Basic Energy Sciences Network Requirements Review Final Report

March – September, 2022
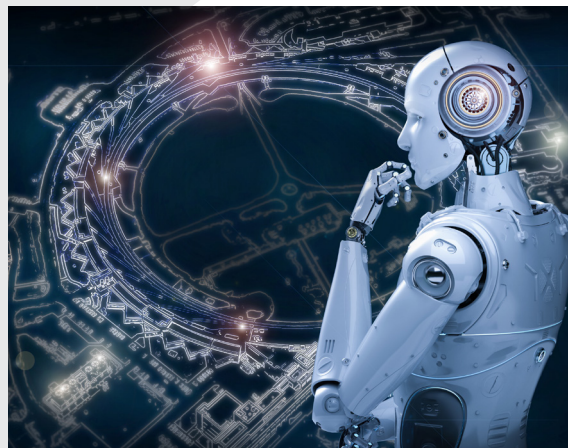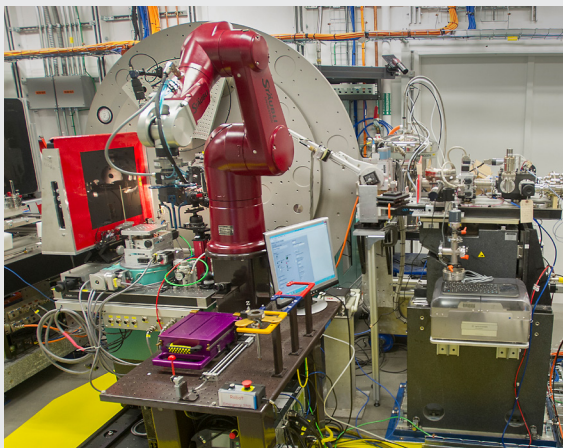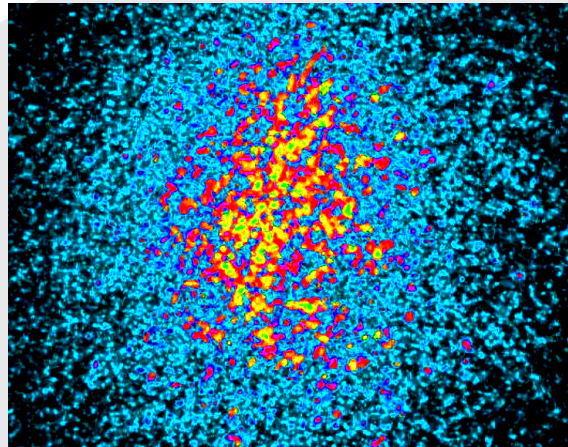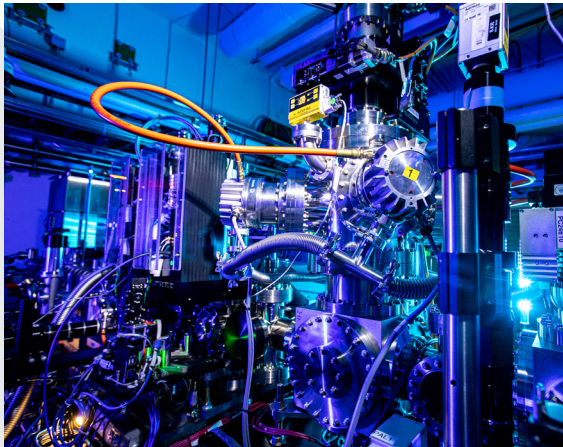

BERKELEY LAB

U.S. DEPARTMENT OF
ENERGY

Office of Science

![ESnet — ENERGY SCIENCES NETWORK]

Basic Energy Sciences Network Requirements Review Final Report

March – September, 2022

Office of Basic Energy Sciences, DOE Office of Science Energy Sciences Network (ESnet)

Disclaimer

Cover Images:

Top left: TMO 1.1 Hutch, located in the Near Experimental Hall in LCLS. (Jacqueline Ramseyer Orrell/SLAC National Accelerator Laboratory)

Top right: L. Shen, M. Seaberg, E. Blackburn and J. J. Turner, "A snapshot review-Fluctuations in quantum materials: from skyrmions to superconductivity", MRS Adv. 6, 221-233 (2021); https://link.springer.com/article/10.1557/s43580-021-00051-y

Bottom left: Hard X-ray Nanoprobe (Brookhaven National Laboratory)

Bottom right: Image by Shutterstock/Phonlamai Photo and Henry Chan.

[1] https://escholarship.org/uc/item/3jj0h54n

Participants and Contributors

Daniel Allan, *Brookhaven National Laboratory*

Taylor Antonich, *Sandia National Laboratories*

Anthony Avarca, *Argonne National Laboratory*

Edward Barnard, *Lawrence Berkeley National Laboratory*

Kipton Barros, *Los Alamos National Laboratory*

Cristian Batista, *University of Tennessee*

Matt Bedynek, *Oak Ridge National Laboratory*

Hassina Bilheux, *Oak Ridge National Laboratory*

Charles Black, *Brookhaven National Laboratory*

Willem Blokland, *Oak Ridge National Laboratory*

Vinny Bonafede, *Brookhaven National Laboratory*

Stuart Campbell, *Brookhaven National Laboratory*

Dale Carder, *ESnet*

Henry Chan, *Argonne National Laboratory*

Maria Chan, *Argonne National Laboratory*

Emory Chan, *Lawrence Berkeley National Laboratory*

Laurent Chapon, *Argonne National Laboratory*

Tanny Chavez, *Lawrence Berkeley National Laboratory*

Sugata Chowdhury, *Howard University*

Paul Cianciulli, *Brookhaven National Laboratory*

James Clifford, *Los Alamos National Laboratory*

Leilani Conradson, *SLAC National Accelerator Laboratory*

Sara Cousineau, *Oak Ridge National Laboratory*

Rich Crompton, *Oak Ridge National Laboratory*

Eli Dart, *ESnet*

Wibe (Bert) de Jong, *Lawrence Berkeley National Laboratory*

Remi Dingreville, *Sandia National Laboratories*

Peter Ercius, *Lawrence Berkeley National Laboratory*

Richard Fenner, *Argonne National Laboratory*

Jon Fortnoy, *Oak Ridge National Laboratory*

Mark Foster, *SLAC National Accelerator Laboratory*

Giulia Galli, *University of Chicago*

Panchapakesan Ganesh, *Oak Ridge National Laboratory*

Feliciano Giustino, *University of Texas, Austin*

Matthias Graf, *Department of Energy, Office of Science*

Stephen Gray, *Argonne National Laboratory*

Sinéad Griffin, *Lawrence Berkeley National Laboratory*

Corey Hall, *Argonne National Laboratory*

Steve Hartman, *Oak Ridge National Laboratory*

Chris Harwood, *Lawrence Berkeley National Laboratory*

Geoffroy T. F. Hautier, *Dartmouth College*

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

Alexander Hexemer, *Lawrence Berkeley National Laboratory*

Susan Hicks, *Oak Ridge National Laboratory*

John Hill, *Brookhaven National Laboratory*

Aaron Holder, *Department of Energy, Office of Science*

Jennifer Hollingsworth, *Los Alamos National Laboratory*

Martin Holt, *Argonne National Laboratory*

Matthew Horton, *Lawrence Berkeley National Laboratory*

Dale Huber, *Sandia National Laboratories*

Patrick Huck, *Lawrence Berkeley National Laboratory*

Dylan Jacob, *ESnet*

Anubhav Jain, *Lawrence Berkeley National Laboratory*

Jason Jed, *Lawrence Berkeley National Laboratory*

Steve Johnston, *University of Tennessee*

Kyle Kelley, *Oak Ridge National Laboratory*

Cathy Knotts, *SLAC National Accelerator Laboratory*

Rob Knudson, *Oak Ridge National Laboratory*

Jeeem Kohl, *Oak Ridge National Laboratory*

Wilko Kroeger, *SLAC National Accelerator Laboratory*

David Leibfritz, *Argonne National Laboratory*

Eliane Lessner, *Department of Energy, Office of Science*

Yingwai Li, *Los Alamos National Laboratory*

Mike Lilly, *Sandia National Laboratories*

Mark Lukasczyk, *Brookhaven National Laboratory*

Phil Maffettone, *Brookhaven National Laboratory*

Christopher Mayes, *SLAC National Accelerator Laboratory*

Paul McIntyre, *SLAC National Accelerator Laboratory*

Dylan McReynolds, *Lawrence Berkeley National Laboratory*

Apurva Mehta, *SLAC National Accelerator Laboratory*

Dylan Merrigan, *Los Alamos National Laboratory*

Ken Miller, *ESnet*

Jeff Nelson, *Sandia National Laboratories*

Kristin Persson, *Lawrence Berkeley National Laboratory*

David Prendergast, *Lawrence Berkeley National Laboratory*

Thomas Proffen, *Oak Ridge National Laboratory*

Christopher Ramirez, *SLAC National Accelerator Laboratory*

Timmy Ramirez-Cuesta, *Oak Ridge National Laboratory*

Shelly Ren, *Oak Ridge National Laboratory*

Adam Rondinone, *Los Alamos National Laboratory*

Cody Rotermund, *ESnet*

Thomas Russell, *Department of Energy, Office of Science*

Anjana Samarakoon, *Argonne National Laboratory*

Alec Sandy, *Argonne National Laboratory*

Subramanian Sankaranarayanan, *Argonne National Laboratory*

Andreas Scholl, *Lawrence Berkeley National Laboratory*

Nicholas Schwarz, *Argonne National Laboratory*

Athena Sefat, *Department of Energy, Office of Science*

Brandon Siegel, *Argonne National Laboratory*

Richard Simon, *Lawrence Berkeley National Laboratory*

Eli Stavitski, *Brookhaven National Laboratory*

Rune Stromsness, *Lawrence Berkeley National Laboratory*

Bobby Sumpter, *Oak Ridge National Laboratory*

Nicholas Talbot, *Brookhaven National Laboratory*

Alan Tennant, *Oak Ridge National Laboratory*

Vivek Thampy, *SLAC National Accelerator Laboratory*

Jana Thayer, *SLAC National Accelerator Laboratory*

Sergei Tretiak, *Los Alamos National Laboratory*

Joshua Turner, *SLAC National Accelerator Laboratory*

Bogdan Vacaliuc, *Oak Ridge National Laboratory*

Rama Vasudevan, *Oak Ridge National Laboratory*

Stefan Vogt, *Argonne National Laboratory*

Guimei Wang, *Brookhaven National Laboratory*

Greg Watson, *Oak Ridge National Laboratory*

Andrew Wiedlea, *ESnet*

Theresa Windus, *Iowa State University and Ames National Laboratory*

Linda Winkler, *Argonne National Laboratory*

Ryan Wixom, *Sandia National Laboratories*

Qin Wu, *Brookhaven National Laboratory*

Jie Xu, *Argonne National Laboratory*

Kevin Yager, *Brookhaven National Laboratory*

Wanli Yang, *Lawrence Berkeley National Laboratory*

Dmitri Zakharov, *Brookhaven National Laboratory*

Yugang Zhang, *Brookhaven National Laboratory*

Jianxin Zhu, *Los Alamos National Laboratory*

Jason Zurawski, *ESnet*

Report Editors

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

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

Matthias Graf, *Department of Energy, Office of Science:* Matthias.Graf@science.doe.gov

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

Aaron Holder, *Department of Energy, Office of Science:* Aaron.Holder@science.doe.gov

Dylan Jacob, *ESnet:* dtjacob@es.net

Eliane Lessner, *Department of Energy, Office of Science:* Eliane.Lessner@science.doe.gov

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

Cody Rotermund, *ESnet:* crotermund@es.net

Thomas Russell, *Department of Energy, Office of Science:* Thomas.Russell@Science.doe.gov

Athena Sefat, *Department of Energy, Office of Science:* Athena.Sefat@science.doe.gov

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 US 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 regularly, and 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 and relies on collaboration from other ESnet teams: Software Engineering, Network Engineering, and Network Security. This team meets with each individual program office within the DOE SC every three years, with 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.

Between March and September 2022, ESnet and the Office of Basic Energy Sciences (BES) of the DOE SC organized an ESnet requirements review of BES-supported activities. Preparation for these events included identification of key stakeholders: program and facility management, research groups, and technology providers. Each stakeholder group was asked to prepare formal case study documents about its relationship to the BES 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 BES program supports fundamental research to understand, predict, and ultimately control matter and energy at the electronic, atomic, and molecular levels in order to provide the foundations for new energy technologies and to support DOE missions in energy, environment, and national security.

The research disciplines supported —condensed matter and materials physics, chemistry, geosciences, and aspects of physical biosciences—are those that discover new materials and design new chemical processes. These disciplines touch virtually every aspect of energy resources, production, conversion, transmission, storage, efficiency, and waste mitigation. BES research provides a knowledge base to help understand, predict, and ultimately control the natural world and serves as an agent of change in achieving the vision of a secure and sustainable energy future.

The BES[1] program is one of the nation's largest sponsors of research in the physical sciences. The program funds basic science at nearly 170 universities, national laboratories, and other research institutions in the US. BES has also built and supports a national network of major shared research facilities based at DOE national laboratories and open to all scientists. These user facilities help form the backbone of the nation's research infrastructure. Over 16,000 scientists and engineers make use of these facilities each year.

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

- Advanced Light Source (ALS)
- Advanced Photon Source (APS)
- National Synchrotron Light Source II (NSLS-II)
- Linac Coherent Light Source (LCLS)
- Stanford Synchrotron Radiation Light Source (SSRL)
- High Flux Isotope Reactor (HFIR) and Spallation Neutron Source (SNS)
- Center for Functional Nanomaterials (CFN)
- Center for Integrated Nanotechnologies (CINT)
- Center for Nanoscale Materials (CNM)
- Center for Nanophase Materials Sciences (CNMS)
- The Molecular Foundry (TMF)
- Autonomous Experiment Steering for BES Facilities
- BES Design and Development of Digital Twin Strategies
- Multifacility Experimentation and Analysis Workflows: X-ray Light Source Perspective
- Multifacility Experimentation and Analysis Workflows: Neutron Scattering Perspective

---

[1] https://science.osti.gov/bes

- Multifacility Experimentation and Analysis Workflows: Nanoscale Science Research Center (NSRC) Perspective

- Use of the ESnet for Quantum Simulations of Materials and Molecules

- The Materials Project (MP): Status and Future Directions

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

We achieve these goals by review of the case study documents, discussions with authors, and general analysis of the materials. The resulting output is a set of review findings and recommendations that will guide future interactions between BES, ASCR, and ESnet.

These terms are defined as follows:

- **Findings**: key facts or observations gleaned from the entire review process that highlight specific challenges, particularly those shared among multiple case studies.

- **Recommendations**: potential strategic or tactical activities, investments, or opportunities that are recommended to be evaluated and potentially pursued to address the challenges laid out in the findings.

The review participants spanned the following roles:

- Subject-matter experts from the BES activities listed previously.

- ESnet Site Coordinators Committee (ESCC) members from BES activity host institutions, including the following DOE labs and facilities: Argonne National Laboratory (ANL), Brookhaven National Laboratory (BNL), Lawrence Berkeley National Laboratory (LBNL), Los Alamos National Laboratory (LANL), Oak Ridge National Laboratory, Sandia National Laboratories (SNL), 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 BES.

- ESnet staff supporting positions related to facility leadership, scientific engagement, networking, security, software development, and R&D.

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

- **Facility Management and Readiness:**

  — Although the data generated by individual experiments varies, over the next decade the combined data generation rates for some BES user facilities will reach the exabyte per year range.

    ° Processing power of hundreds of PFLOPs is expected to be available to fully analyze this data, which implies strong networking capabilities to link facility instruments to edge, local, campus, and centralized computing facilities reliably, and with low latency.

    ° Terabit per second networking and beyond will be required to handle the large amounts of data expected in the coming years.

    ° Upgraded facilities will feature a significant increase in the data throughput from

today's 1–5 GBps to 200 GBps. Future planned upgrades are expected to increase the throughput to multiple TBps in the coming years.

