  - A large-scale effort to define and participate in data challenges with Rubin.
  - Ways to publish the DESC data results, and long-term storage (at NERSC, or other locations).
  - The full impacts of COVID-19 delays on Rubin and DESC.
  - The role DESC will play in handling some scientific use cases that Rubin will publish (e.g., transients).

[22] https://www.nersc.gov/systems/perlmutter/

5.2.2 DESC Case Study

5.2.2.1 Background

DESC is the international science collaboration that will make high accuracy measurements of fundamental cosmological parameters using data gathered from the Rubin Observatory project's LSST. LSST will take some 800 nine-square-degree exposures of each location over 10 years covering about 18,000 square degrees of the southern sky. DESC is the international science collaboration that will make high accuracy measurements of fundamental cosmological parameters, and derive cosmological constraints, using data from the LSST.

For static science, user access to the LSST data will be via an annual data release prepared at Rubin's processing centers. Rubin anticipates the primary use of the data originating from object catalogs describing the stars and galaxies found in their initial data processing. Some limited use of the "raw" image data that is used to produce the catalog is expected. After 10 years, the image data are expected to occupy about 500 PB of space, while the object catalogs would occupy about 5 PB. Rubin will provide a complete science analysis platform and encourages users to access the data via dedicated resources at its data centers[23]. It is expected that some collaborations, such as DESC, may require bulk downloads of data to their own computing resources to do their analyses.

DESC is organized around five main probes of dark energy enabled by the LSST data:

1. Weak gravitational lensing (WL) — the deflection of light from distant sources due to the bending of space-time by baryonic and dark matter along the line of sight, which allows a measurement of the growth rate of cosmic structure (and therefore is also sensitive to dark energy).
2. Large-scale structure (LSS) — the large-scale power spectrum for the spatial distribution of matter as a function of redshift. This includes the Baryonic Acoustic Oscillations measurement of the distance-redshift relation.
3. Type Ia (read: "type one-A") Supernovae (SN) — luminosity distance as a function of redshift measured with Type Ia SN as standardizable candles.
4. Galaxy clusters (CL) — the spatial density, distribution, and masses of galaxy clusters as a function of redshift.
5. Strong gravitational lensing (SL) — the angular displacement, morphological distortion, and time delay for the multiple images of a source object due to a massive foreground object.

DESC has defined a Science Requirements Document (SRD)[24] and Science Roadmap[25] to lay out goals for Year 1 Survey performance, and the deliverables (analysis software, simulations, image processing) needed to produce the validated analysis pipelines required to do that science.

In addition to running analysis pipelines on the object catalogs, either via querying a database or reading in parquet files[26], DESC anticipates reprocessing some fraction of the raw image data in order to determine the systematic error budget. This will involve transferring data from the Rubin DFs to NERSC, and running much of the Rubin image processing pipeline on it, while modifying operating parameters to determine the effects on the analysis results. DESC data products will be the responsibility of the collaboration to curate, and generally are value added catalogs (e.g., galaxy shape correlations due to gravitational weak lensing) derived from the original Rubin Observatory catalogs.

[23] In October 2020, the Rubin USDF was identified as SLAC.
[24] https://arxiv.org/abs/1809.01669
[25] https://lsstdesc.org/assets/pdf/docs/DESC_SRM_latest.pdf
[26] https://parquet.apache.org

5.2.2.2 Collaborators

DESC's current analysis model involves bringing collaborators directly to the major processing centers to run at-scale analyses and access collaboration data products. Currently the two main centers are NERSC and CC-IN2P3 in France. Note that CC-IN2P3 is an operational partner of Rubin, so via that path, they maintain a full copy of the survey data that will not require re-staging.

The collaboration currently comprises some 1,000 members, of whom about 220 are full members. These originate in 23 countries and about 50 institutions. Of these, about 75% are from the United States, 17% from France, and 7% from the United Kingdom. About 400 of the members are grad students or postdocs. As Rubin approaches operation, DESC anticipates more countries joining the collaboration.

In the current model, DESC anticipate transfers of the data releases from the Rubin USDF at SLAC to NERSC. These would be fractions, perhaps 10%, of the image data and the full object catalogs as parquet files. Operations on these data would be largely confined to NERSC, but the possibility exists to retrieve the data for computation elsewhere.

DESC is currently developing infrastructure, including starting on data challenges. DESC is exploring using tools like Globus, HTTP, and RSYNC for data mobility. The Rubin Data Management team is investigating Rucio[27] as a data management tool. DESC will keep apprised of that situation, as reusing Rubin infrastructure will simplify operations.

User/collaborator and location	Primary or secondary copy of the data	Data access method	Avg. size of data set	Frequency of data transfer or download	Are data sent back to the source and method	Any known issues with data sharing
NERSC[28]	Partial secondary	Data portal and data transfer	50,000 GB images; 100,000 GB for catalogs	Ad hoc	N	Data challenges pending
THE FRENCH NATIONAL INSTITUTE OF NUCLEAR AND PARTICLE PHYSICS (IN2P3)	Primary	Data transfer	Rubin data on-site	N/A	N/A	Data challenges pending
GRIDPP[29] — SIMULATION ONLY	N	Data transfer	100s of GB to few TB	N/A (simulation trials)	Y, Globus/ RSYNC	Full data challenges pending
ALCF[30] — SIMULATION ONLY	N	Data transfer	100s of GB to few TB	N/A (simulation trials)	Y, Globus/ RSYNC	Full data challenges pending

Table 2: DESC data projections

5.2.2.3 Instruments and Facilities

LSST DESC will be responsible only for analyzing data collected by the Rubin Observatory. The anticipated start of the survey is FY24, barring COVID-19 related delays. DESC is not directly involved in the data-taking process itself, or in the development of the instrument. DESC also has no influence on the facility's computing capabilities. While DESC will be able to possibly carry out some analysis on the Science Platform provided by the facility, it is not fully clear yet how these resources will be used. It is anticipated that DESC will primarily download data via the platform or pre-arranged bulk-data movement tools. DESC will refer to the Rubin Observatory response for further details for this section.

[27] https://rucio.cern.ch
[28] https://www.nersc.gov
[29] https://www.gridpp.ac.uk
[30] https://www.alcf.anl.gov

5.2.2.4 Process of Science

DESC's timeline is driven by the Rubin construction and operations plan. Due to the delays brought on by COVID-19, the Rubin survey is projected to be delayed by a year, to October 2023. Until now, DESC had been preparing through the use of data challenges: DC2 is a five-year simulation of 300 square degrees.

As mitigation for the COVID-19 delay, Rubin is making plans to use the DESC simulated sky as a "Data Preview 0" to exercise its infrastructure, pipelines, and early user experiences. This is anticipated in 2021 and would involve a transfer of about 100 TB of simulated data to the Rubin DF.

Rubin is developing plans for commissioning, using both a small commissioning camera and then the full camera, followed by a short science verification period. The survey is planned to run 10 years. Rubin has not yet released its commissioning plan, nor what or how much data will be accessible by the science collaborations.

As detailed previously, during the survey, DESC anticipates transferring up to about 10% of a single year's data from the Rubin USDF to NERSC — about 5–6 PB for raw images and prepared object catalogs that are released annually.

Our analysis pipelines are described at a high level in the Science Roadmap[31]. A high-level overview follows:

1. The analyses process primarily starts from the annual object catalogs.
2. For weak lensing analyses, galaxy shapes are the main observable used, combined with redshift measurements inferred from multi-band photometry. There are three summary statistics currently envisaged which correlate galaxy shapes with each other; correlate galaxy shapes with positions; and correlate galaxy positions with each other. These are combined into a set of three 2-pt correlations. DESC also anticipates higher order statistics being needed to capture non-Gaussian information.
3. Analysis of supernovae is based on identifying SN1a from the transients' stream, followed by using the apparent brightness and redshift to determine a recession velocity.
4. Strong lensing is based on finding lensed quasars and supernova systems and then measuring the time delays for various paths of the lensed object using multi-year light curves.
5. Finally, clusters of galaxies are used to predict cosmological parameters based on the relationships between cluster properties and halo mass. These parameters are obtained via a likelihood analysis using mass function predictions obtained from simulations.

5.2.2.5 Remote Science Activities

Present to Two Years

Currently, LSST DESC is focusing on working with simulated data and publicly available data from precursor surveys. The simulations for DC2 have been carried out at three sites: NERSC, ALCF, and the UK/French grid. The processing of the resulting images has been performed at CC-IN2P3. All the data products generated in the process (input catalogs to the image simulations, raw image simulations, intermediate data products, and final catalogs) have been transferred back to NERSC. Therefore, remote instruments in this case are Theta at the ALCF, grid resources, and the computing resources at CC-IN2P3. The precursor data have been collected by the Dark Energy Survey (DES) and Hyper Suprime-Cam (HSC)[32]. However, the data that are being analyzed by DESC from these surveys are small and counting the telescopes from these surveys as remote instruments is not necessary.

[31] https://lsstdesc.org/assets/pdf/docs/DESC_SRM_latest.pdf
[32] https://hsc.mtk.nao.ac.jp/ssp/instrument/#hyper

Next Two to Five Years

During the next three years, DESC will continue to prepare for the data arrival and in particular continue to develop its analysis pipelines and workflows for carrying out targeted data processing tasks. In addition, more simulated data will be generated (e.g., supercomputing resources at ALCF and OLCF are used to generate new simulations and extragalactic catalogs currently). DESC will continue to use supercomputing resources for these tasks. For the later years in this time period, DESC will focus on reprocessing the data to understand possible systematics and suggest possible improvements in the analysis algorithms. In the past, DES injected simulated data into real data to study possible systematics. DESC very likely will carry out similar tasks with the data arriving in the first years. This again will require access to supercomputing resources. Given that these tasks are very focused and need only a few DESC members to carry them out, supercomputing resources beyond NERSC can be used easily for these tasks.

Beyond Five Years

It is very likely that DESC will use supercomputing and grid resources during the full 10-year timescale of survey operations. The tasks will continue to include the simulation of test data, partial reprocessing, and data quality assessments. Tasks that can be easily isolated to run on resources that do not need to be accessed by a large number of DESC members will continue to be farmed out. DESC might also carry out analyses in collaboration with satellite missions (e.g., Roman[33] and Euclid[34]) but the data transfers in these cases would be small, and the data would come directly from the connected computing centers holding their data using tools curated by these scientific efforts.

5.2.2.6 Software Infrastructure

In the current period of data challenges, with simulated data totaling a few hundred TB, transfer has largely been via HTTP, RSYNC, and Globus for data sets moving from NERSC to and from CC-IN2P3. DESC understands these to be placeholder tools while the Rubin team settles on tools for data transfer, since the primary paths will be from the Rubin USDF at SLAC and CC-IN2P3. DESC expects the tools to include both the transfer mechanism (e.g., Globus) and the cataloging (e.g., Rucio); DESC has yet to investigate catalog tools during this period.

Rubin will also be running at-scale complex workflows, but has not selected (or written) the tools needed to do this. For DC2, DESC could not wait and prototyped the use of Parsl for workflow management. DESC has good experience with other workflow tools and has been able to work closely with the Parsl team on desired features to run distributed processing with minimal human oversight. Rubin may not choose a tool with significant abilities for heterogeneous, distributed processing, so DESC may be on its own in this area, and continues to evaluate the Parsl tool with DC2 processing.

5.2.2.7 Network and Data Architecture

The Rubin USDF at SLAC was announced October 2020, but the full implications of how this will affect DESC is not fully known at this time. NERSC and SLAC are both connected by ESnet and within the same geographical region, thus site-to-site capacity is not anticipated to be an area of concern. DESC will work with NERSC, ESnet, and SLAC in the coming years on data challenges to fully understand capabilities and bottlenecks.

DESC is settled with the connection with CC-IN2P3, noting that it is also an operating partner of Rubin, processing half the data and storing a full copy. It is a T1 LHC site, and so has very capable networking capacity within Europe and to the United States. Near-term network requirements are for the transfer of DC2 image files to CC-IN2P3 from NERSC and ALCF and the return of processed images and object catalogs from CC-IN2P3 to NERSC. Observed transfer rates are about 1–1.2 GB/s for transfers of ~200 TB data.

[33] https://roman.gsfc.nasa.gov
[34] https://www.jpl.nasa.gov/missions/euclid

DESC's longer-term computing model and split of responsibilities across the three main resources is not set yet. With the current model where the bulk of the compute time would go to reprocessing images multiple times to determine systematic error budgets, DESC would not need to first transfer those files to CC-IN2P3 from NERSC, and only need to transfer back results.

Rubin's plan is to provide the community with annual data releases during the survey, so DESC would be doing its bulk transfer from the USDF on that cadence. DESC expects on the order of 5 PB of images per year, so 10% of that would lead to transfers of 500 TB for each release (it is not clear whether this 10% needs to be per year's data or integrated from the start of the survey). DESC currently estimates that downstream processing is small compared to the reprocessing load and would be run at NERSC (or CC-IN2P3) from object catalogs likely stored as parquet files. Assuming all iterations of the reprocessing are kept, DESC estimates about 1 PB per year of object data.

5.2.2.8 Cloud Services

There is currently no plan within DESC for the use of cloud services, given that both primary centers are large, dedicated facilities. Rubin has plans to set up an interim "cloud" facility to test software and operations, but DESC does not anticipate having to know the details of that. The interfaces should be the same whether the data are in the cloud or at the Rubin USDF at SLAC.

5.2.2.9 Data-Related Resource Constraints

DESC will depend heavily on next-generation supercomputers in the DOE complex. While the Rubin Data Management stack is being written to run on serial high-throughput machines (generally single-threaded and requiring a few GB memory per core), the current understanding is that the Perlmutter machine coming to NERSC would be able to accommodate that code base. CC-IN2P3, as an operating partner, is obliged to be compatible with Rubin Data Management system code. Perlmutter is expected to come online in 2021 and operate for five to six years.

Remaining uncertainties for DESC are stable annual allocation, up time, and queue scheduling. HPC queues are typically customized to large MPI jobs, in order to fill up the machine. DESC's workflow does not fit in this category and has suffered execution delays in the scheduling process.

5.2.2.10 Outstanding Issues

There is nothing to report for this section. Data trials to support operations with the newly named Rubin DF at SLAC will be performed at a later date, and ESnet will be engaged in that process.

5.2.2.11 Case Study Contributors

DESC Representation

- Katrin Heitmann[35], ANL
- Richard Dubois[36], Stanford University

ESnet Site Coordinator Committee Representation

- Linda Winkler[37], ANL
- Mark Foster[38], SLAC
- Damian Hazen[39], LBNL and NERSC
- Tavia Stone Gibbins[40], LBNL and NERSC

[35] heitmann@anl.gov
[36] dubois@stanford.edu
[37] winkler@mcs.anl.gov
[38] fosterm@slac.stanford.edu
[39] dhazen@lbl.gov
[40] tavia@lbl.gov

5.3 DESI

DESI is a scientific research instrument for conducting spectrographic astronomical surveys of distant galaxies. It will utilize the Mayall Telescope[41] (a four-meter telescope), located at KPNO[42] near Tucson Arizona.

The overall process of science is focused on creating a 3D 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 approximately 700 MB for an image, which are 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. 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. 10 Gbps exists today as provided by KPNO, although upgrades and redundancy are stretch goals. The experiment has the ability to buffer data when connectivity is lacking through the use of some local computation and storage and a workflow manager that is controlled at NERSC.

DESI expects a model similar to other astronomical experiments, where most (if not all) user analysis will be done at the location of the data (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 the instances where that is required, DESI will leverage existing NERSC infrastructure (DTNs 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 Discussion Summary

The following discussion points were extracted from the case study and virtual meetings with the case study authors. These are presented as a summary of the entire case study, but do not represent the entire spectrum of challenges, opportunities, or solutions.

- DESI is a cosmology collaboration with the goal of creating a 3D map of the universe over a five-year runtime (starting in 2021).
- 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 715MB of data. After observing, data must be sent to NERSC in semi-real time. Processing will result in a 10 GB data product per exposure. The data products are 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. Data transfer from KPNO to NERSC is done on an approximate 10-minute cadence initiated by NERSC. In some circumstances, NERSC may not be able to handle data ingest from the instruments due to network connectivity problems or lack of computational resources. In these situations, the data can be buffered at the source and will be synchronized during the next window of opportunity.

[41] https://www.desi.lbl.gov/telescope/
[42] https://www.noao.edu/kpno/

- DESI requires network connectivity between KPNO and NERSC to ensure stable operations. Limited buffering space is available in the event of 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) to a certain degree.
- There is a computational facility affiliated with KPNO. It is provided by the NSF National Optical Infrared Astronomy Research Laboratory (NOIRLab), and is operated by the Association of Universities for Research in Astronomy, Inc. (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.
- In the event that sending data to NERSC is not possible, there are some limited processing options on-site at KPNO that can be used for quality assurance (QA) procedures (e.g., helping to adjust telescope position during observational phase).
- User-level analysis will be conducted at NERSC through dedicated compute and storage allocations, along with yearly tasks to reprocess data sets 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 1Gbps for the entire shared facility where there are several other funded-astronomical projects from different agencies. The network is supplied by a commercial networking provider which has established a connection to the University of Arizona in Tucson. From there, the University has ample (e.g., 10 Gbps and 100 Gbps) connectivity provided by the Sun Corridor Network, Internet2, and ESnet to foster the connection to NERSC.
- 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). Redundancy and potentially capacity increase if other users/more intense use cases emerge would help to add stability to the trajectory of the experiment.
- Policy-level differences between operational sites have an impact on tool adoption. The policies of KPNO and NOIRLab, which are NSF-funded facilities, differ from those of NERSC and DESI, which are DOE-funded experiments. This results in the use of software toolchain that is not as efficient as it can be.
- DESI leverages workflow and data-movement tools supported by NERSC, and is integrated into the facilities computational and storage resources.
- 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 the majority 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 DESI Case Study

5.3.2.1 Background

DESI is a DOE-led international cosmology collaboration designed to create the world's largest 3D map of the universe. This five-year survey starting in 2021 will observe the spectra of tens of millions of galaxies, quasars, and stars. These data will be used to construct the map to study the origins of the accelerated expansion of the universe and test models of gravity.

Observations begin at the Mayall 4-meter telescope at KPNO outside of Tucson, Arizona. Every ~15 minutes, DESI obtains spectra of 5,000 astronomical objects, resulting in 715 MB of data per exposure. These data are transferred in semi-real time to NERSC in Berkeley, California, where they are processed in a real-time queue such that results from each night are available to collaboration scientists each morning. Processed data are ~10 GB/exposure. Results from each night may be used to update the observing plan for the following night. Real-Time QA analysis uses computers at KPNO to guide observers throughout the night. Hosting the QA at KPNO allows DESI to be robust to network and NERSC outages.

NERSC is also used as the primary scientific analysis center for the ~900-person international collaboration of scientists, who access the data at NERSC for the cosmology analyses. Yearly "data assemblies" will reprocess all of the data taken to date; these will also be publicly released after the initial set of cosmology papers is published by the DESI collaboration. DESI projects that the total data volume will grow to 10 PB by the end of the survey in 2025. These data are available at NERSC for all collaborators, but in practice many collaborators also prefer to download a subset of the data to use at their local institutions.

5.3.2.2 Collaborators

DESI is an international collaboration with over 600 collaborators from over 70 institutions in 10 countries spanning North and South America, Europe, Asia, and Australia. Although the primary data processing and scientific analysis will occur at NERSC, many DESI collaborators will also download subsets of the data for analysis at their local institutions. Small volumes of data are typically accessed via scp or https, while more expert users download larger data sets using Globus. A mirror of the data is kept at the Astro Data Lab at the NSF's NOIRLab in Tucson, Arizona.

User/collaborator and location	Primary or secondary copy of the data	Data access method	Avg. size of data set	Frequency of data transfer or download	Are data sent back to the source and method	Any known issues with data sharing
KPNO (TUCSON, AZ)	Data origination	rsync	40 TB of the most recent data	~10 minute intervals (NERSC initiated)	Limited amount of calibration data	Incompatible multisite security policies (VPN, MFA, firewalls) required getting backdoor exceptions
NERSC (BERKELEY, CA)	Primary	Users directly log in to use the data at this site. Also serves data via Globus, rsync, and https	Growing to ~10 PB over the next five years	New data from KPNO downloaded every 10 minutes. User access intervals are unknown.	Processed subset sent back via rsync and git/svn repo syncing	Collaborators from certain countries are not allowed to have accounts at NERSC, thus limiting their data access to the primary data site. Downtimes and short mean-time-to-failure affect productivity of collaborators using data directly at this site
NERSC HIGH-PERFORMANCE STORAGE SYSTEM (HPSS) TAPE BACKUP (BERKELEY, CA)	Secondary	htar	Growing to ~10 PB over the next five years	Daily backups of raw and processed data	N/A	
NOIRLAB (TUCSON, AZ)	Secondary	Data portal	Growing to ~10 PB over the next five years	Daily	N/A	
INDIVIDUAL DESI INSTITUTIONS (WORLDWIDE)	Secondary subsets	Download from NERSC using scp, rsync, https, or Globus	Unknown, ad hoc	Ad hoc	N/A	Need better tools

Table 3: DESI data projections

The table is phrased in terms of the centers that host the DESI data, available to (nearly) all DESI collaborators, rather than itemizing the >70 collaborating institutions and how they each access the data.

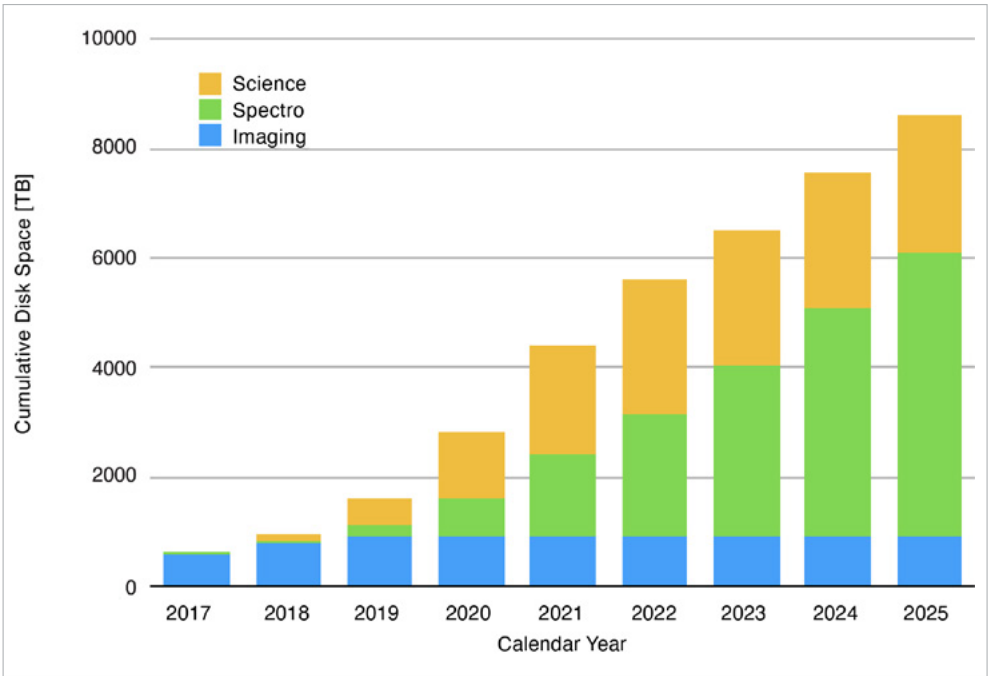


Figure 1: DESI data growth

Figure 1 shows the projected growth of DESI data through 2025, split into three categories: imaging (finishes in 2020, but kept on disk for science analyses), spectroscopy (the core DESI survey data), and science (simulations and analysis outputs).

5.3.2.3 Instruments and Facilities

Data originates from the DESI spectrographs at the Mayall 4-meter telescope at the KPNO near Tucson, Arizona. Raw data are transferred to NERSC for processing and science analysis, with subsets of the processed data downloaded to individual institutions for further customized/ad-hoc analysis. Raw and processed data are backed up to NERSC's HPSS tape system and mirrored to NOIRLab in Tucson, Arizona, as a geographically separated site. Raw data are retained at KPNO until they are verified to be at NERSC disk, NERSC tape, and NOIRLab disk.

5.3.2.4 Process of Science

Raw data are directly downloaded from KPNO to NERSC, which routes via the University of Arizona in Tucson where it routes over the Internet2 backbone network. The bandwidth is limited on the KPNO to Tucson 1 Gbps fiber link which is leased from the Tohono O'odham Utility Authority[43]. However, it is sufficient for the needs of DESI and the other telescopes operating at KPNO. A small cluster of computers at KPNO support observations, but this is not provided as a general resource to the collaboration for data analysis. As a remote mountain site, KPNO has occasional network outages, so a key operational requirement is that observations must be able to continue without network connectivity.

At NERSC, the data volume is projected to grow at a rate of ~1 PB/year from 2020 to 2025, reaching ~8 PB by the end of 2025. The data are stored in a hierarchy of thousands of directories with millions of files.

Public data sets will be curated beyond the timescale of DESI by the AstroLab at NOIRLab. The intention is to also host the public DESI data via the Cosmology Data Repository at NERSC, though a long-term funding commitment for that is not guaranteed.

[43] https://toua.net

DESI uses the full ecosystem of services available at NERSC including DTNs; the real-time queue for rapid processing; the interactive queue for data analysis by collaborators; the regular queue for large reprocessing runs; database servers; web servers for data distribution and visualization; and a Jupyter[44] server for data exploration.

5.3.2.5 Remote Science Activities

The data are obtained at KPNO (near Tucson, Arizona), processed at NERSC (in Berkeley, California), and accessed by a worldwide collaboration, so the workflow is fundamentally multisite. Although the total data volume will grow to be rather large at NERSC (8–10 PB), the real-time bandwidth needs to support the operations are relatively modest (e.g., ~715 MB every 15 minutes from KPNO to NERSC throughout the night). Remote collaboration access to the data is more limited by the convenience of the tools than raw network bandwidth.

5.3.2.6 Software Infrastructure

Data transfer from KPNO to NERSC is performed with rsync spawned by a custom daemon-like Python script that checks for new data every 10 minutes. Backups to NERSC HPSS tape use HPSS Tape Archiver (HTAR). Transfers of processed data from NERSC to the NOIRLab mirror use Globus for the initial bulk transfer, then rsync to pick up any later differences. Transfers of data within the NERSC center use a combination of Globus, rsync, and Unix cp.

Collaborators primarily access the data through direct login to NERSC, or through NERSC's Jupyter server. Many collaborators have expressed a preference for downloading subsets of data to use locally, though DESI lacks developed tools for easy management of that remote analysis workflow.

5.3.2.7 Network and Data Architecture

DESI infrastructure at KPNO shares resources with several other scientific projects. **Figure 2** shows the Layer 2 network organization, and **Figure 3** shows the Layer 3 network organization.

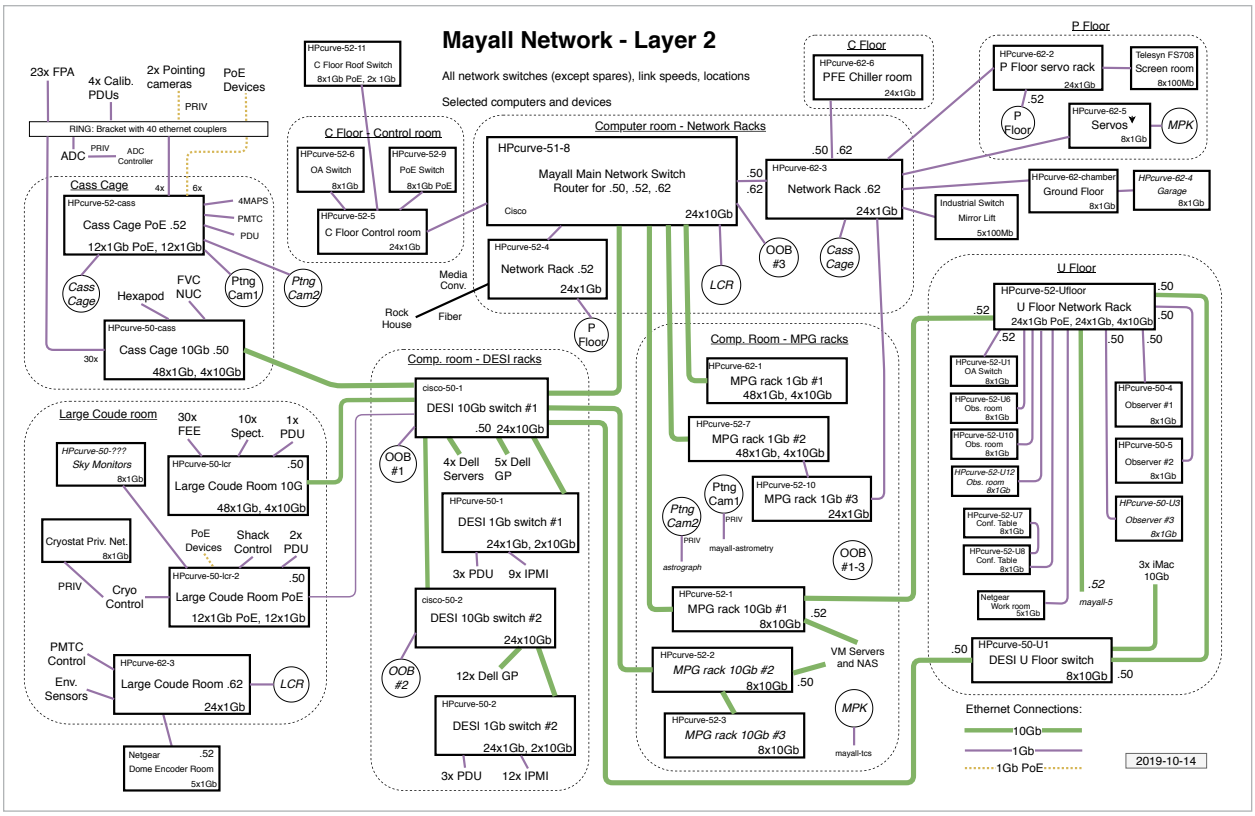


Figure 2: Layer 2 network organization

[44] https://jupyter.org

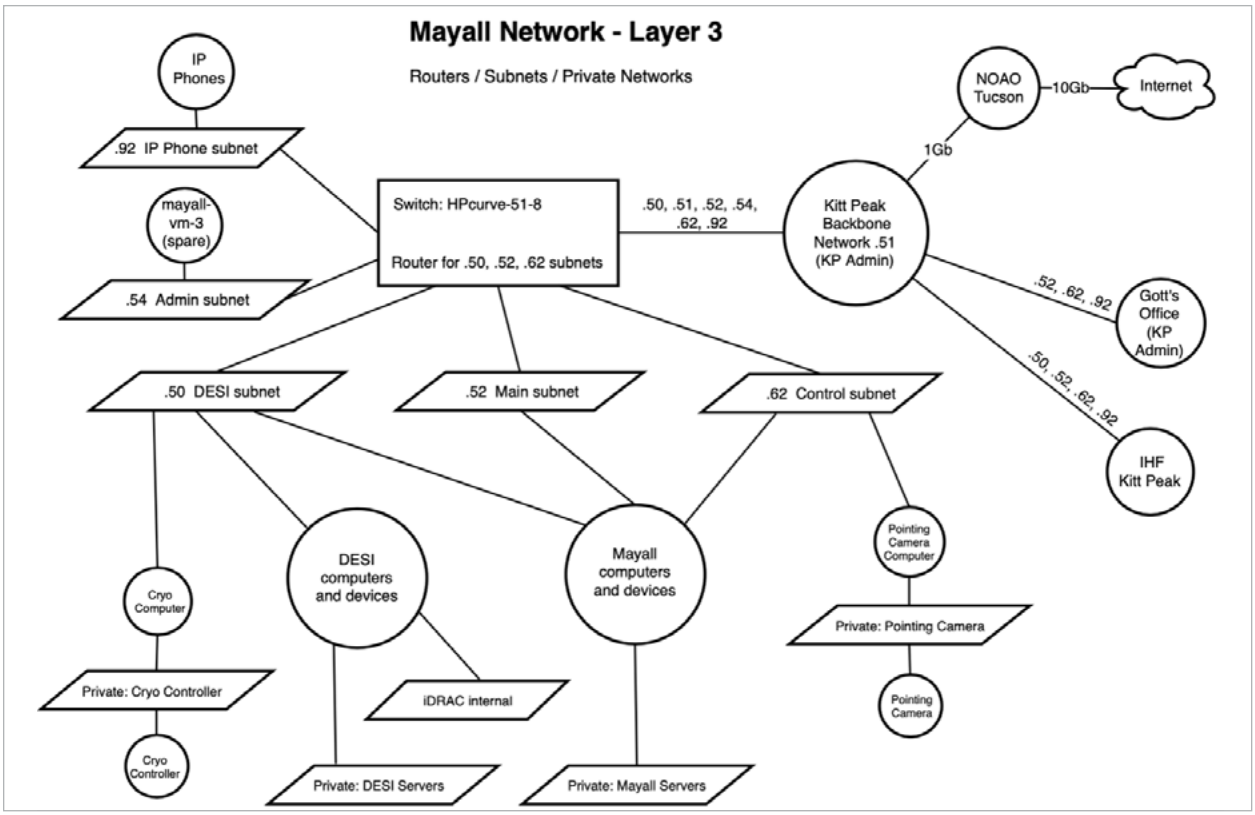


Figure 3: Layer 3 network organization

5.3.2.8 Cloud Services

DESI uses Amazon Web Services (AWS) for hosting its wiki, svn repository, mailing lists, document repository, publication database, project website, and a public-facing version of the imaging data browser[45].

DESI does not currently use cloud services for data processing or serving core data products. On the one- to two-year timescale, it will explore deploying the data processing pipeline on AWS as an alternate site during NERSC outages, though data transfer and hosting costs make this prohibitive as a primary computing solution.

5.3.2.9 Data-Related Resource Constraints

Within the NERSC computing center, DESI is limited by I/O server metadata access (not capacity or bandwidth).

Although astronomy data sets are made publicly available, this is done through https servers providing individual files for download, not through systems designed for accessing and jointly processing entire data sets from different sites, and it is rare for data portals to provide any significant amount of computing resources to process their data by external users. This leaves users on their own to download and process data at different data centers in an uncoordinated manner.

5.3.2.10 Outstanding Issues

Broadly speaking, DESI is not limited by raw bandwidth — organizations like ESnet and Internet2 have stayed ahead of the curve such that bandwidth is not a problem. There are limitations by the convenience of tools for managing cross-site data flow. Globus is a huge step forward for efficient one-off data transfers, but issues such as

[45] http://viewer.legacysurvey.org/

expiring certificates, multifactor authentication, and data discovery (i.e., selecting what needs to be downloaded) limit its convenience for day-to-day random access of huge data sets.

Ideally, a service like Apple's iCloud Photos could be developed which would allow devices to have access to a library of photos much larger than can fit on the device. The device and the iCloud storage would sync back and forth transparently to the user. Ideally computing centers and end users would be able to subscribe to data sets from each other, with automatic efficient syncing of the subsets of the data that are actively being used, with the toolkits managing purging unused data to stay within the local quota limits. The faster the network is, the more viable this becomes, since it reduces the barrier to having to re-download data.

5.3.2.11 Case Study Contributors

DESI Representation

- Stephen Bailey[46] (Data Management Lead), LBNL
- Michael Levi[47] (DESI Director), LBNL
- David Schelgel[48] (Co-Project Scientist), LBNL
- Julien Guy[49] (Co-Project Scientist), LBNL
- Daniel Eisenstein[50] (Co-Spokesperson), Harvard University
- Nathalie Palanque-Delabrouie[51] (Co-Spokesperson), the French Alternative Energies and Atomic Energy Commission (CEA)

ESnet Site Coordinator Committee Representation

- Rune Stromsness[52], LBNL
- Richard Simon[53], LBNL
- Damian Hazen[54], LBNL and NERSC
- Tavia Stone Gibbins[55], LBNL and NERSC

5.4 The Rubin Observatory and the LSST

The Rubin Observatory, previously referred to as the Large Synoptic Survey Telescope (LSST), is an astronomical observatory currently under construction in Chile and the United States. 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 charge-coupled devices (CCD) imaging camera. For the purposes of the DOE, there are several dark energy experiments (notably DESC) that will utilize data produced by Rubin on a yearly basis. The COVID-19 pandemic has stopped some progress, namely the physical construction at the site (restarted November 2020). Work on the camera has proceeded, with some promising early results in a laboratory environment at SLAC.

[46] stephenbailey@lbl.gov
[47] melevi@lbl.gov
[48] djschlegel@lbl.gov
[49] jguy@lbl.gov
[50] deisenstein@cfa.harvard.edu
[51] nathalie.palanque-delabrouille@cea.fr
[52] rstrom@lbl.gov
[53] rsimon@lbl.gov
[54] dhazen@lbl.gov
[55] tavia@lbl.gov

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 USDF at SLAC. 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. An interim DF is underway using the GCP to begin to test software for analysis, as well as operational aspects.

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, transferred from the top of the mountain and to the USDF, and processed using the analysis toolchain; 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 therefore 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 most of the year processing raw data for a yearly data release. This release will then be made available to scientists in the United States, 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.

5.4.1 Discussion Summary

The following discussion points were extracted from the case study and virtual meetings with the case study authors. These are presented as a summary of the entire case study, but do not represent the entire spectrum of challenges, opportunities, or solutions.

- 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-IN2P), for primary and secondary processing and storage.
- Transient events are defined to be short-time window bursts (that are sized approximately 13 GB in two images) of objects that will require special processing.
  - It will be necessary to send the data from Chile to the United States, complete processing, and alert a series of brokers to the existence of the transient so that it may become available to subscribers and potentially get more attention from other observatories within one minute of capture.
  - 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.
  - The Rubin Observatory does not expect extensive "off-platform" data use, and will expect that most analysis will be done by the user community through the provided framework.

- Off-site use for affiliated projects (e.g., DESC) will be organized in a structured manner to allow for bulk-data movement (potentially yearly, to coincide with data-product releases).

- 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-IN2P.
- An interim facility to be used for testing tools, and housing pre-survey operational data, is being staged in the GCP, and will be available in 2021.
- The major data streams will thus be Chile to the USDF at SLAC, and the USDF at SLAC to France.
  - Alerts will also be a use case, but will not comprise the bulk of network volume.
- WAN requirements are focused 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, the United States, and France) to guarantee operational stability.

5.4.2 The Rubin Observatory Case Study

5.4.2.1 Background

Rubin Observatory will carry out the LSST, using the Simonyi Survey Telescope and the Rubin Observatory LSST Camera. It will take repeated images over more than 18,000 deg2 of the southern sky in six broad-band optical filters (ugrizy). The resulting data set will provide a static census of approximately 40 billion objects as well as a dynamic time-domain census over an unprecedented range of timescales and flux limits. The primary deliverables for Rubin Observatory include a real-time stream of "alerts" of transient events, prompt data products, annual data release data products, and a Science Platform specifically developed to reduce the barrier of entry to Rubin Observatory data and to shorten the path to science for the user community.

Rubin expects ~30 TB of image and metadata to flow from Chile to the United States each night. Real-time transfer over dedicated high-bandwidth segments is required to analyze each image at the USDF and broadcast alerts to the community of all sources in the image that have varied in brightness by more than 5 sigma compared to a reference image. Alerts are required to be broadcast via community-developed brokers that receive a stream from Rubin in 60 seconds from the time the camera shutter closes on the mountain top in Chile. Other "prompt data products" are made available to the community within 24 hours.

Nightly data are accumulated and processed with science pipelines at the USDF and a DF in France, producing deep stacked images and source catalogs. These "annual data release" products are served to the user community via a Rubin Science Platform hosted at the USDF, to be located at SLAC.

Rubin expects significant amounts of data will be moved in bulk to large science collaborations for custom analysis. In the case of DESC,[56] this will amount to perhaps 10% of the accumulated images in any given year. Rubin would expect that at least DESC transfers will be made via ESnet. ESnet-based transfer to the French DF is a possibility, too.

All Rubin data are proprietary for two years except the alert stream which is world public. After two years, the data are shareable with anyone, but no specific mechanism for making these data public and serving them is yet defined.

[56] https://lsstdesc.org

5.4.2.2 Collaborators

Rubin Observatory is an NSF-DOE funded partnership. Operations partners are NSF's NOIRLab[57] (managed by AURA[58]) and SLAC[59]. Rubin collaborates with three operations affiliate partners: Chile, host country of the telescope facility, Brazil (providing high-bandwidth network connectivity from Santiago to São Paulo), and CC-IN2P3[60] physics institute in France (providing 50% of the annual data processing capacity).

The science community includes data scientists, astronomers, and physicists in the United States and Chile as well as certain physicists and astronomers in Brazil and France. International members in Europe, Asia, and North America who have approved in-kind contributions to the Rubin science enterprise will also be collaborators with equal access and rights to work with Rubin data. All these collaborators can work independently or through self-governed science collaborations. Presently, Rubin engages with eight such science collaborations that largely drive the discussion about the science needs of the community.

Depending on the size of the in-kind contribution, international members may have agreements with DOE or NSF directly or with AURA or SLAC.

User/collaborator and location	Primary or secondary copy of the data	Data access method	Avg. size of data set	Frequency of data transfer or download	Are data sent back to the source and method	Any known issues with data sharing
USDF (E.G., SLAC)	Primary	Portal for all users (plus redundant high-speed data transfer to/from Chile not for general user access), broadcast of alerts packets to selected brokers	5 PB/yr. With ancillary and other image data products, up to 500 PB by end of survey in 2035.	Nightly northbound transfers for all images and metadata. Alert production to brokers involves two seven second bursts evert minute.	Processed images/ alerts/ data releases/ calibration data sent to the data center in Chile (southbound)	N/A
CHILE	Primary	Data portal for Chilean users	5 PB/yr	Nightly raw images stored. Weekly/ monthly calibration data. Annual data release data products.	N/A	N/A
BRAZIL	Secondary	None, data only transits	5 PB/yr	Same as Chile (same data transits for Brazil)	Same as USDF (same data transits for Brazil)	N/A
CC-IN2P3	Secondary	Portal (possibly), default is USDF	5 PB/yr	Nightly raw images sent at low or modest rate. ½ of annual data release products processed at USDF sent here.	½ data release products processed here returned to USDF	N/A

[57] https://noirlab.edu/public/
[58] https://www.aura-astronomy.org
[59] https://www6.slac.stanford.edu
[60] https://cc.in2p3.fr/en/

User/collaborator and location	Primary or secondary copy of the data	Data access method	Avg. size of data set	Frequency of data transfer or download	Are data sent back to the source and method	Any known issues with data sharing
DESC	Not Rubin supported (10% annual images plus custom derived data products)	Not Rubin supported	Few PB/yr	Modest transfer rate for fraction of annual images	Some custom data products could be sent back to USDF for users generally	N/A
SLAC (CAMERA R&D THROUGH 2023)	N/A	No general data access, only camera servicing activity	Few TB/week	Few TB/week	N/A	N/A
RUBIN OBSERVATORY HQ TUCSON	N/A	No general data access, observatory science and management activity	Few TB/week	Few TB/week	N/A	N/A

Table 4: Rubin Observatory data projections

5.4.2.2.1 North–South Networking

All image data will be carried to the project's Archive Center at the USDF at SLAC. The total northbound data from the observatory to the Archive Center is approximately 20 TB/night. A very demanding application using these flows is Alert Production, where 12.7 GB (uncompressed) of pixel data need to be transferred from the Base Center in La Serena, Chile, to the Archive Center in seven-second bursts. There are two such bursts approximately every minute of nighttime observing. The bursts need to be sustained for approximately 355 days in each year. The required reliability level for these flows is very high, and is seen to require a nearly lossless path-diverse network.

This same network carries substantial traffic from the USDF to the Rubin Observatory Base Center La Serena, in the form of calibration data and annual data releases, but the aggregate volume is not as large as the South–North traffic, nor is there a near-real-time latency requirement for that traffic. Establishing duplex service at the level required for the South–North traffic should more than suffice for the North–South traffic.

Lesser, but still operationally significant, volumes of data will flow from the Base Center to the Rubin Observatory Headquarters site in Tucson, then to SLAC and to other sites involved in LSST operations.

5.4.2.2.2 National and International Networking

The USDF will provide raw and processed data to Rubin Observatory authorized users, including DESC and its international collaborators, which can include science collaborations investigating dark energy in the UK and Europe. The USDF will distribute data to one or more data access centers (DACs) in Europe and elsewhere. The scope of these centers and their reliance on the USDF is currently being worked out. It is worth noting that the data volume of an annual release is very significant, with just the catalogs of detected objects beginning at petabyte-scale.

International networking has a role in supporting Rubin Observatory LSST production processes. LSST annual releases will be computed at the USDF at SLAC in California and at the satellite computing facility at CC-IN2P3 in Lyon, France. This generates a need for networking to support continuous data production, in addition to distribution of data to scientific data analysis. By agreement, CC-IN2P3 is the responsible institution for providing connectivity between CC-IN2P3 and the United States.

5.4.2.3 Instruments and Facilities

The Rubin Observatory Data Management System and Facilities[61] describes the full details of instrument operations. The data originate from the 3.6 gigapixel LSSTCam mounted on the 8.3m telescope located on the summit of Cerro Pachón in Chile. Data will be transported via a long-haul network from the telescope to a Chilean Data Center and then to the USDF at SLAC. A backup stream of data is then routed to the French DF at IN2P3 in Lyon.

The operations of Rubin Observatory are broken into three phases:

- 2019–2023, pre-operations.
- 2024–2033, main survey.
- 2034–2035, post operations.

Pre-operations will run early versions of all the data production elements described previously. In 2021–2023, some of this will be done utilizing the GCP starting in FY21. The goal is to use precursor data sets (simulated or from other telescopes); more information is available in Section 5.4.2.8. Pre-operations will culminate with data production at full-scale testing, ideally using science validation survey data from the telescope and LSSTCam obtained in the last part of commissioning. This is currently forecast in mid-2023 based on delays due to COVID19.

Transition to the USDF in the pre-operations phase is expected to begin in 2022, reaching full-scale readiness before operations hand over in 2023. The full LSST will run for 10 years and final data processing will take place for two years after that. Survey data taking ends in 2033.

5.4.2.4 Process of Science

The Rubin LSST is a public survey, and the Rubin team itself is not responsible for the delivered science. The DOE is chiefly involved in Rubin for dark matter and dark energy science via DESC, which is described in Section 5.2.

For the NSF, the science goals are based on studies of the nature and distribution of solar system bodies, the halo of the Milky Way, galaxy formation and evolution, and the transient and variable universe. Each of these science areas has an associated science collaboration. These are self-governed and only DESC has a direct funding connection to one of the agencies (the DOE).

All the collaborations as well as independent scientists will access and analyze data through the Rubin Science Platform (RSP) at the USDF or possibly at other data centers. There will be a DAC at the base facility in Chile. Chileans will have priority access to this DAC, but others can use it as well. All the collaborations, as well as independent scientists, will access and analyze data through the Rubin Science Platform (RSP) at the USDF or possibly at other data centers. There will be a DAC at the base facility in Chile. Chileans will have priority access to this DAC, but others can use it as well. Discussions are ongoing regarding contributions from various international members in the collaboration to establish other DAC locations around the world, including Europe.

All scientists will have compute and storage resources at SLAC or the Chilean DAC. Tools available next to the data will allow complex queries of the object and source catalogs as well as statistical and other common analyses.

The prompt processing will generate a nightly stream of sources which vary in brightness. This stream will be made available to at least five community brokers. These brokers allow (outside Rubin) access to the alert stream, including the tools needed to filter the stream for specific science cases. At least some of the brokers will be open to any member of the worldwide science community.

Solar system objects will be identified in the nightly processing and new and repeat observations of these bodies will be sent to the Minor Planet Center (MPC)[62] for further orbit determination and cataloging. Scientists can access the latest information from the MPC or the RSP.

A subset of Rubin data will be managed by the Rubin and NOIRLab education teams. This cloud-based resource will support public engagement and citizen science.

[61] https://docs.google.com/document/d/1YW2QMdrmE4MwjDHJw7v8eyYSTGAjfWHYSBXt4bK-nRc/edit?usp=sharing
[62] https://www.minorplanetcenter.net/iau/mpc.html

5.4.2.4.1 Network Use in System Integration and Commissioning

The Rubin Observatory networks are used in several stages of observatory integration and commissioning. Prior to this, the networks undergo routing testing in a pre-verification stage, as described in the LSST LHN End-to-End_Plan.[63] This plan also includes documentation of the extensive use of perfSONAR to instrument and test the networks.

Then, to formally verify the networks against system requirements, in conjunction with observatory integration and commissioning, a series of tests are conducted in increasing scales (data rates, data volumes, devices) as described in LDM-732 Rubin Observatory Network Verification and Validation Plan.[64]

Prior to and during commissioning, Rubin Observatory does a number of data management operations rehearsals, with the goal of exercising the operations processes and rehearsing the interactions of the operational personnel. These also utilize the networks.

Finally, as the observatory moves through commissioning, a series of data previews are performed to accomplish end-to-end data capture, transfer, and processing. Again, the networks are utilized in the data previews.

MILESTONE LEVEL	ID	ACTIVITY	BASELINE FINISH	PROJECTED FINISH	END-TO-END B/W, Cerro Pachon - La Serena	END-TO-END B/W, La Serena - NCSA	Bandwidth Achieved through Demonstration
2	DM-NET-1	Base - Archive Network Functional 1 Gbps	6/11/15	6/11/15	0.5 Gbps	Max: 1G Best Effort	0.5 Gbps (operational)
2, 3	DM-NET-2, -3, 6, DMTC-6800-1310	Mountain - Base Network Functional 2 x 100 Gbps, Summit LAN Installed, Initial Network Ready (Summit), Network Acceptance/Verification Review for Early Integration	3/27/18	6/30/18	2 x 100 (shared AURA DWDM)	Max: 20G Best Effort	46Gbps (LSST First Light demo)
3	DM-NET-4, DMTC-6800-1320	Base LAN Installed, Network Acceptance/Verification Review for ComCam on bench	9/16/20	9/16/20	6 x 100 (dedicated LSST DWDM) + 2 x 100 (shared AURA DWDM)	Max: 20G Best Effort	80Gbps (LSST SC18 demo)
		Auxiliary Telescope Spectrograph on Sky Observing		1/1/20			
		Network Pre-Verification Planning/Progress Review		6/17/20			
		Commissioning Camera on Test Bench Observing		7/31/20			
		DM Operations Rehearsal #2 (AuxTel data)		8/15/20			
3	DMTC-6800-1325	Network Acceptance/Verification Review for Full Integration	9/16/20	9/16/20	6 x 100 (dedicated LSST DWDM) + 2 x 100 (shared AURA DWDM)	Max: 140G Best Effort	
		DM Operations Rehearsal #3 ComCam on bench data)		11/1/20			
2	DM-NET-5	Base - Archive Network Functional 100 Gbps	7/3/19	11/30/20	6 x 100 (dedicated LSST DWDM) + 2 x 100 (shared AURA DWDM)	Max: 140G Best Effort	
		Data Preview 0 Start		12/15/20			
		Commissioning Camera on Sky Observing		4/23/21			
3	DMTC-6800-1330	Network Acceptance/Verification Review for Science Verification	7/2/21	7/2/21	6 x 100 (dedicated LSST DWDM) + 2 x 100 (shared AURA DWDM)	Max: 140G Best Effort	
		Data Preview 1 Start		8/24/21			
		Ful Camera on Sky Observing		10/12/21			
		Data Preview 2 Start		2/15/22			
3	DMTC-6800-1340	Network Acceptance/Verification Review for Full Operations	6/30/22	6/30/22	6 x 100 (dedicated LSST DWDM) + 2 x 100 (shared AURA DWDM)	Max Dedicated: 200G dedicated up to NCSA. 300G at NCSA. Burst: 100G	

Table 5: Rubin Observatory milestones (pre-COVID)

5.4.2.5 Remote Science Activities

By definition, all Rubin science is done remotely from the telescope facility and physically removed from the USDF.

5.4.2.6 Software Infrastructure

The Rubin Observatory Data Management System and Facilities[65] provides an overview of the software and service infrastructure for the Rubin Observatory. The following sections detail this.

[63] https://docushare.lsstcorp.org/docushare/dsweb/Get/Document-14789/Rubin%20Observatory%20Network%20End-to-End%20Testing%20Plan.docx

[64] https://ldm-732.lsst.io/v/DM-25765/index.html

[65] https://docs.google.com/document/d/1YW2QMdrmE4MwjDHJw7v8eyYSTGAjfWHYSBXt4bK-nRc/edit?usp=sharing

5.4.2.6.1 Rubin Observatory Data Management System Architecture

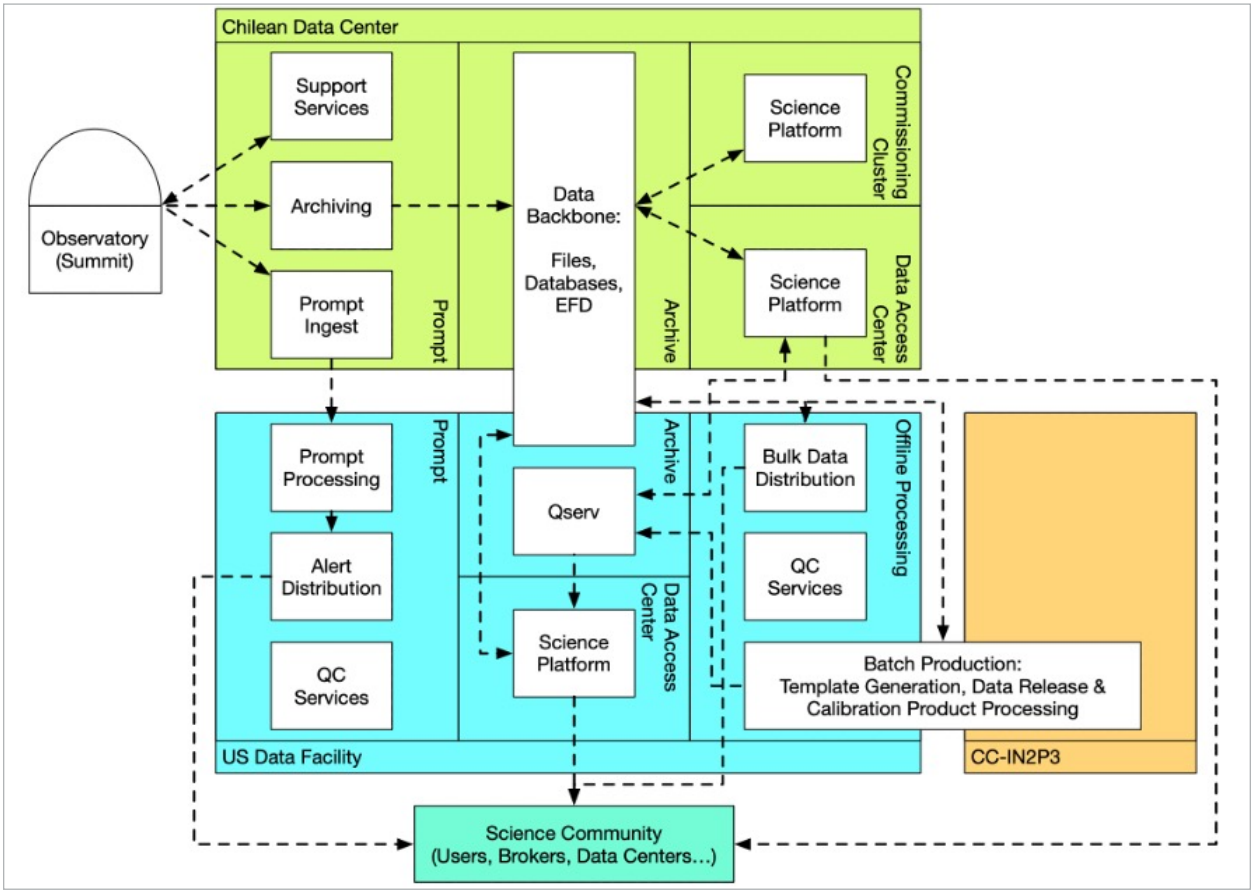


Figure 4: Architectural overview of the Rubin Observatory Data Management System. Colored boxes represent the sites over which the Data Management System is deployed; divisions within them show the deployment enclaves; white boxes indicate particular services.

The Rubin Observatory Data Management System (DMS) is the combined hardware and software infrastructure that will collect data from the Rubin Observatory LSSTCam and other ancillary instrumentation; transport, archive, and process it into science-ready data products; and make the raw data and derived data products available to the community.

The DMS is structured around a number of services, which are deployed and managed within enclaves, and hosted over three separate sites — the Chilean Data Center, the USDF, and CC-IN2P3 — connected by wide-area networks. This architecture is shown schematically in **Figure 4.**

Each site is separated into functional enclaves based on requirements for availability, access, and change management. Enclaves are distinguished by having different users, operational timescales, interfaces, and services. The Chilean Data Center supports four enclaves: Prompt Base, Archive Base, Commissioning Cluster, and Chilean DAC. The USDF provides Prompt USDF, Archive USDF, USDAC, and Offline Processing USDF enclaves. A single Offline Processing Satellite enclave is deployed at IN2P3.

Each enclave supports one or more services. These services can be considered as consisting of four tiers of components:

1. The Science Platform, which provides a user interface and data AFs for use by both Observatory staff and the wider scientific community.

2. Applications software, which provides libraries of data structures and algorithmic code for representing and processing Rubin Observatory data. These libraries are available for use through the Science Platform, and are used to construct the science pipelines which are executed to generate the standard Rubin Observatory data products.
3. Middleware is used to provide the environment within which the science pipelines execute. It provides a layer of isolation and abstraction between the pipeline "payloads" and the underlying infrastructure on which they execute, and provides data access services to both science users and observatory staff.
4. Infrastructure — the compute hardware, storage, and networking, and the low-level services and software necessary to run and manage a data center — underpins and supports all of the previous component tiers.

The various enclaves, and the services hosted within them, are described in subsequent sections.

5.4.2.6.2 Compute and Storage Sizing

The DMS construction team has completed an extensive evaluation of the compute and storage infrastructure necessary to process and release all of the data collected during the Rubin Observatory LSST (the "sizing model"), and have further used that to estimate the total non-labor cost of the system (the "costing model"). The full details of this evaluation are available in technical note DMTN-135[66].

It is worth noting that the sizing and costing models described in this section have been developed based on the prototype DF built by the Construction Project at the National Center for Supercomputing Applications at University of Illinois at Urbana-Champaign (NCSA). The ultimate location of the operational DF has recently been confirmed to be SLAC. As plans are formulated, there may be some changes to the sizing and costing model for computing due to technology choices, the availability of pre-existing infrastructure, and hosting costs. However, the algorithmic approach taken, and the size and type of the resulting data products, will remain unchanged, so these impacts will be limited.

5.4.2.6.2.1 Storage Requirements

The storage sizing model is derived on the basis of six fundamental considerations:

- LSE-81[67], which provides estimates of the total number of sources that will be observed by the Rubin Observatory based on the LSST Science Book[68], the SRD (LPM-17[69]), and the Observatory System Specifications (LSE-30[70]).
- LSE-163[71], the Data Products Definition Document, which summarizes the various types of image that will be made available, and describes the information that will be included in Rubin Observatory catalogs; and LDM-153[72] which maps those catalogs onto a detailed database schema.
- Practical experience of processing precursor data sets with prototype versions of the Rubin Observatory science pipelines, which helps to establish intermediate data products that must be stored temporarily; provides a basis for estimating how effective compression is when applied to Rubin Observatory data; and establishes the size of ancillary data products, like representations of the point-spread function.

[66] https://dmtn-135.lsst.io
[67] https://ls.st/lse-81
[68] https://www.lsst.org/scientists/scibook
[69] https://ls.st/lpm-17
[70] https://ls.st/lse-30
[71] https://ls.st/lse-163
[72] https://ls.st/ldm-153

- Science Platform users will have access to personal storage space for generating and storing their work.
- Access requirements for data products, including ensuring that products where low-latency access is required are stored on fast storage; that all data releases are archived to tape; and that the two most recent data releases are available on spinning disk (LPM-151[73]).
- The approach taken to disaster recovery, which determines how much storage redundancy is required.

Based on these considerations, Rubin identifies five separate classes of storage:

- Fast storage, likely comprised of solid-state devices, is used for the Alert Production Database and the client-facing query management and indexing components of the distributed Qserv database (Section 5.4.2.6.4.2.2).
- Normal storage, corresponding to networked filesystems hosted on enterprise-grade spinning magnetic disks, is used for initial collection of raw data, and intermediate and final data products being generated or modified during the production of a data release, and for user storage.
- Object stores, hosted on consumer-grade spinning disks, are used for providing cheap and scalable access to read-only data. In general, Rubin expects data to migrate from filesystem-based "normal" storage to object stores when processing is completed.
- Qserv storage is dedicated to the distributed database system used to manage data release catalogs. The Qserv system is intrinsically fault tolerant, so consumer-grade spinning disks are used.
- Tape storage is used for archiving and backup of all data products hosted on other storage media.

At any given time, the two most recent data releases are made available on the "live" Qserv and object store system; older data releases are only available from tape, by special arrangement and at substantially greater latency.

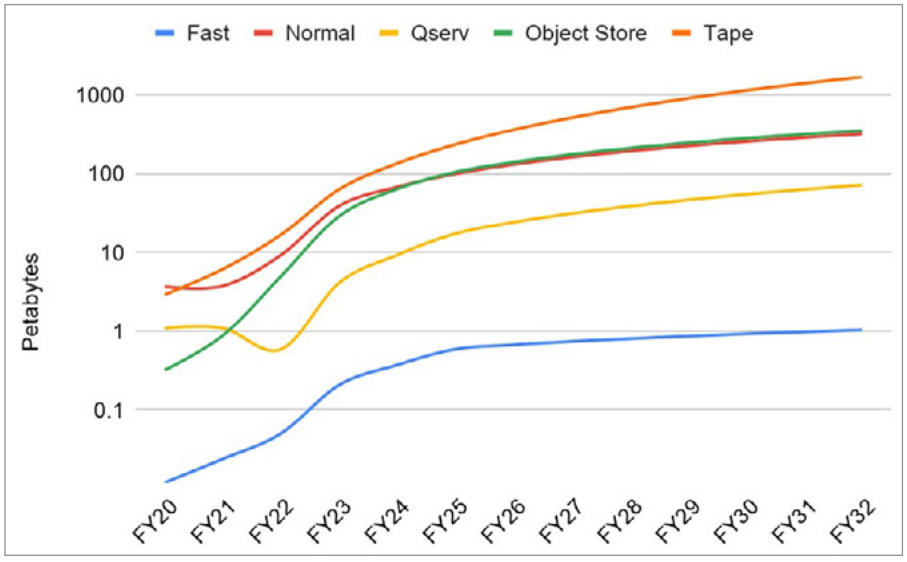


Figure 5: Total storage required at the USDF as a function of time

[73] https://ls.st/lpm-151

The USDF at SLAC will store and make available to the data rights community all Rubin Observatory LSST data products. It will also archive old data releases and maintain backup copies of the data. Finally, it provides personal storage for Science Platform users[74]. Taken together, these imply the total storage requirements for the USDF shown in **Figure 5.** Note that this figure also shows storage required in support of construction, commissioning, and pre-operations activities before the start of full operations. Storage used by the construction project will be purchased by the construction budget, and will (unless retired) transition into the operations project. However, the bulk of storage required in operations will be purchased under the operations budget.

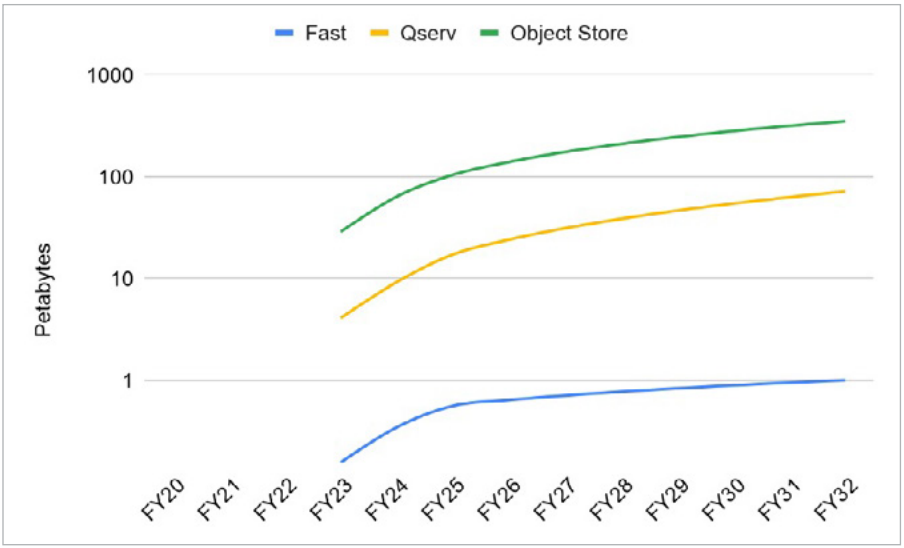


Figure 6: Total storage required at the Chilean DAC as a function of time

Storage is also required to support the activities of the Chilean DAC (Section 5.4.2.6.3.4). The Chilean DAC is not required to support any activities before the start of full operations; no storage is required before until the first data release is made available at the start of FY23. Further, the model assumes that all data in Chile are hosted on fast storage (Qserv query management and index), Qserv storage (the bulk Qserv database), or object storage (for image and file-based data products). The expected evolution of storage at the Chilean DAC is shown in **Figure 6.**

5.4.2.6.2.2 Compute Requirements

Compute requirements are estimated separately for prompt processing (Section 5.4.2.6.4.1.1), data release processing, and the Science Platform. The Science Platform is made available for users at the USDAC (Section 5.4.2.6.4.4), the Chilean DAC, and for internal staff use; Rubin considers each of these use cases separately. This analysis does not consider other processes, such as periodic template generation or calibration products processing, which are assumed to be negligible.

Estimated compute requirements for the science pipelines are derived from processing precursor data — primarily taken from Hyper Suprime-Cam on the Subaru Telescope and DECam on the Victor M. Blanco Telescope[75] — through existing pipeline prototypes, and scaling the results based on the source counts, derived from scientific considerations, documented in LSE-81[76]. Compute times are measured on existing hardware and converted into core hours on fiducial hardware[77]. This estimation methodology incorporates all I/O,

[74] Rubin assumes 5,000 users each having access to 0.4 TB in the first year of operations, scaling to 7,500 users having access to 1.3 TB by the end of the survey.
[75] https://www.darkenergysurvey.org/the-des-project/instrument/
[76] https://ls.st/lse-81
[77] For conversion between different architectures, the ratio of industry-reported achievable FLOPS is used.

memorybandwidth, cache miss, and other overheads into the core-hour measurement. Note that the fiducial hardware[78] is not assumed to evolve with time. In practice, Rubin expects that continued technology evolution will result in a reduction of the number of core hours required late in the survey relative to the estimates presented here.

In some cases, Rubin Observatory algorithms are still under active development, and are not yet appropriate for making informed estimates of actual compute cost. In these cases, core-hour figures obtained for comparable jobs from the DES[79] have been used in their place. Where equivalent DES and Rubin codes are available (e.g., in single visit processing), run times have been shown to be well matched, which gives us confidence in this basis of estimate.

Based on the previous considerations, Rubin requires a constant 1,188 CPU cores throughout the survey period to perform prompt processing. These cores will be used during the night for alert generation, and during the day for catch-up (where necessary) and solar system processing. Each core will have access to 5 GB of memory, and all of these cores are located at the USDF.

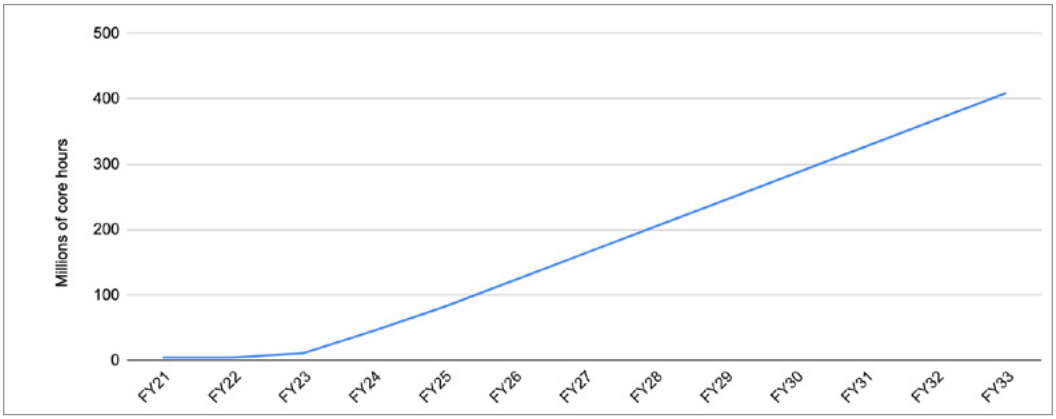


Figure 7: Fiducial core hours required for data release processing as a function of time

The core hours required for data release processing increase as the volume of data collected grows, as shown in **Figure 7.** One-third of these cores will have access to 5 GB of memory; the remainder, access to 20 GB; jobs will be allocated to high- or low-memory core depending on their requirements (e.g., operations on a single CCD will generally be executed on a low-memory system, while combining multiple detectors or deblending complex scenes will require a high-memory system). These cores will be divided between the USDF and the Satellite Data Processing (Section 5.4.2.6.5) to enable split-site processing.

Cores are allocated to user-driven processing in the Science Platform as a ratio of the total available compute system, following the construction SRD (LPM-17[80]). On this basis, Rubin assumes that:

- USDF user computing is sized as 10% of the total data release processing compute, amounting to over 500 cores at the start of operations.
- Chilean DAC computing is sized as 20% of USDF user computing.
- Project staff computing is sized as 10% of USDF user computing.

It will be possible to dynamically reallocate CPU cores between services at the USDF. In particular, it will be possible to reallocate cores from data release processing to the Science Platform at times of high demand, assuming that the long-term average level of data release compute is adequate to meet the release schedule.

[78] Intel Xeon E5-2680v3 at 2.50 GHz
[79] https://www.darkenergysurvey.org
[80] https://ls.st/lpm-17

5.4.2.6.3 Chilean Data Center

The Chilean Data Center is hosted at the La Serena Base Computing Center. This center is a megawatt facility with redundant power service provided by on-site gas generators. The Chilean Data Center will have dual 100 Gbps network links to the summit facility — to receive streamed data from the telescope and to enable remote control — and will have dedicated dual point of presence (POP) and path-diverse 100 Gbps networking to SLAC as well as connectivity to dedicated R&E networks and the commodity (or open commercial) internet.

Through an agreement with the other AURA centers on the NOIRLab Recinto[81], the computing center will be used to consolidate computing and network servers from CTIO, SOAR, and AURA operations (and likely Gemini in the future).

The Chilean Data Center hosts four DMS enclaves, as described later in this case study. In addition, it will host a backup of the entire Rubin Observatory data set.

The Chilean Data Center will be operated by the Observatory Information Technology Center (ITC) team, under the supervision of the Observatory ITC Manager. This team is part of the Observatory Operations Department. Most services deployed within the various enclaves will operate software which is developed and maintained by the Data Production Department (primarily, the Science Users Middleware and Science Platform and Reliability Engineering groups).

5.4.2.6.3.1 Prompt Base

The Prompt Base enclave exists to support services that must interact with the Observatory Control System (OCS) or the Camera DAQ. In several cases, these services are closely related to counterparts in the Prompt USDF enclave (Section 5.4.2.6.4.1). Services in this enclave are, where possible, designed to operate autonomously and with a high degree of fault tolerance; the enclave is designed to be functional whenever the Observatory is taking data.

The Prompt Base enclave supports the following services.

5.4.2.6.3.1.1 Archiving

The Archiving Service captures raw images taken by the Observatory's cameras (both the main LSSTCam and the Auxiliary Telescope spectrograph), including data from wavefront and guide sensors. These are augmented by metadata and telemetry captured by the Header Service, which forms part of the Archiving Service. Image data and metadata are then supplied to the Observatory Operations Data Service (OODS; Section 5.4.2.6.3.1.4) and staged for ingestion into the Data Backbone (Section 5.4.2.6.6).

The Archiving Service also includes a "catch-up" capability, which can stage any data to the OODS and Data Backbone that were missed by the primary archiving system due to network or other outages.

5.4.2.6.3.1.2 Planned Observation Publication

The Planned Observation Publication Service retrieves telemetry from the OCS describing the location of the next visit and the scheduler's predictions of future visits. It makes these available as a globally accessible web page for human inspection, and via machine readable web APIs.

5.4.2.6.3.1.3 Prompt Processing Ingest

The Prompt Processing Ingest Service captures images from the LSSTCam data system together with selected metadata from the OCS and forwards them directly to the Prompt Processing Service in the Prompt USDF enclave (Section 5.4.2.6.4.1.1).

5.4.2.6.3.1.4 Observatory Operations Data

The OODS provides access to files and metadata to Observatory systems. It is designed to provide lower latency than the Data Backbone (Section 5.4.2.6.6) to facilitate fast turnaround analyses by the team in Chile.

[81] https://noirlab.edu/public/about/contacts/aura-recinto/

5.4.2.6.3.1.5 OCS Driven Batch

The OCS Driven Batch Service provides a batch computing service capable of executing science payloads. It is used for processing daily calibration data. Data are read from and written to the Data Backbone (Section 5.4.2.6.6).

5.4.2.6.3.1.6 Telemetry Gateway

The Telemetry Gateway returns information from the Prompt USDF enclave (Section 5.4.2.6.4.1) to the OCS. In particular, this service is used to provide current status on prompt processing to the OCS, including both which images have been processed and an accompanying set of data quality metrics. The complete set of telemetry available is described in LSE-72[82].

5.4.2.6.3.2 Archive Base

The Archive Base enclave provides a single service: an endpoint for the Data Backbone. The Data Backbone is described in Section 5.4.2.6.6.

5.4.2.6.3.3 Commissioning Cluster

The Commissioning Cluster enclave provides a single service: an interactive computing environment for rapid turnaround of human-driven ad-hoc analysis of data during any recommissioning of the system. In addition, this enclave supports human-driven quality control activities undertaken in Chile.

The Commissioning Cluster runs an instance of the LSST Science Platform with low-latency access to the OODS (Section 5.4.2.6.3.1.4).

5.4.2.6.3.4 Chilean DAC

The Chilean DAC is responsible for providing science-user-facing data access and analysis services to the Chilean community. It will contain all raw data on disk, all file-based data release products, and databases subject to intense query (e.g., data release catalogs and the Engineering and Facility Database, Section 5.4.2.6.6). The primary user access and interface to the DAC is through an instance of the LSST Science Platform. The DAC will enforce authorization and authentication policies set by the Rubin Observatory for access to compute and data resources.

5.4.2.6.4 USDF

The USDF at SLAC acts as the primary processing and distribution site for all Rubin Observatory data products. This includes responsibility for prompt processing and alert distribution, 50% of data release processing (the remainder of the processing is carried out at the French DF at CC-IN2P3, as described in Section 5.4.2.6.5), archiving the results of processing, and making them available to the data rights community through a DAC. The USDF will also provide for user generated data-product production.

The USDF is configured with a high-reliability IT service model that ensures professional, secure, and consistent operation. Systems at the USDF have been designed to allow for dynamic reallocation of resources between Observatory operations, batch processing, and DAC subsystems in response to shifting peak demand.

The USDF will be operated by the Infrastructure and Support team within the Data Production Department, under the supervision of the USDF Lead. Generation of data products using DF infrastructure will be carried out by the Data Production Department's Execution team, under the leadership of the Lead Production Scientist. These groups will work closely with the Science Users Middleware and Algorithms and Pipelines teams, which will provide the execution middleware and the science payloads to generate data products, and with the Science Platform and Reliability Engineering team, which will develop and support the Science Platform. This Platform provides the primary mechanism by which Observatory staff and members of the data rights community will access and work with data stored in the facility.

[82] https://ls.st/lse-72

The USDF hosts four DMS enclaves.

5.4.2.6.4.1 Prompt USDF

The Prompt USDF enclave is responsible for compute-intensive processing for all near-real time operations and other operations closely tied with the Observatory. The services provided by this enclave feed back status information to the OCS through the Telemetry Gateway (Section 5.4.2.6.3.1.6); as such, the services form an important part of the overall Rubin Observatory system, and are generally expected to be available whenever observing is ongoing[83].

5.4.2.6.4.1.1 Prompt Processing

The Prompt Processing Service receives images and accompanying metadata from the Prompt Processing Ingest Service (Section 5.4.2.6.3.1.3), executes one of a number of possible algorithmic payloads upon them, and stores resulting data products to the Data Backbone (Section 5.4.2.6.6).

Algorithmic payloads include:

- Alert Production pipelines: science alerts are passed to the Alert Distribution Service (Section 5.4.2.6.4.1.2) for distribution. In this mode, data from two visits are processed simultaneously to meet the throughput and latency requirements resulting from the duration of each visit being 30 seconds, and alert distribution taking place within 60 seconds of visit completion.
- Calibration Product Production pipelines: these components operate on the data taken by the LSSTCam, as well as ancillary information, and are used to generate calibration products which include bias, flat, and dark frames. These products are used to calibrate the telescope during Alert Production processing.
- Exposures may be grouped and processed by specially configured pipelines, similar to what is used for Alert Production or Calibration Product Production, when deep-drilling fields are being observed.

5.4.2.6.4.1.2 Alert Distribution

The Alert Distribution Service receives alert packets generated by the Prompt Processing Service (Section 5.4.2.6.4.1.1) and distributes them to consumers. A number — at least five — of community-provided "brokers" will receive all alerts generated (expected to amount to approximately 10,000,000 per night). In addition, the Alert Distribution Service provides a limited capacity and capability Alert Filtering Service, which will allow individual members of the data rights community to receive alerts directly from the Observatory.

5.4.2.6.4.1.3 Prompt Quality Control

The Prompt Quality Control Service will monitor the execution of the Prompt Processing Service (Section 5.4.2.6.4.1.1) and will post-process data products written to the Data Backbone (Section 5.4.2.6.6) to generate additional measurements. Warnings are provided to the responsible Observing Specialists and DF Production Scientists when these measurements cross pre-defined thresholds.

5.4.2.6.4.2 Archive USDF

The Archive USDF enclave provides bulk storage for file and catalog data used by other enclaves within the USDF. It provides the following services.

5.4.2.6.4.2.1 Data Backbone Endpoint

The Archive USDF enclave provides an endpoint for the Data Backbone. The Data Backbone is described in Section 5.4.2.6.6.

[83] Prompt USDF systems may not be unavailable for more than 24 hours while observing is ongoing, and no more than one day of unplanned maintenance is expected per year.

5.4.2.6.4.2.2 Qserv Distributed Database

Over the course of the operational period, data release production will generate catalogs consisting of tens of trillions of rows and tens of petabytes of data. These will be made available for data rights holders to access and query through the Science Platform. Catalogs will be stored and queried using Qserv, a custom massively parallel relational database system (LDM-135[84]). Qserv makes intelligent use of replication, chunking, vertical partitioning, and shared scans to provide a system that is fault tolerant and capable of scaling to meet the variety and complexity of anticipated queries ranging from simple object lookups to complex full-sky correlations over billions of elements. Although Qserv is being developed directly to meet Rubin Observatory's needs, it builds upon the mature open-source technologies MariaDB[85] and XROOTD[86].

The Qserv instance is hosted in the Archive USDF and provides the primary source of truth for released data — this is where catalogs are ingested when data release processing is complete — and provides query capabilities to users of the US DAC (Section 5.4.2.6.4.4). Additional Qserv instances may be deployed locally by other DACs as required.

5.4.2.6.4.3 Offline Production USDF

The Offline Production USDF enclave is responsible for all long-period data processing carried out at the USDF. This includes the largest and most complex science payloads executed by the Rubin Observatory system, used to generate calibration products and data releases. This enclave also includes systems for monitoring and quality control of ongoing processing jobs, and for bulk distribution of data to partner institutions.

Services in the Offline Production USDF enclave are expected to manage long-running processing jobs — spanning weeks or months of real time — efficiently and reliably. They track the execution of millions or billions of individual tasks, and ensure output data is collected and saved to the Archive USDF enclave. Services are designed for autonomous operation where possible, although provisions are made for manual intervention where appropriate.

In general, low-latency processing is not critical in this enclave. However, because the enclave is sized to provide appropriate capacity for generating data releases on schedule, downtime must be minimized to ensure data are released on time.

The Offline Production USDF enclave provides the following services.

5.4.2.6.4.3.1 Batch Production

The Batch Production Service executes processing "campaigns" where each campaign consists of a given science pipeline, configuration, and set of inputs and outputs. Various different campaigns can be executed on different cadences, including:

- Periodic calibration product generation.
- Annual data release processing.
- Daily solar system processing.
- As needed alert catch-up processing and special programs processing.

More than one campaign can be executed simultaneously. In particular, data release processing will largely be scheduled as a single campaign that runs for much of the year, while other, smaller campaigns execute in parallel with it.

This service can coordinate with its peer at the IN2P3 Satellite Data Processing Center (Section 5.4.2.6.5; also, potentially, with other facilities) to enable split-site processing of data releases.

[84] https://ls.st/ldm-135
[85] https://mariadb.org
[86] http://xrootd.org

5.4.2.6.4.3.2 Offline Quality Control

The Prompt Quality Control Service will monitor the execution of the Batch Production Service (Section 5.4.2.6.4.3.1) and will post-process data products written to the Data Backbone (Section 5.4.2.6.6) to generate additional measurements. Warnings are provided to the responsible Observing Specialists and DF Production Scientists when these measurements cross pre-defined thresholds.

5.4.2.6.4.3.3 Bulk Distribution

- The Bulk Distribution Service is used to provide Prompt and Data Release products to major facilities and partners such as science collaborations. It extracts data products from the USDF Archive and transmits them over high-bandwidth connections to designated recipients. It is not available for direct access by science users, who will, instead, access data through the DAC (Section 5.4.2.6.4.4).

5.4.2.6.4.4 US DAC

The US DAC is responsible for providing science-user-facing data access and analysis services to the worldwide Rubin Observatory data rights community. It will contain all raw data on disk, all file-based data release products, and all databases. The primary user access and interface to the DAC is through an instance of the LSST Science Platform. The DAC will enforce authorization and authentication policies set by the Rubin Observatory for access to compute and data resources.

5.4.2.6.5 IN2P3 French DF

A memorandum of agreement (MOA, Agreement-51) has been established by IN2P3, LSSTC, Rubin Observatory, and the University of Illinois for the IN2P3 Computing Center (CC-IN2P3) in Lyon, France, to provide 50% of the computing capacity needed for annual data release processing during LSST operations, as estimated at the time of the MOA, and to host a complete backup of the Rubin Observatory data archive. CC-IN2P3 is a highly data-intensive supercomputer center with an established history of data management for HEP experiments, including the LHC at CERN, and in which large-scale astronomy experiments are a future strategic priority. The MOA requires that France will provide the TA network connectivity to the United States needed to support data transport to and from CC-IN2P3.

During the current construction phase, this collaboration is managed by a Joint Coordination Committee chaired by the NCSA. This committee is charged with preparing for combined processing, which will take place during survey operations.

In operations, Rubin envisions the French DF as an integrated part of the Data Production Department. Computing and storage will be managed locally at IN2P3, but an advisory council with membership from IN2P3 and the USDF will meet regularly with the Data Production AD to ensure smooth co-processing of data. This structure is illustrated in **Figure 8.** Further, staff members from IN2P3 will be integrated with the Data Production Work Breakdown Structure, providing excellent communication and awareness throughout the department.

The French DF hosts a single DMS enclave.

5.4.2.6.5.1 Offline Processing Satellite

The Offline Processing Satellite enclave provides a single service.

5.4.2.6.5.1.1 Batch Production

The Batch Production Service offered at the Satellite Data Processing Center offers a similar range of capabilities to that at the USDF (Section 5.4.2.6.4.3.1). It will be used to perform 50% of the annual data release processing.

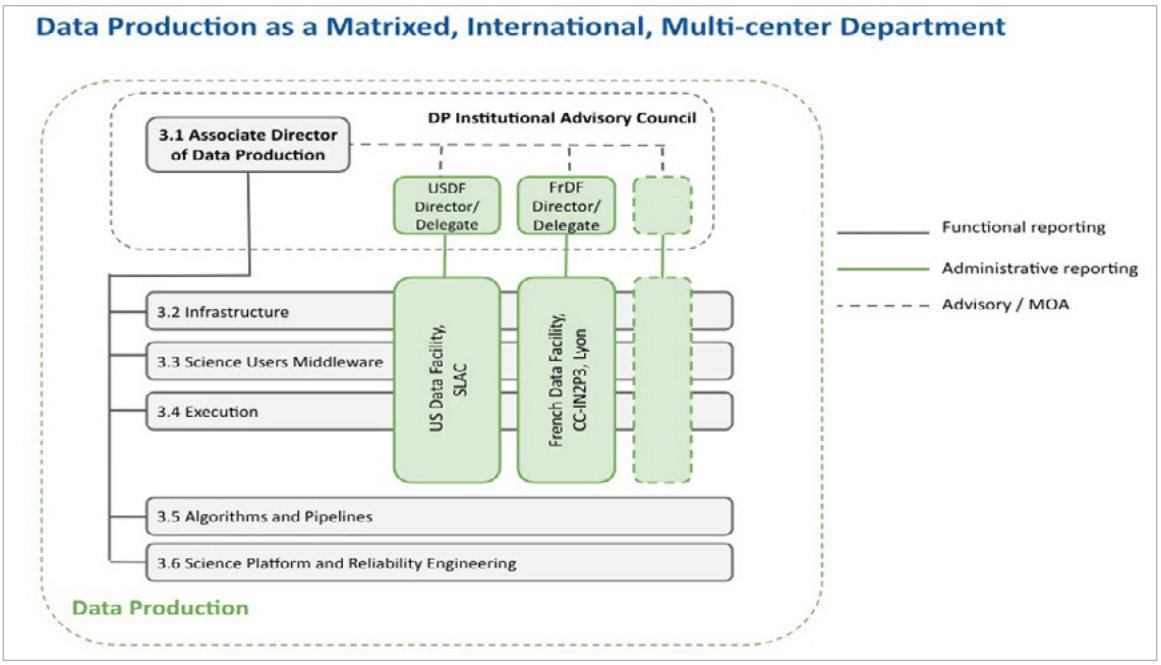


Figure 8: Data Production organization. The USDF and the French Satellite Data Processing Center are matrixed in the Data Production Department. An advisory council advises the Dual-Phase (DP) Associate Director.

5.4.2.6.6 Backbone Services

The Data Backbone is responsible for data storage, transport, and replication, including transferring files between the constituent facilities of the Rubin Observatory system. Specifically, the Data Backbone is responsible for:

1. Management of files and data resident in designated database management systems in Chile and at the USDF.
2. Enforcing data retention and redundancy policies set by Rubin Observatory for data products at all levels.
3. Providing disaster recovery capabilities for data.
4. Transporting data as needed between the observatory sites.
5. Meeting data service levels specified for all users (e.g., members of the scientific community) and other consumers of Rubin Observatory data services (e.g., alert brokers).

To meet these responsibilities, the Data Backbone provides policy-based replication of files (including images) and databases (including metadata about files as well as other miscellaneous databases, but not including the Qserv data release database) across multiple physical locations. It provides for caching of data at each endpoint, and automatic persistence of data to long-term archival storage when appropriate.

The Data Backbone has "endpoints," where data may be stored or retrieved, at the USDF (Section 5.4.2.6.4.2.1) and the Chilean Data Center (Section 5.4.2.6.3.2).

5.4.2.7 Network and Data Architecture

The network portions of the Rubin Architecture are excerpted here, after being originally described in the Rubin Observatory DMS and Facilities[87] documentation.

5.4.2.7.1 Network Requirements

The Rubin Observatory's wide-area networks (**Figure 9**) are responsible for transferring data and control information among the distributed observatory sites, and also provide connectivity to data centers for scientists and nonscientists alike. The requirements are therefore complex, being drawn not just from the Observatory System Specifications (LSE-30[88]) as flowed down through the DMS Requirements (LSE-61[89]), but also from MOAs between the Observatory and key partners. These requirements are summarized in the Network Design Document (LSE-78[90]) for convenience.

These requirements necessitate data transfers on a wide range of timescales, including:

- Order of milliseconds within the OCS and the control systems of each subsystem (the Telescope, Camera, and DMS).
- Order of seconds in acquiring and processing the raw image stream from the Camera through Data Production in order to create transient alerts.
- Order of days in archiving the raw image stream and other metadata for subsequent use in the production of astronomical catalogs and other data release products.
- Order of weeks to months in producing and deploying data necessary for calibration of the Telescope and Camera.
- Order of years in producing and deploying data releases of processed images and catalogs.

In addition, there are large geographical distances between the various sites, which imply non-trivial latency in certain data transfers. There are also data transfer requirements associated with supporting user access in DACs in Chile and the United States, and in the Education and Public Outreach Data Center (EDC)[91] in the United States. In the case of Chile, these requirements are derived from MOAs between Rubin Observatory and AURA, and between AURA and Chile that require providing a DAC in Chile in return for authorization to site Rubin Observatory in that country.

Considering continuity and criticality, Rubin has explicitly planned for certain levels of outages and failures, and has replaced or augmented existing systems to enhance reliability. For example, before the start of construction a microwave link was the only connection between the mountain and the base site. This did not meet the reliability requirements for Rubin Observatory operations, and so the project has installed fiber optic-based networks while retaining the microwave system as a backup. The system also has redundancy and diverse paths in every link from Santiago to the USDF at SLAC.

[87] https://docs.google.com/document/d/1YW2QMdrmE4MwjDHJw7v8eyYSTGAjfWHYSBXt4bK-nRc/edit?usp=sharing
[88] https://ls.st/lse-30
[89] https://ls.st/lse-61
[90] https://ls.st/lse-78
[91] https://www.lsst.org/content/education-public-outreach

5.4.2.7.2 Network Architecture and Design

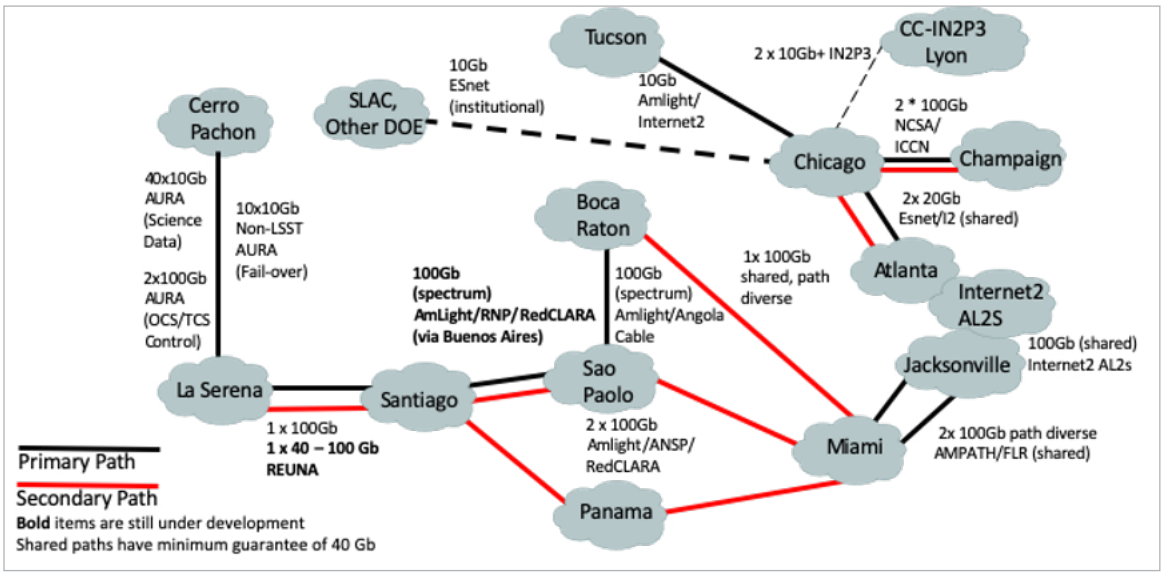


Figure 9: Rubin Observatory network paths and capacities. This diagram is consistent with the current Construction era DF being located at NCSA in Illinois. The Operations DF will be at SLAC and the detailed network path from Florida to SLAC will be worked out in 2021.

5.4.2.7.2.1 Summit to Base Link

The key driving requirements for the Summit to Base communications are the bandwidth and reliability required to transfer the image data and associated Engineering and Facility Database metadata for alert processing, to transfer the raw image data to the Base Center, and to handle OCS command and control traffic (see **Figure 10**).

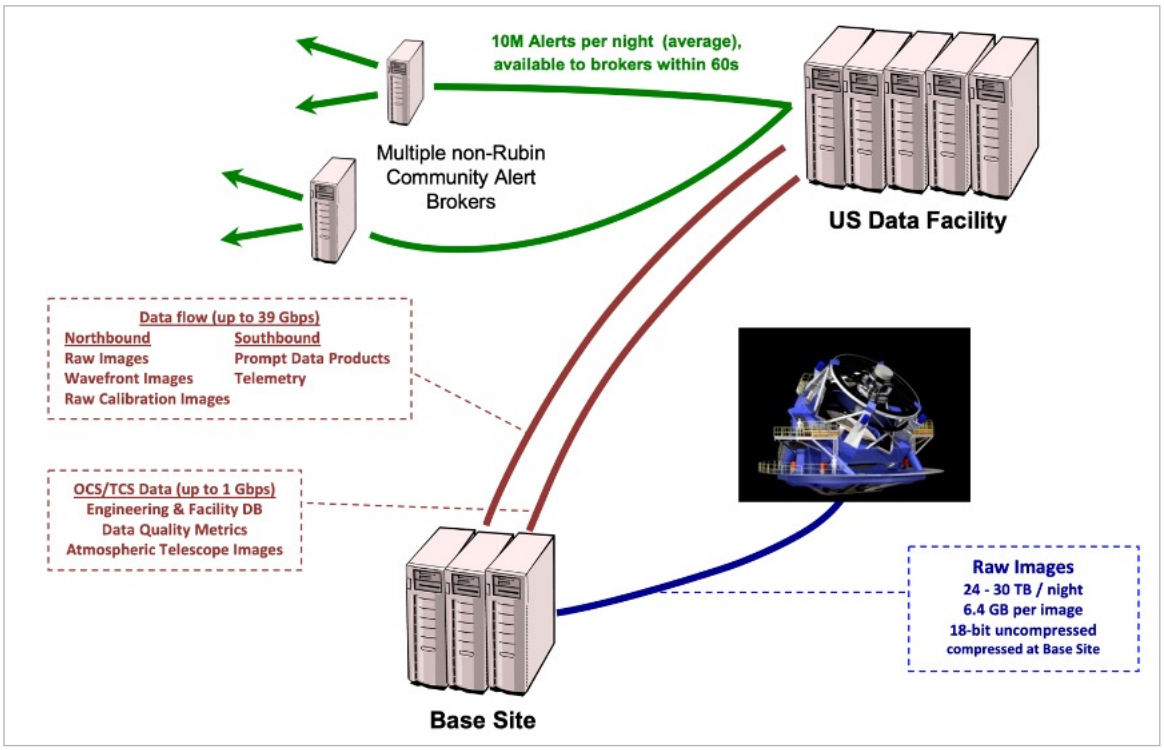


Figure 10: Nightly flow of data to support prompt processing

Rubin Observatory has installed an optical fiber pair between the Summit and Base Center utilizing dense wavelength division multiplexing technology to provide a minimum of two 100 Gbps lambdas with end nodes that will be owned and operated for Rubin Observatory by NOIRLab COS-ITO.

This capacity is divided into a minimum of 100 Gbps for image data transfer and 100 Gbps for OCS command and control data. A second pair of fibers carrying non-Rubin AURA traffic will be available for fail-over.

The historical data for fibers in north-central Chilean mountainous areas suggests a mean time between failures of ~two years and a mean time to repair of one week. The availability estimate on this segment is 98%. Therefore, the summit facility is capable of storing a week's worth of data to mitigate network downtime events between the summit and the base. During such downtime, the existing microwave backup link will be used which is sufficient for voice, email, video, and web traffic (but not for data transfer).

5.4.2.7.2.2 Base to USDF Link

The required peak bandwidth from the Base Center in La Serena to the USDF is established by the need to transfer the images for prompt processing (see **Figure 10**).

In order for the USDF to release alerts of transient and variable sources within 60 seconds of image readout from the camera electronics, seven seconds of this time budget are allocated to data transfer, including compression/decompression. Two seconds of that budget are in the Summit to Base transfer, and the other five seconds the Base to Archive Transfer and compression/decompression. This equates to a minimum bandwidth of 40 Gbps.

Current agreements between NOIRLab and other South American observatories have secured dual, POP- and path-diverse leased 100 Gbps links between Chile and the United States and an additional 100 Gbps spectrum link. The latter is expandable with purchased equipment upgrades at Rubin Observatory's discretion. The non-spectrum links are shared and have a guaranteed minimum for Rubin of 2 x 40 Gbps during observing and x 20 Gbps during non-observing times. These agreements have been achieved at the equivalent cost of a single 10 Gbps link if acquired via non-partner commercial leased services.

Networking between the Base and the USDF rests on provisioning initially provided by the Major Research Equipment and Facilities Construction project and is sustained and enhanced as needed by the operations project. This networking includes:

- One 100 Gbps lambda and one 40–100 Gbps lambda on physically distinct footprints from the Base Center to Santiago, Chile.
- One 100 Gbps lambda running up the west coast of South America, landing in a nuclear- and hurricane-hardened facility in Miami, Florida.
- One 100 Gbps lambda from Santiago up the eastern side of South America, landing in a nuclear- and hurricane-hardened facility in Miami, Florida.
- One 100 Gbps spectrum[92] from Santiago up the eastern side of South America, landing in Boca Raton, Florida.
- Networking provided by Florida LambdaRail to a demarcation point for national backbone research networks.
- Networking provided by the ESnet national research backbone network in the United States.
- Two 100 Gbps lambdas over physically distinct paths from Chicago to the USDF.

[92] "Spectrum" refers to the right to use a given frequency band on a fiber, using end equipment of one's choosing, such that the capacity of the link can be increased by the deployment of more capable equipment. This choice is up to the "owner" of the spectrum.

All the network links are being tested by Rubin Observatory and network partners at multiple levels during construction and commissioning, including installer tests supervised by Rubin and partners, demonstrations and end-to-end tests, verification tests, and commissioning tests.

This infrastructure provides general research connectivity to the United States for Rubin Observatory. The networking supports creation of bandwidth-protected channels that provide a guaranteed minimum rate with extra idle bandwidth available to the channel. Networking to the computing center in France is provided by the USDF to ESnet, and by France from the United States to Lyon.

5.4.2.7.2.3 Archive to DAC and Education and Public Outreach Center to User Links

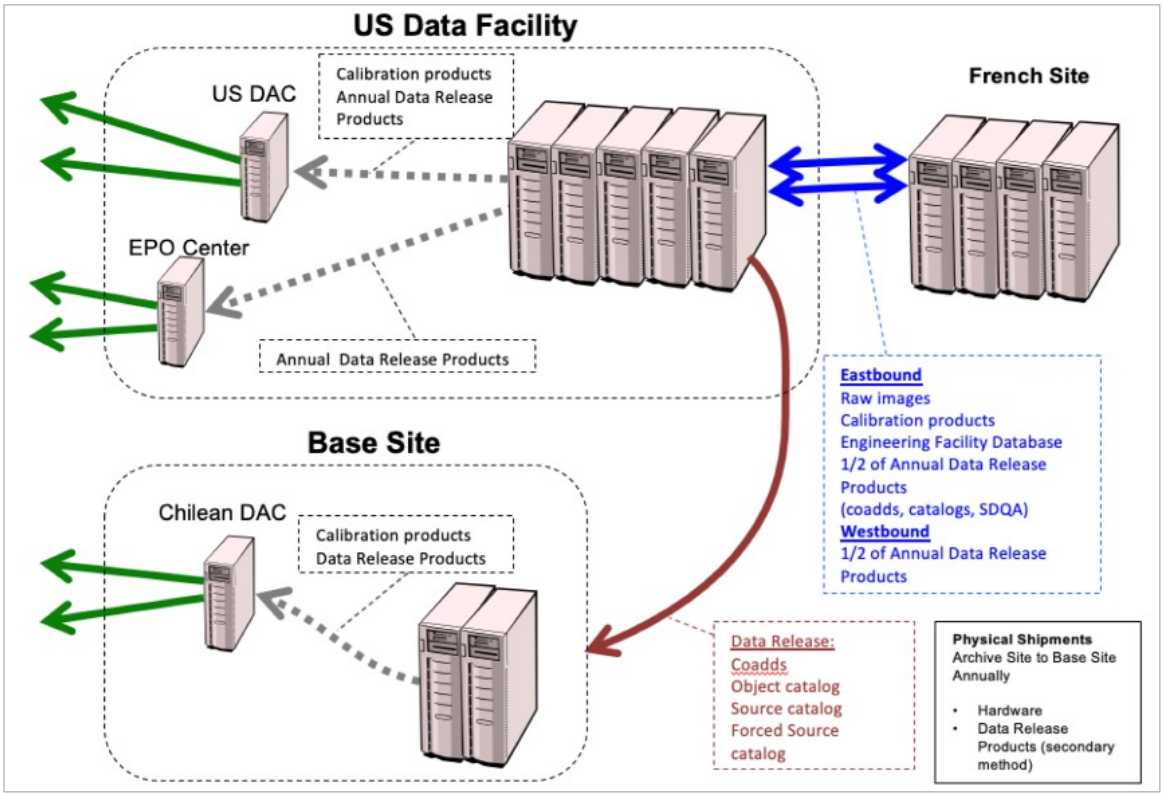


Figure 11: Data flows during processing and distribution of calibration and data release products

Rubin Observatory must provide data access to the US, Chilean, and international contributor scientific communities and to a worldwide community of non-specialist users whose curiosity may be triggered by Rubin Observatory education and public outreach (see **Figure 11**). User access to the DACs and Education and Public Outreach Center will be via public and R&E network connections (e.g., Internet and Internet2[93], XSEDE[94], ESnet[95]), and the aggregate bandwidth will be limited only by the connectivity of the hosting and using institutions. In cases of stand-alone DACs or science centers funded outside the project, the entity developing and operating the center will be responsible for providing network connectivity to Rubin Observatory USDF to enable data transfer.

Rubin Observatory network architecture will include identity management, cybersecurity management, and privilege limitations on usage of any resources (e.g., computing, storage, networks, software, and data) in a resource management layer on top of the network. Rubin Observatory network management will include network taps for passive traffic monitoring, firewalls, and other security mechanisms to enable this resource management.

[93] https://internet2.edu
[94] https://www.xsede.org
[95] http://www.es.net

5.4.2.8 Cloud Services

As described previously, in pre-operations Rubin Operations will work in the commercial cloud. A so-called IDF is being set up on the GCP in FY21. Rubin data production software and precursor data sets will be deployed on commercial cloud services to allow Rubin staff to develop processes and train for operations readiness. This includes serving users the precursor data sets to enable user familiarity with Rubin systems.

The IDF is used to train operations staff and the community in the development and use of LSST data. The IDF is a three-year planned program to deploy pre-operations data sets known as data previews. Data previews are early previews of what operations annual data releases will look like. They include user support and documentation. The first data preview is to be based on simulated data from DESC's DC2. Subsequent previews will include data obtained in commissioning and science verification.

5.4.2.9 Data-Related Resource Constraints

The choice of the USDF site was made known in October of 2020. Since the construction era processing facility is at the NCSA, and the Operations USDF will be at SLAC, SLAC will ensure that redundant, high-speed network connections to the nearest ESnet POP will be included in implementation of the USDF.

The IDF located in the GCP mentioned previously allows the ops team to continue to work toward operational readiness while the implementation design for the USDF at SLAC is in progress. The expected transition of Rubin systems from the cloud-based IDF to the SLAC USDF will begin in FY22.

5.4.2.10 Outstanding Issues

Final location of the USDF was made in October of 2020. An implementation plan will be made in 2021.

5.4.2.11 Case Study Contributors

Rubin Observatory Representation

- Jeff Kantor[96], Rubin Observatory
- Will O'Mullane[97], Rubin Observatory
- Bob Blum[98], Rubin Observatory
- Phil Marshall[99], SLAC
- Amanda Bauer[100], Rubin Observatory

ESnet Site Coordinator Committee Representation

- Mark Foster[101], SLAC
- Phil DeMar[102], Fermilab
- Andrey Bobyshev[103], Fermilab

[96] jkantor@lsst.org
[97] womullan@lsst.org
[98] rblum@lsst.org
[99] pjm@slac.stanford.edu
[100] abauer@lsst.org
[101] fosterm@slac.stanford.edu
[102] demar@fnal.gov
[103] bobyshev@fnal.gov

5.5 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 LBNL being the lead institution on the DOE side, and the University of Chicago leading for the NSF. When it is complete, there will be three large and 18 small telescopes deployed between two sites: the South Pole and Chilean Atacama Desert. Each site has a specific use case:

- The South Pole will specialize in drilling down on a single ~5% sky patch with large and small telescopes.
- The 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. There are therefore efforts to ensure that operation can proceed with limited (or severed) resources, with goals of increasing the available connections where possible.

5.5.1 Discussion Summary

The following discussion points were extracted from the case study and virtual meetings with the case study authors. These are presented as a summary of the entire case study, but do not represent the entire spectrum of challenges, opportunities, or solutions.

- 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.
- The underlying scientific ideas and use of technology are understood on a logical level, and physical instantiation will take several years to plan, execute, and complete. The remote nature of the observation sites will compound this.
- There are unique opportunities for the science 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 of the subsequent science analyses. How these science analyses will be supported is still to be determined, although it is expected to 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 on a daily basis, 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.
- CMB-S4's science output will be in the form of object catalogs and sky maps, similar to other surveys like LSST, together with a range of data products derived from them, including angular power spectra and cosmological parameter likelihood functions.
- Physical infrastructure at the two sites (South Pole and Chile) may be limited. Due to this limitation, 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.

5.5.2 CMB-S4 Case Study

5.5.2.1 Background

The CMB consists of the photons created in the Big Bang, propagated through the universe until detected today. As a consequence, the photons that make up the CMB trace the entire history of the universe and provide a unique window on fundamental physics and cosmology. In addition, CMB data sets are extraordinarily sensitive mm-wave surveys, incidentally supporting a wealth of astrophysics and astronomy. The full scope and scale of CMB science is laid out in the CMB-S4 Decadal Survey Report[104].

CMB-S4 is the "Stage 4" ground-based CMB experiment, bringing the entire US community together to build a single experiment capable of detecting the faintest CMB signals. The four design-driving science cases are:

1. To detect (or very tightly constrain) primordial gravitational waves to test theories of inflation.
2. To detect (or very tightly constrain) additional species of light relic particles.
3. To map the matter in the universe and elucidate the formation and evolution of galaxy clusters.
4. To detect mm-wave transients and add mm-wave data to multi-messenger astronomy.

Note that meeting these will necessarily also enable the full scope of other CMB science.

CMB-S4 is a joint DOE and NSF project, recommended by the Particle-Physics Project Prioritization Panel (P5)[105] in 2014 and the Astronomy and Astrophysics Advisory Committee [106] in 2017. It adds the capacities and capabilities of ANL, Fermilab, LBNL, and SLAC under the DOE Office of HEP to the longstanding NSF university program spanning the Division of Astronomical Sciences (MPS/AST) and the Division of Physics (MPS/PHY) in the Directorate of Mathematical and Physical Sciences and the Office of Polar Programs in the Directorate of Geosciences. Other partners include NASA's Jet Propulsion Laboratory and Goddard Space Flight Center (GSFC), the Department of Commerce's National Institute of Standards and Technology, Associated Universities Incorporated, and the Smithsonian Astrophysical Observatory. Collaborative discussions are

[104] https://arxiv.org/abs/1907.04473
[105] https://www.usparticlephysics.org
[106] https://www.nsf.gov/events/event_summ.jsp?cntn_id=299907&org=NSF

underway with the Cerro Chajnantor Atacama Telescope (CCAP-prime) and the Simons Observatory and its supporting Simons and Heising-Simons foundations. CMB-S4 also anticipates international partners and in-kind contributions, especially from Australia, Canada, Europe, and Japan.

Here and throughout the document, the current project plan, to be presented to DOE Critical Decision 1 and NSF Preliminary Design Review in spring 2021, is described. While some details may evolve, CMB-S4 does not expect significant changes. Note that while CMB-S4 anticipate their availability, none of the off-project network and computing resources that are planned for use have yet been secured as of November 2020.

To meet the CMB-S4 science goals (and all the other science that this will be enable) three coordinated surveys will be conducted from 2028 to 2035:

1. A wide/deep survey of ~70% of the sky from the Chilean Atacama Desert with O(270,000) detectors spanning 6 mm-wave frequencies on 2 x 6m telescopes with a daily cadence.
2. An ultra-deep survey of ~3% of the sky from the South Pole with O(150,000) detectors spanning 8 mm-wave frequencies on 18 x 0.5m telescopes.
3. An ultra-deep survey of the same ~3% of the sky from the South Pole with O(120,000) detectors spanning 7 mm-wave frequencies on 1 x 5m telescope.

In operations, CMB-S4 will gather 1.2 x 108 samples per second in Chile and 6.6 x 107 samples per second at the South Pole, corresponding to compressed data rates of 1.3 and 0.74 Gbps, respectively. Chilean data will be transferred in real time over fiber networks to the primary data center at the DOE's NERSC and copied to the secondary data center at the DOE's ALCF. A small fraction of the South Pole data will be transferred daily over the Tracking and Data Relay Satellite System (TDRSS)[107] to the White Sands Test Facility[108], and then over fiber networks to NERSC and ALCF. Given the limited bandwidth along this path, the bulk of the data will be stored locally and shipped to the United States in the austral summer. All data are archived at both NERSC and ALCF.

Each day's data are immediately reduced to a map of the observed sky, using US agency computing resources (DOE HPC, NSF HTC, and potentially FABRIC[109]) for the Chilean data and using dedicated on-site project computing resources at the South Pole. These maps are analyzed to identify transient sources, with alerts issued to the entire scientific community via the Community Alert Broker, and for data quality and telescope health checks.

The bulk data from both sites are analyzed on US agency resources to:

1. Characterize the instruments (e.g., determine their exact pointing, measure their beams and band-passes, etc.).
2. Identify systematic effects in and develop software mitigations sufficient for their residuals to be less than the tiny CMB signals.
3. Reduce them to maps at each observing frequency for a multitude of different data cuts over time and/or detectors, together with a statistical characterization of each map (e.g., a noise correlation matrix or a MC suite of simulated raw data sets reduced to equivalent sets of maps).

These well-characterized frequency maps are then passed to the various collaboration analysis working groups for the full range of scientific analyses.

Given the size and complexity of the CMB-S4 data set (~100 PB), collecting, transporting, and reducing the raw data from timestreams of detector samples to well-characterized sky maps is the responsibility of the

[107] https://www.nasa.gov/directorates/heo/scan/services/networks/tdrs_main
[108] https://www.nasa.gov/centers/wstf/index_new.html
[109] https://fabric-testbed.net

CMB-S4 project. Given the extraordinary range of science these maps support and their relatively small size (~100 TB),analyzing these maps is the responsibility of the CMB-S4 collaboration and the scientific community at large.

5.5.2.2 Collaborators

The CMB-S4 collaboration currently consists of 236 members at 93 institutions in 14 countries and 21 US states. The institutions include DOE, NASA, and Department of Commerce laboratories, as well as a host of universities.

Currently all data are simulated, and the primary site for creation, storage, sharing, and analysis of these is NERSC, although this will expand to ALCF and OSG/XSEDE as the construction project proceeds.

In operations that raw data will be archived at NERSC and ALCF; kept spinning at NERSC and ALCF; reduced to daily maps at NERSC, ALCF, OSG/XSEDE and possibly FABRIC (Chilean data) or on dedicated on-site resources (South Pole); and bulk-reduced at NERSC, ALCF, and OSG/XSEDE. Simulated data will be generated at NERSC and ALCF and typically reduced to maps on the fly to avoid IO and storage overheads. In both cases, the resulting maps will be archived at NERSC and ALCF; kept spinning at NERSC; and analyzed at NERSC and on collaboration members' local resources.

Published data (primarily maps and derived scientific data) will be made available to the scientific community at NERSC and Legacy Archive for Microwave Background Data Analysis (LAMBDA)[110] located at the NASA GSFC[111]. The raw data and the software used to characterize and reduce them will also be made available to the scientific community at NERSC.

User/collaborator and location	Primary or secondary copy of the data	Data access method	Avg. size of data set	Frequency of data transfer or download	Are data sent back to the source and method	Any known issues with data sharing
TELESCOPE SITE: SOUTH POLE	Primary	Data transfer / portable hard drive	200 GB–8TB (daily), 3 PB (annual)	Daily (burst), annually (physical shipment)	No	Insufficient satellite network capacity
TELESCOPE SITE: CHILEAN ATACAMA	Primary	Data transfer	14 TB	Daily (continuous)	No	N/A
PRIMARY DATA CENTER: NERSC AT LBNL	Primary	Data portal, data transfer, portable hard drive (receive/load)	14.2 TB (daily), 3 PB (annually)	Daily and annually	No	N/A
SECONDARY DATA CENTER: ALCF AT ANL	Secondary	Data transfer	14.2 TB (daily), 3 PB (annually)	Daily and annually	Data transfer	N/A
GRID COMPUTING: OSG/XSEDE (DISTRIBUTED)	Secondary (transient)	Data transfer	14.2 TB (daily), 3PB (annually)	Daily and annually	Data transfer	To be evaluated
ARCHIVE: LAMBDA AT GSFC	Tertiary	Data transfer	600 TB	Single use	No	To be evaluated
COLLABORATION MEMBERS (SEE ABOVE)	Tertiary	Data portal, data transfer	< 600 TB	Intermittent	Data transfer	To be evaluated

Table 6: CMB-S4 data projections

[110] https://lambda.gsfc.nasa.gov
[111] https://www.nasa.gov/goddard

5.5.2.3 Instruments and Facilities

5.5.2.3.1 South Pole Site

Present/Next

South Pole Observatory pathfinder experiment with ~10% of the CMB-S4 data volume.

Beyond

1 x 5m + 18 x 0.5m telescopes to be deployed starting in 2026:

- Gathering 3 PB/year.
- File size/count/distribution TBD.

On-site computing resources to be deployed in 2025/26 sufficient to:

- Store one year of data (3 PB).
- Generate daily maps and identify transients (Tera-Scale).

Potential significant agency/international upgrade to data path, either fiber or satellite to McMurdo Station and beyond — unknown timescale.

5.5.2.3.2 Chilean Atacama Site

Present/Next

Simons Observatory pathfinder experiment with ~10% of the CMB-S4 data volume.

10 Gbps dedicated fiber link from the site to the Atacama Large Millimeter Array (ALMA) being installed by Simons Observatory now, with ESnet support.

Beyond

2 x 6m telescopes to be deployed starting in 2026:

- Gathering 5 PB/year.
- File size/count/distribution TBD.

On-site computing resources to be deployed in 2025/26 sufficient to:

- Store one month of data (420 TB)

5.5.2.3.3 NERSC/ALCF/OSG/XSEDE

Present/Next

Sufficient cycles and storage for experiment design studies, data management subsystem development, data analysis pipeline validation, and verification.

NERSC disk and data portal for data distribution to the collaboration (Tera-Scale).

The exact distribution of cycle/storage resources between these systems will depend on the agencies and their allocations. The overall driver in the construction project will be a series of four data challenges intended to:

- Validate and verify the experiment design at the degree needed at each DOE/NSF review gate.
- Validate and verify the project data management and collaboration data analysis pipelines.
- Demonstrate scaling the data volume the simulation/reduction tools can process to achieve 100% by 2026.

- Demonstrate porting and sufficient optimization of the data management software stack to each new generation of HPC and HTC architecture deployed by the DOE/NSF.

An estimate of the resources required for each challenge and the target HPC system (to be augmented by HTC resources) is in the following table.

Challenge	DC1	DC2	DC3	DC4
YEAR	2021	2022	2023	2026
DATA FRACTION	12.5%	25%	50%	100%
SAMPLES	9.59E+14	1.92E+15	3.84E+15	7.67E+15
TOTAL CYCLES (EFLOP)	3.0	6.0	12.0	24.0
PEAK MEMORY (PB)	2.5	4.9	9.9	19.8
PEAK SCRATCH (PB)	1.3	2.7	5.4	10.7
SYSTEM	CORI	PERLMUTTER	AURORA	NERSC-10
PEAK CYCLES (PFLOP/S)	30	150	1000	1000
TOTAL MEMORY (PB)	1	3	5	5
TOTAL SCRATCH (PB)	30	150	200	200
FULL-SYSTEM RUNTIME (HRS)	27.79	11.12	3.33	6.67
PEAK MEMORY FRACTION	247%	165%	198%	395%
PEAK SCRATCH FRACTION	4%	2%	3%	5%

Table 7: CMB-S4 computational challenge summary

Note that CMB-S4 exceeds the peak memory available in all cases, limiting us to annual maps (1/7 of the requirement) rather than the full data maps at this stage.

Beyond

Sufficient cycles and storage for operations data management and data analysis.

NERSC disk and data portal for data distribution to the collaboration (Tera-Scale).

In operations, CMB-S4 anticipates a 10–1000x increase in the integrated computing requirements over DC4, progressively growing between 2027 and 2035, but no increase in the peak requirements. These requirements — which will include a mixture of very many small analyses and a small number of very large analyses — will be distributed across all of the available systems, although CMB-S4 anticipate that the small analyses will primarily be run on HTC, and the large on HPC, systems.

Along with the raw data files listed under the sites, the primary data products will be the frequency maps:

- Daily Chile (7 years x 365 days): 140 TB in 15,000 files.
- Daily Pole (7 years x 365 days): 10 TB in 36,000 files.
- MC Chile+Pole (1,000 realizations x 7 years): 450 TB in 150,000 files.

The full CMB-S4 computational requirements were last enumerated in September 2020[112].

[112] https://docs.google.com/spreadsheets/d/1I3SccH3rVEaT2DzQtY4bscxU73gqV9rwtGPZ4YQb8XM/edit#gid=1066998237

5.5.2.4 Process of Science

Present/Next

During the construction phase, all of the significant data will be simulated (although small amounts of laboratory data will be generated, primarily at the national laboratories, some of which may be transferred to NERSC).

Simulations of raw data will be run at NERSC or ALCF and the data copied between them, peaking at O(10) PB in 2026. Reductions to maps will be run at NERSC, ALCF, OSG/XSEDE, and possibly FABRIC. Maps will be made available to the collaboration at NERSC, with the bulk of the analyses performed there. Small amounts of data may be moved to collaborators' local resources for analysis.

Beyond

During the operations phase, Chile data and some South Pole data will be transferred to NERSC and ALCF over the network, with the exact mechanics (including failsafe processes for system downtimes) still to be determined. Bulk South Pole data will be stored on disk and shipped to NERSC annually.

Daily reductions to maps will be run at NERSC, ALCF, OSG/XSEDE, and possibly FABRIC for the Chilean data, and using dedicated on-site hardware at the South Pole. Bulk-data reductions to maps will be run at NERSC, ALCF, and OSG/XSEDE.

MC simulations will be used for uncertainty quantification and debiasing of the instrument data. Such simulations will be generated at NERSC and ALCF and reduced to maps on the fly. The maps will be archived at both sites and made available to the collaboration at NERSC.

5.5.2.5 Remote Science Activities

Both of the observing sites (Chilean Atacama, South Pole) are remote.

Present/Next

Precursor experiments at both sites (Simons Observatory in Chile, South Pole Observatory at the South Pole) will exercise many of the systems CMB-S4 plans to use, albeit at ~10% of the data volume.

Beyond

In operations, the Chilean telescopes will generate 1.3 Gbps, and will be connected with a robust 10 Gbps network to the United States. The South Pole telescopes will generate 0.75 Gbps, but current bandwidth only supports O(0.06) Gbps to the United States. Current plans are to transfer the South Pole data annually on disk, but a significant increase in the available bandwidth would reduce the dependence on-site computing and increase the robustness and speed of the analyses of those data.

5.5.2.6 Software Infrastructure

5.5.2.6.1 Data Movement

Present/Next

During construction, occasional transfers of simulated data between NERSC and ALCF, OSG/XSEDE (TBD); access to data products by collaboration through a NERSC data portal.

Beyond

During operations, continuous network transfer of data from Chile to NERSC/ALCF is possible; daily network transfer of data from South Pole to NERSC/ALCF (TBD); annual disk transfer of data from South Pole to NERSC and then to ALCF is possible; occasional transfer of data from NERSC to OSG/XSEDE is possible; access to data products by collaboration through a NERSC data portal.

Overall, the data-movement mechanisms/software are yet to be defined; CMB-S4 will be working with ESnet to do this.

5.5.2.6.2 Data Processing

Present/Next/Beyond

TOAST (HPC)[113] and SPT-3G (HTC)[114] open-source CMB data processing frameworks, with ongoing work on their interoperability.

5.5.2.7 Network and Data Architecture

CMB-S4 network planning is in the nascent stages, and does not have a full picture of what capabilities may exist when the project starts. The following diagram is thus used for planning purposes, utilizing current technology offerings:

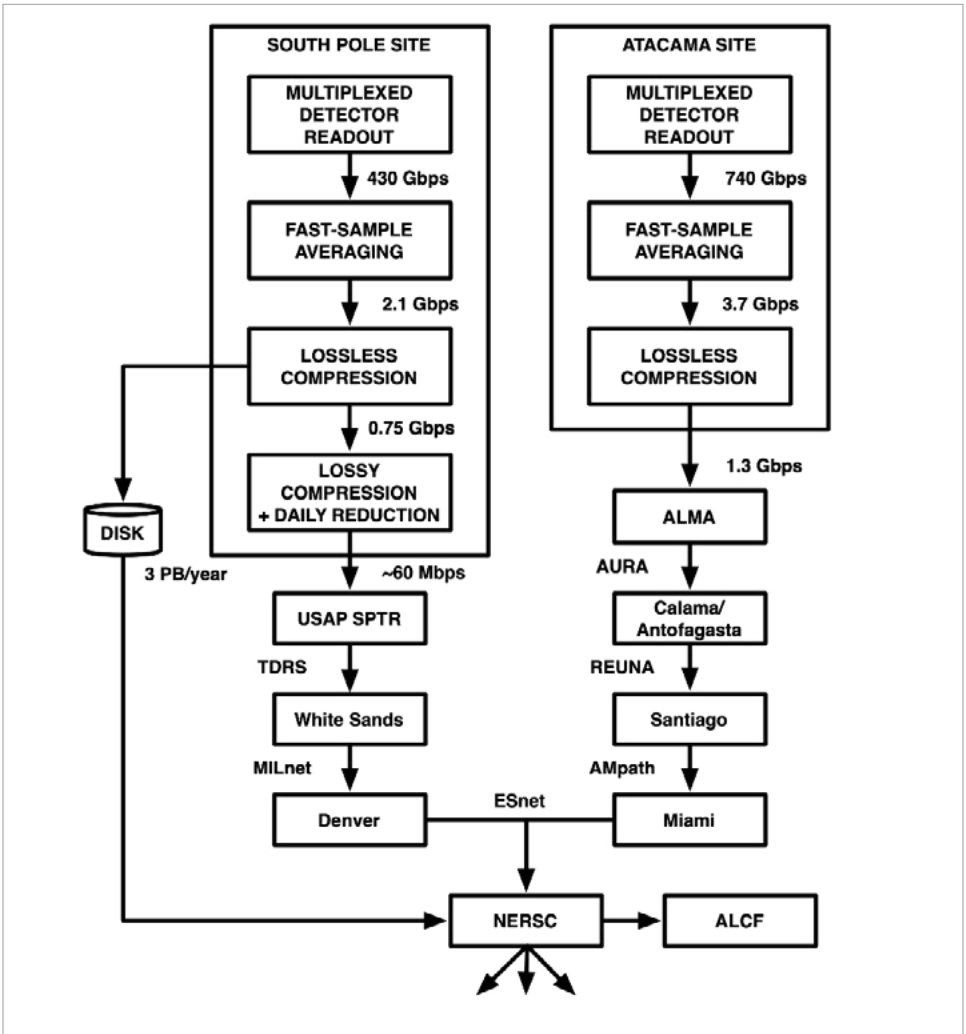


Figure 12: CMB-S4 logical network map

[113] https://www.nersc.gov/news-publications/nersc-news/science-news/2017/a-toast-for-next-generation-cmb-experiments/
[114] https://astro.fnal.gov/science/cmbr/spt-3g/

5.5.2.8 Cloud Services

Cloud services are not anticipated at this time.

5.5.2.9 Data-Related Resource Constraints

The major networking constraint is bandwidth from the South Pole, where the total capacity is less than 10% of the requirement and must be shared by all the experiments located there. TDRSS is theoretically capable of supporting 800 Mbps of capacity, but this is highly affected by conditions and concurrent use cases. A very significant increase in bandwidth would be a huge benefit to CMB-S4 either via other satellite options or terrestrial network cables.

The biggest overall data-related challenge is the capacity and capability of HPC and HTC systems, especially in their total system memory and their available cycles and the efficiency with which CMB-S4 can use them (especially in the post Moore's law epoch of energy-constrained computing).

5.5.2.10 Outstanding Issues

There are no outstanding issues to report at this time beyond continuing to work with ESnet on wide-area/international bandwidth strategies that will become a factor in the future.

5.5.2.11 Case Study Contributors

CMB-S4 Representation

- Julian Borrill (Co-Spokesperson)[115], LBNL and University of California, Berkeley
- John Carlstrom (Co-Spokesperson)[116], University of Chicago and ANL
- Jim Yeck (Former Project Director)[117], University of Wisconsin, Madison
- John Corlett (Interim Project Director)[118], LBNL
- Gil Gilchriese (Interim Deputy Project Director)[119], LBNL

ESnet Site Coordinator Committee Representation

- Rune Stromsness[120], LBNL
- Richard Simon[121], LBNL
- Damian Hazen[122], LBNL and NERSC
- Tavia Stone Gibbins[123], LBNL and NERSC

5.6 LZ Dark Matter Experiment

The LZ Dark Matter Experiment is located at SURF in South Dakota, and is managed primarily by LBNL. The scientific focus is on dark matter direct detection through the use of DAQs deployed within SURF, with analysis being performed at NERSC after the data are streamed. The experiment has a long five-year runtime (i.e., it does

[115] jdborrill@lbl.gov
[116] jc@kicp.uchicago.edu
[117] jhyeck@gmail.com
[118] jncorlett@lbl.gov
[119] mggilchriese@lbl.gov
[120] rstrom@lbl.gov
[121] rsimon@lbl.gov
[122] dhazen@lbl.gov
[123] tavia@lbl.gov

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 of LZ's technology workflow has been developed and deployed using container technology (CVMFS), which gives a layer of protection and redundancy to cope with resource constraints that may exist at NERSC due to maintenance.

5.6.1 Discussion Summary

The following discussion points were extracted from the case study and virtual meetings with the case study authors. These are presented as a summary of the entire case study, but do not represent the entire spectrum of challenges, opportunities, or solutions.

- LZ will explore dark matter through the use of a detector that is 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 are expected to begin in the autumn of 2020, and will operate in stable and continuous condition for five years (i.e., 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 approximately 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.

5.6.2 LZ Dark Matter Experiment Case Study

5.6.2.1 Background

LZ is a next-generation dark matter experiment. LZ has been selected by the DOE and the NSF as one of the three "G2" (for Generation 2) dark matter experiments. In the spring of 2015, LZ passed the "Critical Decision Step 1" or CD-1 review, and became an official DOE project. LZ is an experiment funded and operated under DOE's HEP and an important experiment in the Cosmic Frontier program.

The LZ detector consists of a DP Time Projection Chamber (TPC) filled with liquid xenon (LXe) and instrumented by a top and a bottom array of photomultipliers tubes (PMT). When a dark matter particle interacts with the target, it produces a xenon Nuclear Recoil which excites and ionizes the medium. The excitation results in the emission of scintillation photons, which form the so-called S1 prompt signal. The ionization electrons are

drifted toward the liquid phase thanks to the application of an electric field. Once the electrons reach up the top of the liquid phase, they are amplified in the gas phase of the TPC with a higher electric field. This results in the emission of photo-ionization photons, which constitute the late so-called S2 signal. A typical dark matter signal is then composed by one S1 and one S2 signal. The light signal collected by the PMTs is then converted into an electrical signal, which is digitized by the DAQ. The DAQ records the waveforms of all the PMTs and assembles the waveforms into events.

The LZ experiment is currently in its final construction phase in the Davis cavern of the SURF underground laboratory. The inner core of the detector, a Time Projection Chamber, is fully assembled and has been installed underground at SURF.

LZ will acquire data at SURF, then transfer data to NERSC. Once at NERSC, data will be transferred to the UK data center. Then the NERSC copy will be processed at NERSC to produce reduced quantities files. US users are going to access the processed data stored at NERSC essentially using NERSC computing resources.

The archived data and the data on spinning disks will have to be kept for 5–10 years, depending on the operations schedule. The processed data will be kept on spinning disks until superseded by a new version of the processing software, at which point it will be archived to tape.

5.6.2.2 Collaborators

LZ is a collaboration mainly spread between many US facilities and the UK. The LZ data sets are going to be accessible to users exclusively into two data centers: the United Kingdom Data Center (UKDC) (GridPP[124] resources, a majority existing at Imperial College London) and NERSC. At NERSC, the LZ collaboration has 180 active users from US and UK institutions.

The estimated data volume is 1 PB/year including calibrations data.

User/collaborator and location	Primary or secondary copy of the data	Data access method	Avg. size of data set	Frequency of data transfer or download	Are data sent back to the source and method	Any known issues with data sharing
SURF IN LEAD, SD	Temporary storage only		~1 PB/year, 1-2 GB/file	Copy file to the surface lab after creation	No	N/A
SURF SURFACE FACILITY (SURF) IN LEAD, SD	Temporary storage only		~1 PB/year, 1–2 GB/file	Receive data from under-ground lab and transfer to NERSC	No	N/A
NERSC, BERKELEY CA	Primary and secondary	Login/DTN node, tape archive	~1 PB/year, 1–2 GB/file	When any data file is ready to transfer	No	N/A
GRIDPP, UK	Secondary	XROOTD, batch system	~1 PB/year, 1–2 GB/file	When any data file is ready to transfer	No	N/A

Table 8: LZ Dark Matter data projections

[124] https://www.gridpp.ac.uk

5.6.2.3 Instruments and Facilities

The LZ detector is located in the Davis cavern of the SURF underground laboratory. The detector consists of a DP liquid xenon TPC equipped with an optical readout. The optical readout consists of a top and a bottom arrays of 264 photomultipliers. The detector is currently almost fully commissioned and the project is expecting to start data taking in September to October 2020. The intention is to operate the detector in stable conditions, with no hardware modification that would affect the data flow.

The data files are built on the underground lab and then transfer to the surface lab. The hardware responsible for the data building is:

- 15 data collectors (DC) with partial events:
  - 1 TB solid-state drive (SSD) data disk/machine + 3.5 TB HDD/machine.
  - DCs are connected via 10 GB network to Event Builders.
- 5 event builders:
  - 14 TB/machine to store root-based event files.
  - 70 TB of total disk space is sufficient for 16 days of Kr data or 37 hours of Deuterium-Deuterium (D-D) data.

On the surface lab, there exists 260 TB of Redundant Array of Independent Disks (RAID) space for temporary data storage. This disk space allows 93 days of buffer. Once the data are on the RAID disk, the data are transferred to NERSC using SPADE.

5.6.2.4 Process of Science

Once data are transferred to NERSC and the transfer has been validated, the data files are removed from the surface lab disk. LZ will not rely on SURF -> NERSC data transfer for the data analysis: the analysis will happen at NERSC and UKDC.

Data are going to be processed both at NERSC and UKDC to extract reduced quantities by using the collaboration software LZap. This will produce .RQ files. During this data processing, LZ will extract from the PMTs waveforms various quantities that are going to be useful for the analysis, such as pulse area, pulse start, and pulse width. After the data processing, the analyzers are going to apply a set of cuts to select only the relevant data for dark matter searches. In parallel, LZ will generate at NERSC a complete background model by using simulation software. After that, LZ will compare the reduced data with simulated background models, with a profile likelihood ratio test statistic. LZ anticipates that the analysis tools will remain roughly stable over the lifetime of the experiment.

5.6.2.5 Remote Science Activities

The main network connections are going to be between:

- SURF underground → SURF surface lab.
- SURF surface lab → NERSC or SURF surface lab → UKDC (as backup).
- NERSC → UKDC.
- Users → NERSC.

In addition to that, LZ has some database replication that will require network usage.

5.6.2.6 Software Infrastructure

To manage the data transfer between SURF, NERSC, and UKDC, LZ is using a tool named SPADE[125], an open-source tool developed for IceCube and also used on Dayabay[126]. It uses the following transfer protocols: scp[127], bbcp[128], gridFTP[129], and Globus[130].

The data transfer at NERSC will trigger the processing of the data using LZap. The file transfer is confirmed by using a checksum.

For the data analysis, the uniformization of the software is ensured thanks to software distribution through the CERN Virtual File System (CVMFS)[131] and containerization technologies.

The data analysis is also relying on databases that are replicated at NERSC and at the UKDC using standard MySQL tools.

5.6.2.7 Network and Data Architecture

The SURF and LZ network infrastructure can be summarized as follows:

- Direct connection of LZ Online underground (-4850ft) and surface:
  - We own and manage the connection endpoints (using existing dark fiber provided by SURF).
  - Single mode fibers, in 10km range for Long Reach (LR) optics.
  - Fully redundant 4 x 10Gbit point-to-point connections using bidirectional 10GbE LR transceivers:
    - Two strands per shaft (Ross, Yates).
    - Up and running since December 2018 no unscheduled downtime.
    - No PLC dedicated emergency access (sufficient redundancy in main network).
    - Redundant pairs of Netgear m4300 24x24F Layer 3 switches both at surface and underground.
      - Very good experience with these switches.
- Fully redundant 4 x 10Gbit point-to-point connections to SURF IT core switches.
- Infrastructure and application servers (surface and underground) have redundant connections to their respective switch stacks.
- External network connectivity:
  - 10 Gbit/s via REED / Internet 2.
  - 1 Gbit/s via commercial regional internet service provider (ISP) (backup and "non-research" access).
    - Not intended for LZ bulk-data transfer.
  - Tertiary backup via commercial regional ISP:
    - Not usable for LZ bulk-data transfer.

[125] http://nest.lbl.gov/projects/spade/html/index.html
[126] https://dayabay.lbl.gov
[127] https://www.ssh.com/ssh/scp/
[128] https://www.nics.tennessee.edu/computing-resources/data-transfer/bbcp
[129] https://gridcf.org
[130] https://www.globus.org
[131] https://cernvm.cern.ch/fs/

- Bulk (DAQ) data transfer to NERSC:
  - Average data production rate well below 3Gbit/s but the systems are capable of >10 GBit/s peak (two storage servers at the surface).
- The overall performance of the network is done by using perfSONAR. A server commissioning has been delayed at the UKDC because of COVID-19.

5.6.2.8 Cloud Services

We have no plan to use cloud computing services for data analysis.

5.6.2.9 Data-Related Resource Constraints

We do not have any current or anticipated constraints that would affect productivity. Network loss events to SURF are a factor, but can be managed due to the on-site technical capabilities.

5.6.2.10 Outstanding Issues

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. LZ is not the greatest user of site bandwidth at SURF, but relies on stable and robust connectivity. Having alternative paths established (to ensure connectivity), along with ESnet peering capability, will ensure that the critical link between SURF and NERSC is maintained during LZ operation.

5.6.2.11 Case Study Contributors

LZ Representation

- Simon Fiorucci[132], LBNL
- Quentin Riffard[133], LBNL
- Maria Elena Monzani[134], SLAC

ESnet Site Coordinator Committee Representation

- Rune Stromsness[135], LBNL
- Richard Simon[136], LBNL
- Damian Hazen[137], LBNL and NERSC
- Tavia Stone Gibbins[138], LBNL and NERSC

5.7 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. Simple 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 (SM).

[132] sfiorucci@lbl.gov
[133] qriffard@lbl.gov
[134] monzani@slac.stanford.edu
[135] rstrom@lbl.gov
[136] rsimon@lbl.gov
[137] dhazen@lbl.gov
[138] tavia@lbl.gov

Muon g-2 is currently operating at Fermilab and has finished Run 3 of a planned five runs (with expected end time in 2022). 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 AI/ML, both of which may influence future operations for Mu2e.

Mu2e is under construction, and will go into operation in 2024 with a five-year run cycle. It is expected that it 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. The majority of cycles will be provided by Fermilab, with some use allocated to other participating sites (a minority of the expected computation and storage power).

The use of HPC resources is not currently large, although the workloads would convert to the use case if there were resources to convert and adapt software (at current time, this is not a high priority).

5.7.1 Discussion Summary

The following discussion points were extracted from the case study and virtual meetings with the case study authors. These are presented as a summary of the entire case study, but do not represent the entire spectrum of challenges, opportunities, or solutions.

- Muon G minus two (g-2) and the Muon-to-electron-conversion experiment (Mu2e) are two experiments at Fermilab involving the study of muon particles. These are distinct, but share some common components.
- The g-2 experiment started in 2018, and is currently operating (expected through 2022). Mu2e is under construction (with the first beam scheduled for 2022/2023, 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 g-2 experiment relies on a detector and two DAQs that capture events. Data are captured and then assembled on-site. Analysis can be performed by the Fermilab team, or individuals that are collaborating via the distributed grid/software infrastructure.
- Additional runs of g-2 are possible if the start of Mu2e is delayed, as the experiments will overlap in terms of resources utilized.
- The 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 Mu2e experiment will rely on a series of sensors and a DAQ that will stream raw data to computing and storage systems at Fermilab. The raw data will go through a processing step before being released for analysis work (with an expected latency that is not real-time).
- The Mu2e experiment will also utilize grid resources at Fermilab for analysis, as well as other distributed resources among collaborators.
- The Mu2e experiment is estimated to produce approximately 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 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 g-2 and Mu2e experiments utilize OSG[139] software (with modifications) when applicable, which facilitates a majority grid-computing use case. Migration to some forms of HPC is probable, but due to project timelines/funding is not necessarily feasible.
- Data movement is typically handled as streaming, and coordinated through tools like XROOTD[140] and Rucio[141].

5.7.2 Muon g-2 and Mu2e Science Background

Muon g-2 and Mu2e are two precision experiments involving the study of muon particles at Fermilab. The experiments themselves are distinct in terms of their collaborations, science goals, and stages in their timeline, as well as computing, data, and networking needs. In this document, each experiment will be described in separate subsections. Where there are commonalities, those will be discussed in the "top" section, like here.

For both experiments, the stakeholders are the host laboratory, Fermilab, and the members and institutions of their respective collaborations, described in section 2 of this report.

Raw and reconstructed data from both experiments, as well as simulated events, are stored on tape using the Enstore system, cached to disk by dCache, and managed by the SAM data handling cataloging system. Mu2e is likely to transition from SAM to Rucio, a more modern data transfer system that is just now being deployed at Fermilab. These systems are supported by the Fermilab Scientific Computing Division (SCD).

Raw data, final analysis-format data sets ("n-tuples" or their evolution), and selected intermediate data sets will be digitally archived at Fermilab for several years after each experiment ends. The exact duration is yet to be determined. The experiments will follow the recommended policy for migration of tape-based data sets to new generation media and will work with SCD to ensure that their software can be built and will run correctly on the archived data sets so long as those data sets are retained.

5.7.3 Muon g-2 Case Study

5.7.3.1 Background

The Muon g-2 experiment at Fermilab aims to measure the anomalous magnetic moment of the muon to an unprecedented level of precision of 140 parts per billion and compare it to the prediction from the SM of particle physics. A statistically significant discrepancy would be an indication that the SM is incomplete and perhaps there are particles waiting to be discovered. More information may be found at the project website[142].

Data are created by the DAQs of the experimental apparatus and from MC simulations. There are two distinct DAQs at Muon g-2. The main DAQ takes data from the detectors that measure the decay positrons from the muons injected into the storage ring. A much smaller data set is taken by the Field DAQ, which takes data from fixed nuclear magnetic resonance magnetic probes as well as probes on a trolley that travels around the storage ring every few days. Each data set is processed into a reconstructed data set. The detector data are by far the largest and most complicated to reconstruct. Subsequent processing of the detector data occurs by analysis groups and individual physicists.

The experiment started taking data with a commissioning run in the spring of 2018, immediately followed by the first data-taking run (Run 1). Muon g-2 completed Run 3 in late March 2020, when Fermilab ceased accelerator

[139] https://opensciencegrid.org
[140] https://xrootd.slac.stanford.edu
[141] https://rucio.cern.ch
[142] https://muon-g-2.fnal.gov

operations due to COVID-19. The first three runs amount to 35% of the total data volume design goal. The remaining 65% of the data are planned to be taken in FY21 and FY22, a significant increase in the amount of data per year compared to the past runs.

5.7.3.2 Collaborators

The Muon g-2 collaboration[143] features a total of 203 collaborators from 35 institutions in 7 countries. The geographic locations of universities and laboratories are throughout the United States, the UK, Germany, Italy, Russia, South Korea, and Shanghai, China. The UK and Italy (INFN) have made available large computation farms for the use case. Elsewhere, and especially within the United States, Muon g-2 uses OSG resources. Data transfers are managed by the Fermilab SAM data handling system and generally use XROOTD (perhaps with https under the hood — the protocol is not visible to end users) as the transfer mechanism. All raw and reconstructed data as well as most analysis data are resident in the Fermilab data storage complex.

Detector and field data originate from Fermilab. MC data may be produced anywhere, but large samples are generated with resources on the OSG and at Fermilab itself. Muon g-2 is just now (August 2020) starting to produce simulation samples at NERSC.

Remote (non-Fermilab) institutions generally request copies of data at the "n-tuple" level, which are typically ROOT trees. These data sets are highly compressed and much smaller than the raw and reconstructed data sets (by a factor of 10 to 100). Each institution is responsible for storing the data it requests locally.

User/collaborator and location	Primary or secondary copy of the data	Data access method	Avg. size of data set	Frequency of data transfer or download	Are data sent back to the source and method	Any known issues with data sharing
OSG (ANALYSIS)	Primary at FANL	XROOTD	~ 2 TB in; ~ 0.2 TB out	On demand	XROOTD	N/A
OSG (PRODUCTION)	Primary at FANL	XROOTD	~ 20 TB in; ~ 20 TB out	On demand	XROOTD	N/A
NERSC (PROPOSED MC PRODUCTION)	N/A	Copy (typically XROOTD)	0 in; ~ 50 TB out	N/A	XROOTD	N/A
DEDICATED INFN (ITALY) AND UK RESOURCES	Secondary	Data transfer (typically XROOTD)	~ 100 TB in; ~ ~ 50 TB out	Minimal per year	XROOTD	N/A

Table 9: Muon g-2 data projections

We run three different types of jobs:

- **Production:** These jobs process raw data into reconstructed data suitable for analysis by physicists. Muon g-2 runs production jobs mostly on Fermilab resources (93% of jobs) but also opportunistically uses OSG (7%). Raw data large input files are read in (~ 2 GB) and large reconstructed files are output (also ~ 2 GB).
- **Analysis:** These jobs process the reconstructed data files and produce small n-tuples and plot files, or they process previously produced n-tuple files into more reduced n-tuples or plots. Most run using Fermilab resources with 10% of these jobs run on the OSG.
- **MC production:** These jobs create simulated events at the reconstructed level. They have no inputs and produce large output files (~2 GB). These jobs run mostly at Fermilab with ~ 10% running on the OSG. Muon g-2 has just started to run at NERSC, and may expand to ALCF.

In general, Muon g-2 does not specify the destination for jobs (except for NERSC, which currently is a special case). The use of the OSG (~10% of jobs) is governed by the availability of opportunistic resources.

[143] https://muon-g-2.fnal.gov/collaboration.html

Note that the data set size estimates are for the current data sample. That sample will grow by a factor of ~3 in the next two years.

We plan to use dedicated farms at INFN (Italy) and in the UK for specialized production purposes.

In **Table 9,** the average data set size is for an average campaign of jobs. For production, Muon g-2 may have anywhere from one to four campaigns running simultaneously. The number of simultaneous analysis campaigns running varies widely from a few to more than 10.

5.7.3.3 Instruments and Facilities

The Muon g-2 apparatus is described in detail in technical reports[144]. The readout data volume is completely dominated by the 24 calorimeter stations each containing a 9x6 array of PbF_2 crystals for a total of 1,296 crystals. The signal from each crystal is digitized with analog-to-digital converters (ADCs) at 800 Megasamples-per-second. Other devices that add to the readout are two stations of straw trackers, beam position monitors, and nuclear magnetic resonance probes to measure the magnetic field.

The Maximum Integrated Data Acquisition System (MIDAS) based DAQ collects the data from the devices for each "fill" of the storage ring when muons are injected by the Fermilab accelerator complex. There are 16 fills every 1.4 seconds. For a fill, the devices' signals are recorded for 700 microseconds. The DAQ employs GPUs to find islands of interesting signals corresponding to measuring a positron from a muon's decay in the calorimeter crystal readout. While there are no plans to alter or expand the readout devices, steps are being taken to improve the efficiency for muons to be stored. Such an improvement would increase the number of discovered islands in the calorimeter data and would, correspondingly, increase the data size.

Figure 13 shows the current and anticipated data sets. The units are the number of muon decays recorded as compared to the previous incarnation of the experiment at BNL. The goal is to measure 20 times the number of muon decays as BNL.

The raw data from the first three runs that have been collected amounts to approximately 4 PB. Reconstructed, analysis, and simulation data add another 5 PB.

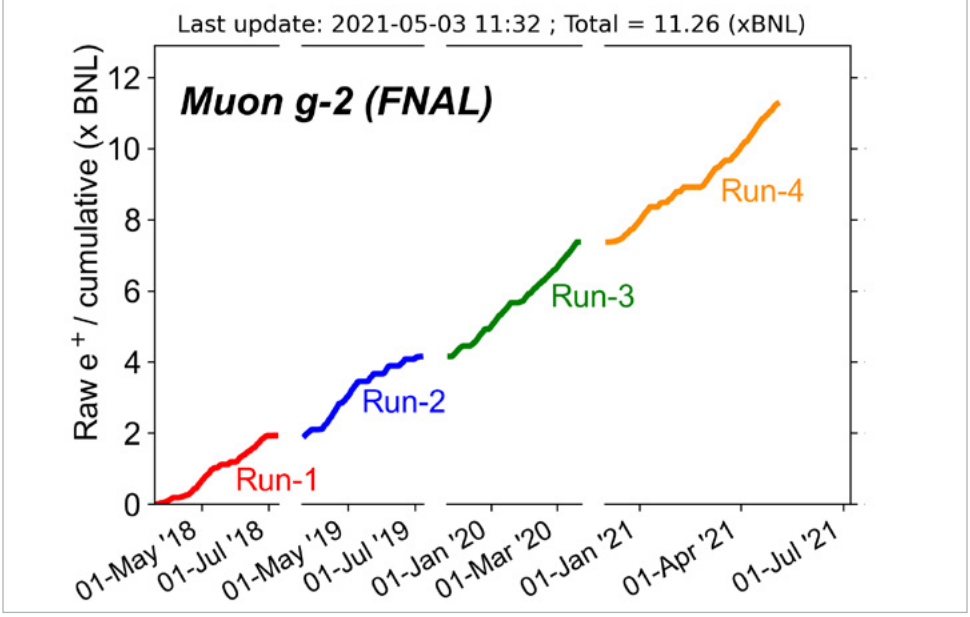


Figure 13: Muon g-2 data set projections

[144] https://arxiv.org/abs/1501.06858

For FY21, Muon g-2 expects to add 9 PB of raw, reconstructed, simulated, and analysis data for Run 4. For FY22, 7 PB of data for Run 5 could be added.

Currently, there are no plans to run past FY22. Production activities may extend into FY23 and analysis activities may well continue for a few more years after that to determine and publish the final result of the experiment.

5.7.3.4 Process of Science

The experiment expects to produce ~25 PB of data. This can be broken down into ~10PB of raw data, ~10PB of reconstructed and analysis data, and ~5PB of simulation data. These data need to be input and processed in order to obtain and publish the final result (we will publish at least one intermediate result along the way). Much of that processing happens with Fermilab resources. But as mentioned, Muon g-2 will take advantage of opportunistic OSG resources at the 10% level (this happens automatically), and is also starting to take advantage of NERSC and perhaps will branch out to ALCF (likely not OLCF as the algorithms are CPU based). A large bandwidth between Fermilab and NERSC/ALCF is especially beneficial as Muon g-2 would need to return output files back to Fermilab, though there will be an attempt to limit the amount of data returned. In parallel, data are also populating locations in Italy and the UK (actual locations are yet to be determined), so taking advantage of large bandwidth pipes would be beneficial as well, though the use of them would be rare and sporadic (just a few times a year).

Data reduction involves processing the reconstructed data files and outputting much smaller ROOT trees. Muon g-2 may in the future output HDF5 files as well.

Data analysis involves processing the ROOT trees and outputting even smaller ROOT trees or histogram files. The networking needs are not significant.

Our workflows are quite simple; production and analysis jobs read files sequentially. Some workflows are multistep and produce intermediate files that are not returned. All workflows create an output file, either of similar size to the input (for production) or much smaller (data reduction/analysis). Production jobs do require communication with a database server at Fermilab, though the amount of data transferred is very small. MC jobs have no input files and can produce large output (equivalent to production). The jobs are embarrassingly parallel and do not use interprocess communication.

While data taking is expected to end in FY22, Muon g-2 will likely continue with production and analysis for several (perhaps two to four) years afterwards in order to produce a final result and publication. There may be a need to reprocess all of the data with a final version of the reconstruction program. That task would require significant computing resources and networking bandwidth if processed off-site. Muon g-2 could decide to reprocess at an HPC center (we would need to populate the HPC center with the data). The need for this activity is unclear at this point, but experience suggests that it is something experiments typically do at the end of data taking and before a final result.

5.7.3.5 Remote Science Activities

As written earlier, the majority of the processing occurs with Fermilab resources. The main tasks that would require significant networking bandwidth are processing the raw data, reprocessing that data, and generating simulation samples. A significant reprocessing campaign may need to occur on remote sites if the Fermilab resources are too busy, but such a task is likely two to three years away if the collaboration decides it is necessary.

Normal use of remote sites has already been stated (10% OSG, special processing in UK and Italy). Reduced data samples may be transported to collaborating institutions and the final data sample in reduced form may be several hundred terabytes to one petabyte. Significant bandwidth may be required to move those data. In practice, however, the bulk of the data analysis so far occurs using Fermilab resources. After data taking ends in FY22, such transfers may be the bulk of network usage by Muon g-2.

As noted earlier, Muon g-2 is starting to use NERSC for generating simulation samples and may use ALCF as well. For simulation, data population is not required, but there will a need to be a step to return generated data back to Fermilab. The overall goal is to try to make this as small as possible (e.g., instead of generating full data sets, generate smaller n-tuples or plots).

5.7.4 Mu2e Case Study

5.7.4.1 Background

Mu2e at Fermilab [145] will search for the neutrinoless decay of a muon into an electron in the Coulomb field of an atomic nucleus. Observation of this process would provide unambiguous evidence for physics beyond the SM. The projected single-event sensitivity of Mu2e is 2.87 x 10-17, approximately four orders of magnitude more sensitive than the previous best experiment, SINDRUM-II. For the quoted sensitivity, the projected background level is about 0.4+/-0.2 events. More information is available in the Mu2e Technical Design Report (TDR)[146].

Mu2e raw data will flow from sensors in five subsystems, through the DAQ and trigger to storage on tape in the Fermilab central computing facility. The raw data will be processed into analysis-format data sets in a multistep process with a latency of a few weeks. The analysis-format data sets will be small and widely distributed to collaborating institutions; the intermediate data sets will be used by members of the collaboration to improve calibrations and algorithms. From time to time, Mu2e will reprocess all data using the most up-to-date calibrations and algorithms; when this is done, intermediate data sets older than a few iterations will be retired. For more information, see the Mu2e Data Management Plan[147].

Mu2e is still under construction. Using the pre-COVID schedule, Mu2e will start taking cosmic-ray commissioning data early in calendar 2022 and first beam data late in calendar 2023. Additional information is available in section 3, below.

In recent years, the computing needs of Mu2e were dominated by simulation campaigns to provide data sets of simulated events for the purpose of value engineering the design, validating the design, and providing input for the development of trigger, reconstruction, calibration, and analysis algorithms. A typical scale is 25 million core hours once every few years plus 10 million hours per year for developing the code to produce these data sets and extracting results from these data sets. This will remain the computing driver until the start of beam data in late calendar 2023.

At first the simulations were done on OSG, including Fermilab, but in recent years Mu2e has used resources on Theta at ALCF. The output of each of these simulation campaigns is on the scale of 250 TB.

5.7.4.2 Collaborators

The Mu2e collaboration[148] features a total of 242 scientists from 40 institutions in 6 countries: China, Germany, Italy, Russia, the UK, and the United States.

The primary data store for raw data and intermediate data sets will be in the computing centers on the Fermilab site. Most of the calibration jobs, and all of the reconstruction jobs, will be run on Fermigrid or use opportunistic cycles on OSG. Mu2e has no collaborators with access to dedicated cycles on OSG.

There have been preliminary discussions with groups at INFN in Italy and in the UK for them to provide computing resources at the level of a few million CPU hours per year that would be used for the calibration of subsystems for which they have the primary responsibility. At this time, there are no firm commitments.

[145] https://mu2e.fnal.gov
[146] https://mu2e-docdb.fnal.gov/cgi-bin/ShowDocument?docid=4299
[147] https://mu2e-docdb.fnal.gov/cgi-bin/ShowDocument?docid=5993
[148] https://mu2e.fnal.gov/mu2e_collaboration_list.shtml

Should these plans be realized, Mu2e estimates that outbound data sets will be on the scale of 0.5 PB/year to each location and that inbound data sets will be small, at most a few tens of TB/year. The inbound data sets will be calibration sets and data quality monitoring information.

Mu2e expects that many institutions will request copies of analysis-format data sets and that these will be stored locally at each institution. At this time, Mu2e estimates that one copy of the analysis-format data set for the full experiment might be, at most, on the scale of 25 TB; it may be considerably smaller.

Simulations will be run on Fermigrid, OSG, NERSC, and ALCF. The simulated events and reconstructed simulated events will be stored at Fermilab. Mu2e expects that many institutions will request copies of the analysis-format data sets derived from the simulated events; these data sets are likely to be about a few times the volume of the corresponding experimental data sets.

In this table, "in" refers to traffic from Fermilab into the location and "out" refers to traffic from the location to Fermilab. See the figure in 5.7.4.4 for details of the time dependence.

User/collaborator and location	Primary or secondary copy of the data	Data access method	Avg. size of data set	Frequency of data transfer or download	Are data sent back to the source and method	Any known issues with data sharing
OSG	Primary is at Fermilab	XROOTD	~5 PB in ~1 PB out	Yearly	Y, method TBD	N/A
HPC FACILITIES FOR MC PRODUCTION (NERSC, ALCF)	N/A	Copy (method TBD), out only	~0.5 PB out	Yearly	Y, method TBD	N/A
INFN (ITALY)	Secondary	Copy (method TBD) in/out	~0.5 PB in ~0.02 PB out	Yearly	Y, method TBD	N/A
UK	Secondary	Copy (method TBD) in/out	~0.5 PB in ~0.02 PB out	Yearly	Y, method TBD	N/A

Table 10: Mu2e data projections

Mu2e will run three classes of jobs:

Data production is a multistage process: select events for calibration, perform calibrations, perform main reconstruction pass and optional secondary reconstruction passes, and create analysis-format data sets. From time to time, the secondary reconstruction will be repeated and perhaps once every few years the full data set will be reprocessed; in both cases the goal is to take advantage of improved calibrations and algorithms. Mu2e plans to run these jobs using dedicated Fermigrid resources and opportunistic OSG resources. The breakdown is not known at this time. For purposes of this study, Mu2e will assume that 50% of the work is done on Fermigrid and that other 50% is done on OSG. At this time, there are no plans to use HPC resources for data production.

End-user analysis: the input to this work is the analysis-format data sets that are the final product of data production. This work will likely be distributed across Fermigrid, opportunistic OSG, and resources at remote institutions.

MC production: these jobs create simulated events and process them through the same reconstruction chain as for experimental data. These jobs have no data inputs, only configuration input. An unusual feature of Mu2e compared with other HEP experiments is that the output of these data sets is small per unit CPU. A simulation campaign of 10 million hours will produce somewhere between 50 TB and 250 TB, depending on what processes are simulated and what level of intermediate results are retained. The reason is that simulation of the signal is trivial and does not require significant resources. What takes time is the simulation of background and pileup processes. Mu2e has exploited all of the opportunities to sculpt these processes at the generator level, but the reality is that there is a need to include many corners of generator phase space that produce very rare, but still

important, backgrounds. Therefore, most simulated events fail the selection criteria and are not written to the output file. The dominant workload in these jobs is Geant4, and Mu2e has demonstrated the ability to use multi-threaded Geant4 on KNL-based HPC resources. Mu2e has run these jobs on Fermigrid, on OSG, on Theta and BeBop at ALCF, and on Cori I and II at NERSC. The analysis-format data sets generated from these simulations are on the scale of 2 TB per 10 million hours of computing.

5.7.4.3 Instruments and Facilities

For full information about the Mu2e apparatus, see the Mu2e TDR[149].

The parts of the apparatus that generate data are a straw tube tracking system, a caesium iodide (CSI) calorimeter, a plastic scintillator-based Cosmic-Ray Veto System, a pixel telescope-based Beam Extinction Monitor System, and a Stopping Target Monitor comprising two crystals: one high purity Germanium and the other LaBr. Other major subsystems include the proton delivery system, including an extinction insert, and the Muon Beamline comprising the three large superconducting magnets, the production target, a collimator system, the stopping target, and the muon beam stop. One can also consider the detailed arrangement of shielding materials as a subsystem in its own right.

All of these subsystems are located in the Muon Campus at Fermilab.

The operational cycle of Mu2e is 1.4s, divided into an on-spill period of approximately 0.45 s and an off-spill period of approximately 0.95 s. During the on-spill period, a proton pulse hits the production target every 1695 ns and the apparatus is configured to record one event for every proton pulse. During the off-spill period, the apparatus will be configured to record cosmic rays and to perform in-band calibration tasks; off-spill events will have a duration of 100 μs, which is a tradeoff between several competing requirements. Data from the five data-generating subsystems flow into a streaming, deadtime-less DAQ and from there into a software-based trigger system that will inspect every event. During on-spill data taking, the trigger will accept approximately 1 in 400 events. During off-spill data taking, the trigger will select cosmic-ray events in the tracker and calorimeter at a rate of order 10 Hz and will prescale other events to collect unbiased samples that are needed to optimize the time dependent trade-offs between efficiency and fake rates.

At the design rate of protons on target, the Mu2e Triggering and Data Acquisition (TDAQ) system will produce 7 PB/year of raw data, which will be copied to the Fermilab computer center for processing and long-term storage.

Mu2e is currently under construction. Using the pre-COVID schedule, Mu2e will start taking cosmic-ray commissioning data early in calendar 2022 and first beam data late in calendar 2023. When the proton beam data begin, the rate of protons on target (POT), will be about 55% of the design rate and the experiment will ramp up toward the design POT rate during early calendar 2025. Starting in summer 2025, there will be a two-year shutdown of Fermilab accelerator operations to allow building the LBNF beamline. Mu2e will resume operations at full design POT rate in the fall of calendar 2027 and run until summer 2030.

Accelerator operations for Mu2e will be managed by the Fermilab Accelerator Division while the operation of the Mu2e apparatus will be managed by staff from the collaboration and two Fermilab divisions: the Particle-Physics Division and Technical Division. Operation of computing for Mu2e will be managed by the collaboration, drawing on resources, services, and staff supplied by the Fermilab Computing Sector.

For the Snowmass 2013 planning exercise, members of Mu2e performed preliminary design studies for an upgrade of the apparatus, named Mu2e-II. This work has been extended several times since then, and there is currently an effort underway to take a big step forward as part of the Snowmass 2021 planning exercise. The goal is to identify the minimal set of modifications to the apparatus that will improve the physics reach by a factor of 10, using the proton beam that will be available from the Fermilab PIP-II Linac. The technically limited schedule would be to begin commissioning Mu2e-II about two years after the end of Mu2e.

[149] https://mu2e-docdb.fnal.gov/cgi-bin/ShowDocument?docid=4299

Throughout all of these periods, Mu2e expects to use computing resources supplied by the SCD. On the timescale of now to five years from now, Mu2e expects that the primary resource for data production will be Fermigrid and OSG. For MC production there is an expectation to use HPC facilities that are available, including Haswell-like cores and KNL cores.

The comments in the previous paragraph also apply to the Mu2e-II design work that will be done during this period.

At this time, Mu2e has no short-term plans to run reconstruction code on HPC resources. The reason is that the reconstruction code is not thread-safe and making it so would divert critical effort that is required to prepare for and commission the detector and computing operations. Once operations are established, Mu2e will have effort available for making the code thread-safe, porting selected components to GPU-based architectures, and reimplementing algorithms to exploit resources that will then be available or in the pipeline. This effort could begin in calendar 2025 and is likely to be well underway during the LBNF shutdown of fiscal 2026 to 2027. Once Mu2e has this capability, a large fraction of the reconstruction work will be well suited to HPC centers.

One could also imagine upgrading the trigger hardware to a new architecture with enough additional power that it is possible to greatly improve the trigger rejection without loss of good events. The return on this investment would be a greater physics reach accompanied by a reduction in the resources required for reconstruction, including network resources. It is very likely such a trigger system could be in place for Mu2e-II and one could imagine testing components of that system during the later stages of Mu2e. At this time, it seems unlikely that this work would converge on time for an upgrade of the Mu2e trigger compute servers during the LBNF shutdown.

In summary, on the two-year time horizon Mu2e's computing needs will be dominated by simulations with needs similar to recent years. Toward the end of that period, Mu2e will begin to take cosmic-ray data, which will mean a modest increase in data processing needs. In the period two to five years from now, Mu2e will complete commissioning and will have two years of physics data taking, most of it at 55% of nominal POT. Toward the end of that period, Mu2e will begin the effort of porting code to run on new architectures. On the timescale of five to 10 years from now, Mu2e will have a three-year run at nominal POT and, if Mu2e-II is funded, construction of Mu2e-II will begin. The Mu2e-II trigger will be in an excellent position to exploit then-available technologies.

5.7.4.4 Process of Science

At design POT rate, the Mu2e experiment will produce about 7PB/year of raw data, including physics data, calibration data, and special runs. Approximately half of this will have two copies stored on tape and the rest will have only a single copy. The details of how many events will be selected for calibration, how many events will be written out by reconstruction, and the size per event of these data sets are not well understood at this time. The current estimate is that an additional 5 PB/year will be needed for intermediate data sets, repeat processing, and for storage of simulated events. This brings the total tape needs to about 15 PB/year. While this number may seem small, the discussion below explains it.

When data arrives at the computer center, the goal is to complete the first processing step with a turnaround of less than six hours 95% of the time. This step will select events to be used for calibration and will produce near-line monitoring of data quality. It may also split the event stream into several sub-streams based on trigger information. The data sets used as input to calibration will be quite small. First, most events will not be interesting for calibration. Second, the output events will be much smaller than the input events. In a typical Mu2e on-spill event, the duration of the live window is about 1 μs, during which time the tracker will record about 3,000 hits; most triggered events will contain a single reconstructable track and only a small fraction will contain more than one. A typical reconstructed track will have 30 to 50 hits on it, and the time that a track is "live" in the apparatus is about 100 ns. The vast majority of the hits are not associated with any reconstructable track, and are not relevant for most calibration purposes. Therefore, the calibration data sets will be stripped of uninteresting hits, making them much smaller per event than the raw data. Mu2e can perform a similar

compression on the hits in the calorimeter and CRV. A subset of the events will be written out without this compression so that they can serve as input to other studies.

Once the calibration workflows are complete, the raw data will then be processed through the main reconstruction pass. There are many open questions here: will Mu2e read the full raw data or one of the sub-streams of the raw data? How will Mu2e prescale events that have good reconstructable tracks that are of no value for physics but have value for monitoring? For what fraction of the data set will uninteresting hits be removed? The net result is that Mu2e expects the size of the output of the main reconstruction pass to be smaller than the size of the raw data.

For the first two years of data taking, the POT rate will be about 55% of nominal. For some of this time Mu2e may open up some prescales to study events that would normally be discarded, but there is a belief that for most of this period, the computing needs will scale with the POT rate.

Note that not all intermediate data sets will be written to tape. For example, the data sets produced by iterations over the same data several times in order to derive a calibration set will likely be disk resident only and be discarded once the calibration information is extracted.

As discussed previously, the present plan is that it would be possible to process all of these data on Fermigrid, but for planning purposes Mu2e assumes that about half of it will be processed on Fermigrid and half on opportunistic OSG. Discussions have taken place that may lead to some calibration work being done in the UK or at INFN, using resources provided by their funding agencies. For this exercise there is an assumption that about 0.5 PB/year will be transferred to each location.

Figure 14 shows an estimate of the time dependence of Mu2e network bandwidth needs from now until the end of analysis.

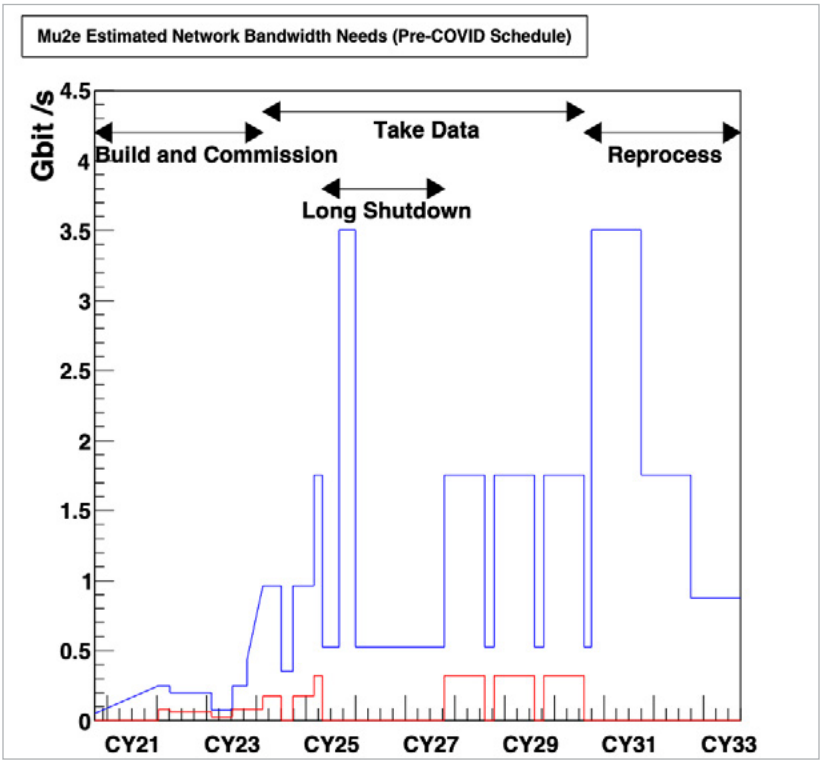


Figure 14: Mu2e estimated ban

In **Figure 14,** the blue curve shows the estimated bandwidth needs for Mu2e for running jobs on OSG; the notations at the top of the figure show the main periods in the lifecycle of the experiment. The spike in bandwidth

during the long shutdown represents reprocessing needs during this shutdown. The bandwidth will be used roughly five-sixths for traffic from Fermilab to OSG sites and one-sixth for return traffic. The curves represent the average bandwidth that is needed on a monthly basis. It is still possible that usage within a month may exceed the expected average by as much as five times, but would be short lived. The red curve shows the bandwidth that may be needed to send data to INFN and the UK; the curve represents the sum of both contributions. Not shown in this figure is the bandwidth needed for major simulation campaigns which will occur at times that are not yet known; each campaign will require on the order of 0.5 Gbit/s for a period of a month. There will be three or four such campaigns each year. The assumptions going into this plot are discussed in the text.

5.7.4.5 Remote Science Activities

As discussed earlier, Mu2e can do all of its data production using only Fermigrid resources but will exploit opportunistic cycles on OSG should they be available. For planning purposes, assume that about 50% of data production work will be done on OSG. On the two-year time horizon, the data volume is not well known, but a high-end estimate is in the range of 0.5 PB/year outbound and 0.1 PB/year inbound.

On the two- to five-year horizon, the estimate is better understood and expected to be about 3.0 PB/year outbound and 0.6 PB/year inbound. On a timescale of 5 to 10 years from now, Mu2e estimates 5 PB/year outbound and 1 PB/year inbound.

There is a possibility that some data may be copied to INFN or to the UK. If those plans come to fruition, very little data will be transferred within the next two years. On the timescale of three to five years from now, Mu2e projects that networking requirements will be 0.3 PB/year outbound and 0.01 PB/year inbound. On the timescale of 5 to 10 years from now, the network needs will be 0.5 PB/year output and 0.02 PB/year inbound.

Every year, Mu2e plans to use some mix of ALCF or NERSC to perform simulations. The amount of work each year may vary; there are estimates that in a year with a lot of simulation work, transferring 0.5 PB from one of these centers to Fermilab is possible.

Mu2e projects that the analysis-ready data sets for each year of data will be on the scale of 5 TB for experimental data and perhaps 15 TB for simulated events. These will be copied to remote institutions for use there.

5.7.5 Shared Software Infrastructure

All of the data movement and management for the Intensity Frontier Fermilab experiments, which include Muon g-2 and Mu2e, are handled by the Fermilab Fabric for Frontier Experiments (FIFE) project[150]. Components include SAM for data management, transport, and cataloging; Jobsub/HEPCloud for job submission; and Fermi-FTS for automated and robust file transfer management[151]. These components use standard open-source tools such as XROOTD[152].

FIFE and subcomponents are significant projects at Fermilab and are managed by the Fermilab SCD. Note that there is a plan to replace SAM with Rucio[153] and some experiments, such as ProtoDUNE, are using Rucio components. Currently, Muon g-2 and Mu2e are using SAM proper. Given the limited life of Muon g-2, that experiment is likely to stay with SAM. Mu2e is following the progress of Rucio and will make a decision in the next year whether to adopt it promptly or to wait until the LBNF shutdown.

Both experiments use Geant4[154] to simulate the interaction of particles with materials and are following the development of improvements to Geant4 and the development of GeantV. Mu2e has already demonstrated theability to use multi-threaded Geant4, thereby opening up running on memory-poor machines, such as KNL-based HPC machines.

[150] https://cdcvs.fnal.gov/redmine/projects/fife/wiki
[151] Warning: Some of these names, like SAM and FTS, are used in other projects and may lead to confusion. See the FIFE web page for the meanings.
[152] https://xrootd.slac.stanford.edu
[153] https://rucio.cern.ch
[154] https://geant4.web.cern.ch

Both experiments write their own C++ code for processing raw data into processed products (by the reconstruction program). Muon g-2 and Mu2e write their code within the art event processing framework[155]. At this time, both experiments write out data as ROOT tree files. Muon g-2 may experiment with HDF5[156] in order to perform some analyses at HPC centers, though this has not started yet.

Many experiments are moving to Python-based analysis codes. This choice is generally left to the individual analyzers.

The general evolution of these software tools is an important open discussion for the intensity frontier experiments and especially DUNE. Muon g-2 is likely to remain with its current system of art and C++ code given the lifetime of the experiment. Mu2e will likely take advantage of progress spearheaded or in collaboration with DUNE.

5.7.6 Fermilab Network and Data Architecture

The Mu2e and Muon g-2 physics experiments utilize shared scientific computing resources at Fermilab, including the general dCache disk storage and the Enstore tape storage facilities. With the exception of the US-CMS Tier 1, all of the other HEP experiments hosted at Fermilab also utilize those storage resources. While the storage facilities have direct access to special-purpose science data WAN paths, such as LHCOPN, LHCONE, OSCARs circuits, etc., neither Mu2e nor Muon g-2 currently anticipate requiring those types of customized WAN services. In terms of expected WAN bandwidth requirements, Mu2e and Muon g-2 are expected to contribute only a negligible part of Fermilab's aggregate WAN bandwidth needs. For the foreseeable future, Mu2e and Muon g-2 WAN requirements should be satisfied by the general routed Internet Protocol (IP) service that ESnet provides. However, if preferential network services emerge as part of ESnet's routed IP network services, such services could find use by the Mu2e and Muon g-2 experiments.

Figure 15 depicts the local data storage facilities used by Mu2e and Muon g-2, as they interface to ESnet.

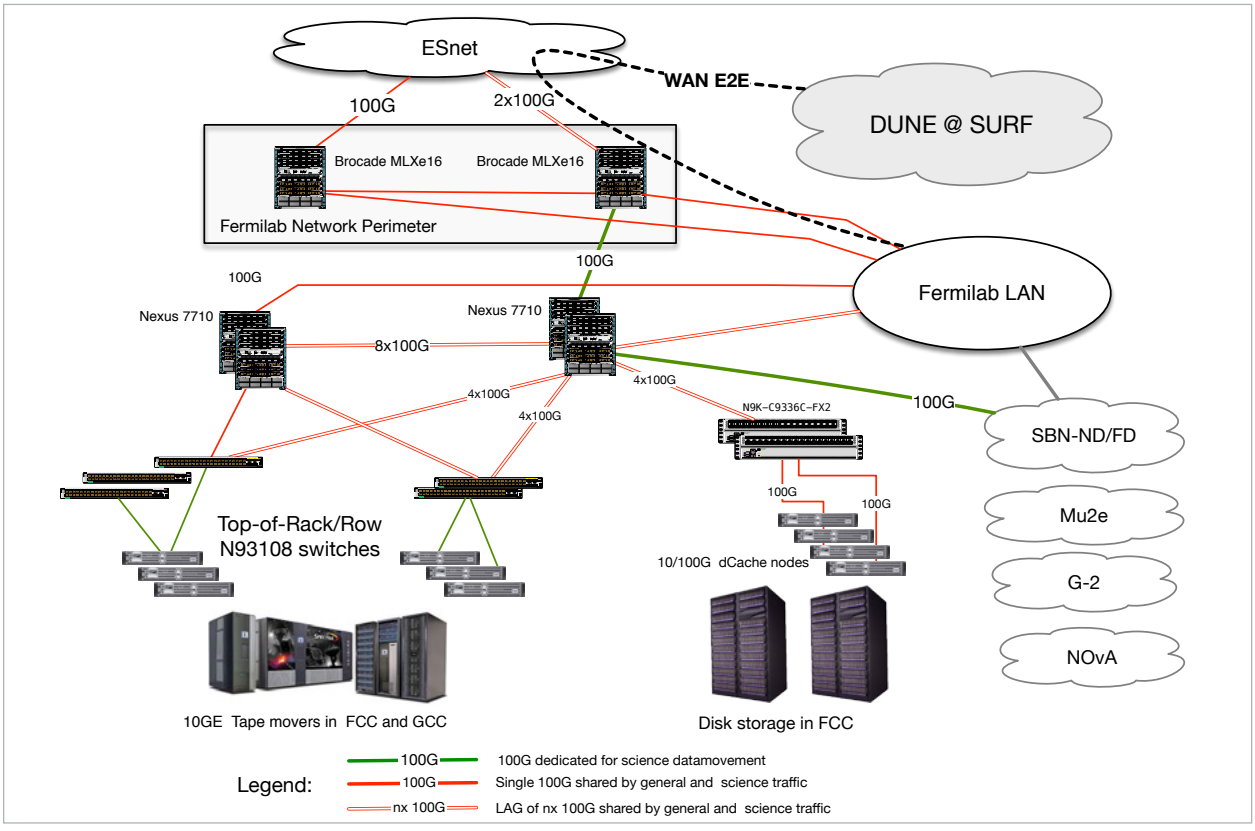


Figure 15: Fermilab data architecture

[155] http://art.fnal.gov
[156] https://support.hdfgroup.org/HDF5/

In terms of planned network upgrades that will affect computing resources used by Mu2e and Muon g-2, upgrade to 400GE network technology for the internal LAN infrastructure supporting the storage systems, as well as Fermilab's WAN infrastructure, is expected in the 2022 to 2023 time frame.

5.7.7 Shared Cloud Services

In general, use of cloud services is not a decision made by the experiment, but rather by Fermilab SCD if a proposed processing task is deemed suitable and other resources are not available or not compatible. Currently, and for the foreseeable future, Mu2e and Muon g-2 will not use cloud resources explicitly. A future feature of HEPCloud is that it may choose to use cloud resources if the pricing is competitive and other resources are full.

5.7.7.1 G-2 Cloud Services

G-2 does not anticipate having its own funding to purchase cloud services. The only use of cloud services will be when SCD chooses it as the most cost-effective solution. G-2 will structure its work so that this option is always open.

5.7.7.2 Mu2e Cloud Services

Mu2e does not anticipate having its own funding to purchase cloud services. The only use of cloud services will be when the SCD chooses it as the most cost-effective solution for a particular job; in such cases, funding will flow through the SCD. Mu2e will structure its work so that this option is always open.

5.7.8 Muon Experimentation Data-Related Resource Constraints

HEP experiments depend on data processing to achieve physics results and knowledge discovery. A significant future constraint in processing capability, including a networking constraint that reduces the use of off-site resources, would obviously have a negative impact on the physics output of the experiments. The projects are not aware of any such explicit future constraints. The computing and networking bandwidth needs for both Muon g-2 and Mu2e are small in comparison with the HL-LHC experiments and DUNE. That being said, if resources are not enough for HL-LHC and DUNE, then smaller experiments will likely suffer as well. Furthermore, the aggregate activity of smaller experiments may generate enough load to affect the entire community.

5.7.9 Outstanding Issues

There are no additional issues to track in this section.

5.7.10 Case Study Contributors

Mu2e and g-2 Representation

- Elizabeth Sexton-Kennedy[157], Fermilab
- Adam Lyon[158], Fermilab
- Robert Kutschke[159], Fermilab

ESnet Site Coordinator Committee Representation

- Phil DeMar[160], Fermilab
- Andrey Bobyshev[161], Fermilab

[157] sexton@fnal.gov
[158] lyon@fnal.gov
[159] kutschke@fnal.gov
[160] demar@fnal.gov
[161] bobyshev@fnal.gov

5.8 Belle II Experiment

Belle II is a third generation "B meson" experiment located at KEK. It is expected to operate through 2030, and is a worldwide collaboration (of which BNL is a major supplier of computation and storage). Belle II uses a grid paradigm for computation, and the analysis of data is fully distributed around the world. The first step in this process relies on the use of data mobility tools to migrate the raw output to facilities that create more compact formats that are then distributed again to allow for analysis activities. 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 that are 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.

5.8.1 Discussion Summary

The following discussion points were extracted from the case study and virtual meetings with the case study authors. These are presented as a summary of the entire case study, but do not represent the entire spectrum of challenges, opportunities, or solutions.

- BNL is a major supplier of computation and storage to the overall collaboration. The Belle II raw data consists of two copies: one copy at KEK and the second copy at BNL. Beginning 2021, the second copy will be 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 (e.g., 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.
- Computing operations use a grid paradigm, where analysis is fully distributed around the world, and operations rely on data movement to migrate raw output to centers that can convert into more usable analysis formats.
- 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 will be migrating to Rucio which will be operated by BNL. During the transition, and during operation, the experimental operations staff will be watching latency-based interactions between the United States 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 HTC functions. This infrastructure features diverse 100 Gbps paths to ESnet, and averages multiple PBs of data transferred monthly.
- Belle II's success relies on transpacific networking capacity via the R&E community provided by NSF-funded links (e.g., TransPac, PacWave) as well as those provided by the Japanese science collaborations (SINET). These links provide sufficient capacity and fail-over for a number of projects that collaborate between the Asia-Pacific region and the United States.

- BNL participates in LHCONE for use in both LHC experiments (LHC Tier 1 for ATLAS), and Belle II. One of the most complex areas in operating this type of network infrastructure is the adherence to the LHCOPN/LHCONE AUPs. In a multipurpose lab utilizing a unified network perimeter, this becomes exponentially complex as scientific programs want exclusivity over a VPLS or L3VPN circuits while utilizing BGP (e.g., LHCONE, LHCOPN, or the possibility of a Multi-One deployment).