- **Scientific Data Management:**
  - Data movement and data processing are critical for scientists and researchers, particularly once they have performed their experiments, computations, or simulations. Steps include:
    - Data acquisition via computational simulations or experiments.
    - Data reduction using edge processing.
    - Data movement, cataloging, and archiving.
    - Data analysis and visualization.
    - Fitting to models and analysis of the processed data for structure, function, or dynamics information.
  - The increase in brightness and advances in detector data rates at BES user facilities will generate substantially more data than contemporary equivalents, e.g., multiple orders of magnitude more data than is generated today. These advancements will require step-change improvements in networking, controls and data acquisition, computing, workflow, data reduction, and analysis tools to operate effectively.
  - As BES user facilities increase data volumes generated through experimentation, the effective utilization of high-rate data streams and real-time management of experimental conditions will become more critical. This process will involve:
    - Improved methods to cope with raw data management (e.g., high rates and volumes from instruments).
    - Developing suites of tools to optimize facility usage that focus on integrated data management, analytics, and simulation.
    - Integration of AI/ML approaches to automatically steer experimental progress.
    - Tightly coupled operation between experimental facilities, analysis and computational resources, and networks.
    - Adoption of digital twins to improve the overall design and utilization of limited instrument access.
  - The use of multimodal data requires more sophisticated data processing, and will require increases in computing capabilities, which can include training AI/ML models as well as real-time analysis and feedback for autonomous experiment steering. The BES light sources are exploring the utilization of edge-computing resources, coupled closely with detectors and instruments, to facilitate AI/ML data reduction algorithms
  - Upgraded BES light source data generation rates may be too large for traditional file-based workflows, which will facilitate a move to streaming-based workflows directly to computer system memory requiring robust facility and Laboratory networking, and increases in data processing capabilities throughout the DOE scientific complex.
  - Access to data generated by BES user facilities is typically governed by the user agreement, a legally binding document. In many (most) cases, the data is owned by the users, not the facility. In some instances, data is open and available, given there are no restrictions on the collected data beyond understanding how it was taken and managed.
    - Data sharing is extremely difficult. Many users come through the facility, and

keeping track of whose data is where is far too time consuming. None of this has been automated, as it takes a dedicated team of professionals to keep such a system running.

- ° For large datasets, data storage and movement are bottlenecks in current workflows.

— Key elements of a future data management strategy for the light sources include a common API for accessing network and computing resources, parallel data-transfer tools, high-fidelity data transfer, network performance monitoring, reservations, and dynamic network provisioning.

- **Scientific Workflow:**

  — Consistent with the user agreements that typically includes data ownership by the users, users of BES facilities are assigned allocations to use instruments for experimentation, but no explicit long-term computation or storage at DOE HPC facilities is provided as a part of this process. Users are given access to their data when on-site, and encouraged to take it with them, as there are rarely resources that can support long-term custodial storage.

  — Future adaptations to the generalized workflow used to support BES science will involve:

    - ° Integrated software to manage data movement, processing, and analysis.

    - ° More automated methods to handle adaptive/autonomous experiment steering (e.g., integration with simulation capabilities).

    - ° Categorizing and sharing of data with other national resources.

    - ° Increases in data volume due to the upgraded capabilities.

  — A number of BES user facilities are limited in the scope and capabilities of resources they can provide after the experimental phase has completed:

    - ° Data is often stored at the facility for a short period of time, provided local storage resources are available. Backup facilities (e.g., institutional or cloud-based) are not guaranteed.

    - ° Users are encouraged to take a copy of their data; most do through cloud storage, external media transfer, or high-performance Data Transfer Nodes (DTNs) and Globus, when available.

    - ° Local computational resources vary, and can range from multiprocessor workstations with a few analysis tools, to larger clusters that run complex analysis frameworks. A number of BES user facilities have taken steps to integrate portions of their experimental workflow with DOE HPC resources using an institutional allocation.

    - ° The lack of a uniform data pipeline standard that spans facilities causes issues, such as differences in metadata associated with different data types and different levels of data manipulation. To reach a stable and universal environment, standard data formats, software development frameworks, sample tracking, metadata capturing, and labeling are required.

  — The BES instrumentation landscape remains complex, with a number of experimental components unable to be operated as integrated components of a larger facility. Many instruments are operated stand-alone; in some cases, the instruments do not have sufficient storage needs and must be manually integrated into institutional or national-level storage and computation resources. This often reveals significant networking challenges in ensuring ample capacity and low latency.

- **Computational and Storage Requirements:**
  - BES user facilities face three primary challenges in the coming years related to computation, storage, and networking:
    - Data storage is currently sufficient but will become a significant concern as facilities upgrade and data volumes increase.
    - Access to computation is particularly limited for experimental and theory groups. When it cannot be used locally, it must be found either within the DOE ecosystem or at commercial providers.
    - Transfer speeds are also currently sufficient, but will need to keep pace with data volumes and a growing need to access external computation.
  - The current storage capacities at BES light source facilities are often insufficient to handle the expected data volumes for future upgrades. The most data-intensive facilities are expected to have steady-state data rates approaching 10 PB/year, and peak data rates during burst acquisition of >1 TB/hour.
  - After a number of BES light source upgrades, it will be impractical and unreasonable to support the scale of computing required with local resources only. BES user facilities are increasingly partnering with DOE HPC facilities to deliver a new model of computing, tightly coupling experiment instruments and supercomputers to accelerate scientific discovery.
  - BES user facilities of the future will rely on a pipeline that facilitates data transfer from experimental source to different layers of computation. These computational capabilities could be edge computers close to experiments that will facilitate AI and ML tasks, as well as midcapacity and HPC infrastructure in national labs to produce simulations and operate analysis workflows. These resources will guide experimental investigation and accelerate materials as well as fundamental physics discoveries.
  - DOE HPC facility allocations are not automatic and must be applied for on cycles. When they run out, it can sometimes be hard to request more. Having a more stable way to rely on HPC resources at a given facility (or several interchangeable facilities) is desirable. BES would like to explore having a permanent allocation for use, instead of having each facility (or principal investigator (PI)) have to request each time. In the future, BES may have fewer dollars to spend on storage and computation. Having a permanent way to address computing challenges would alleviate some of this pressure off the workflow. It would also lend itself to simplified use cases, using standard APIs, and less bespoke ways to address the overall challenge of data reduction, analysis, or storage approaches.
- **Remote Collaboration and Operational Requirements:**
  - Current remote science activities remain primitive, and reflect the results of a rapid deployment process to cope with the COVID pandemic. As these approaches adapt, remote engagement among scientific collaborators is necessary, and future facilities will need to support remote progress. Outstanding challenges include:
    - Developing and deploying tools to facilitate remote data acquisition and analysis.
    - Providing a remote instrument control experience commensurate with that one can obtain while local to the instrument.
    - Providing remote data access through the facility, or linked to cloud storage solutions within the DOE facility space, or through the use of commercial cloud vendors.

- Increasing the capabilities to support streaming data analysis, when the size and time constraints of bulk-data movement are not possible due to users without access to dedicated computation or storage resources, or a lack of sophisticated software tools.

- Finding future network solutions that integrate seamlessly and provide high-speed connections to instruments, storage, and computation.

- **Multifacility Computational Workflows:**

  — Users at BES facilities often leverage the complete ecosystem of available DOE capabilities. The process of science often involves combining and analyzing data from multiple sources. As data rates and complexity continue to increase, sufficient networking connectivity, bandwidth, and reliability is required to connect measurement facilities, a computing facility, and user home institutions to enable effective data management and analysis.

  — BES use of DOE HPC facilities for computational needs does have some limitations:

    - Tying a workflow to resources that are not directly controlled by a BES user facility or experiment means that if the resources are not available (e.g., due to scheduled or unscheduled downtime), it is often not possible to migrate allocation time, research data, or code to execute at the last minute. This could be mitigated with the ability to more easily port allocations between facilities or resources.

    - There is no universal DOE HPC API, which means code written for one facility may not work at others natively.

    - Allocation resources are finite, and if they expire and are not renewed, data must be migrated elsewhere.

    - Having a more stable way to rely on HPC resources at a given facility (or several interchangeable facilities) is desirable.

  — BES light source facilities use several DOE HPC facilities for storage and processing. Manual, or automated data-movement processes, can move data (bulk or streamed) for real-time and offline analysis. For the most demanding computational problems, large-scale computing facilities must be used, including the Argonne Leadership Computing Facility (ALCF), the National Energy Research Scientific Computing (NERSC) Center, and the Oak Ridge Leadership Computing Facility (OLCF). Sending data from a light source facility to the HPC facility requires a data path of sufficient bandwidth.

- **Domestic Networking for Local and Wide-Area Data Mobility:**

  — As data rates and volumes continue to grow, greater demands will be placed on facility, DOE national laboratory, and national networks. Each facility is investing considerable effort, and working with laboratory IT staff, to re-design the data architecture components. BES is moving toward models where complex analysis pipelines are being considered, with BES users becoming interested in exploring how to schedule network capabilities, and real-time computational resources, to be available during experimentation.

  — BES users experience complications when sharing data across administrative boundaries due to security policies within and between BES user facilities. Identity management continues to be a friction point, and is exacerbated in a multifacility workflow model where a user may need to authenticate to multiple locations, using multiple identities, on a regular basis. A unified way to handle identity is still highly desirable.

  — BES user facilities that leverage DOE HPC resources will continue to expand in the coming years, placing heavy emphasis on ESnet connectivity to manage traffic across

the country. DOE HPC facilities must keep pace with upgrades at BES user facilities, particularly the light sources, to ensure ample capacity to support the scientific mission.

— By the end of the decade, data aggregated at rates of multiple terabits per second (100–300 Gbps within three years and exceeding 1 Tbps in five years) may flow via ESnet from any of the DOE light sources to any of the DOE HPC facilities. There is an urgent need to upgrade the light source to ESnet connections at some of the laboratories to 400 Gbps and above to match the input bandwidth at ASCR computing facilities within the next two years.

— All DOE national laboratories supporting BES user facilities are connected to ESnet with at least 100 Gbps, with most making plans to increase this to multiples of 100 Gbps, or 400 Gbps, in the two- to five-year timeframe. These laboratories also feature Science DMZ architectures. These vary in implementation, but have commonality in offering high-performance network connections to support data transfer, dedicated DTN hardware, availability of Globus for transfer, and participation in regular perfSONAR testing.

— Use of cloud storage solutions to support the long-term storage of research data is increasing, but is still done in an ad-hoc manner. DOE SC must focus on addressing questions regarding the long-term archival storage of facility data, and if it will be stored at each facility, in a central DOE HPC center, or in commercial clouds. If facilities are to be responsible for long-term storage of user data, then consideration and commitments regarding scientific mission, funding for storage resources, and increases in-network capacity for access will be needed.

- **Emerging Needs:**

    — There is limited subject-matter expertise to address technical challenges in the use of HPC resources within the BES community. BES leverages HPC expertise from each of the facilities when possible, but finds it hard to also grow this within the community of users who also know the underlying science use case.

    — Autonomous experiment steering at BES light source facilities will be crucial to future experiment success, and requires that a number of gaps in capabilities and infrastructure be addressed:

        ◦ Sufficient and reliable bandwidth.

        ◦ Sufficient and sustainable data storage resources.

        ◦ On-demand access to large-scale computing systems.

        ◦ Transparent access to facilities and systems through federated identity and shared data protocols.

        ◦ Software tools and infrastructure to facilitate the development of scientific data workflows.

    — The rapid advancement in AI/ML algorithms, improved shared workflows, and the advent of exascale computational resources make it possible to create a physically informed virtual platform to perform experimentation: digital twins. Digital twins are computational environments designed to mimic a physical experiment, and will help guide a research question from conception to synthesis and measurement. Digital twins will:

        ◦ Facilitate the creation of a virtual environment to exhaustively explore experimental controls.

        ◦ Import experimental read-outs at any time to provide instant synthetic read-outs.

- ° Create data that can iteratively be used as training for model improvements.
  - ° Allow for a small subset of simulated experiments to be explored in actual experiments at the SUFs.
- — The further development and application of digital twin simulations supporting experimental design and operation requires that a number of gaps in capabilities and infrastructure be filled, including sufficient and reliable networks, data storage resources, computing systems, transparent access to facilities via federated identity and shared data protocols, and software tools to operate across the DOE's distributed resource landscape.
- — A number of high-performance computation and quantum simulation users and collaborations currently use multiple facilities and locations to accomplish science goals that cross the DOE, commercial, and university complex.
  - ° Calculations and simulations of quantum systems, materials, and molecules use a variety of platforms for hybrid computing, encompassing midsize compute cluster and storage systems owned by universities, and high-performance architectures at DOE national laboratories.
  - ° A future hybrid computing strategy will require tighter integration of classical and quantum resources for an overall multitiered approach to computing. In this case, a quantum processor becomes a co-processor on which certain tasks are offloaded, but will require real-time data movement. The data transfer would be small, but real-time transfer will be critical between resources.
  - ° Data accessibility, searchability, and usability is a key issue for quantum simulation, and improvements to enable high-throughput scientific workflows will be required.
  - ° Data sharing is primarily challenged by data formats and searchability, and for certain classes of simulations the size of the data to be transferred remains a significant issue to solve.
  - ° There are significant needs for software that performs well on heterogeneous computers. In addition, cloud computing could be an area of need in the future in terms of resources and interoperability.

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

- **Facility Management and Readiness:**
  - — ESnet will work with BES user facilities and host DOE national laboratories on strategies and implementations of advanced networking to address pending capability upgrades. These discussions will focus on:
    - ° Increasing network capacity to address growing data volumes between BES user facilities, collaborators, and DOE HPC resources.
    - ° Implementing advanced services that can address scientific requirements for data management, mobility, and remote operation.
    - ° Understanding and implementing peering policy to reach networks and resources of interest, including commercial cloud providers.
- **Scientific Data Management:**

- ESnet, ASCR, and BES will continue to work to understand the implications of increased data rates from facilities through yearly engagements. The estimates in this report show that within 10 years, aggregated data rates of multiple terabits per second flowing across ESnet from any of the light sources to any of the DOE HPC facilities may be possible. There is an urgent need to upgrade the light source to ESnet connections at some of the laboratories to 400 Gbps and above, to match the input bandwidth at ASCR computing facilities within the next two years.

- BES will continue to work on the ongoing design, development, and adoption of the Bluesky framework: a way to streamline data acquisition (DAQ) and ease the process of collecting and comparing data. Given that a number of BES light sources plan to implement the Bluesky databroker/tiled software for data management, in order to move to a consistent "data API," future integration with network-layer services may be possible when using remote DOE HPC resources across ESnet.

- **Scientific Workflow:**
  - The five BES light source facilities, coordinated through the five-way LSDCSC, will engage with ESnet on networking topics. ESnet can advise this group as it works toward a unified vision for the distributed data infrastructure to enable user science, the DISCUS.

- **Computational and Storage Requirements:**
  - ESnet, the BES light source facilities, the ALCF, NERSC, and the OLCF must continue to collaborate on network upgrade strategies.
    - After a number of BES facility upgrades, it will be impractical and unreasonable to support the scale of computing required with local resources only. BES user facilities are increasingly partnering with DOE HPC facilities to deliver a new model of computing, tightly coupling experiment instruments and supercomputers to accelerate scientific discovery.
    - BES user facility's use of DOE HPC resources will continue to expand in the coming years and is estimated to top aggregated data at rates of multiple terabits per second. This will place heavy emphasis on ESnet connectivity to manage traffic across the country.
    - BES user facilities of the future will rely on a pipeline that facilitates data transfer from initial source to different layers of computation. These computational capabilities could be edge computers close to experiments that will facilitate AI and ML tasks, as well as midcapacity and HPC infrastructure in national labs to produce simulations and operate analysis workflows. These resources will both guide experimental investigation and accelerate materials and fundamental physics discoveries.
    - Networking must be able to robustly connect BES user facility instruments to remote computing facilities with low latency.

- **Remote Collaboration and Operational Requirements:**
  - ESnet will explore mechanisms to support the following aspects of remote use of BES facilities:
    - Supporting the network to enable the remote operation of instruments, either fully automated or through communication over telecommunications tools.
    - Building high bandwidth/low latency paths to support computational and storage needs

between facilities.

- ° Enabling peering to cloud providers to support communication and collaboration tools.
- ° Encouraging the deployment of user-focused data portals to allow for data mobility.

- **Multifacility Computational Workflows:**

  - — ESnet will continue to advise BES as the use of multimodal data increases, and crosses the boundaries between user facilities and DOE HPC resources. Multimodal data requires more sophisticated data processing and increases in computing capabilities.

- **Domestic Networking for Local and Wide-Area Data Mobility:**

  - — ESnet will continue to support R&D efforts within the BES community that are advancing use of network technologies to support scientific workflows. For complex analysis pipelines or high data rate experiments, it is anticipated that future users of BES facilities will have to schedule network capabilities and real-time computational resources to be available during the time of experiments. The BES community is coordinating with ESnet and DOE HPC facilities on possible solutions.