5.8.2 Belle II Experiment Case Study

5.8.2.1 Background

The Belle II Experiment collects data at the SuperKEKB asymmetric electron-positron collider at KEK in Tsukuba, Japan. SuperKEKB operates at and near the $\Upsilon(4S)$ resonance to produce pairs of B mesons. Other heavy flavor particles, such as charm mesons and tau leptons, are also produced. Belle II precision measurements of rare decays and charge conjugation parity (CP) violation (matter-antimatter asymmetries) in heavy quarks and leptons provide a unique probe of new physics beyond the SM.

The goal of Belle II is to accumulate approximately 50 times more collisions than the predecessor experiment (Belle) by 2030. Belle II began operation in 2019 and has collected about one-tenth as much data as Belle. The plan is to increase the amount of data acquired per year to reach the goal in 2030. The estimated data volume to be stored at BNL through 2027 is provided in Section 5.8.3. The Belle II collaboration comprises 1,034 members from 119 institutions from 26 countries around the world. There are 127 US collaborators from 18 institutions, including BNL. The US contribution to Belle II is funded by the DOE OS HEP.

Belle II has a grid-computing model and uses a three-level hierarchical structure of computing sites: raw data centers, regional data centers, and MC (simulated data) production centers (See **Figure 16**). Raw data centers are also regional data centers and MC production centers. Regional data centers are also MC production centers. The raw data coming out of the DAQ at KEK are permanently stored, calibrated, and processed at raw data centers. **Figure 17** shows the raw data distribution scheme among data centers starting 2021. All raw data centers are connected to LHCONE.

Two copies of the raw data are stored; one copy at KEK and the second copy at BNL. Beginning in 2021, the second copy will be distributed between BNL (30%), Canada (15%), France (15%), Germany (20%), and Italy (20%). The fully reconstructed events, coming from the raw data processing step, are stored in the miniDST (mDST) format. MC events are simulated and reconstructed using the same software used to process detector events and then also stored in mDST format. Detector and MC mDST are stored in regional data centers. User analysis is generally performed on filtered subsets of the mDSTs dubbed uDSTs. Generally, two replicas of mDSTs and uDSTs are stored. In addition, Belle II retains replicas of the results of processing with a previous major software release, which are expected on a yearly basis.

The use and allocation of Belle II computing resources is re-evaluated each year, taking into account past performance and experience as well as expected changes in data rates. The computing model is periodically updated to ensure efficient usage of resources.

Belle II plans to upgrade the DAQ in 2020 to 2021 and replace the innermost vertex detector and a portion of a particle identification detector in 2022. An upgrade of SuperKEKB is currently anticipated for 2026. No concomitant upgrade of the Belle II detector is currently planned. The data volume and resource estimates in this report take these upgrades into account.

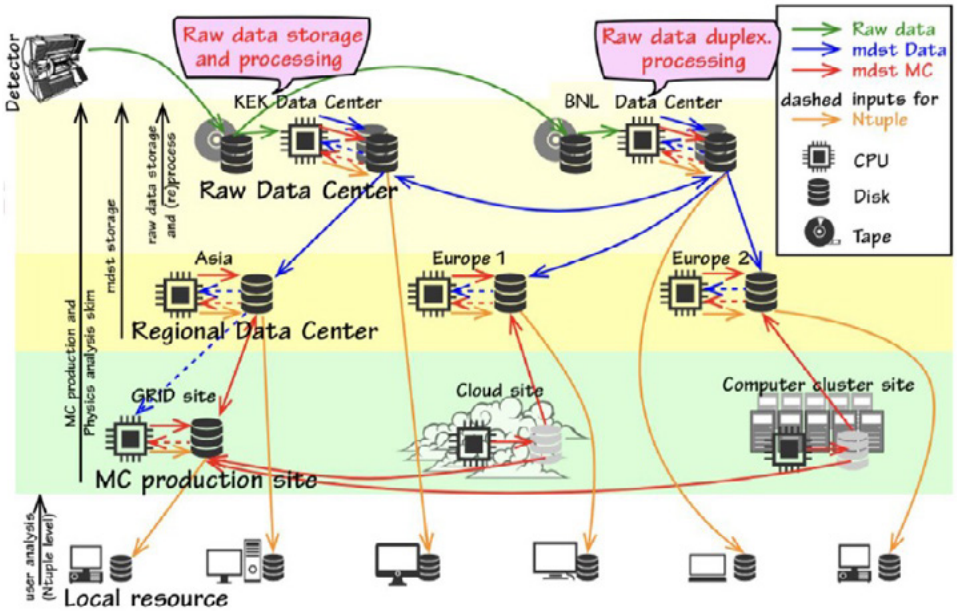


Figure 16: Belle II computing sites organization

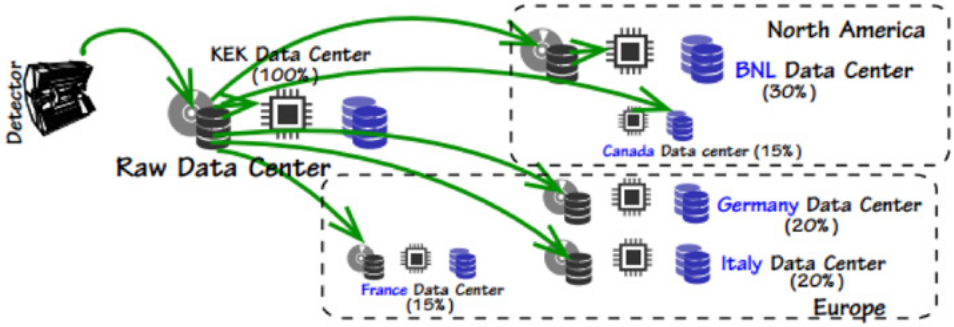


Figure 17: Belle II raw data distribution scheme

5.8.2.2 Collaborators

Estimated resource levels for 2021 are provided in the table. Note that secondary and primary copy storage is listed separately (e.g., Italy holds both a primary and secondary copy and appears twice). Transfer to primary storage is constant during data taking (up to nine months per year) and minimal otherwise. Transfer to secondary storage occurs during production of simulation data or processing of raw data which nominally occurs throughout the year with an ~83% duty factor. Distributed data management techniques (discussed later in the case study) enable "grid" access.

User/collaborator and location	Primary or secondary copy of the data	Data access method	Avg. size of data set	Frequency of data transfer or download	Are data sent back to the source and method	Any known issues with data sharing
CANADA	Primary	Grid	0.18	Constant during running	No	None
FRANCE	Primary	Grid	0.18	Constant during running	No	None
GERMANY	Primary	Grid	0.25	Constant during running	No	None
ITALY	Primary	Grid	0.25	Constant during running	No	None
KEK	Primary	Grid	1.62	Constant during running	No	None
USA	Primary	Grid	0.76	Constant during running	No	None
ARMENIA	Secondary	Grid	0.04	Constant during production	No	None
AUSTRALIA	Secondary	Grid	0.16	Constant during production	No	None
AUSTRIA	Secondary	Grid	0.12	Constant during production	No	None
CANADA	Secondary	Grid	0.23	Constant during production	No	None
CHINA	Secondary	Grid	0.57	Constant during production	No	None
CZECH	Secondary	Grid	0.08	Constant during production	No	None
FRANCE	Secondary	Grid	0.21	Constant during production	No	None
GERMANY	Secondary	Grid	1.36	Constant during production	No	None
INDIA	Secondary	Grid	0.26	Constant during production	No	None
ISRAEL	Secondary	Grid	0.08	Constant during production	No	None
ITALY	Secondary	Grid	0.98	Constant during production	No	None
KEK	Secondary	Grid	1.66	Constant during production	No	None
KOREA	Secondary	Grid	0.40	Constant during production	No	None
MALAYSIA	Secondary	Grid	0.04	Constant during production	No	None
MEXICO	Secondary	Grid	0.12	Constant during production	No	None
POLAND	Secondary	Grid	0.16	Constant during production	No	None
RUSSIA	Secondary	Grid	0.57	Constant during production	No	None
SAUDI ARABIA	Secondary	Grid	0.04	Constant during production	No	None
SLOVENIA	Secondary	Grid	0.18	Constant during production	No	None
SPAIN	Secondary	Grid	0.04	Constant during production	No	None
TAIWAN	Secondary	Grid	0.14	Constant during production	No	None
THAILAND	Secondary	Grid	0.08	Constant during production	No	None
TURKEY	Secondary	Grid	0.04	Constant during production	No	None
UKRAINE	Secondary	Grid	0.08	Constant during production	No	None
USA	Secondary	Grid	1.35	Constant during production	No	None
VIETNAM	Secondary	Grid	0.02	Constant during production	No	None

Table 11: Belle II data projections

5.8.2.3 Instruments and Facilities

Figures 18 and **19** show the amount of usable disk and tape resources provided by the Belle II Tier 1 site at BNL (values in FY18–20 range are actual delivered values, and values in the range FY21–27 are from the most up-to-date projection as of August 2020).

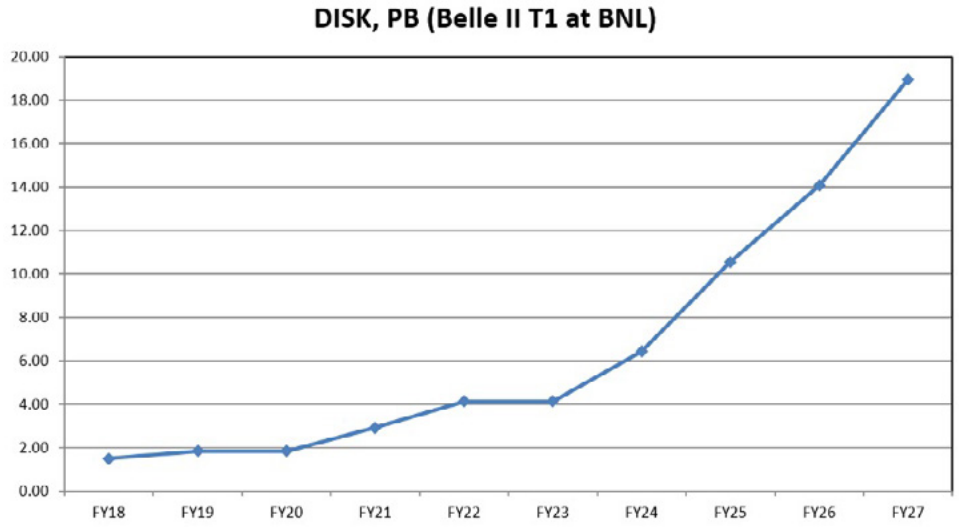


Figure 18: Belle II disk resources at BNL

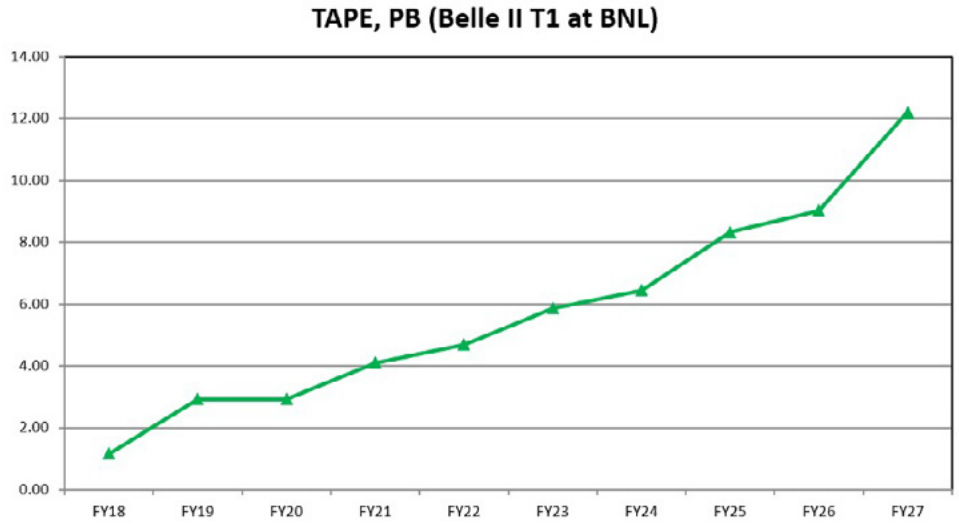


Figure 19: Belle II tape resources at BNL

5.8.2.4 Process of Science

A copy of the raw data is transferred to the raw data centers using the Distributed DMS (see next section for the software). A fraction of this raw data is needed for calibrating the detector (calibration skims, raw data format). Until summer 2020, the calibration was performed at the KEK Central Computer System (KEKCC) and from fall 2020 the calibration skims will be sent to BNL. The calibration procedure will be managed by an Airflow service at Detaches Electronic-Synchrotron (DESY), Germany, which submits jobs to the HTCondor queues at BNL. The calibration proceeds in several stages with an important reusable intermediate data product, the cDST, being a byproduct: many re-calibrations can reuse this data format. The raw data are approximately 70 kB/event, while a cDST is approximately 120 kB/event but is produced on only a fraction of the data. Calibration constants are uploaded to the Conditions database at BNL.

Once calibration constants are available, the raw data can be reconstructed to produce the mDST that can in principle be used for physics analysis; it has a size of approximately 15 kB/event. As the retention rate of most analyses is on the order of 1%, a subsequent processing step produces uDSTs for several event selections. The event size of this data format is approximately 20 kB/event as information is added to the mDST information. Physicists process the uDST files with a grid-based analysis framework (gbasf2) to produce n-tuples that they download to their local resources for subsequent analysis.

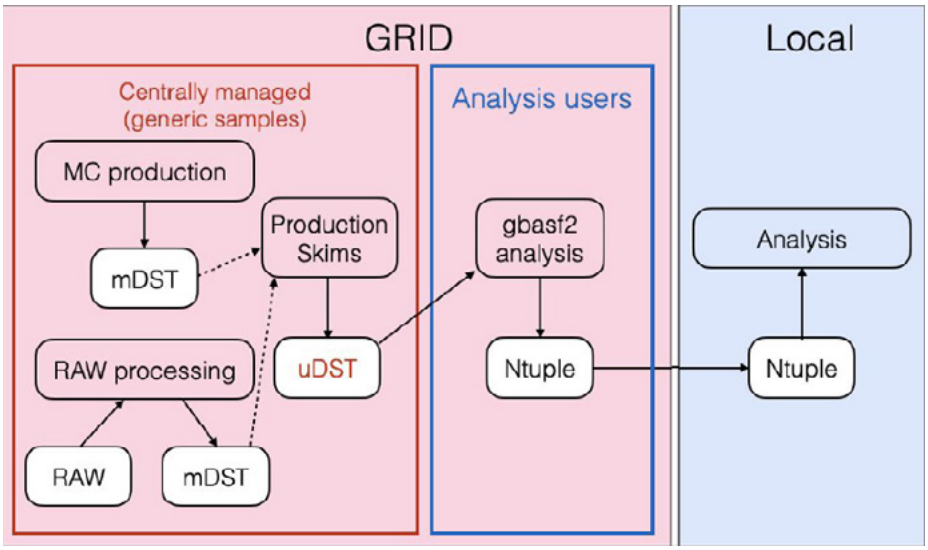


Figure 20: Belle II computing model

In the Belle II computing model (see **Figure 20**), raw data processing can only occur at raw data centers. The output mDST is distributed across seven regional data centers. uDST files are distributed across many more sites. For simulation, the picture is similar but simpler: mDST is produced directly across the more numerous MC production sites, and uDST distribution follows. A key input to simulation are samples of background data taken from real data. These background files must be distributed across the MC production sites to respect the MC production policy.

5.8.2.5 Remote Science Activities

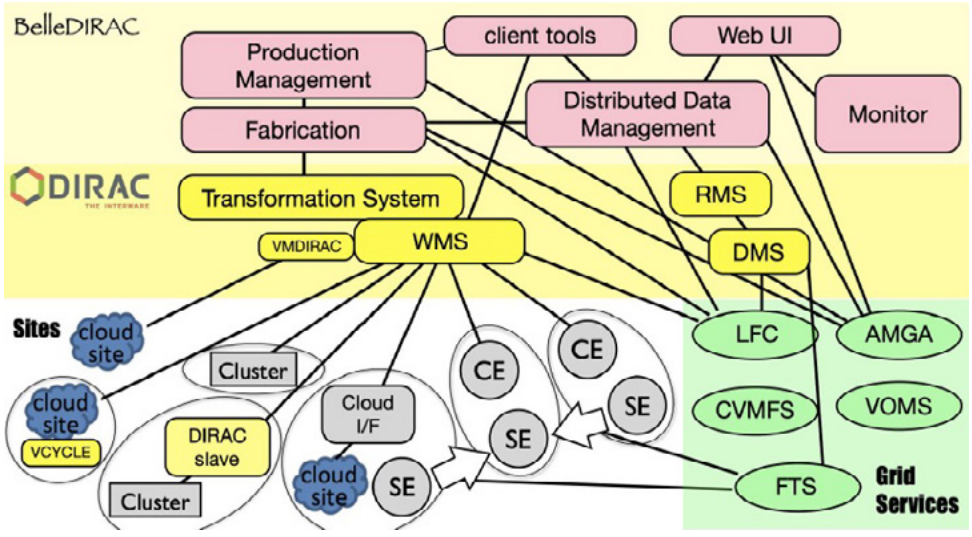


Figure 21: BELLE II DIRAC operations

Belle II uses an extension of the DIRAC framework, called BelleDIRAC, to manage its distributed computing needs (see **Figure 21**). DIRAC allows a distributed infrastructure that Belle II utilizes. While most systems run at KEKCC, several run at BNL, the most important of which are the Distributed DMS and the FTS service. In September 2020, the file catalog (currently the WLCG File Catalog as shown in the diagram below) migrated to using Rucio, and this will also operate at BNL. A lot of attention is needed to the production software to ensure that the network latency between KEKCC and BNL does not lead to bottlenecks, e.g., avoiding file-by-file requests to a remote service.

Another important service running at BNL is the Conditions Database (CDB), shown in **Figure 22.** The main goal of the CDB is to support Belle II computing grid data processing and interactive users. In addition, as part of the detector's calibration cycle, the CDB serves as persistent storage for selected data collected by the Belle II DAQ at KEK which is properly formatted for CDB injection.

The CDB system is composed of two sub-services: the payload service and the metadata service. The payload service is accessed if the client-side software cannot find a copy of the payload on CVMFS (payloads are migrated to CVMFS with a latency of hours) or on the client's local caches such that the load on the CDB payload service at BNL for reading is typically modest. Writing new calibration data payloads, which can only proceed via the CDB service at BNL, is also typically modest as the CDB service is a mostly read service. The CDB metadata service provides the list of which conditions data payloads are needed for a given data processing job. This service also has a selective backup on CVMFS but, unlike for payloads, the metadata service is the primary source. The data transfer to or from this (metadata) service per query is also modest. Nevertheless, bursts of requests to the CDB system have been observed and successfully supported.

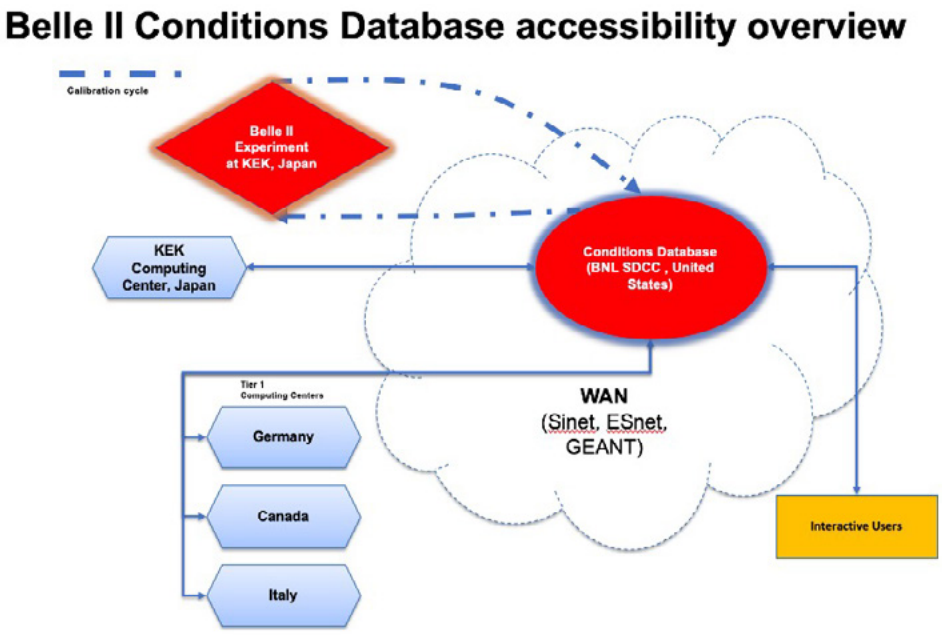


Figure 22: BELLE II conditions database

5.8.2.6 Software Infrastructure

In September 2020, Belle II moved to using Rucio as its high-level data management software, with FTS as the backend and also hosted at BNL. dCache is the primary data storage at BNL.

The legacy data management software, bespoke software based on DIRAC, will be decommissioned, but some features of the underlying DIRAC DMS will continue to be used for low-level operations in the distributed computing environment. These are expected to eventually be replaced by Rucio, in particular for end-user data access.

5.8.2.7 Network and Data Architecture

5.8.2.7.1 Domestic Connectivity

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

The HTSN has five key components:

1. Network Perimeter
    a. Two (soon to be three) diverse 100Gbps circuits that peer with ESnet in New York City. These circuits are utilized by all scientific and administrative communities at BNL. All traffic to and from BNL flows through either of these circuits.
    b. The BNL network perimeter transfers on average 7–8 PB of data monthly, with spikes up to ~12 PB.
2. Science DMZ
    a. Supports open, high-speed WAN/internet access for all scientific collaborations throughout the BNL campus.
3. Science Core
    a. A Tbps Science and Data Center Interconnect for data-intensive collaborations at BNL. This Science Interconnect enables high-speed connectivity between collaborations such as ATLAS, STAR, PHENIX, CAD, CFN, NSLS-II, HPC Clusters and the SDCC.
    b. Intelligence and routing policies are applied within the Science Core to restrict or grant access to specific resources within the SDCC.
4. Spine
    a. A Tbps network Spine that interconnects all Leaf switches. Leaf switches can consist of top of rack (ToR) or chassis-based switches that connect compute, storage, or general infrastructure service servers.
    b. The responsibility of the Spine is fast packet forwarding and flexibility, not policy insertion or server termination.
    c. External Border Gateway Protocol (eBGP) is utilized throughout the HTSN. EBGP was chosen for its ability to immensely scale and to create modularity and fault domain isolation down to the rack level. Each Spine group shares the same Autonomous System Number (ASN) but does not have Internal BGP (iBGP) peering between them. Each Leaf or pair of Leaves will require its own ASN.
5. Storage Core
    a. A redundant terabit per second switching block that aggregates high-performance storage services.

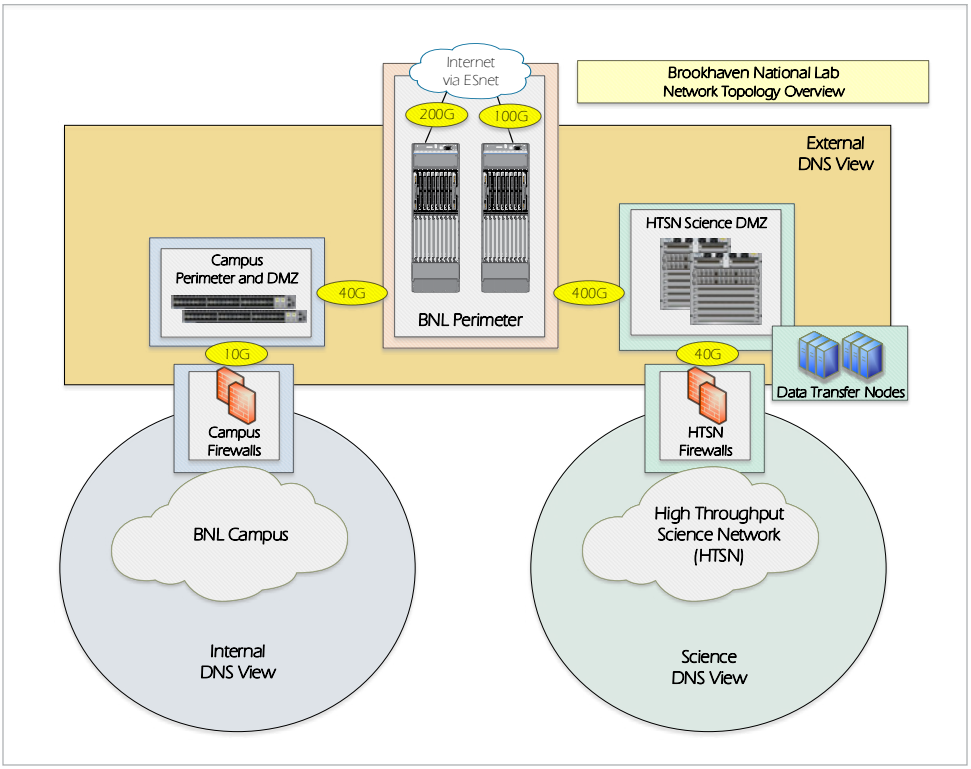


Figure 23: High-level overview of the BNL network perimeter and domain name service (DNS) architecture

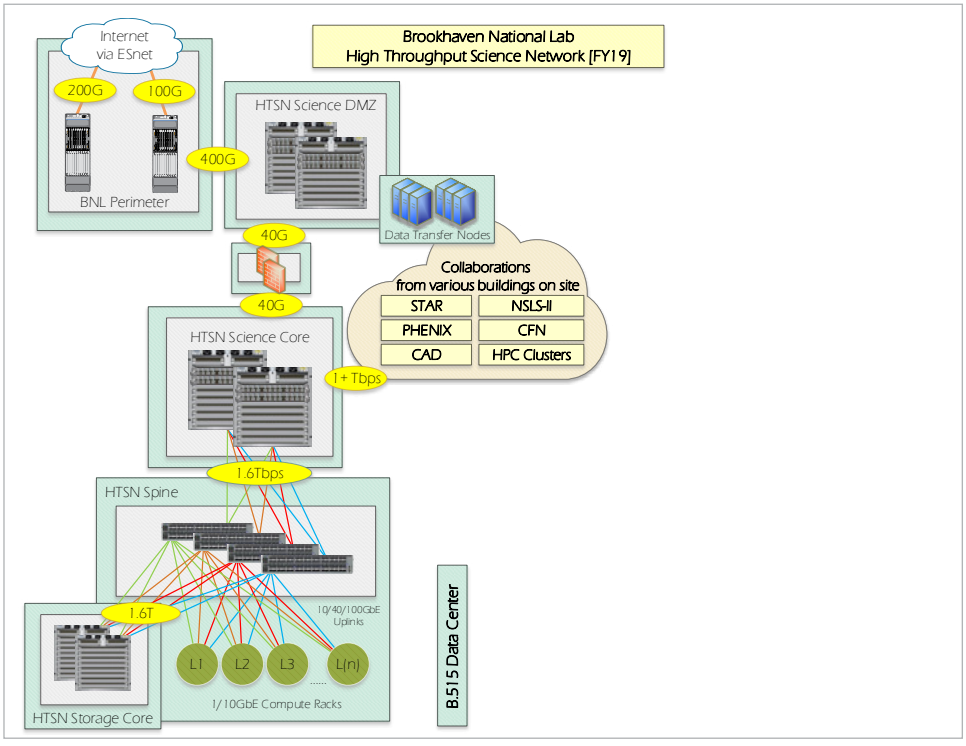


Figure 24: High-level overview of the BNL HTSN (FY19)

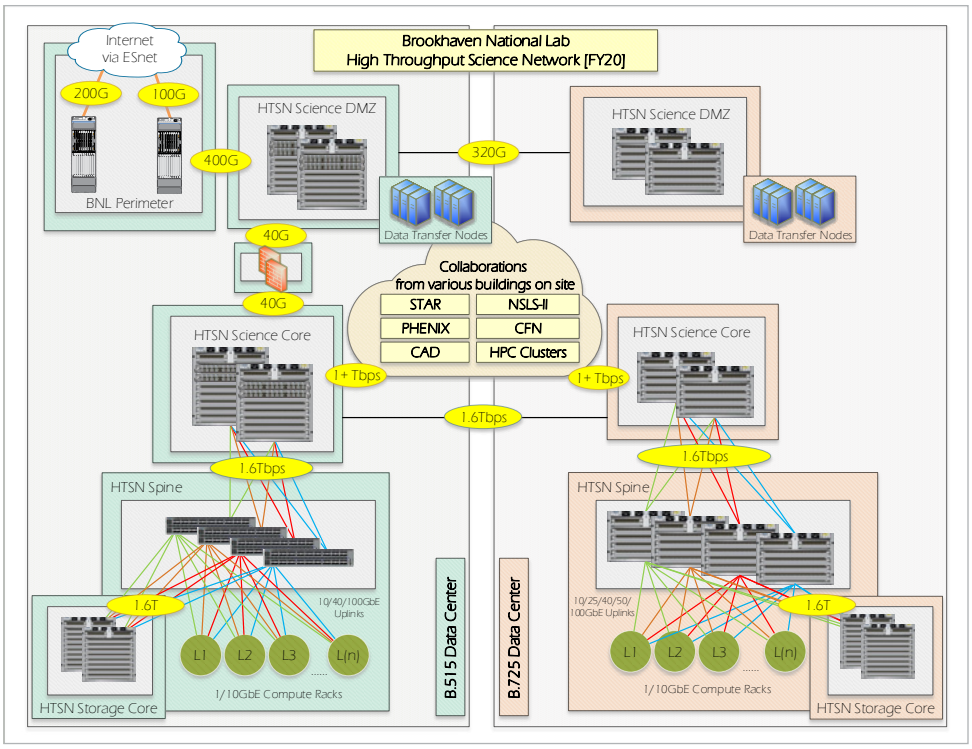


Figure 25: High-level overview of the BNL HTSN in FY20 (includes HTSN expansion into new data center, right hand side)

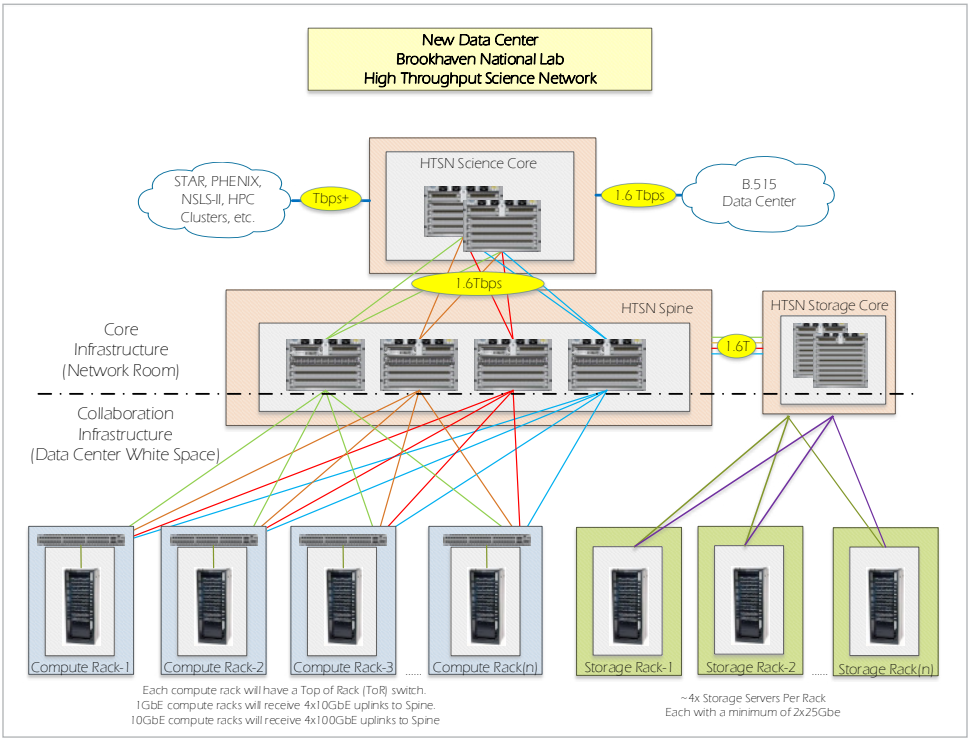


Figure 26: High-level overview of the BNL HTSN, demarcation points between core and collaboration infrastructures

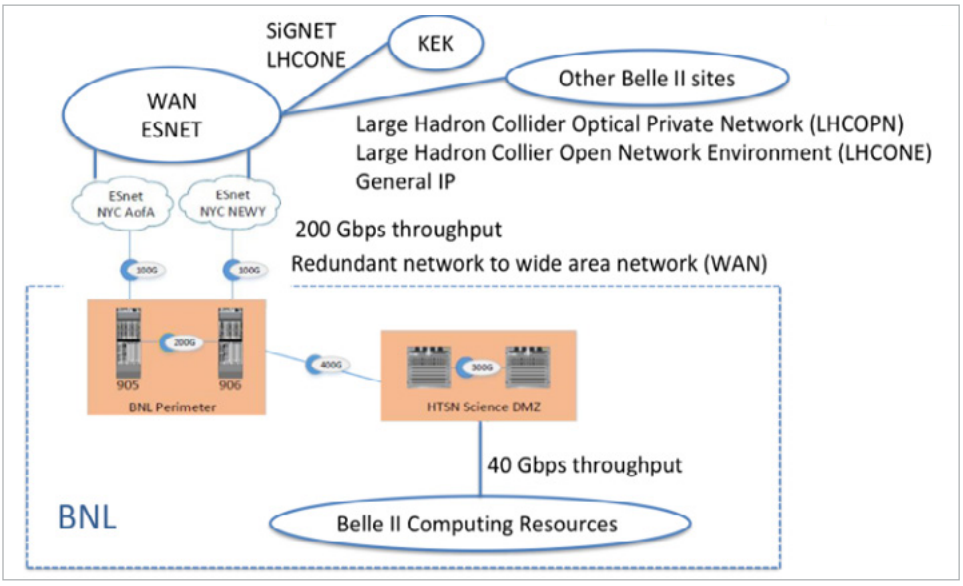


Figure 27: Network diagram Belle II Tier 1 site at BNL as of Aug 2020

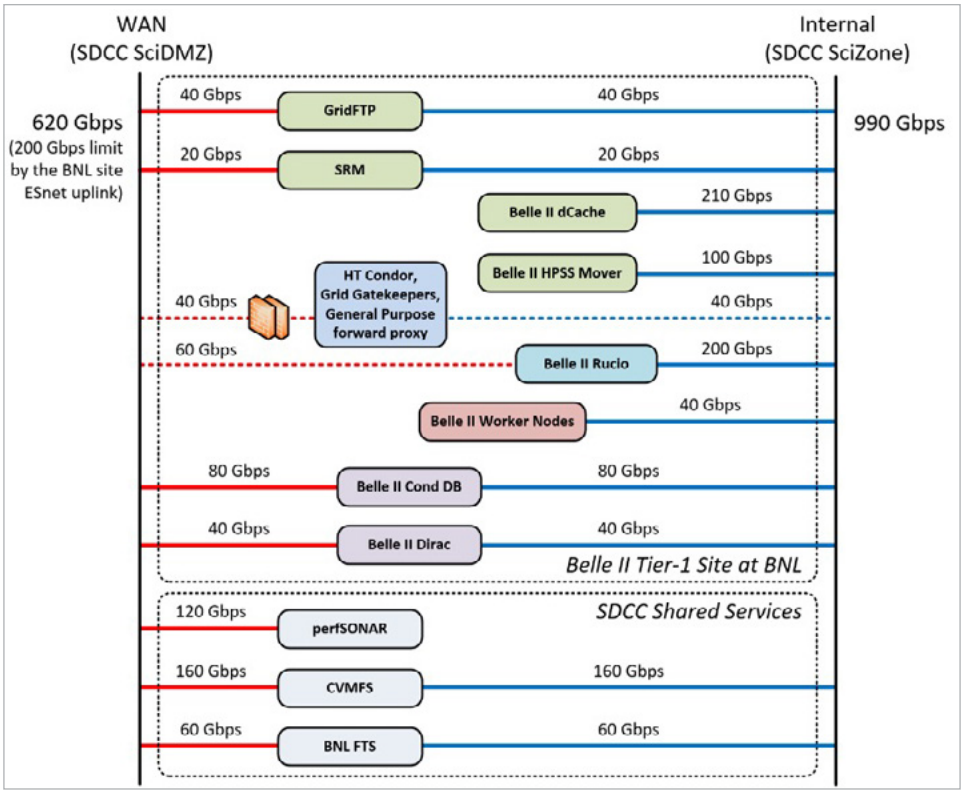


Figure 28: Belle II service capacity (LAN and WAN)

Reliability and performances of network between main data centers are monitored with perfSONAR[162].

[162] http://maddash.aglt2.org/maddash-webui/index.cgi?dashboard=Belle%20II%20Mesh%20Config

Evolution of network capacity requirements between BNL and Japan and Europe are illustrated in **Figure 29.**

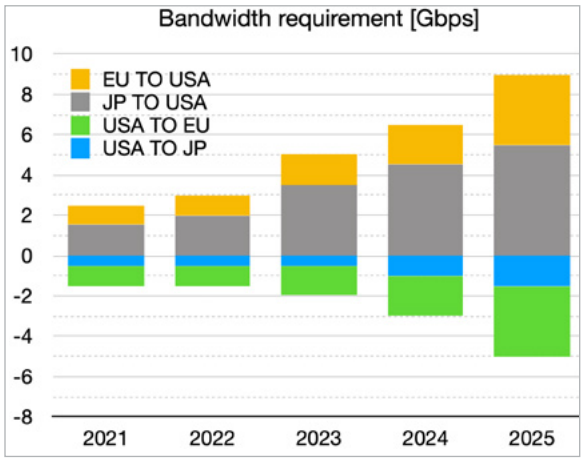


Figure 29: International bandwidth requirements for Belle II

5.8.2.7.2 International Connectivity

The success of Belle II relies on transpacific networking capacity provided by a number of partners that participate in the LHCONE overlay network. Belle II relies heavily on network access, thus the partnerships that underpin LHCONE are critical for ongoing operations. LHCONE delivers service using physical resources operated by a number of different entities:

- Belle II is connected to SINET[163], a Japanese academic backbone network that works in collaboration with APAN[164].
- SINET is connected to the United States via multiple direct and indirect paths:
  - 100 Gbps is available via the NSF-funded PacificWave (PacWave)[165] project to Los Angeles, California.
  - 100 Gbps is available via the NSF-funded TransPac link, operated by Indiana University International Networks[166] to Seattle, Washington.
  - 100 Gbps via the National Institute of Communications Technology R&D testbed network: JGN[167] that connects to the United States via Hong Kong and Guam. This connectivity is operated in partnership with the PIREN project[168] and TransPac.
  - 100 Gbps to Singapore operated with JGN, where alternate peering arrangements can function as backups to carry traffic to the United States via Guam or Australia.
- The PacWave exchange points in the United States are located in Seattle ,Washington; Los Angeles, California; and Sunnyvale, California. Domestic networks such as Internet2, ESnet, and CENIC[169], as well as other international partners, maintain peering here.

[163] https://www.sinet.ad.jp
[164] https://apan.net/about
[165] http://pacificwave.net
[166] https://internationalnetworks.iu.edu/initiatives/transpac
[167] https://testbed.nict.go.jp/jgn/english
[168] https://www.hawaii.edu/piren
[169] https://cenic.org

5.8.2.8 Cloud Services

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

5.8.2.9 Data-Related Resource Constraints

A network data challenge completed in February 2020 showed that all centers holding a primary copy of the data (Canada, France, Germany, Italy, KEK, and the United States, see the table in Section 5.8.2.2) could meet or surpass the maximum performance requirement of 42 TB/day rate expected in 2027.

5.8.2.10 Outstanding Issues

Over the last year, there has been much discussion about expanding the LHCONE architecture into Multi-ONE's for new and existing HEP collaborations. Today, BNL already struggles with adhering to the current LHCONE/LHCOPN AUPs and the creation of Multi-ONE's would only further increase the CAPEX and OPEX complexities at end sites.

BNL hosts numerous data-intensive collaborations from different parts of the DOE complex: Basic Energy Sciences, Biological and Environmental Research, HEP, and Nuclear Physics. Architecting and implementing a centralized, fault tolerant, scalable, terabit per second, unified HTSN infrastructure/service has been challenging on many levels. One of the most complex areas in operating this type of network infrastructure is the adherence to the LHCOPN/LHCONE AUPs. In a multipurpose lab, utilizing a unified network perimeter becomes exponentially complex as scientific programs want exclusivity over a VPLS or L3VPN circuits while utilizing BGP (e.g., LHCONE, LHCOPN or the possibility of a Multi-One deployment). Some of the complexities that are introduced are:

- Implementation of separate routing tables.
- Policy-based routing (PBR, source-based routing).
- Asymmetric routing.
- Increase cost in network perimeter routers.
- Increases operational complexities (2 a.m. rule goes out the window).
- Limits the number of vendors to choose from.

Adhering to the current LHCOPN/LHCONE AUPs while implementing new resources such as Belle II or other new programs creates a multitude of complexities such as:

- IP addressing:
  - Collaborations cannot share existing IPv4/IPv6 address space at the end site. This leads to each collaboration procuring its own dedicated direct assignment address space for both IPv4 and IPv6.
  - This complexity also leads to system administrators needing to follow a complex matrix to determine which subnet their DTNs should be assigned to.
- Creation of multiple virtual routing and forwarding (VRF) instances:
  - Increases operational complexity by requiring additional EBGP, iBGP, and interior gateway protocol peerings.
  - Requires PBR (source-based routing) to make sure data are transported into the correct VRF instance.
  - Time to deploy is increased because of operational complexity.
  - Locks end sites into very expensive networking hardware.

- Use point solutions per scientific collaboration:
  - Provide dedicated equipment, circuits, and peerings.
  - Add significant costs for procurement, maintenance, and management.
  - Longer time to deploy.
  - Move in the opposite direction of a unified network to reduce costs.
  - Who is supposed to pay for this?

At what point does the cost and complexity of these solutions no longer justify strict compliance to these AUPs? These AUPs are forcing BNL to migrate away from an "end site" architecture and begin to mimic a service provider environment. Given the fiscal climate and the growth of additional scientific programs at BNL, these AUPs and the creation of Multi-ONEs will only continue to increase operational complexities and drive hardware costs higher. BNL's aspiration is to remove the complexities of source-based routing and revert to destination-based routing, which will eliminate operation complexity and reduce costs.

5.8.2.11 Case Study Contributors

Belle II Representation

- Eric Lancon[170], BNL
- David Jaffe[171], BNL
- Paul Laycock[172], BNL
- Alexandr Zaytsev[173], BNL

ESnet Site Coordinator Committee Representation

- Vincent Bonafede[174], BNL
- Mark Lukasczyk[175], BNL

5.9 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. The two experiments profiled are examples of this: the SBN and DUNE.

Both focus on the study of neutrino oscillations, and use a similar set of scientific technology for observation. It is also expected that both experiments will share similar approaches to their implementations of computational and storage infrastructures at Fermilab, and their approach to wide area networking. 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 SURF 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 wide-area network 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.

[170] elancon@bnl.gov
[171] djaffe@bnl.gov
[172] laycock@bnl.gov
[173] alezayt@bnl.gov
[174] bonafede@bnl.gov
[175] mlukasczyk@bnl.gov

5.9.1 Discussion Summary

The following discussion points were extracted from the case study and virtual meetings with the case study authors. These are presented as a summary of the entire case study, but do not represent the entire spectrum of challenges, opportunities, or solutions:

- The SBN Program
  - The program will rely on a chain of three particle detectors: ICARUS, MicroBooNE, and 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.
  - The program will measure and inform behaviors that will influence the future long-baseline neutrino experiment DUNE.
  - The program's event data are dominated by the data from the detector instrumentation: each has the ability to sample and record behaviors during events. All of 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 roughly for a 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 to 7PB 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 to 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).
- DUNE
  - DUNE is an international neutrino experiment that will be conducted with the international LBNF at Fermilab and SURF.
  - DUNE is under active design and construction, and will come online in the late 2020s (estimated to be 2025 or 2026). Two 5% scale prototypes, ProtoDUNE-SP and dual phase, have run at CERN and plan a second run in 2022.
  - DUNE operation will span several US states: the beam will originate at Fermilab in Illinois using a pulse rate of approximately once ever second, and operate 24 hours per day. During operating periods, this will result in approximately 15 million pulses per year. The far end in at SURF in South Dakota will house the detection equipment. Due to the rarity of neutrino interactions that are anticipated, the design team has planned for the ability to handle 7,500 interactions per year. Given the rarity of scientific observation, the accumulated data during these limited number of interactions makes the computational, storage, and networking technology linking the facilities critical for experimental success.
  - 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 approximately 13 PB of data per year per module, with the project expecting to retain 30 PB of this per year on Fermilab storage. 30 PB/year on the wide-area network is an average of 1.3 GB/sec, less than the rates already demonstrated for protoDUNE acquisition and storage.
- Supernova candidates pose a unique problem for data acquisition and reconstruction. The supernova triggers involve short, very large bursts of data 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 be on the order of 200 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, putting an extreme need for reliable and predictable networking.
- 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 to the overall success of the workflow. Data volumes could reach or exceed 30 PB per year.
- The operations of the DUNE experiment will proceed in three major phases. ProtoDUNE at CERN will run again between 2021 and 2022, and will continue active data analysis until 2025. Installation and commissioning of the far and near detectors in South Dakota and at Fermilab will occur over the period between 2025 and 2029. Physics operations running with both the near and far detectors will occur between 2028 and 2040, and possibly beyond.
- The continued operation of ProtoDUNE will require transatlantic networking capabilities due to the location at CERN, and use of Fermilab resources. Data rates of between 2 and 3 GB/s are possible during runs. A total of 2–10 PB of data will be generated during the 2021–2022 operational phase.
- DUNE reconstruction and analysis will be a constant process during operation, with the majority of 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 workflow will also be grid-based and look similar to 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. By late in the 2020s we anticipate a need to sink 100 GBs from the SURF site to Fermilab and redistribute those

data simultaneously to sites worldwide. This will require improvements in the SURF to Fermilab link.

- For rapid response data such as supernova data, it would be useful to have both a higher network QoS as well as a faster storage QoS.

5.9.2 SBN Case Study

5.9.2.1 Background

The SBN Program[176] will study properties of neutrinos and neutrino interactions using beams of accelerator-generated neutrinos at Fermilab. In total detectors — the SBND, MicroBooNE, and the short-baseline far detector (ICARUS T600) — will detect neutrinos from Fermilab's Booster Neutrino Beamline (BNB), with the major goal of conclusively identifying or ruling out anomalous neutrino oscillations hinted at by many other measurements. The detectors are all Liquid Argon TPCs (LArTPCs), which combine great position and calorimetric resolution to enable precise event reconstruction and particle identification. In addition to searching or ruling out "new physics" in neutrino oscillations, the SBN detectors will measure important properties of neutrino interactions on argon that will be critical for the future long-baseline neutrino experiment DUNE.

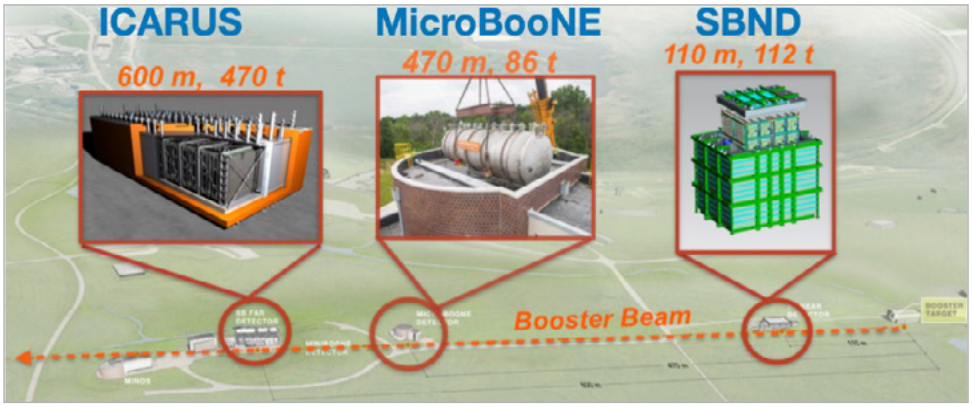


Figure 30: Layout of the SBN Program detectors along the BNB at Fermilab

All three detectors are located on Fermilab's campus at distances of a few hundred meters apart, and record interaction events at an approximately 1 Hz rate. Event data are dominated by the data from the TPCs themselves: each detector has thousands to tens of thousands of sensing wires that sample and record charge signals over a period of milliseconds for each interaction event. In each detector, the TPC data are combined by the DAQ with data from other detector systems that record scintillation light signals and the passage of cosmic rays through the detectors, and then written to a local computing cluster at each detector building site. The complete raw data is then transferred to Fermilab's central computing, where it is staged and then written to archival tape.

Raw data must then be processed for analysis. A simple view of the processing chain is as follows. Raw data go through a signal processing and hit-finding stage to filter out noise, deconvolve electronics and electric field response, and identify areas of detected charge. Then, higher-level reconstruction algorithms perform pattern recognition and image analysis, combining data across all views of the detector to reconstruct 3D charged particle tracks and electromagnetic showers and perform calorimetric measurements. Finally, those higher-level objects are used in analysis to identify neutrino interactions and reconstruct their topologies and final-state particles.

Collected data are typically compared to simulations to assess reconstruction efficiencies and systematic uncertainties, and ultimately to test predicted models of neutrino interactions. Simulated data go through

[176] SBN Proposal, https://arxiv.org/abs/1503.01520

the same processing chain as described previously, but first must go through a generation process to produce equivalent raw data. Neutrino interactions are modeled using calculations available in software packages, particle interactions in matter are then modeled using GEANT4, and then detector modeling is applied through custom-written software. Because each detector is located near the surface of the earth, cosmic-ray backgrounds are a constant presence: they can be simulated through software, but for final analyses the detectors will take data events taken out-of-time with the beam which offers a random snapshot of activity induced by cosmic rays in the detector, and then neutrino interaction simulations will be overlaid on top of this raw data. A library of "out-of-time" data events will be collected to allow for this overlay simulation.

The SBN detectors will detect millions of neutrino interactions, will collect many more events that are simply cosmic-ray-induced background interactions, and will also need to simulate tens of millions of neutrino interactions to achieve SBN's physics goals. Typically, simulation and the two stages of reconstruction (signal processing and 3D reconstruction) are run centrally by the experiment data production teams, and produce files containing 3D reconstruction information with a size on the order of 50 kB per event. These files are stored centrally at Fermilab, then accessed by analyzers across the collaboration to perform the final physics measurements.

As improvements in the simulation and reconstruction are always ongoing, the SBN experiments plan 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. Simulation and signal-processing outputs will be produced roughly once per year, while 3D reconstruction outputs will be produced roughly twice per year (using the previous stages as inputs). The output of 3D reconstruction will then be accessed through the year for the performance of physics analysis. SBN currently has a data lifetime model of saving the bulk of derived data for two years, and the much smaller analysis data used for publications permanently.

5.9.2.2 Collaborators

The SBN international collaborations comprise approximately 250 scientists from more than 50 institutions, and include members from across the United States, CERN, Italy, the UK, Switzerland, Spain, Turkey, and Brazil. Member institutions critically include both scientific laboratories and universities. While the center of activity is typically at Fermilab, where many collaborators (particularly young researchers) locate themselves for some period of time, analysis and data activities span these institutions, with significant/large clusters of researchers and activity centered around the UK, CERN, and Italy.

Typically, the collaborations have relied on the computing infrastructure at the host lab (Fermilab) for data storage and the bulk of computing. Dedicated backups of raw data are planned to be sent and stored at the INFN CNAF facility in Bologna, Italy, and investigations on a possible additional backup location will be underway soon. Computing resources at institutions have been and will continue to be added to be used in central processing much like sites are added to the OSG: those sites will be used to augment computing resources as much as possible, and will likely be used for dedicated purposes on top of general computing (for example, the generation of dedicated or specialized data sets).

CNAF will host ICARUS data backups, which are expected to be roughly 1.2 PB per year. These will be collected and sent on a continuous basis during beam data-taking periods, which generally occur during a nine-month continued period of operation during the year. Transfers will be managed via Rucio subscriptions, and be validated through checksums. While data should be transferred in a timely fashion (to avoid a pileup of data needing to be backed up), there is not currently a strict latency requirement.

User/collaborator and location	Primary or secondary copy of the data	Data access method	Avg. size of data set	Frequency of data transfer or download	Are data sent back to the source and method	Any known issues with data sharing
1) ICARUS (SBN FAR DETECTOR)	No	Create/transfer to Fermilab storage	~3.5 PB/year	Continuous	No	
2) SBND	No	Create/transfer to and from Fermilab	~ 3 PB/year	Continuous	No	
3) FERMILAB DATA CENTER	Yes	Ingest from 1 to 2 and streaming transfer out for prompt processing on Fermigrid/OSG/WLCG	Up to 6.5 PB/year (most done on Fermigrid)	Continuous	Yes	Ingest
4) FERMILAB DATA CENTER	Yes	Transfer out to secondary tape/disk archives (e.g., infn cnaf)	~2 PB/year	Continuous	No	Egress from Fermilab
5) FERMILAB DATA CENTER / US HPC SITES	No	Transfer to/from HPC facilities	Up to 6.5 PB (still under development)	~2–4 times per year	Yes	
(6) FERMILAB DATA CENTER OSG/COMPUTING SITES	No	Transfer to/from osg sites in large-scale production	~2.5 PB	~twice yearly	Yes	
(7) COLLABORATING INSTITUTIONS	No	Generally small data sets received data streaming/transfer, largely for analysis development work	Up to ~4 PB/year, roughly ~1-100 GB at a time	Ad hoc	Yes	

Table 12: SBN data projections

5.9.2.3 Instruments and Facilities

Present to Two Years:

The MicroBooNE detector is currently taking and storing data at Fermilab. While its future data-taking plans are still unknown, the data it has taken will continue to be hosted and stored centrally at Fermilab, with processing of that data being largely focused there.

The ICARUS and SBND detectors ("far" and "near" detectors, respectively) will be coming online and begin data-taking in the next two years. The current expected raw data rate from the detectors is 3.3 PB and 2.8 PB, respectively. Raw data will be stored at Fermilab, with a backup for a significant portion (roughly 20%) of beam-on ICARUS data planned to be stored at CNAF. For large-scale processing, Fermilab computing and OSG sites will be used.

Two to Five Years:

The ICARUS and SBND detectors will execute the bulk of their data-taking and analysis during this period. In addition to use of Fermilab central computing, SBN will onboard computing at CNAF, CERN, and the UK to further facilitate general production and analysis work. SBN also plans to make use of HPC facilities like Theta at

ANL and Cori at NERSC/LBNL to further enable general production processing (e.g., event simulation, signal processing, and event reconstruction at Theta) and final-fits/sensitivity estimation (e.g., Feldman-Cousins analysis at NERSC). There is no major change expected in the detector infrastructure or basic computing envisioned, though there are hopes to continue to evolve the ability to rely less on tape for data storage, and have a greater availability of disk for data that has an expected lifetime less than a few years (e.g., much of the bulky derived data from simulation and first-stage reconstruction output).

Beyond Five Years:

The SBN Program will likely be finished with data-taking at this timeline, and be in the final stages of physics exploitation. This will hopefully reduce significant computing load for taking new data, but larger scale production and analysis are likely to continue for some time in this period. It is unlikely that significant changes in data sources or workflows will occur in this time.

5.9.2.4 Process of Science

The general process of science is described previously in Section 5.9.2.1. Data are generated either from the detector ("raw data") or are simulated. The former is managed by the online DAQ and data management team in the collaboration, while the latter is managed by the offline production team. Data then proceed through two stages of reconstruction (signal processing and 3D reconstruction), managed by the offline production team. The output of the reconstruction includes slimmed analysis tuples that will be read by analyzers across the collaboration.

Data reduction is largely accomplished through the reconstruction processing. After signal processing, algorithms to identify regions of interest are employed to reduce the search region for hit reconstruction in waveforms. For example, ICARUS data are expected to be ~140 MB per event at the raw data stage, but then only 40 MB / event after first-stage reconstruction. The second stage of reconstruction will increase that size by approximately a factor of two (largely to hold detailed charged particle track trajectories), but then high-level reconstruction objects can be summarized to then create a final analysis tuple with the goal of a size of ~50 kB per event.

Present to Two Years:

For normal data-taking from the detector, raw data will arrive from ICARUS at a steady rate of about 370 GB/hr (~100 MB/s), and from SBND at about 320 GB/hr (~90 MB/s). These data will be delivered to Fermilab dCache and Enstore tape storage on-site, and be made available for immediate processing which will, for the most part, access data via XROOTD to run on-site at Fermigrid, but can be accessed and run on OSG sites as available. For immediate processing, data should still be in disk cache and so not need to be accessed from tape. Raw data backups will be ~75 GB/hr (~20 MB/s) to CNAF for ICARUS, and 130 GB/hr (~35 MB/s) for SBND (location not yet identified). During commissioning periods, the overall data rate from the detector can be much larger, but generally these data are not stored permanently and not all of them processed, so the previous rates remain reasonable expectations.

For reconstruction processing of data, ICARUS computing jobs will need to have data delivered at a consistent average rate of up to 2.35 TB/hr to satisfy demands during data production campaigns (~60 days in length). Data write rates back to Fermilab for storage will be ~1 TB/hr. Data will need to be delivered to wherever that production is running: the expectation is this will be to largely Fermilab computing resources, but will include some fraction of jobs to available OSG sites across the United States and Europe, and to CNAF computing in Italy. For SBND, these numbers will peak at approximately 1 TB/hr for data reading and ~0.8 TB/hr for data write-back in this similar two-year period.

In this two-year period, ICARUS simulation should begin using off-beam data events for its base simulation (see sec 1.2), which will require a data delivery rate of ~0.7 TB/hr to computing nodes over a production campaign period (~60 days), with a similar rate for data to write back to send back to Fermilab storage. SBND may not be

doing these overlay simulations as a default at this point, and so would see a much-reduced need for reading data when starting simulation generation; however, data write-back rates will still be large: approximately 1.25 TB/hr during a production campaign period.

ICARUS and SBND production campaigns are likely to overlap with each other in time, and will typically include data reconstruction processing alongside simulation. ICARUS and SBND production campaigns are likely to overlap with each other in time, and will typically include data reconstruction processing alongside simulation. Thus, the total data delivery rates coming out from Fermilab data storage for the porous of reconstruction will achieve a sustained rate of 4 TB/hr over a period of 60 days, and with simulation data being written back to storage at a rate of ~3.8 TB/hr over a similar period. These estimates assume that the storage hardware can keep up with the computational and network demands.

While the collaboration will push for most analysis to use the slimmed analysis tuples (hosted at Fermilab, but small and able to be copied easily elsewhere), there will be a need for calibration and event selection tasks to access reconstruction data sets. These will be several 100 TB data sets staged from central storage locations and delivered to computing nodes at Fermilab or OSG sites.

Two to Five Years:

Generally, the situation for the present to two years is repeated for the two- to five-year case. Detector operations/data acquisition will continue at the steady rates presented there.

General needs for production processing will increase year-to-year in a generally stable way: data reconstruction and simulation (and its corresponding reconstruction) scale with the total amount of data taken/neutrino interactions collected. The expectation is that data IO rates will not be network bound, but rather tape/disk I/O limited, and so the peak rates described previously should not increase by more than a factor of two, and can be offset by longer production campaign timelines (e.g., moving from 60 to 90 days)

In this period, slimmed analysis tuples should be almost exclusively the files that most collaboration data analyzers will access (necessary to keep total data I/O rates under control).

New features in this period related to networking are related to further use of HPC centers, like Theta at ANL, for performing simulation and data reconstruction general processing. Development of these workflows is still in R&D, but it is possible that significant fractions of the data inputs for simulation and reconstruction could be sent to Theta for more efficient processing. Following the numbers provided previously, this could amount to a sustained average rate of ~4 TB/hr to and from that facility over a production campaign period (~60 days).

Beyond Five Years:

At this time period, data acquisition will stop, and this time period will see general production processing slow down as the reconstruction and simulation reaches its final state. Most activity will be centered around analysis, which will depend on many analyzers accessing small analysis tuples.

5.9.2.5 Remote Science Activities

Please see Section 5.9.2.4 for much of this information. In addition to the information presented there, an additional remote science activity is to look at "event displays" of the data for particular events. This involves making available raw, noise-filtered, and/or deconvolved data waveforms to remote users on an event-by-event basis, for further study and development of reconstruction and analysis algorithms. The software for doing this will be available via a docker container, and access to data through XROOTD will be from either Fermilab (most cases, and the default) or from CNAF or other remote sites (in the case the data are known to be at those sites). The overall rate of data for this will be small and user-driven (so highly periodic), but having data delivered with as minimal latency as possible will be important to aid analysis development.

5.9.3 DUNE Case Study

5.9.3.1 Background

5.9.3.1.1 DUNE Science Background

DUNE will begin running in the late 2020s. The goals of the experiment include (1) studying neutrino oscillations using a beam of neutrinos originating from Fermilab in Illinois and directed at the Homestake mine in Lead, South Dakota, (2) studying astrophysical neutrino sources and rare processes, and (3) understanding the physics of neutrino interactions in matter.

The neutrino beam from Fermilab will consist almost entirely of muon-type neutrinos when produced. Neutrinos are known to come in (at least) three flavors, which can be distinguished by their interactions: electron type neutrinos produce electrons when they interact via charged currents, muon-type neutrinos produce muons, and tau type neutrinos, tau particles. But these flavors do not correspond to fixed mass states as is represented by the illustration in **Figure 31.** All three flavors of neutrinos are mixtures of mass states, much as light in the x direction can be considered a superposition of x' and y' polarizations along alternate axes rotated by 45 degrees. When a neutrino propagates through space, it is the mass state that sets its wavelength, and if the neutrino goes far enough, the multiple mass states corresponding to the initial flavor state will get out of phase. When the mixture is later probed for its flavor, it may give a different answer than the neutrino that started out initially. This phenomenon is neutrino oscillation and has been shown to exist in multiple experiments since it was first confirmed in 1998 by the SuperKamiokande experiment in Japan.

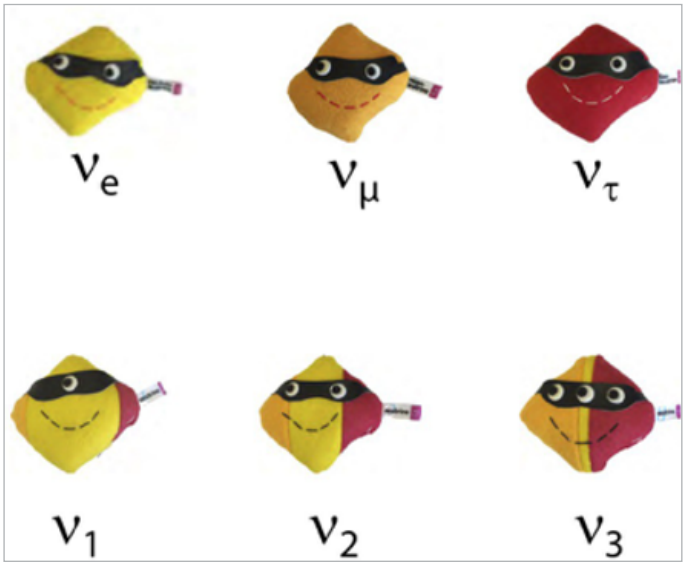


Figure 31: Illustration of the neutrino flavor and mass states. The mass states are a superposition of the flavor states. Courtesy the particlezoo.net.

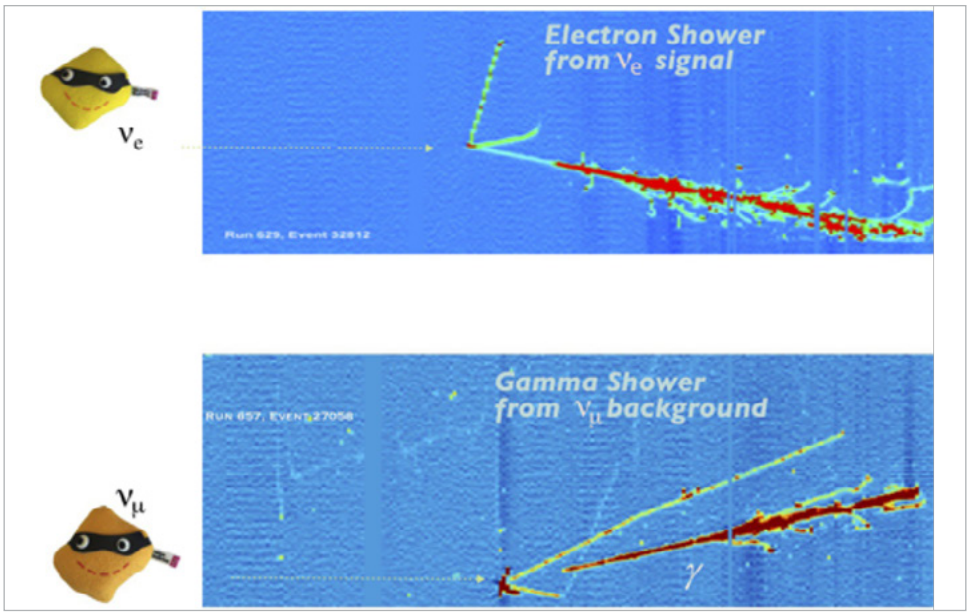


Figure 32: Event displays of electron neutrino appearance signal (top) and background (bottom) as seen in the ArgoNeut experiment[177]. In the appearance signal, an electron is seen emerging from the primary neutrino vertex and then showering. In the background interaction, a muon neutrino produces a final-state muon along with a photon, which propagates some distance before showering.

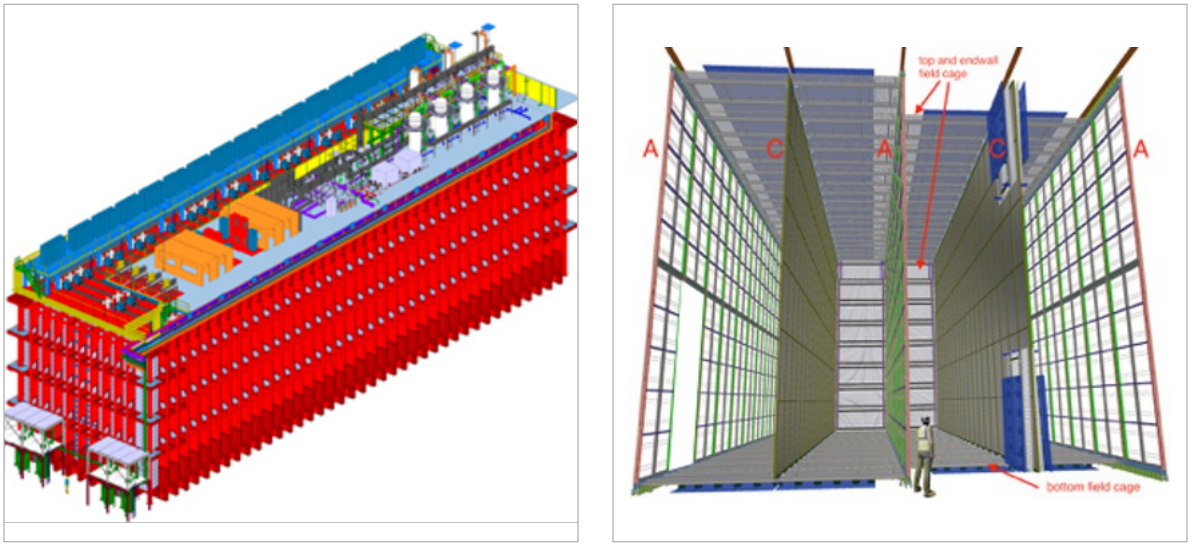


Figure 33: (Left) A far detector cryostat that houses a 10 kT far detector module. (Right) A 10kt DUNE Far-Detector SP module, showing the alternating 58 m long (into the page), 12 m high anode (A) and cathode (C) planes, as well as the field cage that surrounds the drift regions between the anode and cathode planes. The blank area on the left side was added to show the profile of a single anode plane assembly (APA). **Figure 33 (right)** of a person indicates the scale.

DUNE, in particular, wishes to understand the conversion of muon neutrinos created in Illinois into electron neutrinos using a far detector in the Homestake mine in South Dakota and compare that conversion rate between neutrino and anti-neutrino beams. The location of the far detector and energy of the neutrino beam were chosen to maximize the oscillation effect. A difference in the conversion rate for neutrinos and anti-neutrinos could be evidence for matter-anti-matter asymmetry in the neutrino sector, a phenomenon called CP violation.

[177] R. Acciarri et al. (ArgoNeuT), Phys. Rev. D95, 072005 (2017), 1610.04102

To make these measurements, it is necessary to distinguish electron neutrino interactions appearing in the muon neutrino beam from the dominant muon neutrino interactions one would expect in the absence of oscillations. Doing this requires not only a very large detector, as neutrino interactions are intrinsically rare, but an extremely fine-grained one as well. Noble liquid time projection detectors, which read out large transparent volumes of liquid by drifting electrons from interactions to charge detectors through strong electric fields, have the needed capabilities of extremely large scale and fine-grained resolution. The proposed DUNE far-site detector will instrument four 14 x 12 x 58 m3 volumes of liquid argon with readout granularity of 0.5 cm. The detectors will be located 4,850 ft below the surface to reduce the rate of cosmic rays traversing the detector by orders of magnitude compared with a detector on the surface. This low cosmic-ray rate will allow sensitivity to very low energy solar and astrophysical neutrinos as well as the higher energy neutrinos produced at Fermilab.

The neutrino beam from Fermilab will be pulsed approximately once per second, 24 hrs per day during running periods with approximately 15 million pulses per year. Because neutrinos interact extremely rarely, the expectation is to detect on the order of 7,500 neutrino interactions/year in each of four 10 kT detector modules located at the far detector site in South Dakota.

Construction of the detector halls and infrastructure for the large 10 kT fiducial volume far detector modules is starting now, as are design and construction of detector readout modules. A full TDR for the far detector program has recently been completed and is available in the references[178]. The DUNE neutrino oscillation experiment will receive first beam late in this decade with commissioning of the DAQs for the first far detector module expected to start in 2025–2026.

5.9.3.1.2 ProtoDUNE Tests at CERN

Building an experiment of this size requires an extensive period of prototyping. The Argoneut[179], MicroBooNE[180], and ICARUS[181] collaborations have demonstrated the capabilities of large liquid argon time projection chambers (LArTPCs) for neutrino detection on scales between 1- and 500-ton fiducial mass. In preparation for the DUNE experiment, a campaign testing proposed DUNE components in 700-ton detectors in the EHN1 hadronic test beam was launched at CERN in 2018. Both SP and DP prototypes were constructed and tested.

5.9.3.1.2.1 ProtoDUNE-SP

The ProtoDUNE-SP experiment began taking data at CERN in late 2018. ProtoDUNE-SP uses SP LArTPC technology where ionization electrons are collected directly from the liquid argon. The readout system consists of APAs that each have three layers of wires arranged in different directions. Each layer contains 800–1,200 wires spaced 0.5 cm apart. Electrons drift from the original interaction in the argon, through a strong electric field, to the wire planes and induce signals on the wires. The location in the plane of hit wires gives one coordinate of ionized electrons, and the time the signal takes to drift to the wire from the original interaction measures a second coordinate. The third coordinate is derived by combining information from overlaps of signals in the three different wire layers. Signals are amplified electronically and then digitized. **Figure 34** illustrates the operation of a generic LArTPC.

[178] B. Abi et al. (DUNE) (2020), 2002.02967, 2002.03005, 2002.03008, 2002.03010
[179] R. Acciarri et al. (ArgoNeuT), Phys. Rev. D99, 012002 (2019), 1810.06502
[180] R. Acciarri et al., Journal of Instrumentation 12, P02017 (2017)
[181] F. Varanini (ICARUS), EPJ Web Conf. 164, 07017 (2017)

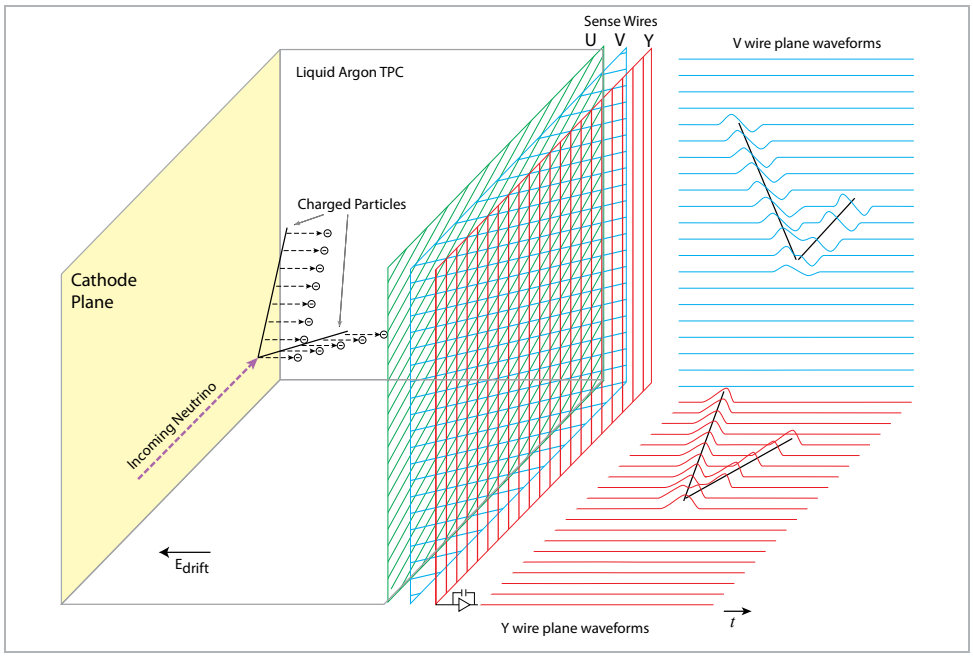


Figure 34: Diagram from[182] illustrating the signal formation in a LArTPC with three wire planes[183]. For simplicity, the signal in the first U induction plane is omitted in the illustration.

The ProtoDUNE-SP detector consists of a 700-ton volume of liquid argon with a cathode plane in the center and three APAs mounted on both edges of the liquid volume. The drift distance is 3 m with a nominal voltage of 180kV across that distance. Each APA has 2,560 channels and each channel reads out a 12-bit ADC every 0.5 μsec. For ProtoDUNE-SP, the readout time appropriate for a 3 m drift was set to 3 msec, resulting in 6,000 12-bit samples per channel. The total data size for 6 APAs is thus 140 MB with additional header and data from photon and external tagging systems bringing the nominal event size up to approximately 180 MB. For part of the ProtoDUNE-SP data taking, lossless compression of the TPC readout was implemented in the data acquisition system, resulting in a final compressed event size of approximately 75 MB.

Data Acquisition

Prior to the beam run, several data challenges were performed with simulated data. Data were shipped from the DAQ machines to data centers at CERN and Fermilab. Rates of over 2 GB/s were achieved in these tests and in the actual data taking.

The test beam delivered particle bunches at rates of up to 25 Hz over a period of six weeks in late 2018 with beam momenta between 0.5 and 7 GeV/c. Time of flight and Cherenkov counters provided beam flavor tagging. Around 8M total physics events were written, with approximately 3M having beam tag information. In total, 570 TB of raw test beam data were written, along with 1 PB of commissioning and cosmic-ray data. The data were written in 8 GB files, each containing 100–130 events.

These data were successfully transferred from the experimental hall to the CERN computing facility, cataloged, and written to tape storage at both CERN (the CERN Advanced STORage manager [CASTOR]) and Fermilab (Enstore tape backed dCache) at rates of up to 2 GB/s via the CERN (and Fermilab) File Transfer Systems. Files were generally available on disk at Fermilab within 30 minutes of being cataloged at CERN, but transfers to tape sometimes took one to two days depending on tape drive ability at Fermilab.

[182] R. Acciarri et al. (MicroBooNE), JINST 12, P08003 (2017), 1705.07341
[183] R. Acciarri et al. (MicroBooNE), JINST 12, P02017 (2017), 1612.05824

Downstream Processing

Thanks to significant prior effort in the liquid argon (LAr) computing and algorithms community, reconstruction software was ready to process ProtoDUNE-SP data, and the first reconstruction pass began soon after data taking started. It was complete within two weeks of the end of data taking. Those results were extremely useful in demonstrating the capabilities of the detector and are summarized in Volume II of the DUNE far detector TDR[4]. A second pass, with improved treatment of instrumental effects ranging from stuck bits to 2-D deconvolution to correction for space charge effects, was completed in late 2019. **Figure 35** shows test beam data after hit reconstruction. A beam interaction and approximately 40 cosmic-ray traces can be seen for this 3 msec exposure.

While LArTPCs benefit from fine granularity and a uniform detector medium, diffusion, argon purity, fluid flow, and the buildup of space charge in the active medium can all introduce distortions into the detector response. These effects have all been simulated and tested in the ProtoDUNE-SP data.

Compressed raw input event records were of order 75 MB in size and took 500–600 seconds to reconstruct, of which approximately 180 sec was signal processing and the remainder high-level reconstruction dominated by 40–60 cosmic rays per readout. Memory footprints for individual processing jobs ranged between 2.5 and 4 GB. Output event record sizes were reduced to 22 MB by dropping the raw waveforms after hit finding. Reconstruction campaigns took on the order of four to six weeks (similar to the original data taking) and utilized up to 15,000 cores on OSG/WLCG resources. Job submission was done through the Production Operations Management Service (POMS)[184] job management system developed at Fermilab. POMS supports submissions to Fermilab dedicated resources and selected OSG and WLCG sites. **Figure 35** shows the site distribution of wall hours used for reconstruction in 2019. CPU estimates are based on typical times on a Fermilab/OSG/WLCG grid core, which average to approximately 12 HS06 units.

For reconstruction, data were streamed via XROOTD[185] from dCache storage at Fermilab to the remote sites. Despite individual processing jobs taking 15–30 hrs to complete, network interruptions rarely caused job failures.

5.9.3.1.2.2 ProtoDUNE-DP

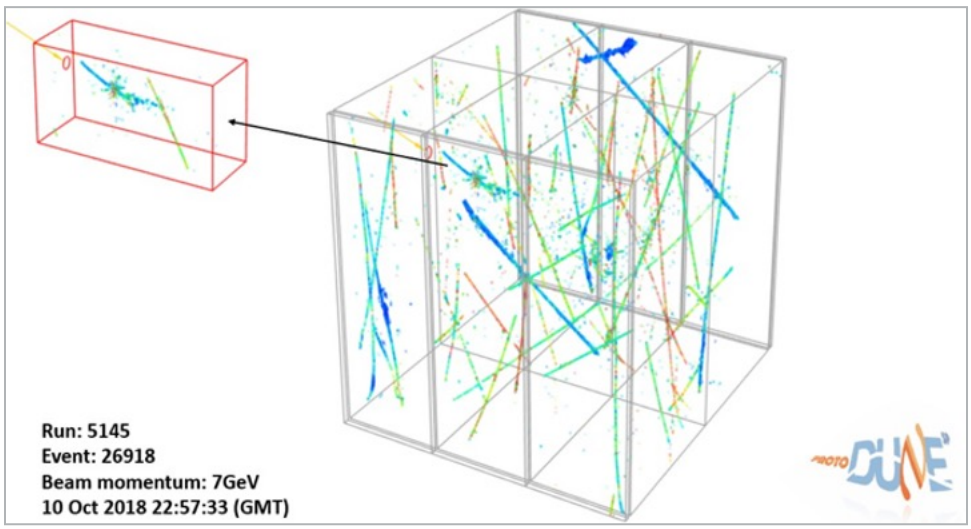


Figure 35: ProtoDUNE-SP detector (gray box) showing the direction of the particle beam (yellow line on the very far left) and the outlines of the six APAs. Cosmic rays can be seen throughout the white box, while the red box highlights the beam region of interest with an interaction of the 7 GeV beam. The 3D points are obtained using the Space Point Solver reconstruction algorithm.

[184] https://cdcvs.fnal.gov/redmine/projects/prod_mgmt_db
[185] G. Behrmann, D. Ozerov, T. Zangerl, J. Phys. Conf. Ser. 331, 052021 (2011)

The ProtoDUNE-DP detector began taking data using cosmic rays in August 2019. Thanks to preceding data challenges, those data have been successfully integrated into the full data cataloging and reconstruction chains and are now being reconstructed as they arrive. The ProtoDUNE-DP technology locates the readout systems above a thin layer of argon gas above the liquid argon surface. This gas layer allows an external electric field to accelerate the electrons and produce gas amplification. The result is a substantial increase in signal-to-noise in the resulting signals, at the cost of longer electron drifts from the bottom of the liquid volume.

Process	Rate/time	Size/trigger (MB)	Total size (TB)
SP COMMISSIONING	25 Hz	36 MB	773
SP BEAM EVENT	25 Hz	70-200 MB	856
SP COSMIC EVENTS	1-2 Hz	50-150MB	2310
SP RECONSTRUCTED DATA	500 s	20-120 MB	2524
SP SIM	2700 s	280 MB	415
DP COMMISSIONING		50-150 MB	129
DP COSMIC EVENT	1 Hz	50–300 MB	140
DP RECONSTRUCTED DATA	200 s	23 MB	22
OTHER SIMULATION	600 s	10–200 MB	2178
Total			**9346**

Table 13: Data taken as of 9/2020 with the protoDUNE detectors. Beam was available for six weeks, but cosmic-ray testing has been running since 10/1/2018.

5.9.3.1.2.3 Conclusions from Prototype Tests

ProtoDUNE prototype runs with cosmic rays are ongoing and will continue through beam tests in 2022 at CERN. Data cataloging, movement, and storage techniques were tested before the start of the ProtoDUNE-SP and ProtoDUNE-DP runs and were able to handle the full rate of the experiments. Reconstruction algorithms were also in place on time and were able to produce early results that led to increased understanding of the detector and improved calibrations for a second iteration. An additional run of both ProtoDUNE-SP and ProtoDUNE-DP is planned for 2022, allowing further development and testing of the computing infrastructure before the full detector comes online in the late 2020s.

5.9.3.1.3 On to Full DUNE

The full DUNE far detector will begin with one SP module to be installed at SURF starting in the middle of this decade. High-intensity neutrino and anti-neutrino beams should arrive after a year or so of commissioning of the detector and LBNF beamline. This first module will largely resemble a scaled-up version of ProtoDUNE-SP with 150 APAs distributed two deep at the center and long edges of the cryostat. The argon volume will be 15 x 14 x 62 m3 with a fiducial mass of 10kT. **Table 14** summarizes the expected event rates and data volumes for one such detector module. Additional detector modules, likely one DP, another SP, and one with novel technology, will be added. For now, assume that data volumes and rates coming from other technologies will be less than or equal to the SP values.

Process	Rate/module	Size/instance	Size/module/year
BEAM EVENT	41/day	6 GB	47 TB/year
COSMIC RAYS	4,500/day	6 GB	9.7 PB/year
SUPERNOVA TRIGGER	1/month	115 TB	1.4 PB/year
CALIBRATIONS	2/year	750 TB	1.5 PB/year
Total	**12.9 PB/year**		

Table 14: Data sizes and rates for different processes in each far detector module. Uncompressed data sizes are given. As readouts will be self-triggering, an extended 5.4 ms readout window is used instead of the 3ms for the externally triggered ProtoDUNE-SP runs. Assume beam uptime of 50% and 100% uptime for non-beam science. These numbers are derived from references[186-187].

The regular data rates will be dominated by 4,500 cosmic rays expected per module/day. These events are vital for monitoring and aligning the detector. The beam-initiated data are a small subset of the cosmic-ray triggers but will have similar size and processing characteristics.

The next most significant source of data will be supernova triggers (see Section 5.9.3.1.3.1) and calibration campaigns with radioactive and neutron sources and lasers. These, in contrast to the steady cosmic-ray rate, are expected to occur one to two times/month and involve data volumes of up to 200 TB of compressed far detector data.

In all cases, the goal is to gather data from the full volume of the detector with as fine a granularity as possible. Beam interactions themselves are expected to be quite rare, occurring in only 1/2000 beam gates. Extraction of oscillation parameters will require both the powerful electron background rejection discussed in the previous section and precise calibration of the energy scale of the experiment, hence the much larger calibration samples.

5.9.3.1.3.1 Supernova Candidates

Supernova candidates pose a unique problem for data acquisition and reconstruction. Supernova physics in DUNE is discussed in some detail in the TDR[4] and only summarized here. A classic, core-collapse supernova 10 kpc away would be expected to yield approximately 3,000 charged-current electron neutrino interactions across four detector modules. The interplay of supernova and oscillation physics is not fully understood and can result in significant modulations of the event rates for different neutrino types over the few tens of seconds of the burst. DUNE's fine-grained tracking should allow significant pointing power with the most optimistic scenario of four modules and high electron neutrino fraction yielding pointing resolutions of less than five degrees. The ability to produce a reasonably fast pointing signal would be extremely valuable to optical astronomers doing follow-up, especially if a supernova was in a region where dust masks the primary optical signal. The need to be alert to supernovae and to quickly transfer and process these data imposes significant requirements on triggering, data transfer, and reconstruction beyond those imposed by the more regular beam-based oscillation physics. For example, a compressed supernova readout of all four modules will be on the order of 184 TB in size and take a minimum of four hrs to transfer over a 100Gbps network, and then take on the order of 130,000 CPU-hrs for signal processing at present speeds. If processing takes the same time as transfer, a peak of 30,000 cores would be needed.

The supernova triggers (and some large calibrations) 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. Supernova/calibrations and normal beam/cosmic running are accounted for separately in the transfers specified in **Table 15.**

[186] https://docs.dunescience.org/cgi-bin/private/ShowDocument?docid=16028
[187] https://docs.dunescience.org/cgi-bin/ private/ShowDocument?docid=14983

Quantity	Value	Explanation
Far detector beam: SINGLE APA READOUT SINGLE APA READOUT APAS PER MODULE FULL MODULE READOUT BEAM REP. RATE SIGNAL PROCESSING CPU TIME/APA SIGNAL PROCESSING CPU TIME/INPUT MB MEMORY FOOTPRINT/APA	 41.5 MB 16.6 MB 150 6.22 GB 0.83 Hz 40 sec 2.5 sec/MB 0.5-1 GB	 Uncompressed 5.4 ms ×2.5 compression Uncompressed 5.4 ms Untriggered from MC/ProtoDUNE compressed input ProtoDUNE experience
Supernova: SINGLE CHANNEL READOUT FOUR MODULE READOUT TRIGGER RATE	 300 MB 460 TB 1 per month	 Uncompressed 100 s Uncompressed 100 s (assumption)

Table 15: Useful quantities for computing estimates based on the current design for SP readout of a far detector module. For sparse far detector events, the pattern recognition phase, which scales with occupancy, is expected to be substantially faster than the signal-processing phase, which scales with detector size.

Summary

Overall, a bottom-up estimate yields data volumes of approximately 13 PB/year/module. While lossless compression and experience operating the detector should reduce this volume, additional modules will add additional trigger readouts that will increase these rates. A maximum data volume transferred to permanent storage of 30 PB/year across all four modules and modes of operation has been specified and agreed upon by the collaboration.

We will note that 30 PB/year is an average of 1.3 GB/sec, less than the rates already demonstrated for protoDUNE acquisition and storage. In principle, at 2.5 CPU sec/MB of compressed input, 2,000–3,000 cores could keep up with these data rates, but this throughput must be maintained over many years. In addition, supernova candidates will require bursts of much higher acquisition and processing rates. **Table 15** summarizes the computational characteristics expected for far detector data.

5.9.3.1.4 Near Detector

High-precision oscillation physics requires a near detector system to allow measurement of the original neutrino flux and improved understanding of neutrino interaction physics. The DUNE collaboration is proposing a suite of near detectors optimized for these two goals.

The near detectors will be located in an enclosure on the Fermilab site 574 meters from the target and will be exposed to the DUNE neutrino beam. Interaction rates per spill (at 0.83 Hz) are expected to be very large, with 40–60 neutrino and other interactions produced per beam spill, including muons originating from interactions in material upstream of the fiducial volumes. **Figure 37** shows the beamline and location of the near detectors on the Fermilab site. There are three major subsystems. A pixel readout liquid argon detector, ND-LAr, is the most upstream of the three sub-detectors shown in **Figure 38,** where the beam propagates from right to left. Immediately downstream of ND-LAr is the gaseous liquid argon detector, ND-GAr, which serves ND-LAr as a muon spectrometer and allows more detailed study of neutrino interactions that occur within its gas volume. Beyond ND-GAr is the SAND component of the near detector that acts as a beam monitor.

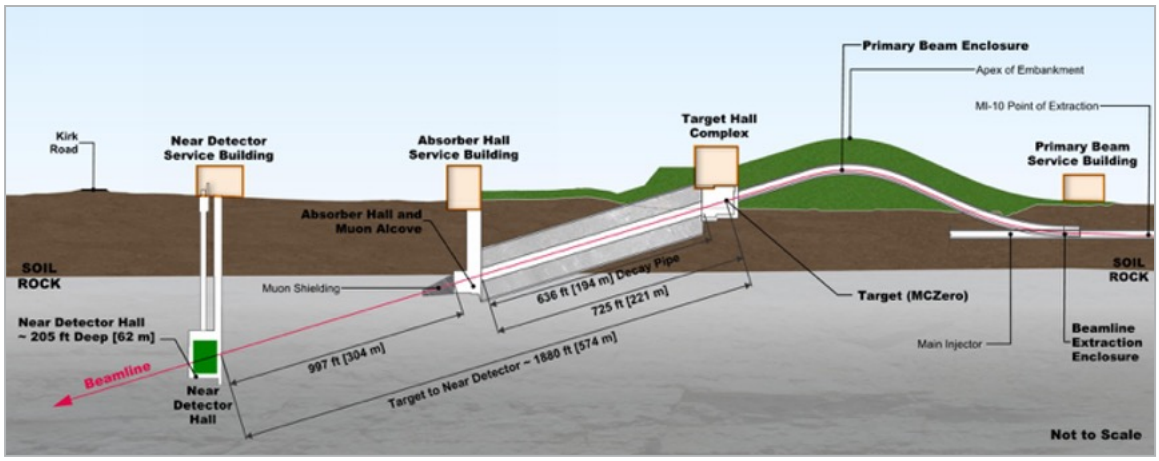


Figure 37: LBNF neutrino beamline on the Fermilab site. The near detectors will be situated 574 m from the target and 62 m below grade.

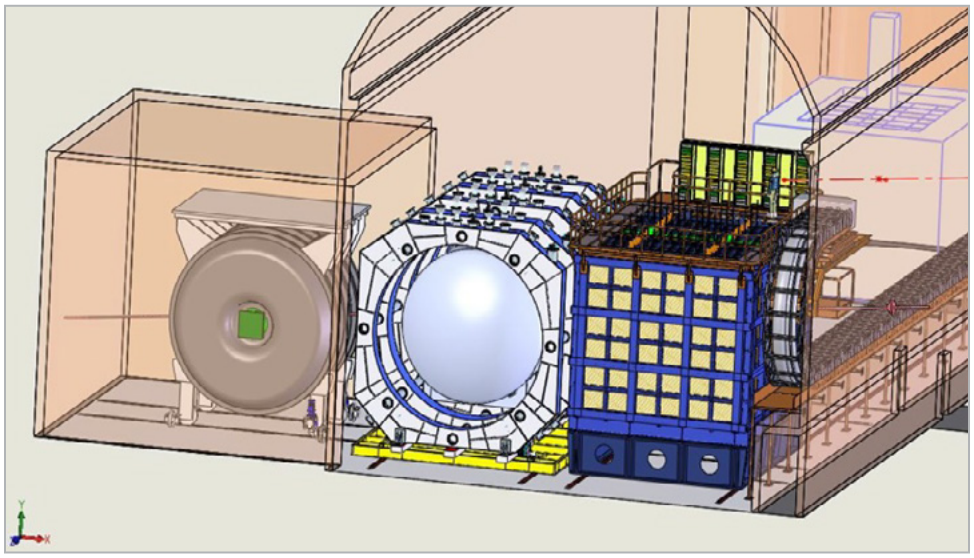


Figure 38: Near detector systems in an on-axis configuration. The beam enters from the lower right in this view. The SAND scintillating beam monitor remains at beam center while the pixel ND-LArTPC detector and gaseous ND-GAr TPC detectors can be moved off-axis to make detailed studies of the neutrino flux at multiple angles.

5.9.3.1.4.2 Near Detector CPU Needs and Simulation

Table 16 summarizes the expected data sizes from the near detectors. Due to the much higher data density in the near detector, CPU times/beam spill are expected to be much higher and are estimated to be 300 CPU/sec/spill using current processors for 1.5 X 107 spills/year. Simulated data samples will need to be an order of magnitude larger and thus require at least 10 times the CPU power. This leads to a rough estimate of CPU needed of approximately 3,000 core-years for each year of data collected from the near detector. As the near detector is located on the Fermilab site, there is no need for high-bandwidth external network connections for data acquisition and archival, in contrast to the far detector.

A Conceptual Design Report for the Near Detector systems is in preparation and the near detector computing efforts are being integrated with the existing far detector and protoDUNE efforts.

Subdetector	In-spill data	Out-of-spill cosmics	Calibration	Annual total (terabytes)
ND-LAR	144	16	16	176 TB
ND-GAR	52	10	6	68 TB
SAND	4	1	1	6
Total	**200**	**27**	**23**	**250**

Table 16: Annual data sizes in the three sub-detectors for different processes in the near detector. Uncompressed data sizes are given. Assume beam uptime of 50% and 100% uptime for non-beam science. These numbers are derived from the DUNE Near Detector Conceptual Design Report.

5.9.3.2 Collaborators

There were 34 countries, 207 active institutions, and 1,209 active collaborators as of July 20, 2020.

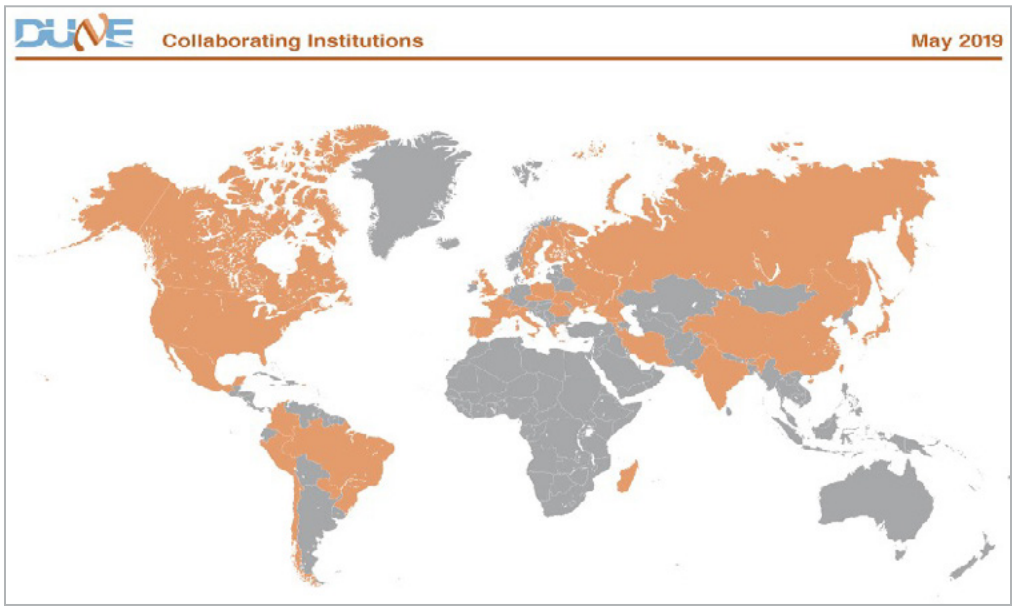


Figure 39: This map is a graphical display of the locations of DUNE collaborating institutions.

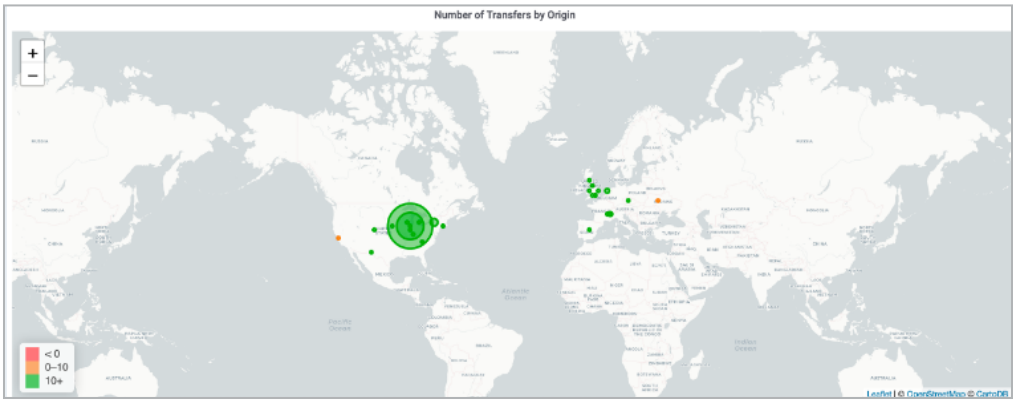


Figure 40: Access patterns for DUNE data over the past year[188]

[188] https://fifemon.fnal.gov/monitor/d/aGGlQY5Wz/dcache-transfer-overview-map?orgId=1&var-filter=storage_group%7C%3D%7Cdune&-from=now-1y&to=now

User/collaborator and location	Primary or secondary copy of the data	Data access method	Avg. size of data set	Frequency of data transfer or download	Are data sent back to the source and method	Any known issues with data sharing
1) FAR DETECTOR UNDERGROUND (DAQ[189])	No	Create/transfer to Fermilab	30 PB/year	Continuous	No	One 2 GB readout every 10–100 seconds
1.5) FAR DETECTOR UNDERGROUND (TEMPORARY DAQ DATA[190])	No	Create/transfer to Fermilab	<5 PB/year	Ad hoc	No	Raw data that may be transferred for processing without long-term storage
2) FAR DETECTOR CONTROL[191] (DAQ, DETECTOR, CRYOGENICS)	No	Create/transfer to and from Fermilab	???	Continuous	Yes	Control/DB needs very high availability
3) FAR DETECTOR SUPERNOVA AND CALIBRATIONS	No	Create/transfer to Fermilab	200 TB at Once	4 hrs once per month	No	Needs to be very fast
4) NEAR DETECTOR	No	Create/transfer within Fermilab	1–2 PB/year	Continuous	No	
5) FERMILAB DATA CENTER	Yes	Ingest from 1–4 and streaming transfer out for processing on OSG/WLCG	30 PB/year	Continuous	Yes	Ingest
6) FERMILAB DATA CENTER	Yes	Transfer out to secondary tape/disk archives in Europe/Asia	30 PB/year	Continuous	No	Egress from Fermilab
7) FERMILAB DATA CENTER	No	Transfer to HPC facilities	200 TB at once	4 hrs once per month	No	Needs to be very fast
8) OSG SITES	Partial	Transfer output back to storage at DUNE sites	12 PB/year	Continuous	Yes	
9) US HPC SITES	No	Transfer back to storage at DUNE sites	12 PB/year	Continuous	Yes	Staged transfer and not streaming
10) CERN DATA CENTER	Partial	Transfer to Fermilab primary store and European sites	5 PB/year	Continuous	Yes	ProtoDUNE
11) EUROPEAN SITES	Partial	Provide data via streaming to DUNE compute sites and institutions	12 PB/year	Continuous	Yes	
12) REST OF WORLD	Small	Provide data via streaming/transfer	6 PB/year	Continuous	Yes	
13) COLLABORATING INSTITUTIONS	Small	Receive data via streaming/transfer	30 PB/year	Continuous	Yes	

Table 17: DUNE data projections

[189] The DAQ will have enough storage capacity to continue taking data for one to two weeks, if the bulk data transfer to Fermilab is interrupted, but the control path of the DAQ remains available.

[190] Since the computing capacity at SURF is limited, DUNE needs to preserve the ability to transfer data to Fermilab that are not meant for permanent storage, but need to be processed (e.g., for monitoring or calibration purposes), within the bandwidth provided by ESnet available to absorb the peak traffic rates of use-case 3). A scenario made possible if this additional network bandwidth is available is to loosen the data selection, transfer data in excess of 30 PB/year, and have a first stage of quasi-online data reduction at Fermilab, where installing and managing compute power is easier than at SURF.

[191] Ability to remotely view and control the status of the experiment is mandatory at all times; effort is being put to have redundant networking components, under uninterrupted power, for those parts of the system that require it at SURF.

The DUNE far detector at SURF needs two links, one for large data uploads and one for smaller control and database access.

Control Path

Need high-reliability connection between Sanford Lab and Fermilab for experiment control, monitoring, and database access.

Normal Data Paths

Up to 30 PB/year of cosmic ray and data will be taken per year by the DUNE far detector and sent to Fermilab for storage to tape. These data will then need to be distributed worldwide for processing and the resulting reduced samples will be stored on distributed disk and tape at Fermilab and at secondary sites in other countries. These reduced samples are likely to be a few PB in size. There is a plan to also maintain a secondary tape copy of physics raw data at a secondary site.

In parallel, the near detector, located at Fermilab, will be producing several PB/year which must also be distributed worldwide for processing with the derived data sets stored on distributed disk.

Supernova/Calibration Data Paths

Supernova interactions and calibration runs are likely to produce up to 200 TB of information over a period of a few minutes to a few hours. These data are buffered at the SURF site but, in the case of supernovae, need to be moved and analyzed as quickly as possible in the event of a real supernova. As these data provide pointing information, a timescale of hours for delivery is needed. It should be noted that the supernova will be sensitive once photon detectors are commissioned and the first far detector module is filled with liquid argon in approximately 2028.

5.9.3.3 Instruments and Facilities

Present to Two Years:

The current instruments actively being used by DUNE right now are primarily located at CERN or Fermilab. The largest of these instruments are the SP and DP ProtoDUNE detectors located at the CERN Neutrino Platform facility and were described in Section 5.9.3.1.2.2. The ProtoDUNE detectors will be utilized for prototyping DAQ, detector components, and detector configuration (high voltage, xenon doping, etc.) for at least the next two years, but likely through 2025. There are additional small-scale prototype instruments located at Fermilab that are testing novel light detection and pixelated readout from LArTPCs.

DUNE currently utilizes extensive computing and storage facilities at Fermilab, CERN, and distributed computing across the WLCG and OSG. At Fermilab, the dCache and Enstore facility provides primary storage of all raw and derived data sets from ProtoDUNE and other prototypes. Additionally, raw ProtoDUNE data are also stored at CERN through EOS and CASTOR storage facilities. The total DUNE storage utilization as of Aug 2020 through Rucio data management was close to 13 PB of raw and derived data stored across more than 15 Rucio Storage Elements.

The production processing of DUNE simulation and ProtoDUNE data and simulation, along with user analysis computations, take advantage of processing resources at Fermilab, CERN, HPC in the United States, and resources that are part of the WLCG/OSG. DUNE offline production workflows have utilized more than 30 computational sites in more than 7 countries. These resources have been a combination of opportunistic and pledged resources, and it is anticipated that the breadth of resources will continue to grow for the next decade.

DUNE has recently been able to access and successfully produce ProtoDUNE simulation at the NERSC Cori facility. The ability to utilize HPC is seen as an important aspect of DUNE's computing model. During the next two years, the availability of HPC resources is seen as supplemental but not essential to the success of DUNE computing.

Two to Five Years:

With 200 collaborating institutions worldwide, the DUNE Computing Consortium anticipates growing considerably from the current infrastructure in terms of both processing sites and storage elements. The formation of a Computing Contribution Board and an updated computing model based on the knowledge gained from ProtoDUNE operations will see the implementation of a more complete conglomeration of resources. It is anticipated that more than 30 Rucio Storage Elements and more than 50 computational sites will become part of the DUNE computing infrastructure. Additionally, the instantiation of new HPC facilities, such as Aurora at ANL or Perlmutter at NERSC, will create opportunities for unique processing workflows for DUNE data and simulation. The access and staging of data to and from these facilities, along with network resource needs, will involve extensive R&D in the coming years.

The large archival storage facilities at Fermilab and CERN will continue to be the largest providers of raw and derived data set storage through 2025.

Beyond Five Years:

Once the DUNE far detector is commissioned in the second half of the decade, the DUNE computing resource needs will start to match the requirements for which the DUNE Computing Model was designed. It is anticipated that collaboration contributions of computing infrastructure across the world will have the processing and storage capabilities needed to allow the accomplishment of the physics goals of DUNE. Transition to full operations will begin with the start of data acquisition at the far detector at SURF in South Dakota with an anticipated rate of 30 PB per year. The replication and processing of this data volume is the largest anticipated steady state resource need for DUNE. The primary archival storage being located at Fermilab and secondary archival storage replicated from those resources. The DUNE near detector operation at Fermilab is anticipated shortly after commissioning of the first far detector module, and will produce data sets on the order of 1 to 2 PB per year.

Figure 41 shows estimated tape storage needs through 2030, assuming two copies and a lifetime of 10 years for derived samples. **Figure 42** shows estimated disk needs, assuming two copies of important samples with lifetimes from 6–24 months.

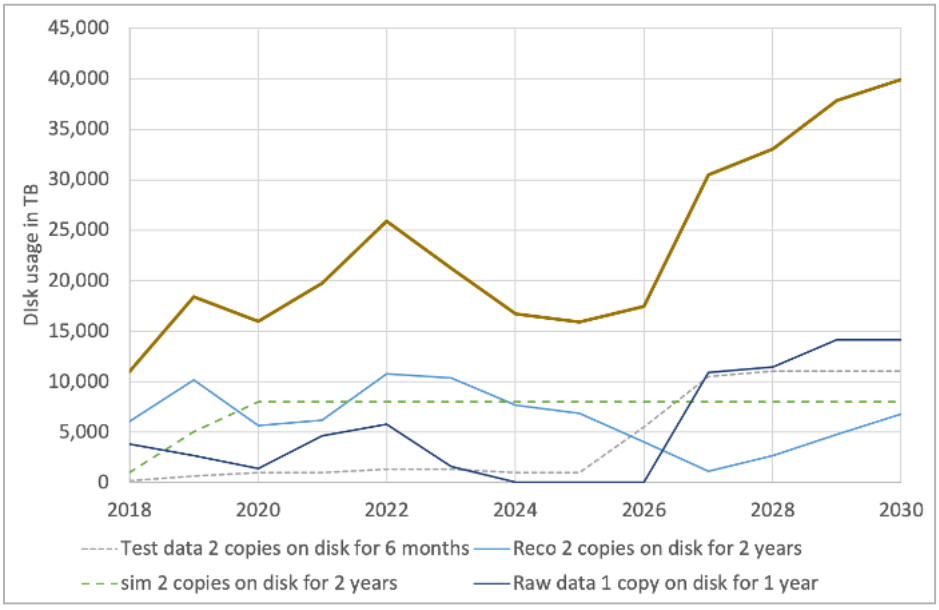


Figure 41: DUNE tape storage needs in TB, assuming two copies of raw data and one of derived samples. This assumes that approximately half of data taken starting in 2026 are for commissioning tests, and are not retained permanently on tape.

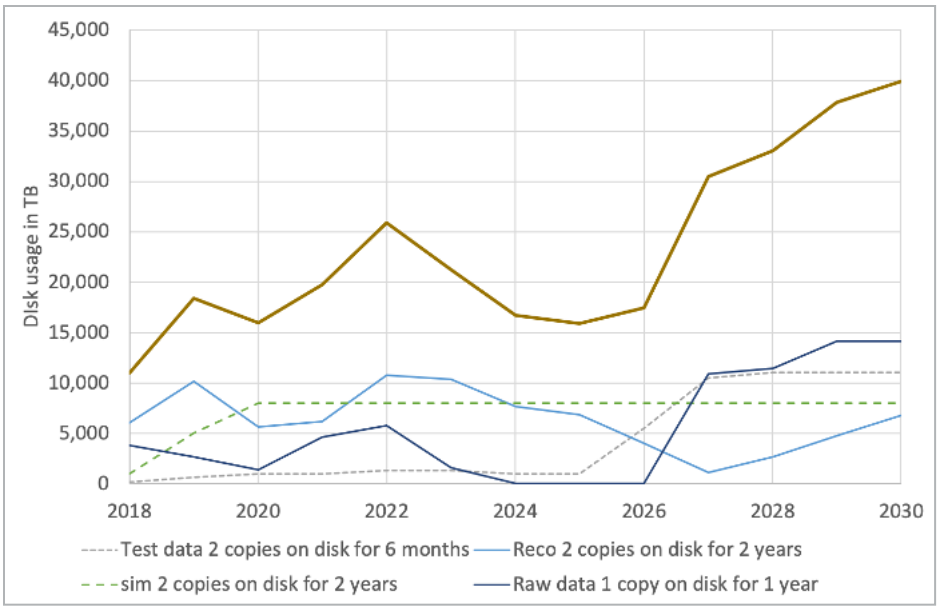


Figure 42: Estimated disk needs for DUNE. There is an early peak for ProtoDUNE data and analysis, followed by the commissioning of the far detector systems starting in 2026–2027.

5.9.3.4 Process of Science

The operations of the DUNE experiment will proceed in three major phases:

1. ProtoDUNE 2 runs at CERN are scheduled for 2021–2022 and may potentially continue to operate until 2025.
2. Installation and commissioning of the far and near detectors in South Dakota and at Fermilab over the period 2025–2029.
3. Physics running of the near and far detectors from 2028–2040 and possibly beyond.

Each of the operations has several data processing phases, each with different characteristics:

1. Data taking and storage.
2. Data reconstruction.
3. Simulation.
4. Data analysis.

Phase 1: ProtoDUNE Data Taking at CERN

Data taking: The protoDUNE prototypes at CERN will run in the EHN1 test beams. During beam running, average data rates can reach 25Hz, which leads to data rates from each of the two experiments of up to 2–3 GB/s. The data will be stored temporarily in the EHN1 site and then moved to permanent cataloged tape storage at CERN and Fermilab. As the data originate at CERN, this step requires use of the TA networks. Beam runs are typically one to two months long, but the detector also takes data on cosmic-ray interactions for performance study at lower rates over longer times. The resulting data sample will be in the 2–10 PB size range.

Data reconstruction: Data reconstruction uses OSG/WLCG facilities. In the current model, data are staged from tape to disk at Fermilab. Reconstruction jobs are started on grid nodes across the DUNE Virtual Organization (VO) and those jobs stream 2–8 input GB files via the XROOTD protocol. The resulting output files (reduced by a factor of four in size) are then copied back to tape storage. Two copies of the output file are kept in the

Rucio DMS, one in Europe and one in the United States. Processing tends to run in bursts, as fast turnaround is useful for feedback to the running detector and to allow improved algorithms and calibrations. The goal is to do a first reconstruction of data with a short delay from when it is taken and to keep up with data taking. Once the data taking is done, data are processed, calibrations are done, and algorithms tuned, and then data are typically reprocessed on a six-month cycle for several years.

Simulation: Simulation is similar to data processing except that, for parts of the simulation chain, little input data are needed and coupling to the data-taking schedule is absent. This makes simulation easier to schedule and more optimal for HPC centers such as NERSC. Simulation CPU and storage needs are similar to the needs for raw data.

Analysis: For protoDUNE most data analysis is detailed calibration or algorithm development and requires the full reconstructed event records, which are data samples of several 100 TB in size. These are accessed via streaming by a group of 20–50 experts using OSG/WLCG grid resources. Aggregate data access rates can exceed those for reconstruction as many more users are accessing the data and may do so simultaneously. There are plans to maintain redundant data copies in Europe and the United States, which should reduce dependence on TA networks.

Phase 2L 2025–2029 Commissioning of the Near and Far Detectors

This phase is similar to Phase 1, except that commissioning of the far detector at SURF becomes the main data producer. The control links between SURF and Fermilab become very important, and tests of high-speed data transfer and storage (up to 8 GB/s) begin. Large-scale data challenges to test the entire data taking and distribution system will be performed. Likely a first data challenge will be done to test the software and hardware infrastructure outside of SURF early in this time period with a full test with hardware at SURF timed later to avoid buying hardware too early.

Phase 3: 2028–2040, Stable Physics Operation of DUNE

Phase 3 of DUNE will involve stable operation of the far detector at SURF and the near detector at Fermilab. A first far detector module is expected to come online in 2028–2029 with three additional modules being commissioned over the next few years. As a result, stable operations will run in parallel with commissioning of new detectors.

Data taking: Steady data taking from the far detectors will consist of ~ 4,000 cosmic-ray interactions per module/day and 10–20 beam interactions/module/day. In addition, as described previously, supernova alerts (including tests) and calibration runs will generate 200–500 TB of data over shorter time periods which need to be exported, cataloged, stored, and processed quickly. In parallel with the main data stream, which will be buffered locally, a highly reliable connection is needed for control and to generate fast supernova alerts for the SuperNova Early Warning System (SNEWS). Data will be stored to tape at Fermilab with a second copy distributed among collaborating institutions.

The far detectors will run continuously, doing calibration and astrophysics even when the beam is off. The aggregate data volume for commissioning, cosmics, calibrations, and supernova alerts are expected to be 30 PB/year. The near detector at Fermilab will generate large amounts of data only half of the time that the beam is running, with an aggregate raw data volume in the 1–5 PB range

Data reconstruction: Data reconstruction will use worldwide OSG/WLCG or successor facilities. In the current model, data will be pre-staged from tape to disk and reconstruction jobs will run on grid nodes across the DUNE VO. Input files will be 2–8 GB in size and likely correspond to a single readout of a far detector module or several minutes of near detector data. The resulting output files (reduced by a factor of up to 100 in size due to the lower occupancy in the underground detector) are then copied back to tape storage.

Two copies of output files will be kept on disk in the Rucio DMS, one in Europe and one in the United States. The goal is to do a first reconstruction of data with a short delay from when it is taken and to keep up with data taking.

Processing will be a mix of steady processing of the cosmic-ray and beam events as they come in and bursts as fast turnaround is needed for calibrations and supernova triggers. Steady processing of far detector data will occupy a few thousand cores at current processing speeds but bursts may require up to 100,000 cores for short periods. Near detector data will likely be smaller but require more CPU time per byte due to the complexity of the detectors. Current estimates indicate that near detector processing will also require a few thousand cores to keep up with the data.

Full supernova readouts will consist of up to 100,000 5 msec readouts. The processed outputs will need to have sufficient metadata to allow a full picture in time and space of the supernova event to be built from the thousands of individual processed output files.

As in stage 1, data will typically be partially reprocessed on a six-month cycle for several years. Near detector data will likely be smaller but require more CPU time per byte due to the complexity of the detectors.

Simulation: Simulation will be similar to data processing except that, for parts of the simulation chain, little input data are needed and coupling to the data-taking schedule is absent. This makes simulation easier to schedule and more optimal for HPC centers such as NERSC. Simulation CPU and storage needs are similar to the needs for raw data. One possible issue is the need to overlay real data as part of the detector simulation phase. This requires delivery of reasonably large data inputs to simulation jobs.

Analysis: Data analysis will consist of both detailed calibration and algorithm development by a group of 50–100 experts at 10–20 sites and analysis of derived samples by a much large (250–500) group of collaborating physicists at up to 100 distributed sites worldwide. Aggregate data access rates will exceed those for reconstruction, as many more users will be accessing the data and may do so simultaneously.

DUNE is still developing a long-term computing model, largely based on the HEP Software Foundation (HSF) DOMA framework with data and CPU resources distributed across collaborating countries. **Figure 43** shows the DOMA model for compute centers. Fermilab and a few large sites (Archive Sites in DOMA language) will provide the tape archive with a large number of national entities pledging significant CPU and dedicated disk resources as disk and compute (DCC) centers. Smaller institutes will serve as compute centers and ingest data via streaming or copies to local cache. This model is highly dependent upon good quality network connectivity and the ability to monitor and control access to network resources.

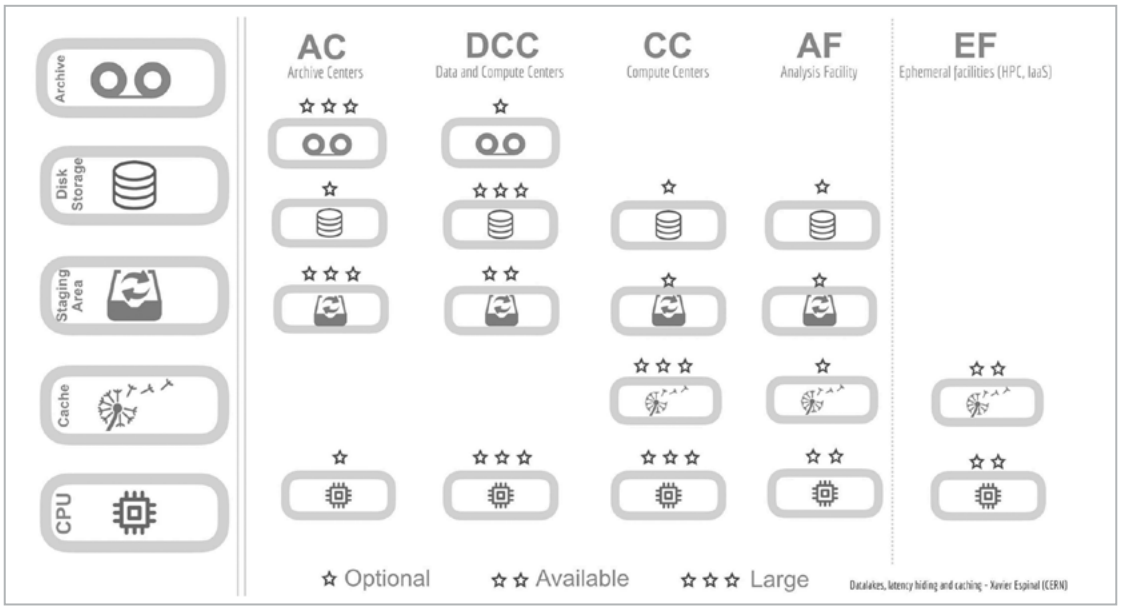


Figure 43: DOMA MODEL for compute sites. Fermilab and CERN currently serve as ACs while US labs and several European countries (UK, FR, CZ) host DCCs and others (OSG, ES, NL, RU) mainly contribute CPU resources.

5.9.3.5 Remote Science Activities

As described previously, the DUNE experiment is highly distributed.

In the short (zero- to three-year) time frame, DUNE will be running the protoDUNE experiments at CERN, storing the data on tape at Fermilab and CERN, and analyzing those data worldwide through WLCG/OSG grid resources. ProtoDUNE data taking can be up to 2–3 GB/s for extended (several week) bursts and needs to be transferred to Fermilab in close to real time and then exported for processing. This was done successfully for several two-week long data-taking runs in late 2018. As the beam will not be running full time, use of network links is likely to be variable on timescales of weeks.

In the medium time frame (2025–2028), simulation work will continue worldwide and commissioning of detector systems in South Dakota and at Fermilab will commence. This will require both reliable control links for instrument control and monitoring and a high-speed data link capable of handling up to 10 GB/s during detector commissioning, calibration, and supernova data challenges. During commissioning, usage is likely to be variable, with high rates during tests and little activity at other times.

In the long time frame, the near and far detectors will be fully operational. The far detectors will be live > 95% of the time and delivering data to Fermilab steadily at rates of order 1 GBs for normal beam/cosmic-ray operations. Once a month, a dump of up to 200 TB over a period of a few hours will occur for supernova tests and calibration runs. A disk buffer capable of storing one week of normal data or several supernova/calibration dumps will be provided at SURF in South Dakota.

The end users of all of these data are distributed worldwide and access the data through streaming via XROOTD and individual file transfers. This data distribution is spread more widely but in the aggregate is likely to exceed the transfer rates for the original raw data.

5.9.4 Shared Software Infrastructure

SBN experiments use much of the resources offered through FIFE[192] at Fermilab. SBN is looking to onboard with Rucio in the next year, but otherwise will follow the software infrastructure being developed for general experiments and for DUNE (including SAM, CVMFS, XROOTD, dCache).

DUNE has not finalized its software stack, but will use best practices from SBN, as well as the LHC experiments at Fermilab.

5.9.5 Fermilab Network and Data Architecture

As described in Section 5.9.3.4, there will be four phases of data processing for DUNE: data taking, reconstruction, simulation, and analysis. DUNE networking is expected to have three distinct components to facilitate those various phases of computation:

1. A private backend network from the far detector site at SURF to Fermilab. This backend network will provide dedicated network path(s) for the movement of raw data from the far detector to the Tier 0 archive at Fermilab. The backend network will be a critical component in the data-taking phase of DUNE computing. In addition to supporting the movement of raw data from the detector, the backend network will support controls, monitoring, and other management traffic required to sustain remote operations of the far detector. A dedicated high-bandwidth primary path and a geographically diverse secondary path are envisioned for the backend network. Currently, targeted bandwidth for those paths is 100 Gb/s and 10 Gb/s, respectively. During supernova events, the backend network will be used exclusively for supernova-related traffic.

[192] https://cdcvs.fnal.gov/redmine/projects/fife/wiki/Introduction_to_FIFE_and_Component_Services

These are items 1–3 in Table 13:

2. Special network services across the R&E network infrastructure, customized specifically for DUNE data movement. The level of customization is expected to vary according to the requirements of the type of data movement involved. Current expectations include special network services for the following types of DUNE data movement:

    a. Raw data to secondary (T1?) archival site(s). A private network type of service (analogous to LHCOPN) with bandwidth guarantees, redundancy, and tightly controlled access. The number of sites involved will be very small (1–3?). Bandwidth expectations are expected to be at the 100 Gb/s level. This traffic is also expected to include TA links.

This is item 6 in the Table 13:

    b. Production data movement for the reconstruction and simulation phases of DUNE computing, as well as placement of data within DUNE federated storage. A DUNE collaboration type of service (analogous to LHCONE) is envisioned. The sites involved will be designated DUNE computing and storage locations, numbering in the tens. Use of TA links will be an element of the overall data movement. No site-to-site bandwidth guarantee requirements are expected, but an aggregate bandwidth guarantee for the experiment will likely be desired. ProtoDUNE traffic from CERN to Fermilab is currently being used to prototype this type of service.

This is items 7, 8, 9, 11, 12, 13 in Table 13:

    c. Dynamic point-to-point services with bandwidth guarantees for situations such as supernova events that require fast processing. These services are expected between the T0 (Fermilab) and sites where significant processing is immediately available. The remote sites may be other DUNE collaboration sites, cloud resources, or HPC facilities.

Items 5,6,7 in Table 13:

3. General network services available from the global R&E network infrastructure. This general-use type of service is expected to meet experiment needs that do not require the customized network services required for DUNE's structured data movement. This general network service, however, should not be limited to just conventional best effort. As preferential network services emerge as component of general R&E network services, such services should be expected to find wide use for analysis-related data movement within DUNE computing.

The primary facilities around which this data movement revolves are Fermilab, where the T0 will be located, and far detector facility at SURF, where most of DUNE's experiment data are generated. A description of Fermilab's local network facilities that are currently supporting DUNE is listed below, followed by the conceptual design (as it exists today) of the DUNE local network facilities at SURF.

DUNE Network Facilities and Resources at Fermilab

DUNE currently utilizes the general dCache and Enstore storage facilities at Fermilab for ProtoDUNE-generated data and simulation. There are also smaller storage facilities at CERN and in European labs. The Fermilab storage resources are shared with other experiments. The local network infrastructure consists of fully redundant, high-performance switching fabric. The general dCache fabric has the following characteristics:

- Currently based on 100 GE network technology, with 10 GE and 100 GE connectivity available for data server/mover connectivity;
- Use of a link aggregation group (LAG) to scale bandwidth to N x 100 Gb/s where needed;
- PBR techniques to route ProtoDUNE traffic to/from CERN over special-purpose networks (currently LHCOPN; soon to be ProtoDUNE-specific).

For computing, DUNE currently makes use of a broad spectrum of grid-computing resources. Approximately 40% of its computing is done locally. Approximately half of the computing is performed overseas, mostly involving data movement that traverses TA links.

Fermilab's WAN architecture is based around separating its high-impact science data traffic from its general internet traffic. Conceptually, this design is analogous to the Science DMZ architecture. It is expected that most DUNE traffic into and out of Fermilab will be via the science data path(s). The private backend network to SURF will represent one class of special-purpose network path for DUNE. Services similar to the LHCOPN and LHCONE are expected to emerge for DUNE that will provide a separate class of special-purpose network paths for DUNE production data movement.

In terms of aggregate WAN capacity out of the Tier 0, Fermilab currently has three 100 Gb/s links to ESnet, all carried over a geographically redundant optical network ring. Two of those 100 Gb/s links are used to support the special-purpose network paths for science data movement. The two 100 Gb/s links traverse opposite sides of the optical ring for resiliency, but are link-aggregated into a single 200 Gb/s logical connection to ESnet at layer-2. This 200 Gb/s link is currently shared with LHC data movement. Fermilab's third 100 Gb/s link supports the Laboratory's general R&E network traffic. The 2x100 Gb/s special-purpose networks connection and 100 Gb/s general internet connection serve a redundant function for each other. **Figure 44** 'x' depicts the interface of DUNE storage and archival resources at the Tier 0 (Fermilab) with ESnet WAN services.

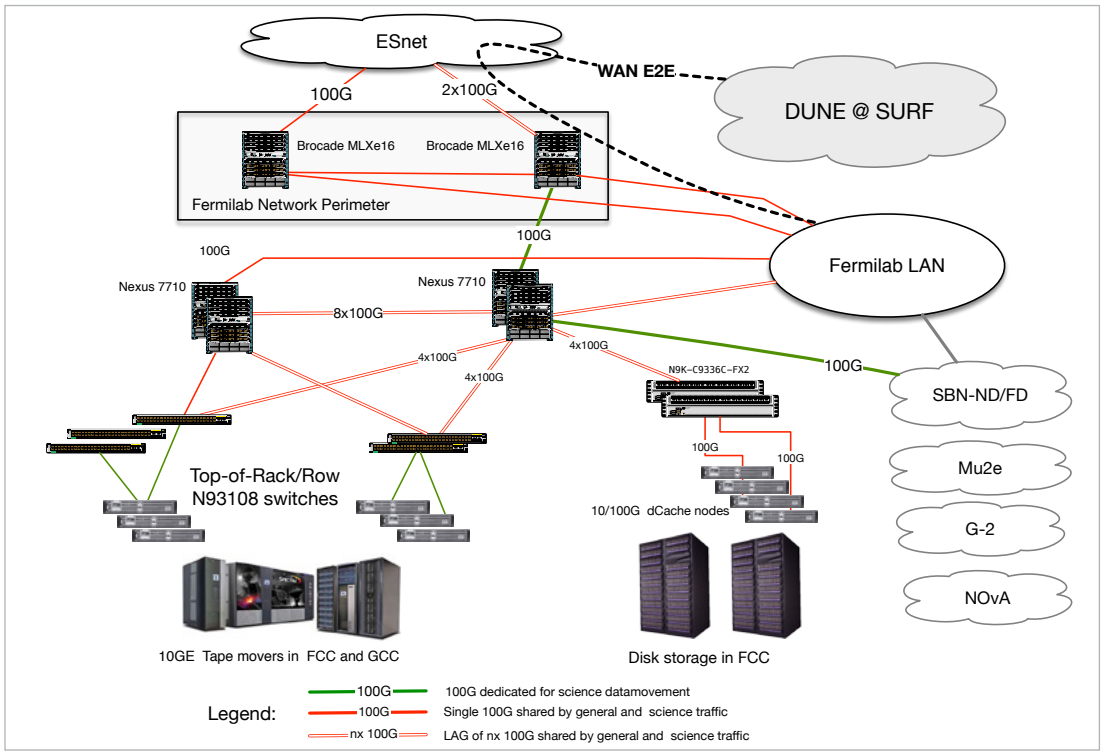


Figure 44: Integration of the DUNE networking into the Fermilab architecture

In terms of future enhancements to the Fermilab network infrastructure:

- Upgrade to 400 GE technology is expected in the 2022–2023 time frame, for both the internal LAN infrastructure that supports DUNE, and for the Laboratory's WAN infrastructure.
- Additional WAN capacity to ESnet, in the form of more 100 GE WAN links or potentially a 400 GE WAN link, will be added as WAN traffic needs require. It is worth noting that having an optical network infrastructure in place makes adding bandwidth capacity only an incremental hardware cost.

DUNE Far Detector Network Design (Work in Progress)

The DUNE local network facilities at SURF are still in the design phase. When finalized, that design is expected to adhere to the following principles:

- Logically within the network perimeter of Fermilab.
- Fully redundant, high-performance switching fabric, both down in the cavern and up on the surface based on 100 GE/400 GE network technologies for inter-switch connectivity, with redundant 10GE and 100GE connectivity available for host system connectivity.
- Fully redundant and geographically separated fiber connectivity between the cavern and the computing facilities up on the surface in Ross.
- Redundant essential network services (DNS, NTP, DHCP).
- A redundant network perimeter, with border devices in both the Ross and Yates buildings.

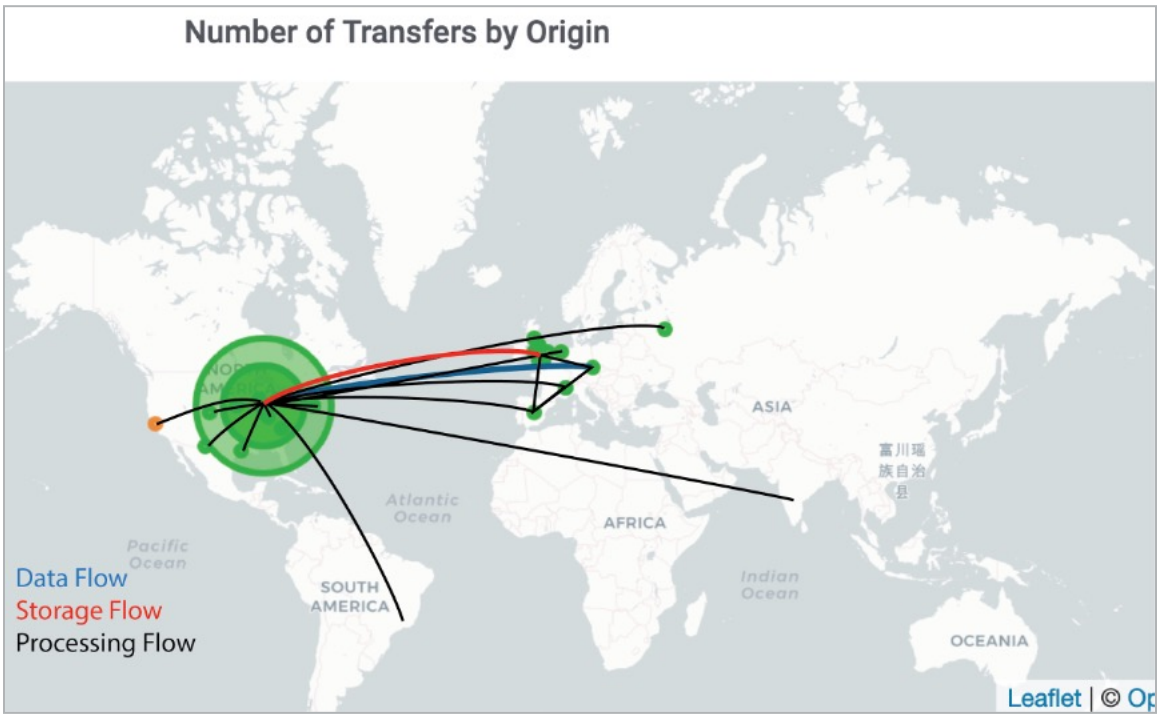


Figure 45: Illustration of current data path as of October 2020. Blue indicates the flow of data from CERN to Fermilab. Red shows storage movements back to Europe. Black shows data flows to processing centers. Data are served widely, but storage is localized at a few large sites.

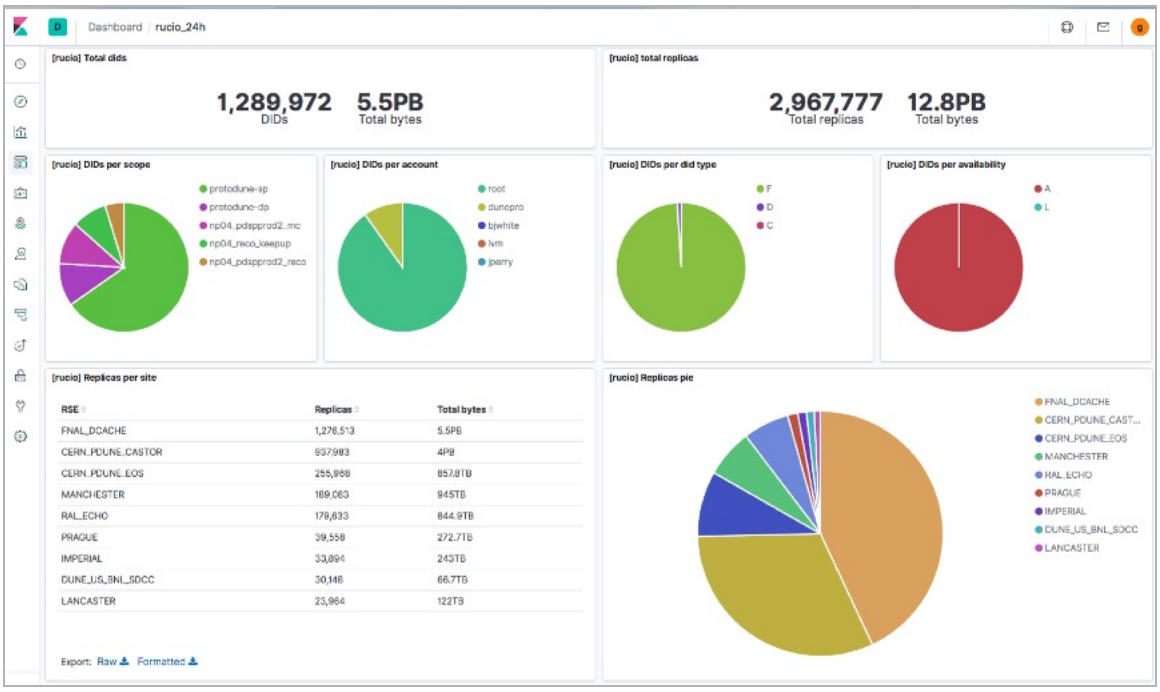


Figure 46: Rucio data location summary

Figure 47: Rucio data location by type dashboard as of 9/13/2020

5.9.6 Shared Cloud Services

Cloud services are available to DUNE and other experiments via the HEPCloud facility at Fermilab. HEPCloud provides access to Amazon EC2 and the GCP at limited levels. To date, these have been used by DUNE to test various machine-learning scenarios at limited scale. There is not any commercial cloud usage in the current production of data and MC for DUNE, nor is there any anticipated in the next two-year period. There have

been some limited tests using field-programmable gate arrays (FPGAs) and GPUs on Google as remote inference servers, and it is expected that these will continue. DUNE does have access to the VAC/VCycle private cloud facilities of GridPP but has not used them in bulk as well. All of these use cases require minimal network ingress and egress from the compute clouds, and DUNE does not expect a sustained bandwidth capacity would be necessary.

The longer-term technology horizon remains to be determined, and will follow the larger strategic direction of Fermilab. The low-level reconstruction of DUNE is computing that is well suited to GPU processing, and although the commercial cloud is not seen as the first source of this hardware for DUNE, it may be necessary.

5.9.7 Data-Related Resource Constraints

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. By late in the 2020s, DUNE anticipates a need to sink 100 Gbps from SURF to Fermilab and redistribute those data simultaneously to sites worldwide. This will require both improvements in the SURF-> Fermilab link and in the ability of the disk systems at SURF and Fermilab to handle those data rates.

For DUNE's use case of rapid response data, such as supernova data, it would be useful to have both a higher network QoS as well as a faster storage QoS. Both would facilitate faster operational readiness.

To date, DUNE has not come close to filling the cross-Atlantic network pipe that was allocated, and will likely continue to be a small fraction of TA traffic. DUNE needs to develop a comprehensive model of how many IOPS are needed to sink the incoming data, to copy the data to elsewhere, and to serve production and analysis users. But part of that model will be situating copies of important data samples on both sides of the Atlantic to reduce the need for streaming transfers between regions.

SBN's major concern, based on MicroBooNE production experiences, is the performance of data storage I/O, rather than networking. SBN currently relies on tape storage for the bulk of the data storage needs, and staging back and forth from tape can generally be the main obstacle for data production workflows. SBN is working to update much of the computing model. As these updates occur, some changes to the calculations presented in previous sections are expected, placing more strain on networking resources. SBN and Fermilab computing experts will need to keep a close eye on how modifications in production workflows have effects on the whole system.

5.9.8 Outstanding Issues

The DUNE data ecosystem will be very complex and will need sophisticated methods to monitor and track the data movement between sites. Rucio may be able to track the movement of large data samples, but access to data by CPUs for processing will require some other method of monitoring.

A major development question for us is the tradeoff between using fast networks to locate data semi-permanently in the most efficient way and then access it via streaming wherever CPU power is available or, if due to network constraints that make streaming impractical, DUNE needs to develop systems that temporarily move data closer to CPU resources on demand. DUNE much prefers the simplicity of the first model, but it depends on high-quality networking and the ability to monitor activity to avoid bottlenecks.

Integrating network quality of service and available resources into the data management and batch systems is a development project worth pursuing, likely in collaboration with DOMA and ESnet efforts.

5.9.9 Case Study Contributors

Neutrino Experiments at Fermilab Representation

- Heidi Schellman[193], Oregon State University
- Tim Bolton[194], Kansas State University
- Steve Brice[195], Fermilab
- Mike Kirby[196], Fermilab
- Wes Ketchum[197], Fermilab
- Kurt Biery[198], Fermilab
- Nicole Avila[199], University of Chicago
- Steve Timm[200], Fermilab
- Ken Herner[201], Fermilab
- Stuart Fuess[202], Fermilab
- Kate Scholberg[203], Duke University
- Ramon Pasetes[204], Fermilab

ESnet Site Coordinator Committee Representation

- Phil DeMar[205], Fermilab
- Andrey Bobyshev[206], Fermilab

5.10 LHC Experimentation and Operation

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 experiments running at the LHC are exploring the fundamental structure of matter and the forces that govern structures of matter's interaction.

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

193 Heidi.Schellman@science.oregonstate.edu
194 bolton@phys.ksu.edu
195 sbrice@fnal.gov
196 kirby@fnal.gov
197 wketchum@fnal.gov
198 biery@fnal.gov
199 navila@uchicago.edu
200 timm@fnal.gov
201 kherner@fnal.gov
202 fuess@fnal.gov
203 kate.scholberg@duke.edu
204 rayp@fnal.gov
205 demar@fnal.gov
206 bobyshev@fnal.gov

The two main general-purpose collaborations at the LHC are ATLAS and CMS. 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.

5.10.1 ATLAS Experiment Notes

This section focuses on the ATLAS detector and collaboration. ATLAS is the largest of four particle detectors that measure and record the particle collisions at the LHC. The primary scientific goal is to quantitatively measure and discover properties of the SM of particle physics. A major emphasis of the upcoming Run 3 at the LHC will be the precision measurement of Higgs properties.

The LHC collides protons more than a billion times every second, out of which ATLAS selects interesting collisions for recording at a rate of a thousand times a second. With over two thousand hours of data collection every year when the LHC is running, ATLAS has a huge data sample for physicists to analyze from the completed Run 1 and Run 2. Run 3 is scheduled to start data collection in 2022.

Several core areas of discussion emerged out of this section:

- 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.
- Data sets in ATLAS are collections of files organized by category/workflow. Data sets are the fundamental units in ATLAS, and vary largely in size:
  - Raw data sets are in the range of 1 to about 50 TB.
  - AOD data sets are in the range of 1 GB to about 50 TB.
  - DAOD data sets are in the range of 1 GB to about several TB.
  - HITS data sets are in the order of several TBs.
- The ATLAS grid infrastructure consists of the Tier 0 computing site at CERN, 11 Tier 1s, 70 Tier 2s, 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 AOD files is available at the Tier 0 and Tier 1 sites.
- As a result of delays inflicted on the Long Shutdown 2 program by the COVID-19 pandemic, in June 2020 the CERN Directorate issued a revised plan for the start of Run 3. This plan foresees the re-start of LHC operations in February 2022, assuming that ATLAS can install its second New Small Wheel (NSW-C) during 2021. Run 3 will last until the end of 2024.
- All of the equipment needed for the HL-LHC, the LHC's successor, and its experiments will be installed during Long Shutdown 3, between 2025 and mid-2027. The HL-LHC is scheduled to come into operation at the end of 2027.
- The WLCG collects resources worldwide and enables their usage by the LHC experiments as a distributed computing facility. The mission of the WLCG is to provide global computing resources to store, distribute, and analyze the ~50–70 PB of data expected every year of operations from the LHC.
- The US ATLAS Tier 1 is hosted at BNL's SDDC. 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.
- There are four ATLAS Tier 2 centers in the United States: 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 by 2022, at the start of Run 3.

- ATLAS computational activities process data sets with sizes between 10 GB and 50 TB consisting of 10–10k event data files and file sizes up to 15 GBs each.
- In Run 3 starting in 2022, the number of derivation production formats will be significantly reduced and most of the analysis will use the common DAOD_PHYS format. Also, the number of AOD file replicas stored on disk will be reduced and read back in on demand from tape in derivation production. This will have a significant impact on the amount of data that must be sent across networks.
- ATLAS has a long history of successfully using HPC resources during Run 2 at the LHC. From 2016–2020, US-based HPC resources supported 10–25% of ATLAS simulation production. European HPC resources were also used, though mostly through grid interfaces. In the future, 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 its tier, and 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 data set catalog (for all data the experiment produces). Rucio stages data between facilities based on processing requests; approximately 1–2 PB per day are migrated worldwide in this manner. Rucio leverages other tools (FTS, etc.) to physically transfer the data sets.
- The 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 Tbps HTSN, which serves as the primary network transport for all data-intensive collaborations at BNL. It provides high-throughput connectivity to all HPC and HTC collaborations, and supports the timely transfer of large amounts of scientific data via the internet.
- 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. However, commercial cloud resources are being evaluated for specialized usage by analyzers. There are currently two proof of concept projects. If these projects are successful, ATLAS will require good network pipe between grid sites and commercial clouds. In this model, the network needs will be similar to university-based US Tier 2 sites. ATLAS expects a few PB of data transfers to cloud sites on a daily basis starting in 2021.
- 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 in 2022, it is anticipated that network needs will grow gradually. Increasing network capacity and performance will be needed at US ATLAS Tier 1, Tier 3, and Tier 3 AFs.

5.10.2 CMS Experiment Notes

The CMS experiment at the LHC is designed to probe new phenomena at the energy frontier. The LHC operates by streaming beams of protons which "collide," and then are observed via the affiliated detector experiments (of which CMS is one). The CMS collaboration is made up of more than 3,000 members from more than 50 countries. Researchers at US institutions comprise about 30% of the collaboration.

CMS is currently in a shutdown phase (2020–2022) and will resume Run 3 operations. These typically last for eight months of a year, for the time period between 2022 and 2024. Another shutdown will last between 2025 and 2027, and Run 4 (also known as the HL era where the LHC will receive major upgrades) will being in 2028 and run until 2030. Run 4 will see increases in data sizes by orders of magnitude beyond the prior runs.

CMS as a collaboration is focused heavily on research efforts to cope with the data demand, and is constantly looking into new ways to improve the core components of the research workflow (analysis, simulation, data sharing). It is expected that upcoming software will be adaptive to the challenges of the increase in data volumes, both by trying to use new file formats that are compact as well as leveraging both streaming and bulk-data movement approaches to cleanly and efficiently use network resources. Computation has traditionally followed a grid-computing model that is distributed worldwide at hundreds of sites, and will continue to do so into the future. Emerging use cases to leverage HPC facilities are very attractive, provided that some fundamental areas of friction can be addressed: porting of software, availability of network resources to support streaming workflows, and allocation of cycles that can be tied to the timelines of experimentation.

Several core areas of discussion emerged out of this section:

- CMS is divided into tiers of operation: CERN is considered the 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.
- CMS distinguishes three types of data: data from collisions in the accelerator, data from simulations, and user-produced data. The first two are centrally produced, while the third is produced by the physicists themselves as part of their analysis workflow.
- The United States operates one Tier 1 facility (Fermilab), which is responsible for 40% of CMS Tier 1 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 27PB of active disk available for use.
- The United States 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 approximately 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 (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.

- All CMS researchers can access storage via streaming or by grid analysis jobs. This is called any data, anywhere, anytime within CMS. Originally the tiers were hierarchical and flowed between adjacent levels only. This rigid concept has been eliminated over the last decade as network capacity and capability progressed. Data now flow across the full mesh, among all sites of all tiers.
- 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 today 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.
- CMS collision data operate at a typical cadence of three running years, followed by two shutdown years. Within a running year, collisions start approximately March to May, and end in November or December. The LHC provides data to CMS in roughly 10-hour data-taking periods with minimal downtime (typically a few hours) between periods, meaning there is a roughly constant stream of data.
- Run 2 produced approximately 45 PB of total data during the four years of operation, and a roughly similar set is expected for Run 3 as there were no major technology upgrades beyond changes to file formats on the analysis side. Run 4 will usher in a new era of scientific technology, and will see expectations of 350 PB per year starting in 2028.
- Near the end of a year, a "reprocessing" phase is typically performed, where raw data are repeated and run through the most recent software and analysis infrastructure to recreate experimental results. This is also performed at the end of the Run cycle, coinciding with experimental shutdown.
- CMS has produced simulations of roughly two to three times as many collisions and plans to continue this practice during Run 3. CMS has about 140 PB of Run 2 MC simulation data sets stored on tape (over four years).
- The centrally produced CMS data come in multiple formats ranging from the most versatile and complete (raw and AOD) to the easiest, smallest, and fastest to use (MiniAOD and NanoAOD). CMS introduced the NanoAOD format, a data format designed for interactive end-user analysis. It is expected that the adoption of this format will grow in Run 3 with the goal that 50% of CMS analyses are able to use the NanoAOD as their primary data tier by the end of Run 3, with another 40% or more to be based on MiniAOD. At HL-LHC scales, CMS may not be able to afford to keep AOD on disk anymore, given its size. In that scenario, access to AOD would require retrieval from archival storage.
- Data formats differ in the level of detail stored per collision. Raw data size is approximately 1 MB currently, and will grow to 6.5 MB during Run 4. AOD format data are reduced to approximately 400 KB, but will be approximately 2 MB during Run 4. MiniAOD is approximately 60 KB currently, and will grow to 250 KB during Run 4. Lastly, NanoAOD is approximately 1 KB in size, and will grow to 2 KB during Run 4.
- CMS uses "top-down data placement" at Tier 1 and Tier 2 centers combined with applications specifying the data set they need and being automatically routed and executed at the sites that have it. In this mode, all data access is local to the site via the site's LAN.
- CMS also performs "bottom-up data placement," as is implicit in caching. Here the applications are routed to sites with caches, applications access the cache locally, and cache misses are handled by the CMS XROOTD Data Federation (also referred to as the AAA federation).
- 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 any data, anywhere, anytime.

- CMS has adopted a dynamic DMS that attempts to keep "useful" data available based on-site availability, site resources, recently used data samples, and other experiment policies for data replication and data cleanup. This approach has added and deleted more than 40 PB of data from sites per month. As the rate of new subscriptions and deletions are similar, most of this operation consists of moving a data set from one site to another for operational reasons. Understanding and reducing unneeded data set transfers is important for CMS.
- FTS is used to manage scheduling and file transfer. For bulk transfers, CMS has historically used PhEDEx to handle transfers at the data set (i.e., groups-of-files) level. In November of 2020, CMS will switch to using Rucio instead of PhEDEx and Dynamo to manage data set storage and data set transfers (while still relying on FTS underneath).
- US CMS is in the process of retiring the use of gridFTP and replacing it with TPC https, implemented via XROOTD servers. Sites typically have multiple such servers that each provide 10 Gbps, and all have access to the same filesystem. Large bandwidth transfers are thus accomplished by orchestrating very 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.
  - Most traffic into and out of the US-CMS T1 is via the science data path(s). For CMS, those science data paths mean the LHCOPN and LHCONE. Fermilab's LHCOPN connectivity supports movement of raw data from the T0 (CERN), as well as production data movement with some of the other CMS T1s. Fermilab's LHCOPN configuration consists of three OSCARs circuits (primary, secondary, and tertiary) to CERN, which provide levels of redundancy with differing bandwidth guarantees for that traffic. LHCONE supports production data movement between the Fermilab Tier 1 and most CMS Tier 2s, as well with CMS T1s that do not use LHCOPN for T1-T1 data movement.
  - In terms of aggregate WAN capacity out of the site, Fermilab currently has three 100 Gb/s links to ESnet, via a geographically redundant metro ring. Two 100 Gb/s links are used to support the science data network paths, including LHCOPN and LHCONE.
  - Upgrade of the Laboratory network perimeter infrastructure to 400 GE technology is expected in the FY21 to FY22 time frame, likely to be aligned with availability of 400 GE services from ESnet. Additional WAN capacity from ESnet, either in the form of additional 100 GE WAN links or
    a 400 GE WAN link, will likely be needed as Run 3 commences.
- US CMS delegates network performance measurement collection to OSG. CMS expects all Tier 2s and the Tier 1 to keep up their perfSONAR instrumentation with the bandwidth requirements for the sites. All sites participate in perfSONAR measurements, which are archived by OSG at least 10 Gbps.
- 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. A notable exception would be the case of needed resource bursts for either CPU or network. The resources available for either CPU or TA networking in the cloud far exceed those available to CMS. 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 data volume that can be handled by the networks within and coming out of the CMS detector facility far exceeds what can be handled offline within current CPU, storage, and networking infrastructures.

- CMS is currently a major user of the TA 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 TA link is allowed (even if discouraged). If the current growth rate in TA link use by CMS continues, the size of the TA link becomes a major limitation already in Run 3.
- Failures in data streaming are a large source of job failures in CMS.
- Reliable and high-capacity streaming of input data, either raw or pileup simulation, would considerably reduce the disk requirements of CMS at HPC and other non-dedicated computing facilities.
- 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 approximately 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 are part of ongoing R&D.
- CMS has started internal efforts to validate and improve the transfer accounting in our software layers. The extent to which this should include traffic tagging and/or flow tagging is unclear to us at this point.
- CMS is participating in efforts including SENSE and AutoGOLE on how to transition to managed network usage for our production operations.
- CMS has traditionally treated the global network of sites as a mesh with identical links when it comes to bulk transfers. The data lake model makes clean regional distinctions. We expect that at least the existence of the Atlantic Ocean will become an architectural feature of our data distribution architecture.

5.10.3 LHC Operations Notes

This section will focus on the shared operational infrastructure of the experiments in the immediate to short term, namely the computational and storage infrastructure, software components, networking approaches, and R&D efforts underway during LS2 and those planned for Run 3. There are two main portions to this section:

- A list of infrastructure software products and tools that are relevant to data movement and/or access is presented. The list indicates how the various tools relate to the process of science in ATLAS and CMS as described in other parts of this report (Section 5.10.5 and Section 5.10.6) now and into the future (Section 5.10.8).
- An understanding of how network use is scaling is presented, contrasting it with what is deemed "affordable." A significant gap between needs projections based on past experience and projections of natural growth based on past investments into networking infrastructure expansion has been identified. The work in this area leads to a conclusion to fundamentally rethink the use of networking resources.

Several core areas of discussion emerged out of this section:

- 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 intra-resource 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 comes nominally for reasons of simplification: 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.
- Within CMS the majority of data orchestration over the wide-area network is performed with FTS or XROOTD (which utilize other tools such as PhEDEx and Dynamo management systems), and the Globus software toolkit/GridFTP for data movement. There will be a transition to Rucio data management tool (which will use HTTPS) in 2021.
- Data management in ATLAS is orchestrated by Rucio Distributed Data Management (DDM) system via FTS. Globus is used (as lower level to Rucio) for data transfer from HPC centers to US ATLAS sites.
- The challenges identified in operational monitoring of the network has led to the creation of a working group focused on Packet Marking. The goal is to be able to mark network packets by owner and purpose, enabling identification and accounting of traffic anywhere along the network path.
- The data from perfSONAR, as well as additional network-related data, are being gathered by OSG/WLCG and sent to an analytics platform at the University of Chicago. The data are stored in Elasticsearch and publicly accessible via Kibana dashboards.
- The experiments are performing R&D for situations with constrained network resources and potentially intelligent network services. The 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.
- Up to now, the experiments have treated the wide-area network as an appliance with almost infinite capacity, the only counterexamples being known poor connections to computing centers in isolated areas. Network capabilities have become more and more an integral part of the computing model via the use of tools that can stream data whenever needed (e.g., any data, anywhere, anytime).
- 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.

5.10.4 HL Era of the LHC Notes

This section focuses on the shared vision for the LHC, and its associated experiments, as they undergo a major upgrade in the next six years, leading to HL-LHC operations around 2027.

The LHC collides protons more than a billion times every second, out of which ATLAS and CMS will select interesting collisions for recording at a rate of 10,000 times every second. With over 2,000 hours of data collection every year when the HL-LHC starts running in 2027, 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 of LHC running as the entire period of running of the LHC has produced up to the start. This implies that the science capabilities are expected to be roughly equivalent to the data-taking from 2010–2024, or runs 1, 2, and 3 combined. The entire HL-LHC era will last for 10 years of data taking, with 12–24 month maintenance periods interspersed roughly every 3 years.

Several core areas of discussion emerged out of this section [ATLAS]:

- The current members of the ATLAS collaboration are not expected to change in a significant manner for the HL-LHC. The collaborators will continue to be distributed worldwide, which puts strong requirements on global networking to accomplish computation.
- The exact computing model for the HL-LHC has not been finalized. It is assumed that the current ATLAS computing model will be the baseline model, with minor improvements. There will be no fixed hierarchy of computing sites for most data processing and data access services. In order to improve usage efficiency, sites will be primarily categorized by size, service level, and capability. Hence the location of data sets and users is not deterministic.
- ATLAS expects worldwide distribution of all resources and users. ATLAS also expects about seven large sites in the United States, which will all be required to have a full range of distributed computing capabilities. They will store both primary and secondary data, will provide access to hundreds of users, and will participate in continuous data transfers. These sites will include the BNL Tier 1, the current Tier 2s (Great Lakes, Midwest, Northeast, and Southwest), SLAC, and a few HPCs. It is expected that a few hundred Gbps links will be needed between them when the HL-LHC starts. BNL Tier 1 will need additional capacity to handle worldwide traffic. The network capacity should be provisioned to match the scale of the available resources at each site.
- The impact of HL-LHC on ATLAS 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 HPC will have an impact on the compute resources (storage and networking). These HPC centers are increasing in computing power and several exaflop 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 ATLAS data centers for subsequent processing.
- The HL-LHC (i.e., Run 4) will start in 2027. The physics events that drive the experiment will be collected at a rate ten times more than during previous runs. There will be challenges involved in collecting, storing, reconstructing, and analyzing the data volume; it is expected that MC simulation events will need to be produced in similar numbers in the preceding years.
- Networking has been fundamental to the success of ATLAS and 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.

- As part of this optimization, the ability to mark network packets to identify sources of traffic, and to route traffic to control speed and cost, may become vital.
- Economizing storage is an important goal for HL-LHC computing. Unlike CPU, storage needs will continue to increase during the lifetime of the HL-LHC. 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.
- Remote data delivery reliant on powerful networks will in general be essential to minimize data replication and disk storage footprint while fully utilizing distributed processing resources. In order to most efficiently use bandwidth and minimize latencies, ATLAS is developing new services and workflows to deliver across the network only the data needed by the consuming workload.
- The ATLAS Distributed Computing (ADC) system and organization today is a sophisticated ensemble of software systems, computing facilities, and people.
- The HL-LHC directed capabilities and workflows are being developed on the foundation of the ADC system, particularly the PanDA workload management system, ProdSys production management system, and the Rucio DDM system.
- 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 multiple 100 G scale (200–800 Gbps). Network use will be at least a factor of 10 larger than Run 2.
- For HL-LHC, four main requirements have been identified:
  - **Capacity:** Run 3 is moving to multiple 100G 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.
- While PanDA+Rucio can use commercial cloud resources interchangeably with grid-based WLCG resources, currently ATLAS has no plans to use clouds for the HL-LHC. The baseline plan is to use grid and HPC resources. If some grid resources are set up as cloud resources, they will also be used. However, commercial cloud resources are being evaluated for specialized usage by analyzers.
- Joint ATLAS and CMS use of Rucio for DDM 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 QoS or deadline requirements.

Several core areas of discussion emerged out of this section [CMS]:

- Each data-taking year, the experiments, ATLAS and CMS combined, are expected to accumulate roughly 1 EB of new data.
- Both experiments make the same assumptions around how often data must be processed and reprocessed over the course of a year.
- 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 lakes model." A 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.
- US CMS currently assumes that all disk storage at the Tier 1 and Tier 2s will be part of a single US data lake.
- 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.
- 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 shared (set of) 100 Gbps link(s), shared with the entire institution. It is common that the US CMS LHC program dominates the WAN link use at the Tier 2 institutions. This may change in the future.
- Tier 2s that are part of the data lake origin will be required to provide guaranteed, managed, and possibly scheduled burst capacity at up to 400 Gbps to support large-scale ingests over the course of hours to a day (400 Gbps for a day is roughly 4 PB of data).
- US CMS Tier 2s each provided 5 PB of usable storage in 2020. By 2028, 4 PB of storage is likely to be a minor fraction of the origin storage of a Tier 2 that provides origin storage to the US data lake.
- In 2020, all US CMS Tier 2s provided roughly the same functions and capacity. This may no longer be the case for the HL-LHC era. It is thus conceivable that not all Tier 2s in US CMS will provide origin storage for the US CMS data lake during the HL-LHC era.
- The LHC experiments are planning for 100 Gbps sustained network use for all Tier 2s, with occasional bursts to 400 Gbps, throughout the first run of the HL-LHC.
- A consideration of network needs will be those required to support distributed physics analysis, which is expected to be centered at a variety of AFs. These are dedicated pieces of infrastructure designed to provide access to large data sets and computational resources that enable rapid iterative analysis of physics data.
- 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 oceans. 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.
- The Tier 1 at Fermilab will require Tbps burst capabilities. Steady state network bandwidth consumption is expected to be between 200–300 Gbps, at a minimum.
- Tier 2s will require 400 Gbps burst capabilities. Steady state network bandwidth consumption is expected to be approximately 100 Gbps.

- The large exascale HPC centers funded by the DOE will require Tbps burst capabilities in order for CMS 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.
- As Tier 3 systems are smaller in scale, or CMS allocations on big systems are smaller in scale, networking requirements are expected to be more modest.
- US CMS wants to be able to account for the bulk of the usage of LAN and WAN networking resources. For WAN resources, there is a desire to reason at a high level about why the network is used, at what capacity, and when; there is a desire to plan and manage bandwidth use. The overall goal is to understand the requirements related to capacity and capability in a manner that is compatible with other network use, both in the core and at the edges (at the Tier 1 and Tier 2s).
- US CMS expects computational nodes to be connected at 10 Gbps, data nodes at up to 100 Gbps (depending on size), campus networking to institutional boundaries at Tier 2s to reach multiple 100 Gbps, and for the Tier 1 at Fermilab to reach 500 Gbps to Tbps.
- It is expected that the Chicago MAN-link will provide Tbps to Starlight, and ideally Tbps across the Atlantic to CERN.
- 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 of US CMS.
- It is expected that these facilities will share bandwidth in LHCONE in between each other, and to the Fermilab Tier 1.
- US CMS will continue to collaborate and share with ATLAS, as well as other science projects. Sharing network bandwidth with ATLAS and other science projects is expected. Given the large burst needs, network management will be a core concern and area of research.
- All relevant network traffic that will be accounted for will be performed by either FTS or XROOTD infrastructure software. Any transfers between data lakes, as well as all output handling of central production workflows, will involve FTS. All data streaming to applications from either caches or data origins inside the lake will involve XROOTD.
- US CMS delegates network performance measurement collection to OSG, and expects to continue to do so. This is done with perfSONAR measurement; all Tier 2s and the Tier 1 will keep up their perfSONAR instrumentation.
- We have evaluated the use of commercial cloud both for processing and for TA transfer. US CMS finds that both are not cost-effective, at present. Experimentation has shown that making large-scale use of cloud resources if the cost structure were to change is possible.

5.10.5 ATLAS Experiment Case Study

5.10.5.1 Background

The LHC at the European Laboratory for Particle Physics 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.

ATLAS is the largest of four particle detectors that measure and record the particle collisions at the LHC. The primary scientific goal is to quantitatively measure and discover properties of the SM of particle physics. ATLAS

took a giant step in this with the discovery of the Higgs boson in 2012, which led to the 2013 Nobel Prize in Physics for Peter Higgs and Francois Englert. In addition to the paper announcing the discovery of the Higgs boson, ATLAS has published over a thousand new results in particle physics. A major emphasis of the upcoming Run 3 at the LHC will be the precision measurement of Higgs properties.

The SM of particle physics is considered to be an incomplete theory. In order to explain many observations and measurements like dark matter and Higgs mass hierarchy, new phenomenology remains to be discovered experimentally. ATLAS physics goals combine a strong program of SM measurements with search for new phenomena like Supersymmetry.

5.10.5.2 Collaborators

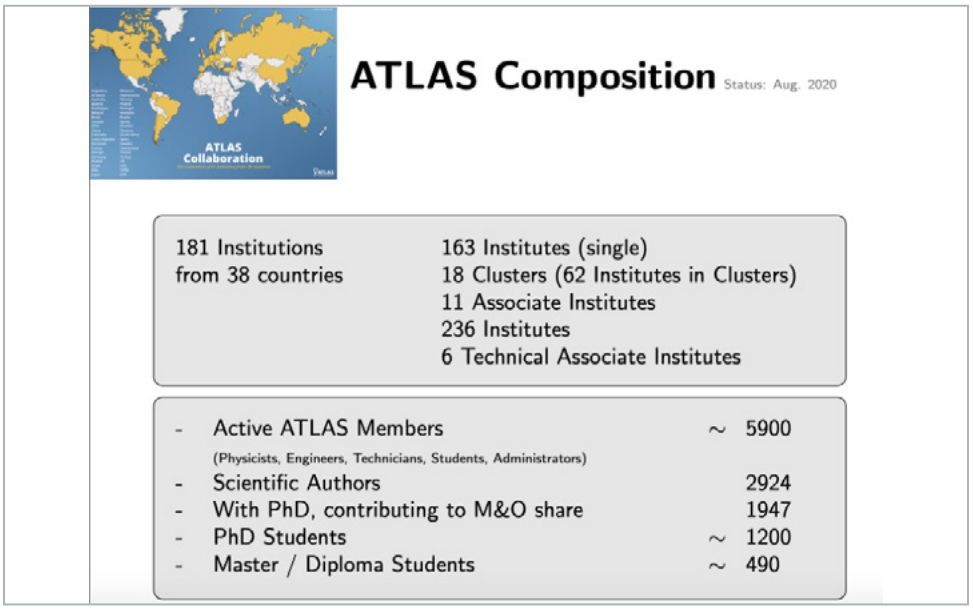


Figure 48: ATLAS composition

ATLAS composition, as shown in **Figure 48,** is worldwide. A full list of ATLAS collaborators is as follows:

- Algeria
  - Department of Physics, University of Jijel, Jijel, Algeria
- Argentina
  - Departamento de Física, Universidad de Buenos Aires, Buenos Aires
  - Instituto de Física La Plata, Universidad Nacional de La Plata and CONICET, La Plata
- Armenia
  - Yerevan Physics Institute, Yerevan
- Australia
  - Department of Physics, University of Adelaide, Adelaide School of Physics, University of Sydney, Sydney
  - School of Physics, University of Melbourne, Victoria

- Austria
  - Institut für Astro- und Teilchenphysik, Leopold-Franzens-Universität, Innsbruck
  - Fachhochschule Wiener Neustadt, Wiener Neustadt
- Azerbaijan
  - Institute of Physics, Azerbaijan Academy of Sciences, Baku
- Belarus
  - B.I. Stepanov Institute of Physics, National Academy of Sciences of Belarus, Minsk Research Institute for Nuclear Problems of Byelorussian State University, Minsk
- Brazil
  - Brazil Cluster: Departamento de Engenharia Elétrica, Universidade Federal de Juiz de Fora (UFJF), Juiz de Fora; Universidade Federal do Rio De Janeiro COPPE/EE/IF, Rio de Janeiro; Universidade Federal de São João del Rei (UFSJ), São João del Rei; Instituto de Física, Universidade de São Paulo, São Paulo
- Canada
  - Department of Physics, Simon Fraser University, Burnaby BC
  - Department of Physics, University of Alberta, Edmonton AB
  - Department of Physics, McGill University, Montreal QC
  - Group of Particle Physics, University of Montreal, Montreal QC
  - Department of Physics, Carleton University, Ottawa ON
  - Department of Physics, University of Toronto, Toronto ON
  - Department of Physics, University of British Columbia, Vancouver BC
  - TRIUMF, Vancouver BC; Department of Physics and Astronomy, York University, Toronto ON
  - Department of Physics and Astronomy, University of Victoria, Victoria BC
- CERN
  - European Organization for Nuclear Research, Geneva, Switzerland
- Chile
  - Chile Cluster: Departamento de Física, Pontificia Universidad Católica de Chile, Santiago; Universidad Andres Bello, Department of Physics, Santiago; Instituto de Alta Investigación, Universidad de Tarapacá; Departamento de Física, Universidad Técnica Federico Santa María, Valparaíso
- China
  - China IHEP-NJU-THU Cluster: Institute of High-Energy Physics, Chinese Academy of Sciences, Beijing; Physics Department, Tsinghua University, Beijing; Department of Physics, Nanjing University, Nanjing
  - China USTC-SDU-SJTU Cluster: Department of Modern Physics and State Key Laboratory of Particle Detection and Electronics, University of Science and Technology of China, Hefei; Institute of Frontier and Interdisciplinary Science and Key Laboratory of Particle Physics and Particle Irradiation , Shandong University, Qingdao; School of Physics and Astronomy, Shanghai Jiao Tong University, KLPPAC-MoE, SKLPPC, Shanghai; Tsung-Dao Lee Institute, Shanghai Hong Kong Cluster: Department of Physics, Chinese

University of Hong Kong, Shatin, N.T., Hong Kong; Department of Physics, University of Hong Kong, Hong Kong; Department of Physics and Institute for Advanced Study, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

- Colombia
  - Colombia Cluster: Facultad de Ciencias y Centro de Investigaciónes, Universidad Antonio Nariño, Bogotá; Departamento de Física, Universidad Nacional de Colombia, Bogotá, Colombia
- Czech Republic
  - Palacký University, RCPTM, Joint Laboratory of Optics, Olomouc
  - Charles University, Faculty of Mathematics and Physics, Prague
  - Czech Technical University in Prague, Prague
  - Institute of Physics of the Czech Academy of Sciences, Prague
- Denmark
  - Niels Bohr Institute, University of Copenhagen, Copenhagen
- France
  - LAPP, Université Grenoble Alpes, Université Savoie Mont Blanc, CNRS/IN2P3, Annecy
  - LHC Physics Center (LPC), Université Clermont Auvergne, CNRS/IN2P3, Clermont-Ferrand
  - IRFU, CEA, Université Paris-Saclay, Gif-sur-Yvette
  - LPSC, Université Grenoble Alpes, CNRS/IN2P3, Grenoble INP, Grenoble
  - CPPM, Aix-Marseille Université, CNRS/IN2P3, Marseille 2
  - IJCLab, Université Paris-Saclay, CNRS/IN2P3, 91405, Orsay
  - LPNHE, Sorbonne Université, Université de Paris, CNRS/IN2P3, Paris
- Georgia
  - Georgia Cluster: E. Andronikashvili Institute of Physics, Iv. Javakhishvili Tbilisi State University, Tbilisi; High-Energy Physics Institute, Tbilisi State University, Tbilisi
- Germany
  - Institut für Physik, Humboldt Universität zu Berlin, Berlin
  - Physikalisches Institut, Universität Bonn, Bonn
  - Lehrstuhl für Experimentelle Physik IV, Technische Universität Dortmund, Dortmund
  - Institut für Kern- und Teilchenphysik, Technische Universität Dresden, Dresden
  - Physikalisches Institut, Albert-Ludwigs-Universität Freiburg, Freiburg
  - II. Physikalisches Institut, Justus-Liebig-Universität Giessen, Giessen
  - II. Physicalists Institute, Georg-August-Universität Göttingen, Göttingen
  - DESY, Hamburg and Zeuthen
  - Kirchhoff-Institut für Physik, Ruprecht-Karls-Universität Heidelberg, Heidelberg; Physikalisches Institut, Ruprecht-Karls-Universität Heidelberg, Heidelberg
  - Institut für Physik, Universität Mainz, Mainz
  - Fakultät für Physik, Ludwig-Maximilians-Universität München, München

- Max-Planck-Institut für Physik (Werner-Heisenberg-Institut), München
- Department Physik, Universität Siegen, Siegen
- Fakultät für Mathematik und Naturwissenschaften, Fachgruppe Physik, Bergische Universität Wuppertal, Wuppertal
- Fakultät für Physik und Astronomie, Julius-Maximilians-Universität Würzburg, Würzburg

- Greece
  - National Centre for Scientific Research "Demokritos," Agia Paraskevi
  - Physics Department, National and Kapodistrian University of Athens, Athens
  - Department of Physics, Aristotle University of Thessaloniki, Thessaloniki
  - Physics Department, National Technical University of Athens, Zografou
- Israel
  - Department of Physics, Technion, Israel Institute of Technology, Haifa
  - Department of Particle Physics, Weizmann Institute of Science, Rehovot
  - Raymond and Beverly Sackler School of Physics and Astronomy, Tel Aviv University, Tel Aviv
- Italy
  - INFN Sezione di Bologna; INFN Bologna and Università di Bologna, Dipartimento di Fisica INFN e Laboratori Nazionali di Frascati, Frascati
  - INFN Sezione di Genova; Dipartimento di Fisica, Università di Genova, Genova
  - INFN Sezione di Lecce; Dipartimento di Matematica e Fisica, Università del Salento, Lecce INFN Sezione di Milano; Dipartimento di Fisica, Università di Milano, Milano
  - INFN Sezione di Napoli; Dipartimento di Fisica, Università di Napoli, Napoli
  - INFN Sezione di Pavia; Dipartimento di Fisica, Università di Pavia, Pavia
  - INFN Sezione di Pisa; Dipartimento di Fisica E. Fermi, Università di Pisa, Pisa
  - INFN Gruppo Collegato di Cosenza, Laboratori Nazionali di Frascati; Dipartimento di Fisica, Università della Calabria, Rende
  - INFN Sezione di Roma; Dipartimento di Fisica, Sapienza Università di Roma, Roma
  - INFN Sezione di Roma Tor Vergata; Dipartimento di Fisica, Università di Roma Tor Vergata, Roma
  - INFN Sezione di Roma Tre; Dipartimento di Matematica e Fisica, Università Roma Tre, Roma
  - INFN-TIFPA; Università degli Studi di Trento, Trento
  - INFN Gruppo Collegato di Udine, Sezione di Trieste; ICTP, Trieste; INFN Gruppo Collegato di Udine, Sezione di Trieste; Dipartimento Politecnico di Ingegneria e Architettura, Università di Udine, Udine

- Japan
  - Research Center for Advanced Particle Physics and Department of Physics, Kyushu University, Fukuoka
  - Graduate School of Science, Kobe University, Kobe
  - Faculty of Science, Kyoto University, Kyoto
  - Kyoto University of Education, Kyoto
  - Department of Physics, Shinshu University, Nagano
  - Graduate School of Science and Kobayashi-Maskawa Institute, Nagoya University, Nagoya Graduate School of Science, Osaka University, Osaka
  - Department of Physics, Tokyo Institute of Technology, Tokyo
  - Graduate School of Science and Technology, Tokyo Metropolitan University, Tokyo International Center for Elementary Particle Physics and Department of Physics, University of Tokyo, Tokyo
  - Ochanomizu University, Otsuka, Bunkyo-ku, Tokyo
  - Waseda University, Tokyo
  - Division of Physics and Tomonaga Center for the History of the Universe, Faculty of Pure and Applied Sciences, University of Tsukuba, Tsukuba
  - KEK, High-Energy Accelerator Research Organization, Tsukuba
- Morocco
  - Morocco Cluster: Faculté des Sciences Ain Chock, Réseau Universitaire de Physique des Hautes Energies — Université Hassan II, Casablanca; Faculté des Sciences, Université Ibn-Tofail, Kénitra; Faculté des Sciences Semlalia, Université Cadi Ayyad, LPHEA-Marrakech; Faculté des Sciences, Université Mohamed Premier and LPTPM, Oujda; Faculté des sciences, Université Mohammed V, Rabat
- Netherlands
  - Nikhef National Institute for Subatomic Physics and University of Amsterdam, Amsterdam Institute for Mathematics, Astrophysics and Particle Physics, Radboud University Nijmegen/Nikhef, Nijmegen
- Norway
  - Department for Physics and Technology, University of Bergen, Bergen Department of Physics, University of Oslo, Oslo
- Poland
  - Institute of Nuclear Physics Polish Academy of Sciences, Krakow
  - AGH University of Science and Technology, Faculty of Physics and Applied Computer Science, Krakow; Marian Smoluchowski Institute of Physics, Jagiellonian University, Krakow
- Portugal
  - Portugal Cluster: Laboratório de Instrumentação e Física Experimental de Partículas — LIP, Lisboa; Departamento de Física, Faculdade de Ciências, Universidade de Lisboa, Lisboa; Departamento de Física, Universidade de Coimbra, Coimbra; Departamento de Física, Universidade do Minho, Braga; Departamento de Física Teórica y del Cosmos, Universidad de Granada, Granada (Spain); Dep Física and CEFITEC of Faculdade de

Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica; Instituto Superior Técnico, Universidade de Lisboa, Lisboa

- Romania
  - Romania Cluster: Transilvania University of Brasov, Brasov; Horia Hulubei National Institute of Physics and Nuclear Engineering, Bucharest; National Institute for Research and Development of Isotopic and Molecular Technologies, Physics Department, Cluj-Napoca; Department of Physics, Alexandru Ioan Cuza University of Iasi, Iasi; University Politehnica Bucharest, Bucharest; West University in Timisoara, Timisoara
- Russia
  - D.V. Skobeltsyn Institute of Nuclear Physics, M.V. Lomonosov Moscow State University, Moscow
  - Institute for Theoretical and Experimental Physics named by A.I. Alikhanov of National Research Centre "Kurchatov Institute", Moscow
  - National Research Nuclear University MEPhI, Moscow
  - P.N. Lebedev Physical Institute of the Russian Academy of Sciences, Moscow
  - Novosibirsk State University Novosibirsk; Budker Institute of Nuclear Physics and NSU, SB RAS, Novosibirsk
  - Institute for High-Energy Physics of the National Research Centre Kurchatov Institute, Protvino
  - Konstantinov Nuclear Physics Institute of National Research Centre "Kurchatov Institute", PNPI, St. Petersburg
  - Tomsk State University, Tomsk
  - JINR
  - Joint Institute for Nuclear Research, Dubna, Russia
- Serbia
  - Institute of Physics, University of Belgrade, Belgrade
- Slovak Republic
  - Slovak Republic Cluster: Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava; Department of Subnuclear Physics, Institute of Experimental Physics of the Slovak Academy of Sciences, Kosice
- Slovenia
  - Department of Experimental Particle Physics, Jožef Stefan Institute and Department of Physics, University of Ljubljana, Ljubljana
- South Africa
  - South Africa Cluster: Department of Physics, University of Cape Town, Cape Town; Department of Mechanical Engineering Science, University of Johannesburg, Johannesburg; University of South Africa, Department of Physics, Pretoria; School of Physics, University of the Witwatersrand, Johannesburg; iThemba Labs, Western Cape

- Spain
  - Institut de Física d'Altes Energies (IFAE), Barcelona Institute of Science and Technology, Barcelona
  - Departamento de Física Teorica C-15 and CIAFF, Universidad Autónoma de Madrid, Madrid
  - Instituto de Física Corpuscular (IFIC), Centro Mixto Universidad de Valencia — CSIC, Valencia
- Sweden
  - Fysiska institutionen, Lunds universitet, Lund
  - Department of Physics, Stockholm University; Oskar Klein Centre, Stockholm Physics Department, Royal Institute of Technology, Stockholm
  - Department of Physics and Astronomy, University of Uppsala, Uppsala
- Switzerland
  - Albert Einstein Center for Fundamental Physics and Laboratory for High-Energy Physics, University of Bern, Bern
  - Département de Physique Nucléaire et Corpusculaire, Université de Genève, Genève
- Taiwan
  - Department of Physics, National Tsing Hua University, Hsinchu Institute of Physics, Academia Sinica, Taipei
- Turkey
  - Ankara Cluster: Department of Physics, Ankara University, Ankara; Istanbul Aydin University, Application and Research Center for Advanced Studies, Istanbul; Division of Physics, TOBB University of Economics and Technology, Ankara
  - Bogazici Cluster: Bahcesehir University, Faculty of Engineering and Natural Sciences, Istanbul; Istanbul Bilgi University, Faculty of Engineering and Natural Sciences, Istanbul; Department of Physics, Bogazici University, Istanbul; Department of Physics Engineering, Gaziantep University, Gaziantep
- United Kingdom
  - School of Physics and Astronomy, University of Birmingham, Birmingham
  - Department of Physics and Astronomy, University of Sussex, Brighton
  - Cavendish Laboratory, University of Cambridge, Cambridge
  - Department of Physics, University of Warwick, Coventry
  - Particle Physics Department, Rutherford Appleton Laboratory, Didcot
  - SUPA — School of Physics and Astronomy, University of Edinburgh, Edinburgh
  - Department of Physics, Royal Holloway University of London, Egham
  - SUPA — School of Physics and Astronomy, University of Glasgow, Glasgow
  - Physics Department, Lancaster University, Lancaster
  - Oliver Lodge Laboratory, University of Liverpool, Liverpool
  - Department of Physics and Astronomy, University College London, London
  - School of Physics and Astronomy, Queen Mary University of London, London

  - School of Physics and Astronomy, University of Manchester, Manchester
  - Department of Physics, Oxford University, Oxford
  - Department of Physics and Astronomy, University of Sheffield, Sheffield
- United States of America
  - Physics Department, SUNY Albany, Albany, NY
  - Department of Physics and Astronomy, University of New Mexico, Albuquerque, NM, Department of Physics and Astronomy, Iowa State University, Ames IA
  - Department of Physics, University of Massachusetts, Amherst, MA
  - Department of Physics, University of Michigan, Ann Arbor, MI
  - High-Energy Physics Division, ANL, Argonne, IL
  - Department of Physics, University of Texas at Arlington, Arlington, TX
  - Department of Physics, University of Texas at Austin, Austin, TX
  - Physics Division, LBNL and University of California, Berkeley, California
  - Department of Physics, Indiana University, Bloomington, IN
  - Department of Physics, Boston University, Boston, MA
  - California State University, California
  - Laboratory for Particle Physics and Cosmology, Harvard University, Cambridge, MA
  - Enrico Fermi Institute, University of Chicago, Chicago, IL
  - Ohio State University, Columbus, OH
  - Physics Department, Southern Methodist University, Dallas, TX
  - Department of Physics, Northern Illinois University, DeKalb, IL
  - Department of Physics, Duke University, Durham, NC
  - Department of Physics and Astronomy, Michigan State University, East Lansing, MI
  - University of Iowa, Iowa City, IA
  - Department of Physics and Astronomy, University of California Irvine, Irvine, California
  - Nevis Laboratory, Columbia University, Irvington, NY
  - Department of Physics, University of Wisconsin, Madison, WI
  - Department of Physics and Astronomy, Tufts University, Medford, MA
  - Department of Physics, Yale University, New Haven, CT
  - Department of Physics, New York University, New York, NY
  - Homer L. Dodge Department of Physics and Astronomy, University of Oklahoma, Norman, OK
  - Institute for Fundamental Science, University of Oregon, Eugene, OR
  - Department of Physics, University of Pennsylvania, Philadelphia, PA
  - Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, PA
  - Physics Department, University of Texas at Dallas, Richardson, TX
  - Louisiana Tech University, Ruston, LA
  - Santa Cruz Institute for Particle Physics, University of California, Santa Cruz, Santa Cruz, California

- Department of Physics, University of Washington, Seattle, WA
- SLAC, Stanford, California
- Department of Physics, Oklahoma State University, Stillwater, OK
- Departments of Physics and Astronomy, Stony Brook University, Stony Brook, NY
- Department of Physics, University of Arizona, Tucson, Arizona
- Physics Department, BNL, Upton, NY
- Department of Physics, University of Illinois, Urbana, IL
- Department of Physics, Brandeis University, Waltham, MA

User/collaborator and location	Primary or secondary copy of the data	Data access method	Data access method	Frequency of data transfer or download	Are data sent back to the source and method	Any known issues with data sharing
INTERNATIONAL (OVER 100 SITES IN MORE THAN 30 COUNTRIES)	Yes	Data transfer	In range of 10 GB to 50 TB	Continuously	Yes, see note below	Long tails from distributed transfers and access
US — BNL TIER 1 BNL UPTON	Primary	Asynchronous data transfer, direct access	Same	Continuously	Same	
US — AGLT2 (U MICHIGAN ANN ARBOR, MICHIGAN STATE LANSING)	Primary	Asynchronous data transfer, direct access	Same	Continuously	Same	
US — MWT2 (U CHICAGO, U INDIANA BLOOMINGTON, U ILLINOIS URBANA)	Primary	Asynchronous data transfer, direct access	Same	Continuously	Same	
US — NET2 (BOSTON U, HARVARD U)	Primary	Asynchronous data transfer, direct access	Same	Continuously	Same	
US — SWT2 (UT ARLINGTON, OKLAHOMA STATE)	Primary	Asynchronous data transfer, direct access	Same	Continuously	Same	
US — SLAC STANFORD (SHARED TIER 3 AF)	Secondary	Asynchronous data transfer, direct access	Same	Continuously	Same	
US — SUNY ALBANY	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US — U NEW MEXICO ALBUQUERQUE	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US — IOWA STATE	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US — UMASS AMHERST	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US — ANL, ARGONNE	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US — UT AUSTIN	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US — LBNL, BERKLEY	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US — CALIFORNIA STATE	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US — OHIO STATE, COLUMBUS	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US — SOUTHERN METHODIST DALLAS	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US — NORTHERN ILLINOIS, DEKALB	Tertiary	Data transfer (pull)	Same	Occasionally	Same	

User/collaborator and location	Primary or secondary copy of the data	Data access method	Data access method	Frequency of data transfer or download	Are data sent back to the source and method	Any known issues with data sharing
US — DUKE, DURHAM	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US — UC IRVINE	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US — COLUMBIA, NY	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US — WISCONSIN-MADISON	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US — TUFTS, MEDFORD	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US — YALE NEW HAVEN	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US — NYU NEW YORK	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US — U OKLAHOMA NORMAN	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US — OREGON EUGENE	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US — U PENNSYLVANIA PHILADELPHIA	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US — U PITTSBURGH	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US — U TEXAS DALLAS	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US — LOUISIANA TECH RUSTON	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US — UC SANTA CRUZ	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US — U WASHINGTON SEATTLE	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US — OKLAHOMA STATE STILLWATER	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US — SUNY STONY BROOK	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US — U ARIZONA TUCSON	Tertiary	Data transfer (pull)	Same	Occasionally	Same	
US — BRANDEIS U WALTHAM	Tertiary	Data transfer (pull)	Same	Occasionally	Same	

Table 18: ATLAS data projections

Data sets in ATLAS are collections of files organized by category/workflow. **Table 18** shows the breadth and depth of ATLAS data set sizes. Due to the distributed nature of computation, these figures represent an average view of data transfer, storage, and computation, since it is not possible for every site to have a complete view of the ATLAS data catalog. Data sets are the fundamental units in ATLAS, fully integrated into the Rucio DMS, and used for workflow, transfers, and user analysis. Individual data sets vary largely in size:

- Raw data sets are in the range of 1 to about 50 TB.
- AOD data sets are in the range of 1 GB to about 50 TB.
- DAOD data sets are in the range of 1 GB to about several TB.
- HITS data sets are in the order of several TBs.