  - — ESnet will work with BES user facilities to increase the adoption of the Science DMZ paradigm, and the affiliated technologies that support network measurement (perfSONAR, DTNs) and data mobility (Globus). A number of BES user facilities and their parent DOE national laboratories have adopted these tools, but there remain gaps in automated approaches to use them regularly to support scientific workflows.

- **Emerging Needs:**

  - — ESnet will work with BES user facilities, and DOE HPC facilities, as autonomous experiment steering is more widely adopted. Given the implicit use of networks to join experimental facilities with DOE HPC resources, ESnet can provide high bandwidth, low latency network connections to ensure experimental success.

  - — Coupling digital twin simulations with experimental steering will become an increasingly relied-upon capability at the light sources as the decade progresses. Due to the high computational cost associated with data processing, reduction, and analysis, and of model training on such large datasets, the light sources will need to employ the use of geographically distributed computing facilities and data repositories.

  - — ESnet will continue to work with BES on a strategy to support quantum computing and hybrid computing requirements from a network perspective:

    - ° It is expected that a hybrid quantum-classical model, where part of the calculations will be carried out on classical architectures and part on quantum hardware, will persist for a number of years, implying a transfer between the two platforms.

    - ° A number of high-performance computation and quantum simulation users and collaborations currently use multiple facilities and locations on a variety of platforms for hybrid computing to accomplish science goals that cross the DOE, commercial, and university complex. Data sharing will be required in this environment.

2 Review Findings

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

- Facility Management and Readiness
- Scientific Data Management
- Scientific Workflow
- Computational and Storage Requirements
- Remote Collaboration and Operational Requirements
- Multifacility Computational Workflows
- Domestic Networking for Local and Wide-Area Data Mobility
- Emerging Needs

2.1 Facility Management and Readiness

- Although the data generated by individual experiments varies, over the next decade the combined data generation rates for some BES user facilities will reach the exabyte per year range. [Section 5]
  - Processing power of hundreds of PFLOPps is expected to be available to fully analyze this data, which implies strong networking capabilities to link facility instruments to edge, local, campus, and centralized computing facilities reliability, and with low latency.
  - Terabit per second networking and beyond will be required to handle the large amounts of data expected in the coming years.
  - Upgraded facilities will feature a significant increase in the data throughput from today's 1–5 GBps to 200 GBps. Future planned upgrades are expected to increase the throughput to multiple TBps in the coming years.
- The ALS is a user facility with 40 different beamlines and 2,100 scientific users per year from all over the world. The biggest change to our facility will be the transition to a diffraction-limited light source (ALS-U) starting in 2025. As complexity and data rates ramp up, preparing and testing data streaming workflows will be critical. ALS-U will need access to compute and analytics resources and the ability to schedule the required resources in advance. [Section 5.1]
- The APS is a scientific user facility that serves approximately 6,000 unique users per year via 68 beamlines. The APS upgrade (APS-U) project will replace the entire APS storage ring. A one-year shutdown is required, and is scheduled to begin in April 2023. [Section 5.2]
- NSLS-II began user operations in 2015 and has hosted approximately 1,800 distinct researchers prior to the pandemic, with approximately 1,400 seen during the pandemic years. The NSLS-II currently operates 28 beamlines, each with its own unique data generation characteristics. Depending on the beamline and measurement, beam time may range from a few hours to more than a week in duration. [Section 5.3]
- The LCLS, located at the SLAC National Accelerator Laboratory, features 10 specialized instruments. LCLS has had over 13,000 scientific user visits, and more than 3,000 unique users

over the course of operation. On average, LCLS has around 1,000 users, and 700 unique users, every year. Unlike at a synchrotron where X-rays can be delivered to multiple instruments at once, the linear accelerator can deliver X-rays to only one instrument at a time, although it is possible for other experiments to run parasitically using the X-rays that pass through the upstream instrument. As a result, there are usually one or two experiments during a shift. [Section 5.4]

- The SSRL has 32 operational beam line stations on SPEAR3, of which 26 can operate simultaneously. SSRL provides facilities to approximately 1,700 unique scientists per year (pre-pandemic). [Section 5.5]

- The SNS and the HFIR are both DOE BES scientific user facilities at Oak Ridge National Laboratory (ORNL). Between the two user facilities, there are 2,000 people who wish to access data remotely per year, and typical beam time requests range from 2 to 16 days. [Section 5.6]

- The CFN is a NSRC operated for the US DOE at BNL. Scientific projects at CFN fall under three nanoscience themes: nanomaterial synthesis by assembly, accelerated nanomaterial discovery, and nanomaterials in operando conditions. CFN is not a single collaboration. The CFN facility includes advanced instrumentation in nanolithography, materials preparation, electron and photon probes, and computational resources. [Section 5.7]

- The Center for Integrated Nanotechnologies (CINT) is a DOE Office of Science NSRC operating as a national user facility. As a vibrant partnership between LANL and SNL. CINT has users from most US states (42) and many (26) foreign countries. In a typical pre-COVID year, CINT would host approximately 800 users both on-site and remotely. [Section 5.8]

- The CNM at ANL is one of five NSRCs operated by the DOE Office of Science. During FY21, the CNM hosted 349 on-site and 353 remote user researchers from academia, DOE national laboratories, and industry. [Section 5.9]

- The CNMS at ORNL is a nanoscience user facility that provides services to users in areas of nanofabrication, synthesis, theory, and characterization. The CNMS serves over 600 unique users each year. In FY2021, 70% of our users were nonlocal, with 120 US institutions from 36 states and territories, and 43 foreign institutions from 24 foreign countries. [Section 5.10]

- TMF is one of five national user facilities for nanoscale science that serves roughly 1,000 academic, industrial, and government scientists around the world each year. Users come to the Foundry to perform multidisciplinary research beyond the reach of their own laboratory. [Section 5.11] There are seven facilities at TMF:

  — Imaging and Manipulation of Nanostructures

  — Nanofabrication Facility

  — Theory of Nanostructured Materials Facility

  — Inorganic Nanostructures Facility

  — Biological Nanostructures Facility

  — Organic and Macromolecular Synthesis Facility

  — National Center for Electron Microscopy (NCEM)

- Some NSRCs are located at higher-security facilities, and because of this must use in-house resources. Security compliance can be tricky to manage at a number of BES user facilities, and some struggle with the need to accommodate the requirements, but also make data accessible. It can be a struggle to migrate even small data sets to users in ways that do not violate cyber security requirements. This event has resonated by showing lots of national-scale resources are

available, and it is a worthwhile exercise to explore what options may be available to leverage HPC facilities and ESnet for future collaborative efforts. [Section 6]

- The Materials Project (MP) provides open web-based access to computed information on known and predicted materials, as well as powerful analysis tools, and now serves a community of over 200,000 registered users. [Section 5.18]

2.2 Scientific Data Management

- Each BES user facility has different requirements surrounding data handling: the relative data volumes, the processing steps required, and the analysis and interpretation of data vary depending on the facility, measurement technique, detector(s) utilized, and scientific goal. The size of the data may vary from a few megabytes to hundreds of terabytes. Some form of data processing or reduction is usually performed immediately after data is collected. [Section 5]

- BES-funded light sources often house multiple beamlines and instruments. Most operate 24/7 for eight to nine months a year with multiple user groups running concurrently. Experiments are diverse in their data size with a daily rate ranging from 10 MB (X-ray spectroscopy) to many TB (imaging and tomography). In many cases, data is transferred from the host facilities to the users for analysis using portable drives or data-transfer tools. [Section 5]

- The increase in brightness and advances in detector data rates at BES user facilities will generate substantially more data than contemporary equivalents, e.g., multiple orders of magnitude more data than is generated today. These advancements will require step-change improvements in networking, controls and data acquisition, computing, workflow, data reduction, and analysis tools to operate effectively. [Section 5]

- Upgraded BES light and neutron source data generation rates may be too large for traditional file-based workflows, which will facilitate a move to streaming-based workflows directly to computer system memory, requiring robust facility and Laboratory networking, and increases in data processing capabilities throughout the DOE scientific complex. [Section 6]

- BES light source scientists and users actively develop and support data management tools that assist during the full data lifecycle from acquisition to storage and analysis. These tools include: [Section 5.14]

  — Software to detect, package, and transfer data.

  — Database systems that register data and metadata, to allow search and organization.

  — Workflow tools that allow both user-driven and automated processing that leverage local and remote HPC resources.

  — Web portals that allow users to view or download results or seamlessly move data into an interactive analysis environment for further processing.

  — Access to storage and computing local to the facility.

  — An increasing availability of common APIs for accessing network and computing resources, parallel data-transfer tools, high-fidelity data transfer, network performance monitoring, reservations, and dynamic network provisioning.

- Data rates and volumes for BES light sources can range from several MB per month to 3 GBps currently. Datasets can be as small as 1 GB, or as large as 1–10 TB in size. Future light source upgrades may produce a PB per shift, or several PBs per experiment. [Section 5.14]

  — Many BES user facilities feature automated methods to manage data volumes to use local computation and storage resources, or data mobility tools to facilitate transfer to DOE HPC

facilities reached via ESnet and other research and education (R&E) network providers.

- The use of multimodal data requires more sophisticated data processing, and will require increases in computing capabilities, which can include training AI/ML models as well as real-time analysis and feedback for autonomous experiment steering. The BES light sources are exploring the utilization of edge-computing resources, coupled closely with detectors and instruments, to facilitate AI/ML data reduction algorithms. [Section 5.12, Section 5.13]

- Access to data generated by BES user facilities is typically governed by the user agreement, a legally binding document. In many (most) cases, the data is owned by the users, not the facility. In some instances, data is open and available, given there are no restrictions on the collected data beyond understanding how it was taken and managed. [Section 5.11]

  — Data sharing is extremely difficult. Many users come through the facility, and keeping track of whose data is where is far too time consuming. None of this has been automated, as it takes a dedicated team of professionals to keep such a system running.

  — For large datasets, data storage and movement are bottlenecks in current workflows.

- Data movement and data processing are critical for scientists and researchers, particularly once they have performed their experiments or simulations. Steps include: [Section 5]

  — Data acquisition via computational simulations or experiments.

  — Data movement, cataloging, and archiving.

  — Data analysis and visualization.

  — Fitting to models and analysis of the processed data for structure, function, or dynamics information.

- The SNS First Target Station (FTS) at ORNL is constructing the Versatile Neutron Imaging Instrument (VENUS). It will require 10 Gbps networking to support the expected data volumes; initial estimates are that VENUS will produce 20–30 TB of raw data per day. Reduced data will be a fraction of that amount, but the intention is to keep all raw data. The next-generation detector in development (five+ years) will have approximately 16 times the data requirements. [Section 5.6]

  — The SNS Second Target Station (STS) will add a new suite of instruments to the SNS during the next 10 years. The data rates and storage needs are not well defined at present, but it is safe to assume that the STS will add another 30% to networking and storage requirements over the current FTS plus another VENUS-like instrument.

- A planned collaboration between the CFN at ANL with LBNL to operate a high frame rate camera, capable of generating data rates of 50 GBps, suggests that about 1 PB per year will be generated. [Section 5.7]

  — In the next two to five years, additional direct detection capabilities will likely be added with acquisition of new Scanning/Transmission Electron Microscopes (STEMs) which could drive the data generation rate up to as much as 4 PB per year.

- The BES NSRCs still struggle with data management approaches due to a number of technological gaps. [Section 5.16]

  — Data is collected by users and staff either through network transfer (email, ftp) or encrypted USB drive as required by host DOE national laboratory security policies.

  — Many instrument computers are not networked due to patching and reboot requirements, so data must be transferred via USB drive. The dataset sizes at most of the NSRCs are

1–10 GB.

- — Some collaboration and data sharing are allowed through cloud resources, primarily Google Drive.

- — Limited data movement capabilities, which though not currently an operational inhibitor, will become so as data volumes increase.

- Most microscopy dataset collections are in the 1–10GB range, with more for electron microscopy (can be >10GB for a single dataset in that instance). [Section 5.16]

  - — The difficulty usually lies not in storage but in analysis, so that there is enough feedback for how to optimize measurement parameters before the user departs. For large files, data transfer can be an issue if the institutions involved do not possess Globus endpoints.

  - — STEM devices typically have data rates of approximately 100Mbps, and 4D-STEM can increase this to greater than 100Gbps. Data reduction using edge processing is critical to handle the volumes of the latter.

- NCEM at TMF has nine electron microscopes, of which four are large data generators: [Section 5.11]

  - — TEAM I generates datasets that range in size from 10–50 GB over the course of 10 minutes, and several can be generated per day.

  - — TEAM 0.5 generates data at 480 Gbps, with maximum data sets being 700 GB over a 15-second run. To support this use case, the data must either be reduced locally or transferred directly to NERSC (via dedicated 100 Gbps campus networking). It can produce 20 datasets (14 TBs) per day, but this is currently limited by the data reduction time more than the ability to use the instrument. Data generation will increase by 25% in the one-year time frame, after an upgrade.

  - — TitanX generates data in the 100 MB–2 GB range. Several of these can be acquired per day.

  - — ThemIS generates datasets in the range of 10–50 GB in a few minutes. Several of these are acquired per day.

  - — A new transmission electron microscope (TEM), similar to TEAM I, will arrive in one to two years.

  - — A new TEM, similar to TEAM 0.5, will arrive in three+ years.

- The near real-time interpretation of structure revealed by X-ray diffraction requires significant computational resources. The analysis pattern is characterized by bursts of short jobs, requiring very short startup time. Current data collection rates are about 15 TB/day for high compute experiments. Within three to five years, it is expected that this rate will increase to 500 TB/day and in five+ years to > 1 PB/day. [Section 5.14]

  - — 4D cameras can operate at 350 Gbps, and can produce 5 PBs of data per year. These types of data rates will only increase in electron microscopy due to the increased proliferation and continued development of high-speed electron detectors.

  - — Most of the data from microscopy collaboration is first reduced before it is shared over cloud storage. Globus is often used to transfer files to local storage to run analysis on HPC systems.

  - — The size of datasets for frontier electron microscopy experiments (tomography, 4D, high frame rate) makes it infeasible for users to download the full dataset, and impractical to browse the dataset looking for information of interest. This points to the need for remote

interfaces for data browsing and analytics.

- Key elements of a future data management strategy for the light sources include a common API for accessing network and computing resources, parallel data-transfer tools, high-fidelity data transfer, network performance monitoring, reservations, and dynamic network provisioning. [Section 5]

- Scientific characterization tools have been pivotal to increasing our understanding of nanoscience. Substantial bottlenecks exist between the data streams emanating from these tools, and the feedback from theoretical and simulation insights. Directly coupling the microscope data to simulations in real time to assist the experimenter is of crucial importance in the quest towards automation and autonomous materials and physics discovery platforms. [Section 5]

- The software being used across facilities is currently highly heterogeneous. There is not currently a single, integrated plan for the software systems that will be deployed. [Section 5.16]

  - NSRCs host commercial tools from a multitude of vendors. This makes inevitable interacting with a large number of different software systems, many of which use proprietary data formats. It remains a significant challenge to integrate these tools with automated workflows.

  - Cross-facility integrated datasets face a substantial challenge with respect to acquiring and preserving suitable metadata. Improved software tools for acquiring, transferring, and browsing/searching through metadata are necessary.

- Data rates for next-generation electron microscopes exceed typical local infrastructure capabilities. As microscope data rates will increase even further in future iterations, it is necessary for future network and storage infrastructure to be up to the task of handling these data rates. [Section 5.16]

- First principles molecular dynamics (FPMD) simulations generate datasets of tens to hundreds of terabytes. [Section 5.17]

  - The diverse and heterogenous data generated may not reside on the HPC facilities where they have been created for future access, curation, and reuse, which necessitates data transfer between collaborators.

  - Globus is used through ESnet for the transfer of datasets at different stages of investigation.

  - Most of the data storage and transfer speeds are appropriate for quantum simulation, and this is unlikely to change within the next several years.

- There is not a common practice on how BES user facilities are handling data storage requirements. Some facilities attempt to store all data in perpetuity, others have tried to put sensible policies in place to cycle out the oldest data as space becomes limited, and others still may rely on time-based allocations that delete based on age alone. [Section 6]

- It is desirable to consider the use of DOE SC facilities for experimental data storage, but moving toward this model will require addressing some challenges within BES. Export controls, and other security policy differences between facilities, remain a large hurdle to a universal solution. A secondary problem is the lack of a unified data storage and metadata cataloging framework; these issues remain challenging since automating this is often hard, and experimental users often skip this step in the rush to acquire and publish research results. [Section 6]

- The amount of data produced by some BES instruments is growing beyond current capabilities. Thus, a standard approach to deal with current (and future) technology with unified software, computation, and storage is desirable. The NSRCs and neutron sources currently employ a

heterogenous approach, which must change to be more uniform. This heterogenous comes from having environments where a number of different types of instruments, each with a different software stack (some of which may be vendor specific), are integrated into a user facility. This environment does not lend itself to creation of a standard workflow without a lot of custom software and APIs. [Section 6]

- Data needs from MP users range from highly frequent requests for common objects (hundreds of GB to TB sized downloads), to less frequent, but larger, data objects (tens to hundreds of TB). [Section 5.18]

  — The need to transfer data to and from MP is expected to increase as third-party contributions of data grow.

  — Transferring data between DOE facilities remains cumbersome; MP is trying to move toward accessing other facilities from the cloud, and would like to see templates developed in this use case to allow other facilities to set up cloud resources and then connect on the Virtual Private Cloud (VPC) level within Amazon Web Service (AWS) regions.

- BES would like to investigate the emerging use case that allows for streaming data during analysis. One concept that has stopped BES, as well as a number of other communities, has been the challenges in having HPC worker nodes stream data directly from a remote repository. The typical use case has always been to perform a bulk-data movement to scratch storage, since worker nodes often do not have the ability to perform a wide-area transfer directly. Ongoing research by DOE HPC facilities to enabling different approaches that may help this problem. These approaches are being explored by other DOE communities, including facilities and experiments from High Energy Physics (HEP). [Section 6]

To better align future upgrades of technology components, understanding how the relationship between upgrade capabilities (e.g., detector area captured, readout rate, etc.) will influence the requirements for networking, central processing unit (CPU) memory and bandwidth requirements, and overall storage volumes is desirable. Currently there is not a one-to-one relationship between any of these items; the BES community notes that any technology upgrades that can be provided by ESnet or the HPC facilities will ultimately improve the scientific process until the next bottleneck is reached. [Section 6]

2.3 Scientific Workflow

- There are two primary data flow paths for a generalized BES workflow: [Section 6]

  — Acquisition from an instrument, with data then flowing to local storage

  — Migration from local storage to long-term storage, an analysis facility, or sharing with collaborators

- Consistent with the user agreements that typically includes data ownership by the users, users of BES facilities are assigned allocations to use instruments for experimentation, but no explicit long-term computation or storage at DOE HPC facilities is provided as a part of this process. Users are given access to their data when on-site, and encouraged to take it with them, as there are rarely resources that can support long-term custodial storage. [Section 5]

  — A number of BES user facilities are limited in the scope and capabilities of resources they can provide after the experimental phase has completed: [Section 5]

  — Data is often stored at the facility for a short period of time, provided local storage resources are available. Backup facilities (e.g., institutional or cloud-based) are not guaranteed.

  — Users are encouraged to take a copy of their data; most do via the use of cloud storage,

external media transfer, or high-performance DTNs and Globus when available.

- — Local computational resources vary, and can range from multiprocessor workstations with analysis tools, to larger clusters that run dedicated workflow tools. A number of facilities have integrated with DOE HPC resources that may offer limited resources under an institutional allocation.

- — The lack of a uniform data pipeline standard that spans facilities causes issues, such as differences in metadata associated with different data types and different levels of data manipulation. To reach a stable and universal environment, the BES community must create standard data formats. Software development framework, sample tracking, metadata capturing, and labeling are required.

- The BES instrumentation landscape remains complex, with a number of experimental components unable to be operated as integrated components of a larger facility. Many instruments are operated stand-alone; in some cases, the instruments do not have sufficient storage needs and must be manually integrated into institutional or national-level storage and computation resources. This often reveals significant networking challenges in ensuring ample capacity and low latency. [Section 5]

- There are a number of different sources/sinks for BES user facility data, depending on the workflow: [Section 5.14]

  - — BES user facility to DOE HPC facility.

  - — BES user facility to user facility or DOE national laboratory.

  - — BES user facility to university.

  - — BES user facility to cloud computing/storage.

- BES Synchrotron facility to an DOE HPC facility is the most heavily used for streaming analysis to process raw data or do post-experiment analysis, as well as archive historical data, train and retrain AI/ML, and use simulations to inform experimental process. [Section 5.14]

- The BES workflows between DOE national laboratories are exercised when multimodal analysis is desired, or if a lab/university has a specialized local computing resource used for analysis. [Section 5.14]

- The BES user facility to university use case is mostly used to do the slow transfer of data sets post-experiment to a users' local storage/computing resources. [Section 5.14]

- The BES user facility to cloud computing/storage is not used by the facility but may be employed by users. [Section 5.14]

- Data acquisition collects all aspects and metadata. A single measurement is called a run, and experiments are composed of many runs. Collections of runs may be grouped into a reduced data set. Stored procedures are executed on computing clusters at the end of every run such that reduced data can be provided to the users as soon as possible following the end of data collection. [Section 5.15]

- A typical BES light source facility workflow can be summarized as: [Section 5.14]

  - — Application and execution of facility time for experimentation; this can be done in person or remote.

  - — Generally, experimental data is first stored on a computer attached to the acquisition detector or some other facility storage system.

- Use of integrated management software tools during data acquisition phase.
- Immediate analysis using local resources, or possible data mobility to remote resources such as DOE HPC facilities.
- Data reduction and analysis rely heavily on the use of HPC, graphical processing units (GPUs), edge devices, and distributed computing environments to obtain results with near real-time completion.
- The final analysis and interpretation of the data for publication is generally carried out by the experiment team at their home institutions, is very experiment specific, and may take months or even years to perform.
- Publication of results.

- Future adaptations to the generalized workflow used to support BES science will involve: [Section 5.14, Section 5.15, Section 5.16]
  - Integrated software to manage data movement, processing, and analysis.
  - More automated methods to handle adaptive/autonomous experiment steering (e.g., integration with simulation capabilities).
  - Categorizing and sharing of data with other national resources.
  - Increases in data volume due to the upgraded capabilities.

- BES light source facilities are coordinating data management technology strategies through the five-way LSDCSC. This group produced a unified vision for the distributed data infrastructure to enable user science, the DISCUS. [Section 5.14]

- The BES NSRCs are experimenting with ways to automate their workflows. These may resemble the following: [Section 5.16]
  - Experiments involving synthesis or analysis.
  - When instruments are involved, data is collected on local control device.
  - Data is typically analyzed on the associated local control computer using proprietary software, if the instrument is proprietary.
  - Data may be analyzed on a separate workstation if the data is nonproprietary.
  - Transfer of data between computers or off-site takes place using encrypted USB or general network services. Methods for moving data can be dictated by host DOE national laboratory IT security policies, and may not be automated.

- There is a need for common standards and shared workflows for data across the NSRCs which will not only provide an effective data solution, but will also enable cross-center data sharing, augmentation, and manipulation. NSRCs have different schemes and different levels of implementation for acquiring, labeling, storing, and providing access to the heterogeneous data generated. The lack of data pipeline standards causes issues, such as differences in metadata associated with different data types and different levels of data manipulation. Top address this, many levels of technical details need to be worked out: [Section 5.16]
  - Data formats.
  - Software development framework.
  - Sample tracking.
  - Metadata capturing.

- — Labeling.
- — Datasets from correlated measurements, and the corresponding simulations, need to be handled in a coordinated manner to extract synergistic information.

- As BES user facilities increase data volumes generated through experimentation, the effective utilization of high-rate data streams and real-time management of experimental conditions will become more critical. This process will involve: [Section 5.14]

  - — Improved methods to cope with raw data management (e.g., high rates and volumes from instruments).
  - — Developing suites of tools to optimize facility usage that focus on integrated data management, analytics, and simulation.
  - — Integration of AI/ML approaches to automatically steer experimental progress.
  - — Tightly coupled operation between experimental facilities, analysis and computational resources, and networks.
  - — Adoption of digital twins to improve the overall design and utilization of limited instrument access.

- A number of BES light sources plan to implement the Bluesky databroker/tiled software for the initial data management layer to enable a data acquisition system consistent with the other light sources. This has been developed as a collaborative open-source project. This system aims to provide a consistent "data API" rather than prescribe a given on-disk data format. [Section 5.14]

- Future lab instruments should expose a standard API to accommodate automation of data acquisition. This will allow users to devise integrated experimental workflows and leverage resources that span institutional boundaries. [Section 6]

2.4 Computational and Storage Requirements

- BES user facilities face three primary challenges in the coming years related to computation, storage, and networking: [Section 5.14, Section 5.15, Section 5.16]

  - — Data storage is currently sufficient but will become a significant concern as facilities upgrade, and data volumes increase.
  - — Access to computation is particularly limited for experimental and theory groups. When it cannot be used locally, it must be found either within the DOE ecosystem or at commercial providers.
  - — Transfer speeds are currently sufficient, but will need to keep pace with data volumes, and a growing need to access external computation.

- BES light source facilities adopt a graded approach to resource utilization: [Section 6]

  - — Multicore processors and GPU units are typically available local to beamlines.
  - — Most BES light sources feature a set of on-site computing and storage cluster resources that can be used as needed.
  - — Laboratories maintain computing resources as a part of their institutional computing programs.
  - — DOE HPC facilities (e.g., the ALCF, the NERSC Center, and the OLCF) are available via allocation procedure.

- Data at BES light source facilities is transferred to a local beamline workstation, the facility

computing cluster, a computing center located on the intuitional campus, or resources such as DOE HPC facilities for processing. [Section 5, Section 6]

- The computational resources at BES light source facilities are limited, and meant to be used for temporary storage and limited processing jobs on CPU and graphics processing unit (GPU) machines that can be used for analysis on a first come first served (FCFS) basis. These are expected to grow as instruments are upgraded, but are still not able to meet the demand of all users. It is estimated that 50 to 80% of real-time analysis can be done locally, but that the remainder must use other DOE HPC resources. [Section 5.14]

- The current storage capacities at BES light source facilities are often insufficient to handle the expected data volumes for future upgrades. The most data-intensive facilities are expected to have steady-state data rates approaching 10 PB/year, and peak data rates during burst acquisition of >1 TB/hour. [Section 5.14]

- Typically, BES light sources can support storage requirements in the range of hundreds of TB of short-term retention, and beyond 10 PB for tape storage. Given the limited nature of the storage, retention is limited to weeks or months. Most are in the process of expanding networking capacity, and data-movement platforms, to better utilize DOE HPC facility resources. [Section 5.14]

- DTNs are often available at BES light sources for reliable, high-speed data movement. Users can download data at their home institutions by using Globus or other data mobility tools. Users often utilize their own "cloud storage" accounts (e.g., Google Drive, Dropbox, etc.) to transfer data. [Section 5.14]

- Most BES light source users complete their offline analysis in the four months following an experiment using computing at a facility, or through a DOE HPC allocation. On average, a typical user will rerun over their entire dataset up to 10 times. [Section 5.14]

- Taking advantage of available (local or remote) compute and analysis resources will become an integral part of BES experimental planning in the very near future. In order to deploy efficient analysis and ML approaches at the time of the experiments, workflows will have to be established, and ML algorithms will have to be trained on existing or simulated data. [Section 5.12, Section 5.13]

- After a number of BES facility upgrades, it will be impractical and unreasonable to support the scale of computing required with only local resources. BES user facilities are increasingly partnering with DOE HPC facilities to deliver a new model of computing, tightly coupling experiment instruments and supercomputers to accelerate scientific discovery. [Section 5.14]

- BES user facilities of the future will rely on a pipeline that facilitates data transfer from experimental source to different layers of computation. These computational capabilities could be edge computers close to experiments that will facilitate AI and ML tasks, as well as midcapacity and HPC infrastructure in national labs to produce simulations and operate analysis workflows. These resources will both guide experimental investigation and accelerate materials and fundamental physics discoveries. [Section 5.12, Section 5.13]

- DOE HPC facility allocations are not automatic, and must be applied for each resource cycle. When an allocation runs out, it is sometimes a challenge to request more. Having a predictable mechanism to rely on HPC resources at a given facility (or several interchangeable facilities) is desirable. BES user facilities would like to explore having a permanent allocation for use, instead of having each facility (or PI) have to request each time. In the future, BES may have fewer dollars to spend on storage and computation; thus having a permanent way to address computing challenges will help to simplify the workflow. It will also lend itself to simplified use

cases, standard APIs, and less bespoke ways to address the overall challenge of data reduction, analysis, or storage approaches. [Section 6]

- The BES community lacks subject-matter expertise to explore new ways to address computational challenges. The BES community leverages HPC expertise from each of the facilities when possible, but finds it hard to grow this within the community of users who also know the underlying science use case. BES community members do not want to be experts in software or networking engineering. and relish adopting BCPs that the ASCR community can recommend. [Section 6]

- The NOMAD Repository is used to help with the public dissemination of output from the Materials Project, since it is able to provide a platform and means to publicly share and transfer large amounts of data that scales to a higher degree to meet user demands. [Section 5.18]

2.5 Remote Collaboration and Operational Requirements

- BES user facilities have adapted to the COVID pandemic by shifting to modalities that facilitate some remote scientific capabilities, away from more traditional models with an expectation of users being on-site. The BES user facilities anticipate that remote access will continue for some users and use cases, and thus is now a critical infrastructure component to support and grow from an experimental, computational, and networking perspective. Observations include, but are not limited to: [Section 5]

  - A mail-in process for the initial user sample analysis.

  - Remote operation of instruments; either fully automated or through communication over telecommunications tools (e.g., VNC, Zoom, Teams, etc.) to facility staff.

  - Ongoing advances in robotic sample handling, automated data acquisition and data management, AI/ML to reduce human intervention, and other technology improvements to light sources and detectors.

  - Deployment of remote access tools that enable experimenters to run measurements remotely.

  - Remote login infrastructure that provides remote users access to web services, data management systems, and a data analysis infrastructure that allow for virtual access to enable on and off-site analysis.

  - Associated cybersecurity measures.

  - Automated data mobility tools that include a set of high-performance DTNs as well as high-performance software that enables transfer and sharing such as Globus.

- Current remote science activities remain primitive, and reflect the results of a rapid deployment process to cope with the COVID pandemic. As these approaches adapt, remote engagement among scientific collaborators is necessary, and future facilities will need to support remote progress. Outstanding challenges include: [Section 5]

  - Developing and deploying tools to facilitate remote data acquisition and analysis.

  - Providing a remote instrument control experience that is commensurate with that one can obtain while local to the instrument.

  - Providing remote data access through the facility, or linked to cloud storage solutions within the DOE facility space, or through the use of commercial cloud vendors.

  - Increasing the capabilities to support streaming data analysis, when the size and time

constraints of bulk-data movement are not possible due to users without access to dedicated computation or storage resources, or a lack of sophisticated software tools.

— Developing future network solutions that integrate seamlessly and provide high-speed connections to instruments, storage, and computation.

2.6 Multifacility Computational Workflows

- BES multifacility workflows follow a common pattern: [Section 5.14, Section 5.15, Section 5.16]

  — Experimental and simulation data, computer resources, and collaborators are located in different locations, requiring remote use.

  — Simulations that can be used to guide and explain the observations encountered in the experiment are critical, and require fast processing and network access to be fully realized.

  — When the compute infrastructure is used to inform experimental decisions (whether fully or semi-autonomously) the network transfer speed can be a significant limiting factor on experimental throughput.

  — During analysis execution at a remote HPC site the data stream can be about 1 Gbps, but is likely to increase by at least a factor of 10x within three years as more sophisticated methods of visualizing intermediate results are implemented.

  — Significant computational needs beyond edge computing will be required in the future, including GPUs.

- Users at BES facilities often leverage the complete ecosystem of available DOE capabilities. The process of science often involves combining and analyzing data from multiple sources. As data rates and complexity continue to increase, sufficient networking connectivity, bandwidth, and reliability are required to connect measurement facilities, a computing facility, and user home institutions to enable effective data management and analysis. [Section 5]

- BES use of DOE HPC facilities for computational needs does have some limitations: [Section 6]

  — Tying a workflow to resources that are not directly controlled by a BES user facility or experiment means that if the resources are not available (e.g., due to scheduled or unscheduled downtime), it is often not possible to migrate allocation time, research data, or code to execute at the last minute. This could be mitigated with the ability to more easily port allocations between facilities or resources.

  — No universal DOE HPC API exists, which means code written for one facility may not work at others natively.

  — Allocation resources are finite, and if they expire and are not renewed, data must be migrated elsewhere.

  — Having a more stable way to rely on HPC resources at a given facility (or several interchangeable facilities) is desirable.