The ATLAS grid infrastructure consists of the Tier 0 computing site at CERN, 11 Tier 1s, 70 Tier 2s, 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 AOD files is available at the Tier 0 and Tier 1 sites. The US Tier 1 and Tier 2 sites are described in Section 5.10.5.3. Lots of computing usage broken down by activity/workflow (data processing, MC, analysis etc.) is shown in Section 5.10.5.4.1.

The setup and scope of Tier 3 sites vary largely. They are dedicated AFs with all grid components at SLAC and BNL. Similar activities at a reasonable fraction of a Tier 2 site can be expected here. Tier 3 sites at most universities are usually smaller local CPU clusters with several TBs of disk storage attached. At these sites individual end-user analysis happens without any production activity.

Depending on the size of the site (irrespective of tier label), there are 100 to 20k CPUs and 0.5 to 20 PB of disk storage available. In total, ATLAS had a pledged disk space of 230 PB and a tape space of 320 PB at all sites worldwide combined in 2020. There was a CPU pledge of approximately 3500 kHS06 (about 400k batch cores) for the full year 2020. ATLAS requests additional CPU resources "beyond pledge" from all sites to generate simulated events which are crucial for physics studies. On average, a substantial amount (~1600 kHS06) is delivered as beyond pledge resource; 25% of this came from HPC sites in the past year. In terms of number of cores, the baseline is approximately 400k CPU cores constantly used worldwide. Further details about the breakdown of the number of cores used concurrently by different workflows at US facilities is presented in Section 5.10.5.4.

Figure 49 shows the total data set volume on disk and tape by data type.

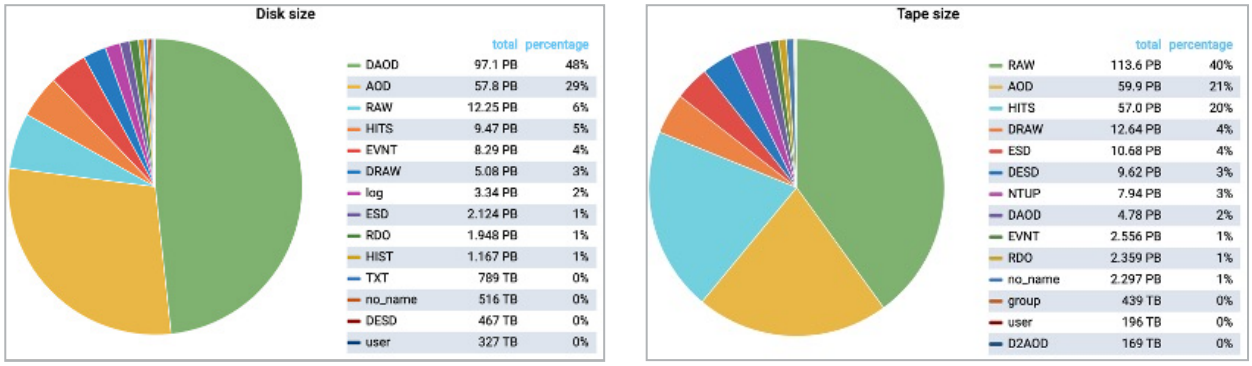


Figure 49: ATLAS data set volume by storage type

5.10.5.3 Instruments and Facilities

5.10.5.3.1 LHC

The LHC collides protons more than a billion times every second, out of which ATLAS selects interesting collisions for recording at a rate of thousand times every second. With over 2,000 hours of data collection every year when the LHC is running, ATLAS has a huge data sample for physicists to analyze from the completed Run 1 and Run 2. Run 3 is scheduled to start data collection in 2022. **Table 19** shows typical LHC parameters.

Quantity	Number
Circumfernce	26659 m
Dipole operting temperature	1.9 K (-271.3°C)
Number of magnets	9593
Number of main dipoles	1232
Number of main quadrupoles	392
Number of RF cavities	8 per beam
Nominal energy, protons	6.5 teV
Nominal energy, ions	2.56 TeV/u (energy per nucleon)

Quantity	Number
Nominal energy, protons collisions	13 TeV
Number of bunches per proton beam	2808
Number of protons per bunch (at start)	1.2×10^{11}
Number of turns per second	11245
Number of collisions per second	1 billion

Table 19: LHC parameters

As a result of delays inflicted on the Long Shutdown 2 program by the COVID-19 pandemic, in June 2020 the CERN Directorate issued a revised plan for the start of Run 3. This plan foresees the re-start of LHC operations in February 2022, assuming that ATLAS can install its second New Small Wheel (NSW-C) during 2021. Run 3 will last until the end of 2024. All of the equipment needed for the HL-LHC, the LHC's successor, and its experiments will be installed during Long Shutdown 3, between 2025 and mid-2027. The HL-LHC is scheduled to come into operation at the end of 2027.

In terms of computing, the main impact is the absence of any data taking in 2021, and more data taking in 2022 than was previously envisaged. About 2.5 months of stable beam time are expected for 2022 (with a duty cycle of 50%), with average pileup to be $<\mu>=35$, leading to at most 70 fb-1 of delivered integrated luminosity. Additionally, about two weeks of heavy-ion collisions are expected. In 2022 the average pileup is expected to be close to that of Run 2, while it may rise significantly in 2023–2024. For the heavy-ion running, the same running conditions as in 2018 are assumed.

5.10.5.3.2 ATLAS

ATLAS[207] is the largest of four particle detectors that measure and record the particle collisions at the LHC. Selected highlights from the ATLAS website are shown below. The primary scientific goal is to combine a strong program of SM measurements with search for new phenomena like Supersymmetry:

- The four major components of the ATLAS detector are the Inner Detector, the calorimeter, the Muon Spectrometer, and the Magnet System. Integrated with the detector components are the TriDAS, a specialized multi-level computing system, which selects physics events with distinguishing characteristics; and the computing system, which develops and improves computing software used to store, process and analyze vast amounts of collision data at computing centers worldwide.
- ATLAS is designed to observe up to 1.7 billion proton-proton collisions per second, with a combined data volume of more than 60 million megabytes per second. However, only some of these events will contain interesting characteristics that might lead to new discoveries. To reduce the flow of data to manageable levels, ATLAS uses a specialized two-level online event selection system — the trigger system — which selects events with distinguishing characteristics that make them interesting for physics analyses.
- The ATLAS trigger system carries out the selection process in two stages. The Level-1 hardware trigger, constructed with custom-made electronics, works on a subset of information from the calorimeter and muon detectors. The decision to keep the data from an event is made less than two-and-half microseconds after the event occurs, and the event is then retrieved from pipelined storage buffers. The Level-1 trigger can save at most 100,000 events each second for the High-Level Trigger (HLT).

[207] http://atlas.cern

- The HLT is a large farm of CPUs (i.e., a software-based trigger) which refines the analysis of the hardware-based Level-1 trigger. It conducts a very detailed analysis either by performing overall examination of the whole event for selected layers of the detector (for example, calorimeters, trackers, muon detectors), or, by utilizing the data in smaller and isolated regions of the detector. About 1,000 events per second are selected by the HLT analysis and are fully assembled into an event record. These events are passed on to a data storage system for offline analysis.
- The ATLAS Computing System analyses the data produced by the ATLAS detector, developing and improving computing software used to store, process, and analyze vast amounts of collision data.
- Data from the ATLAS detector are calibrated, reconstructed, and automatically distributed all around the world by the ATLAS DMS. The ATLAS Production System then filters through these events and selects the ones needed for a particular type of analysis. This brings the data set down to a manageable size for someone doing an analysis on their laptop.
- All members of the ATLAS collaboration have equal access possibilities to all ATLAS data, independently of their geographical location, thanks to the WLCG. ATLAS computing infrastructure and software are constantly evolving with the help of members of the collaboration.
- ATLAS has over 130 computing centers worldwide — located on every inhabited continent — nursed around the clock by members of the collaboration.

5.10.5.3.3 WLCG

The WLCG collects resources worldwide and enables their usage by the LHC experiments as a distributed computing facility. Information from the WLCG web site[208] is quoted below.

- The mission of the WLCG project is to provide global computing resources to store, distribute, and analyze the ~50–70 Petabytes of data expected every year of operations from the LHC at CERN on the Franco-Swiss border.
- The scale and complexity of data from the LHC is unprecedented. These data need to be stored, easily retrieved, and analyzed by physicists all over the world. This requires massive storage facilities, global networking, immense computing power, and, of course, funding.
- CERN does not have the computing or financial resources to crunch all of the data on-site, so in 2002 it turned to grid computing to share the burden with computer centers around the world. The WLCG builds on the ideas of grid technology initially proposed in 1999 by Ian Foster and Carl Kesselman.
- WLCG is coordinated by CERN. It is managed and operated by a worldwide collaboration between the experiments (ALICE, ATLAS, CMS, and LHCb) and the participating computer centers. It is reviewed by a board of delegates from partner country funding agencies, and scientifically reviewed by the LHC Experiments Committee. The WLCG is partnered with EGI (European Grid Infrastructure), OSG, and NeIC (Nordic e-Infrastructure Collaboration)

5.10.5.3.4 US ATLAS T1 at BNL

The US ATLAS Tier 1 is hosted at BNL's SDDC. 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.

The infrastructure of the US ATLAS Tier 1 site at BNL consists of the following three large blocks, currently physically located in the existing SDCC datacenter based on B515 building at BNL site:

[208] https://wlcg.web.cern.ch/

- ATLAS Tier 1 Linux Farm, including HTCondor and Grid CE / Gatekeeper infrastructure (CPU resource).
- ATLAS Tier 1 dCache storage system, including Ceph testbed and GridFTP / SRM / XROOTD DTNs (disk resource).
- Oracle SL8500-based part of the SDCC HPSS complex devoted to ATLAS (tape resource).

A new state of the art data center is being established in the former NSLS I building on the BNL site (B725) in anticipation for the increased computing needs of supported programs, in particular ATLAS for the LHC Run 3. It is expected the building will be delivered to ATLAS by summer 2021. The set of racks containing equipment related to ATLAS T1 is to be migrated from the old datacenter to the new building. This equipment consists of 16 CPU racks that are to be moved as is without changing the compute node layout or network equipment in the racks, but with replacing Cooling Distribution Units to match the power distribution infrastructure in the new building. This migration is to be performed in three subsequent interventions involving a maximum of six racks at a time in the July to August 2020 timeframe. This move will result in only a temporary reduction of a CPU integral available under the US ATLAS Tier-1 site. The first 20k slot tape library and the first rack of ATLAS dedicated HPSS movers are to be deployed for ATLAS in the Tape Room of the new building before the end of FY21 (based on the current projections for the construction schedule for the new datacenter and COVID-19 countermeasures in place). The HPSS Core servers are to be moved to the new datacenter before the end of FY21 as well. The remaining ATLAS CPU and disk equipment located in the old datacenter is expected to be gradually decommissioned in FY21–23 as part of normal lifecycle handling. The infrastructure components of the US ATLAS Tier 1 site are expected to be replaced with hardware-based systems in the new datacenter in the FY22–23 period as well. The Oracle SL8500 tape silos and the associated HPSS movers are to be left in the old datacenter in the operational state for the lifetime of the Oracle SL8500 tape silos. All new equipment to be deployed for the US ATLAS Tier 1 site starting from 2021 Q3 is expected to be placed in the new datacenter.

The network architecture and the high-level service layout of the US ATLAS Tier 1 site are described in Section 5.10.5.7.1.

Figures 50, 51, and **52** show the amount of usable CPU, disk, and tape resources provided by the ATLAS Tier 1 site at BNL (values in FY18–20 range are actual delivered values, and values in the range FY21–27 are from the most up-to-date projection as of Aug 2020).

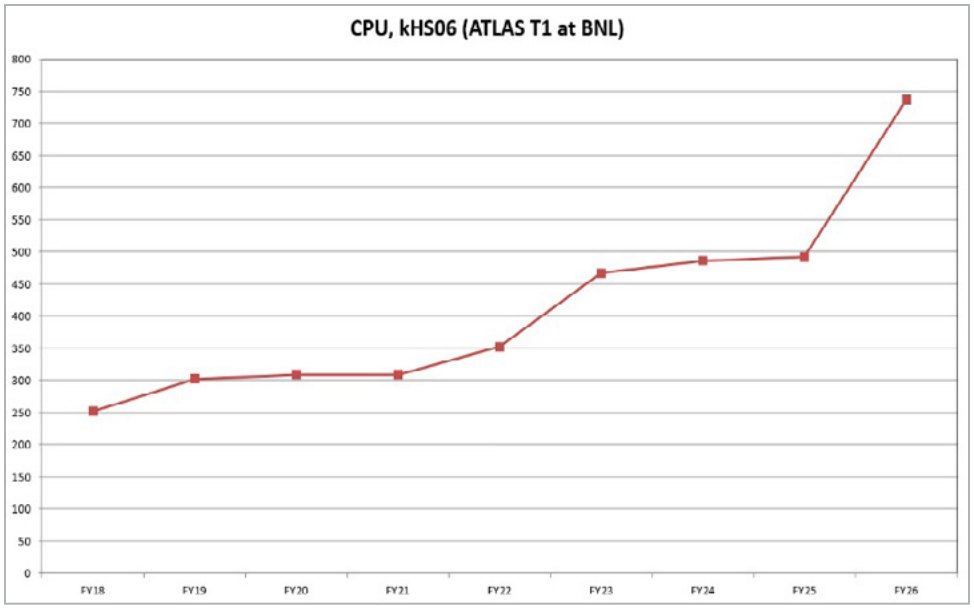


Figure 50: ATLAS Tier 1 at BNL CPU

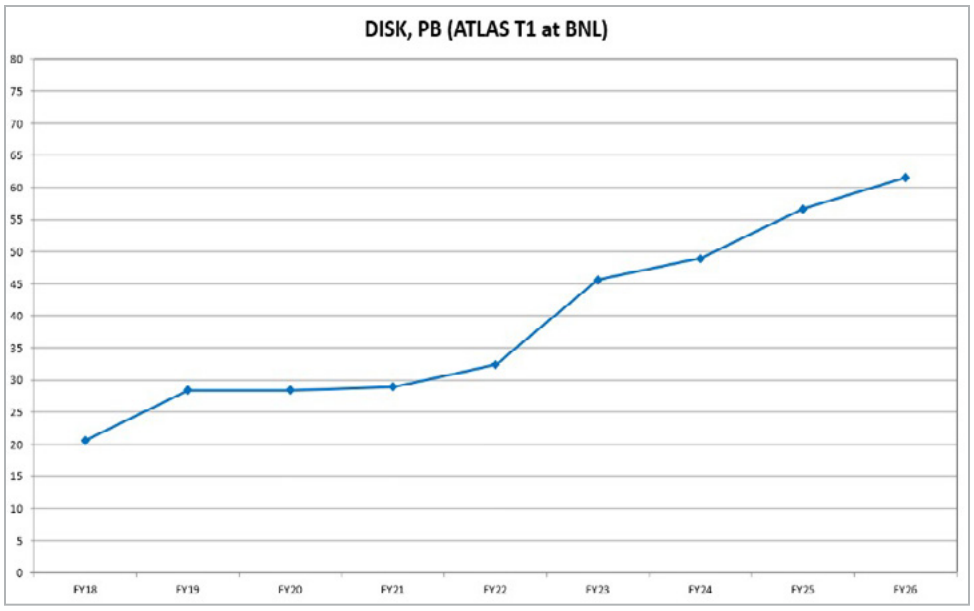


Figure 51: ATLAS Tier 1 at BNL disk

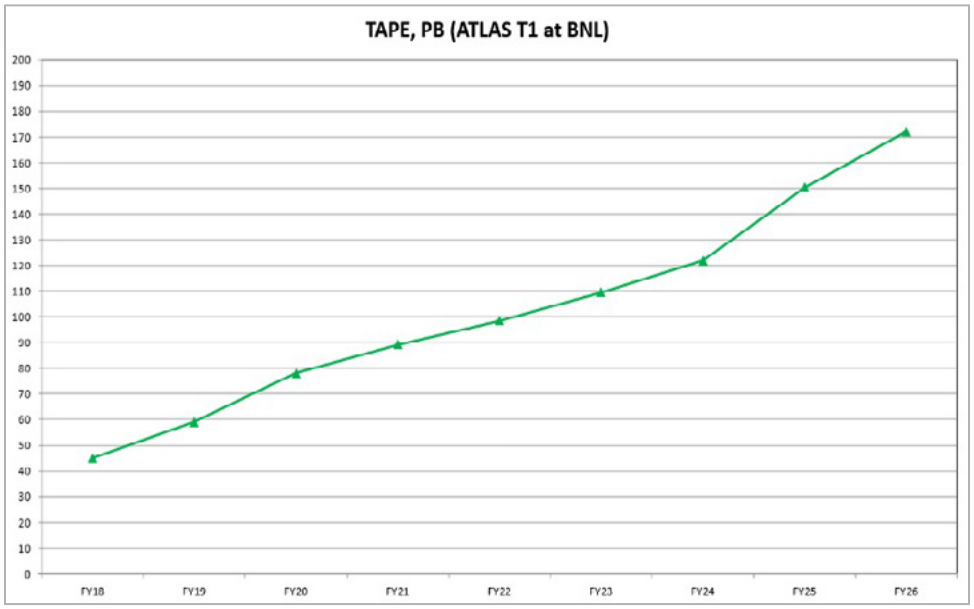


Figure 52: ATLAS Tier 2 at BNL tape

5.10.5.3.5 US ATLAS T2 Infrastructure

There are four ATLAS Tier 2 centers in the United States: 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. The NorthEast Tier 2 center hosts two clusters for Boston University and Harvard University at a common location, the Massachusetts Green HPC Center (MGHPCC), in Holyoke, Massachusetts. The Great Lakes Tier 2 clusters are located at the University of Michigan and Michigan State University. The MidWest Tier 2 clusters are at the University of Chicago, Indiana University, and the University of Illinois at Urbana-Champaign. Finally, the SouthWest Tier 2 center has clusters at the University of Texas at Arlington and Oklahoma University. The MWT2 is funded at a level 50% higher than the other Tier 2 sites in the United States, and therefore provides 50% more resources.

The resources provisioned at each Tier 2 center are based on the WLCG pledges, which are set by the ATLAS/WLCG Resource Review Board and Computing Resource Scrutiny Group annually. These pledges provide the baseline for the minimal CPU and storage resources to be made available by each Tier 2 center in ATLAS. The pledges are set in April of every year, and are tracked monthly[209]. In addition, ATLAS officially requests additional CPU cycles from all sites, in order to accommodate MC simulations in support of physics publications, equivalent to the pledged amount. Finally, an extra 20% of cycles are added to accommodate the physics analysis of US users, dedicating a higher share to analysis queues than the ATLAS average. Combining all of these factors, the US Tier 2 sites aim to provide 220% of the CPU pledge, and 100% of the storage pledge.

We show CPU data in **Table 20** from the EGI accounting portal for the period of 1/2019 to 12/2019. Values shown are Normalized CPU time in units of HEPSPEC06:

Tier 2	2019 pledged power	Pledged wall-clock work (days)	Delivered power	Total work delivered by Tier 2 (days)	Delivered % of pledge
US-AGLT2	64,250	23,460,000	113,761	41,522,649	177%
US-MWT2	96,500	35,235,000	215,930	78,814,426	224%
US-NET2	64,250	23,460,000	83,466	30,465,226	130%
US-SWT2	64,250	23,460,000	128,920	47,055,895	201%

Table 20: Tier 2 CPU metrics for Jan to Dec 2019 in units of HEPSPEC06

For 2019, US Tier 2 sites delivered between 130–225%, thereby falling somewhat short of the average goal of 220% of pledge for CPU cycles as described earlier. However, all sites are above the minimum pledge of 100%.

Table 21 shows the available storage metric for each Tier 2 site as obtained from WLCG reporting for Dec 2019 compared with the April 2019 pledges:

Tier 2	Total online storage (GB)	Disk pledge (GB)	Disk ratio %
GREAT LAKES ATLAS T2	6,130,000	5,500,000	111%
MIDWEST ATLAS T2	8,100,000	8,300,000	98%
NORTHEAST ATLAS T2	4,200,000	5,500,000	76%
SOUTHWEST ATLAS T2	5,480,000	5,500,000	100%

Table 21: Total storage at US Tier 2 sites in Dec 2019

All Tier 2 sites are required to provide a minimum of 10 Gbps connectivity. However, all US Tier 2 sites currently provide 20–100 Gbps. The US goal is to achieve 40 Gbps links at all Tier 2 sites by 2022, at the start of Run 3.

Figure 53 shows the number of slots used at all ATLAS Tier 2 sites during the three-year period from January 2017 to December 2019. Over this period, MWT2 and SWT2 provided the highest number of slots to ATLAS among all Tier 2 sites worldwide, with AGLT2 at number 4, and NET2 at number 8. The average number of slots was 17k at MWT2, and about 8–10k at the other sites.

[209] https://wlcg-rebus.cern.ch/apps/capacities/pledge_comparison

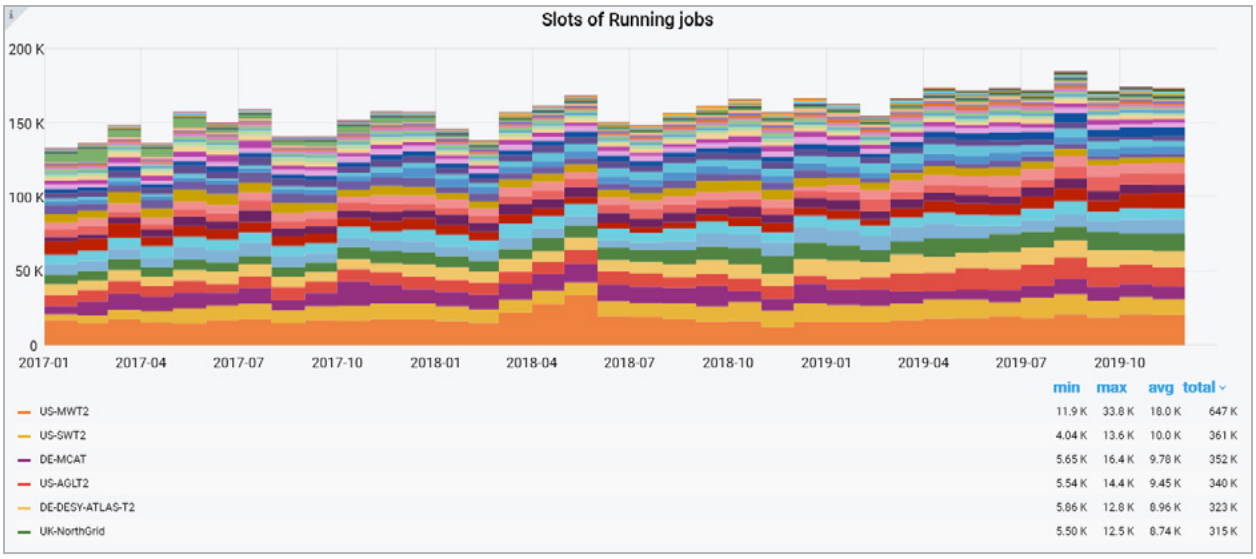


Figure 53: ATLAS job slots

5.10.5.3.6 US ATLAS T3 AFs

Final stages of analysis (statistical inference, weighting, signal/background calculations, plots, etc.) are not usually done on the Tier 1 and Tier 2 sites using grid tools. US physicists typically copy the selected smaller data products from the grid to local facilities for final analysis leading to publications. These local analysis sites are called Tier 3 (T3) AFs.

Many universities and laboratories maintain their own T3 AFs on locally procured resources. Given the complexity of ATLAS software systems, this requires a lot of dedicated effort. Funding is not always available to duplicate these facilities at every US ATLAS institution. US ATLAS maintains three common and shared T3 AFs, available for all US ATLAS physicists. These sites require good network bandwidth to WLCG sites, including US Tier 1 and Tier 2 sites.

The two T3 AFs currently operational are located at BNL and SLAC. A third T3 AF is being set up at the University of Chicago, to be operational by the end of 2021. Each T3 AF consists of a few PB of storage and about 1,000 cores provisioned as a local batch system.

5.10.5.4 Process of Science

5.10.5.4.1 Computation

The main activities in terms of CPU usage are:

- **MC event generation.** Different event generators are used to repeatedly generate proton-proton or heavy-ion collision events and particularly interesting particle decays. There are usually no or very small input file sizes, in the order of GBs, required, and the output file sizes are in the order of a few MBs per job.
- **MC data simulation.** The outputs of the event generation are used to simulate the particle decays in a detailed or parameterized ATLAS detector simulation. These are the most CPU-demanding tasks with event processing times in the range of 1 to 30 minutes. The output file sizes are in the order of several 100 MBs per job.
- **MC simulation reconstruction.** Individual physics objects like electrons, muons, photons, etc. of the previously simulated particle decays in the collision events are reconstructed for a later analysis and stored in AOD files. The processing times per event are in the range of 10–60 seconds and the output file sizes are in the range of 1–10 GB per job.

- **Detector data reconstruction.** Physics objects are reconstructed from raw detector data and stored in AOD files. The processing times per event are in the range of 10–60 seconds. The input file sizes are in the range of 2 to 5 GB and the output file sizes are in the range of 1 to 10 GB per job.
- **Production of derived analysis formats** (derivation production) as inputs for physics and performance analysis. The physics object files are augmented with additional calibrated information. Events are filtered based on interesting event patterns and stored in DAOD files. About 80 different individual DAOD formats are written. The input and output file sizes are in the range of 100 MB to 10 GB per job.
- **Individual and group analysis.** Individual physicists are processing the filtered physics objects files to filter out highly specialized information. The input and output files are in the range of 100 MB to 10 GB per job.

All of the above activities process data sets with sizes between 10 GB and 50 TB consisting of 10–10k event data files and file sizes up to 15 GBs each. The processing of these data sets is broken down into individual jobs by the workflow management system PanDA and each job processes only a fraction of the data sets.

In Run 3, starting in 2022, the following changes are planned in the simulation and data processing:

- The number of derivation production formats will be significantly reduced and most of the analysis will use the common DAOD_PHYS format
- The number of AOD file replicas stored on disk will be reduced and will be read back in on demand from tape in derivation production.

Figure 54 shows the number of running job slots (cores accessible through PanDA) at all US grid sites in the year 2019, as a function of the activity type.

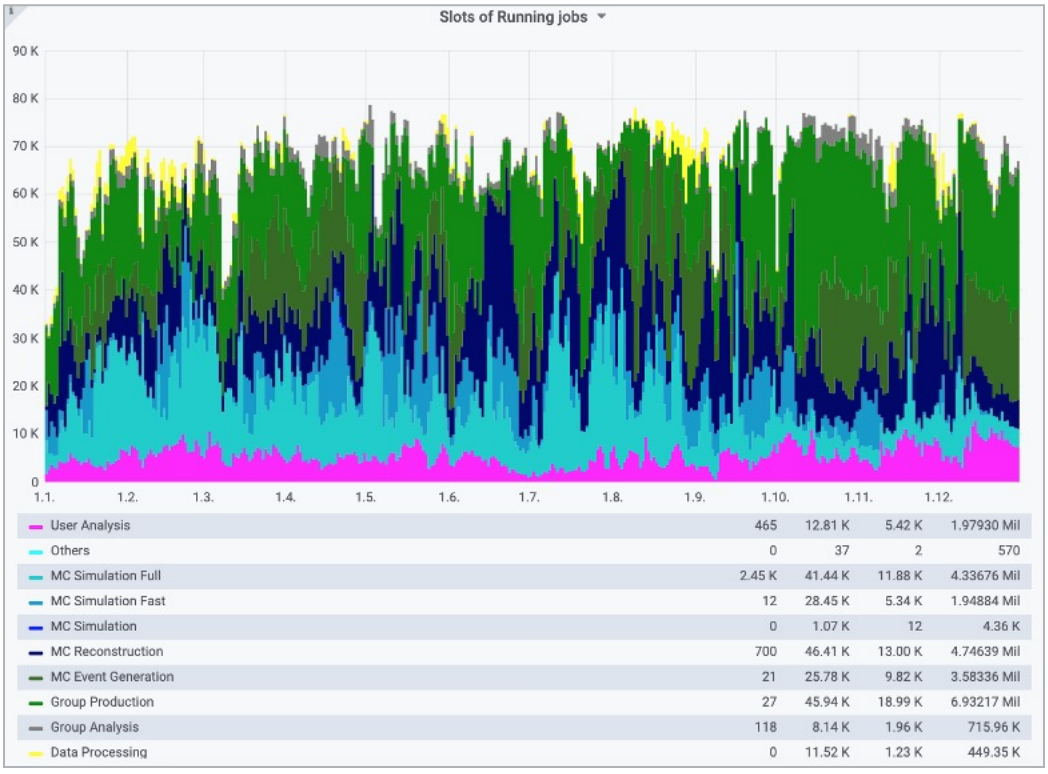


Figure 54: US ATLAS job slots (2019)

Figure 55 shows the number of running job slots in the year 2019 at the BNL Tier 1 as a function of activity type.

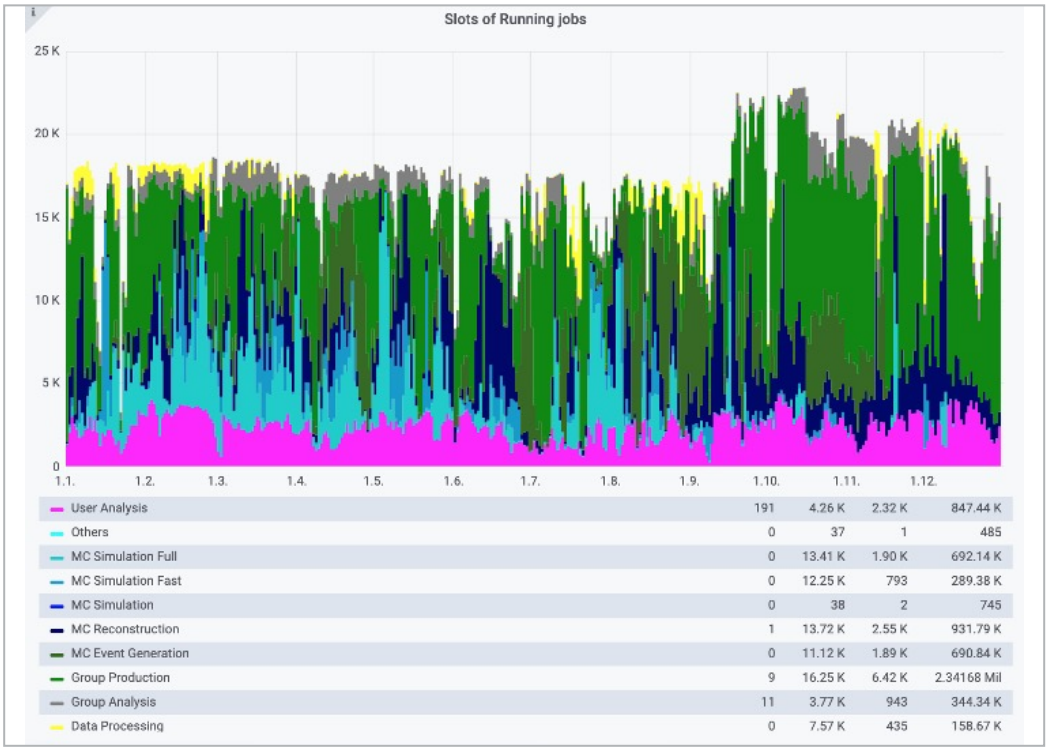


Figure 55: US ATLAS Tier 1 job slots (2019)

Figure 56 shows the number of running job slots in the year 2019 at the US Tier 2 sites AGLT2, MWT2, NET2, and SWT2 as a function of activity type.

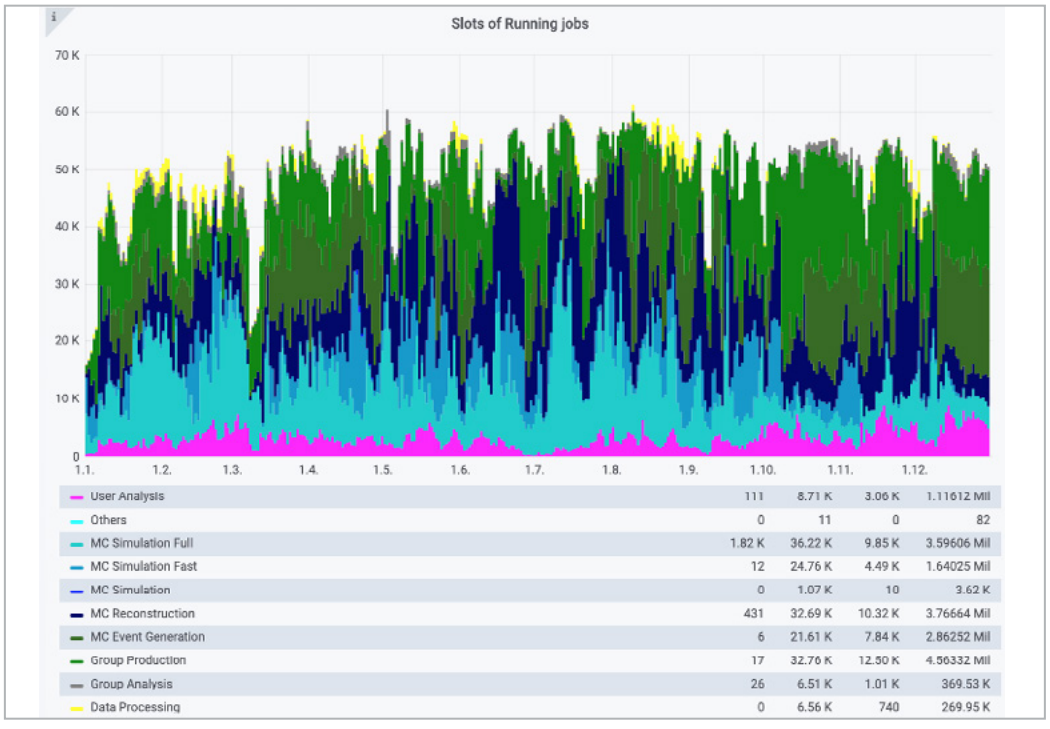


Figure 56: US ATLAS Tier 2 (aggregate) job slots (2019)

Figure 57 shows in addition the number of running job slots in 2019 with the HPC Cori at NERSC added to the T2 sites. Note that the HPC sites are mainly used for MC simulations.

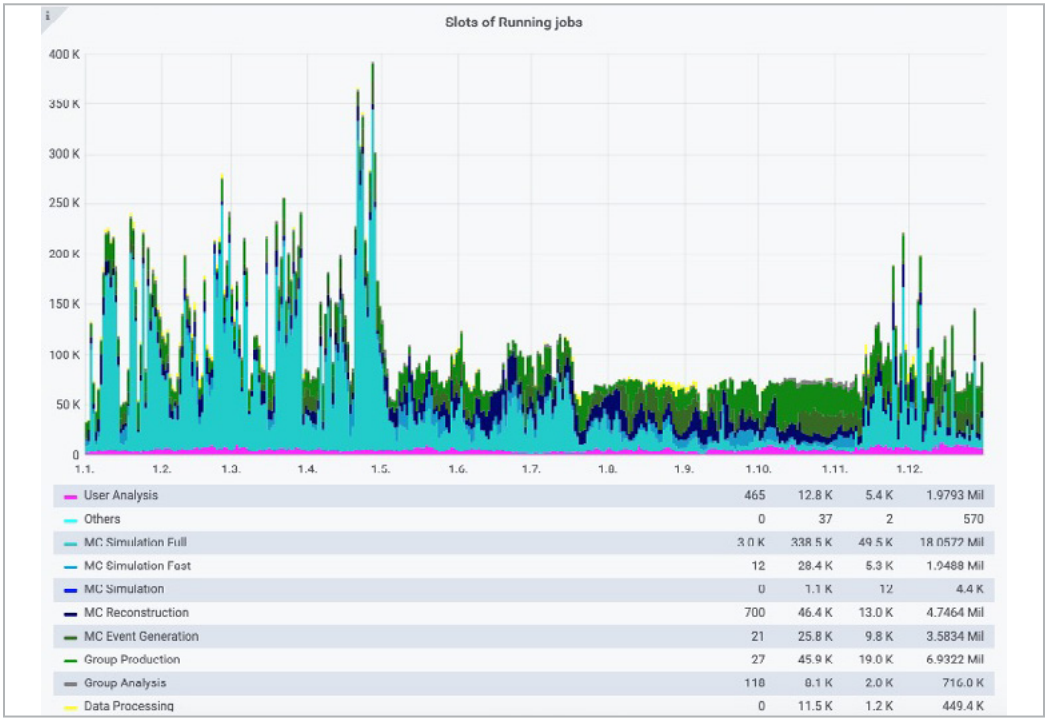


Figure 57: US ATLAS Tier 2 (aggregate) and NERSC CORI job slots (2019)

5.10.5.4.1.1 HPC

ATLAS has a long history of successfully using HPC resources during Run 2 at the LHC. From 2016–2020, US-based HPC resources supported 10–25% of ATLAS simulation production. European HPC resources were also used, though mostly through grid interfaces. US resources at ANL, ORNL, NERSC, and XSEDE required special edge services through Harvester to run ATLAS jobs. Only simulations were carried out since they are not data intensive. In the future, ATLAS plans to run all types of workloads at High Performance Computing Centers. This will put much higher demands on networking.

Figure 58 shows the data transfer rate from NERSC and SLAC combined in the past 12 months. The data that are processed at NERSC in ATLAS production jobs is staged in and out directly from SLAC. The average and peak rates for production are hardly larger than 50 MB/s.

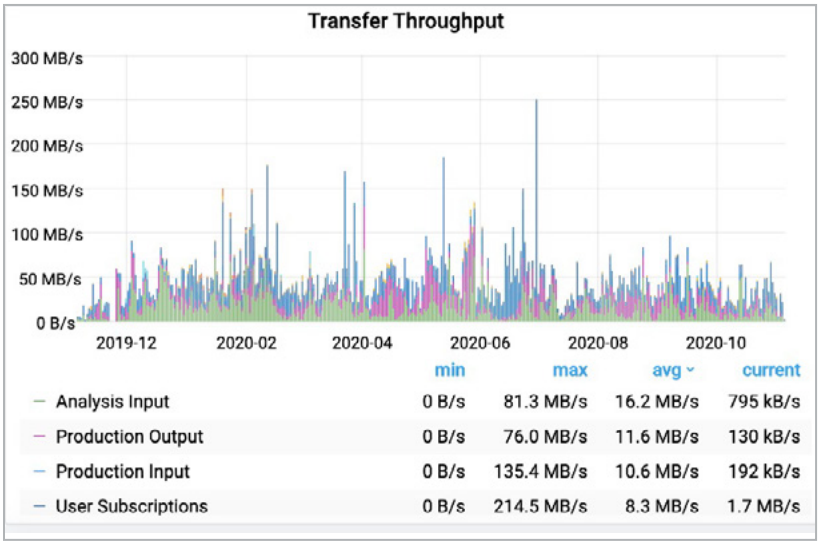


Figure 58: ATLAS from NERSC and SLAC transfer rate

Similarly **Figure 59** shows the data transfer rate to NERSC and SLAC combined in the past 12 months. The rates for production input are at a similar scale as for the transfer rate to the site.

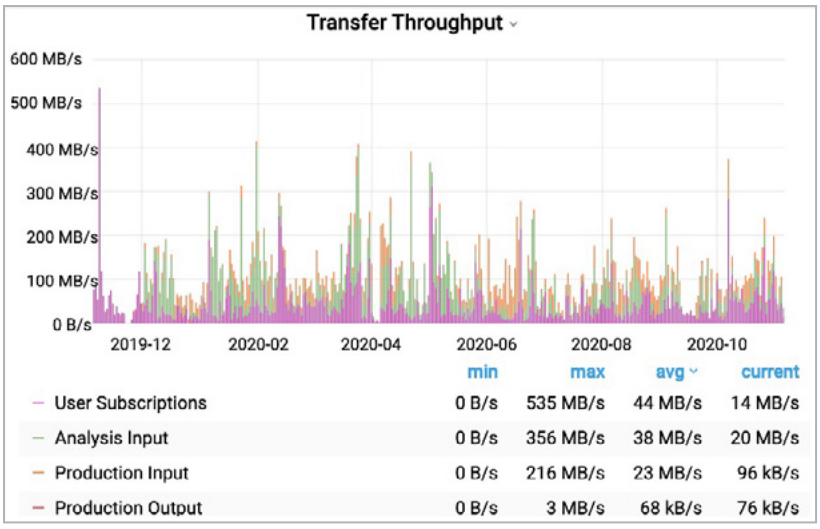


Figure 59: ATLAS to NERSC and SLAC transfer rate

5.10.5.4.1.2 Distributed Computing

ATLAS computing is fully distributed as soon as the raw data are transferred out of CERN. The first step in raw data distribution to the 10 Tier 1 sites is planned according to MOU pledges. This allows for a complete second copy of the raw data. After this archival step, all processing and reprocessing of the data is managed by PanDA, based on the current availability of resources. Rucio is used for data cataloging and data transfer. All computing activities are free to occur at any site, irrespective of their tier and 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.

5.10.5.4.2 Storage

There are four distinct storage types used by ATLAS computing environments: disk, tape, cloud, and cache storage.

- Disk storage is the main bulk of the storage used by ATLAS. Currently, 850M files with the volume of 230 PB are stored in the disk storage. In the US, 160M files with a total volume of 53 PB are stored in the disk storage.
- Tape storage is used as a main data archival system. Currently, 225M files with a total volume of 280 PB are stored in the tape archive system. In the US, 40M files with a total volume of 45 PB are stored in the BNL Tier 1 HPSS system.
- S3 type storage is used for temporary output of event service jobs. The files in S3 are small and will be merged to produce large files before they are written to the disk storage described previously.
- XROOTD XCache storage is used for unmanaged disk data cache for input of user analysis jobs. These small cached data are created on demand by user jobs. And, their data are copied from the disk storage described previously and stay in the cache in the duration of the user analysis jobs.

The main disk storage is split into a few different areas. The largest fraction (~90%) of the disk storage is assigned to DATADISK space where the production system uses it for inputs and outputs of the jobs. GROUPDISK area is used for Physics group production while SCRATCHDISK area is used mainly by inputs and outputs of user analysis jobs.

Files are categorized as primary or secondary depending on if they are the main copy of the data or duplicate copy of the data. Roughly, 75% of the data in the disk storage are primary while 25% of the data are secondary. In addition, files are classified by their data types such as raw, ESD and AOD and DAOD, etc. Raw data are unprocessed data from the detectors. They are stored mainly in tape storage for archiving purposes. But they can be brought to disk storage for production of AOD. AOD are further processed, and resulted in DAOD, Derived AOD. AOD and DAOD data types occupy the largest fraction of the disk storage with ~30% and ~20% respectively.

Storage usage at US T1 and T2 sites is shown in the following figure.

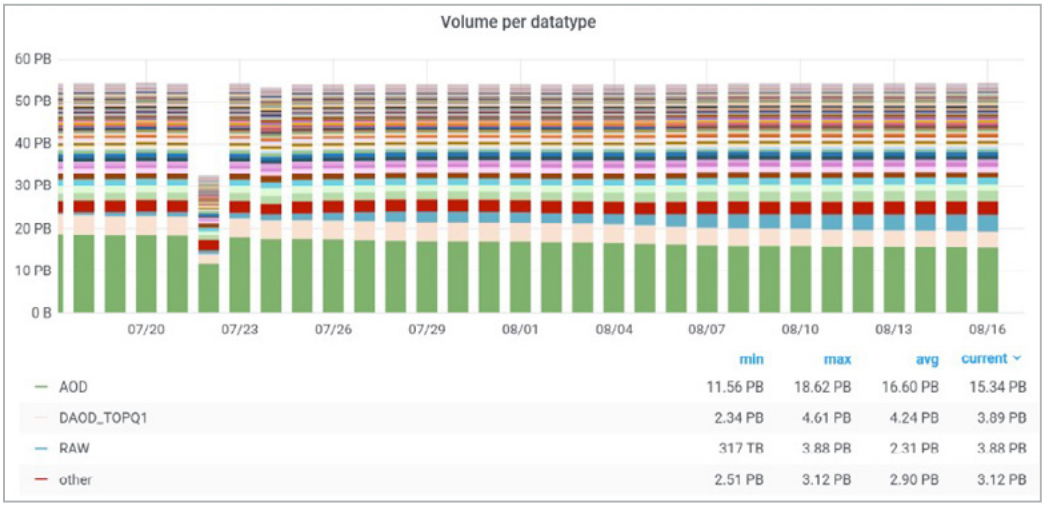


Figure 60: Disk storage usage in US T1/BNL by type

Disk storage usage in US T1/BNL categorized by primary and secondary data is shown in **Figure 61.**

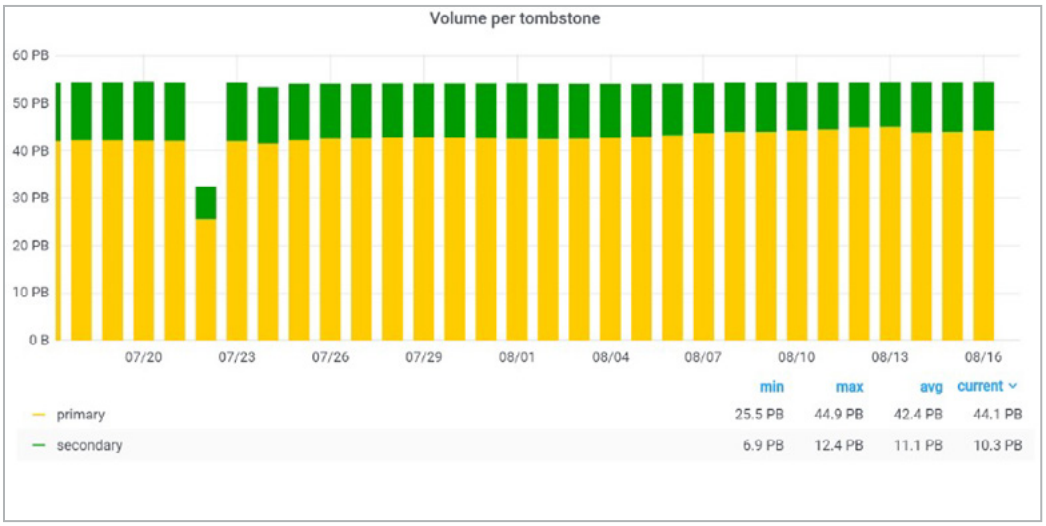


Figure 61: Disk storage usage in US T1/BNL by primary/secondary

Tape storage usage in US T1/BNL is categorized by data type in **Figure 62.**

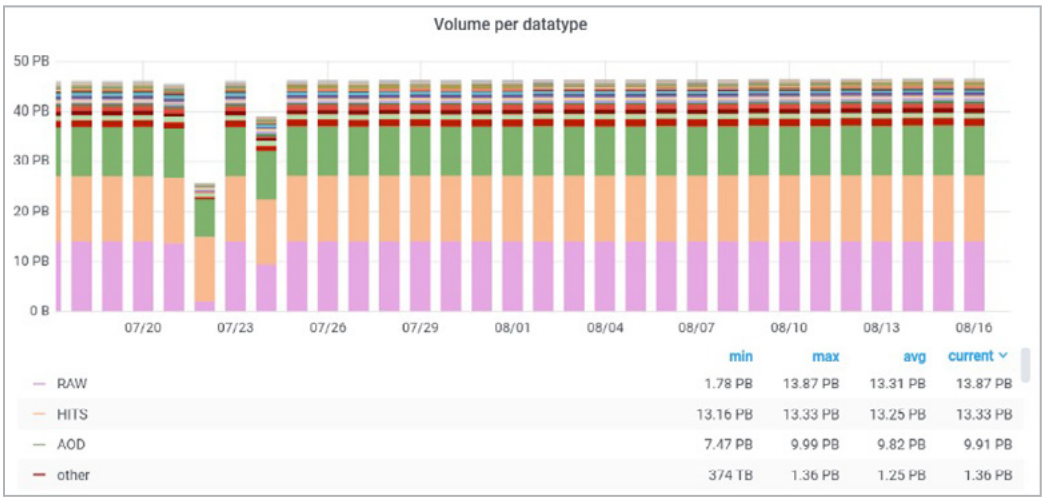


Figure 62: Tape storage usage in US T1/BNL by type

Storage usage in US T2s is categorized by data type in **Figure 63.**

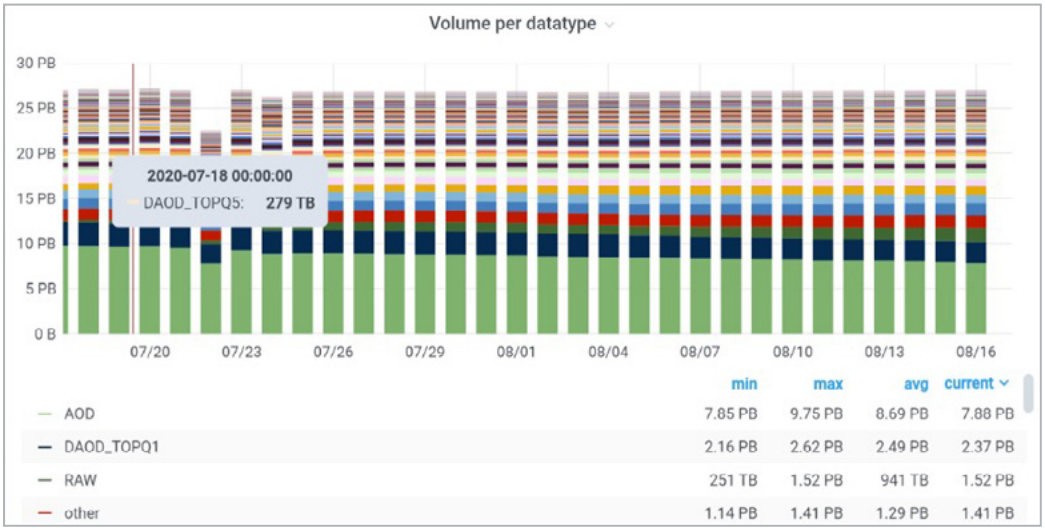


Figure 63: Storage usage in US T2s by type

Storage usage in US T2s is categorized by primary and secondary data type in **Figure 64.**

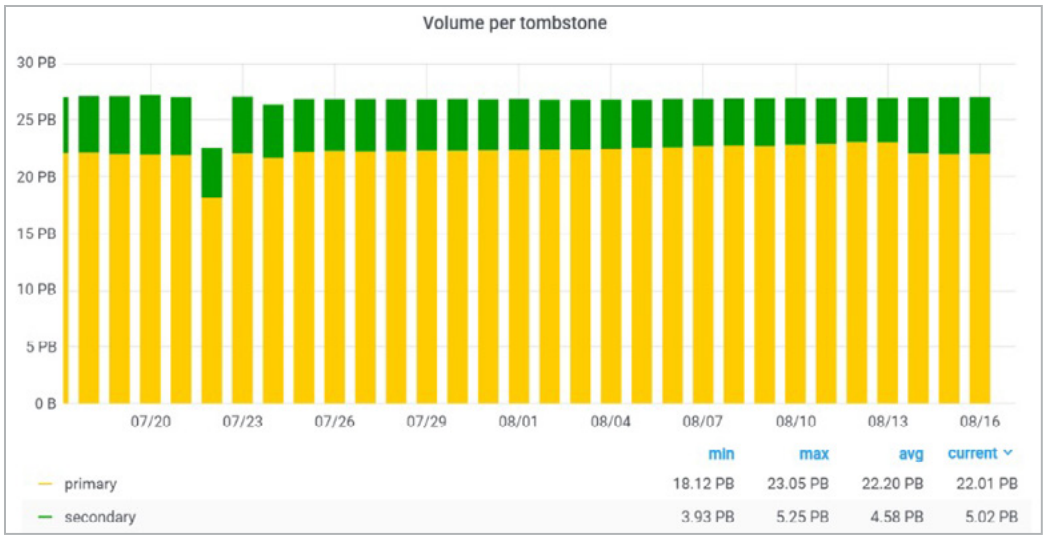


Figure 64: Storage usage in US T2s by primary/secondary

It is expected that the AOD data type which occupies about 30% of the disk storage will be transferred to the tape storage, providing 30% saving in disk space. AOD data will be transferred to the disk area on demand. This process has been heavily tested by the data carousel project, which has closely examined the throughput capabilities of the tape system at the T1 tape sites.

5.10.5.4.3 Network Use Cases and Data Flow

The network use cases and data flows are described in Sections 5.10.5.3 and 5.10.5.4. ATLAS provides some additional plots to quantify data transfer rates and volumes. **Figure 65** shows the transfer volume per day worldwide in 2020 so far:

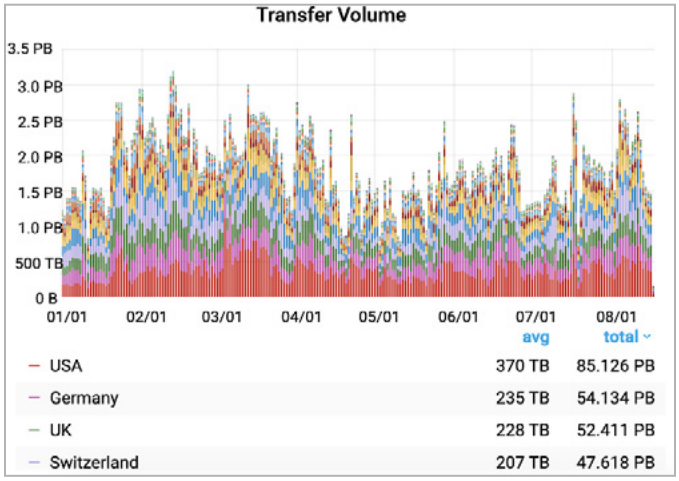


Figure 65: ATLAS transfer volume worldwide (2020)

Figure 66 shows the transfer rate per day worldwide in 2020 so far.

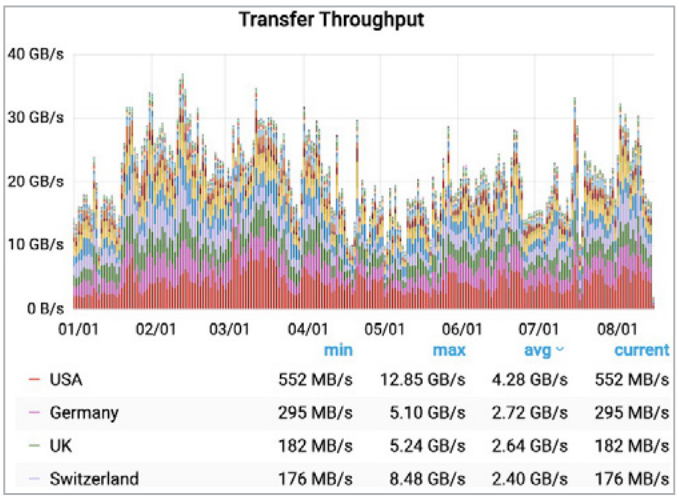


Figure 66: ATLAS transfer rate worldwide (2020)

Figure 67 shows the transfer volume in 2020 so far by US grid site.

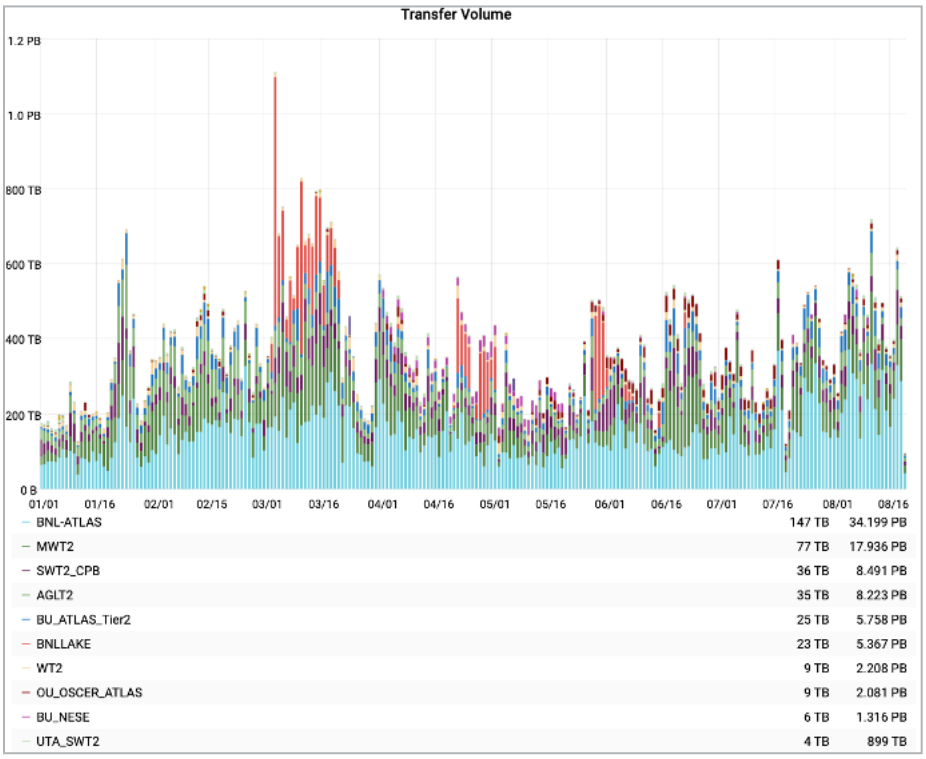


Figure 67: US ATLAS transfer volume by site (2020)

Figure 68 shows the transfer rate in 2020 so far by US grid site.

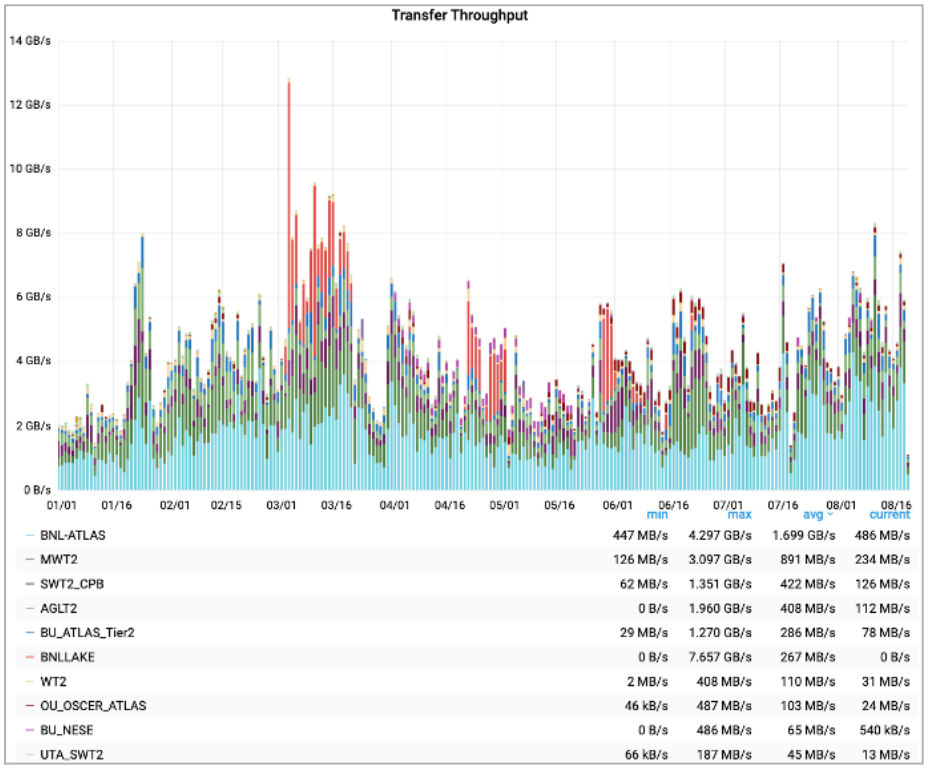


Figure 68: US ATLAS transfer rates by site (2020)

Figure 69 shows the transfer volume in 2020 in the United States by activity.

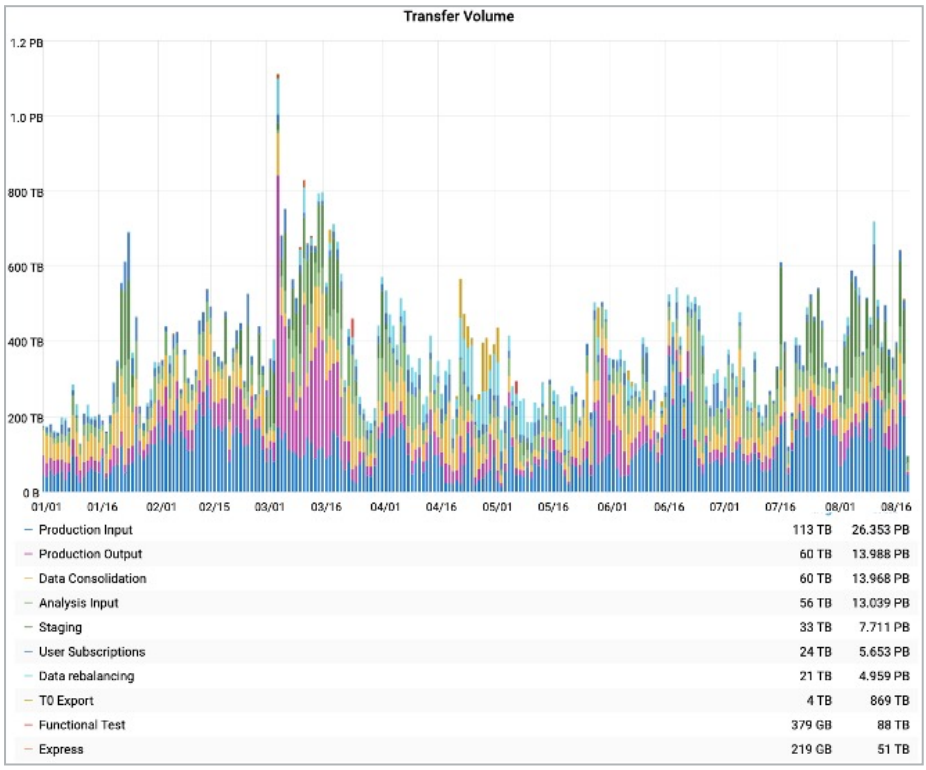


Figure 69: US ATLAS transfer volume by activity (2020)

Figure 70 shows the transfer rate in 2020 in the United States by activity.

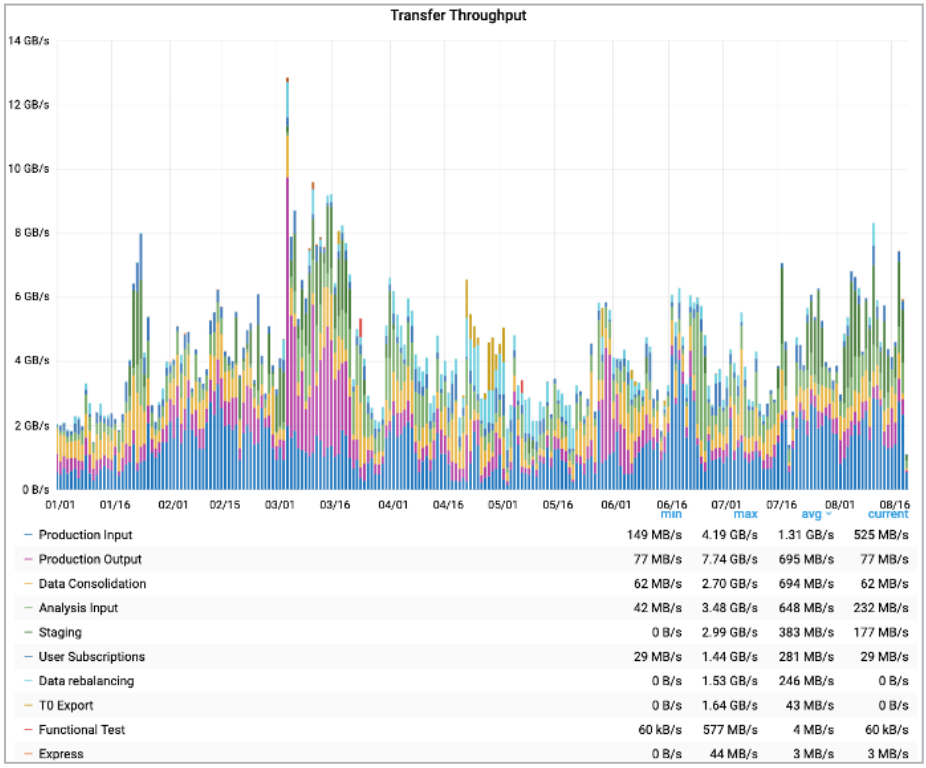


Figure 70: US ATLAS transfer rates by activity (2020)

Figure 71 shows the transfer volume in 2018 in the United States by activity. Note here the significantly larger T0 export activity compared with 2020.

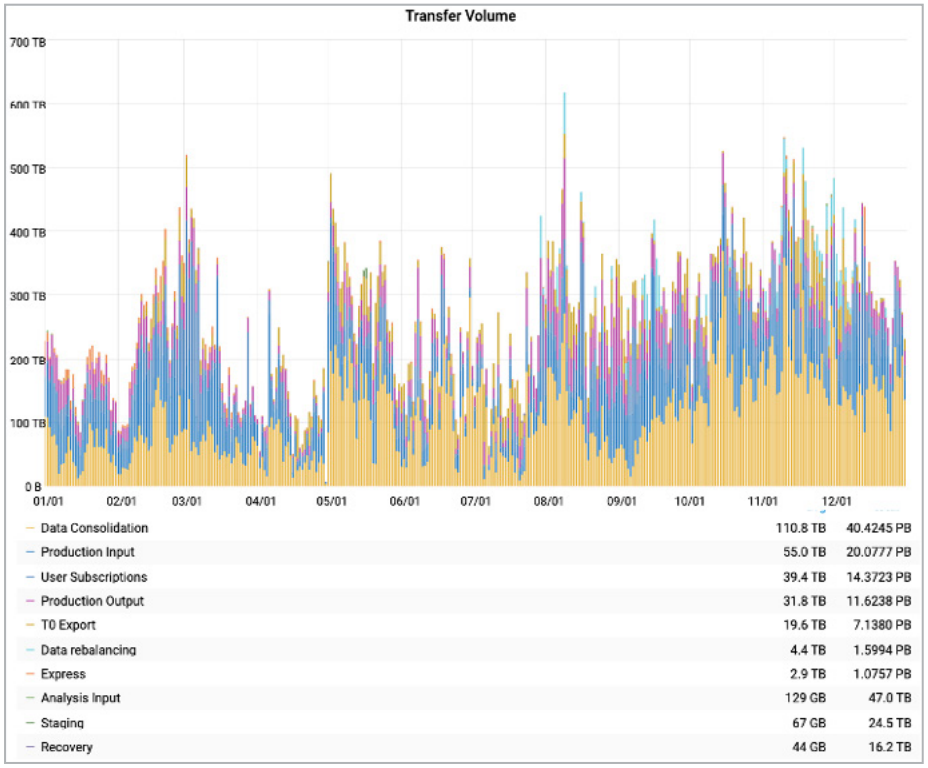


Figure 71: US ATLAS transfer volume (2018)

Figure 72 shows the transfer rate in 2018 in the United States by activity.

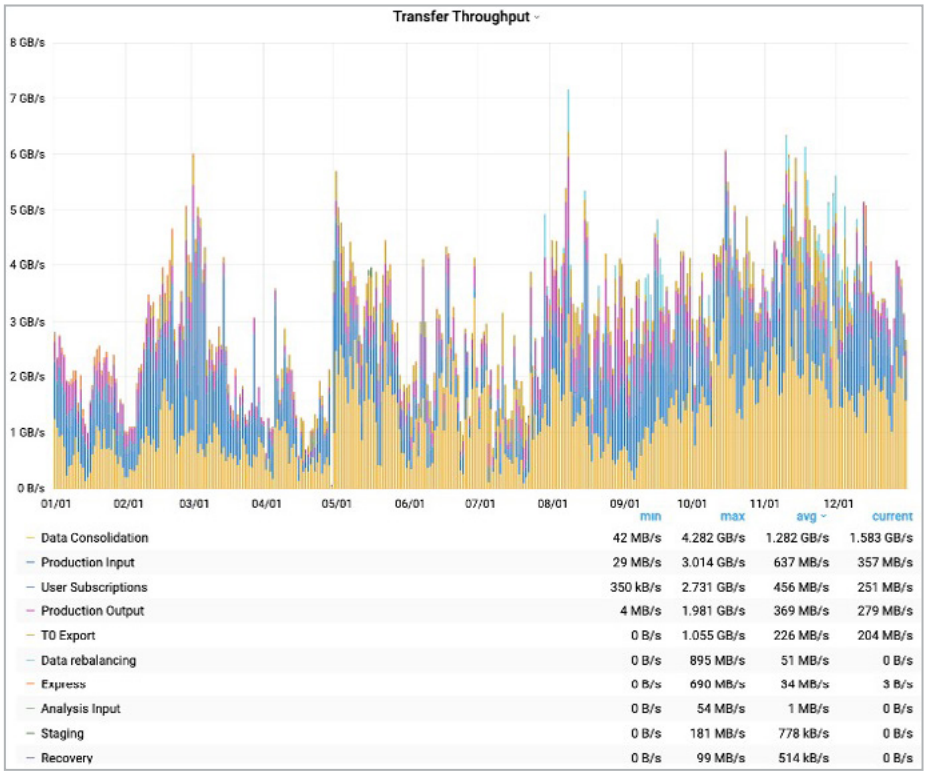


Figure 72: US ATLAS transfer rates by activity (2018)

Figure 73 shows data transfer (PB per week) over the last five years with US ATLAS grid sites as destination for each site category (Tier 1, Tier 2s, Tier 3).

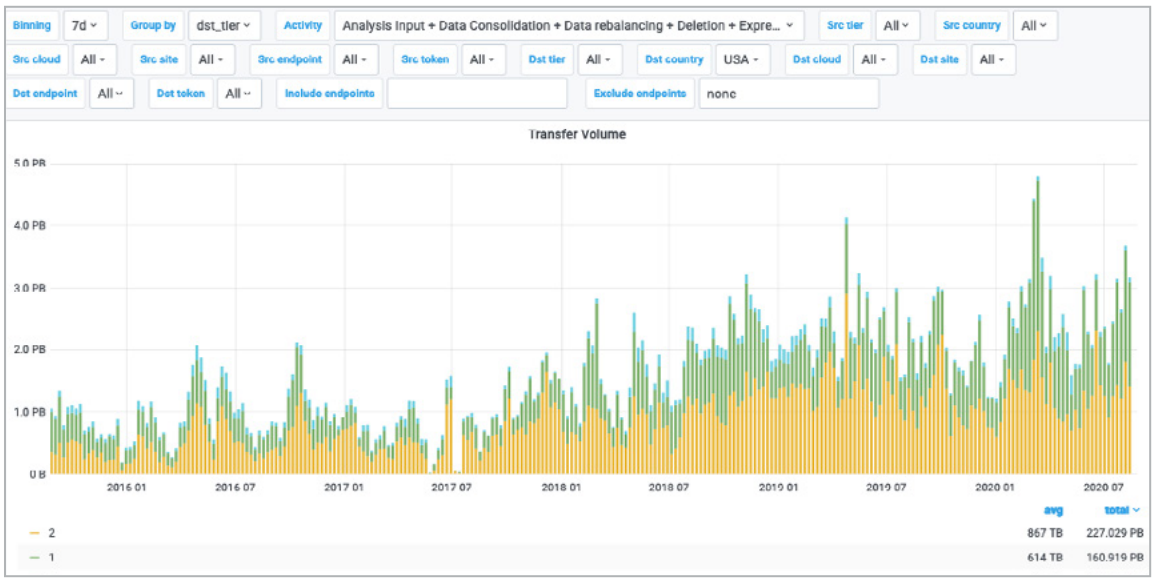


Figure 73: US ATLAS data transfer (destination)

Figure 74 shows data transfer (PB per week) over the last five years with US ATLAS grid sites as origin for each site category (Tier 1, Tier 2s, Tier 3).

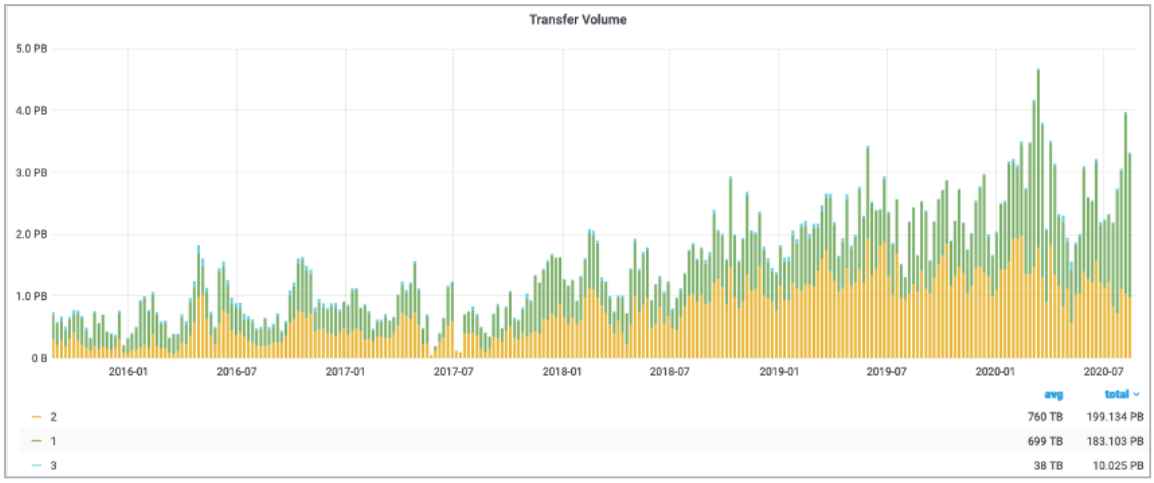


Figure 74: US ATLAS data transfer (origin)

5.10.5.5 Remote Science Activities

As described in previous sections, ATLAS computing is fully distributed, and therefore remote from CERN. The raw data are collected exclusively at CERN. All other activities leading to the final published products are done remotely.

5.10.5.6 Software Infrastructure

5.10.5.6.1 Rucio

The open-source software framework Rucio is used to organize, manage, and access the ATLAS data. The data are distributed across heterogeneous data centers at widely distributed locations worldwide. Rucio is built in different components:

- A central database hosted at CERN contains the "data set catalog". This catalog groups individual files of a detector run or MC simulation process into data sets together with their file location information. These data sets are stored on disk or tape storages at more than 100 sites worldwide.
- Data set files are not accessed remotely, but are moved between storage and sites asynchronously before a processing job request. Currently ATLAS is moving approximately 1–2 PB per day aggregated between sites worldwide.
- The middleware to establish direct storage to storage transfers over the network, commonly called third-party copy, is provided by the FTS. FTS establishes connections between storage systems using the required protocols and ensures that the files are correctly transferred over the networks. Rucio decides which files to move, groups them in transfer requests, submits the transfer requests to FTS, monitors the progress of the transfers, retries in case of errors, and notifies the clients upon completion. If there are multiple FTS servers available, Rucio is able to orchestrate transfers among them for improved parallelism and reliability.
- The input files for individual jobs are accessed through the LAN from a worker node to the local disk storage.
- Remote access to analysis job inputs with network latency is currently studied at several ATLAS grid sites in the context of the WLCG DOMA access project.
- The outputs of individual user analysis jobs in the order of 1 GB to 10 TB is downloaded to university clusters.

5.10.5.6.2 PanDA

The 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 and with built in AI to optimize over a large number of distributed sites deployed worldwide. The main components of PanDA include ProdSys which translates the physics workflows to executable tasks. Deft provides an intelligent user interface to create workflows. Job Execution and Definition Interface (JEDI) transforms tasks into jobs that can run on single cores, single nodes or a single HPC. The execution steps managed by JEDI range from simulations, data processing, to distributed ML. Harvester provides an integrated interface to computing resources, with edge services that can transparently manage execution of workflows across all types of resources: grids, clouds and HPCs. Finally, the most important element of job execution is the PanDA pilot system, which manages job executions that are orchestrated by PanDA. All of these components are deeply integrated with the ATLAS DMS, Rucio.

The highly customizable PanDA system not only manages clusters and storage, but also optimizes workflow over existing networks. Over the years, many capabilities to optimize network performance have been built into PanDA and Rucio. However, direct integration with network layers have remained elusive over the past decade.

5.10.5.6.3 Frontier

The ATLAS model for remote access to database resident information relies upon a limited set of dedicated and distributed Oracle database repositories complemented with the deployment of Frontier system infrastructure on the WLCG. ATLAS clients with network access can get the database information they need dynamically by submitting requests to a Squid server in the Frontier network which provides results from its cache or passes new requests along the network to launchpads co-located at one of the Oracle sites (the master Oracle database at CERN or one of the Tier 1 Oracle database replicas). Since the beginning of LHC Run 1, the system has evolved in terms of client, Squid, and launchpad optimizations, but the distribution model has remained fundamentally unchanged. At the end of Run 3 the direct database access through Frontier will be served from CERN only, and not through replicas.

5.10.5.6.4 CVMFS

The ATLAS software for simulation and reconstruction is distributed via the CernVM files system (CVMFS). CVMFS is implemented as a Portable Operating System Interface (POSIX) read-only file system in user space (a Filesystem in Userspace [FUSE] module). Files and directories are hosted on standard web servers and mounted in the universal namespace /cvmfs. Internally, CVMFS uses content-addressable storage and Merkle trees in order to maintain file data and metadata. CVMFS uses outgoing HTTP connections only. It transfers data and metadata on demand and verifies data integrity by cryptographic hashes.

In addition, ATLAS is currently exploring the usage of docker containers for software distribution. These containers are distributed via the CVMFS files system.

5.10.5.6.5 HTCondor

PanDA uses the HTCondor system for job submission to a majority of computing resources. This allows the Harvester to use a reliable and well-defined HEP standard. While Cream, ARC, native cloud interfaces, or direct interface to batch systems are sometimes used, HTCondor is used for the vast majority of job submissions.

5.10.5.6.5 ROOT

ROOT is a software framework with building blocks for:

- Data processing
- Data analysis
- Data visualization
- Data storage

ROOT is written mainly in C++ (C++11/17 standard) and has bindings for Python available as well. It is highly adopted in HEP and other sciences but also in industry. About 1 EB of data is stored in ROOT format.

5.10.5.7 Network and Data Architecture

5.10.5.7.1 US ATLAS T1 at BNL

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

The HTSN has five key components:

1. Network Perimeter
   - Two (soon to be three) diverse 100 Gbps circuits that peer with ESnet in New York City. These circuits are utilized by all scientific and administrative communities at BNL. All traffic to and from BNL flows through either of these circuits.
   - The BNL network perimeter transfers on average 7–8 PB of data monthly, with spikes up to ~12 PB.
2. Science DMZ
   - Supports open, high-speed WAN/internet access for all scientific collaborations throughout the BNL campus.
3. Science Core
   - A Tbps Science and Data Center Interconnect for data-intensive collaborations at BNL. This Science Interconnect enables high-speed connectivity between collaborations such as ATLAS, STAR, PHENIX, CAD, CFN, NSLS-II, HPC Clusters, and the SDCC.
   - Intelligence and routing policies are applied within the Science Core to restrict or grant access to specific resources within the SDCC.
4. Spine
   - A Tbps network Spine that interconnects all Leaf switches. Leaf switches can consist of ToR or chassis-based switches that connect compute, storage, or general infrastructure service servers.
   - The responsibility of the Spine is fast packet forwarding and flexibility, not policy insertion or server termination.
   - eBGP is utilized throughout the HTSN. EBGP was chosen for its ability to immensely scale and to create modularity and fault domain isolation down to the rack level. Each Spine group shares the same ASN but does not have Internal BGP (iBGP) peering between them. Each Leaf or pair of Leaves will require its own ASN.
5. Storage Core
   - A redundant terabit per second switching block that aggregates high-performance storage services.

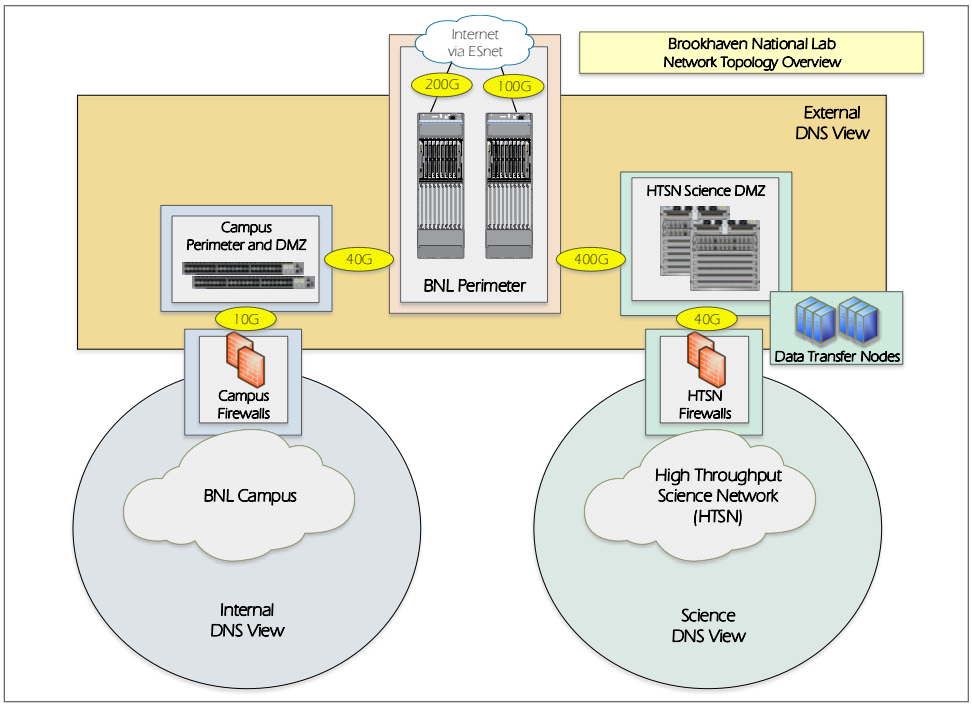


Figure 75: High-level overview of the BNL network perimeter and DNS architecture

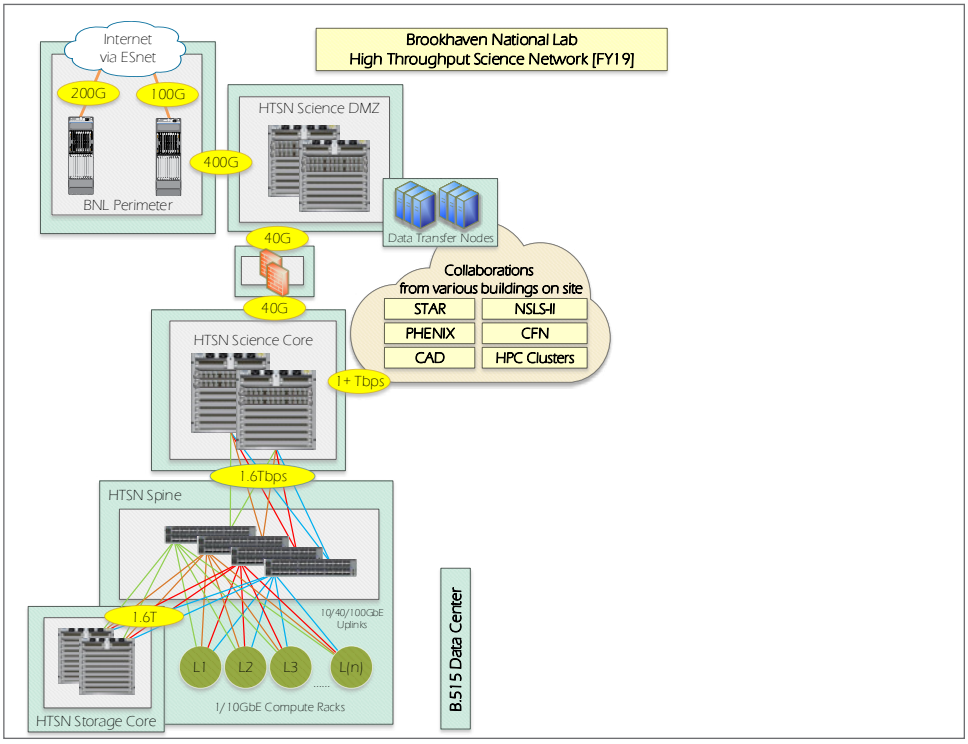


Figure 76: High-level overview of the BNL HTSN (FY19)

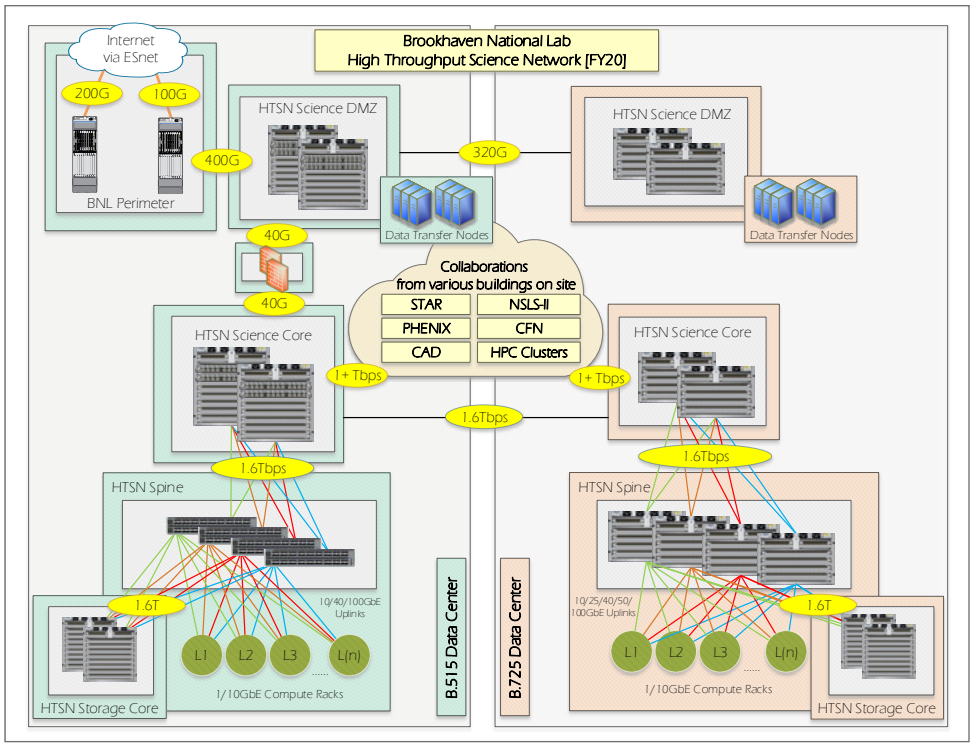


Figure 77: High-level overview of the BNL high-throughput science network in FY20 (includes HTSN expansion into new Data Center, right hand side)

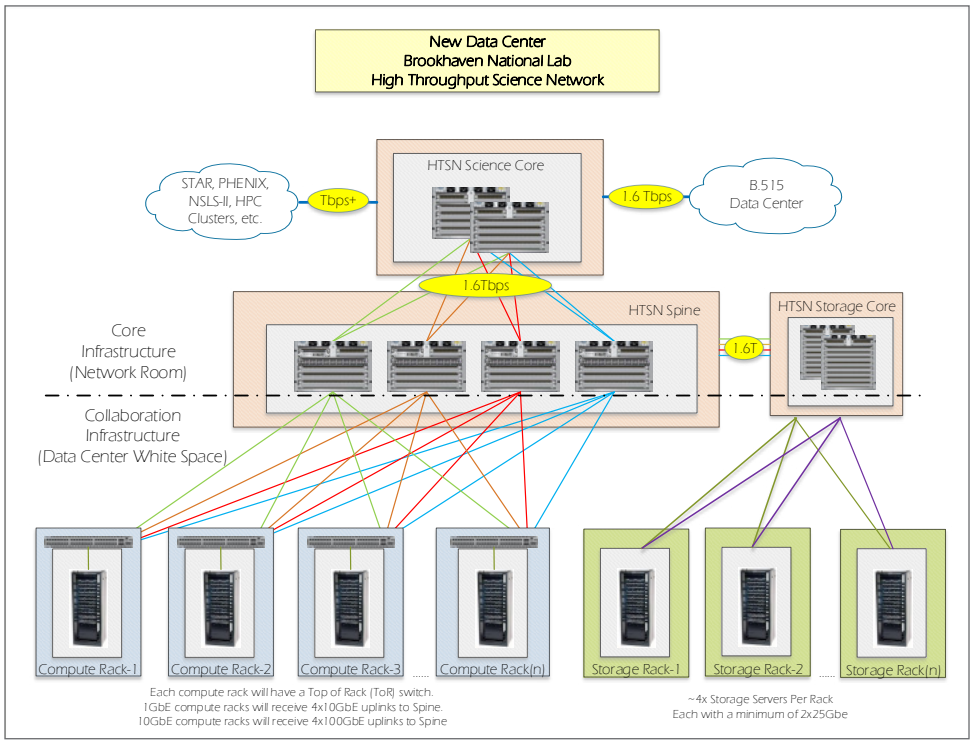


Figure 78: High-level overview of the BNL HTSN and demarcation points between core and collaboration infrastructures

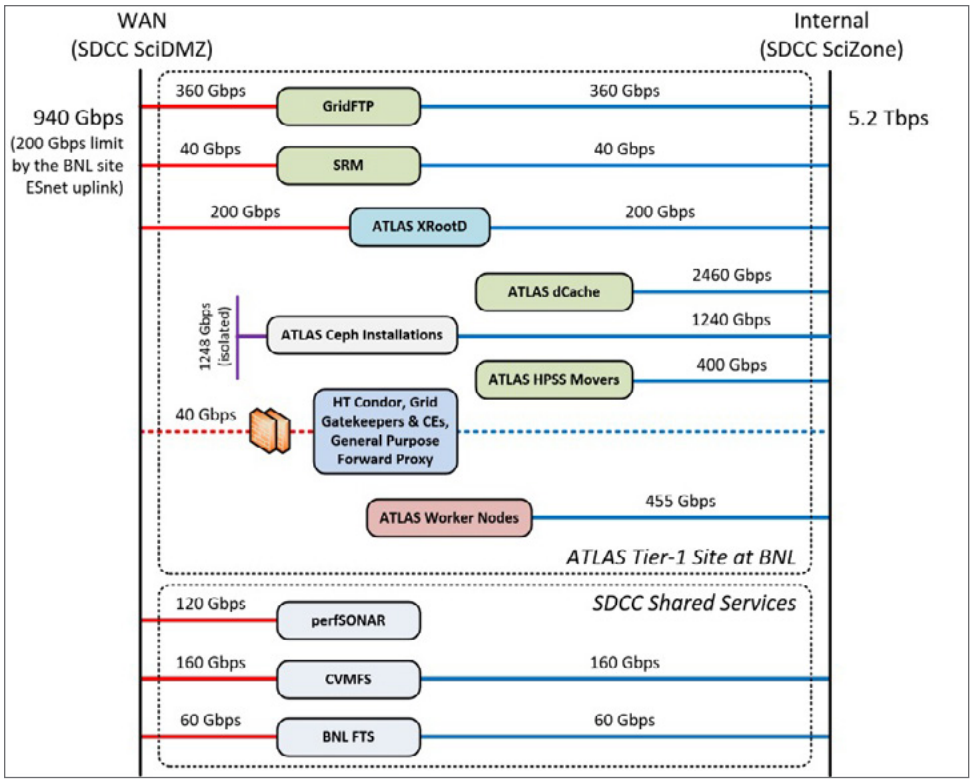


Figure 79: High-level network layout of the ATLAS Tier 1 site at BNL site

5.10.5.7.2 US ATLAS T2 Infrastructure

Each Tier 2 site has unique LAN/WAN architecture developed in coordination with local and regional network managers. ATLAS will show some of the representative architectures here. The three network diagrams below are for the various sites in the MWT2 starting with the University of Chicago site. MWT2 is the largest Tier 2 in US ATLAS, about 1.5 times the size of the other Tier 2s.

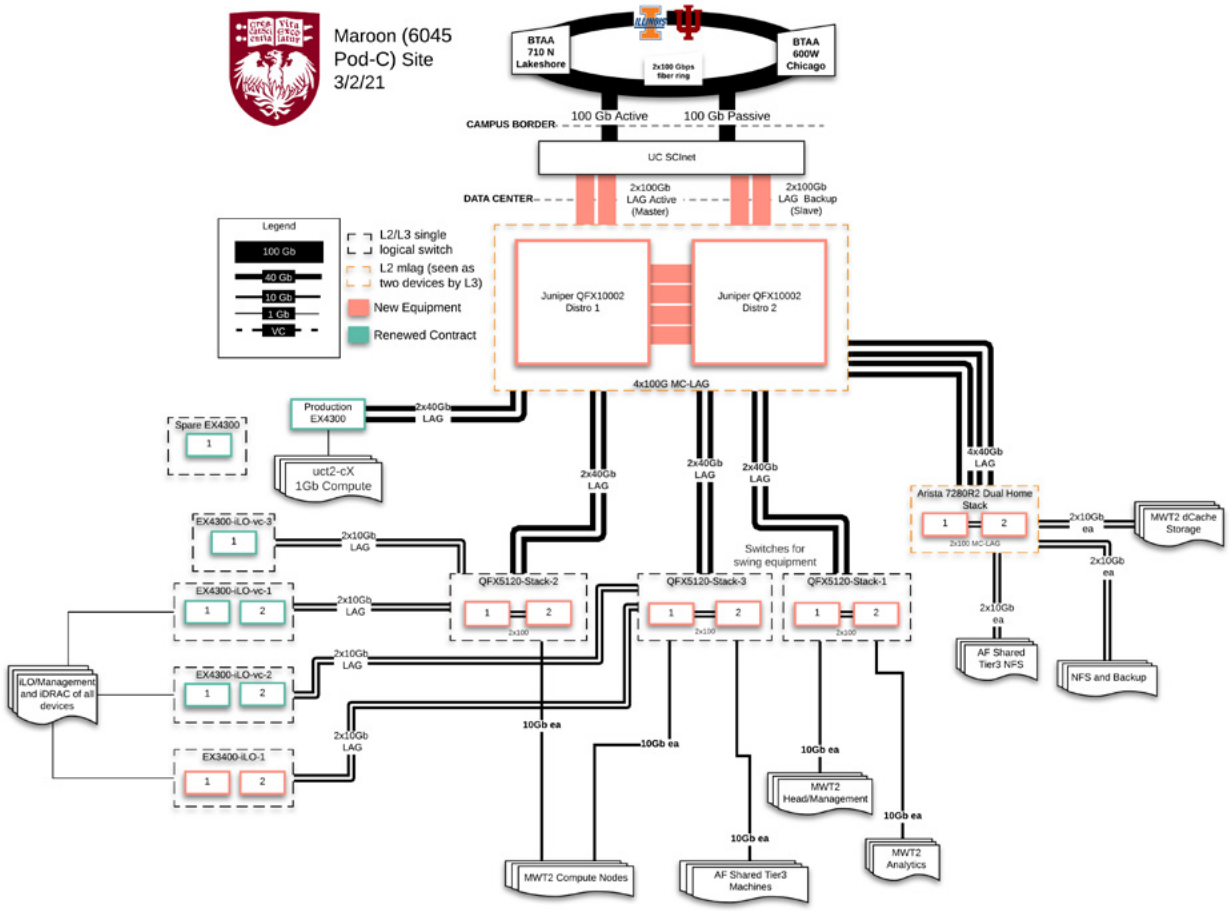


Figure 80: MWT2 at the University of Chicago

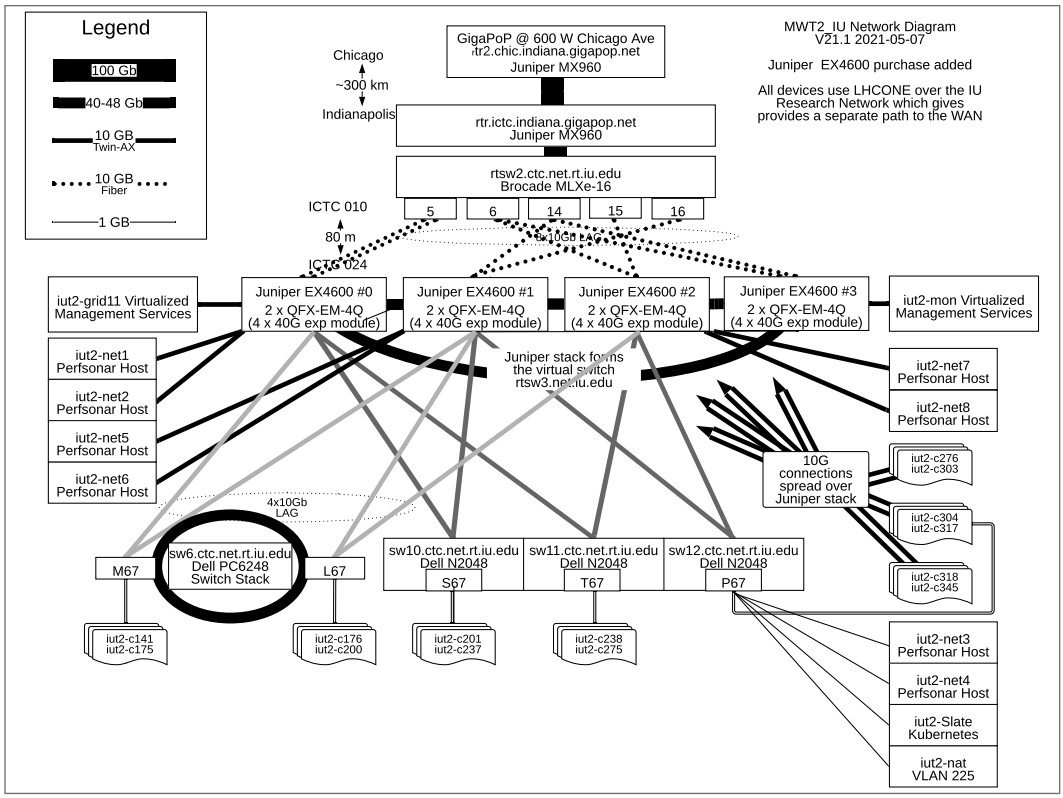


Figure 81: MWT2 at Indiana University

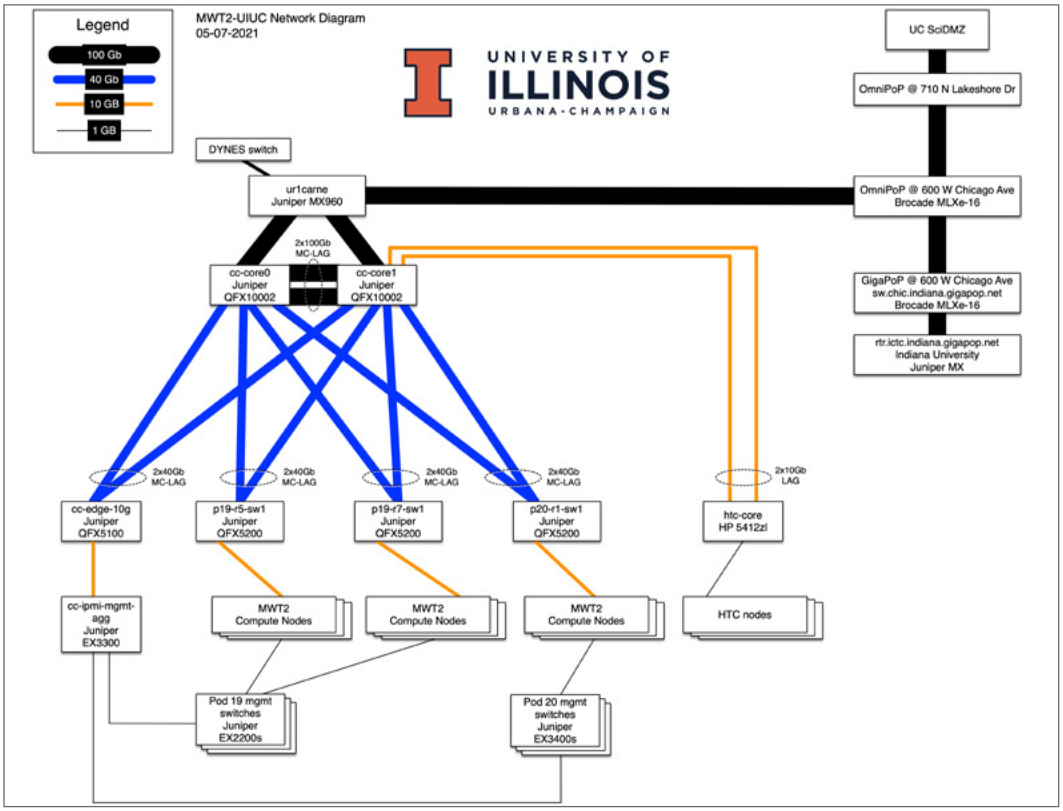


Figure 82: MWT2 at the University of Illinois Urbana-Champaign

5.10.5.8 Cloud Services

PanDA+Rucio can use commercial cloud resources interchangeably with grid-based WLCG resources, though such resources are currently not available in HEP. However, commercial cloud resources are being evaluated for specialized usage by analyzers. There are currently two proof of concept projects. If these projects are successful, ATLAS will require a good network pipe between grid sites and commercial clouds. In this model, the network needs will be similar to university-based US Tier 2 sites. ATLAS expects a few PB of data transfers to cloud sites on a daily basis starting in 2021.

- Google: ATLAS is testing the use of GCP+GCS for end-user analysis. The project is funded for two years, with a decision expected in late 2021.
- Amazon: ATLAS is testing a virtual analysis center on AWS. This project is funded till summer 2021. A decision is expected soon after.

5.10.5.9 Data-Related Resource Constraints

A primary concern is lack of sufficiently fine-grained network monitoring and performance information to help us debug and automate data transfers in real time between various sites.

5.10.5.10 Outstanding Issues

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 in 2022, it is anticipated that network needs will grow gradually, following the patterns described in this document. Increasing network capacity and performance will be needed at US ATLAS Tier 1, Tier 3, and Tier 3 AFs. The exact magnitude of future network provisioning is expected to be determined jointly by US ATLAS, ESnet, and other involved parties.

5.10.5.11 Case Study Contributors

ATLAS Operations Representation

- Alexei Klimentov[210], BNL
- Robert Gardner[211], University of Chicago
- Shawn Mckee[212], University of Michigan
- Paolo Calafiura[213], LBNL
- Kaushik De[214], University of Texas at Arlington
- Johannes Elmsheuser[215], BNL
- Wei Yang[216], SLAC
- Eric Lancon[217], BNL
- Srini Rajagopalan[218], BNL

[210] aak@bnl.gov
[211] rwg@uchicago.edu
[212] smckee@umich.edu
[213] pcalafiura@lbl.gov
[214] kaushik@uta.edu
[215] johannes.elmsheuser@cern.ch
[216] yangw@slac.stanford.edu
[217] elancon@bnl.gov
[218] srinir@bnl.gov

- Chris Bee[219], Stony Brook University
- Alexandr Zaytsev[220], BNL

ESnet Site Coordinator Committee Representation

- Vincent Bonafede[221], BNL
- Mark Lukasczyk[222], BNL

5.10.6 CMS Experiment Case Study

5.10.6.1 Background

The current document is one of three documents that the US CMS Software and Computing Operations Program drafted as input to the ESnet requirements process in 2020. The first drafts of these documents were finalized at the end of July 2020.

This document functions as a high-level introduction describing the process of science within the larger context of the global CMS collaboration: how data are produced, how science is derived from these data, CMS's computing facilities, and its requirements process. This document also provides more details about the networking implications of performing LHC particle physics specifically at the Tier 1 site at Fermilab, the seven Tier 2 sites at US universities, and the numerous other "sites" used to process data and to perform analysis. CMS researchers also utilize campus or research group facilities at about 50 US institutions, the largest of which are Tier 3 sites, as well as individually managed computing resources of each of the approximately 700 researchers (e.g., laptops) within US CMS. Beyond the seven Tier 2 facilities discussed here, the Vanderbilt Tier 2 for DOE-NP went through its requirements process in 2019, and is not included in this document.

The technical aspects of the LHC operations and the R&D activities toward LHC will be described in three separate case study documents:

1. CMS Experiment (this document).
2. Technical aspects of LHC Operations.
3. R&D activities toward HL-LHC.

All of these documents are written from the perspective of networking and data. They should thus not be thought of as "comprehensive documents" that describe the entire computing model, but rather as overview documents to highlight the role networking plays today, and in the future. Across all documents, three periods are identified: now until Run 3, Run 3, and Run 4. These map roughly on the time periods 2020–2021, 2022–2024, and 2028–2030. The time period from 2024 to 2028 is a transition period that is presently hard to predict. It will probably include some large-scale data challenges that are discussed in the third case study.

The CMS experiment at the LHC is designed to probe new phenomena at the energy frontier. Since starting operations in 2010, CMS has collected nearly 200 fb^{-1} of luminosity at center-of-mass energies between 7 and 13 TeV.

The CMS collaboration is made up of more than 3,000 members from more than 50 countries. Researchers at US institutions comprise about 30% of the collaboration. The collaboration has published more than 1,000 papers with scientific findings across a broad physics program enabled by all collaborators having access to the entire data sample for their work. No previous HEP experiment has produced this many publications. The success of the physics program of CMS depends on the availability of sufficient computing resources to store, process, and analyze the data in an efficient fashion.

---

[219] christopher.bee@stonybrook.edu
[220] alezayt@bnl.gov
[221] bonafede@bnl.gov
[222] mlukasczyk@bnl.gov

The LHC provides proton-proton bunch crossings (where each crossing produces many proton-proton interactions) in CMS at a rate of nearly 40 MHz at a duty cycle of about 50% during eight months of a typical running year. CMS commonly uses the word "event" to refer to all the data that correspond to a single bunch crossing, both for collision data and simulated data. The interactions that result from each beam crossing are independent of those from other beam crossings. This means that CMS can reduce, model, and analyze each event as being independent from all other events.

40 MHz worth of detector data (called raw data in this document) cannot all be saved for analyst use, in part because raw events are nearly 1 MB in size. CMS uses a two-step filtering system ("Level-1" and High-Level Trigger [HLT]) to reduce the data rate down to approximately 1 kHz of events to be processed and made available to the collaboration for analysis.

Part of this online data reduction process is to broadly categorize events into "data sets" according to the physics signatures observed during the filtering process. Data sets are also separated according to the data-taking period to ensure that any large changes in LHC conditions, CMS detector operations, or software used can be treated separately by analysts. A typical CMS publication requires analysis of a few data sets of detector data, and a few dozen simulation data sets for each year of data taking it includes. Larger endeavors may require up to 50% of the entire detector data and corresponding simulations. Analysts nearly always aim to analyze as much of the available integrated luminosity as possible (e.g., using as much run time as possible). There are more than a hundred analyses ongoing in CMS at any given time, each engaging between a handful and up to dozens of researchers who are actively involved in the analysis of data (either directly from production, or using samples derived from it).

CMS executes a variety of tasks on its distributed computing infrastructure for reconstructing collision data recorded by the detector, simulating collision data, and analyzing both. The CMS computing system relies on infrastructure distributed worldwide, and as such relies heavily on excellent network connectivity among its dedicated computing centers, and increasingly its connections to shared facilities at universities, HPC centers, and potentially cloud resources when economically viable. Beyond the substantial real-time facility for data taking that is co-located with the CMS detector, CERN facilities are used to provide the initial data reconstruction to support detector commissioning work as well as physics analysis work. Then globally distributed Tier 1 and Tier 2 facilities are responsible for data archiving (at the Tier 1 only), simulated data generation, analysis data storage, and physics analysis activities (primarily at the Tier 2s). The United States operates one Tier 1 facility (Fermilab) and seven Tier 2 facilities (Univ. of Florida, UCSD, Caltech, MIT, Univ. of Wisconsin, Univ. of Nebraska, and Purdue University). Data also move from these facilities to other universities as analysis data sets are reduced and refined during the analysis process.

In a nutshell, one may think of the science publication process of the CMS experiment as starting with detector operations to collect data, then central production to process and do initial data reduction and validation, followed by a decentralized (but coordinated) data analysis process that derives results by using progressively improved reconstruction and analysis methods, as well as new approaches and ideas that come from across the collaboration. This is followed by a convergence of those results through multiple stages of discussion and sometimes refinements of the analysis, leading to publications via a centrally organized internal review process. Publications are then passed onwards to the peer review processes of the respective journals.

The CMS production and analysis process is quite storage and data-movement intensive. Data sets are replicated, moved, or deleted according to their usage patterns, site and resource availability, and experiment priorities. Individual analysts or analysis groups create and manage further derived data sets for their analysis workflows. Analyses are done by small and large groups that share data either at a CMS Tier 2 center, at a local Tier 3, or on other local computing facilities. An "average" event collected by CMS is processed 50–100 times by analysts. That means that CMS analysts are processing hundreds of petabytes each year just to carry out the initial analysis steps which process data created by the production system.

CMS devotes 30% of its disk storage to derived data sets, or approximately 50 PB. These data are typically used much more frequently, as the size for any given analysis is smaller, and can be more quickly processed as the analysis process evolves. In addition to these resources, final analysis steps are done locally on the laptops, desktops, or computing clusters of individual researchers or groups. Reliable distribution of data to these researchers is essential to the CMS analysis process. It enables researchers to work outside of a centrally managed and maintained system, and it thus reduces the complexity and management difficulty that the central system would otherwise have to face.

The infrastructure to support this process is largely developed and operated by CMS experts, including a significant number of computing professionals as well as physicists who have specialized in software and computing technologies. However, the experiment strives to be sufficiently open in the infrastructure such that computing and storage resources at individual universities can be integrated and used in both the central production and decentralized analysis parts of this process.

As the entire CMS data flow and analysis process relies on collaboration and data sharing, the CMS research program and method are discussed in the next section.

5.10.6.2 Collaborators

The CMS experiment is designed, built, and operated by a collaboration of close to 200 institutions across more than 50 countries, and comprises roughly 3,000 members, of which close to two-thirds are physicists with authorship privileges on all CMS physics papers. The United States makes up about 30% of the authors at about 50 institutions.

The collaboration as a collective produces official data samples via a central production team. Applications composed of algorithms developed by the collaboration convert and reduce the raw detector data into derived formats ready for analysis. These data are distributed across the CMS computing facilities on reliable disk, with topical and popular data sets having the greatest accessibility (e.g., largest number of replicas across the distributed system). Tools and support are provided to facilitate the use of these data. All members of the collaboration have access to all centrally produced data, and are free to pursue any kind of physics analysis that is supported by this data. Researchers use everything from local computing facilities to laptops to interact with these organized resources typically via custom approaches developed according to research group interests, abilities, and prior experience.

Typically, physicists self-organize into small groups of a few to a few dozen researchers to collaborate on one or more physics publications. Here again reduced data samples are shared among members of these groups, or often between groups with related interests. Due to the sheer size of the data CMS, the processing of these reduced data samples often requires substantial resources both to produce and to store. CMS nominally devotes 30% of its disk resources to data samples derived by data analysis groups.

Researchers target conference results and subsequently journal publications for discussion and peer review of their results. Conferences bring a set of seasonal deadlines that affect the way resources are used in the collaboration. Most researchers have a peak of activity for the main spring (i.e., March) and summer (i.e., July) conferences. Production campaigns are therefore often aimed to make data available with the best available algorithms and calibrations in advance of these deadlines.

Results are peer reviewed via a process that is agreed upon by the entire collaboration, and in which every member of the collaboration is asked to participate. Typically, there are multiple stages of work and review. Analyses are initially discussed and reviewed by physicists who are expert in related analyses (e.g., analysis working groups), experts in various reconstruction, event selection and signal-vs-background separation methods, experts specialized in "statistics" (hypothesis testing), and subsequently more broadly by the CMS analysis community. The resulting publications that come out of this process carry the names of all members with authorship privileges. There is a well-defined process for somebody new to the collaboration to obtain and maintain authorship privileges.

CMS researchers collaborate broadly on technical and physics research topics. For example, CMS works closely with software authors, infrastructure providers, and the theoretical particle-physics community. However, as CMS data are largely internal to the collaboration, the impact on network use of these collaborations is minimal. The vast majority of network use is thus networks that connect the close to 200 collaborating institutions within CMS.

User/collaborator and location	Primary or secondary copy of the data	Data access method	Avg. size of data set	Frequency of data transfer or download	Are data sent back to the source and method	Any known issues with data sharing
CERN TIER 0 FACILITY (Geneva, Switzerland) Supports all CMS researchers	Partial primary copy (includes all raw data)	Data transfers Data streaming Access via on-site CPU	26 PB active disk for CMS	Constant (7/24 support)	No	
FERMILAB TIER 1 FACILITY Supports all US CMS researchers. All CMS researchers can access storage via streaming or by grid analysis jobs (5% of CPU devoted to user requests)	Partial primary copy (raw and analysis data)	Data transfers Data streaming Access via on-site CPU	27 PB active disk for CMS + US user facility resources	Constant (7/24 support)		
OTHER CMS TIER 1 FACILITIES: CCIN2P3 (Lyon, France) RAL (Oxford, UK) KIT (Karlsruhe, Germany) JINR (Dubna, Russia) CNAF (Bologna, Italy) PIC (Barcelona, Spain) All CMS researchers can access storage via streaming or by grid analysis jobs (5% of CPU devoted to user requests)	Partial primary copy (raw and analysis data)	Data transfers Data streaming Access via on-site CPU	3–11 PB active disk for CMS	Constant (7/24 support)		
US-CMS Tier 2 facilities: University of Nebraska (Lincoln, NE, USA) University of Wisconsin (Madison, WI, USA) Purdue University (West Lafayette, IN, USA) University of Florida (Jacksonville, FL, USA) MIT (Bates Laboratory) (Middleton, MA, USA) Caltech (Pasadena, CA, USA) UCSD (La Jolla, CA, USA) Each supports about 100 researchers with local account access with compute and storage. All CMS researchers can access storage via streaming or by grid analysis jobs.	Partial primary copy (analysis data)	Data transfers Data streaming Access via on-site CPU	Each site has 3 PB active disk for CMS + additional resources for US community (~3 PB in addition)	Constant (5/8 or better support)		

User/collaborator and location	Primary or secondary copy of the data	Data access method	Avg. size of data set	Frequency of data transfer or download	Are data sent back to the source and method	Any known issues with data sharing
OTHER CMS TIER 2 FACILITIES Each supports regional researchers with local account access with compute and storage. All CMS researchers can access storage via streaming or by grid analysis jobs.	Partial primary copy (analysis data)	Data transfers Data streaming Access via on-site CPU	Storage varies from 100 TB to 10 PB	Constant (5/8 or better support)		
LPC TIER 3 FACILITY AT FERMILAB Supports analysis activities of local researchers, collaborators, and US-based researchers in general	Partial secondary copy	Data streaming Access to Fermilab Tier 1 disk				
US TIER 3 FACILITIES (about 50% of US universities collaborating in CMS) Supports local researchers and collaborators (size varies from few to O(100)). Some sites also provide resources to the greater CMS research community, if storage is accessible by streaming or by grid analysis jobs.	Partial secondary copy	Data transfers Data streaming Access via on-site CPU	Storage varies from 10 TB to 1 PB	Ad hoc, but some facilities have constant data transfers when running production		
OTHER US UNIVERSITIES (about 30% of US universities collaborating in CMS) Supports local university group (size varies from few to dozens)	Primary copy of user-derived data	None	Variable active storage (< 1PB)	Ad hoc		
CMS RESEARCHERS (at CMS collaborating institutions worldwide including ~50 institutions and ~700 researchers in the USA)	Primary or secondary copy of user-derived data	None	Variable active storage(<100 TB)	Ad hoc		
HPC CENTERS (see Section 5.10.6.3, "HPC Facilities," below)	None	N/A	O (1 PB) of storage for transient samples needed for production	Ad hoc (constant when running CMS applications)		

Table 22: CMS data projections

Data streaming is defined to mean remote file open, and direct reads to remotely opened files. Data transfer is defined to mean transfer of entire files, most often in bulk.

5.10.6.3 Instruments and Facilities

The LHC computing and storage infrastructure is organized in a tiered computing model including four tiers, 0–3. Originally, when the connectivity among the sites was extremely limited, the tiers were hierarchical in that Tier-N+1 was connected only to an associated Tier-N above and associated Tier-N+2 below. Data were not allowed to flow from Tier-N+2 to Tier-N. This rigid concept has been eliminated over the last decade as network capacity and capability progressed.

Data now flow across the full mesh, among all sites of all tiers. Today, and in the future, the global CMS collaboration together with WLCG and OSG will define services that sites perform. Today the service profile defines the site to belong to a certain tier. For the HL-LHC, even that might change. In the future, centers of a given tier today 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. These changes are part of an ongoing R&D program toward an updated computing model for HL-LHC, and will be discussed in the HL-LHC (Section 5.10.8) in more detail.

In the following subsections, the current services profile for each of the tiers is described, starting with the high-level trigger compute cluster (HLT).

5.10.6.3.1 HLT

The HLT computing system is a roughly 40,000-core x86 cluster that will have GPUs added during Run 3. Its primary purpose is to perform the last stage of the online filtering of events. All events that pass this filter will be stored forever in the archive as raw data. It is located at the site of the CMS detector roughly 10 miles away from the main CERN site, on the other side of the LHC ring. The HLT is part of TriDAS, the CMS trigger and data acquisition system that is responsible for all of the real-time data reduction and collection for CMS. Its size, architecture, and components are reviewed annually, independently of the annual offline and computing requirements process discussed further below.

5.10.6.3.2 Tier 0

The Tier 0 is at CERN and has two primary functions. It is responsible for "prompt calibration and reconstruction" and archiving of all the raw data at CERN. When the LHC is not running, the Tier 0 compute cluster may be used for other processing by CMS.

5.10.6.3.3 Tier 1

CMS globally has seven Tier 1s, including Fermilab in the United States. The collection of the Tier 1s is responsible for operating a distributed archive of all centrally produced data, including a second copy of the raw. The copy of the raw across the Tier 1s is considered the "active copy" (i.e., the raw copy at CERN is considered strictly a backup). In addition, the Tier 1s provide computing resources to the experiment that are primarily used for centrally organized processing. Historically, each Tier 1's computing was responsible for all the reprocessing of the data it archives. The computing and storage capacity of a given Tier 1 center was thus directly related to the amount of data it pledged to archive. More information on pledges follows.

5.10.6.3.4 Tier 2

CMS has roughly 50 Tier 2s globally. The collection of the Tier 2s is responsible for all data analysis and some simulations. Tier 2s also provide disk space to host data for analysis, and a moderate amount of disk space for staging in data for processing.

5.10.6.3.5 CMS Computing Capacity and Planning

The Tier 0, 1, and 2 resources are subject to an annual needs planning process. CMS offline and computing management makes a needs assessment every year that projects out two years into the future. After internal review, the annual needs are reviewed externally by the LHC Coordinating Committee (LHCC) twice a year. After sign-off by the LHCC, the needs determined in this way are used to guide the funding agencies in the various member countries, leading to annual computing resource pledges. In the United States, the LHC Operations programs, jointly funded by the DOE and NSF, are responsible for the US pledges, and coordinate them annually with the funding agencies.

Throughout the year, the actual delivered computing capacity is measured from all Tier 1s and Tier 2s via the WLCG accounting system. For the United States, the propagation of usage into this accounting system is the responsibility of OSG. This will be described in case study #12.

5.10.6.3.6 Tier 3

Tier 3 sites are typically university-hosted clusters that are smaller than Tier 2 sites and are typically without dedicated support to ensure high resource availability. Those in the United States are not funded through the Operations Program. Tier 3 sites can be part of the DMS of CMS and receive official data sets, or can access data through streaming. These sites can also provide disk space to hold user data as required by the local community. However, they are not part of the annual requirements review process. As such, the total Tier 3 resources of CMS are presently modest compared to the other tiers.

5.10.6.3.7 HPC Facilities

CMS also uses compute resources beyond the dedicated resources located in one of the tiered computing centers. These include HPC centers in the United States. At present, the CMS usage of shared HPC facilities funded by the DOE and NSF do not amount to a significant fraction of the overall computing budget (approximately 1% of the CPU used by CMS globally), and are not part of any pledges. Recent allocations include 104 million core hours (Mhours) at NERSC, 1.2 Mhours on Theta and Cori (via a shared proposal with ATLAS), and 25Mhours hours at the Pittsburgh Supercomputing Center (PSC), TACC, and San Diego Supercomputing Center. CMS expects that this model will change for the HL-LHC to include larger allocations on HPC facilities. The size and locations of future facilities are not fully predictable; however, CMS would rely on excellent network connectivity to any HPC facility from its Tier 1 and Tier 2 facilities. For example, NSF facilities typically take one or two years from proposal acceptance to being in production. For more details on this see the HL-LHC case study.

Today, HPC centers funded by the DOE and NSF as shared facilities across multiple science domains do not generally provide the full functionality required from a Tier 2 or Tier 1 center (e.g., they normally do not provide any substantial amount of storage under the control of the experiment). Data access at HPC centers relies on either temporarily staging in input data on storage at the HPC site, or streaming access if the site setup allows for access to the wider network. In general, HPC centers only run compute dominated applications that have modest IO/CPU requirements (e.g., the central production workflows involved in central processing and simulations). Output is archived back to the Tier 1 sites, either directly from the application or through other mechanisms.

5.10.6.3.8 Data Flows

This infrastructure setup defines the data flows. Raw and official data sets flow from Tier 0 to the Tier 1s for archival storage. Simulations are produced on all tiers and are transferred from everywhere to the Tier 1s for archival. Data sets that are needed for input to analysis are distributed across the Tier 1 and Tier 2 disk resources. The distribution is automatic and follows rules defined by the collaboration. The automatic distribution has the ability to increase the number of replicas of data sets to allow for more processing resources to have access to the data, as well as to decrease replication when demand for data sets is decreasing. All disk resources in CMS are accessible through the AAA federated data access system, implemented via XROOTD, and can stream data to remote applications. CMS production jobs have a typical output rate of 1 MB/sec per processing hyperthreaded; compute nodes with 128 Hyperthreads (HT) or more will require more than just a 1 Gbps LAN connection to the worker nodes. Accordingly, most worker nodes at Tier 1 and Tier 2 are connected at 1 Gbps or more to their LAN.

Table 23 shows the most recent LHCC-reviewed needs and pledges for the CMS Tier 0, Tier 1, and Tier 2 facilities. The LHC has its own performance unit to standardize performance of different CPU processors: HS06. For reference, a Dual AMD EPYC 7451 system (2x24 cores, or 96 hyperthreads per node) provides roughly 1,100 HS06, and a US CMS Tier 2 provides upwards of 100 kHS06 in aggregate. As is probably obvious from the description, it is possible to have more or less pledged than needed, and more or less consumed than pledged. In addition, resource providers commonly provide resources to their local researchers beyond what

is pledged to CMS. Both the US Tier 2s and the Fermilab LPC provide US researchers with CPU and disk resources beyond the US share of the CMS resource pledges. The documentation that is prepared for the LHCC review does include detailed quantitative discussion about all of this. It is the primary reconciliation process of what is deemed needed, what was pledged, and what was actually consumed. The United States contributes approximately its authorship share in Tier 1 and Tier 2 pledged resources to CMS.

Resource	Site	2020 CMS approved request (spring 19)	2020 pledges to CMS
CPU (KHS06)	T0+CAF	423	423
	Tier 1	650	693
	Tier 2	1000	985
	Total	**2073**	**2101**
DISK (PB)	T0+CAF	26.1	26.1
	Tier 1	68.0	67.5
	Tier 2	78.0	76.8
	Total	**172.1**	**170.4**
TAPE (PB)	T0+CAF	99	99
	Tier 1	220	194
	Total	**319**	**293**

Table 23: CMS pledges

CMS also projects these needs into the future, including the first running years of HL-LHC, as shown in **Figure 83** for CPU and **Figure 84** for disk. These projections are made assuming little change in the CMS computing or analysis model. The ongoing R&D described in the HL-LHC case study will have a big impact on these projections.

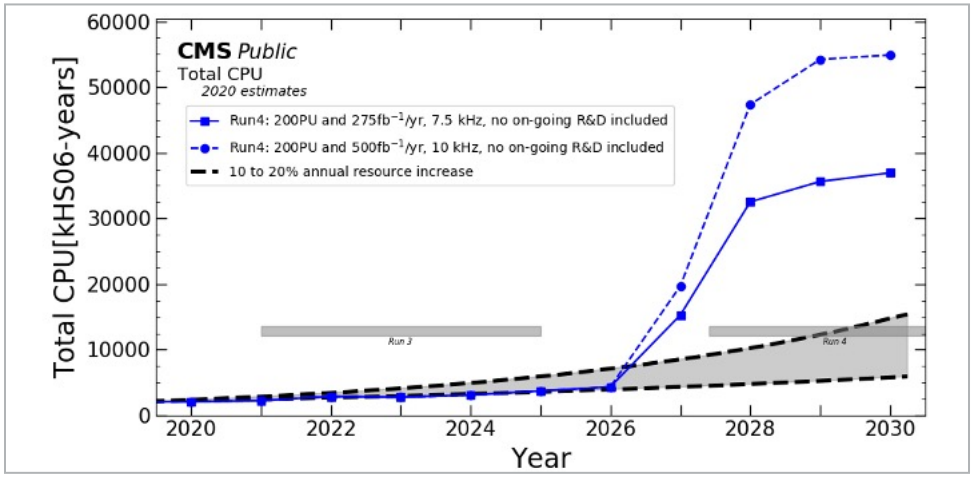


Figure 83: CMS CPU

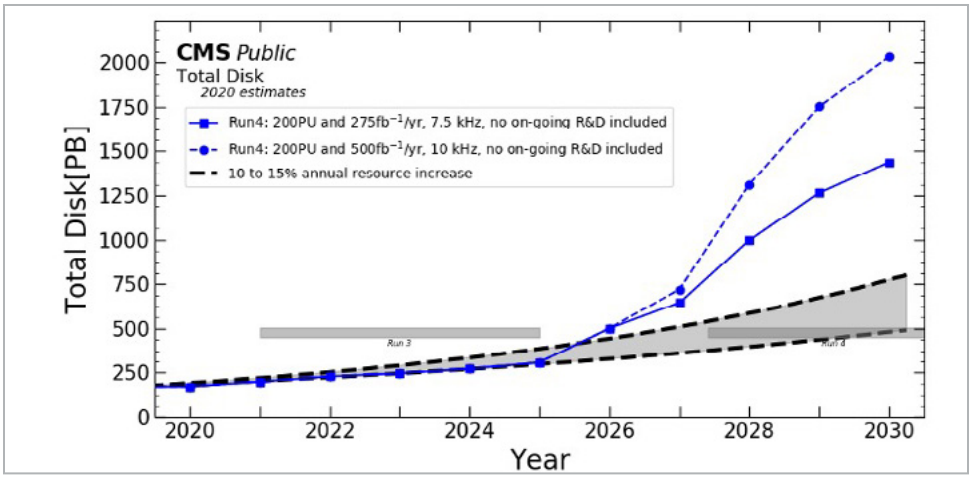


Figure 84: CMS disk

5.10.6.4 Process of Science

For the purpose of this discussion, CMS distinguishes three types of data: data from collisions in the accelerator, data from simulations, and user-produced data. The first two are centrally produced, while the third is produced by the physicists themselves as part of their analysis workflow. In the following subsections these three types of data will be described to illuminate their fundamental differences.

5.10.6.4.1 Collision Data

The LHC operates at a typical cadence of three running years, followed by two shutdown years. Within a running year, collisions start approximately March to May, and end in November or December. Software releases are prepared for the beginning of data taking each year. Changes from one year to the next within a running period are typically more modest than changes from one multi-year running period to the next.

The LHC provides data to CMS in roughly 10-hour data-taking periods with minimal downtime (typically a few hours) between periods. Therefore, a roughly constant stream of data comes out of the TriDAS, the combined trigger and data acquisition system. Event rates of approximately 1 kHz of physics events are typical today and for Run 3, and 7.5 kHz is expected for HL-LHC. Detector level correction (calibrations) processes are run within 48 hours to allow for a prompt reconstruction to start. Beyond the baseline program, there are also targeted data-taking periods with higher data-taking rates. This translates into raw data volumes of:

- Run 2: 45 PB written during four years (2015–2018).
- Run 3: ~45 PB (2022–2024, with current 2022 program plan, and baseline program for 2023+2024).
- Run 4: 350 PB/yr (only baseline program).

It is important to note that to capture the total CMS data volume or flow, these figures are to be multiplied by several factors. For example, data are typically processed by multiple software versions, multiple replicas of important data are kept on disk, and multiple formats for analysis data are provided to researchers.

During the annual running period, bug fixes to the software are introduced into both the online and the prompt reconstruction. In some years, these changes are significant due to the discovery of features of the CMS detector that were not expected nor foreseen (e.g., things break occasionally, requiring adjustments in software because the instrument is inaccessible until shutdowns).

Toward the end of the annual running period, a final software release with the best possible calibrations, alignments, and reconstruction methods is prepared, and the entire year's worth of data are processed a second time. During the shutdown between two running periods, the entire data from that period are then processed one more time to arrive at a data set that is as consistent as possible across the three-year running period.

Data analysis may start at any moment in this cycle, and thus may be performed on any of the three versions of the collision data. A given version of the data thus has a "lifecycle" from the time it is produced until it is retired. All centrally produced data are archived in a distributed tape archive across the Tier 0 and the Tier 1 centers from the time they become available until they are retired. Only raw data and the most recent reprocessing are archived forever, and raw data are archived at the Tier 0 and the Tier 1 sites to achieve two physically separated copies. CMS refers to the last reprocessing as the "legacy" version of the data.

The retirement date is subject to negotiations within the global collaboration with a tension between wanting to retire data to free tape resources, and wanting to retain versions necessary to support ongoing analyses until they are accepted for publication.

There are about 50 "data streams" coming out of the TriDAS. These streams are designed to take into account online trigger decisions and minimize overlaps among streams. The target is to allow at most 10% overall overlap among all streams, i.e., the sum of all collisions in these streams is no more than 10% larger than the sum of all collisions exiting the TriDAS. These data streams vary in size by about a factor of 70 from the largest to the smallest.

In addition, depending on the stability of the detector throughout the annual running period, there may be three to five distinct "epochs" within a year. Epochs are typically similar in duration, one to a few months, varying by at most a factor of five in duration. During an epoch, the raw data from a given data stream are transferred consistently to a particular Tier 1 site for archival purposes. Rebalancing of archival sites happens only at epoch breakpoints. Multiplying the 50 trigger streams by the three to five epochs by the three versions, CMS observes that there are O(500) unique data sets from collisions for a given year's data collection. Each of these is managed separately across the CMS computing infrastructure.

The collaboration has developed two additional data collection modes that allow it to increase the effective trigger rate without the need for proportionately scaling up the complete computing infrastructure.

1. The "scouting" stream consists of only properties calculated during the trigger decision and does not contain the full information content of all detector hits. It is thus much smaller in size per event, allowing many more events to be saved for the same total bandwidth and saved volume of data. Therefore, the scouting stream can be written at much higher rates than the physics streams. The scouting stream can be stored on minimal disk and tape resources and analyzed directly by the collaboration. Reprocessing of scouting stream data is not possible as not enough information per event is kept.
2. In addition to the scouting stream, the collaboration can "park" data produced by TriDAS directly, while skipping the following processing steps. The raw data are stored on tape and recalled at a later time, most commonly in between the running periods when the accelerator and detectors are maintained and upgraded. At that point, the full processing is done on these data and the output is provided to the collaboration for analysis just like any other data produced by the CMS detector. Parked data significantly increase the total tape archiving requirements but have less of an impact on the total computing and storage needs until they are processed. In addition, CMS has traditionally kept parked data only on tape at the CERN Tier 0 in order to minimize the impact they have on the distributed computing system.

5.10.6.4.2 Simulation Data

In addition to data from the detector, CMS has produced simulations of roughly two to three times as many collisions and plans to continue this practice during Run 3. This means 20 billion events are produced during

a typical year. For the HL-LHC, it is currently envisioned to have roughly the same number of events from simulation as from the detector (we expect the ratio of simulated to data events to decrease but the overall number of data events to increase substantially). Simulated data serves all aspects of the CMS scientific process. Simulations are needed to commission detectors, develop detector calibration and alignment procedures, design algorithms that define higher-level physics objects from lower-level instrument readings, determine their efficiency and resolution, and develop data analysis strategies for obtaining results. A mix of simulation and carefully designed measurements based on detector data are used to determine the backgrounds to the expected or hypothesized signal for any given physics analysis. Simulations are thus of central importance to the physics program of CMS.

In total, CMS has about 140 PB of Run 2 MC simulation data sets stored on tape. While some intermediate clean up campaigns have taken place, this is representative of the total four-year production of simulation. The parameters that drive the evolution of data volumes from Run 2 to Run 3 to the HL-LHC are discussed later in this case study.

As the nature of the known, yet to be detected, and hypothesized new physics processes explored at the LHC is very diverse, spanning more than ten orders of magnitude of production rate between common and rare processes, the simulations are correspondingly diverse. This is a fundamental difference between the collision data described previously and simulations. For collision data, TriDAS decides how data sets are put together, while for simulations it is the configuration of the physics generator, the software that produces the simulated energy-momentum four vectors of the particles in an event at or near their point of origin, that determines a data set. As a result, there are 10,000 to 20,000 distinct data sets in a typical annual simulation campaign compared with about 50 distinct data sets from the detector. The size of the simulation samples varies by orders of magnitude from roughly 10,000 to 100 million simulated collisions per sample. While 70% of all simulation data sets in 2016 (a typical data-taking year during Run2) had fewer than 100,000 collision events in them, 80% of all the simulated collision events were in data sets with more than 10 million events per data set.

Simulation data has the same data lifecycle issues for the same reasons as detector data. In fact, CMS typically produces mostly the same set of physics processes for every software release used in detector data processing. Simulated samples that correspond to the legacy processing are kept in archive.

5.10.6.4.3 User Data

As they are the starting point of all physics publications, the data collections or "streams" within the centrally produced CMS data are defined to ease and increase the efficiency of data access for the analyses. A specific analysis typically uses only a fraction of the events collected by the experiment; it uses only certain data streams, and only a subset of the objects that characterize the events within those streams that are available from central production. The first step of most analysis efforts is thus the extraction and formation of user-defined data sets that will become the basis of further analysis. This is most often a massive downselect of petabytes of official data into terabytes of user data. It also often involves a format conversion from formats that are relatively slow to process, e.g., an event processing rate of a few to 10 Hz per CPU core, to a "custom n-tuple" that is afterwards processed at multiple kHz. This supports the nature of end-user analysis which is interactive in nature and needs rapid repetition of making plots and tables.

Until recently, it has been up to the user community to produce and manage its custom "n-tuples." CMS has made a first step toward simplifying this process by introducing the NanoAOD, a data format designed for interactive end-user analysis, and produced centrally. CMS expects adoption of this format to grow in Run 3 with the goal that 50% of CMS analyses are able to use the NanoAOD as their primary data tier by the end of Run 3. For the HL-LHC era, additional R&D is being pursued, especially within the context of IRIS-HEP[223], to further simplify and streamline this process. For more details, see the HL-LHC use case document.Managing the disk space to host the user data is only partly within the scope of the global CMS collaboration. Twenty percent of the disk resources pledged to global CMS are for user data, under the assumption that national entities provide more user disk space to enable the local community. The United States exceeds this as detailed below in the section on

[223] https://iris-hep.org

Tier 2 facilities. User data are not necessarily registered in the central CMS file catalog, or replicable via the CMS DMS. However, this is expected to evolve during Run 3 after CMS moves to Rucio for its DMS.

While CERN provides some user analysis CPU resources, member countries are expected to provide the bulk of these resources for their national communities. The US CMS collaboration chose to place extra resources at the US CMS Tier 1 facility and Tier 2 centers to support user data and user analysis as described in the Tier 1 and Tier 2 sections below. Each US CMS member institution is nominally assigned to one of the US CMS Tier 2 centers, or the Fermilab LPC (a Tier 3 co-located with the Fermilab Tier 1), for both CPU and disk resources.

5.10.6.4.4 Event Reconstruction and Analysis Data Creation

This figure summarizes the typical process for going from raw data through to analysis data in CMS (i.e., all of the centrally run production processing steps that are performed).

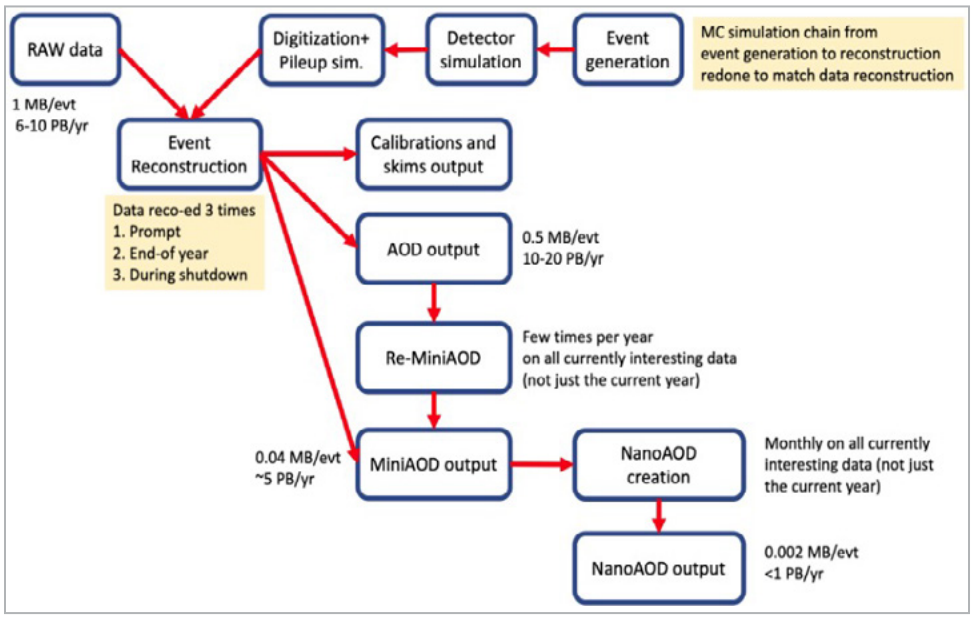


Figure 85: CMS format creation

This multistep process differs slightly for data and simulation processing:

- Prompt reconstruction: 48 hours after a data-taking fill, a period used for deriving calibrations, CMS makes a first pass through the raw data to derive physics objects from the detailed detector information. This process is called reconstruction and happens at CERN on the Tier 0 facility. The reconstruction produces analysis data which are distributed to Tier 1 and Tier 2 disk and Tier 1 tape.
- Simulation production: Simulated samples are created across the CMS Tier 1 and Tier 2 facilities. There are four processing steps: event generation, detector simulation, detector digitization and pileup simulation, and then event reconstruction. The way this has run in CMS has varied with time. Currently the generator step (which models the physics process and produces the particles to be simulated) is run first and saved (a few kB per event). The remaining steps are then run on a single compute node sequentially. This node reads the generator input, typically by streaming it from the site on which it was generated, and produces the analysis data formats. In addition to the event generator input, this process reads a second input file for pileup simulation. The pileup samples are pre-generated and then read via streaming as part of the digitization process. Currently these samples are 2.7 MB/event for Run

2 (and Run 3), and are expected to grow to 13 MB/event for HL-LHC. The pileup samples are each approximately 1 PB and are thus located only at one US site and at CERN. The simulation production takes about 60 seconds on a typical CPU, so the streaming IO need is approximately 50 MB/second/1,000 processing threads. For HL-LHC, this rate is currently estimated to be reduced given the large increase in reconstruction time per event. The simulation production process produces analysis data which are distributed to Tier 1 and Tier 2 disk and Tier 1 tape.

- Re-reconstruction: after final software and detector conditions are derived, CMS reperforms the event reconstruction of all data collected during a year. This processing is much like the prompt reconstruction; however, it relies on the Tier 1 and Tier 2 facilities instead of the Tier 0. Raw data must be recalled from tape for this process (currently done by pre-staging campaigns to recall the data files from tape and distribute them to appropriate sites). It can also be streamed as input (1 MB/event for Run 2 and Run 3 and 7 MB/event for HL-LHC). The reconstruction produces analysis data which are distributed to Tier 1 and Tier 2 disk and Tier 1 tape.
- Analysis-format reduction (MiniAOD and NanoAOD production): as the full event reconstruction takes significant resources, CMS has also developed ways to revise its smaller analysis formats (MiniAOD and NanoAOD), described later in this case study, using information saved in its larger analysis formats. In practice this processing behaves a lot like a re-reconstruction in that it requires its input data to be distributed on Tier 1 and Tier 2 disk, and analysis data formats are produced as output (however, only a subset of them are produced). Outputs are distributed to Tier 1 and Tier 2 disk pools and saved to Tier 1 tape.

Table 24 shows high-level estimates for a typical (or idealized) way of doing event processing in CMS.

Task	Events to process	Input data volume	Inputs streamed?	Inputs staged from tape?	Output data volume
PROMPT RECONSTRUCTION	9 billion/yr	9 PB (1 MB/evt)	No	No	(1 MB/evt)
SIMULATION PRODUCTION	22 billion/yr	59 PB (2.7 MB/evt)	Yes	No	(0.5 MB/evt)
RE-RECONSTRUCTION	6 billion/yr	6 PB (1 MB/evt)	Sometimes	Yes	(0.5 MB/evt)
RE-MINIAOD	16 billion/yr	2.4 PB (0.4 MB/evt)	Sometimes	Yes	(0.05 MB/evt)
RE-NANOAOD	16 billion/yr	0.8 PB (0.05 MB/evt)	Sometimes	No	(0.002 MB/evt)
CENTRALLY RUN ANALYSIS	5 billion/yr	0.25 PB (0.05 MB/evt)	Sometimes	No	Varies

Table 24: CMS processing breakdown

5.10.6.4.5 Analysis Data Formats

The centrally produced data from CMS come in multiple formats ranging from the most versatile and complete (raw and AOD) to the easiest, smallest, and fastest to use ones (MiniAOD and NanoAOD).

Data resulting from central processing are represented in memory (aside from the NanoAOD format) by objects, which are made persistent in files for analysis use. Data are represented in a two-dimensional structure, where one dimension is the collision number (recorded or simulated) and the other dimension is the objects that represent the information content for each collision. The latter can include the raw data, the simulation output, the lower-level reconstructed objects, such as tracks and jets, and the higher-level reconstructed objects, such as electrons, muons, quark-jets, etc. Persistence of data in CMS is based on the ROOT persistency implementation. CMS makes extensive use of ROOT's compression capabilities on storage, and thus data need to be uncompressed during retrieval.

Data formats differ in the level of detail stored per collision. The table below shows that accordingly, the average size of the data stored per collision differs by multiple orders of magnitude between raw and NanoAOD. The HL-LHC numbers are presently largely event size targets, as the actual formats are not yet finalized. The table also shows some values for the typical processing time per event.

	Run 2	Run 3	HL-LHC
LHC ENERGY (TEV)	13	14	14
AVERAGE PILE-UP (PU)	35	55	200
INTEGRATED LUMI (FB^{-1}/YR)	~50	100	480[224]
LIVETIME/YR (10^6 SEC)	6.5	6.5	6.5
PROMPT HLT RATE (KHZ)	1	1	7.5
PARKED HLT RATE (KHZ)	0.4	0.4	0
COLLECTED EVENTS/YR ($X10^9$)	9	9	56
MC EVENTS/YEAR ($X10^9$)	22	31	64
GENERATION/SIMULATION/DIGITIZATION HS06-SECONDS/EVENT	600	625	2050
RECONSTRUCTION HS06-SEC/EVENT	250	370	5000
MINIAOD CREATION HS06-SEC/EVENT			
NANOAOD CREATION HS06-SEC/EVENT	2	2	5
EVENT SIZE FOR RAW	0.9 MB	1 MB	6.5MB
EVENT SIZE FOR AOD	350 kB	400 kB	2 MB
EVENT SIZE FOR MINIAOD	35 kB	60 kB	250 kB
EVENT SIZE FOR NANOAOD	1 kB	1 kB	2 kB

Table 25: LHC parameters for a standard production year in different runs

The use of the various data tiers has evolved considerably over the lifetime of CMS and is expected to continue doing so. It has become possible to use more compact event forms for an increasing fraction of the analyses, as the experiment's software base and the objects used in analysis matured. Run 1 analyses were primarily based on AOD samples. MiniAOD (10x smaller) was introduced for Run 2, and is by now well established as the primary analysis tier, used in about 90% of all analysis activity. MiniAOD includes information on physics objects sufficient to perform various analysis-specific optimizations, and even development of physics object selections, but is not sufficiently detailed to redo CPU intensive aspects of the event reconstruction, for example. Physics analyses for long-lived objects that require specialized tracking algorithms thus need to fall back to AOD. The standardized physics objects that most analyses use are derived from information in MiniAOD rather than being stored directly in the MiniAOD format, and may require a fair amount of processing. Typical processing rates for MiniAOD today range from a few Hz to a few tens of Hz. The MiniAOD format is optimized to support remote reading, as well as some partial file access. An analysis executable typically accesses O(10)% of the data in a MiniAOD file.

More recently, NanoAOD was introduced, which as shown in the previous table is more than an order of magnitude smaller than MiniAOD. Standard physics objects are pre-computed and stored in a flat columnar format. Processing speeds are thus orders of magnitudes faster than MiniAOD. This is achieved via a combination of pre-computed standard objects, smaller event sizes, and more aggressive support of partial file reads.

The goal for Run 3 is for 50% of all physics analysis activities to be based on NanoAOD, with another 40% or more to be based on MiniAOD. The use of AOD for analysis is hoped to be minimal given its large footprint. At HL-LHC scales, CMS may not be able to afford to keep AOD on disk anymore, given its size. In that scenario, access to AOD would require retrieval from archival storage.

[224] For Run 4, this number is scaled down to 275 fb^{-1}/year, according to models with a slower start of HL-LHC operations.

5.10.6.4.6 Data Access

CMS applications and the infrastructure that distributes and launches them support multiple types of data access:

- Traditional top-down data placement at Tier 1 and Tier 2 centers combined with applications specifying the data set they need, and being automatically routed and executed at the sites that have data. In this mode, all data access is local to the site via the site's LAN. The figure below shows the global transfer rate for data set placement by CMS in 2019. Levels of 5 to 6 GB/second were typical during this non-data taking year.
- 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 any data, anywhere, anytime within CMS. CMS considers this a reasonable access modality when the application requires very little IO per CPU. It is used as part of the centrally organized production at some HPC centers simply because simulation, digitization, and reconstruction are all very CPU intensive, leading to small IO/CPU ratios. It is also used as a fail-over when the storage at a site is down while jobs are still running.
- Bottom-up data placement, as is implicit in caching. Here the applications are routed to sites with caches, applications access the cache locally, and cache misses are handled by the CMS XROOTD Data Federation (also referred to as the AAA federation). CMS allows access in this fashion for all of the MiniAOD and NanoAOD formats, but neither AOD nor raw. Raw is on tape only, and AOD is accessible only via top-down placement. Caching is expected to become the dominant data access for end users of MiniAOD and NanoAOD formats.

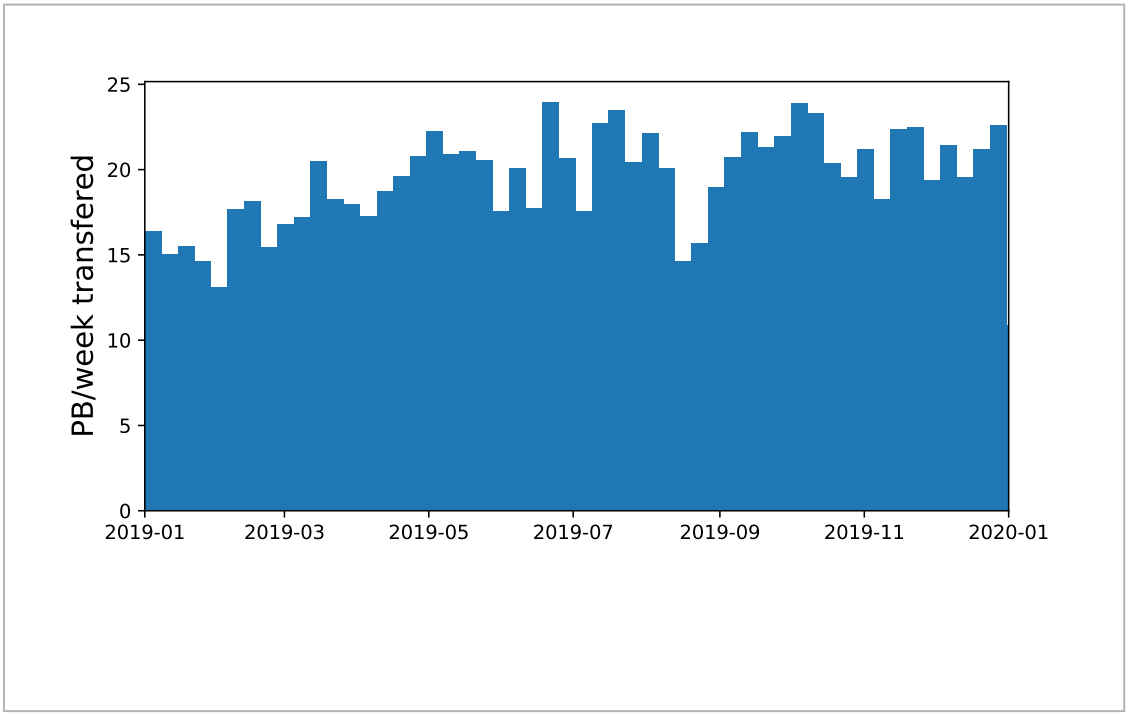


Figure 86: 2019 statistics for data set transfers in CMS for all sites globally

As a consequence, central production workflows use top-down data access and streaming access for workflows that have a very small data to CPU ratio. Caching access is not (yet) used in central production.

Managing data access and data distribution for a distributed storage system is different from a system with one or a handful of sites. One challenge faced by CMS is the need to keep the available disk of each of the

O(50) Tier 2 facilities full of useful data that analysts want. CMS has adopted a dynamic DMS that attempts to do this based on-site availability, site resources, recently used data samples, and other experiment policies for data replication and data cleanup. This system handles data as they are produced by the Tier 0, or by the production system (data reconstruction, MC simulation, MiniAOD, NanoAOD creation, etc.).

In recent years, it has added and deleted more than 40 PB of data from sites per month. As the rate of new subscriptions (newly placed data set replicas at a site) and the rate of deletions (removal of a data set replica from a site) are similar, most of this operation consists of moving a data set from one site to another for operational reasons. It is not clear that this level of growth scales up to the HL-LHC level, where CMS has many more events and larger analysis data formats. Understanding and reducing unneeded data set transfers is important for CMS.

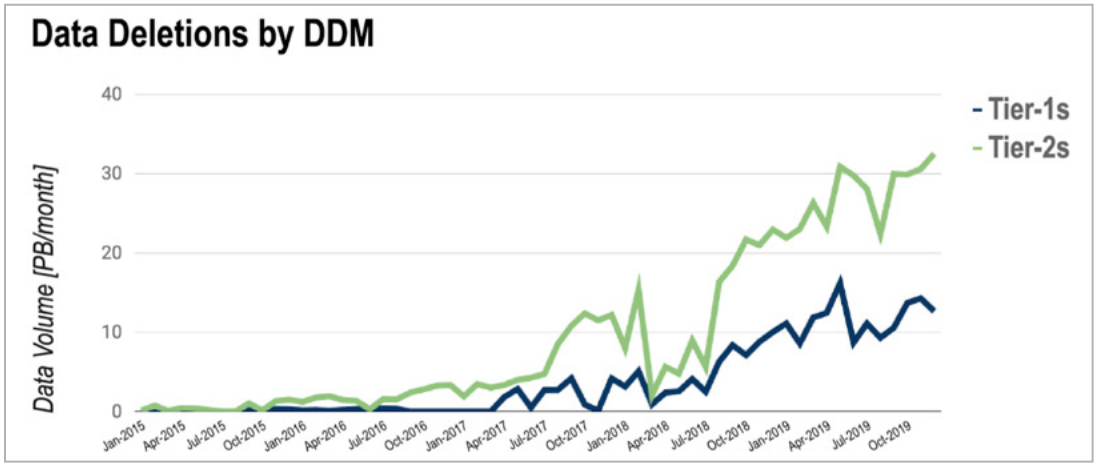


Figure 87: CMS data subscriptions

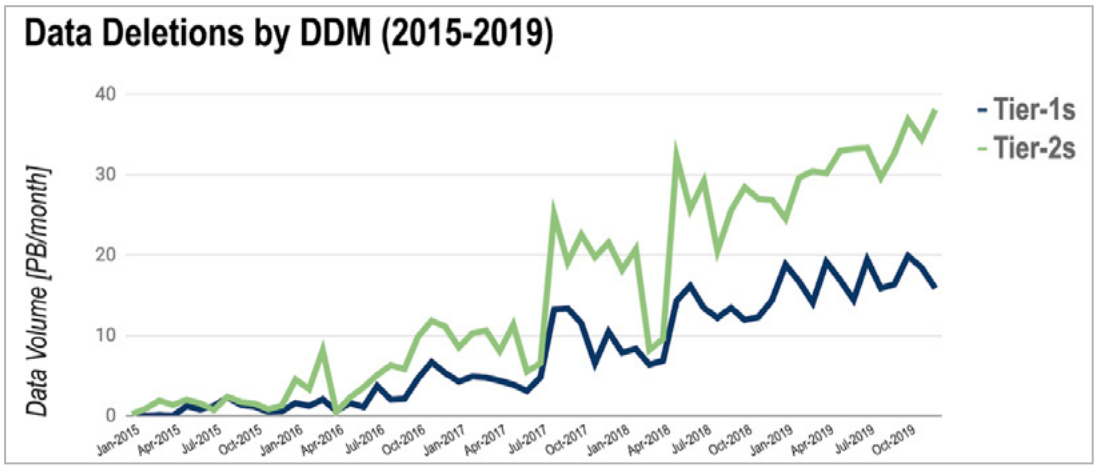


Figure 88: CMS data deletions

User data may be accessed through any of these modalities, although as user data are most often n-tuples that are by design very fast to analyze, they are unlikely to be streamed. In addition, some users reduce the size of their n-tuples to something small enough to be taken outside the CMS computing infrastructure, onto SSDs in laptops, or single stand-alone servers in people's offices or departmental server rooms.

5.10.6.4.7 Analysis Group and Researcher Workflows

So far, this section has focused primarily on the extensive central production system of CMS. In addition to this, an average of about 50,000 cores across Tier 1 and Tier 2 facilities (e.g., 25% of the CMS CPU resource) are used by individual researchers or research groups for data analysis and other compute-intensive work. CMS does not place significant restrictions on the methods and tools used for analysis; therefore, groups have different ways of doing their work. One workflow that will be common for Run 3 analysis work is shown in **Figure 89.**

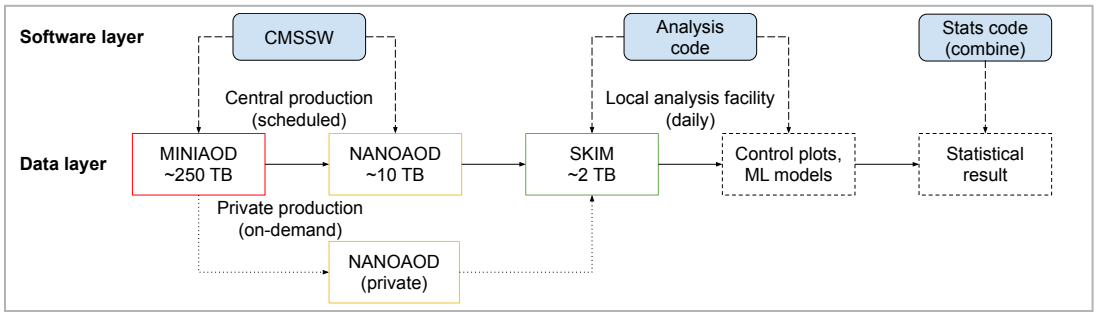


Figure 89: CMS workflow

Data in either MiniAOD or NanoAOD format from the production system are used as input to the analysis workflow. Analysts then reduce these data by removing (aka skimming) events that are deemed uninteresting for the specific physics processes being studied, and by saving only the information in the original data format that is needed for further processing steps. These further processing steps are then performed outside of the distributed computing system, in many cases by transferring data sets to local clusters or other nearby computing systems where interactive access is possible.

Local compute access and smaller data formats provide a more interactive environment for researchers to evaluate different approaches, create proper graphical representations of their data, perform systematic studies, and derive results. This is typically a many step process, with different codes being used through the process. A challenge for CMS through the HL-LHC is to provide researchers with tools that both enable this interactive research at a much larger volume of data and also reduce the need for intermediate processing steps that currently consume significant storage and compute resources.

5.10.6.5 Remote Science Activities

CMS relies on a fully distributed computing system. In that sense, nearly all scientific activities are remote in nature. The use and evolution of these resources are described in the previous sections. Here the CMS network usage and needs with respect to HPC facilities are discussed. As all science in CMS is remote, HPC is by no means special in its remoteness.

Having reliable mechanisms to transfer data into HPC facilities from the CMS Tier 1 and Tier 2 centers is of growing importance. HPCs are different from other resources used by CMS primarily because they provide significant CPU resources but very limited disk resources, and because network access to or from HPC compute nodes is typically limited or not allowed. CMS[225] and the broader LHC community[226] have considered how to best utilize these resources within operational constraints. Unlike most HPC users, CMS cannot concentrate its efforts on using one or a handful of HPC facilities, but is instead asked to use HPC facilities worldwide. Network access and data management are important aspects of this.

Today, the only common solution found by CMS to use US HPCs is to pre-stage any input data. For MC production, the biggest piece of this is the premixing library data which are 2.7 MB for about 60 CPU seconds of processing. Therefore, pre-staging requires short bursts of I/O in order to have data available for jobs. It also

[225] https://cds.cern.ch/record/2707936/files/NOTE2020_002.pdf
[226] https://zenodo.org/record/3647548

means having to plan ahead for what jobs are to be processed on the HPC, which is distinct from the operational model for other resources that CMS uses. For example, pre-staging means either planning far in advance the precise jobs to be run, which is unnatural in a global system, or to be able to transfer data at a rate which is fast compared with the processing time. Instead, if CMS were able to use streaming for HPC input data, this would imply about 4 Gbps per 10,000 cores (e.g., a significant fraction of an HPC facility).

5.10.6.6 Software Infrastructure

The CMS data management software tools are more completely described in case study 12 on infrastructure and tools. A short summary follows.

There are four sources of data transfers of CMS data:

- Bulk transfers due to top-down placement.
- Output handling from analysis or central processing workflows.
- XCache fetching data to handle cache misses.
- Streaming data via XROOT protocol from remote servers to applications.

The first two use cases use CMS-developed tools that sit on top of FTS[227], which is standard across the CERN community. The last two use cases use the XROOT protocol[228].

FTS is used to manage scheduling and file transfer. For bulk transfers, CMS has historically used PhEDEx[229] to handle transfers at the data set (i.e., groups-of-files) level. CMS manages data at the data set level, or for large data sets at the level of "blocks" of files within a data set.

Disk replicas available for analysis are optimized across CMS sites by another tool, Dynamo[230], according to usage patterns and site resource availability. In addition, workflow tools orchestrate the movement of data for production purposes via PhEDEx (recall from tape storage, replicating data to sites with available CPU, etc.).

In November 2020, CMS switched to using Rucio[231] instead of PhEDEx and Dynamo to manage data set storage and data set transfers (while still relying on FTS underneath). Rucio has become a community solution for data management. CMS anticipates using it through Run 3 and beyond.

For bulk transfers via PhEDEx/Rucio and executed by FTS, US CMS is in the process of retiring the use of gridFTP, and replacing it with TPC https, implemented via XROOTD servers. Sites typically have multiple such servers that each provide 10 Gbps, and all have access to the same filesystem. Large bandwidth transfers are thus accomplished by orchestrating very many flows across many servers.

Data streaming uses the XROOT protocol to optimally support partial file reads. XCache handles immediate data transfers, and also uses the XROOT protocol, to refresh the target cache following a cache miss. This is handled so as to minimize latency and thus minimize idle CPU when cache misses occur. The XCache server invokes an XROOT client that initially fetches only the vector of bytes requested by the application, and then later fills in the rest of the file when the server is not too busy.

New concepts of caching are currently being established in CMS. A cache provides access to all or a subset of the official CMS data or user data without the need for organized data movement. The AAA federation is providing data access for the caches. Caches at the level of 1 PB are currently in place at Caltech and UCSD. These are expected to remain approximately the same during Run 3 as current experience with data access patterns shows that this is sufficient for the expected Run 3 CMS data volume.

[227] https://fts.web.cern.ch
[228] https://xrootd.slac.stanford.edu
[229] https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhEDEx
[230] http://t3serv001.mit.edu/~paus/dynamo-documentation/
[231] https://rucio.readthedocs.io/en/latest/

To account for the complete set of all transfers, it is thus sufficient to instrument and monitor XROOTD and FTS within the overall transfer software stack. CMS has recently completed an analysis of the accuracy of the monitoring tools for both of these, and has found issues with both. Thus, a program has started to improve the monitoring capabilities of both in order to arrive at reliable accounting of data transfers. CMS expects this to be complete, including deployment on at least the US Tier 1 and Tier 2 sites, toward the end of 2021.

Fully instrumenting XROOTD software for both the HTTPS and XROOT protocols will provide some redundancy in accounting as FTS will soon use HTTPS as implemented via XROOTD servers at all US-CMS Tier 2s. CMS expects this to be useful to achieve long-term reliable accounting of transfers. Moreover, FTS accounting allows transfer accounting per end-to-end link, while XROOTD servers at both source and destination will have accounting for their outgoing and incoming traffic. Joint work with ESnet staff to regularly exchange network usage accounting information will be particularly useful as these monitoring improvements are put in place.

CMS also relies on high availability for software distribution and distribution of detector conditions and calibration data to production and analysis applications. To avoid redundant data transfers, CMS uses a Squid-caching system, built around Frontier[232] for distributed database caching, and CVMFS[233] for efficient global distribution of software.

5.10.6.7 Network and Data Architecture

5.10.6.7.1 US-CMS Tier 1

The US-CMS Tier 1 Center at Fermilab is the largest of the CMS Tier 1 facilities, currently providing 40% of the total CMS Tier 1 capacity. The Fermilab Tier 1 consists of dedicated computing infrastructure, which includes:

- 260 kHS06 (approximately 27k CPU cores) of compute nodes.
- 27.2 PB of dCache distributed disk storage.
- 88 PB of archival tape storage.

The CMS facility also supports the LPC AF, supporting several hundred physicists with interactive computing nodes, an additional 5000 cores of batch compute, and about 5 PB of distributed EOS storage.

The US-CMS Tier 1's local network infrastructure consists of fully redundant, high-performance switching fabric distributed across multiple Fermilab data centers. That switching fabric has the following characteristics:

- Currently based on 100 GE network technology for inter-switch connectivity, with 10 GE and 100 GE connectivity available for host system connectivity.
- Extensive use of LAG to scale bandwidth for inter-switch connectivity. At the present time, the US-CMS Tier 1 LAN provides approximately 3 Tb/s of network capacity in total.
- PBR techniques to route CMS high-volume traffic over special-purpose networks (such as LHCOPN, LHCONE) utilized for WAN data movement.

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. Most traffic into and out of the US-CMS T1 is via the science data path(s). For CMS, those science data paths mean the LHCOPN and LHCONE. Fermilab's LHCOPN connectivity supports movement of raw data from the T0 (CERN), as well as production data movement with some of the other CMS T1s. Fermilab's LHCOPN configuration consists of three OSCARs circuits (primary, secondary, and tertiary) to CERN, which provide levels of redundancy with differing bandwidth guarantees for that traffic. LHCONE supports production data movement between the Fermilab Tier 1 and most CMS Tier 2s, as well with CMS T1s that do not use LHCOPN for T1-T1 data movement. Like the LHCOPN, connectivity to the LHCONE is via geographically redundant (primary/fail-over) paths. US-CMS Tier

[232] http://frontier.cern.ch
[233] https://cernvm.cern.ch/portal/filesystem

1 WAN traffic that does not utilize either the LHCOPN or LHCONE paths traverses the laboratory's general internet path instead.

In terms of aggregate WAN capacity out of the site, Fermilab currently has three 100 Gb/s links to ESnet via a geographically redundant metro ring. Two 100 Gb/s links are used to support the science data network paths, including LHCOPN and LHCONE. The two science data network links traverse opposite directions of the metro ring, but are link-aggregated into a single 200 Gb/s logical connection at layer-2. The third 100 Gb/s link supports the laboratory's general internet traffic, which includes CMS traffic to/from sites not connected to LHCONE. The 2x100 Gb/s special-purpose networks connection and 100Gb/s general internet connection serve a redundant function for each other.

Future network enhancements include:

- Upgrade of the core US-CMS Tier 1 LAN infrastructure to 400 GE technology will be starting in the end of FY20.
- Upgrade of the Fermilab network perimeter infrastructure to 400 GE technology is expected in the FY21–FY22 time frame, likely to be aligned with availability of 400 GE services from ESnet.
- Additional WAN capacity from ESnet, either in the form of additional 100 GE WAN links or a 400 GE WAN link, will likely be needed as Run 3 commences.

5.10.6.7.2 US-CMS Tier 2s

Current experience at US Tier 2 facilities suggests the following level of network use for various infrastructure components:

- FTS transfers to and from the site: 3–4 Gbps spread across a number of servers (e.g., 10 at a typical site).
- I/O from compute nodes: 1.5 Mbps per thread, corresponding to an average of 15 Gbit/s at a current Tier 2 site.
- XCache servers: 2 Gbps (in or out) per server. As an example, Caltech maintains 360 TB over two cache servers.
- Squids (for CVMFS content): 100 Mbps/squid. Each site runs at least two squids.
- HTCondor gatekeepers: 2M bps/s per gatekeeper. Sites typically have three gatekeepers.
- Other traffic: up to 1 Gbit/s for user interactive access, monitoring, sync with users' desktops and laptops, etc.

CMS has a LAN requirement of 1 Megabyte/HT; i.e., nodes with 128 Hyperthreads (HT) and more will require more than just a single 1 Gbps LAN connection to the worker nodes. Accordingly, most worker nodes at Tier 1 and Tier 2 are connected at 1 Gbps or more to their LAN. Different CMS centers feature networking hardware from a variety of different manufactures. This includes different varieties of LAN aggregation techniques ranging from cluster-wide capabilities to ToR switches that are aggregated into cluster switches via multiple 10 Gbps links. Generally, much of the LAN and WAN networking infrastructure to and within a Tier 2 site has been an institutional contribution to the US CMS Operations Program; i.e., it has not for the most part been directly funded by the program.

The Tier 2 LAN is then typically connected up to the university border switch routers via one or more 100 Gbps links. Depending on location, this campus network connectivity may be shared with others. Most sites then share their 100-Gbps outgoing connection with other Science DMZ customers on campus. Most sites have a regional network that must be traversed before connecting to ESnet. In all cases, that regional network is shared with others. The current theoretical bandwidth of each center to ESnet is thus 100 Gbps, but the actual usable bandwidth is generally less, and will vary over time depending on other activities.

Figure 90 shows the ESnet network map as available on the ESnet website with the US CMS HEP Tier 1 and Tier 2 sites overlaid. Not shown is the Vanderbilt Tier 2 as it is dedicated to heavy-ion physics, and participated in the nuclear physics ESnet requirements process in 2019.

All Tier 2 sites are roughly the same in terms of CPU and storage capacity, and should thus be thought of as having roughly the same networking requirements. Based on this map, CMS can identify specific links that could be used to estimate the growth in networking as a function of time:

1. **SALT — ECHO:** Covers the ESnet bandwidth from UCSD and Caltech to all other Tier 2s and Fermilab, assuming that traffic from Southern California to the East Coast and Midwest is routed via this link.
2. **KANS — STLO:** Adds traffic from University of Nebraska, Lincoln (UNL) to Caltech and UCSD on its way east.
3. **BOST — ALBA:** Traffic to and from MIT west should travel this route.
4. **JACK — ATLA:** Traffic from and to University of Florida Tier 2 should travel this route.

Wisconsin and Purdue connect to ESnet at Starlight at 100Gbps without traversing ESnet. There are thus no obvious ESnet links to characterize traffic from these two sites.

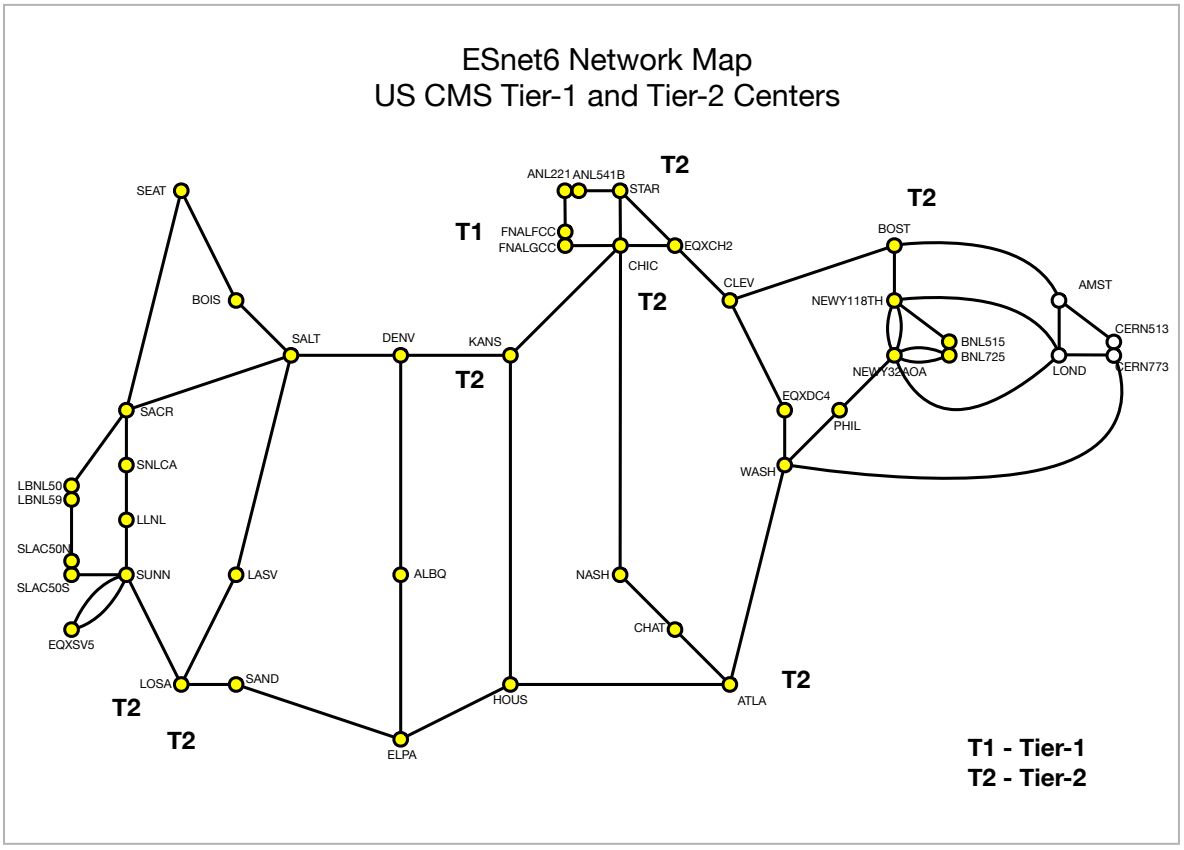


Figure 90: ESnet network map with US-CMS Tier 1 and Tier 2 centers overlaid

US CMS delegates network performance measurement collection to OSG, and expects to continue to do so. At present, Caltech, UCSD, and UNL have 100 Gbps perfSONAR hosts in a MaDDash operated by the Pacific Research Platform (PRP) project. Other locations within this mesh include various Internet2 backbone nodes in Chicago, Manhattan, Kansas City, etc. A corresponding mesh is starting up as part of the OSG-LHC networking

activities. Long term, CMS expects all Tier 2s and the Tier 1 to keep up their perfSONAR instrumentation with the bandwidth requirements for the sites.

All sites participate in perfSONAR measurements, which are archived by OSG at 10 Gbps. A 100 Gbps MaDDash using EHTR is being deployed at present, but it is as yet undecided when the US CMS Operations Program will support, or even require, network performance measurements at this level.

Caltech, UCSD, and UNL have 100 Gbps perfSONAR nodes deployed[234].

5.10.6.8 Cloud Services

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. A notable exception would be the case of needed resource bursts for either CPU or network. The resources available for either CPU or TA networking in the cloud far exceed those available to CMS. At this point, CMS tools are generally able to use cloud services, typically via infrastructure at one of the tiered sites, but CMS does not have plans to use cloud services extensively in the near or longer term. This can evolve depending on largely external factors (e.g., cost evolution or government policies).

5.10.6.9 Data-Related Resource Constraints

We discuss several examples of resource constraints or perceived resource constraints for CMS researchers and their science:

- CMS data volume limitations: the data volume that can be handled by the networks within and coming out of the CMS detector facility far exceeds what can be handled offline within current CPU, storage, and networking infrastructures.
- TA link: CMS is currently a major user of the TA network link. Activities including the archival of raw data, initial copies of analysis data samples, and user-derived data sets using CERN or other non-US facilities are clear use cases for TA network usage. The raw data transferred to Fermilab alone are expected to average more than 10 GB/second during HL-LHC operations. Today the raw data are a small part of the TA network usage by CMS. CMS tools do not prioritize site proximity (in the networking sense) when scheduling data transfers. Streaming data across the TA link is allowed (even if discouraged). If the current growth rate in TA link use by CMS continues, the size of the TA link becomes a major limitation already by Run 3.
- Streaming reliability: failures in data streaming are a large source of job failures in CMS.
- Streaming to sites beyond the CMS infrastructure: reliable and high-capacity streaming of input data, either raw or pileup simulation, would considerably reduce the disk requirements of CMS at HPC and other non-dedicated computing facilities.
- Disk versus network trade-offs: 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 approximately 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. Cache network needs will be typically bursty because users read full data sets rather than just a file or two from one. This suggests that aggregate network measures are a poor metric in this use case. Understanding caching use cases and needs is part of ongoing R&D.

[234] https://perfsonar.nautilus.optiputer.net/maddash-webui/index.cgi?grid=Nautilus%20Mesh%20-%20Throughput%20100G%20-%20Throughput

5.10.6.10 Outstanding Issues

Traditionally, CMS has treated the network as a free and infinite resource. As the experiment prepares for the "Exabyte per year" era of the HL-LHC, CMS envisions to both make more aggressive use of the network and account and manage usage much more carefully:

- Via R&D activities in caching, there is an expectation to be in a position to trade investments into disk space at Tier 2s against network utilization in the long term. Collaboration with ESnet is welcomed to understand the best uses of caching. At present, the main R&D effort in this area is the production cache deployment across Caltech and UCSD, which includes a cache in Sunnyvale on ESnet hardware. The hardware owned by ESnet and operated by the UCSD Tier 2 team is an integral part of the production cache.
- In response to the regular meetings with ESnet on transfer accounting, CMS has started internal efforts to validate and improve the transfer accounting in the software layers. Long term, CMS would like to be in a position to match network layer accounting with higher-level accounting to gain confidence in understanding network usage. The extent to which this should include traffic tagging and/or flow tagging is unclear at this point.
- CMS would like to explore with ESnet and R&D projects, including SENSE and AutoGOLE, how to transition to managed network usage in production operations.
- Traditionally, CMS has treated the global network of sites as a mesh with identical links when it comes to bulk transfers. The XROOTD data federation was designed from the beginning to be cognizant of the TA link being limited, but treated links within the United States as identical. The data lake model currently discussed in WLCG makes clean regional distinctions. CMS expects that at least the existence of the Atlantic Ocean will become an architectural feature of the data distribution architecture.
- We would like to develop a program of transfer tests both to benchmark the methods at increased capacity and integrate new functionality into CMS methods. CMS would like to do such tests in collaboration with ESnet and FABRIC[235]. More details can be found in Section 5.10.7 and Section 5.10.8.
- We believe that national and international collaboration bringing together researchers, data management experts, and networking experts is important for making better use of network resources as usage levels of research networks increase. In HEP, fora for these collaborations include the WLCG Networking Throughput Working Group[236] or more broadly groups including the Global Network Advancement Group[237].