- BES light source facilities use several DOE HPC facilities for storage and processing. Manual or automated data-movement processes can move data (bulk or streamed) for real-time and offline analysis. For the most demanding computational problems, large-scale computing facilities must be used, including the ALCF, the NERSC Center, and the OLCF. Sending data from a light source facility to the HPC facility requires a data path of sufficient bandwidth. [Section 5.14]

- During operation of the upgraded BES user facilities, it is expected that local computing capabilities will still account for 80% of experiments and will require processing resources

on the order of 5 PFLOPS and storage capacities above 100 PB by 2027. These facilities must continue to use DOE HPC facilities for the other 20% of experiments with the largest computing demands, and will heavily leverage ESnet network resources to facilitate data mobility. [Section 5]

- Multifacility workflows contain a deep tie to networking resources, given that many facilities are separated by geographical location. As BES workflows become more time dependent, particularly those that may try to utilize computational output in a real-time fashion, networking is a critical component that must perform well consistently. If a primary link becomes saturated, or fails, the workflow could also fail. [Section 6]

- In typical closed-loop experiments, raw data is collected, handled, and stored at the location generated using instruments maintained by the experimental teams at their home institutions. Depending on the computational power required for data analysis, preprocessed data are either analyzed locally or shared across divisions/facilities that might be physically located at great distances. Data from multiple sources might need to be gathered at a centralized location for multimodal analysis, and transferred to facilitate AI/ML based decision-making. This data is also archived for long-term access at database facilities, which ideally should provide data access control management. [Section 5.12, Section 5.13]

2.7 Domestic Networking for Local and Wide-Area Data Mobility

- As data rates and volumes continue to grow, greater demands will be placed on facility, DOE national laboratory, and national networks. Each facility is investing considerable effort, and working with laboratory IT staff, to re-design the data architecture components. BES is moving toward models where complex analysis pipelines are being considered, with BES users becoming interested in exploring how to "schedule" network capabilities, and real-time computational resources, to be available during experimentation. [Section 5, Section 6]

- Typical BES user facility networks are designed to handle internal flows at lower rates of speed, and enable a pattern where data flows from the instruments to a centralized data management system: [Section 5]

  — In some instances, the facility networks are equipped with tools that directly facilitate external requirements via the Science DMZ approach, perfSONAR monitoring, or Globus transfer.

  — In other cases, facilities may be integrated to larger laboratory infrastructure that can support these.

  — In a few instances, none of these tools may be available due to security policy. The availability of tools and approaches to support wide-area use cases would be incredibly helpful to the process of science.

- With the growing data volumes, local networking must also be upgraded: typically, the 1 Gbps control networks must be increased to 10 Gbps to allow for more seamless operation. As this occurs, the possibilities of integrating to national-level resources (such as HPC facilities) is also within reach for the majority of BES user facilities. [Section 6]

- BES facility use of DOE HPC resources will continue to expand in the coming years, placing heavy emphasis on ESnet connectivity to manage traffic across the country. DOE HPC facilities must keep pace with upgrades at BES user facilities, particularly the light sources, to ensure ample capacity to support the scientific mission. [Section 5, Section 6]

- By the end of the decade, data aggregated at rates of multiple terabits per second (100–300

Gbps within three years and exceeding 1 Tbps in five years) may flow via ESnet from any of the light sources to any of the DOE HPC facilities. There is an urgent need to upgrade the light source to ESnet connections at some of the laboratories to 400 Gbps and above to match the input bandwidth at ASCR computing facilities within the next two years. [Section 5.14]

- BES users experience complications when sharing data across administrative boundaries due to security policies within and between BES user facilities. Identity management continues to be a friction point, and is exacerbated in a multifacility workflow model where a user may need to authenticate to multiple locations, using multiple identities, on a regular basis. A unified way to handle identity is still highly desirable. [Section 6]

- All DOE national laboratories supporting BES user facilities are connected to ESnet with at least 100 Gbps, with most making plans to increase this to multiples of 100 Gbps, or 400 Gbps, in the two- to five-year timeframe. These laboratories also feature Science DMZ architectures. These vary in implementation, but have commonality in offering high-performance network connections to support data transfer, dedicated DTN hardware, availability of Globus for transfer, and participation in regular perfSONAR testing. [Section 5]

- Local area network (LAN) architectures to support BES user facilities vary in size and scope: [Section 5]

  — A number of the NSRCs may have LAN capabilities between 1 Gbps and 10 Gbps linking instruments and local computational resources. Some NSRCs must still traverse the institutional enterprise network to reach ESnet.

  — The neutron scattering and light source facilities are currently connected at LAN capabilities between 1 Gbps and 100 Gbps for their instruments and local computing capabilities, with plans to move to large capacities in the coming years when upgrades occur at the facilities.

  — BES user facilities maintain strong ties with DOE HPC facilities that may share a parent DOE national laboratory. It is common to see solutions where new BES instruments are integrated directly into DOE HPC facility storage and computation, to ensure ongoing synergistic use of resources. This model is successful, and is encouraged to continue.

- Networking must be able to robustly connect BES user facility instruments to remote computing facilities with low latency. Terabit per second networking and beyond will be required to handle the large amounts of data expected towards the end of the decade. There is a need to upgrade the facility and DOE national laboratory connections to ESnet to keep pace with the expected data rates [Section 5]

  — SLAC expects LCLS-II to require 200 Gbps in 2023 and 1 Tbps by 2028, enabling streaming data transfer from LCLS to ASCR computing facilities.

  — LBNL will be upgrading the laboratory ESnet connection from 2x100 Gbps to more than 1x400 Gbps during the 2025–2026 time period to support use cases such as ALS-U.

  — The APS network will refresh hardware in the two- to five-year time frame to support 400 Gbps in support of the APS-U upgrade.

  — The NSLS-II facility at BNL is connected via a 400 Gbps to the BNL High-Throughput Science Network (HTSN), and BNL will upgrade network capabilities in the one- to two-year time frame to support 400 Gbps connectivity to ESnet.

  — ORNL connects to ESnet via redundant border routers at 100 Gbps. The expectation is that these connections will soon be upgraded to 400 G connections.

- — An emerging pattern for many BES user facilities is the use of streaming data reduction, which can reduce data volume in order to be able to write to disk and stream to DOE HPC facilities over ESnet. Throughput from the detectors could exceed 1 TBps by 2028, and local data reduction will reduce by a factor of 10 to about 1 Tbps over ESnet.

- CINT's most pressing network constraint originates in host laboratory security policy. Currently, CINT utilizes the host laboratory networks for connectivity of instruments and office computers. If an additional layer could be added that would satisfy host laboratory IT requirements (or not be subject to them) and allow easier movement of small data sets between the CINT facilities at the two host laboratories (or remote access to certain experimental tools to off-site, non-badged users), that would be a benefit to CINT. [Section 5.8]

- BES user facilities are increasingly looking into using commercial cloud computing, but there is no official policy or approach. Efficiency of workflow, and cost of service, remain factors in the choice to adopt this approach to computing and storage. A barrier to this adoption is often related to networking: [Section 5]

  - — Ample peering capacity between DOE national laboratories, ESnet, and commercial providers remains a critical service.

  - — The ability to create cybersecurity-approved linkages between commercial cloud enclaves, and that of BES user facilities located at DOE national laboratories, remains an important consideration. For example, enabling secure and persistent direct access and data transfer between commercial cloud resources and on-premise resources at other DOE facilities would significantly reduce the burden of contributing and linking data.

  - — High bandwidth/low latency connections between commercial providers and DOE resources will be important to ensure that data can be ingested into the cloud resources in a timely manner at all DOE facilitates, delivered via ESnet connectivity.

  - — Increased demand from the BES user community to integrate cloud services for both file transfer (e.g., Dropbox, Google Drive, etc.) and for communication (e.g., Slack) into all areas of the data lifecycle and compute workflows.

- Use of cloud storage solutions is increasing, but is still done in an ad-hoc manner. In the long term, the DOE complex will need to answer questions about the long-term archival storage of facility data. If facilities are to be responsible for long-term storage of user data, then there will need to be consideration and commitments regarding facility mission, funding for storage, and network infrastructure for access.[Section 5]

- The MP infrastructure has invested heavily in the use of cloud services. Initially supported by LBNL and NERSC compute and data infrastructure, it is now transitioning to AWS cloud resources to serve the growing user base and usage patterns. [Section 5.18]

  - — MP leverages DOE infrastructure for high-throughput computations and management: the core subset of the data is available via cloud resources, with less commonly used data handled directly through DOE infrastructure. A networking pipeline between the two environments is facilitated by ESnet peering relationships.

  - — No private connectivity exists between NERSC and AWS resources, which presents a significant roadblock in transitioning to a seamless cloud experience. Establishing a high-speed path, along with a security enclave, would significantly reduce the burden of contributing and linking data.

  - — Cloud resources can be cost prohibitive in storing and indexing very large datasets, or performing long-running computations in batch. The MP model leverages DOE infrastructure for high-throughput computations and large-scale data management services:

the core subset of the data is then made available via cloud resources, while less commonly used related data or source raw data can be handled directly through DOE infrastructure.

— Integrating cloud and DOE infrastructure in a seamless manner is possible, and can serve as a model for other projects that may need to scale infrastructure beyond the needs of DOE compute centers, while still making optimal use of DOE resources.

2.8 Emerging Needs

- Networking is critical to achieve autonomous and "smart" characterization tools, and for physics discovery. [Section 5.12]

  — Data captured at synthesis and characterization tools can be processed and the experimental conditions updated.

  — A sufficient amount of computational and networking resources to keep up with experiment flow are required to achieve this vision.

  — As these approaches are computationally expensive, data needs to be transferred from the instrument to computation, and back, to provide instructions during autonomous data capture.

- Autonomous experiment steering at BES light source facilities will be crucial to future experiment success, and requires that a number of gaps in capabilities and infrastructure be addressed: [Section 5.12]

  — Sufficient and reliable bandwidth.

  — Sufficient and sustainable data storage resources.

  — On-demand access to large-scale computing systems.

  — Transparent access to facilities and systems through federated identity and shared data protocols.

  — Software tools and infrastructure to facilitate the development of scientific data workflows.

- Experimental facilities have a challenging task. Equipment acquisition rates are increasing along with the complexity and quantity of scientific questions; thus it is less possible to address questions without parallel advances in experimental design. [Section 5.12, Section 5.13]

  — Scientific instruments, combined with the use of robotics and high-throughput workflows for sample preparation/loading, can acquire measurements at ever-increasing rates and resolutions.

  — Autonomous experiment steering leverages these advantages by addressing the bottlenecks associated with data processing and automatic decision-making by integrating data analytics and advanced AI/ML approaches.

  — Autonomous and automated experiments will become heavily adopted in the coming years, due to increases in efficiency and the elimination of analysis and decision-making bottlenecks.

- The future of neutron scattering science and the study of quantum and functional materials will require large volumes of data, from multiple locations, comprised of simulations and experimental results, needing to be co-analyzed and used in ML. Networks are becoming increasingly important to this overall process, and the shift toward more automatic operation will also require 'programmatic' ways to search and retrieve large amounts of data. [Section 5.15]

- — Future multifacility neutron experiments will require: [Section 5.15]

- — The ability to compute and transfer simulations and training data that may exceed 100 TB.

- — Capabilities to use pre-trained AI to automatically direct experimental outcomes.

- — The ability to support real-time analysis during experimentation.

- — Capabilities to support streaming, bulk movement, or datasets from instruments to reach remote computation and storage.

- — The addition of specialized computational resources and codes.

- — The ability to perform co-analysis with data sets from other experimental and computational facilities.

- — Connectivity via ESnet to bridge multifacility gaps.

- Multifacility workflows that span the NSRCs are developing and deploying methods and tools based on AI/ML to analyze electron scattering information. The expected increases to data velocities and volumes present significant challenges for moving, storing, and processing data. [Section 5.16]

- Policy differences between facilities within BES and ASCR can form large barriers to the adoption of autonomous steering approaches because they impart friction into a process that requires agility. BES user facilities will not share common definitions of security and policy with that of ASCR computing centers, which complicates the design and prototype development activities discussed during the requirements review. The common ground of requiring a baseline form of AAA implementation that spans facilities is desired, and it should apply at a facility, instrument, computation, storage, and data mobility level to ensure friction-free operation. This will prevent circumvention of security and allow for a more automated environment so the community can move away from offline forms of operation. [Section 6]

- BES will employ more use cases for AI and ML in the future, and has experimented with edge computing both in or close to the border of the instrumentation and centrally located at large computing facilities. ESnet maintains a testbed that is exploring edge-computing options, as well as the aforementioned research into smarter APIs to interact with network functions. [Section 6]

- As networking and computational resources become more widely available, BES facility users respond by trying to double the amount of experimentation and data movement that may be accomplished in the same time frame. This scalability will have limitations. however, based on the previous discussions on workforce availability. Scientific output will track closely with technology upgrades, and the overall quality of experimentation will increase as well. [Section 6]

- The rapid advancement in AI/ML algorithms, improved shared workflows, and the advent of exascale computational resources make it possible to create a physically informed virtual platform to perform experimentation: digital twins. Digital twins are computational environments designed to mimic a physical experiment, and will help guide a research question from conception to: [Section 5.12, Section 5.13]

  - — Facilitate the creation of a virtual environment to exhaustively explore experimental controls.

  - — Import experimental read-outs at any time to provide instant synthetic read-outs.

  - — Create data that can iteratively be used as training for model improvements.

  - — Allow for a small subset of simulated experiments to be explored in actual experiments at the BES user facilities.

- Coupling digital twin simulations with experimental steering will become an increasingly relied upon capability at the light sources as the decade progresses. Due to the high computational cost associated with data processing, reduction, and analysis, and of model training on such large datasets, the light sources will need to employ the use of geographically distributed computing facilities and data repositories. [Section 5.13]

- The further development and application of digital twin simulations supporting experimental design and operation requires that a number of gaps in capabilities and infrastructure be filled, including sufficient and reliable networks, data storage resources, computing systems, transparent access to facilities via federated identity and shared data protocols, and software tools to operate across the DOE's distributed resource landscape. [Section 5.13]

- Along with the ongoing development of digital twins, it is expected that HPC systems will be one of the major data sources for future experimentation. Simulations are used in autonomous experiments for synthetic data generation to supplement experimental data. Distributed experiments have been proposed that leverage simulations using HPC infrastructure and parallel data acquisition at multiple facilities internationally. [Section 5.13]

- Autonomous intelligent decision-making (instead of automated decision-making) will be the one limiting factor of future self-driving labs. Most successful approaches are based on HPC-driven uncertainty quantification (UQ); this will require readily accessible allocation of HPC resources and communication infrastructure. [Section 5.12]

- Future usage of computing will require interaction between multimodal experimental and multifidelity theoretical approaches. The urgent need is for "on-demand" computing, interoperable smart workflows across various computing platforms, and long-term cloud storage solutions. [Section 5.12, Section 5.13]

- A number of high performance computation and quantum simulation users and collaborations currently use multiple facilities and locations to accomplish science goals that cross the DOE, commercial, and university complex. [Section 5.17]

  – Calculations and simulations of quantum systems, materials, and molecules use a variety of platforms for hybrid computing, encompassing midsize compute cluster and storage systems owned by universities, and high-performance architectures at DOE national laboratories.

  – A future hybrid computing strategy will require tighter integration of classical and quantum resources for an overall multitiered approach to computing. In this case, a quantum processor becomes a co-processor on which certain tasks are offloaded, but will require real-time data movement. The data transfer would be small, but real-time transfer will be critical between resources.

  – Data accessibility, searchability, and usability is a key issue for quantum simulation, and improvements to enable high-throughput scientific workflows will be required.

  – Data sharing is primarily challenged by data formats and searchability, and for certain classes of simulations the size of the data to be transferred remains a significant issue to solve.

  – There are significant needs for software that performs well on heterogeneous computers. In addition, cloud computing could be an area of need in the future in terms of resources and interoperability.

  – ML surrogates will require high-performance computations of quantum systems. Training

these ML models will require training data collected through massively distributed simulations on leadership-class supercomputing facilities. The dataset sizes, and bandwidth requirements, will require efficient movement of GB to TB of data.
[Section 5.12, Section 5.13]

— Simulations on small clusters generate initial structure information for larger simulations. Data is on the order of 10 MB per file, with tens of thousands of files.

— Simulations at HPC facilities can range from 1 GB to 10 TBs, with the number of files and directories can range from one to tens of thousands, with total data set size around tens of TBs.

— Once exascale computers become more widely available to more users, the size of all of simulation data will likely be multiplied by 100 to 1,000-fold.

3. Review Recommendations

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

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

- Facility Management and Readiness.
- Scientific Data Management.
- Scientific Workflow.
- Computational and Storage Requirements.
- Remote Collaboration and Operational Requirements.
- Multifacility Computational Workflows.
- Domestic Networking for Local and Wide-Area Data Mobility.
- Emerging Needs.

3.1 Facility Management and Readiness

- ESnet will work with BES user facilities and host DOE national laboratories on strategies and implementations of advanced networking to address pending capability upgrades. These discussions will focus on: [Section 5, Section 6]
    - Increasing network capacity to address growing data volumes between BES user facilities, collaborators, and DOE HPC resources.
    - Implementing advanced services that can address scientific requirements for data management, mobility, and remote operation.
    - Understanding and implementing peering policy to reach networks and resources of interest, including commercial cloud providers.
- ESnet and DOE HPC centers will work with the following BES user facilities and host laboratories that have identified new scientific data requirements: [Section 5]
    - **NSLS-II at BNL:** over the next decade, data generation rates will reach the exabyte per year range, coupled with the need to utilize hundreds of PFLOPps of computing. BNL will upgrade network capabilities in the one- to two-year time frame to support 400 Gbps connectivity, and beyond.
    - **LCLS-II at SLAC:** the LCLS-II upgrade is scheduled to come online in 2023, and will increase data throughput to 200 GBps. Future planned upgrades are expected to increase the throughput to multiple TBps by 2028. SLAC anticipates upgrading ESnet connections to keep pace with the expected data rates.
    - **ALS at LBNL:** the new ALS-U storage ring installation will start in 2025, and LBNL will be upgrading the laboratory ESnet connection from 2x100 Gbps to more than 1x400 Gbps during the 2025 to 2026 time period.

- **APS at ANL:** the APS-U project will replace the APS storage ring beginning in April 2023. The ANL and APS networks will refresh hardware in the two- to five-year time frame to support 400 Gbps in support of the upgrade.

- **HFIR and SNS at ORNL:** The SNS-FTS and STS will add new suites of instruments capable of a baseline expected data volumes of 20–30 TB of raw data per day, with orders of magnitude increases in the 5- to 10-year time frame. ORNL connects to ESnet via redundant border routers at 100 Gbps. The expectation is that these connections will soon be upgraded to 400 G connections.

- **CFN at BNL, with NCEM at LBNL:** A planned collaboration between CFN and LBNL will explore a multifacility workflow operating advanced camera technology generating data rates of 50 GBps and requiring 1 PB per year storage. New STEMs could require as much as 4 PB per year.

- **CINT at SNL:** data volumes of new instruments will push the boundaries of current capabilities for sharing data from the facility to collaborators and DOE HPC facilities. ESnet will collaborate with CINT to review security policies and instrument network design, and offer sensible approaches to facilitate data mobility.

- **CNMS at ORNL:** instruments within the facility do not feature advanced and automated pipelines to utilize storage, computation, and data mobility approaches. ESnet will work with ORNL to suggest strategies to integrate and accelerate workflows.

- **TMF at LBNL:** there is currently not a facility-wide compute and storage strategy to link instruments to users and collaborators; this has exacerbated the problems related to data sharing and curation. ESnet will work with TMF and LBNL to suggest strategies to integrate and accelerate workflows.

- **The Materials Project at LBNL, NERSC, and AWS:** to further the mission of integrating cloud and DOE infrastructure in a seamless manner, ESnet will continue to work with LBNL on network peering and service approaches to improve the scientific workflow. This includes exploring high bandwidth/low latency connections between AWS and LBNL/NERSC, as well as understanding performance bottlenecks that exist between the infrastructure and the users.

- ESnet will continue to work with major facilities to establish backup networking options to ensure ongoing operation of multifacility workflows. Future considerations could be establishing overlay networking, similar to other science communities such as HEP, to separate out critical science traffic and allow for more finite control over traffic destined for BES collaborations. [Section 6]

3.2 Scientific Data Management

- ESnet, ASCR, and BES will continue to work to understand the implications of increased data rates from facilities through yearly engagements. The estimates in this report show that within 10 years aggregated data rates of multiple terabits per second flowing across ESnet from any of the light sources to any of the DOE HPC facilities may be possible. There is an urgent need to upgrade the light source to ESnet connections at some of the laboratories to 400 Gbps and above, to match the input bandwidth at ASCR computing facilities within the next two years. [Section 5, Section 6]

- ESnet will work with BES user facilities as they design, develop, and implement common APIs for accessing network and computing resources, parallel data transfer tools, high-fidelity data transfer, network performance monitoring, reservations, and dynamic network provisioning. [Section 6]

- BES should work to define a uniform data pipeline standard that spans DOE user facilities. In doing so, it will be possible to eliminate friction caused by differences in metadata associated with different data types and different levels of data manipulation. To reach a stable and universal environment, standard data formats, software development framework, sample tracking, metadata capturing, and labeling are required. [Section 5.14, Section 5.15, Section 5.16]

- BES will continue to work on the ongoing design, development, and adoption of the Bluesky framework: a way to streamline DAQ and ease the process of collecting and comparing data. Given that a number of BES light sources plan to implement the Bluesky databroker/tiled software for data management, in order to move to a consistent "data API," future integration with network-layer services may be possible when using remote DOE HPC resources across ESnet. [Section 5.14]

3.3 Scientific Workflow

- The five BES light source facilities, coordinated through the five-way LSDCSC, will engage with ESnet on networking topics,. ESnet can advise this group as it works toward a unified vision for the distributed data infrastructure to enable user science, the DISCUS. [Section 5.14]

- ESnet will continue to participate in synergistic community activities that span DOE SC as user facilities leverage the complete ecosystem of technology resources to advance scientific missions. An emerging common pattern in the process of science often involves combining and analyzing data from multiple sources. As data rates and complexity continue to increase, sufficient networking connectivity, bandwidth, and reliability are required to connect measurement facilities, a computing facility, and user home institutions to enable effective data management and analysis. [Section 5]

- ESnet will work with BES user facilities that are still experiencing substantial bottlenecks between the data streams emanating from instruments and the feedback from theoretical and simulation insights. The goal will be to directly couple instrument data to simulations, in real time, to better assist the experimenter and enable automation and autonomous materials and physics discovery platforms. [Section 5]

3.4 Computational and Storage Requirements

- ESnet, the BES light source facilities, the ALCF, NERSC, and the OLCF must continue to collaborate on network upgrade strategies. [Section 5.14]

  — After a number of BES facility upgrades, it will be impractical and unreasonable to support the scale of computing required with local resources only. BES user facilities are increasingly partnering with DOE HPC facilities to deliver a new model of computing, tightly coupling experiment instruments and supercomputers to accelerate scientific discovery.

  — BES user facility's use of DOE HPC resources will continue to expand in the coming years and is estimated to top aggregated data at rates of multiple terabits per second. This will place heavy emphasis on ESnet connectivity to manage traffic across the country.

  — BES user facilities of the future will rely on a pipeline that facilitates data transfer from initial source to different layers of computation. These computational capabilities could be edge computers close to experiments that will facilitate AI and ML tasks, as well as midcapacity and HPC infrastructure in national labs to produce simulations and operate

analysis workflows. These resources will both guide experimental investigation and accelerate materials and fundamental physics discoveries.

— Networking must be able to robustly connect BES user facility instruments to remote computing facilities with low latency.

3.5 Remote Collaboration and Operational Requirements

- ESnet will continue to support BES user facilities, as they expand their support for remote operation of experimentation and computation. The deployment of remote access tools enabled experimenters to run measurements remotely during the COVID pandemic, and led to an overall increase in remote users, making this functionality a critical infrastructure component. [Section 5]

- ESnet will explore mechanisms to support the following aspects of remote use of BES user facilities: [Section 5]

    — Supporting the network to enable the remote operation of instruments; either fully automated or through communication over telecommunications tools.

    — Building high bandwidth/low latency paths to support computational and storage needs between facilities.

    — Enabling peering to cloud providers to support communication and collaboration tools.

    — Encouraging the deployment of user-focused data portals to allow for data mobility.

3.6 Multifacility Computational Workflows

- Multi-facility workflows contain a deep tie to networking resources, given that many user facilities are separated by geographical location. As BES workflows become more time dependent, particularly those that may try to utilize computational output in a real-time fashion, networking is a critical component that must perform well consistently. If a primary link becomes saturated, or fails, the workflow could also fail. [Section 5.14, Section 5.15, Section 5.16]

- ESnet will continue to advise BES as the use of multimodal data increases and crosses the boundaries between user facilities and DOE HPC resources. Multimodal data requires more sophisticated data processing and increases in computing capabilities. [Section 5.12, Section 5.13]

- BES should work to define common standards, and shared workflows, for data across the NSRCs. Doing so will provide an effective data solution as well as enable cross-center data sharing, augmentation, and manipulation. The lack of data pipeline standards affects scientific progress. [Section 5.16]

3.7 Domestic Networking for Local and Wide-Area Data Mobility

- ESnet will continue to support R&D efforts within the BES community that are advancing use of network technologies to support scientific workflows. For complex analysis pipelines or high data rate experiments, it is anticipated that future users of BES facilities will have to schedule network capabilities and real-time computational resources to be available during the time of experiments. The BES community is coordinating with ESnet and DOE HPC facilities on possible solutions. [Section 5]

- ESnet will continue to support the use of commercial cloud resources by BES user facilities and their users. Use of cloud storage solutions has increased, but is still done in an ad-hoc manner. In the long term, BES will need to answer questions about the use of cloud resources for computation, storage, communication, and data sharing. [Section 5]

- ESnet will work with BES user facilities to increase the adoption of the Science DMZ paradigm, and the affiliated technologies that support network measurement (perfSONAR, DTNs), and data mobility (Globus). A number of BES user facilities and their parent DOE national laboratories have adopted these tools, but there remain gaps in automated approaches to use them regularly to support scientific workflows. [Section 5]

3.8 Emerging Needs

- ESnet will collaborate with the BES community and the DOE HPC centers to explore the adoption of digital twins. The further development and application of digital twin simulations supporting experimental design and operation requires that a number of gaps in capabilities and infrastructure be filled, including sufficient and reliable networks, data storage resources, computing systems, transparent access to facilities via federated identity and shared data protocols, and software tools to operate across the DOE's distributed resource landscape. [Section 5.13]

- ESnet will work with BES user facilities, and DOE HPC facilities, as autonomous experiment steering is more widely adopted. Given the implicit use of networks to join experimental facilities with DOE HPC resources, ESnet can provide high bandwidth, low latency network connections to ensure experimental success. [Section 5.12]

- ESnet maintains a field-programmable gate array (FPGA) testbed and is also experimenting with edge-computing approaches, and will be working to figure out ways to deploy these resources close to sites so that it may be possible to perform similar functions for facilities and science that could benefit. The BES community is considering many different approaches to its computational needs and is invited to collaborate on scientific workflows that require custom computing resources (such as Xilinx or Altera) to perform data reduction, using AI and ML to steer experimental direction, or other emerging use cases. [Section 6]

- ESnet will continue to work with BES on a strategy to support quantum computing and hybrid computing requirements from a network perspective. [Section 5.17]

  - It is expected that a hybrid quantum-classical model, where part of the calculations will be carried out on classical architectures and part on quantum hardware, will persist for a number of years, implying a transfer between the two platforms.

  - A number of high performance computation and quantum simulation users and collaborations currently use multiple facilities and locations on a variety of platforms for hybrid computing to accomplish science goals that cross the DOE, commercial, and university complex. Data sharing will be required in this environment.

4 Requirements Review Structure

The requirements review is designed to be an in-person event; however, the COVID-19 pandemic has changed the process to operate virtually and asynchronously for several aspects. The review is a highly conversational process through which all participants gain shared insight into the salient data management challenges of the subject program/facility/ project. Requirements reviews help ensure that key stakeholders have a common understanding of the issues and the potential recommendations that can be implemented 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 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, MAN, and wide-area network [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 BES 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 BES of the DOE SC. These case studies were discussed virtually between March and September 2022.

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:

- Advanced Light Source (ALS)
- Advanced Photon Source (APS)
- National Synchrotron Light Source II (NSLS-II)
- LCLS
- SSRL
- HFIR and SNS
- CFN
- Center for Integrated Nanotechnologies (CINT)
- CNM
- CNMS
- TMF
- Autonomous Experiment Steering for BES Facilities
- BES Design and Development of Digital Twin Strategies
- Multifacility Experimentation and Analysis Workflows: X-ray Light Source Perspective
- Multifacility Experimentation and Analysis Workflows: Neutron Scattering Perspective
- Multifacility Experimentation and Analysis Workflows: NSRC Perspective
- Use of the ESnet for Quantum Simulations of Materials and Molecules
- The MAterials Project (MP): Status and Future Directions

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 Advanced Light Source (ALS)

The ALS is an electron storage ring–based synchrotron radiation facility that is supported by the Department of Energy's Basic Energy Sciences program (DOE BES). The ALS started operation in 1993 and since then has been upgraded continuously to remain one of the brightest soft X-ray sources in the world. The ALS is optimized

for X-ray spectroscopy, microscopy, and scattering using intense beams from soft X-ray undulator sources but also serves a broader community conducting research using hard X-rays, infrared (IR), and vacuum ultraviolet (VUV) radiation from superconducting magnets, conventional dipole magnets, and insertion devices.

5.1.1 Discussion Summary

- The ALS is a user facility with 40 different beamlines and 2100 scientific users per year from all over the world.

- Users may only be granted experimental access on a limited basis (e.g., once per year).

- It is common for users to receive only beamtime once or twice a year. With the complexity of experiments increasing and the data rates ramping up, it will be critical to prepare and test the data streaming and analysis beforehand if possible. A development like this will need access to compute and analysis resources and be able to schedule the required resources in advance.

- The biggest change to our facility will be the transition to a diffraction-limited light source. We currently expect the new ALS-U storage ring installation starting in 2025.

- The ALS Computing Program, beamline scientists, and users, actively develop and support data management tools that assist during the full data lifecycle from acquisition to storage and analysis. These tools include:

  — software to detect, package, and transfer data.

  — database systems that register data and metadata, to allow search and organization.

  — workflow tools that allow both user-driven and automated processing that leverage local and remote HPC resources.

- The ALS web portal allows users to view or download results. The web portal can Alsop seamlessly move data into an interactive analysis environment for further processing.

- The ALS has adapted during the COVID pandemic to move from the primary mode of operating with users on-site, to one that facilities a number of new remote capabilities. As such, the traditional mechanism of copying the data to removable media is no longer an option, and thus more data processing is performed at the ALS or migrated off-site to home institutions, or other computational resources through a variety of means (Globus, Google Doc, file transfer protocol (FTP)).

- Several end stations have adopted Bluesky framework, a way to streamline DAQ and eases the process of collecting and comparing data. It is expected that more will adopt in the future, to match the approach of other BES facilities.

- A typical workflow for a user at the ALS involves:

  — beamtime (in person or remote).

  — use of integrated management tools during data acquisition.

  — analysis using local resources, or data mobility to remote resources.

- Future adaptations to the generalized workflow used to support BES science will involve:

  — integrated software to manage data movement, processing, and analysis.

  — more automated methods to handle adaptive/autonomous experiment steering (e.g., integration with simulation capabilities).

  — categorizing and sharing of data with other national resources.

  — increases in data volume due to upgraded capabilities.

- Data rates and volumes range from several MB per month to 3GBps currently. The largest data producers at the ALS are the scattering beamlines, the tomography beamline, ptychography, and the newly developed X-ray photon correlation spectroscopy (XPCS) beamline. In addition to the currently high data rate beamlines, we anticipate various new detectors at other beamlines.

- The current storage capacity at beamlines is insufficient to handle the data and we are in the process of expanding our data movement platform, based on Globus, to more beamlines to move the data to NERSC for processing and storage.

- The ALS has a small institutional storage cluster (a few hundred TB) that can be used for temporary storage, along with several small compute clusters and GPU machines that can be used for analysis on a FCFS basis. The ALS is working to develop software that better manages computational tasks using these resources.

- The ALS makes use of several DOE HPC facilities for the storage and processing. An automated data movement process can move data (bulk, or streamed) for a limited subset of beamlines to HPC facilities for processing and storage.

- Taking advantage of available (local or remote) compute and analysis resources will become an integral part of experimental planning in the very near future. In order to deploy efficient analysis and machine learning (ML) approaches at the time of the experiments, workflows will have to be established, and ML algorithms will have to be trained on existing or simulated data.

- For complex analysis pipelines or high data rate experiments, the users will have to schedule network capabilities and/or real-time computational resources to be available during the beamtime. We are in discussion with ESnet and NERSC to make this a reality.