All of these are discussed further in Section 5.10.7 and Section 5.10.8. CMS believes that an ongoing collaboration between experiment experts and ESnet will facilitate CMS research, allow CMS to migrate to new network capabilities and technologies, and ensure that CMS uses network resources wisely.

5.10.6.11 Case Study Contributors

CMS Operations Representation

- David Lange[238], Princeton University
- Garhan Attebury[239], UNL

[235] https://fabric-testbed.net
[236] https://twiki.cern.ch/twiki/bin/view/LCG/NetworkTransferMetrics
[237] https://www.gna-g.net
[238] David.Lange@cern.ch
[239] garhan.attebury@unl.edu

- Harvey Newman[240], Caltech
- Kenneth Bloom[241], UNL
- Margaret Votava[242], Fermilab
- Tulika Bose[243], University of Wisconsin-Madison
- Lothar Bauerdick[244], Fermilab
- Dan Marlow[245], Princeton University
- Justas Balcas[246], Caltech
- Elizabeth Sexton-Kennedy[247], Fermilab
- David Mason[248], Fermilab
- James Letts[249], UCSD
- Markus Klute[250], Massachusetts Institute of Technology
- Kevin Lannon[251], University of Notre Dame
- Brian Bockelman[252], University of Wisconsin-Madison
- Michael Hildreth[253], University of Notre Dame
- Frank Wuerthwein[254], UCSD
- Oliver Gutsche[255], Fermilab
- Christoph Paus[256], Massachusetts Institute of Technology
- Andrew Melo[257], Vanderbilt University
- Maria Spiropulu[258], Caltech
- Krista Majewski[259], Fermilab

ESnet Site Coordinator Committee Representation

- Phil DeMar[260], Fermilab
- Andrey Bobyshev[261], Fermilab

---

[240] newman@hep.caltech.edu
[241] kenbloom@unl.edu
[242] votava@fnal.gov
[243] Tulika.Bose@cern.ch
[244] bauerdick@fnal.gov
[245] marlow@Princeton.EDU
[246] jbalcas@caltech.edu
[247] sexton@fnal.gov
[248] dmason@fnal.gov
[249] jletts@ucsd.edu
[250] klute@mit.edu
[251] klannon@nd.edu
[252] BBockelman@morgridge.org
[253] mhildret@nd.edu
[254] fkw@ucsd.edu
[255] gutsche@fnal.gov
[256] paus@mit.edu
[257] andrew.m.melo@vanderbilt.edu
[258] smaria@caltech.edu
[259] klarson1@fnal.gov
[260] demar@fnal.gov
[261] bobyshev@fnal.gov

5.10.7 LHC Operations Case Study

5.10.7.1 Background

The experiments running at the LHC are exploring the fundamental structure of matter and the forces that govern matter's interaction. The two main general-purpose collaborations at the LHC are ATLAS and CMS. Both collaborations have thousands of collaborators distributed around the globe who require access to the data generated by the ATLAS and CMS detectors at CERN and to simulated data generated at sites around the world.

The LHC schedule has operating periods (runs) interspersed with maintenance and upgrade periods (long shutdowns). The first run of the LHC was 2009–2012 followed by long shutdown 1 (LS1) from 2012–2014. The second run (Run 2) was from 2014–2018, followed by LS2 from 2019–2021. Run 3 is now scheduled for early 2022. It is important to note that each succeeding run significantly increases the amount and complexity of the data generated, requiring increased storage, network, and compute capacity to successfully exploit.

For this case study (LHC Operations) it is important to understand the activities underway during LS2 and those planned for Run 3.

We chose to dedicate this section to two objectives.

First, a complete list of infrastructure software products and tools that are relevant to data movement and/or access is presented. The list indicates how the various tools relate to the process of science in ATLAS and CMS as described in Section 5.10.5 and Section 5.10.6, as well as Section 5.10.8. The former two focus on the past and present, while the latter focuses on the future.

Second, the CMS understanding of how network use is scaling is presented, contrasting it with what is deemed "affordable." A significant gap between needs projections based on past experience and projections of natural growth based on past investments into networking infrastructure expansion has been identified. This gap is sizable even ignoring the "step-function" increase in data volume per year that the HL-LHC era is expected to bring. In light of this step-function increase, it will be argued that a historical projection approach may significantly underestimate the actual needs. This argument leads to the conclusion to fundamentally rethink the use of networking resources, and to define a process to engage with ESnet in R&D toward substantial improvements in effectiveness of network bandwidth usage. Some aspects of this R&D are described in 5.10.8. The R&D needs to start now to be ready in time for production. It also needs to be structured as a gradual transition of new capabilities into the production infrastructure for the collaborations (CMS and ATLAS) to gain experience and confidence in new services and their interactions with the infrastructure software and processes.

5.10.7.2 Instruments and Facilities

Our tiered, global computing infrastructure and associated sites are described in detail in case studies #10 and #11. For this case study, the relevant timescale is the present through the next ~five years (through the rest of LS2 and Run 3).

ATLAS currently runs event simulation (based on Geant4) on the following DOE and NSF HPC facilities: DOE's ALCF — Theta HPC, DOE's NERSC — Cori HPC and the NSF' TACC — Frontera HPC. The workflow is expected to continue to be used on these machines into Run 3 for as long as these machines last. ATLAS is integrating and will use OLCF — Summit HPC into its distributed computing infrastructure for ML workloads. During Run 3, NERSC's new machine Perlmutter will come online for production usage. ATLAS expects to use both the CPU only partition and the CPU-GPU partition. Also, during Run 3, ALCF's Aurora machine will be used initially for ML workloads.

ATLAS faces a significant challenge in connecting to large-scale resources that are dynamically accessed (e.g., clouds and HPC centers). For the LHC sites, excellent networking and connectivity to LHCONE (or LHCOPN for the Tier 1s) is typical, but for "outside" resources that are opportunistically accessed, the networking may be challenging because:

- We do not control or define the external connectivity or capacity and must use what is available.
- The AUP for LHCONE makes it a challenge to connect resources that are not dedicated for LHC use.

For item 1, efforts have been made to make do with what is available for the resource. For clouds, the bandwidth and connectivity are typically excellent but the challenge (for commercial clouds) is the cost model, which typically makes moving data out of the cloud very expensive. For HPCs the good news is that wide-area connectivity is becoming a more important capability, but there are often significant bottlenecks in capacity and connectivity and each HPC represents its own unique set of challenges. For DOE and NSF HPC centers, the good news is that they are often directly connected to either ESnet or Internet2, both of which serve as R&E backbones in the United States. Even so, there are often significant impedance mismatches in trying to utilize HPC centers at high bandwidth, especially at the Tbps range. This is an area where data challenges will be taken to determine which bottlenecks exist and to see what tools, techniques, and technologies might help. There is much interest in working with ESnet on the definition and execution of these data challenges.

For more details on item 2, please see the details in the response in Section 5.10.7.5 below. The challenge here is that LHCONE is designed to provide a much more friction-free network environment, bypassing firewalls and devices that can adversely affect network performance, yet connecting to LHCONE requires compliance with an AUP that allows advertising networks, which are primarily used by LHC. This is not the case for commercial clouds, nor for DOE or NSF HPCs. To utilize LHCONE for these resources, there would need to be an effort to explore how to dynamically identify "LHC" activities and how to then connect them to LHCONE while those hosts are doing LHC work. This is an area in which WAN network orchestration could be an eventual solution.

5.10.7.3 Remote Science Activities

We need to consider the WLCG use of LHCOPN and LHCONE in light of the existing policies in place to manage those infrastructures. First, a description of each.

The LHCOPN was constructed before LHC startup by the LHC Tier 1 sites to provide the primary path to and from the Tier 0 (CERN). Each Tier 1 (or an entity on its behalf) pays for a dedicated connection between the Tier 1 and the Tier 0. The LHCOPN is dedicated to guaranteeing that sufficient capacity exists to ensure the timely arrival of data from the detectors at CERN to each of the participating Tier 1 sites, which are typically stewards (by MOU) of some fraction of the original raw data. As the LHC program progressed, some Tier 1 sites arranged to provide backup to one another in case their LHCOPN path to CERN failed. Later, there were discussions in the LHCOPN/LHCONE meetings about other uses of LHCOPN, and it was agreed that Tier 1 to Tier 1 traffic could also traverse the LHCOPN. More recently, there have been discussions about other users of the network (other HEP experiments) being allowed to utilize LHCOPN because the alternative was too complex or difficult to maintain. The problem was that some non-LHC experiments (Belle II, DUNE, etc.) shared a number of sites in common with LHC, including Tier 1s. Traffic between Tier 1s for LHC might be configured to use LHCOPN, and that meant that non-LHC traffic would also utilize that path, in violation of the AUP for LHCOPN. It was decided that this kind of leakage was more acceptable than requiring a significant amount of traffic engineering to separately route the non-LHC traffic. The agreement is predicated upon these other users being a small fraction of the LHC traffic on the links, and reviewing traffic use metrics at LHCOPN/LHCONE meetings to ensure that remains the case.

The LHCONE is an overlay network to provide connectivity between LHC sites, especially those not allowed to use the LHCOPN. It was created partially in response to the needs of the sites and experiments involved in the LHC but also because of a corresponding, strong interest from the international R&E networking community. The original design explored both a global layer-2 network and a routed (layer-3) network, with the layer-3 network design using VRF instances being selected. Participation in LHCONE requires approval of the existing LHCONE community, and requests to join are acted upon in the twice per year collaboration meetings. This group was ceded this power by the WLCG management board. Participation in LHCONE also requires

acknowledgement of the AUP[262]. The goal is to make LHCONE usable by only LHC-related resources, but, as noted in the LHCOPN paragraph, this can be challenging. Connectors to LHCONE should allow only LHCONE advertised networks to be routed into LHCONE. Since routing usually depends only upon destination and not source, this means additional technology needs to be used to ensure acceptable use. When a site joins LHCONE, the site needs to define that set of subnets (IPv4 and/or IPv6) that correspond to LHC resources, and only these subnets should have connectivity via LHCONE. The reason this is important is to minimize the exposure for other participants in the LHCONE network, allowing them to choose to bypass firewall and security devices that can adversely affect the network performance. Monitoring and understanding the LHCONE traffic flows is also important. As shown in the AUP, some non-WLCG experiments that have been allowed to join LHCONE. These experiments were primarily allowed to join LHCONE because they had a significant overlap between their resource sites and WLCG's, making it hard for those sites serving both to comply with the AUP. Part of the agreement is that these other experiments should not take a significant amount of the available bandwidth in the LHCONE, and there needs to be detailed monitoring to track this. Currently, there is a lack of 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.

5.10.7.4 Software Infrastructure

In this section, three types of tools are described: data movement, monitoring, and network management services. Tools in the first two categories are in production use now. Tools in the third category are in development and/or ideas for the future.

5.10.7.4.1 Data Management Tools

Within CMS the majority of data movement over the wide-area network is performed with FTS or XROOTD. In the case of FTS, transfers are primarily directed by the PhEDEx and Dynamo management systems and utilize Globus software toolkit, in particular GridFTP. In the case of XROOTD, transfers are primarily the result of intentional remote read requests from workflows or the result of cache misses. The experiments have started a transition from the PhEDEx/Dynamo combination to the Rucio data management tool, and from Globus GridFTP toward HTTPS implemented in XROOTD server software. This transition is expected to be largely completed sometime in 2021. The XROOTD software will be used in two manners, with HTTPS used for top-down data placement as third-party copy operations managed by Rucio via FTS, and the original XROOTD protocol primarily used by application clients.

Data management in ATLAS is orchestrated by Rucio DDM system via FTS. All information about data movement, data replicas, and replication rules is kept in Rucio Oracle database and exported to Elasticsearch for future monitoring and analysis. Globus Online is used (as a lower level to Rucio) for data transfer from HPC centers to US ATLAS sites.

Accounting for less overall volume but of no less importance are tools related to software distribution, calibration and configuration access, and workflow management. CVMFS is the primary method for establishing a uniform runtime environment (including distributed access to software and libraries, etc.) across all processing facilities, while Frontier does the same for uniform calibration database access. Both use caching via the well-known Squid-caching tool. Workflow management is performed with HTCondor, which enables some data movement not included in the previously tools between sites via its file transfer mechanism.

CMS provides its individual physicists access to processing facilities either via the CMS Remote Analysis Builder (CRAB), a CMS tool that supports a set of rich semantics to do data analysis, or directly via HTCondor submission. The latter requires more knowledge from the user but also provides much more flexibility.

[262] The current AUP is at https://twiki.cern.ch/twiki/bin/view/LHCONE/LhcOneAup and the revised version is being developed at https://twiki.cern.ch/twiki/bin/view/LHCONE/NewLhcOneAup.

ATLAS provides its individual physicists access to processing facilities via ProdSys2/PanDA. ProdSys2/PanDA supports all workflows for the experiment, physics groups, and individual physicists. Individual physicists can also use Rucio client tools to download data from grid sites to local resources for physics analysis. Rucio also keeps traces for all users' actions to transfer data between ATLAS and local sites.

The following provides a complete list of tools in use in 2020 by ATLAS and CMS:

- AAA — XROOTD data federation to provide access to CMS data via remote reads and streaming facilitated by a global, regional, and site redirector hierarchy.
- AGIS/CRIC — a grid information system developed in ATLAS and now a community project known as CRIC (Computing Resources Information Catalog). It is used to store and access information about WLCG and external (clouds, HPC) sites. In ATLAS it is also used to keep information about PanDA queues.
- AMI — ATLAS Metadata Interface. It is used to store and access metadata associated with ATLAS (Rucio) data sets and physics event data.
- ASO — asynchronous stage out, a service that stages output via FTS from user analysis jobs at a grid-computing facility to a target remote site specified by the user.
- CRAB — CMS custom tool to support user submission of data analysis jobs. It uses ASO to manage output data.
- CVMFS — https://cernvm.cern.ch/portal/filesystem. The CVMFS is a user-space caching filesystem for scalable software distribution that uses standard HTTP and web servers as the source of data.
- Dynamo — http://t3serv001.mit.edu/~paus/dynamo-documentation. A DMS used by CMS that makes use of PhEDEx/FTS to perform data transfers.
- DAS — https://cmsweb.cern.ch/das. The CMS Data Aggregation Service provides a single interface to multiple CMS data services including Dataset Bookkeeping Service (DBS) and PhEDEx.
- DBS — the CMS Dataset Bookkeeping Service is used to store and access metadata associated with CMS physics event data.
- dCache — https://www.dcache.org. Distributed storage system used as a disk buffer/cache in front of tertiary storage such as tape. dCache is used at Fermilab and BNL as a buffer in front of their Tier 1 tape archive.
- FTS — https://fts.web.cern.ch. Low-level data movement service responsible for transferring data files between sites.
- XROOTD — https://xrootd.slac.stanford.edu. XROOTD as a software product is used in CMS to implement the global data federation. Any CMS member can authenticate against this federation to access any file on disk worldwide either to stream data (i.e., remote file open), or to copy data to local disk. For HL-LHC, this data federation is envisioned to be used to implement the US data lake (see #13). Bulk transfers are expected to be done using HTTP/DAVS, while streaming (= remote file opens) is expected to be done via XROOT protocol.
- XCache — https://www.epj-conferences.org/articles/epjconf/abs/2019/19/epjconf_chep2018_04008/epjconf_chep2018_04008.html. XCache provides a caching service for data federations that serve one or more VOs based on the XROOTD software.

- HTTP/HTTPS — the protocols are used in three ways:
    - CVMFS
    - Frontier (Squid)
    - XROOTD
- HTTP-TPC — third-party copy mode of HTTP used by ASO and FTS. This is the replacement protocol of choice in the United States for GridFTP. The implementation of choice is XROOTD everywhere except the Tier 1. At the Tier 1, both dCache and EOS will support this. See https://twiki.cern.ch/twiki/bin/view/LCG/ThirdPartyCopy.
- iDDS — ATLAS Intelligent Data Delivery Service, PanDA layer. It is used to stage/transfer individual files. All transfers are done via Rucio/FTS.
- Frontier — http://frontier.cern.ch. Distributed database caching system built on top of the Squid caching tool. Used both for small database distribution and by CVMFS for software distribution.
- Rucio — https://rucio.readthedocs.io/en/latest. Data management service developed by the ATLAS experiment and now a community project. Rucio is a primary data management tool in ATLAS, and it keeps all information about data placement, replication, and management. Rucio was destined to replace CMS PhEDEx in November 2020.
- PanDA — ATLAS Workload Management System. PanDA is tracking all tasks and jobs submission in ATLAS for all workflows.
- PhEDEx — https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhEDEx. Database and management agents tracking and maintaining data set distribution across different sites. Triggers and reacts to transfers as necessary to maintain desired state including subscriptions to data sets and replication policies.
- ProdSys2 — the second generation of the ATLAS workflow management system. ProdSys2 communicates with PanDA and Rucio. It triggers data staging from tape and defines Rucio data transfer rules in particular for the ATLAS data carousel project.
- Unified — a tool used for distributing production processing work across sites according to CPU and storage availability and availability of input data sets. It uses Dynamo/PhEDEx/FTS, and in the future Rucio/FTS, for the actual management of data placement.
- HTCondor — https://research.cs.wisc.edu/htcondor. Workload management system that is the primary scheduler used by HEP projects for both user analysis and production workflows. CMS operates a global HTCondor pool for all its activities. Resources in the pool are provisioned via glideinWMS and/or HEPCloud.
- perfSONAR — https://www.perfsonar.net. A network measurement toolkit deployed at a majority of sites involved in LHC experiments.

5.10.7.4.2 Monitoring Tools

As international and highly complex long-running projects, the LHC experiments have numerous monitoring systems in place that have evolved heavily over the years. Network professionals and operators are not necessarily directly involved with the physics activities, and while networks used at the institutions involved may have excellent monitoring of both utilization and quality, these monitoring systems are generally not integrated with the systems covering job workflows, storage, and the movement of data between sites. In recent years, the CMS and ATLAS experiments have put effort toward understanding how these systems may tie together to both understand the usage of the networks and act as input to data movement decision making.

CMS and ATLAS have been engaged with ESnet to better understand network utilization, particularly on TA links, following indications that there would be imminent link saturation during operational periods. It became clear that application-level data from experiment transfer logs did not necessarily line up with actual network utilization records from packet counters and flow monitoring. A seemingly simple question such as "what physics workflow was responsible for this spike in network utilization across the TA links" has proven very difficult to answer with the current systems even in a post-analysis fashion.

Within ATLAS and CMS, the bulk of data movement activities can be attributed to either FTS or XROOTD activities. Both systems have application-level logging of transfers, which is stored within an Elasticsearch database at CERN and accessible via the associated MONIT services described further in this section. In addition, the ATLAS analytics platform at the University of Chicago also gathers and stores transfer data from FTS, XROOTD, and HTCondor, including CMS-related transfers. Transfer logs include a variety of data, including start and end timestamps, bytes transferred or streamed, and metadata, including file or path name and, in some cases, associated reason for the transfer (e.g., purposeful data placement by a scheduler versus an unpredictable user client requesting data). The base logs are enriched further with metadata, such as site tier, site name, associated LHC experiment, and country location. The metadata makes it possible to conduct post-analysis of data movement between site pairs and countries, and even filter for transfers across the TA links, but in practice it was found the data had incomplete or inaccurate metadata and was not representative of reality. In the additional case of XROOTD, an unintentional software bug resulted in a subset of sites having only a fraction of their transfer activity accounted for, an issue discovered only when comparisons were made between the network view via flow data and the XROOTD transfer logs. Further complicating the issue is that application views do not take into account the paths actually traversed for a given transfer (although correlation with perfSONAR traceroute data might enable this), and that the networking views do not take into account the science activity or experiment owner responsible for causing the network use. Matching network events and utilization with the associated science activity causing them is thus very difficult if not impossible with the current systems and tools in place.

ATLAS and CMS are making efforts to correct the application-level deficiencies present today and will always need a source of truth for continual validation of the accounting. Both groups desire to collaborate with ESnet to develop tighter integration between application monitoring and the networking "truth" ESnet is in position to provide. At a primitive level, byte counters on select links with a simple makeup of traffic (e.g., a US CMS Tier 2 transferring to another US CMS Tier 2) can be compared to the aggregate sum of transfer logs between the same sites during the same time period. As has been discovered, these comparisons are approximations at best but in trivial cases are considered good enough to validate application metrics. These comparisons quickly diverge when more complex sites with a mix of experiments and science activities come into play. To understand network usage across experiments at the global scale, more advanced methods will need to be employed.

The challenges identified in operational monitoring of the network have led to the creation of a working group focused on Packet Marking[263]. The goal is to be able to mark network packets by owner and purpose, enabling identification and accounting of traffic anywhere along the network path. This will also enable direct comparisons of, for instance, XROOTD traffic on the wire with the XROOTD transfer logs in MONIT.

The data from perfSONAR, as well as additional network-related data, are being gathered by OSG/WLCG (see **Figure 91** diagram below) and sent to an analytics platform at the University of Chicago. The data are stored in Elasticsearch and publicly accessible via Kibana dashboards.

The collection and availability of the measured data are of increasing importance as next-generation infrastructure is able to consider the network as a constrained and controllable resource. The SAND[264] project coordinates efforts of the SAND team, OSG Networking team (part of IRIS-HEP), and WLCG Throughput Working Group to develop and maintain an archive of measurement data as depicted in **Figure 91.** Metrics are stored long term

[263] https://docs.google.com/document/d/1aAnsujpZnxn3oIUL9JZxcw0ZpoJNVXkHp-Yo5oj-B8U/edit#heading=h.kjs85ae6lo7a
[264] https://sand-ci.org

at UNL (live copy on disk) and Fermilab (tape archiving), and for a shorter term at the University of Chicago. Metrics are also fed into CERN's MONIT system where they can be used in the analysis capabilities of MONIT described previously.

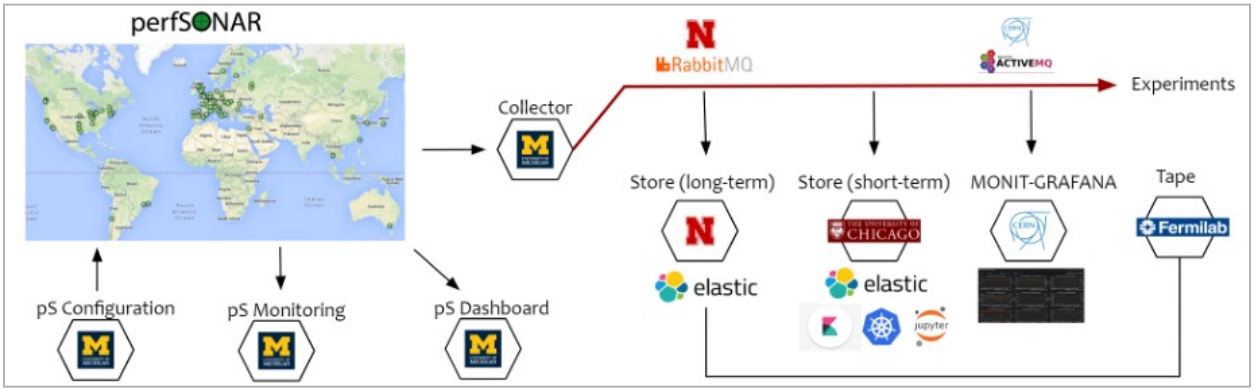


Figure 91: SAND-NMA architecture depicting perfSONAR metric collection and storage

The MONIT[265] system operated by CERN is intended to cover the CERN data center and WLCG project monitoring needs and is positioned to be the central aggregation point of all logging and metric data from the experiments. MONIT is a large system with multiple underlying components for metric acquisition, storage, analysis, and visualization. Metric acquisition is accomplished typically via the Apache Flume[266], ActiveMQ[267], and Kafka[268] frameworks. Storage of metric data is done via a combination of Elasticsearch[269], InfluxDB[270], and Hadoop Distributed File System. Analysis of the stored data can be accomplished with Spark[271] and also via an associated service for web-based analysis known as SWAN[272], which provides Jupyter notebooks with access to data stored in MONIT. Visualization is typically done in the form of dashboards created and presented with Grafana[273], but other experiment-specific and specialized dashboards and views still exist as well. A notable network-related tool feeding into MONIT is perfSONAR, the predominant network measurement toolkit in use by the LHC experiments.

The OSG networking group within IRIS-HEP is presently in the process of establishing a 100 Gbps perfSONAR mesh across US ATLAS and US CMS Tier 1s and Tier 2s. This is an essential next step toward high-capacity network utilization[274].

5.10.7.4.3 Network Management Tools

While the previous sections on data movement and monitoring tools describe tools that are in production use, the next section describes functionality potentially put into production use in the future. As a reference implementation of some of these capabilities, the SENSE software project is referred to.

In the presence of constrained network resources, the intelligent network services provided should allow for the management and best use of the available resources, including the coordination of allocations of network resources with the corresponding computing and storage resources in the context of a set of workflows.

[265] http://monit.web.cern.ch/monit
[266] https://flume.apache.org
[267] http://activemq.apache.org
[268] https://kafka.apache.org/
[269] https://www.elastic.co/elasticsearch
[270] https://www.influxdata.com
[271] https://spark.apache.org
[272] https://swan.web.cern.ch
[273] https://grafana.com
[274] The mesh will soon be visible on the OSG MaDDash instance at https://psmad.opensciencegrid.org/maddash-webui/index.cgi

The basic network services needed should provide the ability to:

- Allocate bandwidth between source and destination with bandwidth guarantees.
- Control the characteristics of a given allocation and an associated transfer, such as immediate versus scheduled, transfer a certain amount of data before a deadline, choose a path between A and B depending on policy or network state and/or performance, etc.

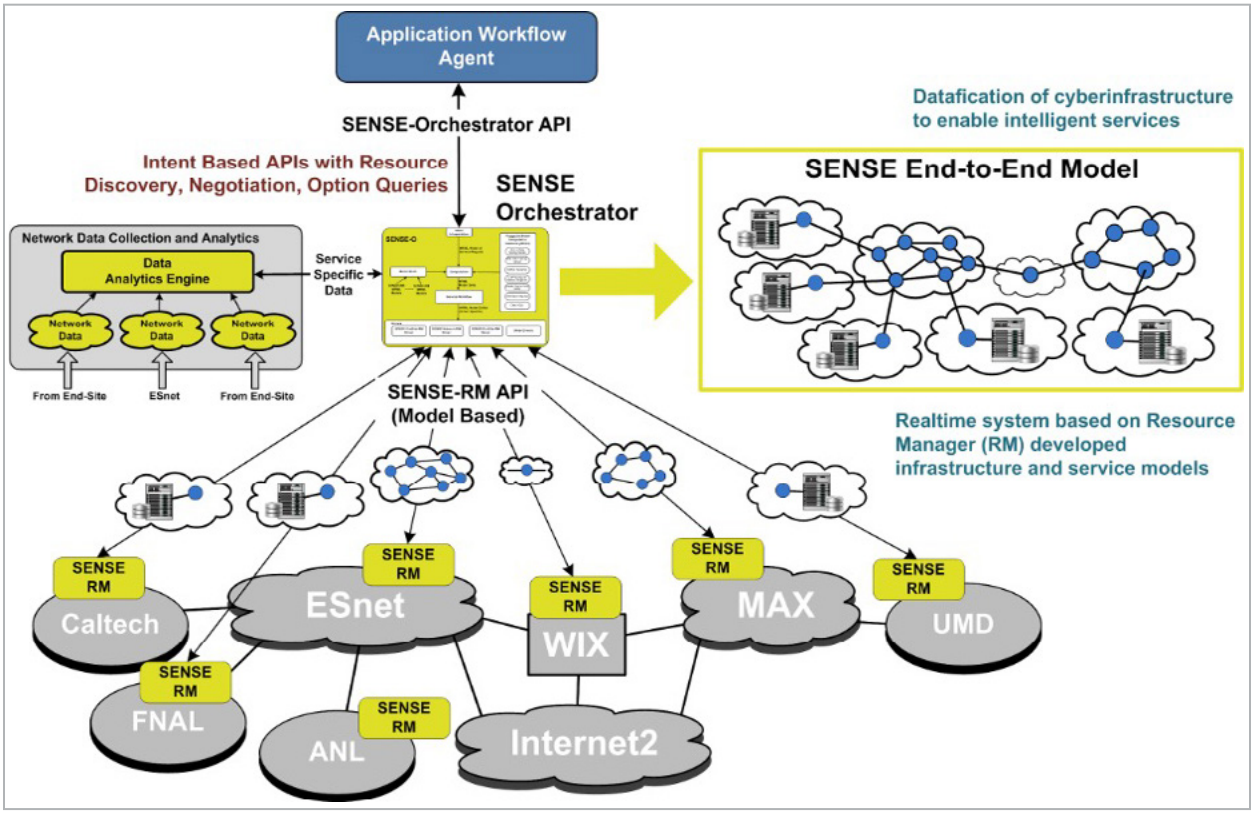


Figure 92: Architecture of SENSE

The 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. In addition, a highly intuitive "intent"-based interface, as defined by the project, allows applications to express their high-level service requirements, and an intelligent, scalable, model-based software orchestrator converts that intent into appropriate network services, configured across multiple types of devices.

With these capabilities, and with development of a sufficient body of network-aware, interactive software by CMS and ATLAS, the experiments will be able to work with ESnet to develop a coordinated workflow system that will manage the network as a first-class, scheduled resource in much the same way as CMS and ATLAS now use computing and storage resources. This, in turn, can enable well-defined and highly tuned complex workflows that require close coupling of resources spread across a vast geographic footprint, in domains including both HEP and other data-intensive sciences.

In preparation for Run 3, US ATLAS and US CMS would like to engage with ESnet and other partners on transitioning some of the SENSE functionality from R&D to production. The experiments would like to identify both appropriate links, and appropriate production-ready functionality, in SENSE, and integrate that into CMS tools for production use. This is an iterative process that scales out over time in both functionality and geographic coverage, with initial targets being chosen together based on a mix of importance, congestion, and convenience.

Links that are neither the most important nor the most congested might be nevertheless appropriate for early transition because of convenience and thus speed with which progress can be made. The process of "starting the transition to managed production networks" is the most important initial step.

We expect that US ATLAS and US CMS, together with ESnet, will define appropriate metrics to measure progress toward the goal of managing all FTS and XROOTD transfers across LHCONE, as well as metrics for success of the individual managed transfers. In addition, there is an expectation that the level of coordination between CPU, storage, and network capacity provisioning and use will increase over time, as the experiments gain operational experience with SENSE and its impact on operations.

To seed these efforts and in response to the requests made by the experiments at the January 2020 LHCONE/LHCOPN meeting at CERN, a Research Networking Technical Working Group has been created[275]. Three broad categories of work are envisioned: Packet Marking, as well as Traffic Shaping and Network Orchestration. Currently, there are more than 83 members from a range of institutions and collaborations, including ATLAS, CMS, and ESnet.

5.10.7.5 Network and Data Architecture

Up to now, ATLAS and CMS have treated the wide-area network as an appliance with almost infinite capacity, the only counterexamples being known poor connections to computing centers in isolated areas. Network capabilities have become more and more an integral part of the computing model, for example in simulation of parasitic pileup collisions, the utilization of data caches, and analysis remote reads through the XCache and AAA CMS XROOTD federation infrastructure. Excellent network interconnection is a prerequisite for data lakes (see Section 5.10.8).

Network bandwidth usage measurements at different points in time over the last several years are used to measure the network bandwidth growth, assuming it is strictly exponential, as past experience indicates. This is done using different types of measurements. The annual growth in network bandwidth used ranges from about 40% to 60% for the various measurements quoted; 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.

This is considered the "steady state" growth rate that is expected through the end of Run 3 (2024). Between Run 3 and Run 4, ATLAS and CMS expect a one-time step-function increase of data volumes in combination with significant changes in computing model necessitated by this large increase in data volumes. This step-function increase is discussed in some detail in case studies #11 and #13. The annual data volume for a single reconstruction version of CMS data and simulations increases at this step function from about 22 PB to 634 PB. For detailed explanations on what drives this increase, see case study #11. That study also shows the corresponding step-function increases in compute and storage needs. At present, no reliable estimates of the corresponding increase in network needs exist. In the absence of such an estimate, the best that can be done is to explore the historical growth in network use, and extrapolate into the future, understanding that this might be an underestimate.

Based on historical measurement of network provisioning costs, it can be assumed that price drops approximately 15% on a yearly basis. The mismatch between 40 to 60% annual growth in use and only 15% annual drop in price leads us to the conclusion that there must be effort invested into R&D toward managing the network use to contain growth.

5.10.7.5.1 Estimating Growth from WLCG Dashboard Data

The typical transfer rates of the last year among all the WLCG sites are shown in **Figure 93** a snapshot taken from the WLCG dashboard, which covers a 12-month period starting in August 2019 along with an indication in the lower left of the traffic level near the start of Run 2 in March 2015 (6 GB/sec). The increase in traffic over the course of Run 2, a factor of 8 in 53 months, corresponds to an annual growth rate of 60%.

[275] https://docs.google.com/document/d/1l4U5dpH556kCnoIHzyRpBl74IPc0gpgAG3VPUp98lo0/edit#heading=h.jc3es9koaa99

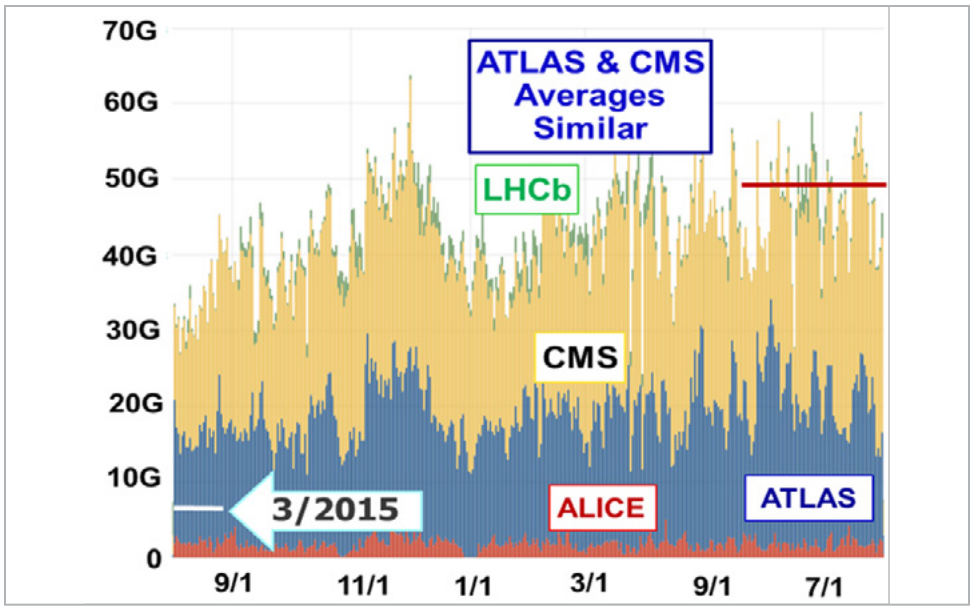


Figure 93: Daily average transfer rates in GB/sec among all the WLCG sites

Figure 93 shows the daily average transfer rates in GB/sec among all the WLCG sites, with the traffic of each of the major LHC experiments marked. The long-term average during the summer of 2019 of 48 GB/sec and the daily average peak of 65 GB/sec during a CMS processing campaign at the end of 2019 are lower than the short-term averages. The arrow and white line indicate the earlier average rate of 6 GB/sec in March 2015 (which is not part of the plot itself).

5.10.7.5.2 Estimating Growth from ESnet and HEP Network Traffic

Figure 94 shows that the monthly ESnet traffic volume reached 100 PB/month (equivalent to an average throughput of 330 Gbps) in early 2020 (before the coronavirus pandemic), and a relatively steady exponential growth rate of 45% per year or a factor of two every two years on average, since 2008. The traffic over LHCONE, which represents the largest component of ESnet traffic, has been growing at an annual rate of 60 to 70% over the last five years. **Figure 95** shows an August 3, 2019, snapshot of ESnet traffic with an average of roughly 300 Gbps and 450 Gbps peaks.

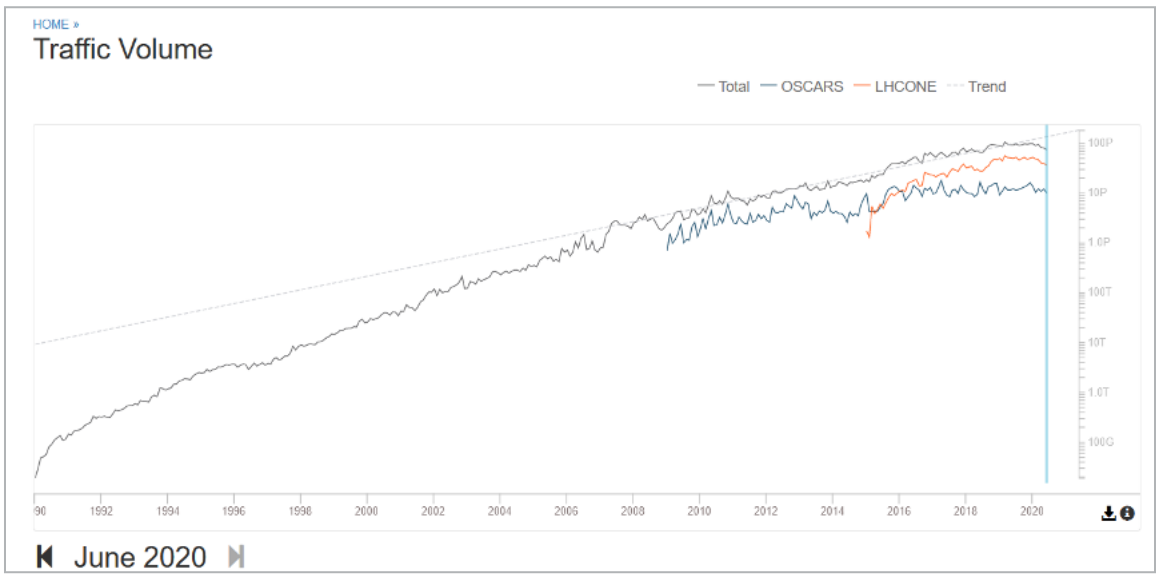


Figure 94: ESnet traffic volume over time

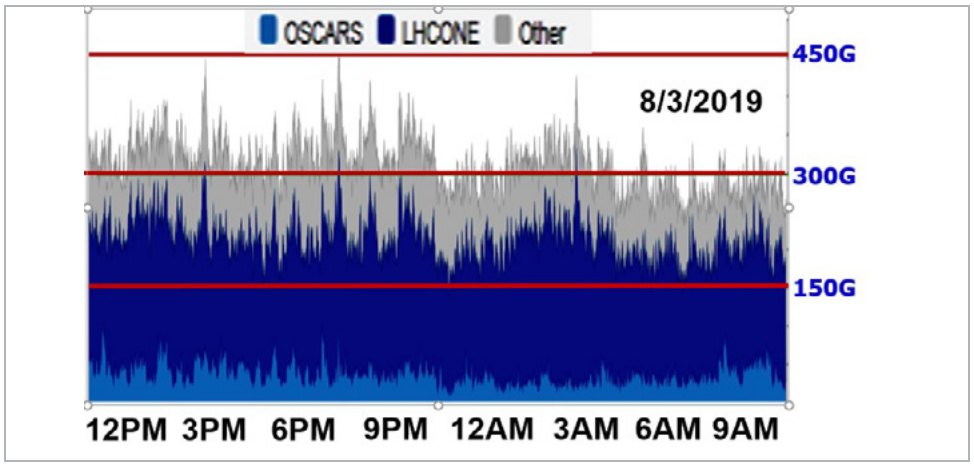


Figure 95: ESnet traffic (August 2019)

The growth of LHCONE traffic in Europe is illustrated in **Figure 96,** showing traffic growth from August 2015 to January 2020 (54 months). The growth factor of 9x in sustained daily peaks corresponds to an annual growth rate of 63%.

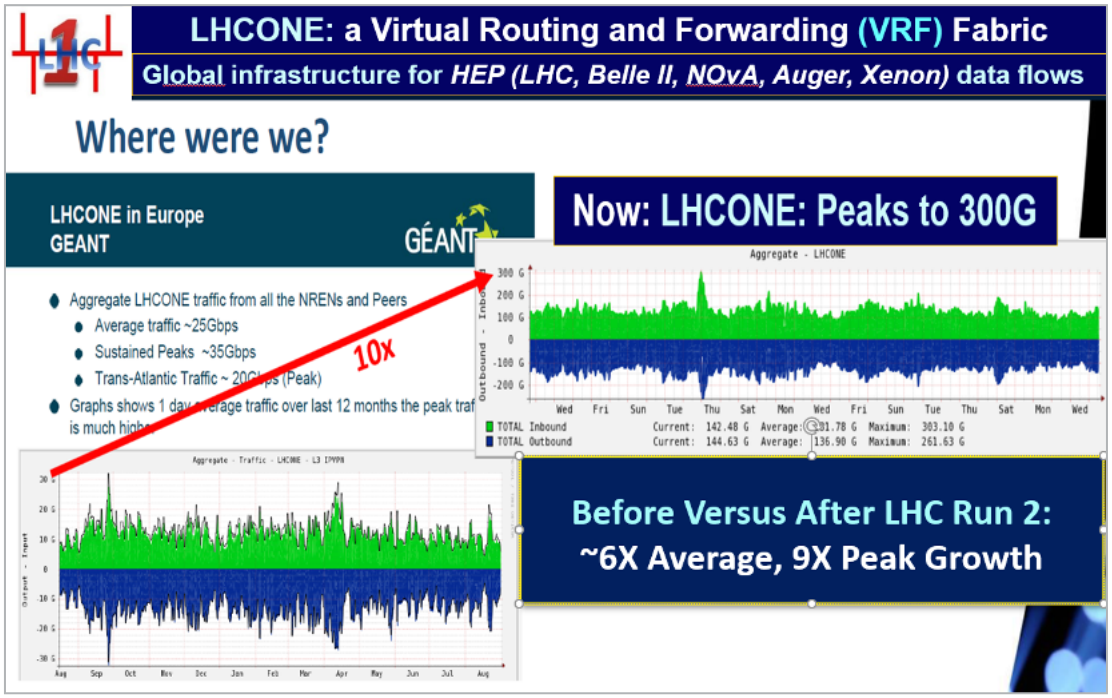


Figure 96: LHCONE traffic comparison

5.10.7.5.3 ESnet TA Traffic

Considering transfers, remote reads for analysis, and pileup mixing, it is likely that HL-LHC computing requires 1 Tbps links (an estimate justified in some detail in case study #13) 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. Measurements of TA links usage show an exponential growth over long periods of time.

In addition, a large difference between peak and average use can be observed also because of the congestion caused by experiments not managing their network use. An extrapolation of this model leads to a possibly critical situation for peak traffic already during Run 3, as illustrated in **Figure 97.**

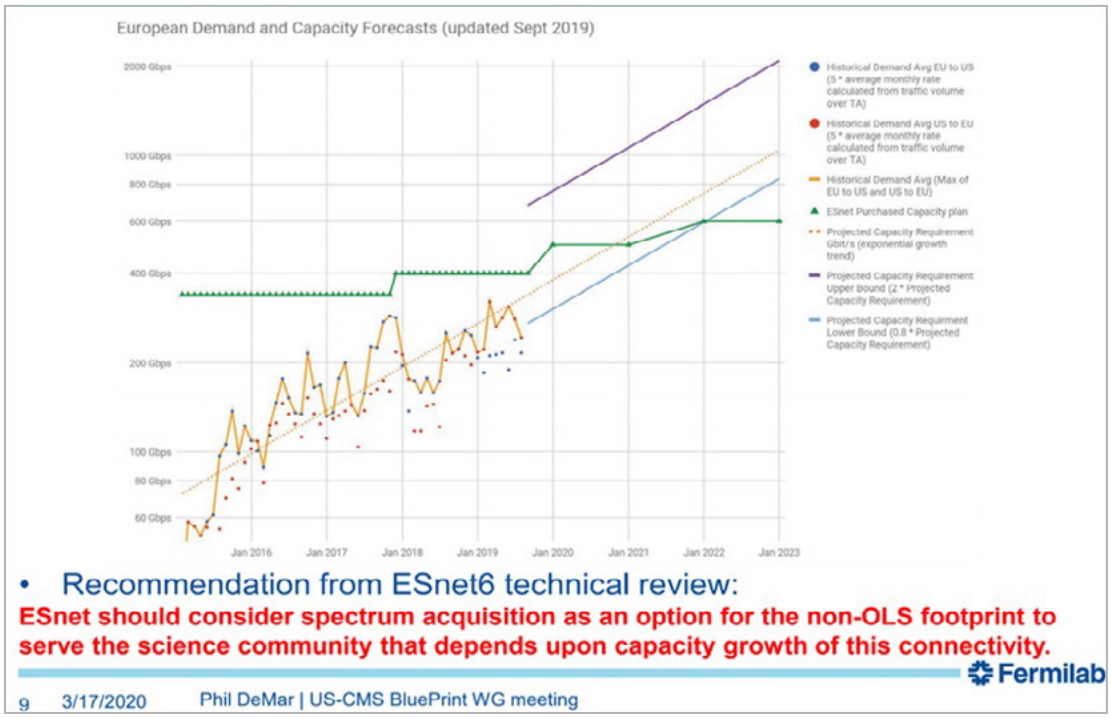


Figure 97: TA network traffic versus time

Figure 97 shows that the projected capacity requirement in ESnet (based on five times the monthly average traffic rate to accommodate short-term peaks, with an uncertainty band of -20% to +100% represented by the blue and purple lines in the figure) is likely to exceed the affordable capacity already by the start of LHC Run 3 in 2022. The growth rate in the figure is approximately a factor of two every two years, or 41% per year, which closely follows the overall growth rate of ESnet traffic. This means that the projected capacity requirement will grow by a factor of 16 between now and mid-2028 if there are no changes to the LHC computing models and operations within the model.

Most of the traffic, about 60% as of this writing, comes from LHCONE, which also has had the largest growth rate. The demand is expressed as five times the average monthly rate to consider the bursty nature of the traffic (peak utilization equal two times the average), long TA provisioning times, and sudden increases of LHC traffic as well as possible low availability experienced by ESnet's links.

5.10.7.5.4 Estimating Growth in Capacity at Constant Cost

Figure 98, taken from Telegeography's Global Bandwidth Research Service Executive Summary[276], using the cost of a 100G link between New York and London as a reference, shows that the 2016 to 2020 compound annual growth rate has been -15%. This corresponds to a 3.7x increase in the affordable capacity within a constant budget by mid-2028, which implies a shortfall in the affordable capacity of approximately a factor of ~4. It must be stressed again that this shortfall does not take into account the step function in annual data volume between Run 3 and the HL-LHC, but is based solely on extrapolations from past use. It is thus likely to be an underestimate.

[276] https://www.dropbox.com/s/jku7pylofbqyffp/GlobalBandwidthResearchService_ExecutiveSummary_TelegeographyJune2020.pdf?dl=0

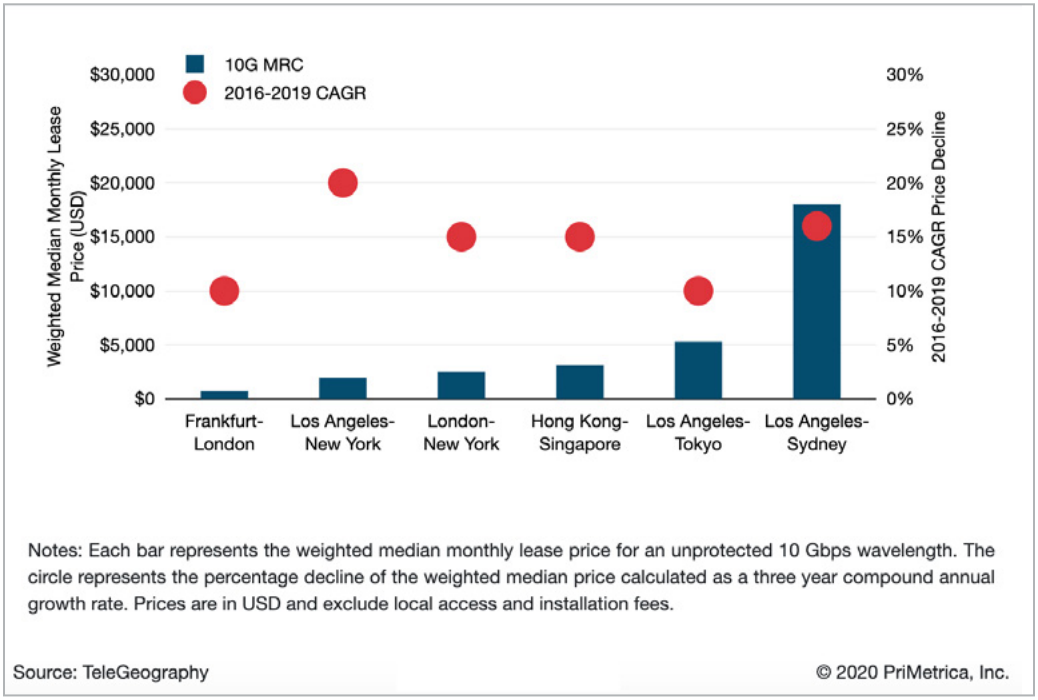


Figure 98: Percentage decline in weighted median price for international route leases

This leads to the conclusion that substantial changes in the computing models of the LHC experiments, and as a consequence, a joint R&D program among ESnet and the other major network providers (Internet2, SURFnet, NORDUnet, et al.), together with the experiments is needed.

5.10.7.6 Cloud Services

We have evaluated the use of commercial cloud both for processing and for TA transfer. In general, we found that both are not cost-effective, at present. The experiments have shown that capability exists of making large-scale use of cloud resources if the cost structure were to change, but if cloud services are used, routing of traffic to and from those services must be done carefully, because varying ingress and egress points can have significant cost implications[277].

5.10.7.7 Outstanding Issues

Organizations including the US ATLAS/CMS operations programs, IRIS-HEP, and the WLCG DOMA working groups are organizing data challenges to build up to the scale projected for HL-LHC. One challenge being developed is to transfer and process 10 PB (e.g., 1 Tbps for a day) of data to an HPC center and back in a day, corresponding to the scale of data movement needed to process one year of HL-LHC data within 100 days. This brings a number of technical challenges, including tape recall, managing disk buffers, and managing network usage. To succeed with the networking component, the data management stack will need to learn how to tag network traffic, to use tools such as SENSE to schedule networks, and to co-schedule networks together with disk and compute resources. A program is being established between ATLAS, CMS, and WLCG to build up tools and processes that enable a progressive set of demonstrations up to the scale of 10 PB in a day between today and the HL-LHC era. It is very likely that the experiments will arrive at a detailed set of data challenges across multiple use cases within the next year or two. Such a program of work would then be executed and refined over the next five years or so. The experiments would very much like to establish a forum with ESnet to discuss progress on these data challenges as they get defined and executed.

[277] As reference for a presentation on using cloud transfer capabilities that includes benchmarking as well as cost information, refer to https://indico.cern.ch/event/923131.

5.10.7.8 Case Study Contributors

LHC Operations Representation

- David Lange[278], Princeton University
- Garhan Attebury[279], UNL
- Harvey Newman[280], Caltech
- Kenneth Bloom[281], UNL
- Alexei Klimentov[282], BNL
- Robert Gardner[283], University of Chicago
- Shawn Mckee[284], University of Michigan
- Margaret Votava[285], Fermilab
- Tulika Bose[286], University of Wisconsin-Madison
- Lothar Bauerdick[287], Fermilab
- Dan Marlow[288], Princeton University
- Justas Balcas[289], Caltech
- Elizabeth Sexton-Kennedy[290], Fermilab
- David Mason[291], Fermilab
- James Letts[292], UCSD
- Markus Klute[293], Massachusetts Institute of Technology
- Kevin Lannon[294], University of Notre Dame
- Brian Bockelman[295], University of Wisconsin-Madison
- Michael Hildreth[296], University of Notre Dame
- Frank Wuerthwein[297], UCSD
- Oliver Gutsche[298], Fermilab

---

[278] David.Lange@cern.ch
[279] garhan.attebury@unl.edu
[280] newman@hep.caltech.edu
[281] kenbloom@unl.edu
[282] aak@bnl.gov
[283] rwg@uchicago.edu
[284] smckee@umich.edu
[285] votava@fnal.gov
[286] Tulika.Bose@cern.ch
[287] bauerdick@fnal.gov
[288] marlow@Princeton.EDU
[289] jbalcas@caltech.edu
[290] sexton@fnal.gov
[291] dmason@fnal.gov
[292] jletts@ucsd.edu
[293] klute@mit.edu
[294] klannon@nd.edu
[295] BBockelman@morgridge.org
[296] mhildret@nd.edu
[297] fkw@ucsd.edu
[298] gutsche@fnal.gov

- Christoph Paus[299], Massachusetts Institute of Technology
- Andrew Melo[300], Vanderbilt University
- Maria Spiropulu[301], Caltech
- Paolo Calafiura[302], LBNL
- Kaushik De[303], University of Texas at Arlington
- Joseph Boudreau[304], University of Pittsburgh
- Johannes Elmsheuser[305], BNL
- Edoardo Martelli[306], CERN
- Simone Campana[307], CERN
- Wei Yang[308], SLAC

ESnet Site Coordinator Committee Representation

- Phil DeMar[309], Fermilab
- Andrey Bobyshev[310], Fermilab
- Vincent Bonafede[311], BNL
- Mark Lukasczyk[312], BNL

5.10.8 HL Era of the LHC Case Study

5.10.8.1 Background

The LHC at the European Laboratory for Particle Physics 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 LHC and its associated experiments will undergo a major upgrade in the next six years, leading to HL-LHC operations around 2027.

As shown in **Figure 99,** the HL-LHC era starting in 2028 will accumulate roughly the same integrated luminosity of data in three years of LHC running as the entire period of running of the LHC has produced up to then. All of that data will be accumulated at a center-of-mass energy of 14 TeV. This implies that the science capabilities of the first years of HL-LHC data taking at design luminosity, in 2028–2030, are expected to be roughly equivalent to the data taken from 2010–2024, or runs 1, 2, and 3 combined.

[299] paus@mit.edu
[300] andrew.m.melo@vanderbilt.edu
[301] smaria@caltech.edu
[302] pcalafiura@lbl.gov
[303] kaushik@uta.edu
[304] boudreau@pitt.edu
[305] johannes.elmsheuser@cern.ch
[306] edoardo.martelli@cern.ch
[307] Simone.Campana@cern.ch
[308] yangw@slac.stanford.edu
[309] demar@fnal.gov
[310] bobyshev@fnal.gov
[311] bonafede@bnl.gov
[312] mlukasczyk@bnl.gov

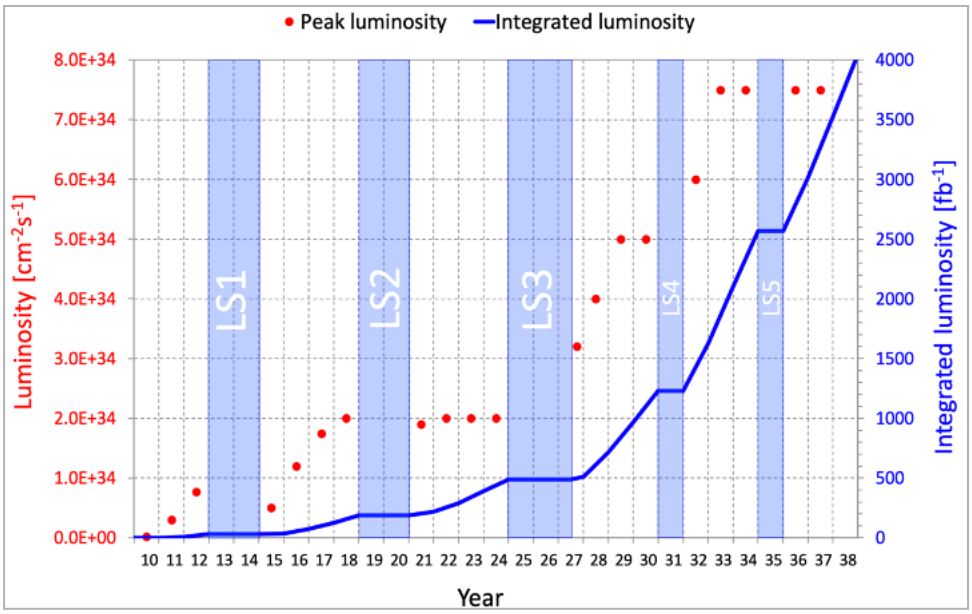


Figure 99: Expected maximum luminosity and integrated luminosity for the LHC as a function of calendar year

The entire HL-LHC era will last for 10 years of data taking, with 12–24-month maintenance periods interspersed roughly every three years[313]. Each data-taking year, the experiments, ATLAS and CMS combined, are expected to accumulate roughly 1 EB of new data. See **Table 26** for more detailed numbers.

5.10.8.1.1 ATLAS

ATLAS is a particle detector that measures and records the particle collisions at the LHC. The primary scientific goal is to quantitatively measure and discover properties of the SM of particle physics. A major emphasis of the HL-LHC will be precision measurement of the properties of the newly discovered Higgs boson.

The SM of particle physics is considered to be an incomplete theory. In order to explain observations and measurements like dark matter and the Higgs mass fine-tuning problem, new phenomenologies remain to be discovered experimentally. ATLAS physics goals combine a strong program of SM measurements with search for new phenomena, like Supersymmetry.

The LHC collides protons more than a billion times every second, out of which ATLAS will select interesting collisions for recording at a rate of ten thousand times every second. With over two thousand hours of data collection every year when the HL-LHC starts running in 2027, ATLAS 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.

5.10.8.1.2 CMS

The CMS computing model is described in Section 5.10.6, which includes definitions of the event data tiers referenced here as well as an overview of the tiered computing structure and the capabilities and uses of each tier. Both experiments make the same assumptions around how often data must be processed and reprocessed over the course of a year. To understand totals as well as differences, it is thus sufficient to just look at the current expectations for the first year of full HL-LHC luminosity, in 2028 or 2029. 5.10.6 includes figures that overlay the CPU, disk, and tape needs versus time computed from the current assumptions. The annual data volumes are summarized in **Table 26.**

[313] https://lhc-commissioning.web.cern.ch/schedule/LHC-long-term.htm

CMS name	CMS annual volume	ATLAS name
RAW	364 PB	Raw
AOD	240 PB	AOD
MINI	30 PB	DAOD
NANO	0.24 PB	DAOD_PHYSLITE

Table 26: Corresponding annual data volumes for CMS. The "raw" data are from the detector only while all others include detector data and simulations.

Differences in some of the parameters between the two experiments result from different choices of analysis procedures and are largely driven by sociology.

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 lakes model. **Figure 100** shows this model as a graphic with two lakes. Clearly, many more lakes will be needed to meet the needs of processing and analysis across all of the different regions worldwide.

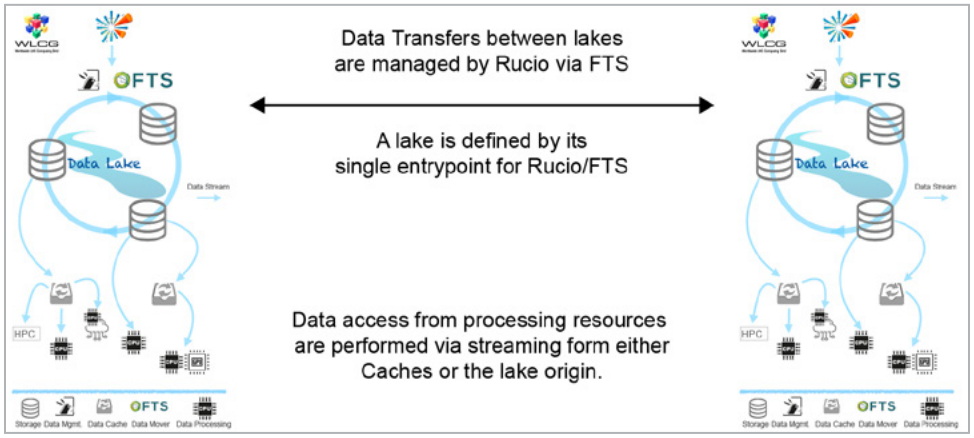


Figure 100: WLCG data lakes model

A lake is defined as a cluster of computing facilities that have a single entry point for interactions outside of the lake. Behind each entry point are generally multiple storage endpoints that are geographically distributed. These storage endpoints will be referred to as "data origins." An entry point may support multiple levels of QoS, depending on the cluster. QoS differences may include data access performance, latency, levels of protections against data loss, etc. The exact definitions of the QoS categories and their expected implementations are still a subject of R&D and discussions. The current structure shown here and the technical details mentioned later can be considered preliminary but provide concrete examples of how data lakes can function.

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. Access to data from processing elements inside the lake is mediated either via caches (implemented via XCache, a subset of XROOTD software) or streaming directly from one of the data lake origins. There is no direct access to data from processing elements across data lake boundaries.

US CMS currently assumes that all disk storage at the Tier 1 and Tier 2s will be part of a single US data lake. The tape archive at Fermilab may be part of that lake, as a separate QoS, or managed as a separate lake. Tier 3s may connect into a lake either via a cache or streaming from a cache. There will be no Tier 3 origins as far as the central management of operations is concerned. Allocations at HPC centers funded by the DOE or NSF, as well as cloud resources, will be treated in the same fashion as T3s, conceptually. Output from processing workflows

will be ingested into the data lake via Rucio/FTS, with the exception of user data produced at AFs. There will be the option to export data from an AF to work outside the data lakes model. Laptops and other personal computing devices are not Tier 3s. There will most likely be an option for accessing data within the lake from a laptop, but the level of access provided this way is negligible compared with all other accesses. This is no different than today.

All of these facility concepts are described in more detail in Section 5.10.8.3.2. Processing workflows and the science process in general are described in Section 5.10.8.4.2. Tools and software used are described in Section 5.10.8.6.2. Section 5.10.8.7.2 provides a summary of all networking requirements across facilities, and describes the necessary networking R&D that CMS would like to engage ESnet in as the infrastructure for the HL-LHC is developed.

5.10.8.2 Collaborators

The collaboration space for both experiments is not expected to radically change between now and the HL-LHC era, or beyond. The basic premise of the experiment (e.g., widely distributed computational and storage infrastructure) that spans countries and continents will remain in place for the foreseeable future.

5.10.8.2.1 ATLAS

The current members of the ATLAS collaboration are described in Section 5.10.5. It is not expected that major changes to the membership will occur for the HL-LHC. The collaborators will continue to be distributed worldwide, which puts strong requirements on global networking. Specifically, it is expected that computing resources will continue to be globally distributed, with no centralized data or computing location. Physicists will need access to the distributed resources from all corners of the world.

User/collaborator and location	Primary or secondary copy of the data	Data access method	Avg. size of data set	Frequency of data transfer or download	Are data sent back to the source and method	Any known issues with data sharing
ATLAS, WORLDWIDE (FOR A BREAK-DOWN OF ALL CURRENT US SITES, SEE SECTION 5.10.6)	Both	All	Range: 10 GB to 50 TB per data set	Continuous	Yes, fully distrib-uted computing	Debugging network transfer problems

Table 27: ATLAS HL-LHC data projections

Since the information by site is available in Section 5.10.5, **Table 27** is not filled in for the 30+ US ATLAS institutions.

The exact computing model for the HL-LHC has not been finalized, given that the LHC Run is not expected to start until 2027. It is assumed that the current ATLAS computing model will be the baseline model, with some small improvements. There will be no fixed hierarchy of computing sites for most data processing and data access services. In order to improve usage efficiency, sites will be primarily categorized by size, service level, and capability. Hence the location of data sets and users is not deterministic: ATLAS expects worldwide distribution of all resources and users. ATLAS also expects about seven large sites in the United States, which will all be required to have a full range of distributed computing capabilities. They will store both primary and secondary data, will provide access to hundreds of users, and will participate in continuous data transfers. These sites will include the BNL Tier 1, the current Tier 2s (Great Lakes, Midwest, Northeast, and Southwest), SLAC, and a few HPCs. It is expected that a few hundred Gbps links will be needed between them when the HL-LHC starts. BNL Tier 1 will need additional capacity to handle worldwide traffic. The network capacity should be provisioned to match the scale of the available resources at each site.

5.10.8.2.2 CMS

The international nature of CMS, location of its facilities, and location of any shared HPC centers is essentially unchanged over time. Refer to Section 5.10.6 for detail

5.10.8.3 Instruments and Facilities

The HL-LHC refers to the upgraded LHC configuration that will be implemented during Long Shutdown 3 (LS3). Luminosity is a measure of the number of collisions per unit area for a given period of time. More luminosity means more particle collisions and more opportunities to measure new, interesting physics. A good way to measure discovery potential is by integrated luminosity, which is measured in inverse femtobarns (fb^{-1}). An inverse femtobarn equates to 100 trillion collisions. The integrated luminosity for all of LHC runs 1 and 2 was about 150 inverse femtobarns of data. The HL-LHC will produce more than 250 inverse femtobarns of data per year and will be capable of collecting up to 4,000 inverse femtobarns.

One of the significant challenges posed by HL-LHC is the increase in the average number of particle collisions each time particle bunches cross each other (which happens approximately 40 million times per second). The Run 2 average number of collisions per crossing is approximately 40, but HL-LHC will increase that to 150–200. It is important to note that the impact of this increase in processing power is exponential and not linear, primarily because of the increased combinatorics involved in assigning particles to collisions.

To upgrade to HL-LHC will involve numerous changes to the accelerator. The LHC particle beam will need to be more intense and more focused than at present. New or upgraded components will need to be installed in various sections of the LHC's 27-kilometer ring. In addition to the accelerator upgrades, the ATLAS and CMS experiments also need to implement significant upgrades and detector replacements to handle the much more challenging environment of the HL-LHC.

5.10.8.3.1 ATLAS

The impact of HL-LHC on ATLAS storage and compute resources is significant. As previously noted, 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 HPC will also have an impact on the compute resources (storage and networking). These HPC centers are increasing in computing power, and several exaflop 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 ATLAS data centers for subsequent processing. Additionally, these machines are being designed to be very efficient at AI/ML. ML training requires large amounts of data transferred into the HPCs for use during the training.

In the HL-LHC Computing Design Report (CDR), ATLAS has chosen to evaluate two scenarios based on how aggressively it is anticipated that the HL-LHC S&C R&D program will deliver improvements. They are useful strawmen to discuss model uncertainties and development costs. These scenarios are measured against an essentially "do no R&D" baseline scenario, not a tenable strategy in itself but a benchmark against which to measure the expected return on R&D investments.

These three scenarios, defined in terms of the R&D program activities described in the next section, are as follows:

- **Baseline:** ATLAS implements the new data formats foreseen by the Run 3 analysis model, the multi-threaded software framework AthenaMT, and updates to the tracking code, but otherwise continues in largely the same way as in Run 2. In particular, the CPU time per event for event generation, detector simulation, and reconstruction are assumed to remain at the level currently achieved by applying the current software to the Phase-II detector simulation, and the mixture of generators and simulation remains the same. This is not a tenable scenario and is provided only as a baseline against which to measure the other two.

- **Conservative R&D:** The research and development activities currently underway for Run 3 are assumed to be successful, including the data carousel, fast track reconstruction, lossy data compression, and most of the detector simulation being performed with fast simulation.
- **Aggressive R&D:** ATLAS implements new developments that very significantly improve the speed and storage volumes of workflows that currently are heavy consumers of resources. For example, these could include porting of high-precision generators to GPUs, sharing events with CMS, or speeding up the full simulation either by software efficiencies or porting parts of the code to GPUs. Almost universal adoption by the physics groups of DAOD_PHYSLITE and development of very high-quality fast simulation that could replace full simulation in almost all cases would also fall into this category. Some R&D activities in the aggressive category cannot yet be quantified in their impact and so are not yet included in the model. An example is ML models for fast simulation and reconstruction.

Figure 101 shows the estimated CPU and storage needs under these different scenarios, to be compared with provisioned resource curves based on flat funding for CPU and storage, and capacity growth per unit cost of 10% and 20% per year. US ATLAS in its current best-estimate budget planning assumes 10% growth (the lower curve) for the years out to HL-LHC, based on an assessment of the technology landscape. Other estimates within the HEP community range from 5% to 20%. These studies demonstrate that ATLAS (and US ATLAS) must pursue vigorously aggressive R&D options to be able to meet the HL-LHC challenges in a sustained budget model.

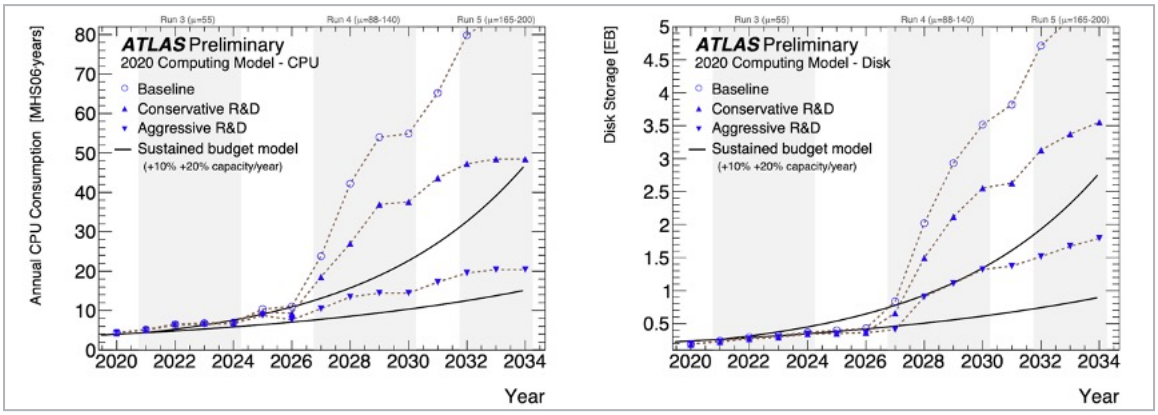


Figure 101: Estimated CPU and disk resources needed for the years 2020 to 2034 under the different scenarios. The solid lines indicate annual improvements of 10% and 20% in the capacity of new hardware for a given cost, assuming a sustained level of annual investment. The US resources are typically 23% of the resources shown here.

5.10.8.3.2 CMS

In the HL-LHC era, it is expected that there will be Tier 1 and Tier 2 facilities fully funded via DOE and NSF, respectively. They are run as coordinated facilities via the US CMS Operations Program for central processing, MC generation, and, at a smaller scale, analysis. CERN will operate the Tier 0, a tape archive, and a Tier 2 for CERN-resident physicists. In addition, it is expected that US CMS will make use of any HPC centers funded by either DOE or NSF that will grant us allocations. CMS expects to have the ability to make use of cloud resources, but at present, budgets to do so are not present, given the current business models of cloud providers. It is unclear whether or not such budgets will exist in the future.

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 (other than the archival tape space at the T1) will be used to support end-user analysis workflows. Both types of workflows have significant data flows, and thus an impact on the networks. More details are given in Section 5.10.8.4.2.

There will be additional resources at universities and elsewhere that participants in the LHC science program will want to integrate into the computation and data federations. These are called Tier 3 resources, or T3. The primary distinction of a T3 resource is that it is owned by somebody other than the US CMS Operations Program, and thus may be available only for special purposes, and to a limited number of people. For example, faculty members may use startup or retention funds to procure resources supporting their own research at the LHC. T3 resources may show up in many different forms, e.g., on-premise hardware aggregated into clusters, allocations on campus owned and/or operated clusters, commercial cloud resources, etc. Together with IRIS-HEP and the OSG, the US CMS Operations Program will support integration of these resources into the global CMS compute and data federations. It is expected that the more permanent of these resources may want to be connected to LHCONE, while the more dynamic will assume peering exists between LHCONE, Internet2, and the regional networks represented in Quilt. Overall, it is assumed that the network bandwidth required for T3 resources is modest compared with a typical T2.

As a rough guide as to the network bandwidth that facilities in US CMS will require in the HL-LHC era, the following estimates were performed:

- In 2020, each US CMS Tier 2 was expected to provide 120 kHS06 of compute capacity. This is roughly equivalent to 8–12k x86 hyperthreads.
- The measured monthly average of IO per hyperthread on US CMS Tier 2s is roughly 1.5 Mbps, with spikes up to x10 larger. LAN capacity will be provisioned to support spikes rather than the long-term average. US CMS Tier 2s thus have either already transitioned, or are presently transitioning, from 1–10 Gbps network connections to their worker nodes.
    - By 2028, it is expected that worker nodes will provide 500–1,000 hyperthreads, requiring 1–2 10Gbps network interface ports or a 25 Gbps port for their LAN connections.
- 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 shared (set of) 100 Gbps link(s), shared with the entire institution. In most cases today, though, the US CMS LHC program dominates the WAN link use at the Tier 2 institutions. This may change in the future.
    - We expect that Tier 2s that are part of the data lake origin will be required to provide guaranteed, managed, and possibly scheduled burst capacity at up to 400 Gbps to support large-scale ingests over the course of hours to a day.
        - As a reminder, 400 Gbps for a day is roughly 4 PB of data. US CMS Tier 2s each provided 5 PB of usable storage in 2020. By 2028, 4 PB of storage is likely to be a minor fraction of the origin storage of a Tier 2 that provides origin storage to the US data lake.
        - While in 2020 all US CMS Tier 2s provided roughly the same functions and capacity, this may no longer be the case for the HL-LHC era. It is thus conceivable that not all Tier 2s in US CMS will provide origin storage for the US CMS data lake during the HL-LHC era.
- For planning purposes, a 15% increase per year in Tier 2 processing capacity is assumed. Over the eight years from 2020 to 2028, this results in a factor of three increase in capacity. Simple scaling would result in 360 kHS06 by 2028, and a 54 Gbps LAN average aggregate bandwidth with spikes of up to 540 Gbps entirely due to processing, and an additional 400 Gbps burst capability to storage if the Tier 2 contributes storage to the data lake origin.

This simplistic extrapolation from Run2 to HL-LHC has very substantial uncertainties in both directions as can be seen from the following:

- The CPU time per event for reconstruction is expected to increase from 250 HS06 per event experienced during Run 2 to 5,000 HS06 per event at the HL-LHC, while at the same time the raw event size is expected to grow from 0.9MB to 6.5MB. The IO rate per HS06 thus is expected to drop by a factor of three for reconstruction.
- Typical analyses on Run 2 data process events in MiniAOD form at rates on the order of 10Hz, while typical analyses of using NanoAOD at HL-LHC are expected to proceed at rates of 1 to several kHz, both per hyperthread. The ratio in event sizes of MiniAOD today to NanoAOD during HL-LHC is 35kB/2kB. There is thus a possible order of magnitude increase in IO per hyperthread expected at HL-LHC, by the time most analyses use NanoAODs, as compared with Run 2, where most analyses have used MiniAOD.
  - See also the AF discussion later in this case study.

There are additional larger uncertainties for WAN IO needs at Tier 2 centers as the experiments will want to trade off investments in disk space in their caching infrastructure against network bandwidth use as follows:

- At the SoCal production cache across Caltech and UCSD, the working set per day, week, and month has been measured with results ranging from a few tens of TB (daily) to a few hundreds of TB (monthly).
- A cache designed to only hold the daily working set can thus be an order of magnitude smaller than a cache designed to hold a month's working set.
- In principle, those measurements can be used to estimate the typical network bandwidth use for different cache sizes and cache-refresh rates. Using the difference in event sizes and data volumes, predictions could be made for network bandwidth needs versus cache size for the HL-LHC era. This work has not been done yet.
- Typical WAN transfers seen in 2020 on the SoCal cache were 10 Gbps with occasional spikes to 25 Gbps. The cache size in 2020 was roughly a petabyte, sufficient to host one-third of the Run 2 data taken from 2015 to 2018, plus corresponding simulations, in MiniAOD format for one version. The total available MiniAOD data in 2020 across all versions was close to 10 PB. In order to extrapolate to HL-LHC, the following assumptions are made:
  - The size of MiniAOD per year increases by a factor of 30 from Run 2 to HL-LHC.
  - With 15% Moore's law scaling per year in terms of the storage capacity per unit cost, a constant annual budget, and assuming that the oldest parts of the storage are retired at the rate of one-eighth per year, the storage capacity per cache would increase only by a factor of two over eight years. The nominal operating point for the caches given constant funding will thus be substantially different in 2028 than in 2020, with a necessarily more frequent cache-refresh rate. As a result, that the Tier 2 caches will need to support an average network utilization of from 100 Gbps to several hundred Gbps by 2028 does not seem unreasonable, if MiniAOD continues to be a significantly used data format.
  - Given that substantial parts of the detector are brand new, and require commissioning, it seems prudent to assume that NanoAOD will not be the dominant data format in 2028/29. In fact, COVID-19 and the schedule delays it produces, and the details of ramp-up of the LHC, understanding of the detector, and commissioning of the data formats impose very substantial risks on the details of any projections at this point.
- At this point, it is probably safe to assume to plan for 100Gbps sustained use for all Tier 2s, with occasional bursts to 400 Gbps throughout the first run of the HL-LHC. The intention is to work with all relevant network providers from campus, through regional and national, to make sure each Tier 2 achieves this goal. It is also probably safe to assume that maybe not all

Tier 2s will be able to provide this capacity at the beginning of the HL-LHC era. Thus, it is necessary to have the software infrastructure for high-level management of activities at different sites depending on their network capacity. Over time, the experiments may make hardware purchasing decisions of CPU and disk at different locations based on the network bandwidth, space, costs, and power at those locations. It is thus conceivable that not all Tier 2s will be close to the same in the future.

At present, the Tier 1 at Fermilab provides 260kHS06 of processing power, 27.2 PB disk space, and 88 PB of tape archiving. In 2020, Fermilab offered 2x100 Gbps WAN network connectivity dedicated to the Tier 1 for CMS, and another 100 Gbps shared with the rest of Fermilab. Typical WAN use averages 50–100 Gbps, with occasional spikes that peg the entire 200 Gbps available WAN bandwidth of the Tier 1. The Tier 1 capacity in 2020 was thus roughly 2xTier 2 in processing, and 5xTier 2 in disk space.

We present a Tier 1 network needs estimate in Section 5.10.8.4.2, after explaining the workflow that is expected to dominate the Tier 1 WAN bandwidth in the HL-LHC era. Section 5.10.8.4.2 also applies a similar logic to a hypothetical HPC center to derive a needs estimate for an exascale HPC system.

A third component of network needs will be those required to support distributed physics analysis, which is expected to be centered at a variety of AFs. These are dedicated pieces of infrastructure designed to provide access to large data sets and computational resources that enable rapid iterative analysis of physics data. A motivation for this concept is the fact that columnar analysis of data formats like the NanoAOD promises to provide a very large increase in data analysis speed, as measured in number of events analyzed per second. Typical analyses of MiniAOD in Run 2 achieve processing rates O(10)Hz while columnar analysis of NanoAOD promises to achieve O(1kHz) or more, both per hyperthread. In addition, an analysis bottleneck in Run 2 MiniAOD analyses is the creation of fast user n-tuples. Several salient features are:

- The AF is a service provided by a Tier 2 and thus fits into the data lakes model of WLCG.
  - It "schedules" an interactive end-user analysis capability onto the Tier 2 resources dynamically; i.e., it uses US CMS Operations Program funded hardware at the Tier 2s.
- A service called "ServiceX" supports user-level n-tuple production (i.e., what would traditionally be done by the end users). It is expected that significant savings in human effort will occur from this as students and postdocs no longer need to develop their own n-tuple frameworks nor operate large workflows to produce their analysis n-tuples.
  - As part of this, ServiceX should support MiniAOD additions to end-user "n-tuples" derived from NanoAOD. Use of AFs will accelerate adoption of NanoAOD by allowing mergers of NanoAOD and pieces or derived quantities using MiniAODs at ingestion into the AF.
  - The computing resource planning for HL-LHC presently foresees:
    - 50% of analyses use NanoAOD.
    - An additional 40% use MiniAOD.
    - And only 10% or less require AOD.
  - The AF concept would lead to 90% of analyses effectively using NanoAOD by merging parts or derived quantities from MiniAOD into NanoAOD automatically, without the intervention of the end user.
    - It is unlikely that AOD could be automatically merged with other formats, similar to the method used for MiniAOD and NanoAOD, as it predominantly resides in tape archives that may be contained in a different data lake (i.e., not be accessible via the caches inside the NanoAOD lake but requires re-staging from tape via Rucio).

- AF supports both interactive and batch mode.
- The data access patterns inside the AF are expected to require low-latency random access media to achieve best performance. It is expected that the columnar data to benefit from SSD or Non-Volatile Memory Express (NVME) hardware-based storage.
    - This implies that an AF does not stretch across site boundaries. All the IO to the columnar data store is local within the AF.
- R&D is ongoing to accelerate ServiceX as a high IO bandwidth service, accelerated via FPGAs or GPUs. The input flow to this service would be scheduled in this scenario.

- AF supports extraction of user-defined data formats to migrate onto laptops, desktops, workstations at home institutions, or at home.
    - This would be a scheduled data transfer.
- Deployment of AF services is envisioned to be done on industry standard platforms like Kubernetes to facilitate deployment within a Tier 3 context.

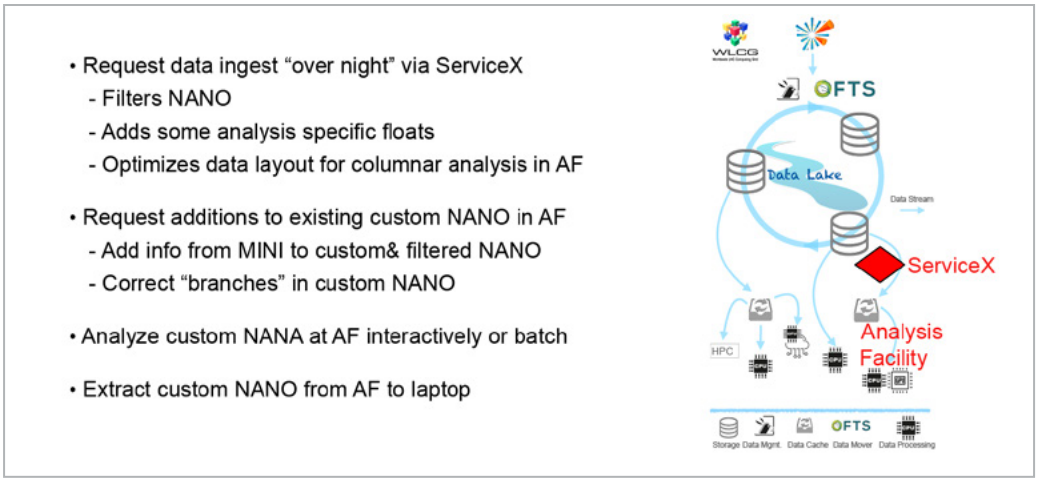


Figure 102: AF concept, how it fits into the Data Lakes Model, and some of its features.

Figure 102 shows how the AF concept fits into the data lakes model. The AF is a new concept discussed both within IRIS-HEP, where there is a significant R&D effort toward it, as well as within the HSF and WLCG. It is conceptually much less mature than anything else discussed in this document, and thus subject to significant changes as R&D continues.

At present, it is assumed that the Tier 2 network requirements as specified previously will be sufficient to support AFs at the Tier 2s. From Run 2 experience, a typical analysis requires between 5–30% of a given data release versions data. As long as the use of ServiceX is structured such that it uses MiniAOD (pieces or derived quantities) as an add-on to NanoAOD rather than as a replacement of NanoAOD, the IO requirements of ServiceX will probably be within 100 Gbps steady state and 400 Gbps burst capacity during the first three-year running period of HL-LHC. Ten percent of 30 PB for an annual MiniAOD version is 3PB. ServiceX will do the job "overnight," say in 30k seconds. 3PB in 30k seconds is 0.1 TB/sec or 1 Tbps. This would not fit into the 400 Gbps envelope. However, if MiniAOD information was added only to a subset of filtered events, or only a few floats per event needed to be fetched via the WAN, then the bandwidth requirements could be very comfortably within the 400 Gbps burst capacity.

5.10.8.4 Process of Science

The primary concern for both ATLAS and CMS in the HL-LHC era will be data volume increases brought on by the upgrades to the experiment. Both are entering into R&D to approach these challenges in a number of ways

5.10.8.4.1 ATLAS

The computing workflows and workloads described in Section 5.10.5 for simulation, data processing, derivations, and analysis will continue to be required, with some significant improvements.

The HL-LHC, commencing with Run 4 in 2027, will deliver unprecedentedly complex events, with up to 200 interactions per proton-proton bunch crossing. These events will be collected at a prodigious rate: ATLAS expects to record data at 10kHz, approximately ten times more than during previous runs. By the end of Run 5 in 2034, the HL-LHC is expected to have delivered an integrated luminosity of up to 2500 fb^{-1}, five times more than all previous runs combined. As well as the challenges involved in collecting, storing, reconstructing, and analyzing such a colossal volume of data, MC simulation events will need to be produced in similar numbers. Taken together, the data and MC requirements of the HL-LHC physics program are formidable, and if the computing costs are to be kept within feasible levels, substantial improvements must be made in both compute and storage.

Networking has been fundamental to the success of ATLAS and 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. HL-LHC data and processing scales are large enough that optimizing for efficient bandwidth use will be essential.

As part of this optimization, the ability to mark network packets to identify sources of traffic, and to route traffic to control speed and cost, may become vital. To this end, ATLAS initiated the HEPiX Network Function Virtualization working group in collaboration with other experiments and national entities, including ESnet. US ATLAS will play a central role in the R&D in this area, drawing on the close relationship with ESnet, the longstanding networking expertise within the facilities, and distributed computing leadership in PanDA and its Rucio integration; shaping network traffic through Rucio and PanDA will be crucial.

Economizing storage is an important goal for HL-LHC computing. Unlike CPU, the requirements for which will become approximately constant once the LHC reaches its design luminosity, storage needs will continue to increase during the lifetime of the HL-LHC. And while opportunistic computational resources exist, opportunistic storage does not. Innovative ideas for optimizing storage by breaking out of the disk/tape paradigm to a finer-grained spectrum of storage cost-reliability-latency are being pursued. In particular, the US-driven ATLAS data carousel project is developing a mechanism to stage data (such as AODs) 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. The data carousel leverages the managed "train production" approach already utilized by ATLAS to drive a tightly integrated orchestration of workflow and workload management that processes data with file-level granularity as soon as it appears from tape, keeping its disk residency to a minimum. With tape facilities at only ~10 sites, and processing resources distributed across over 100 sites, remote processing requiring remote data delivery is an essential element of the approach.

Remote data delivery reliant on powerful networks will in general be essential to minimize data replication and disk storage footprint while fully utilizing distributed processing resources. In order to most efficiently use bandwidth and minimize latencies, ATLAS (largely US ATLAS) is developing new services and workflows to deliver across the network only the data needed by the consuming workload, with fine granularity that avoids the inefficiencies and latencies of data pre-staging and enables the use of processing resources to be highly dynamic while remaining efficient. The intelligent data delivery service (iDDS) provides support for such fine-grained workflows, with its ServiceX component providing intelligent data transformations and filters to deliver just the needed data.

5.10.8.4.2 CMS

As mentioned in Section 5.10.6, it is necessary to distinguish central processing use cases from data analysis use cases. The latter is driven by the end users, and is thus without any central control. As mentioned in Section 5.10.6, a centrally managed set of platforms is provided, including CRAB, direct access to HTCondor, and the concept of an AF, as shown previously. The latter is a very new concept that is not yet mature, as described in Section 5.10.8.3.2.

This section presents a discussion of central processing workflows and the derivation of estimates for network needs for those components that were not outlined previously.

Central processing workflows include:

- Data processing that requires tape recalls. This includes:
  - Raw data processing (364 PB per year) and reprocessing with at least AOD (112 PB) as output, but maybe also MiniAOD (14PB) and NanoAOD (0.1PB).
  - AOD (240 PB per year) processing with at least MiniAOD (30 PB) as output, but maybe also NanoAOD (0.24 PB). Here both detector data and simulations are included.
  - The United States has traditionally provided 40% of CMS's archival capabilities outside of CERN. See Section 5.10.8.1.2 for the evolution of archive size versus time.
  - We presently assume that all primary processing of data archived in the United States is done in the United States, and no data archived outside the United States is processed inside the United States. This was the model in Run 1. In Run 2, a much more dynamic data movement pattern was allowed. At the beginning of HL-LHC, it is expected to go back to more restrictive regional commitments in order to better control the total TA bandwidth, as well as the US total commitment for processing.
  - While a model of enforcing strict processing boundaries can be used for raw data, it is less clear that the same model can be enforced for the AOD format. Some analysis workflows may require access to AOD, which implies that the experiments will either centrally produce a custom NanoAOD from the AOD, or allow the AOD to be recalled from archive and analyzed by end users. While this could all be done regionally, it might be desirable to move AOD across the Atlantic for this kind of purpose.
- Data processing that requires no tape recalls. This includes:
  - MiniAOD (30 PB per year and version, which includes data and simulation) processing to produce NanoAOD (240 TB).
  - Simulation+Digitization+Reconstruction workflow with the PileUp library as input. It is expected the PileUp library will be roughly 10–20 PB in size, and thus to be disk resident for the duration of a processing campaign. Even in the event of multiple concurrent simulation campaigns, the total of the PileUp library will be less than the total of the MiniAOD. The output of simulations is dominated by AOD (128 PB per year and version) with MiniAOD (16 PB) and NanoAOD (0.14 PB) being significantly smaller. Typically, all three formats are produced in the same workflow, leading to a total of roughly 145 PB of output per simulation campaign per year.

As processing cadence, it is presently assumed:

- One complete processing of the raw and production of corresponding simulations at the end of every year of data taking.
- In addition, a second complete raw processing and simulation will be produced at the end of every three-year running period.