- A few ALS beamlines utilize Globus to facilitate data movement. The ALS is in the process of expanding this capability, to enable faster data sharing to other resources (e.g., DOE HPC centers, home institutions), reached via ESnet and other R&E network providers.

- ALS high data rate machines are connected to the facility using 10 Gbps networking, and the less data-intensive beamlines are connected with 1Gbps capacity. Many beamlines still rely on local storage and computation, implying that most network traffic is localized to a specific experimental network. There are plans to upgrade all beamlines to a minimum of 10 Gbps, with some connected at 40 Gbps, to support ALS-U in 2025–2026.

- LBNL will be upgrading the laboratory ESnet connection from 2x100 Gbps to more than 1x400 Gbps during the 2025–2026 time period to support use cases such as ALS-U. Upgrades to the Science DMZ network, include direct connections to user facilities like the ALS, are planned.

- The ALS is looking into using commercial cloud computing but nothing has been developed or decided. The ALS makes use of the NERSC Spin service for hosting containers providing web and other services to its users.

5.1.2 ALS Facility Profile

The 1.9 GeV ring hosts world-class end stations and instrumentation at more than 40 beamlines and serves nearly 2000 users who publish more than 900 publications per year and conduct basic, applied, and industrial research in energy science, earth and environmental science, materials science, biology, chemistry, and physics. Our mission is to advance science for the benefit of society by providing our world-class synchrotron light source capabilities and expertise to a broad scientific community. The ALS is the primary BES-funded soft X-ray facility in the US, and our ambition is to provide the US and the international community with world-leading X-ray capabilities that enable consequential scientific discoveries and lead to a detailed understanding of laws governing natural processes and the properties of engineered systems.
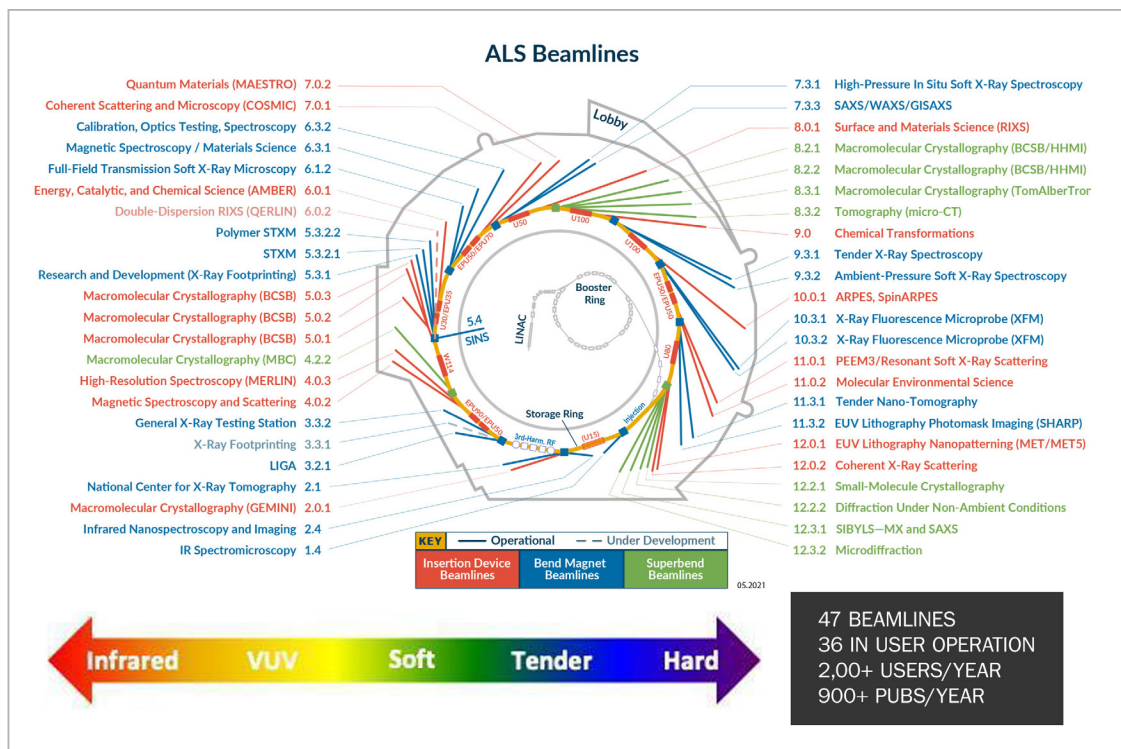
Figure 5.1.1: ALS beamlines instrumentation

5.1.2.1 Science Background

The ALS Computing Program, together with beamline scientists and users, actively develops and supports data management tools that assist users and staff during the full data lifecycle from acquisition to storage and analysis as shown in Figure 5.1.2. These tools include software that runs at the beamline to detect, package, and transfer new data; database servers that register data with its associated metadata to allow subsequent search and organization; and workers, message-passing systems, and workflow tools that allow both user-driven and automated processing to be launched on data sets using a combination of local and remote resources (such as supercomputing centers) as data arrives. Stored metadata will cover information about experiments, the beamline environment, and pointers to raw and derived data that is stored in a variety of locations. An easy-to-use web portal allows users to view or download results during or after their beamtime. The web portal also allows users to seamlessly move their data into an interactive analysis environment (e.g., Jupyter, or other technique-specific web-based tools) for further processing. We expect the data and analysis to be accessed from within the ALS and across the country. In conjunction with the MLExchange project we are developing such capabilities right now. Either the data, the analysis or both will be containerized and moved over the network to the appropriate computational or storage location.
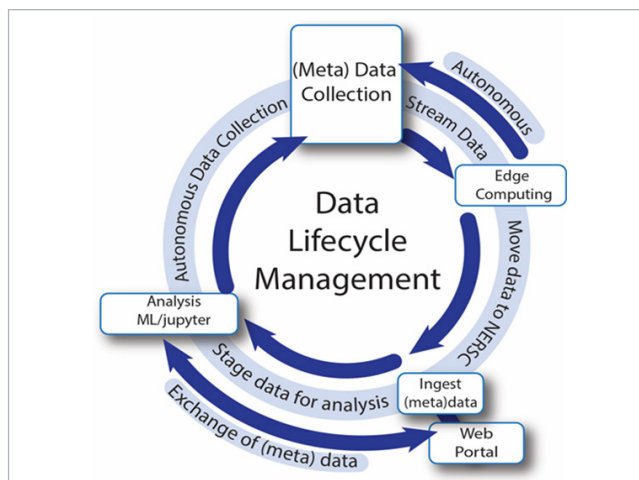
Figure 5.1.2: Data lifecycle currently developed at the ALS in collaboration with other synchrotrons

At various points along the workflow, instrumentation will be added to monitor system performance (network, disk, processing.) A seamless experience for the user will require not only network bandwidth but also fast and accurate feedback about the existing network availability. Facility-based services responsible for orchestrating data transfers and processing jobs will be able to route jobs and data based on the availability of underlying services. It would be interesting if the network itself was smart enough to evaluate compute time, network transfer speed, and availability and create the smartest plan to move the data analysis to the fastest path.

5.1.2.2 Collaborators

The ALS is a user facility with 40 different beamlines and 2100 scientific users per year from all over the world. As shown in Figure 5.1.3, the scientific user community is diverse, ranging from material science to earth and environmental science. On the right side of Figure 5.1.3, the geographic origin of ALS users is shown. A portion of the user community is directly from LBNL, these users are considered local with direct access to local ALS resources and data storage. A large portion of the users are from within California, with a smaller portion of users from the rest of the United States. In addition to national users, the ALS is also the research location of choice for international researchers.
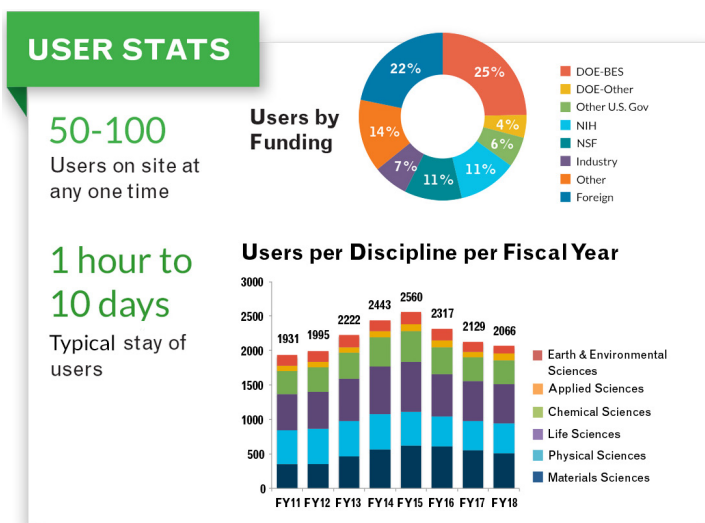
Figure 5.1.3: Users by discipline and geography

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
US UNIVERSITY- BASED PIS	Both	data portal, data transfer (Globus, sftp, cloud storage), portable hard drive	1.MB–10 TB	ad hoc	N	N/A
US NATIONAL LABS BASED PIS	Both	data portal, data transfer (Globus, sftp, cloud storage), portable hard drive	1 MB–10 TB	ad hoc	N	N/A
INTERNATIONAL PIS	Both	data portal, data transfer (Globus, sftp, cloud storage), portable hard drive	1 MB–10 TB	ad hoc	N	N/A

Table 5.1.1: ALS Collaboration Space

In Figure 5.1.4, the distribution of users across the US is shown. The majority of users are coming from California as mentioned before. The rest of the user community is broadly distributed across the US, including a large number coming from the east coast. Scientific users used to exclusively travel to the ALS for beamtime. This process has changed during the last two years due to the COVID pandemic. A significant number of users are using the newly established remote capabilities at the ALS. This also means that the traditional way of copying the data onto thumb drives or hard drives during or directly after the beamtime is not an option anymore for many users. For many beamlines significant data processing is performed at the ALS (whether users are physically at the ALS or using a remote desktop connection to an ALS computer). Whether users are remote or in person, data is made available to them for copying to their home institutions through a variety of means (Globus, Google Doc, FTP).
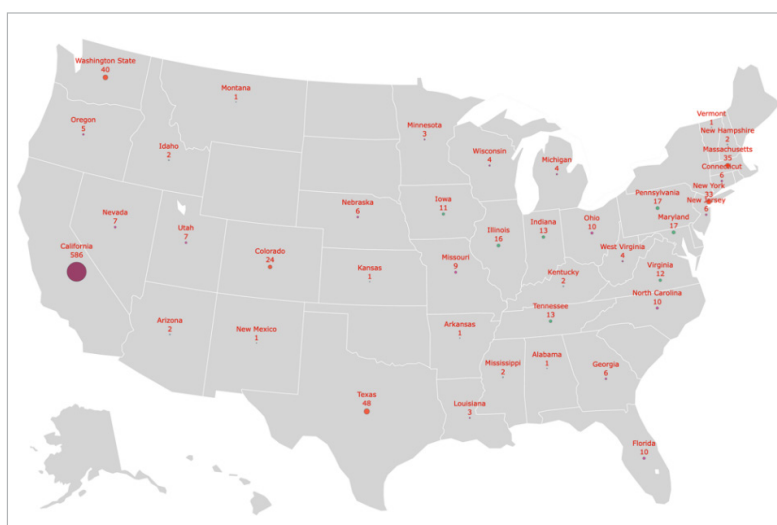
Figure 5.1.4: Geographic map and number of ALS scientific users in the US in 2021

A majority of users are from academia, with the next largest groups at national laboratories. Due to the emphasis on basic research, this is understandable. The ALS has a variety of collaborations and connections with Energy hubs, CAMERA, EFRCs, SBIRs, and other research centers across the country.
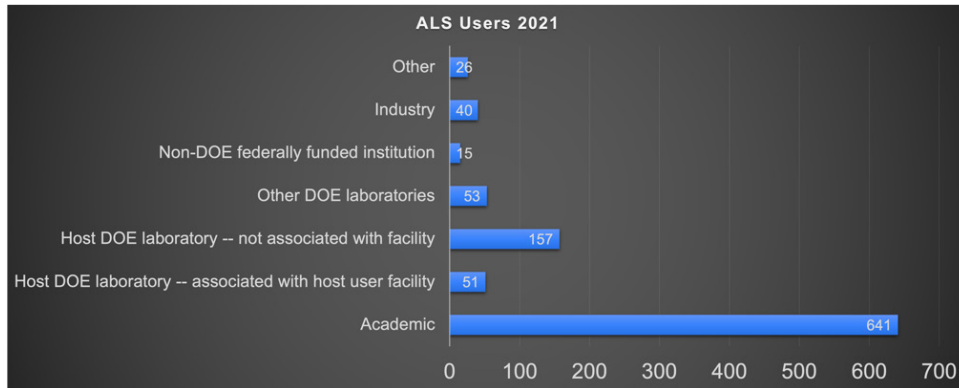
Figure 5.1.5: Distribution of users' different origins

Figure 5.1.6: A number of ALS collaborations and partnerships

5.1.2.3 Instruments and Facilities

The ALS is home to 40 beamlines as described earlier in question 1. Each instrument has a variety of detectors and control computers. The data rates and data volumes range from several MB per month to 3GBps currently. The largest data producers at the ALS are the scattering beamlines, the tomography beamline, ptychography, and the newly developed XPCS beamline. In addition to the currently high data rate beamlines, we anticipate various new detectors at other beamlines. The current and projected data amounts for ALS are shown in Figure 5.1.7.
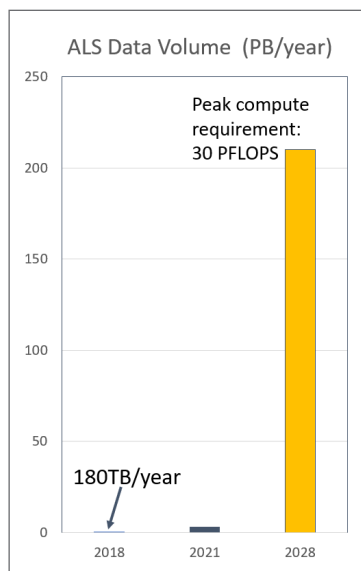
Figure 5.1.7: Data storage and compute estimates for the ALS before and after the upgrade

The current storage capacity at beamlines is insufficient to handle the data and we are in the process of expanding our data movement platform, based on Globus, to more beamlines to move the data to NERSC for processing and storage. At the ALS we have a small storage cluster of a few hundred TB to be used as a buffer. The ALS is also home to several small compute clusters and GPU machines. We are in the process of developing software to distribute smaller compute tasks from individual beamlines to be deployed across these machines. With respect to the network, our high data rate machines are connected to our server room with 10G, and the less data-intensive beamlines are connected with 1G. However, most beamlines still use local beamline storage and beamline compute to preprocess their data directly at the beamline.