- A partial processing of raw and simulation data will be done as the data are being collected. This is referred to as "prompt reconstruction." Some simulations will be done before data taking commences. How large a fraction this is of the total for either raw processing or simulations is not yet decided. A minimum of roughly 5% of the annual total is deemed required for understanding the data quality and verifying proper functioning of the instrument.
- It is still somewhat unclear how often MiniAOD and NanoAOD will be made.
  - We generally expect that NanoAOD will be produced more often than MiniAOD. This is easy to support given that MiniAOD can be disk resident given its size.
  - Producing MiniAOD from AOD is much more difficult, as it requires tape recalls. Given the size of the AOD (240 PB per year and version) it is possible that tape recall bandwidth available will restrict us to only one MiniAOD per year in addition to the AOD processing campaigns.
  - Taking all of this together, a rough estimate of total tape recall for central processing of 364 PB (raw) + 240 PB (AOD) = 600 PB.
- Assuming Fermilab hosts 40%, this implies 240 PB tape recall for central processing per year from the Fermilab tape archive.
- There is very likely additional tape recall required to support various smaller activities. Clear estimates for that are currently lacking.

To estimate the archival bandwidth needs for Fermilab, a slightly different argument is presently used as follows:

- The annual raw processing should complete within 100 days.
- 40% of 364 PB in 100 days plus 240 PB in 300 days adds up to 1.5 PB/day + 0.3 PB/day ~ 2 PB/day archival recall at Fermilab T1 as target goal for HL-LHC.
- Implicit in this estimate is that all other uses for tape recall fit into the days when the tape recall is not completely used as detailed previously.

From this, the outgoing network bandwidth needs of Fermilab T1 can be calculated just for central processing:

- 2 PB/day ~ 200 Gbps at 100% utilization.
- To allow for bursting to fill processing buffers at processing centers, having x5 this bandwidth (or 1 Tbps) in place would be ideal.
  - Note: This implies that not all Tier 2s can attempt to burst traffic at 400 Gbps to/from Fermilab at the same time. Due to the shared nature of the regional and national networking infrastructure, it is a common occurrence that some sites share network infrastructure to reach Fermilab. Any attempt to schedule burst behavior must feature awareness of current utilization patterns, as well as workflow that may be using shared links.
- If 1 Tbps was available, then it is assumed that all other data transfer use cases could be fit into this x5 burst margin, as long as the means to schedule bulk transfers exist, and it is possible to set priorities for different flows. This leads to the networking R&D discussed in Section 5.10.8.7.2.
  - Note: The raw data transfer from CERN to Fermilab is also within this 1 Tbps peak. The numbers documented previously for transfer of raw to a hypothetical large HPC systems apply equally well to transfer of raw data from CERN to Fermilab.
    - 364 PB of raw data are produced per year at CERN when the LHC is running.
    - 40% of that gets archived at Fermilab => 145.6 PB/year.

- The duty cycle of the accelerator is roughly 30%, so ideally, there should be a plan for transferring the 146 PB in roughly 100 days, rather than 365 days.
- That is 1.5 PB per day CERN to Fermilab, or 150 Gbps average, but preferably transferred in scheduled bursts of three to five times that.

Using the same logic, and combining it with an estimate of processing requirements, it is possible to estimate the WAN bandwidth requirement of an exascale HPC system as follows:

- Following the 15% scaling per year from Section 5.10.8.3.2, one might expect that an HPC center compute node in 2028 with two processors might have 128 x 3 = 384 x86 equivalent physical cores, plus some accelerators. Assuming the per physical core processing power is roughly 15 HS06 and a reconstruction requirement of 6000 HS06-seconds would lead to an estimated processing power of 1 event per second per compute node, not accounting for accelerators.
- Assuming that it is possible to find ways for half of the reconstruction software to be offloaded onto the accelerator, the 1 Hz would result in 2 Hz processing per exascale HPC compute node.
- Assuming that a typical allocation would allow us to use 5,000 nodes, such an allocation on an exascale system might be able to reconstruct events at a rate of 10 kHz.
  - Nodes with Dual EPYC 7702 processors are available today. While not yet benchmarked, these can be estimated to provide 256 HTs with 15 HS06 per HT, or 3840 HS06 per node. At 15% Moore's law scaling per year, one would be able to buy single nodes with 11,520 HS06. Furthermore, it can be assumed that half of the processing can be delegated to an accelerator; then a node achieves an equivalent of 23k HS06. Five thousand such nodes would lead to 115 MHS06. The per-event reconstruction CPU power is estimated as 6,000 HS06. Dividing these two numbers leads to a 19 kHz event processing rate. For the purpose of estimation, this can be rounded down to 10 kHz to be somewhat less aggressive in the estimate.
  - It should be obvious from these numbers that there are very large errors in these extrapolations in both directions.
- At a raw event size of 6.5 MB, an exascale HPC system of this type would require a WAN bandwidth of:
  - 10,000 events per second x 6.5 MB per event = 480 Gbps for streaming data. It should be noted that this would exceed the tape recall capabilities of Fermilab, which are estimated at 200 Gbps.
  - If instead a burst, rather than a stream of data, was used into the exascale HPC system, then it might be possible to see 1 Tbps WAN connectivity for exascale systems. A typical workflow would then include five days of tape recall at 200 Gbps, followed by a one-day transfer burst to the exascale HPC system, followed by two days of processing on the HPC system using 5,000 compute nodes. This would imply a 10 PB disk buffer at Fermilab and the HPC system. In this operations mode, a Tbps network connection between Fermilab and the exascale system would be in use one out of every five days, while the 5,000 nodes at the exascale system would be in use for two days out of every five.
  - The output data size for raw processing is roughly one-third that of the input data. As networks are bidirectional, the input to the exascale center determines the required WAN bandwidth.
  - Obviously, this is just a numbers game to illustrate the idea.

We can do a similar exercise for the central simulation workflow as follows.

- Here the input data are 13.5 MB per event of pileup. The output is 2 MB/event of AOD, with MiniAOD and NanoAOD being negligible by comparison.
- CPU time per event for simulation+digitization+reconstruction amounts to 8,000 HS06. IO/CPU on input is thus very similar to raw processing.
- The difference in this workflow is that the total pileup library of events is only O(10 PB) and should thus really be replicated in its entirety to the HPC center, as it is heavily reused for the entire 60 billion event annual campaign. Streaming the pile-up (PU) sample to a large HPC system does not make a lot of sense for this workflow given the large data reuse and the very large processing power of such a system. At a Tbps WAN bandwidth, a 10 PB pileup sample could be transferred in a single day, and then reused for many days after.
  - Note: elsewhere in this document it is argued that streaming might make sense for Tier 2s. In fact, today streaming the PU sample across the network is done. Streaming or replicating really depends on the processing power of the facility as that determines the reuse of the sample.

Decentralized, end-user driven workflows are much more uncertain at present for multiple reasons:

- First, it is unclear how much AOD use is required for end-user analysis. This is especially unclear given that a "commissioning period" is expected for the higher-level data formats like MiniAOD and NanoAOD.
  - At present, it is assumed that CMS will make available only a fraction, maybe less than 10%, of the total AOD data on disk for this commissioning period, and maybe only 1% of raw.
  - At present, it is assumed that commissioning the higher-level data formats will not take more than one year.
  - Both of these assumptions might turn out overly optimistic.
- Second, a transition is expected in the way end-user analysis is done at the LHC from individuals and small groups producing their own user data formats to those formats being produced at ingest in AFs.
  - Details on this transition are presented in Section 5.10.8.3.2.
- Third, the caching and streaming paradigms implemented in the data lake model allow for a tradeoff between disk space at the Tier 2 and networking bandwidth to the Tier 2.
  - Section 5.10.8.3.2 shows how this leads to an order of magnitude difference in network bandwidth use depending on how the caches at the Tier 2s are dimensioned.

5.10.8.5 Remote Science Activities

The LHC accelerator and the ATLAS and CMS detectors are located near Geneva, Switzerland. The storage archive that stores all the data is globally distributed as listed under T1 and T0 facilities previously. The compute and nonarchival storage facilities are even more widely distributed as described previously. In essence, all science activities are remote. As a result, the HL-LHC computing infrastructure will remain distributed. In other sections, ATLAS and CMS have described the LHC computing model, and this basic model is being evolved to meet the needs of HL-LHC. It will certainly augment this infrastructure with additional resources that might be used, like HPC sites, commercial and research clouds, and additional institutional resources.

5.10.8.6 Software Infrastructure

The software infrastructure of both ATLAS and CMS will undergo changes influenced by R&D efforts that are ongoing. Migration to more capable tools (some of which will be in place for Run 3) will help to ensure that all aspects of the workflow can function for the increase in data volumes.

5.10.8.6.1 ATLAS

The ATLAS Distributed Computing (ADC) system and organization today is a sophisticated ensemble of software systems, computing facilities, and people that make it possible for ATLAS processing to run 24x7x365 on about 450k cores across over 100 worldwide sites and a wide range of resource types including grid clusters, HPCs, LCFs, and clouds, processing in total about 1.5 EB of data per year. The new HL-LHC directed capabilities and workflows described in previous sections are being developed on the proven and scalable foundation of this system, particularly the PanDA workload management system, ProdSys production management system, and the Rucio DDM system (which were described in Section 5.10.5). Beginning with Run 1 in the case of PanDA and Run 2 with ProdSys and Rucio, these systems have met the capability and scaling requirements of ATLAS, and have steadily been enhanced and hardened based on running experience. ATLAS has confidence that they have the scalability and robustness to continue to serve ATLAS through Run 4. Hence ATLAS has been able to focus on adding new HL-LHC directed capability to these systems. PanDA, ProdSys, and Rucio are being extended for the demands of HL-LHC in such projects as the data carousel and the intelligent data delivery service (iDDS). Rucio has been adopted as the LHC and WLCG standard. The iDDS service and ServiceX are being developed in a joint effort between ATLAS and IRIS-HEP to support fine-grained workflows with intelligent data transformations and filters to dynamically deliver just the needed data across the network, where and when it is needed. While iDDS and ServiceX both work in close concert with ADC services like PanDA and Rucio, they are implemented as experiment-agnostic services.

5.10.8.6.2 CMS

As described in Section 5.10.6.4.2, the distributed scientific process depends on a number of conceptual elements. In the following bulleted list, the present assumptions are listed as to the software infrastructure products that will be used for each conceptual element, and the architectural relationships between elements where necessary to understand data movement. A much more detailed list of infrastructure software products in use now and in the future is given in Section 5.10.6. Networking R&D that underpins these is described in Section 5.10.6.7.2.

- A consistent runtime environment globally for LHC science applications.
  - Software, configurations, and calibrations are distributed via Apache Squid. The CVMFS is used for software and configurations while the FroNTier system is used for calibrations. Both systems are optimized for very large-scale transaction rates and very modest data volumes.
  - FroNTier is back ended by an Oracle Real Applications Clusters (RAC) system at CERN. Accesses from applications translate into HTTPS accesses to Apache Squid, which translate into SQL on the Oracle RAC.
  - CVMFS has a more federated architecture. Multiple "origins" may serve different parts of the total CVMFS namespace visible to an application. The application performs read-only filesystem accesses that get translated into HTTPS cached in Apache Squid, back ended by a translation back to a local filesystem at one of the federated data "origins."
- Data lake implementation in US CMS: at present, it is assumed that there will be two supported implementations as follows.
  - Data lake implementation for the Tier 1 tape archive at Fermilab is presently planned to be done via dCache. dCache is assumed to provide the disk buffer storage space in front of the Fermilab T1 tape archive.

- Data lake implementation for the Tier 2 caches and origin(s) is presently planned to be implemented via XROOTD. All Tier 2 centers will support some form of XROOTD implemented data access points.
- dCache supports integration into an XROOT data federation as a storage endpoint. It is unclear at this point whether or not the dCache instance at Fermilab T1 will be used as an origin to the US CMS data lake comprising all Tier 2s, and possibly most Tier 3s in the US CMS globally separated tape and disk-based storage endpoints in Run 2. The purpose of distinguishing disk and tape endpoints was to disallow application access to data from tape architecturally. Whether or not those get remerged for the HL-LHC remains to be seen. Within the data lake model, this separation is accomplished most obviously by defining the tape archive with its dCache front end as its own data lake. Data movement between lakes would thus be required for archived data to be accessible to applications.

- Data movement between data lakes.
  - Rucio forms the high-level layer. This includes the global file catalog, including any file-related metadata, any high-level policies on data replication, and any high-level management of that replication. Data lake endpoints are registered in Rucio. Rucio thus understands the structure of the global data federation with all of its lakes. It is expected there will be one Rucio instance each for ATLAS and CMS.
  - Rucio calls FTS to perform the actual data transfers. FTS simply has queues of file transfer requests and manages those queues and the actual transfers. Error handling of transfer requests is performed by FTS. All transfer requests by FTS are third-party (TPC) transfer requests between endpoints. No data are flowing through the FTS server itself. It is expected a few FTS servers will be sufficient to manage all transfers between all lakes.
  - Data lake endpoints speak at least one of two protocols: HTTPS or XROOT. Both support third-party copy. All lake endpoints are expected to support third-party transfers. There may be exceptions to this rule for some T3, cloud, or HPC center resources (i.e., resources outside of the control of the experiment). Entities that do not support TPC will be able to transfer data only with entities that do, as one of the two endpoints engaged in a transfer must handle the third-party copy request from FTS.
  - We expect that US-based processing facilities will be part of US data lakes only. Data lakes do not span the Atlantic nor 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. In particular, CMS has a Tier 2 in São Paulo that is expected to continue to be relevant also in the HL-LHC era. At present, this São Paulo Tier 2 is coordinated with the US-CMS facilities program. Whether or not it makes sense to make São Paulo part of the US-CMS data lake remains to be seen. There are Tier 3s in other places in Latin America.

- Data movement within a data lake.
  - Both caching and streaming are foreseen as use cases for data movement within a lake for lakes that are attached to processing resources.
    - At present, all data movement within the lake prototypes that have been deployed are based on XROOT. This means that any application that streams data does so via the XROOT protocol at present. In principle this could be done with davix, but in practice it has not been done in this manner.
    - Streaming is supported only for applications and between locations where RTT is small enough to guarantee good performance. In particular, today CRAB allows the end user to specify "ignore locality." In that mode, CRAB will place applications anywhere in

the world irrespective of data location. Global streaming is thus conceivable today, but this approach is expected to be forbidden during the HL-LHC era. Within data lakes, streaming will be to caches and/or origins nearby.

- If the Fermilab T1 tape archive was implemented as its own data lake then it might not have any processing resources attached to it.
  - A disk buffer for central processing workflows at a large HPC center might be considered its own data lake in this context (i.e., the endpoint is registered in Rucio, and data are transferred by Rucio into the buffer).
  - Alternatively, an HPC center that is operated entirely via data streaming would be attached to an existing data lake. In principle this could be done either as a processing resource that is part of the Tier 2 lake or the Tier 1 archive lake.
  - All of these kinds of details are still the subject of active R&D.
- An AF focused on columnar data analysis supporting MHz event rates for interactive analysis would be attached to a data lake via ServiceX. Applications on the AF would thus be strictly local, accessing data only via the LAN. This may change as the AF R&D matures in places like Caltech and UCSD that have very small RTT. In general, the AF concepts are the least mature within the HL-LHC computing model at this point.

5.10.8.7 Network and Data Architecture

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 baseline is currently being planned are terabit-scale (1–2 Tbps) backbone networks with the largest resource sites connected at multiple 100 G scale (200–800 Gbps). Network use will be at least a factor of 10 larger than Run 2.

We are also assuming that the global R&E network use will have a different character on the timescale of the HL-LHC compared with the situation the WLCG has experienced in runs 1 and 2. Specifically, the experiments foresee other science domains (astronomy, biology, and engineering) becoming global network users at scales equaling or exceeding the LHC users. In such an environment, it is not wise to assume the experiments can continue to treat the network as an infinite resource, and they need to explore options for effectively operating in a bandwidth-constrained environment.

5.10.8.7.1 ATLAS

For HL-LHC, four main requirements have been identified by ATLAS:

- **Capacity:** Run 3 is moving to multiple 100 G links for large sites, while Run 4 (HL-LHC) is targeting Tbps links.
  - Capacity is fundamental for science at HL-LHC scales.
  - As noted, in a capacity-constrained environment, it will be important to manage the capacity to do the most science possible.
- **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; it will take time.
  - Traffic shaping activities underway in the Research Network Technical Working group are a good capability example.
  - Network orchestration between sites is another good example and something that could help us more effectively exploit available resources.

- **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.
  - Packet marking is viewed as a high priority so that everyone (experiments, sites, and R&E networks) can understand the origin and intent of network flows at any point along their path.
  - We need better mechanisms to coordinate the available monitoring resources from the sites, experiments, and networks, to allow us to have a better understanding of how complex infrastructures are using the network.
- **Testing:** developing, prototyping, and testing network features at suitable scale will be needed.
  - Networks of the future will likely have new features, capabilities, and services that could be leveraged to do more with the resources that are available.
  - Our challenge is identifying which features might be beneficial to try to integrate into operations, noting that such integration can require significant effort to realize.
  - Having at-scale network testbeds will be very important to understand the potential impact changes might have on operations.

These requirements should also motivate and guide network infrastructure upgrades and replacements. It will be critical to understand the bigger picture of R&E networking and its evolution, to ensure the experiments are able to take advantage of capacity and services available.

5.10.8.7.2 CMS

As presented in previous sections, the following networking bandwidth requirements are expected for the first three-year running period of the HL-LHC from CMS:

- The Tier 1 at Fermilab will require Tbps burst capabilities. Steady state network bandwidth consumption is expected between 200–300 Gbps, at a minimum:
  - 200 Gbps to match tape staging described in Section 5.10.8.4.2, and an additional 100 Gbps assuming that the Tier 1 contributions to the US-CMS data lake internal traffic are at roughly the same level as a Tier 2.
- Tier 2s will require 400 Gbps burst capabilities. Steady state network bandwidth consumption is expected to be approximately 100 Gbps, depending on the operational details and use of the various event-formats discussed in the previous sections.
  - There is an expectation that not all Tier 2s of US CMS will achieve 400 Gbps burst capacity on day one of the HL-LHC era. Workflows will need to be matched to network capabilities at the Tier 2s, and possibly even hardware deployments over time, based on networking capacity available at the centers. Tier 2s thus may differ substantially from each other as a result of their network capacity.
- The large exascale HPC centers funded by the DOE will require Tbps burst capabilities in order for CMS to pursue the workflows as described in this document.
- If the NSF were to fund exascale systems in the future, then those would require the same Tbps burst capabilities as the DOE systems mentioned previously.
- As Tier 3 systems are smaller in scale, or CMS allocations on big systems are smaller in scale, networking requirements are expected to be more modest. The discussions in previous sections should sufficiently explain how to scale down requirements appropriately to adjust to the scale of a given Tier 3 system. However, from the technical point of view, transfers in bursts of up to 100 Gbps from any Tier 3 should be possible by 2028.

The core objective of US CMS with regard to networking, as the experiment prepares for the HL-LHC, is to arrive at a system that provides transparency of use, management, and planning. This implies an operational model matched to the available network capacity, but that also implies sufficient capacity matched to the needs. US CMS wants to be able to account for the bulk of the usage of LAN and WAN networking resources. For WAN resources, there is a desire to reason at a high level about why the network is used, at what capacity, and when. For at least some, if not all major network links, there is a desire to plan and manage our bandwidth use. US CMS would like to collaborate with ESnet on understanding implementation options (e.g., bandwidth management via SENSE and AutoGOLE, flow tagging to connect low-level to high-level accounting, and other novel networking features that may not be fully articulated). The overall goal is to understand the requirements related to capacity and capability in a manner that is compatible with other network use, both in the core and at the edges (at the Tier 1 and Tier 2s).

Ideally, engaging in an R&D partnership with ESnet from research, through testing at scale to deployment of new services in production, is desired. US CMS is ready to be an early adopter and collaborative partner in the development of new high-touch features in ESnet6, as well as exploring computer science experimentation and benchmarking on the FABRIC testbed, and its extensions to CERN.

Expected LAN/MAN/WAN Deployments

US CMS expects computational nodes to be connected at 10 Gbps, data nodes at up to 100 Gbps (depending on size), campus networking to institutional boundaries at Tier 2s to reach multiple 100 Gbps, and for the Tier 1 at Fermilab to reach 500 Gbps to Tbps. It is expected that the Chicago MAN-link will provide Tbps to Starlight, and ideally Tbps across the Atlantic to CERN. 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 of US CMS. And it is expected that these facilities will share bandwidth in LHCONE in between each other, and to the Fermilab Tier 1. US CMS will continue to collaborate and share with ATLAS, as well as other science projects. Sharing network bandwidth with ATLAS and other science projects is expected. Given the large burst needs articulated in this document, network management will be a core concern and area of research.

The following paragraphs summarize the salient features as relevant to the networking R&D goals.

Network Use Accounting

All relevant traffic accounting will be performed by either FTS or XROOTD infrastructure software. Any transfers between data lakes, as well as all output handling of central production workflows, will involve FTS. All data streaming to applications from either caches or data origins inside the lake will involve XROOTD. This being said, it is sufficient to instrument only these two software products, validate the accounting information in great detail once, and continue routine validations into the future as new releases get deployed. US CMS has started this program of work and expects it to be complete within a year. There is a desire to engage with ESnet on the "routine validations," meaning that low-level network usage metrics are compared with the high-level view from FTS and XROOTD on a routine basis. This implies sharing of monitoring data. In addition, there is a desire to "tag" flows or packets in some way such that low-level monitoring can identify which high-level activity the flows should get attributed to. This is again an area where joint effort with ESnet is desired, especially in the context of the FABRIC testbed and the ESnet6 rollout.

It is worth mentioning that HTCondor file transfer has seen sufficiently large use in CMS at various times to be noticed. And it is conceivable that managing large-scale production transfers in some cases in the future will be a driving use case. Thus, it is expected that US CMS will work with the HTCondor team to have the same monitoring validation, both one time and continuous, as for FTS and XROOTD.

The LHC community is proposing a partnership with ESnet to advance the R&D effort surrounding traffic tagging from XROOTD caches and origins, FTS servers, and HTCondor file transfers. These projects will better help to understanding traffic accounting on the wide area networks that carry experimental traffic.

Network Management and Planning

We would like to explore a collaboration with ESnet in the context of SENSE and AutoGOLE toward the goal of having all traffic within and among data lakes managed. This implies that all burst transfers between Tier 1, Tier 2, and HPC centers are planned and managed transfers. US CMS is open to suggestions where to start. For us, a reasonable starting point would be the TA link, as that seems likely to be the first production link that requires managed transfers in order to allow reduction of the factor x5 in provisioning that is currently being applied by ESnet. Long term, US CMS is comfortable with an operations principle where some fraction of the bandwidth, say 10–20%, is reserved for unmanaged transfers, while the bulk of the bandwidth on the most important links is managed. It is accepted that unmanaged transfers will generally experience suboptimal performance. The bulk of traffic on LHCONE in the United States and across the Atlantic will also have to be managed, according to the foreseen capacity requirements and the projections discussed previously, including the fact that LHCONE is the largest class of traffic in ESnet, and the one with the largest growth rate as well.

We would be most interested in picking some near-term goals to put managed transfers into production for some links and some use cases to gain operational experience and a better understanding of the relevant concepts and software capabilities. This can be with limited functionality with the understanding that US CMS will collaboratively explore increased functionality as the limited functionality is deployed and operated in production[314].

Network Performance Measurements

US CMS delegates network performance measurement collection to OSG and expects to continue to do so. At present, Caltech, UCSD, and UNL have 100 Gbps perfSONAR hosts in a MaDDash operated by PRP. Other locations within this mesh include various Internet2 backbone nodes in Chicago, Manhattan, Kansas City, etc. A corresponding mesh is starting up as part of the OSG-LHC networking activities. Long term, all Tier 2s and the Tier 1 will keep up their perfSONAR instrumentation with the bandwidth requirements for the sites.

In order to manage network usage as desired, US CMS will require additional monitoring systems at the flow level and in the switches and routers to be able to quickly identify and/or mitigate or avoid problems having to do with capacity limitations relative to the requests, and/or anomalies in switching or routing that may arise. Real-time monitoring systems of this kind, as are being explored in AutoGOLE/SENSE and the PRP for example, could be used to maintain or increase operational efficiency.

5.10.8.8 Cloud Services

Cloud computing use cases for both ATLAS and CMS are still being explored via R&D before any serious production consideration.

5.10.8.8.1 ATLAS

While PanDA+Rucio can use commercial cloud resources interchangeably with grid-based WLCG resources, currently ATLAS has no plans to use clouds for the HL-LHC. The baseline plan is to use grid and HPC resources. If some grid resources are set up as cloud resources, they will also be used. However, commercial cloud resources are being evaluated for specialized usage by analyzers. There are currently two proof of concept projects. If these projects are successful, ATLAS will require good network pipe between grid sites and commercial clouds. In this model, the network needs will be similar to university-based US Tier 2 sites.

- Google: ATLAS is testing the use of GCP+GCS for end-user analysis. The project is funded for two years, with a decision expected in late 2021.
- Amazon: ATLAS is testing a virtual analysis center on AWS. This project is funded until summer 2021. A decision is expected soon after.

[314] A more detailed discussion of the kinds of functionality, and what a staged program of work might include can be found in the following document: https://www.dropbox.com/s/esv876hhw6ohn6h/ComputingandNetworkingOperations_LHCRequirementsandOutlook.docx?dl=0.

5.10.8.8.2 CMS

We have evaluated the use of commercial cloud both for processing and for TA transfer. US CMS finds that both are not cost-effective, at present. Experimentation has shown that making large-scale use of cloud resources if the cost structure were to change is possible[315].

5.10.8.9 Data-Related Resource Constraints

One interesting option to explore on the timescale of HL-LHC is the joint ATLAS and CMS use of Rucio for DDM. This could provide a mechanism to interact with ESnet (and other R&E networks), communicating near-term data movement intents and perhaps negotiating for any required QoS or deadline requirements.

There have been significant efforts made to understand the computing and storage requirements for HL-LHC. The process began by extrapolating current methods forward, accounting for the data increases from HL-LHC and the additional complexity associated with HL operations of the collider. Further assumptions about the initially conservative estimates of computing and storage capacities would be increased for fixed cost and additionally assumed the notion of a "flat budget" for the foreseeable future. Early estimates show significant gaps in both computing and storage (by factors of ~10).

To decrease the gap, significant efforts were made to better optimize how fewer resources could be utilized to do the same amount of science. This has helped decrease the difference between what is needed and what is affordable to factors of three to five instead of ten. While there are potentially some optimistic assumptions in this estimate, it is known that the planning process must change to ensure resource availability on the timescale of HL-LHC. At this point there is still a significant gap in the amount of storage believed to be needed versus what can be provisioned with current budget projections.

Another concern is the potential constraints in R&E network capacity on the timescale of the HL-LHC. One of the options that has been considered in addressing the gap in resources noted previously is taking more advantage of high-performance networks to help reduce storage requirements. However, this implicitly assumes the network continues to be unconstrained and free of cost to the HL-LHC collaborations. As noted elsewhere, network providers serving global HEP collaborations have been very successful in providing network capacities beyond the requirements of existing use cases, giving collaborations whatever bandwidths they have required. This is likely not going to be feasible for the HL-LHC era, not because of the significant growth in HL-LHC data, but rather due to the expected rise in use from other global, data-intensive science domains.

The LHC collaborations have not faced such an environment to date and, if such a situation were to arise, the tools or methodologies to work effectively are not present.

5.10.8.10 Outstanding Issues

One item to note is that the experiments feel that it would be beneficial to discuss what features should be part of network infrastructure acquisitions that the sites make. If a minimum set of capabilities in switches and routers could be identified, to ensure that the next generation of network components are able to both meet the current needs, and able to support prototyping for future capabilities, it would better prepare for HL-LHC operations.

[315] As a reference for a presentation on using cloud transfer capabilities that includes benchmarking as well as cost information, refer to https://indico.cern.ch/event/923131.

5.10.8.11 Case Study Contributors

HL-LHC Representation

- David Lange[316], Princeton University
- Garhan Attebury[317], UNL
- Harvey Newman[318], Caltech
- Kenneth Bloom[319], UNL
- Shawn Mckee[320], University of Michigan
- Margaret Votava[321], Fermilab
- Tulika Bose[322], University of Wisconsin-Madison
- Lothar Bauerdick[323], Fermilab
- Dan Marlow[324], Princeton University
- Justas Balcas[325], Caltech
- Elizabeth Sexton-Kennedy[326], Fermilab
- David Mason[327], Fermilab
- James Letts[328], UCSD
- Markus Klute[329], Massachusetts Institute of Technology
- Kevin Lannon[330], University of Notre Dame
- Brian Bockelman[331], University of Wisconsin-Madison
- Michael Hildreth[332], University of Notre Dame
- Frank Wuerthwein[333], UCSD
- Oliver Gutsche[334], Fermilab
- Christoph Paus[335], Massachusetts Institute of Technology
- Andrew Melo[336], Vanderbilt University
- Kaushik De[337], University of Texas at Arlington

---

316 David.Lange@cern.ch
317 garhan.attebury@unl.edu
318 newman@hep.caltech.edu
319 kenbloom@unl.edu
320 smckee@umich.edu
321 votava@fnal.gov
322 Tulika.Bose@cern.ch
323 bauerdick@fnal.gov
324 marlow@Princeton.EDU
325 jbalcas@caltech.edu
326 sexton@fnal.gov
327 dmason@fnal.gov
328 jletts@ucsd.edu
329 klute@mit.edu
330 klannon@nd.edu
331 BBockelman@morgridge.org
332 mhildret@nd.edu
333 fkw@ucsd.edu
334 gutsche@fnal.gov
335 paus@mit.edu
336 andrew.m.melo@vanderbilt.edu
337 kaushik@uta.edu

- Torre Wenaus[338], BNL
- Doug Benjamin[339], ANL
- Heather Gray[340], University of California, Berkeley
- Matevz Tadel[341], UCSD

ESnet Site Coordinator Committee Representation

- Phil DeMar[342], Fermilab
- Andrey Bobyshev[343], Fermilab
- Vincent Bonafede[344], BNL
- Mark Lukasczyk[345], BNL

[338] wenaus@bnl.gov
[339] dbenjamin@anl.gov
[340] heather.gray@berkeley.edu
[341] mtadel@physics.ucsd.edu
[342] demar@fnal.gov
[343] bobyshev@fnal.gov
[344] bonafede@bnl.gov
[345] mlukasczyk@bnl.gov

6 Focus Groups

A core component of the ESnet Requirements Review process that was displaced by the COVID-19 pandemic was the opportunity to hold impromptu conversations with colleagues. These could occur during the case study review period (and involve topics being presented or stumbled upon), but were also equally likely to occur before, during, or after the meeting. The importance of these interactions cannot be overstated, as they may result in cross-pollination of ideas, collaboration, or other forms of interaction fostered by the organization of the attendees and subject matter. Facilitating these types of interactions was a high priority, despite the challenges of conducting a fully distributed review process.

6.1 Purpose and Structure

In late September 2020, the HEP Requirements Review team convened a series of virtual "focus groups." The general plan for these meetings was to:

- Gather together small groups of case study authors during pre-defined time periods, using virtual tools.
- Prepare the groups by having them review outlines of their case studies and research focus (if they were unfamiliar).
- Structure a conversation such that there was time to review the areas of research, and then seed the conversation with a set of topics that were found to be common across all case studies in the 2020 HEP Requirements Review.

During these two-hour time windows, the HEP Requirements Review team acted as a moderator for the conversation, but let conversation flow organically toward topics of mutual interest. The goals were to:

- Allow emerging projects and facilities to ask questions of the established HEP community, to better prepare for the future.
- Facilitate discussion on known problems and solutions that will guide the process of science, and support from ethnology, in the coming years.
- Establish best practices that span the different parts of the HEP program area.

6.2 Organization

The HEP Requirements Review featured 13 case study groups. Thus the optimal organization for focus groups was to offer four events with three to four groups slotted to attend each. The groupings were as follows:

- Focus Group 1 was held on Tuesday, September 29, and involved the following groups:
  - Case Study #1: Cosmic Frontier Subprogram — Cosmology Computation and Simulation.
  - Case Study #6: Cosmic Frontier Subprogram —LZ Dark Matter Experiment.
  - Case Study #8: Intensity Frontier Subprogram — Belle II Experiment.
  - Case Study #11: Energy Frontier — CMS Experiment.
- Focus Group 2 was held on Thursday, October 1, and involved the following groups:
  - Case Study #2: Cosmic Frontier Subprogram — DESC.
  - Case Study #9: Intensity Frontier Subprogram — Neutrino Research at Fermilab (DUNE at LBNF and SBN Program).
  - Case Study #10: Energy Frontier Subprogram — ATLAS Experiment.

- Focus Group 3 was held Friday, September 25, and involved the following groups:
  - Case Study #4: Cosmic Frontier Subprogram —the Rubin Observatory.
  - Case Study #5: Intensity Frontier Subprogram — "Stage-4" ground-based CMB-S4.
  - Case Study #12: Energy Frontier Subprogram — LHC Operations.
- Focus Group 4 was held Wednesday, September 23, and involved the following groups:
  - Case Study #3: Cosmic Frontier Subprogram — DESI.
  - Case Study #7: Intensity Frontier Subprogram — Muons Research at Fermilab (Mu2e and Muon g-2).
  - Case Study #13: Energy Frontier Subprogram —HL-LHC Research.

The agenda for each event was designed to be simple and dedicated to keeping a majority of the event available toward attendee discussion:

- Brief introductions from the HEP Requirements Review Team, along with meeting purpose.
- "Elevator pitches" from the represented case studies. These could be presented verbally, or with visual aids, but were limited to five minutes to explain the case study background. Emphasis was placed on having the presenter reveal:
  - Structure and membership of the project or facility.
  - Science highlights, goals, and outcomes.
  - Process of science and use of technology.
  - Areas of "pride" for the effort, areas of need, and things worth sharing with outsiders (through the view of science or technology).

After the initial set of discussion, the remainder of the focus group time was allocated to discussion topics. These were defined prior to the meeting (and shared with attendees) by the requirements review team. All topic areas were pulled directly from observations made by case study authors. The topics are as follows:

1. Future networking requirements (capacity, traffic expectations, services).
2. Cloud computing potpourri (experimentation, interest, barriers).
3. Use of HPC (ASCR facilities or other) resources in HEP workflows.
4. Simulation approaches and activities (resource consolidation, etc.)
5. Use of HTC (OSG, etc.) resources in HEP workflows.
6. Data sharing tools/approaches (central versus distributed management, ad hoc).
7. Reprocessing campaigns in relation to experimentation lifecycle.
8. Long-term data management (central, distributed, tools.)
9. Analysis workflows: use of local versus remote versus distributed resources.
10. Fully remote/constant collaboration approaches/tools in 2020 and beyond.

A piece of "polling software" was utilized to gauge the relative interest in each topic area during the meeting. This was done to gain an understanding of what mattered to those who were represented in the room. The interest could be based on things they wanted to hear more about (potentially from other attendees), things they were concerned with implementing, or things they felt they could share experience with. Each focus group came to different conclusions about what topics mattered most, and as a result each focus group's conversation flowed more naturally toward the strengths and weaknesses of those that attended.

6.3 Outcomes

The following sections highlight the areas of discussion and relevant findings and actions that emerged during the talks. Some are directly related to the structured conversation, but others came out of discussion on specific points made by case study authors during the elevator pitches.

6.3.1 Focus Group 1

The following sections outline the discussion and summary of Focus Group 1.

6.3.1.1 Case Study #1: Cosmic Frontier Subprogram — Cosmology Computation and Simulation

The case study authors outlined the fundamental purpose and approach to the creation of cosmological simulations through the use of HPC resources. The products of this research are used for a number of other experiments (some being profiled by the requirements review program), as well as others that are funded through other agencies such as the NSF. The core products of simulation serve as inputs to other scientific experiments, helping to create algorithms, test theories, and select and implement technology that analyzes, transforms, stores, and shares information worldwide.

A fundamental friction that participants in this work face is meeting the demand to efficiently store and share results over time. Surveys that are created may have a life cycle that spans decades, implying that a central location to store, search, and share results would be highly desirable. The lack of such a facility is related to the funding environment: projects have a set life cycle that does not facilitate storage beyond a certain event horizon, and the use cases that span funding agencies are hard to predict and plan for in terms of budget.

The case study authors have a trajectory to continue to utilize large HPC facilities for creation and storage of surveys, and will see data growth as the intricacy and magnitude of surveys they produce must increase to match the observational experiments that consume their products. To date the transfer of these products between HPC facilities has improved and is now routine, but "downstream" sharing is always harder to predict: thus, well-tuned and centralized locations to share are desirable (in addition to the aforementioned reasons that were cited regarding data set value over time).

6.3.1.2 Case Study #6: Cosmic Frontier Subprogram —LZ Dark Matter Experiment

The LZ Dark Matter Experiment is located at SURF in South Dakota, and is managed primarily by LBNL. 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 a long 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 of LZ's technology workflow has been developed and deployed using container technology (CVMFS), which gives a layer of protection and redundancy to cope with resource constraints that may exist at NERSC due to maintenance.

6.3.1.3 Case Study #8: Intensity Frontier Subprogram — Belle II Experiment

Belle II is a third generation "B meson" experiment that is located at 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 that are 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.

6.3.1.4 Case Study #11: Energy Frontier — CMS Experiment

The CMS experiment is one of two particle-physics detectors built on the LHC at CERN. Currently the facility is in a planned "long shutdown" through 2022 for upgrades, and then will enter into a running period (Run 3) for a number of years. Another shutdown period will proceed to the last run (Run 4) which is tentatively scheduled for 2028 (accounting for delays due to the current pandemic). This later run will usher in the era of HL, which will increase data sizes by orders of magnitude beyond the prior runs, and those of the upcoming Run 3. CMS as a collaboration is focused heavily on research efforts to cope with the data demand, and is constantly looking into new ways to improve the core components of the research workflow (analysis, simulation, data sharing).

It is expected that upcoming software will be adaptive to the challenges of the increase in data volumes both by trying to use new file formats that are compact, as well as leveraging both streaming and bulk-data movement approaches to cleanly and efficiently use network resources. Computation has traditionally followed a grid-computing model that is distributed worldwide at hundreds of sites, and will continue to do so into the future. Emerging use cases to leverage HPC facilities are very attractive, provided that some fundamental areas of friction can be addressed: porting of software, availability of network resources to support streaming workflows, and allocation of cycles that can be tied to the timelines of experimentation.

6.3.1.5 Group Discussion

The polling during the meeting produced the following discussion topics that were of interest to the assembled group:

- Use of HPC (ASCR facilities, or other) resources in HEP workflows.
- Long-term data management (central, distributed, tools).
- Future networking requirements (capacity, traffic expectations, services).
- Use of HTC (OSG, etc.) resources in HEP workflows.
- Data sharing tools/approaches (central vs. distributed management, ad hoc).

During this period of discussion, several notable items were brought up:

- Certain communities, such as those affiliated with cosmology simulation production and sharing, have identified a key gap to the long-term success of their work: a lack of a long-term solution for storage and curation of simulated sky surveys. The current environment in which they operate is to utilize storage that is affiliated with major DOE HPC facilities where they have computing allocations. Research products are created, stored, curated, and shared from these single locations in this model, meaning that when funding concludes it is necessary to work out alternative arrangements for storage allocations for critical data sets that are still valuable to the scientific community. This is a common problem in the community, and has resulted in many different surveys (some of which remain very desirable over time) being located in different locations. It is hard to gauge interest over time; thus, older surveys still may have value. This problem is compounded by the funding source (NSF, DOE). Thus creation of a single long-term location to store PBs (scaling to potentially EBs in the coming years) of old survey products will become challenging over time. The adoption of certain services that ESnet is investigating (caching, etc.) may help the distribution problem, and there is a potential to leverage cloud solutions. The long-term home for data of this form (centralized or distributed through a uniform portal interface) will require careful coordination within DOE, and potentially with partner agencies.

- Availability of computing and storage resources can sometimes challenge experiments that are centrally located in their design. Some collaborations (like DESC) rely on a single source (NERSC). When things are operating fully, the environment works as designed. When there are downtimes (either planned for maintenance, or unplanned due to disasters), the progress of science is affected. This area of friction is a known problem. As the number of users at some facilities increases, the availability of spare resources decreases, and unexpected events may affect overall availability. Discussion in this area centered on portability of workflow (through the use of software containers) that may facilitate deployment on other resources at other facilities, or different portions of the same facility. "Fate sharing" between DOE HPC facilities is something that individual experiments may build into their software capabilities, but it is not something that is discussed at the facility level. DESC, LZ, and the Cosmology Simulation community (as users of DOE HPC facilities) would strongly desire to build redundancy into their systems to cope with scheduled and unscheduled downtime, provided there were mechanism to enable migration of data and re-staging of computation cycles.
- CMS uses HPC facilities for simulation workflows currently, but is exploring wider use for analysis and would desire ways to treat the resource like they do for other grid-computing workflows. Ensuring that the HPC facilities are prepared for this use case is a core part of their development road map, and involves several considerations, such as ensuring that network connectivity is widely available to support their data volume needs, along with end-systems that can be used to stream data, or ways to mitigate this through intelligent staging via bulk-data movement mechanisms. The computing model for CMS (and ATLAS) will still follow the grid paradigm, but will leverage large HPC facilities that can support the software that is being developed and enhanced for the HL era.

6.3.2 Focus Group 2

The following sections outline the discussion and summary of Focus Group 2.

6.3.2.1 Case Study #2: Cosmic Frontier Subprogram — DESC

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 (to be named) 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.

6.3.2.2 Case Study #9: Intensity Frontier Subprogram — Neutrino Research at Fermilab (DUNE at LBNF and the SBN Program)

The case study profiles two aspects of the neutrino research program at Fermilab: DUNE at LBNF and the SBN Program. Both focus on the study of neutrino oscillations, and use a similar set of scientific technology for observation, as well as supporting the computation/storage/networking approach. The work of SBN 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 wide-area network 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.

6.3.2.3 Case Study #10: Energy Frontier Subprogram — ATLAS Experiment

The ATLAS experiment is a general-purpose particle detector experiment at the LHC, a particle accelerator at CERN in Switzerland. The process of science is to run particle-on-particle collisions to validate and understand SM physics. ATLAS and the CMS experiment form the core of US involvement in the international LHC efforts. Currently the LHC is not in operation (in Long Shutdown 2 until 2022), meaning that most experimental activity is focused on R&D of new software, reprocessing of old data, and simulation to prepare for Run 3.

LHC collisions and ATLAS observations can be broken down into MB of raw data (events) that are captured and stored on tape archives at CERN, and distributed in portions to Tier 1 facilities around the world: BNL in the United States is the ATLAS Tier 1. From the raw data, there are derivations that produce smaller event sizes, that are then grouped into data sets that are shared around the world for analysis on the WLCG operated at Tier 2 facilities. It is estimated that Run 3 will double the amount of data generated versus prior runs. Thus R&D efforts to handle the increased data load are well underway. In the coming years, the LHC will enter into a shutdown again to prepare for the HL era of operation, which will further increase data requirements at all levels (storage, computation, and networks).

6.3.2.4 Group Discussion

The polling during the meeting produced the following discussion topics that were of interest to the assembled group:

- Future networking requirements (capacity, traffic expectations, services).
- Analysis workflows: use of local versus remote versus distributed resources.
- Use of HPC (ASCR facilities, or other) resources in HEP workflows.
- Long-term data management (central, distributed, tools).
- Use of HTC (OSG, etc.) resources in HEP workflows.

During this period of discussion, several notable items were brought up:

- Across experimentation in HEP (and even beyond), it is hard to center on a succinct definition of the term "data set." This exacerbates the job of the research groups, as they try to accurately depict this unit of measurement for groups that provide computation, storage, or networks. There are also complications of measurements done by producers (e.g., Rubin Observatory) versus consumers (e.g., DESC), who will have different views of the data set sizes that are directly related. The ATLAS definition of a data set is closely tied to the operation of the LHC device and detectors: eight hours of continuous beam operation will produce an entire data set. That data set consists of all of the event files related to the particular run, and may top TBs of raw data. While there is no need for this group to produce a more succinct definition, all the parties understand the core requirements that ESnet wishes to gather: "data usage over time" will give a uniform baseline for potential network use.
- Many of currently running experiments have adopted the approach of "any data anywhere," which is shorthand notation for being able to locate, download, and perform analysis on experimental data wherever capabilities exist. This could mean using widely deployed tools to orchestrate the download and compute at a Tier 3 site from storage resources located elsewhere, or it could also mean allowing the tools to pre-stage data to locations with an abundance of storage and compute, so that users can leverage those resources. For either use case, it becomes desirable to leverage networks to assist in the dissemination of data, as well as use intelligent software paradigms (such as caching) to reduce the overall amount sent via networks when possible. Discussion of caching approaches reveal that "blind caching," i.e., making guesses of data that may get reused without context on prior usage or understanding of value from

analysis workflow systems, can only go so far. LHC R&D has focused heavily on ways to inject intelligence and hints when possible through monitoring how workflow/analysis systems operate on user requests. ESnet will continue to work with interested parties in this area, as caching will be available in the ESnet6 network.

- Some experimentation will migrate away from distributed grid approaches, and leverage large storage and compute allocations at core facilities. DESC is taking this approach for data that come from the Rubin Observatory. After DESC acquires data sets, most (if not all) analyses will occur at NERSC. Mechanisms to perform analysis on the entire data set will occur there, along with discretionary compute and storage resources that will be made available to collaborators. There will be ways to download and compute resources elsewhere, but the hope is that by providing all necessary components in a single location, there will be less migratory use.
- Splitting data sets between major collaborators poses some challenges, although the use of automated software mechanisms to catalog and distribute makes the job easier. ATLAS is accustomed to the mentality of splitting data sets (raw, AOD, etc.) around the world. DESC will rely on a single source (Rubin) but plans to keep everything at NERSC (with a backup at IN2P3). DUNE is evaluating approaches still, and anticipates using a majority of resources at Fermilab, but could leverage collaborators around the world. The tipping point for most experiments is storage space, storage longevity, and distribution of computational load. Operational overhead increases as distributed partnerships grow, which is something for new experiments to consider.
- HPC facilities remain attractive for some use cases (DESC), but still are not a major part of the workflow for distributed models (ATLAS, DUNE). Beyond functional considerations, such as the type of codes used for processing, there are considerations for long-term storage, and the broader user community to consider. Some HPC aspects are appealing: e.g., the ability to reprocess or simulate in a batch environment would speed up portions of experimentation. There is a desire that future HPC architectures can also work to support streaming data flows more efficiently. HTC/grid work has its own set of desirable features, namely the ability to scale the required resources up/down as needed during the course of experimentation.

6.3.3 Focus Group 3

The following sections outline the discussion and summary of Focus Group 3.

6.3.3.1 Case Study #4: Cosmic Frontier Subprogram —the Rubin Observatory

The Rubin Observatory, previously referred to as the LSST, is an astronomical observatory currently under construction in Chile. Its main task will be 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 CCD imaging camera. For the purposes of the DOE, there are several dark energy experiments (notably DESC) that will utilize data produced by Rubin on a yearly basis. The COVID-19 pandemic has stopped some progress, namely the physical construction at the site. Work on the camera has proceeded, with some promising early results in a laboratory environment at SLAC.

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, through local data storage facilities, to the USDF at SLAC. 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. An interim DF is planned using the GCP, starting in FY21, to begin to test software for analysis, as well as operational aspects.

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 and processed using the analysis toolchain; and the processed data then to be made available through a series of brokers that will distribute the data to interested parties. Therefore, 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 most of the year processing raw data for a yearly data release. This release will then be made available to collaborators and the general public. Rubin will follow a model of "bringing people to the data" and will make and 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.

6.3.3.2 Case Study #5: Intensity Frontier Subprogram — 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 LBNL being the lead institution on the DOE side, and the University of Chicago leading for the NSF. When complete, there will be three large and 18 smaller telescopes deployed between two sites: the South Pole and Chilean Atacama Desert. Each site has a specific use case:

- South Pole will specialize into 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 (i.e., 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. There are therefore efforts to ensure that operation can proceed with limited (or severed) resources, with goals of increasing the available connections where possible.

6.3.3.3 Case Study #12: Energy Frontier Subprogram — LHC Operations

The LHC Operations case study is jointly prepared by members of the ATLAS and CMS experiments. The purpose of this case study was to provide perspective on the current and future term use of common experimental features: software, network infrastructure, computational facilities, and joint R&D activities. A separate case study will focus on specifics related to the HL-LHC era.

As explained in the individual case studies, both experiments use the LHC as the major instrument coupled to their detector hardware. Operation adheres to an operational schedule that are split by scheduled "shutdowns" to facilitate upgrades and maintenance:

- Run 1: 2009–2012.
- Shutdown 2: 2012–2014.
- Run 2: 2014–2018.
- Shutdown 2: 2019–2021.
- Run 3: 2022–2025 (est.).

During a shutdown, there are still extensive data exercises that take the form of simulation and reprocessing, as well as R&D on data formats, software infrastructure, computational approaches, and general operational preparedness. The major technology components that assist both experiments are:

- The WLCG, a constellation of grid-computing sites located at the various tiers of the collaborators (e.g., Tier 0 at CERN, Tier 1s that are typically country-scale computing facilities, such as BNL for ATLAS and Fermilab for CMS), and Tier 2s that are regional-scale facilities funded by the NSF and operated by university partners). Tier 3s are also utilized for user-level analysis, but are not funded or deemed a critical part of the core mission of scientific production.
- LHCONE, a collaborative effort to create an "overlay" network between major LHC computing facilities. This enables high-performance access and management of traffic that is designed to isolate and prioritize traffic related to LHC operations. CERN, the Tier 1s, and most of the Tier 2s are connected to this via their regional network.
- A set of shared software packages, developed and maintained through the OSG and some other collaborators. These include data management tools, workflow orchestration, data movement, and monitoring. Current focus areas include the use of automation, understanding network use, and ways to better utilize computational, storage, and networking resources more effectively as the data sets will grow in size and volume in the coming years.

Each experiment has its plate full of priority work that requires assignment of the limited manpower available, making it hard to gain additional effort to explore, prototype, and test significant revisions to the experiment's operational infrastructure. There are open questions on how best to evolve networks (e.g., LHCONE), especially regarding access to commercial and opportunistic resources (e.g., commercial clouds) that do not have access to such networks. To push us forward, we are planning a series of incremental challenges that highlight the use of the network among our existing and potential future resources. The ability to leverage HPCs, FABRIC, LHCONE/LHCOPN, ESnet, Tier 1s, and Tier 2s are all being considered during the challenge phase.

The LHC is expecting that HL-LHC will require the following performance characteristics at the various tiers by 2028 (with early data trials requiring less, but with plans to test and characterize):

- Tier 1s: 200 Gbps, with bursts that could reach 1 Tbps.
- Tier 2s: 100 Gbps, with bursts that could reach 400 Gbps.

Using available data sets, estimates show that there is still a 4x gap between what is needed and what is currently available in terms of networking capacity. The experiments expect 40–60% increase in data volume, which equates to 2x every two years. Due to these expected increases, these groups are looking at many concurrent mitigations:

- Reducing analysis-format sizes to bring down pressure on sending duplicative information where applicable.
- Adding additional intelligence into tools to transfer from more topological friendly locations (e.g., closer, to not utilize long-haul bandwidth unnecessarily).
- Researching ways to leverage caching.
- Looking into pre-staging data (e.g., data lakes) to assist with both data access and network transfer requirements.

No single solution will fix the problem, which is why the R&D efforts now are critical to the future success.

6.3.3.4 Group Discussion

The polling during the meeting produced the following discussion topics that were of interest to the assembled group:

- Future networking requirements (capacity, traffic expectations, services).
- Long-term data management (central, distributed, tools).
- Data sharing tools/approaches (central vs. distributed management, ad hoc).

During this period of discussion, several notable items were brought up:

- The use of commercial clouds within science projects is still limited, but growing as some experiments and projects have conducted testing and are planning for deployments. Costs remain high, which is the largest barrier to adoption. The LHC experiments (particularly CMS) have performed several R&D activities to understand the impacts of network, computation, and storage. Published results[1] have attempted to characterize a number of performance characteristics (ingress and egress data performance, intra-cloud transfers, managing "cloud bursts" for time-time critical applications) which could be useful for future HEP experiments looking to utilize clouds. Other experiments, such as Rubin, are budgeted to utilize cloud services both as an interim solution until their USDF is named (e.g., as a testing platform for software and operations) as well as a continuously used resource that will house project-internal tools.
- A critical component to the success of LHC operations (domestically and internationally) is the use of TA networking. A number of connectivity options are in place today provided by different funding sources. The DOE and ESnet have dedicated multiple 100 Gbps paths to support LHC and other science between the European Union and United States. The NSF funds several links that are general purpose for the R&E community. Other consortia of R&E operators have also collaborated to ensure fate sharing and peering arrangements that make capacity available in the event of link maintenance and failure. Additional capacity is on the roadmap, and prices are dropping significantly in this space (and will continue to do so).
- To ensure equitable use of TA bandwidth, changes will be required to the operational approach of some of the software. For instance, it is possible now through the any data, anywhere model that exists for a Tier 2 facility in the United States to request data sets that may exist in Europe, thus triggering an international data flow that could be large (far greater than 10 Gbps) and could consume significant resources. If this becomes more regular as Tier 2s grow to burst beyond 100 Gbps, bandwidth resources will dwindle and affect production use cases that are constantly staging data from the European Union to the United States. Thus, three main pushes are needed in this space: (a) introducing more compact analysis formats to reduce the required transfer sizes, (b) securing more bandwidth to ensure experimental readiness toward the growing data set size and volume increases, and lastly (c) altering the data staging approaches to leverage more intelligent methods (caching, staging into data lakes, facilitating fetching from more geographically relevant locations).
- While R&D is encouraged to make better use of resources, there will always be capacity available to meet the science mission. The bottleneck may migrate from being the TA piece, and could end up being pushed closer to the Tier 2 and HPC facility layer (e.g., via the US regional and campus networks that connect the computing and storage facilities). Programs like the NSF CC*[2] efforts to upgrade campus infrastructure are critical to ensuring success of science

[1] https://arxiv.org/abs/2002.06667, https://arxiv.org/abs/2004.09492, https://arxiv.org/abs/2005.05836, https://arxiv.org/abs/2002.04568
[2] https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504748

programs, and should be encouraged as a way to help upgrade infrastructure. ESnet will also continue to work with DOE-connected labs and computing facilities to ensure that capacity is not a factor on the wide area, but has minimal control over local area handling of networking to computation and storage infrastructure.

- Connectivity and peering to cloud providers remain important during the R&D lifecycle of experiments such as the LHC and Rubin. Both are still trying to understand the impacts of using these resources: the performance, the costs, and the predictability. The LHC in particular wants to work with ESnet to better understand architectural implications of software and workflows that egress/ingress ESnet from cloud providers. Being able to predict the timing and performance implications are critical, because intelligent workflow managers may use that information to better influence where to schedule resources.
- Some R&D efforts are reaching a maturity stage that requires more coordination with network providers to study effectiveness. The work to institute "packet marking" (i.e., an approach that manipulates the network traffic sent by an application, and can be read/understood by monitoring systems and intelligence network controls) requires coordination on more pieces of the wide-area path. It has been tested by sites, but would need wider deployment on ESnet/LHCONE/other locations to get more experimental results. The LHC community is far along, and seeks to partner with ESnet on near-term timescales regarding this work. Given the potential to affect operations, care must be taken to ensure that risks and mitigations are understood, and that all sides are sharing these. Any research project would have to begin with understanding resources required, what is being offered, milestones, and some other factors. These choices will affect networks like ESnet, as well as the experimental leadership for ATLAS and CMS.
- A common observation from experiments that have traditionally not utilized HPC facilities, but would like to understand how they can in the future, is lack of a clear and understood way to access these resources in a uniform fashion. For example, a workflow that utilizes grid/HTC approaches has a common API and usage pattern no matter where it may run. This is not the case for work that may be launched at NERSC, ALCF, or OLCF unless those facilities are making the same type of interface available. ESnet is in a unique position to work with the HEP experiments and the LCF facilities to open a dialog about ways this can be managed and improved in the future.
- All experiments (CMB-S4, Rubin, and the LHC) have a so-called "2-sided" workflow they manage: instrument to first level of computation/storage, and then a second that involves a network to more plentiful/high-performance storage (the later could be fully distributed on a grid, or to a single HPC/DF). From there, additional "fan out" of data to users may be possible. Automation of these two different sides has increased significantly in recent years in the LHC use case, and is expected to heavily influence the CMB-S4 and Rubin use cases. Both experiments indicate that the handling of the instrument-to-local data flow will be almost fully automatic, with enough captured telemetry to assist any human intervention in the event of a problem. Local-to-remote is anticipated to be almost fully automatic as well, but will rely on monitoring to understand the probability of success due to the limited (and occasionally chaotic) availability of bandwidth. In some cases, limited bandwidth resources could force buffering for CMB-S4. Rubin may choose to hold off or slow the regular data stream during times of limited network bandwidth to ensure that high-priority operations (e.g., transients) have enough resources to function fully. Both experiments believe it will be possible to share best practices on approaches, but common tooling (hardware, software, cybersecurity approaches) may not be shared to support these use cases.

- Long-term data management for emerging experiments (Rubin, CMB-S4) is still under discussion. Both plan on making sure that old data sets remain available through some form of portal, and anticipate making the data available in highly compressed formats to reduce size and transfer requirements. CMB-S4 anticipates that "on project" work, i.e., those pieces of research directly funded, will dominate the need for data use/transfer, with smaller volumes of "off-project" use. Rubin anticipates making data sets available from the single (or small number) of main DFs, but also anticipates that the majority of use cases will compute on/close to these data using the available tools and allocations. The major consumers (such as DESC) are communicating their requirements early and often. Thus there will be mechanisms in place for regular bulk-data movement of the data set when it is released on a yearly basis. Both projects anticipate keeping active data available through fast-access (i.e., disk) media when applicable, and will fall back to tape when space is exhausted. Older catalogs will lose value but will not be deleted.
- Rubin will be entering into operations soon, and is seeking advice from LHC on the nimbleness that is required with respect to data movement and user patterns. The LHC offers some advice, namely that it did not restrict what users would do to start, and as a result was able to observe that some estimates on behaviors were below what was expected. This did not overtax what was available, but was helpful in adjusting the tools used for future planning to be sure there were ample resources to keep up with demand and data growth patterns.

6.3.4 Focus Group 4

The following sections outline the discussion and summary of Focus Group 4.

6.3.4.1 Case Study #3: Cosmic Frontier Subprogram — DESI

DESI is a scientific research instrument for conducting spectrographic astronomical surveys of distant galaxies. It will utilize the Mayall Telescope (a four-meter telescope), located at KPNO near Tucson, Arizona.

The overall process of science is focused on creating a 3D 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 approximately 700 MB for an image, which are 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 collaborations. 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. 10 Gbps exists today as provided by KPNO, although upgrades and redundancy are stretch goals. The experiment has the ability to buffer data when connectivity is lacking through the use of some local computation and storage and a workflow manager that is controlled at NERSC.

DESI expects a model similar to other astronomical experiments, where most (if not all) user analysis will be done at the data's 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 the instances where that is required, DESI will leverage existing NERSC infrastructure (DTNs 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.

6.3.4.2 Case Study #7: Intensity Frontier Subprogram — Muons Research at Fermilab (Mu2e and Muon g-2)

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 SM.

Muon g-2 is currently operating at Fermilab, and has finished Run 3 of a planned five runs (with expected end time in 2022). 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 AI/ML, both of which may influence future operations for Mu2e.

Mu2e is under construction, and will go into operation in 2024 with a five-year run cycle. It is expected that it 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. The majority of cycles will be provided by Fermilab, with some use allocated to other participating sites (a minority of the expected computation and storage power).

The use of HPC resources is not currently large, although the workloads would convert to the use case if there were resources to convert and adapt software (at the current time, this is not a high priority).

6.3.4.3 Case Study #13: Energy Frontier Subprogram —HL-LHC Research

The HL-LHC begins approximately 2027 to 2028 with current estimates. This case study was jointly prepared and presented by members of the ATLAS and CMS experiments. The overall purpose of this case study was to provide perspective on the implications of the data volume growth and how current R&D efforts (into storage, computation, network use, and software) will influence the operational preparedness of the experiments. It is anticipated that there will be an increase of 5x more events to manage, and an event size increase that will range from 5–10x more than current Run 3 observations.

It is anticipated that the various LHC computing tiers will have access to increases in technology (larger networks, more storage, more and after CPUs and GPUs). As a result of this, there are several questions to answer:

- How can the experiments better measure and project needs (particularly those related to networking domestically and internationally)?
- How will the changes to the underlying technology change the outcome of the science process?
- When will the technology be available, and how can it be incrementally adopted?
- How can national providers, like ESnet, become more integrated into the process?
- How can emerging R&D use cases and ideas be prototyped and tested in the wide area?

As discussed in other LHC case studies (ATLAS, CMS, and LHC Operations), the availability of network capacity is a core concern:

- TA capacity that links CERN to the US computing facilities (Tier 1s and Tier 2s).
- Domestic network capacity on ESnet that links Tier 1s and Tier 2s.
- Domestic network capacity on US R&E networks for Tier 2s.
- HPC facility ingress and internal networking capacity.

In addition to raw capacity, efforts to become better users of networks involve:

- Reduction in size of data sets that must be shared (without decreasing fidelity of what they represent):
- More intelligent ways to access data closer to where it is needed, either staged or on demand.
- Reuse for common data sets over time via caching approaches.
- More accurate accounting of what is being used, by whom, and when.
- Improvements to protocols and tools.

6.3.4.4 Group Discussion

The polling during the meeting produced the following discussion topics that were of interest to the assembled group:

- Use of HPC (ASCR facilities, or other) resources in HEP workflows.
- Future networking requirements (capacity, traffic expectations, services).
- Analysis workflows: use of local versus remote versus distributed resources.
- Data sharing tools/approaches (central versus distributed management, ad hoc).
- Long-term data management (central, distributed, tools).
- Use of HTC (OSG Software, etc.) resources in HEP workflows.

During this period of discussion, several notable items were brought up:

- HPC resource use by some users (in particular the muon and LHC experiments) is desirable, but has two barriers: network capacity for expected data needs and inability to support streaming use cases. For the LHC experiments, there is concern that data ingress, as well as internal data architectures, will not keep pace with the volume of data that is expected by the HL-LHC. There are years to address this, but these facilities will not be utilized if they become a bottleneck versus more well-connected grid sites that can provide similar if not better performance. The second factor, ability to handle streaming workflows, is of equal importance due to the current architecture (and expected future use cases) that leverage grid resources and the sometimes-rare commodity of long-term storage. Storage remains a core requirement for use cases like the LHC, and because it can be expensive to acquire and maintain long term, streaming workflows became more common because they could function well on computation and fast networks without requiring lots of local storage. The LHC software stack adopted streaming of data to computation resources in an on-demand fashion as a common use case (bulk-data movement still exists, but is less common in grid environments) because it helped scale to resources more easily. An HPC resource typically does not allow worker nodes to make repeated call outs to the wide area, thus fetching data on demand is made harder. There are workarounds that include pre-staging of data to semi-local sources (e.g., data lakes) that are showing promise, as well as architecture changes to future HPC architectures (e.g., the Cray Slingshot) that will facilitate more use of WAN during computation.
- Software development for scientific use cases remains a challenge. There are two main approaches: using software that is developed/supported by others for the same or similar use cases, or attempting to write one's own (either funded by a project or unfunded). The former is encouraged, and recommended by all represented parties (DESI, Mu2e, Muon g-2, and the LHC experiments); all utilize aspects of the OSG software stack or other pieces that are used/supported by the facilities they utilize (sometimes with minor modifications for use cases that may differ). The latter is not recommended, but sometimes must occur. This can be destructive

for an experiment, because if there was no original budget for software, there typically will not be knowledgeable staff available to steer a successful creation of usable packages. All recommend that software become a first-class citizen for the planning process going forward, as software is now deeply tied to the use cases of computation, storage, and networks.

- Conversion of software to use one major computing paradigm (HTC/grid versus HPC) is problematic. Not only is it time consuming to rewrite for a different use case, it is often done as a last resort that is not funded. In many cases, projects are not budgeted for software; thus using what is available is a first approach, and in rare cases new software may be created to fill gaps. In the latter case, creation of hard to support/non-battle tested tools results, which can make the overall success of the research suffer. Investment to use one or the other must be chosen carefully.
- Network capacity concerns for the future are borne out of the current observations that experiments like those in the LHC are making based on past use, growth patterns, and anticipated outputs. A common pattern that is observed involves a network plateau before a run (associated with lack of live experimental data, but a steady state of reprocessing and simulation), followed by slow growth during (live experimental data, in addition to any reprocessing or simulation), and lastly a new plateau established after. Given the LHC runs are scheduled and regular, this gives a natural three- to five-year cadence to the patterns and allows basic forecasting.
- Experiments like DESI are still grappling with what will be required to support data sharing. Analysis formats make this harder, as there is a desire to ensure that the unit of analysis contains enough information to be useful, but is compact enough to be shared. Once a format is created, there are the issues of the tools used, hardware required, and how it all interacts over the wide area. ESnet is in a unique position due to the close relationship we have to the facilities that share data. Thus, we are a part of the data-sharing equation, and try to encourage the use of intelligent tools and systems to simplify data sharing. Dedicated AFs are a part of this, are used in some experiments already, and are being investigated by others. This would create well connected and supported facilities with the only job of ingesting and egressing large amounts of data directly to ESnet and its connected resources and peers.

Appendix A – International Connectivity

Throughout the 2020 HEP Requirements Review process, the case for international networking needs has come to the forefront to support nearly every case study for aspects of the workflow. These needs can categorized as follows:

- **Instrument/detector source and distributed AFs:** scientific instruments, such as particle accelerators, telescopes, etc., have a single source, and often rely on an AF/AFs that are physically separated. Global collaboration often means that international networks are a critical part of the process of science.
- **Intercollaboration information sharing:** other portions of the scientific workflow (distributed analysis on intermediate formats, production of simulation data, backups, etc.) may involve international collaborators.
- **User-level data sharing:** users of scientific data are worldwide, and are not always known a priori.

The following sections will highlight specific findings from the review, along with supplemental information on international connectivity from the R&E community. Some of the links are funded via the DOE (e.g., ESnet); others come from the NSF and foreign collaborators (e.g., GEANT, RNP, NORDUnet, etc.).

A.1 Current State and Near-Term Plans for the International R&E Circuits

International connectivity for the R&E community is provided by a number of different providers and funding sources, and is delivered through several exchange points located around the country. These facilities feature connectivity to domestic R&E and commercial carriers, which link many of the HEP facilities.

A.1.1 Domestic Exchange Points

There are number of domestically located exchange points where network providers establish peering with each other. This fabric of connectivity allows for a seamless transfer of scientific network traffic between cooperating providers:

- MANLAN: New York, New York[1].
- WIX: Washington, DC.
- Starlight: Chicago, Illinois[2].
- Pacific Wave, Los Angeles, California, and Seattle, Washington[3].
- AMPATH: Miami, Florida[4].

ESnet maintains connectivity to these locations, as well as peering with providers that are present, to ensure that traffic can reach critical international locations.

A.1.2 TA Networking

As of December 2020, there were nine 100 G circuits, providing an aggregate of 900 G of R&E capacity, between the United States and Europe as shown in **Figure A.1.** These links are supported by the DOE, NSF, Internet2[5],

[1] https://internet2.edu/network/global-networks-and-partnerships/man-lan-new-york-and-wix-virginia-exchange-points
[2] http://www.startap.net/starlight
[3] http://pacificwave.net
[4] https://ampath.net
[5] https://internet2.edu/network/global-networks-and-partnerships

CANARIE (Canadian National Research and Education Network [NREN])[6], GÉANT (European NREN)[7], SURF (Dutch NREN)[8], and NORDUnet (Nordic NREN)[9]. During the last quarter of 2020, these links averaged 18.5 Gbps across the suite, and transferred over 1.8 PB of data. Many of these networks collaborate regularly through established consortia[10,11].

In early January 2021, an additional 100 G link will be added between New York and Copenhagen, with backhaul to Amsterdam, provisioned by the NSF and the Networks for European, American, African, and Arctic Research (NEA3R) project[12]. Current plans are to renew the other existing US–EU circuits as needed, while keeping an eye on the used capacity. The biggest addition/adaptation will be a possible 100 G circuit from the Nordic region to Japan, currently called Arctic Connect, which has an earliest operational date of 2023.

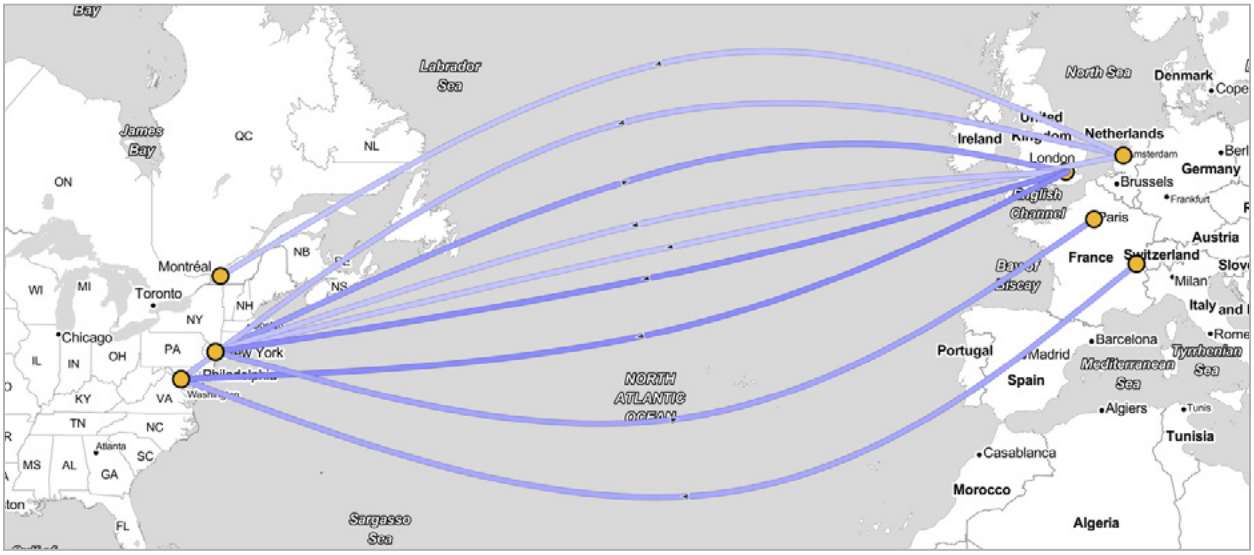


Figure A.1: Current R&E networks between the United States and Europe. Data available live at http://ana.netsage.global

A.1.3 Transpacific Networking

In Asia, the Asia Pacific Ring (APR) Consortium jointly supports connectivity (shown in Figure A.2) for roughly 400 G of capacity between the United States and Asia as well as 10–20 G between Guam and Singapore and Guam and Hong Kong. In late 2020, the SingAREN/Internet2 link between Singapore and Los Angeles was replaced by a SingAREN-managed circuit that runs between Singapore, to Tokyo, and then to Los Angeles (on a different cable than the SINET Tokyo-LA capacity)[13,14]. In early 2021, it is expected that the path between Guam and Singapore will be upgraded to 100 G. Depending on Federal Communications Commission regulators, the Guam–Hong Kong and Sydney–Hong Kong paths may be upgraded to 100 G in 2021 or 2022 as well. Currently, these links are underutilized, but the diversity of paths is needed for redundancy and resilience in the earthquake and tsunami-prone Ring of Fire region.

[6] https://www.canarie.ca/about-us
[7] https://www.geant.org/Networks
[8] https://www.surf.nl/en
[9] https://www.nordu.net
[10] https://internet2.edu/network/global-networks-and-partnerships/advanced-north-atlantic-ana
[11] https://gna-re.net
[12] https://in.iu.edu
[13] https://www.singaren.net.sg
[14] https://www.sinet.ad.jp/en/aboutsinet-en

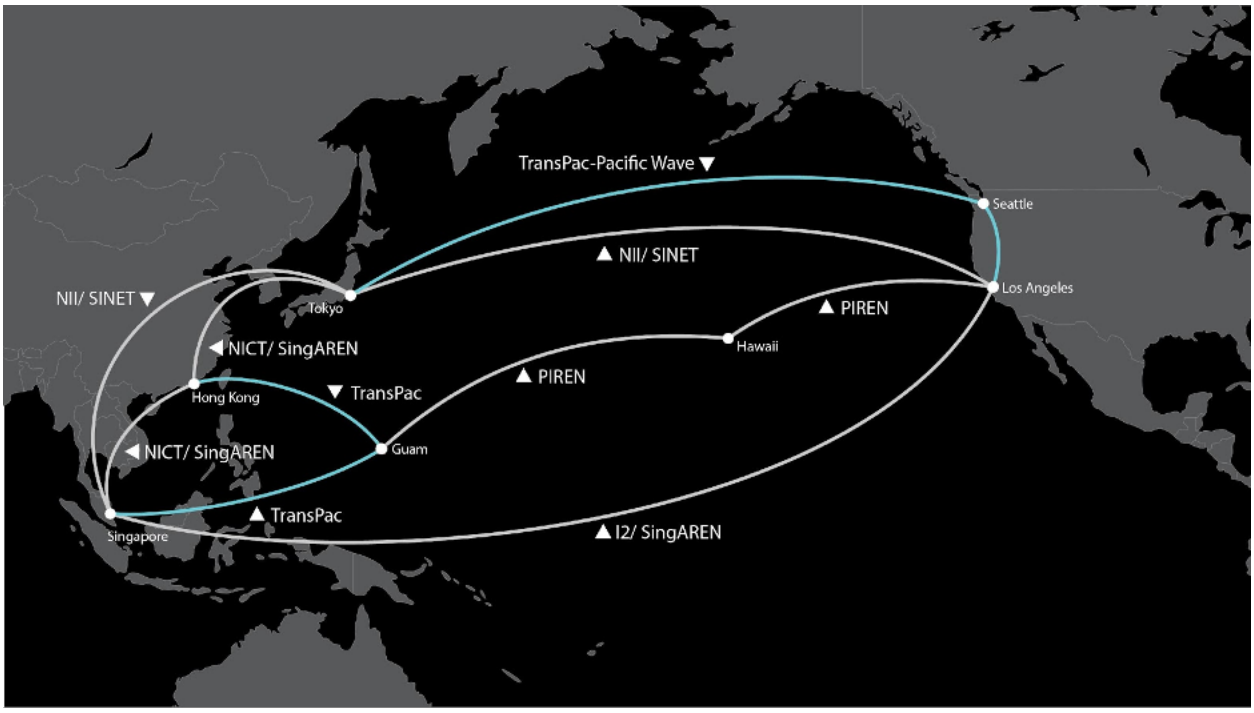


Figure A.1: Current R&E networks between the United States and Europe. Data available live at http://ana.netsage.global

A.1.4 South American Networking

Between the United States and South America, R&E networking is primarily supported via an NSF IRNC award to Julio Ibarra entitled "Americas-Africa Lightpaths Express and Protect (AmLight-ExP)"[15]. Figure A.3 shows the current (2020) production circuits, consisting of the 400G Express spectrum (green) and the 200 G Protect leased (red). There are plans in the next three years for Rede Nacional de Ensino e Pesquisa (RNP)[16], the NREN for Brazil, to activate 200 G between Fortaleza and São Paulo (blue) and for RedCLARA[17], the Latin American NREN, to begin to support an additional 2x100G capacity between Fortaleza-Portugal on the new Express optical platform between Europe & Latin America (ELLA) circuit (magenta).

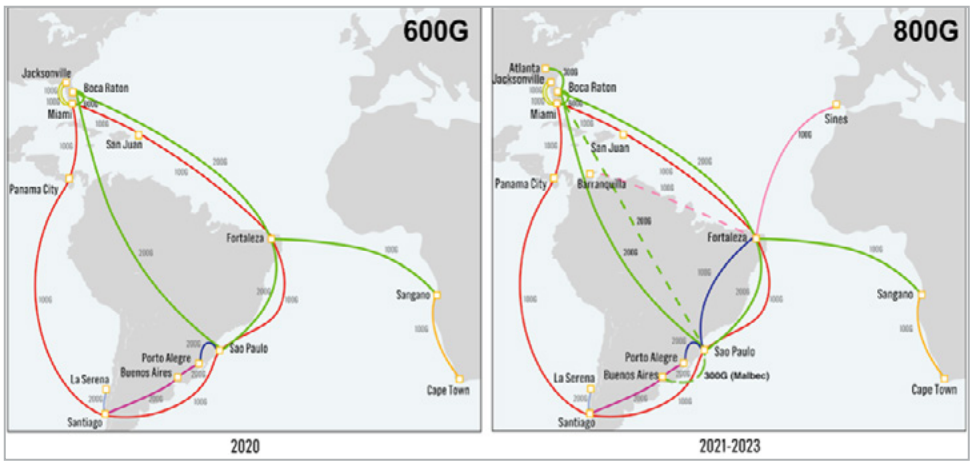


Figure A.3: R&E networks between the United States and South America

[15] https://ampath.net
[16] https://www.rnp.br/en
[17] https://www.redclara.net/index.php/en

A.1.5 Polar Networking

Connectivity to the South Pole relies on satellite connectivity at the current time. The US Antarctic Program (USAP)[18], funded by the NSF, operates the South Pole TDRSS Relay 2, or SPTR2, to communicate with the NASA TDRS satellites. NASA makes a great effort to ensure South Pole Station is granted adequate access time to conduct science and operational communications, but cannot always provide a seamless and consistent access schedule. Online times and durations via TDRS vary daily; however, SPTR2 is typically online and providing the South Pole with communications from between two and four hours each day. During this time, a theoretical transmission speed of up to 275 Mbps is possible for some of the high-rate channels.

A.2 Case Study Findings

A.2.1 Cosmological Simulation Research

The case study does not leverage any international locations for the production of simulations; all data products are produced domestically at DOE computing centers, or affiliated US-based universities.

It is likely that users that are based internationally are downloading the simulations, but fine-grained analysis is not available to give specific examples.

A.2.2 DESC

DESC will be a consumer of data produced by the Rubin Observatory, but will retrieve this location from the domestic DF (e.g., SLAC) for use at NERSC. DESC will have an international collaborator in France (IN2P3) that will receive copies of the processed data from NERSC.

DESC anticipates running simulation production code at a variety of domestic and international locations. For now, this is limited to GridPP in the UK, but could expand to other similarly operated computational grids.

Lastly, it is likely that users that are based internationally could download scientific data from NERSC using bulk-data movement tools. The intention of the DESC computing model is to provide computation and storage via NERSC directly, but the ability to transmit data off-site will be supported.

A.2.3 DESI

DESI will not use any internationally located instruments or data sets during operation, but it is likely that users that are based internationally could download scientific data from NERSC using bulk-data movement tools. The intention of the DESI computing model is to provide computation and storage via NERSC directly, but the ability to transmit data off-site will be supported.

A.2.4 Rubin Observatory and the LSST

The Rubin Observatory is physically located in Chile, and relies on a set of international partnerships to deliver high-speed networking capabilities to the USDF located at SLAC. This network includes portions operated by AURA, REUNA, AMPATH, RNP, RedCLARA, FLR, and ESnet along the entire path.

Once the data are housed at SLAC, there will be periodic backups that are sent to a computing partner located in France (IN2P3) via ESnet connectivity.

Lastly, it is likely that users based internationally could download scientific data from SLAC using bulk-data movement tools. The intention of the Rubin computing model is to provide a scientific platform (e.g., computation and storage), but the ability to transmit data off-site will be supported.

[18] https://www.usap.gov/technology/1971

A.2.5 CMB-S4

CMB-S4 will leverage two remote instrument sets that are located internationally: Chile and the South Pole.

As in the case with Rubin, connectivity to Chile relies on a set of international partnerships to deliver high-speed networking capabilities to the USDF located at NERSC. This network includes portions operated by AURA, REUNA, AMPATH, RNP, RedCLARA, FLR, and ESnet.

As mentioned in Section A.1.5, connectivity to the South Pole relies on satellites provided by a number of collaborating agencies. The availability is sporadic, and may allow only for a (shared) window of several hours per day, with minimal network speeds (e.g., a theoretical transmission speed of up to 275 Mbps is possible for some of the high-rate channels). Data are then transmitted back to the United States via facilities in White Sands, New Mexico, and will use terrestrial networks to reach NERSC.

CMB-S4 anticipates running simulation production and analysis code at a variety of domestic and international locations via the OSG infrastructure. Specifics will be defined in the coming years.

Lastly, it is likely that users based internationally could download scientific data from NERSC using bulk-data movement tools.

A.2.6 LZ Dark Matter Experiment

LZ will not use any internationally located instruments or data sets during operation, but it is likely that users that are based internationally could download scientific data from NERSC using bulk-data movement tools. The intention of the LZ computing model is to provide computation and storage via NERSC directly, but the ability to transmit data off-site will be supported.

LZ anticipates using GridPP in the UK as a backup site for storage and processing, but could expand to other similarly operated computational grids.

A.2.7 Muon Experimentation at Fermilab

The Fermilab muon experiments are both operated, and use a computation and storage, from within the Fermi grid infrastructure. The experiments do leverage OSG resources, so the use of other domestic and international resources is possible, typically in the form of streaming data. Both have relationships with international grid resources in Italy (INFN) and the UK (GridPP), which results in international data exchange.

Lastly, it is likely that users that are based internationally could download scientific data from Fermilab using data-movement tools.

A.2.8 Belle II Experiment

Belle II utilizes an experimental facility located in Japan, and is heavily reliant on transpacific network connectivity for the process of science. Domestic and international partners (e.g., SINET, APAN, TransPac, and PacWave) provide networking resources to support a number of scientific use cases, such as Belle II, and interconnect to ESnet, which delivers the traffic to BNL.

Once the data are housed at BNL, there are periodic data exchanges with international partners (e.g., Canada, France, Germany, and Italy), often using the LHCONE overlay network. This relationship is facilitated due to Belle II operations occurring at LHC T1 and T2 facilities in most cases.

Lastly, it is likely that users based internationally could download scientific data from BNL using data-movement tools.

A.2.9 Neutrino Experiments at Fermilab

The Fermilab neutrino experiments use computation and storage from within the Fermi grid infrastructure, but will also leverage OSG resources located domestically and internationally. SBN's streaming needs will be smaller

than those of DUNE when it comes online. It is expected that the use of international resources will resemble other grid-computing use cases that are focused on analysis and simulation, and will take the form of streaming or bulk-data movement. When DUNE reaches full operational use, integration with LHCONE is also possible, to facilitate international data sharing with other well-positioned HEP computing centers.

DUNE's early experimentation (ProtoDUNE) is located at CERN and will operate from there for a number of years. With this instrumentation being remote, a steady stream of data will travel on the TA connections to reach Fermilab.

Lastly, it is likely that users that are based internationally could download scientific data from Fermilab using data-movement tools.

A.2.10 LHC Experimentation and Operation

The LHC operational pattern is well established; the data sources are located at CERN and the experiments send a steady stream of raw data to the Tier 1 centers. In the United States, these are at BNL (ATLAS) and Fermilab (CMS). These data will grow in the coming years, beyond what the available resources of ESnet or the other R&E providers have available currently. Augmenting capacity will be necessary to keep up with this and other use cases.

The LHCONE overlay network heavily leverages international connectivity, and links T0, T1, and T2 facilities to support HEP use cases (LHC, but also Belle II and potentially others). This overlay is not focused on the raw data transmission, but is used to support the exchange of other analysis and simulation formats.

List of Abbreviations

ADC	Analog-to-digital converters, ATLAS Distributed Computing
AF	Analysis facility
AI	Artificial intelligence
ALCC	ASCR Leadership Computing Challenge
ALCF	Argonne Leadership Computing Facility
ALMA	Atacama Large Millimeter Array
ANL	Argonne National Laboratory
AOD	Analysis object data
APA	Anode plane assembly
APAN	Asia Pacific Advanced Network
APR	Asia Pacific Ring
ASCR	Advanced Scientific Computing Research
ASN	Autonomous System Number
ASO	Asynchronous stage out
ATLAS	A Toroidal LHC ApparatuS
AUP	Acceptable usage policy
AURA	Association of Universities for Research in Astronomy, Inc.
AutoGOLE	Automated GOLE
AWS	Amazon Web Services
BGP	Border Gateway Protocol
BNB	Booster Neutrino Beamline
BNL	Brookhaven National Laboratory

CAPEX	Capital expenses
CASTOR	CERN Advanced STORage manager
CC	Campus Cyberinfrastructure
CCD	Charge-coupled devices
CDB	Conditions Database
CI	Cyberinfrastructure
CMB	Cosmic Microwave Background
CMS	Compact Muon Solenoid
CNAF	National center of INFN
CP	Charge conjugation parity
CPU	Central processing unit
CRAB	CMS Remote Analysis Builder
CRIC	Computing Resources Information Catalog
CSI	Caesium iodide
CVMFS	CERN Virtual File System
DAC	Data access center
DAOD	Derived AOD
DAQ	Data acquisition system
DBS	Dataset Bookkeeping Service
DCC	Disk and Compute Centers
DD	Deuterium-Deuterium
DDM	Distributed Data Management
DES	Dark Energy Survey
DESC	Dark Energy Science Collaboration
DESI	Dark Energy Spectroscopic Instrument
DESY	Detaches Electronic-Synchrotron
DF	Data facility
DIRAC	Distributed Infrastructure with Remote Agent Control
DM	Data Management
DMS	Data Management System
DNS	Domain name service
DOE	Department of Energy
DOMA	Data Organization, Management, and Access
DP	Dual-Phase
DTN	Data Transfer Nodes
DUNE	Deep Underground Neutrino Experiment
EB	Exabyte
EDC	Education and Public Outreach Data Center
EGI	European Grid Infrastructure
ESCC	ESnet Site Coordinators Committee
FIFE	FabrIc for Frontier Experiments

FTS	File Transfer Service
FUSE	Filesystem in Userspace
GCP	Google Cloud Platform
GOLE	Global Lambda Integrated Facility Operators of Lambda Exchanges
GPU	Graphics processing units
GSFC	Goddard Space Flight Center
HEP	High Energy Physics
HEPAP	High Energy Physics Advisory Panel
HL	High Luminosity
HLT	High-Level Trigger
HPC	High Performance Computing
HPSS	High Performance Storage System
HSF	HEP Software Foundation
HTAR	HPSS Tape Archiver
HTC	High-throughput computing
HTSN	High-Throughput Science Network
IDF	Interim data facility
IFAE	Institut de Física d'Altes Energies
IFIC	Instituto de Física Corpuscular
INCITE	Innovative and Novel Computing Theory and Experiment
INFN	Italian Institute for Nuclear Physics
IO	Input/output
IOPS	Input/output operations per second
IP	Internet Protocol
IRNC	International R&E Network Connections
ISP	Internet service provider
IT	Information technology
ITC	Information Technology Center
JEDI	Job Execution and Definition Interface
JGN	Japan Gigabit Network
KEK	Japanese High Energy Accelerator Research Organization
KEKCC	KEK Central Computer
KNL	"Knights Landing" architecture for Intel CPUs
KPNO	Kitt Peak National Observatory
LAG	Link aggregation group
LAMBDA	Legacy Archive for Microwave Background Data Analysis
LAN	Local Area Network
LArTPC	Liquid Argon TPCs
LBNF	Long-Baseline Neutrino Facility
LBNL	Lawrence Berkeley National Laboratory
LCF	Leadership computing facilities

LCG	WLCG
LHC	Large Hadron Collider
LHCC	LHC Coordinating Committee
LHCONE	LHC Open Network Environment
LHCOPN	LHC Optical Private Network
LPC	LHC Physics Center
LR	Long Reach
LSST	Legacy Survey of Space and Time, Large Synoptic Survey Telescope
LZ	LUX-Zeplin
MAN	Metro Area Network
MB	Megabyte
MC	Monte Carlo
MGHPCC	Massachusetts Green HPC Center
MIDAS	Maximum Integrated Data Acquisition System
ML	Machine learning
MOA	Memorandum of agreement
MOU	Memorandum of understanding
MPC	Minor Planet Center
MPI	Message passing interface
NCSA	National Center for Supercomputing Applications at University of Illinois at Urbana-Champaign
NeIC	Nordic e-Infrastructure Collaboration
NERSC	National Energy Research Scientific Computing Center
NREN	National Research and Education Network
NSF	National Science Foundation
NVME	Non-Volatile Memory Express
OCS	Observatory Control System
OLCF	Oak Ridge Leadership Computing Facility
OODS	Observatory Operations Data Service
OPEX	Operating expenses
OSCARS	On-demand Secure Circuits and Reservation System
OSG	Open Science Grid
PB	Petabyte
PBR	Policy-based routing
PIREN	Pacific Islands Research and Education Network
PMT	Photomultipliers tubes
POMS	Production Operations Management Service
POP	Point of presence
POSIX	Portable Operating System Interface
POT	Protons on target
PRP	Pacific Research Platform

PU	Pile-up
QA	Quality Assurance
QoS	Quality of service
R&D	Research and development
R&E	Research and education
RAC	Real Applications Clusters
RAID	Redundant Array of Independent Disks
RSP	Rubin Science Platform
RTT	Round-trip time
SBN	Short-Baseline Neutrino
SBND	SBN near detector
SC	DOE Office of Science
SCD	Scientific Computing Division
SDCC	Scientific Data and Computing Center
SDDC	Scientific Data and Computing Center
SDN	Software Defined Networking
SENSE	SDN for End-to-End Networked Science at the Exascale
SINET	Science Information Network, a Japanese academic backbone network
SLAC	SLAC National Accelerator Laboratory
SM	Standard Model
SNEWS	SuperNova Early Warning System
SP	Single-Phase
SRD	Science Requirements Document
SSD	Solid-state drive
SURF	Sanford Underground Research Facility
TA	Transatlantic
TACC	Texas Advanced Computing Center
TB	Terabyte
Tbps	Terabits per second
TDAQ	Triggering and data acquisition
TDR	Technical Design Report
TDRS	Tracking and Data Relay Satellite
TDRSS	Tracking and Data Relay Satellite System
ToR	Top of Rack
TPC	Third-Party Copy, time projection chamber
TriDAS	LHC's trigger and data-acquisition system
UCPMS	UC Publication Management System
UCSD	University of California, San Diego
UFJF	Universidade Federal de Juiz de Fora
UKDC	United Kingdom Data Center
UNL	University of Nebraska, Lincoln

USDF	US Data Facility
VO	Virtual Organization
VPLS	Virtual Private LAN Service
VRF	Virtual routing and forwarding
WAN	Wide-area networking
WLCG	Worldwide LHC Computing Grid
XSEDE	Extreme Science and Engineering Discovery Environment