The biggest change to our facility will be the transition to a diffraction-limited light source. We currently expect the new ALS-U storage ring installation starting in 2025. The ALS-U is an upgrade of the ALS that will endow the ALS with revolutionary X-ray capabilities. The ALS has been a global leader in soft X-ray science for more than two decades. Recent accelerator physics breakthroughs now enable the production of highly focused beams of soft X-ray light that are at least 100 times brighter than those of the existing ALS. Applying this technology at the ALS will help us to better understand and develop the new materials and chemical systems needed to advance our energy, economic, and national security needs in the 21st century, securing the United States' world scientific leadership for decades to come. The upgraded ALS will occupy the same facility as the current ALS, replacing the existing electron storage ring and leveraging $500 million in existing ALS infrastructure and experimental systems. The new ring will use powerful, compact magnets arranged in a dense, circular array called a multibend achromat (MBA) lattice. In combination with other improvements to the accelerator complex, the upgraded machine will produce bright, steady beams of high-energy light to probe matter with unprecedented detail. The improved capabilities of the upgraded ALS will enable transformative science that cannot be performed on any existing or planned light source in the world. This new science includes 3D imaging with nanometer-scale spatial resolution and measurement of spontaneous nanoscale processes with time scales extending from minutes to nanoseconds—all with sensitivity to chemical, electronic, and magnetic properties. Moreover, the beam's high coherence as shown in Figure 5.1.8, will enable new classes of optical techniques that will provide the groundbreaking sensitivity and precision needed to detect the faintest traces of elements and subtle electrochemical interactions on the scale of nanometers.
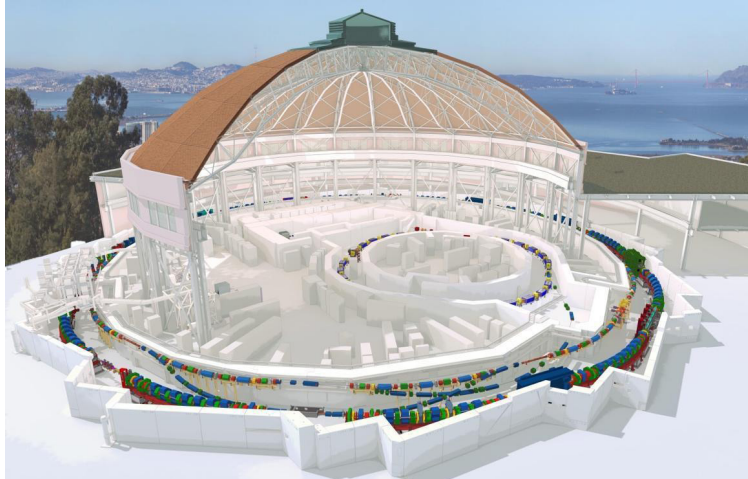
Figure 5.1.8: This cutaway view of the Advanced Light Source shows new equipment (colored rings) that will be installed during the ALS Upgrade project

The new science the ALS-U will require high data rate analysis and processing as shown by the amount of data estimated in Figure 5.1.8. The software framework we are developing is described below.
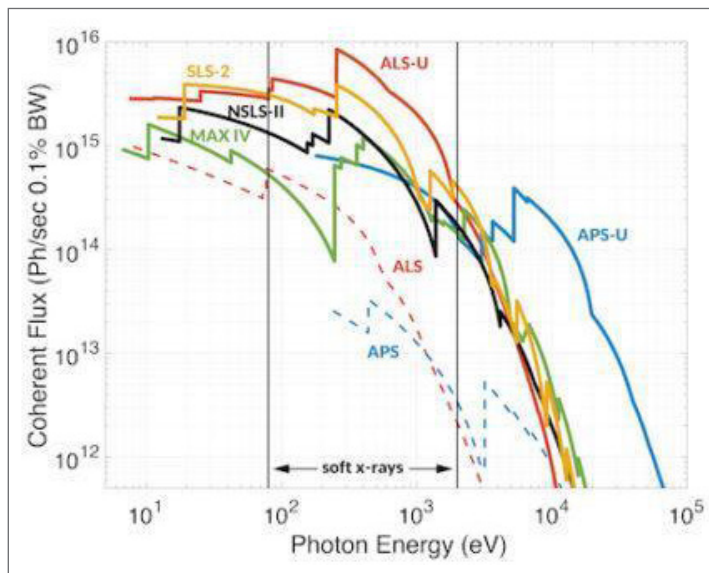
Figure 5.1.9: Coherent flux vs. photon energy of the US and international soft X-ray light sources

5.1.2.4 Generalized Process of Science

The current and future generalized process of science varies depending on the size and volume of data produced at the beamline. However, below is a more or less typical workflow for knowledge discovery which also includes future developments and current projects we are working on.

It is common for users to receive only beamtime once or twice a year. With the complexity of experiments increasing and the data rates ramping up, it will be critical to prepare and test the data streaming and analysis beforehand if possible. A development like this will need access to compute and analysis resources and be able to schedule the required resources in advance.

Workflow for a User at the ALS

- Present to Two Years

  — Beamline users in person and remote.

  — Most beamlines use local resources for storage and processing.

  — Rollout of integrated movement, processing and analysis tools.

  — Rollout of adaptive/autonomous scanning tools.

- Next Two to Five Years

  — Increased rollout of movement, processing and analysis tools.

  — Increased integration with Molecular Foundry.

- Beyond Five Years

  — ALS-U will be live, necessitating increased storage and processing. In addition, flexible use of storage and processing resources will necessitate more movement of data outside of the ALS.

A typical ALS user is assigned 24- to 48-hour shifts per cycle at a beamline. To maximize time, generally multiple people from a research team will be present, taking shifts around the clock. During beamtime, user activities involve:

- Sample preparation and placement
- Beamline alignment
- Scans
- Data analysis of the scan
- Potential realignment and rescan
- Potential additional sample preparation and placement

Remote users rely on beamline staff for sample placement and beamline alignment. Some beamlines allow users to scan remotely, while others depend on beamline staff. Depending on the beamline, data analysis may be performed on computers at the beamline (majority), clusters available at the ALS and at NERSC.

- (Current and future development) Taking advantage of available compute and analysis resources will become an integral part of experimental planning in the very near future. In order to deploy efficient analysis and ML approaches at the time of the experiments, workflows will have to be established, and ML algorithms will have to be trained on existing or simulated data. The ALS is currently developing such a framework as part of the MLExchange project. In Figure 5.1.10 the general process for ML augmented analysis is shown. A network is trained before the beamtime on existing data and then deployed during the actual experiment.

- (Current and future developments) For complex analysis pipelines or high data rate experiments, the users will have to schedule network capabilities and/or real-time compute resources to be available during the beamtime. We are in discussion with ESnet and NERSC to make this a reality.

- During the beamtime, the users are either on-site or remote and deploy analysis and ML processes either locally or remote. The ALS has an integrated solution for remote users as shown below in question 5. The deployment location of the analysis algorithm depends heavily on the complexity and compute requirements of the algorithms. CAMERA together with ALS and MLExchange is developing a framework to allow for optimal deployment of algorithms

on devices ranging from edge-computing hardware to compute clusters at leadership-class compute facilities. The network requirements heavily depend on several key factors: the data rate (file-based vs. streaming), the algorithm (iterative, in situ, post hoc), resource constraints (e.g., FPGA vs. GPU, task vs. data parallel), and location of the deployment workflow.

The ALS is testing a beta version of an integrated processing and analysis application framework based on docker containers. During the experiment, the data is moved to NERSC using Globus. For most experiments at the ALS right now, this means a simple file copy and then deployment of analysis software. Some beamlines, such as tomography and ptychography, have more integrated analysis and processing routines.
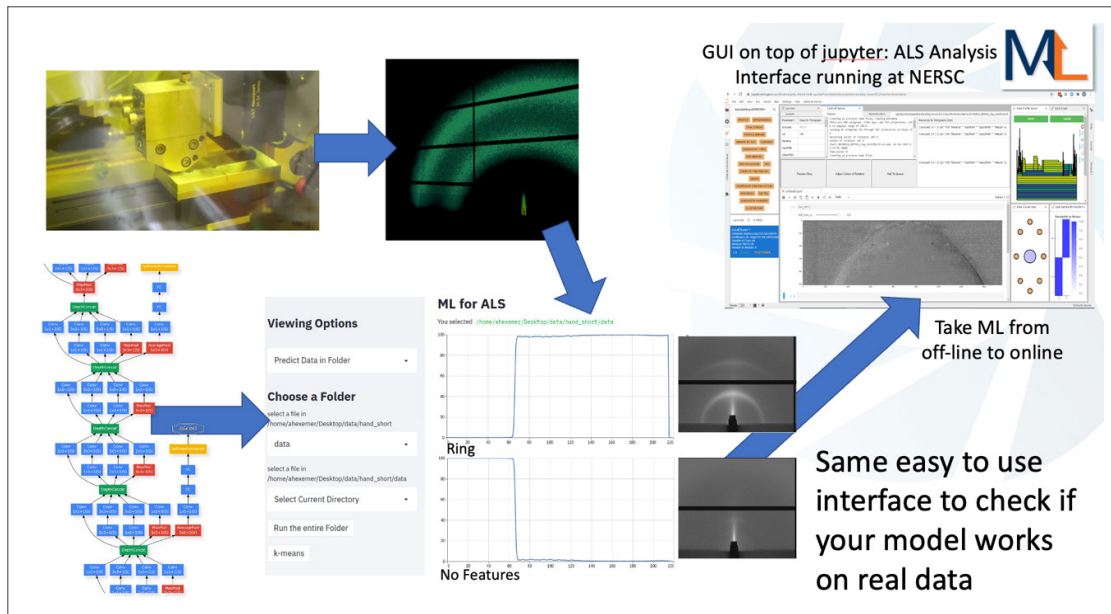
Figure 5.1.10: ML augmented data analysis and processing using MLExchange and CAMERA developments

After the beamtime, most beamlines have their users copy their experimental data and analysis results to hard drives or thumb drives. A few high data rate beamlines at the ALS, such as the tomography beamlines, allow the download of data sets via Globus to their home institute. The ALS is in the process of expanding the number of beamlines that provide better data transfer service. Most users process data at their home institute and will need to move the data using ESnet since the data size will prohibit them from storing on disc. Some users will be able to use Globus to move data to home institutions, while others are prohibited by the security requirements of their institution. Some advanced users leverage the ALS NERSC allocation to further simulate and analyze their data. We are planning to expand such capabilities to other computational centers around the country.
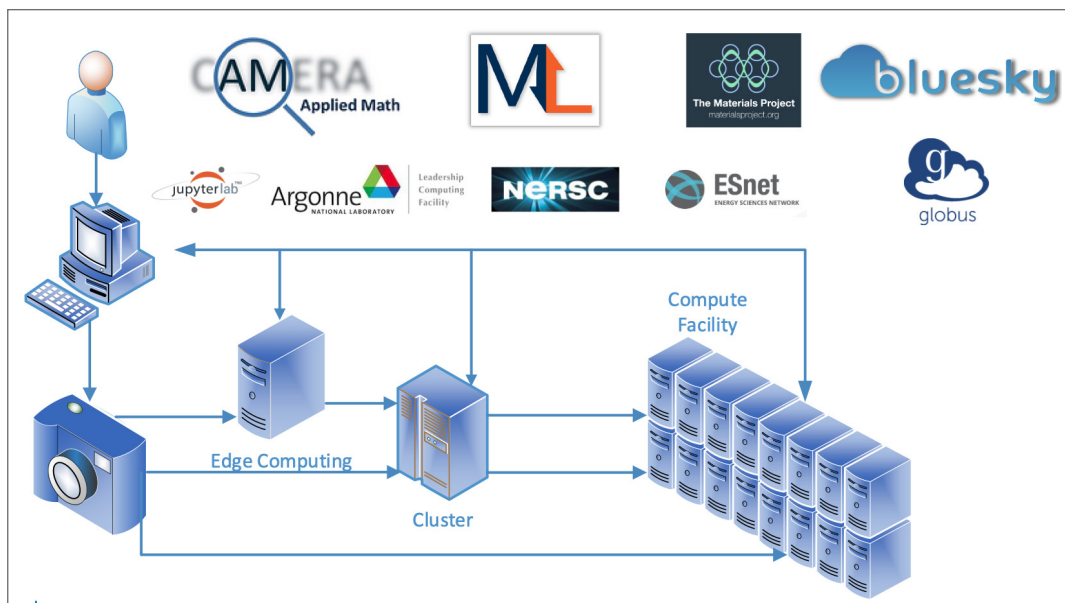
Figure 5.1.11: Different deployment locations for analysis algorithms based on the compute and network requirements of the data and experiment

The U.S. DOE-funded light sources have developed a common vision for computing across the light sources, the DISCUS, and a decade-long roadmap to achieve the vision (see Figure 5.1.12). This vision proposes a transformative computational fabric that covers the full lifecycle of data generated at the BES light sources to accelerate discovery and insight. This vision proposed connecting the 200+ instruments at the light sources to a multitiered computing landscape, including edge, local, campus, and ASCR compute resources, and discoverable data repositories using high-performance, robust feature-rich networks. This fabric would facilitate the full lifecycle of data across the complex, including theory/modeling and simulation, experiment design, data generation at scientific instruments within the light sources, data reduction and processing, analysis and interpretation, and publication and dissemination, serving the 10,000+ light sources users per year.

Figure 5.1.12: DISCUS Vision

5.1.2.5 Remote Science Activities

HPC Facilities

The ALS makes use of several DOE HPC facilities for the storage and processing.

Present–Two Years

An automated data movement process moves data sets for two beamlines (tomography and SAXS/WAXS) to NERSC's Community File System (CFS). This provides storage and the ability to use NERSC processing resources. Over the next two years, more beamlines will be added to this framework covering high-bandwidth techniques such as XPCS, ptychography and dichroism.

Next Two to Five Years

The ALS will use more DOE HPC facilities, including ALCF/OLCF, requiring data movement from ALS to ORNL and ANL. Additionally, workflows involving streaming data to HPC facilities will begin to be implemented in order to provide real-time feedback to ALS users.

Beyond Five Years

The data deluge of the ALS-U upgrade will necessitate the implementation of a wide variety of data reduction techniques requiring more and more reliance on computing resources at beamlines, both within the ALS and at HPC facilities.

Remote Users

The ALS, similar to all the other facilities, has undergone a major transition during the COVID pandemic. As shown in Figure 5.1.13, the total number of remote users has changed significantly from 2019 to 2020.
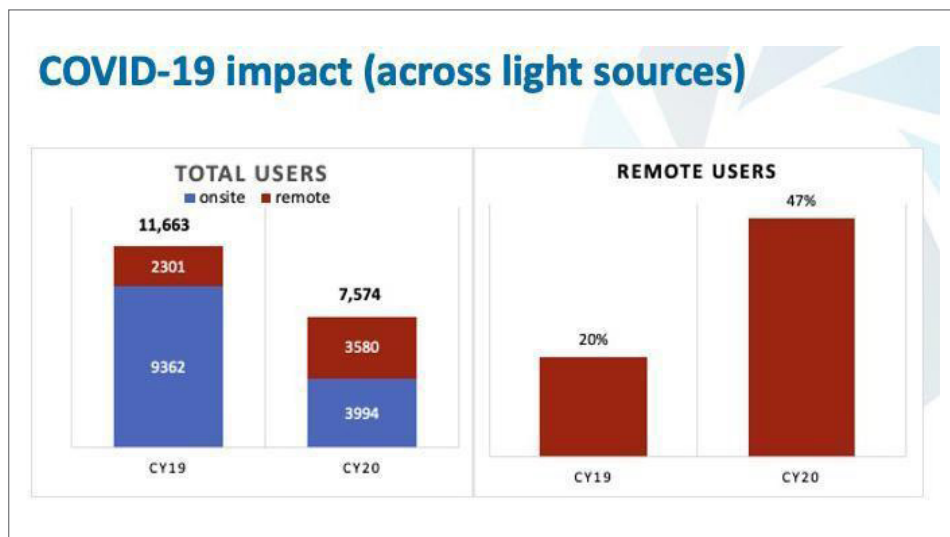
Figure 5.1.13: COVID-19 User Impact at DOE Light Sources

Figure 5.1.14 shows the numbers of ALS users before COVID and now. The ALS has since been developing tools for streamlining remote access to beamline computers and analysis tools.
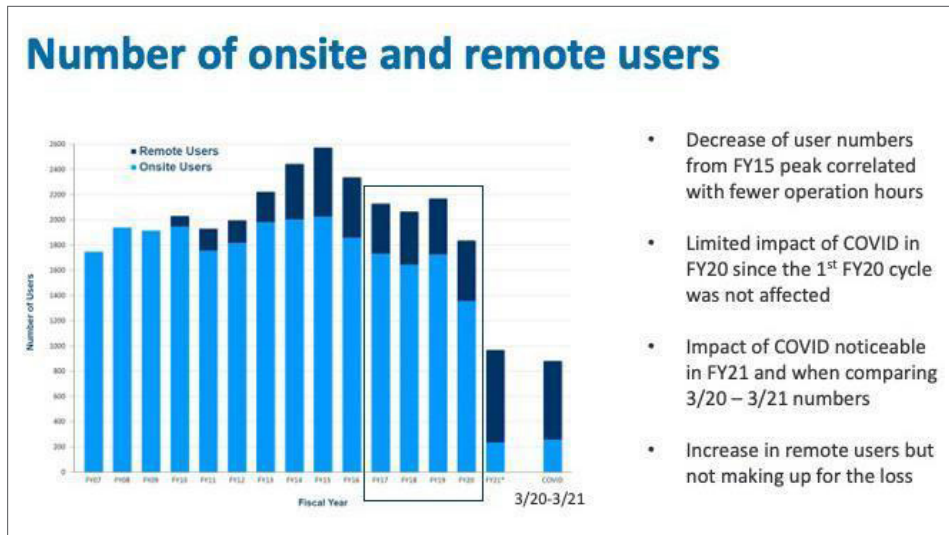
Figure 5.1.14: COVID-19 User Impact at ALS

Scheduling and Remote Desktop

The ALS created a tool that allows beamline scientists to schedule time for users to access beamline computers. It allows users to use VNC to access the beamline computer and control the beamline during their beamtime.
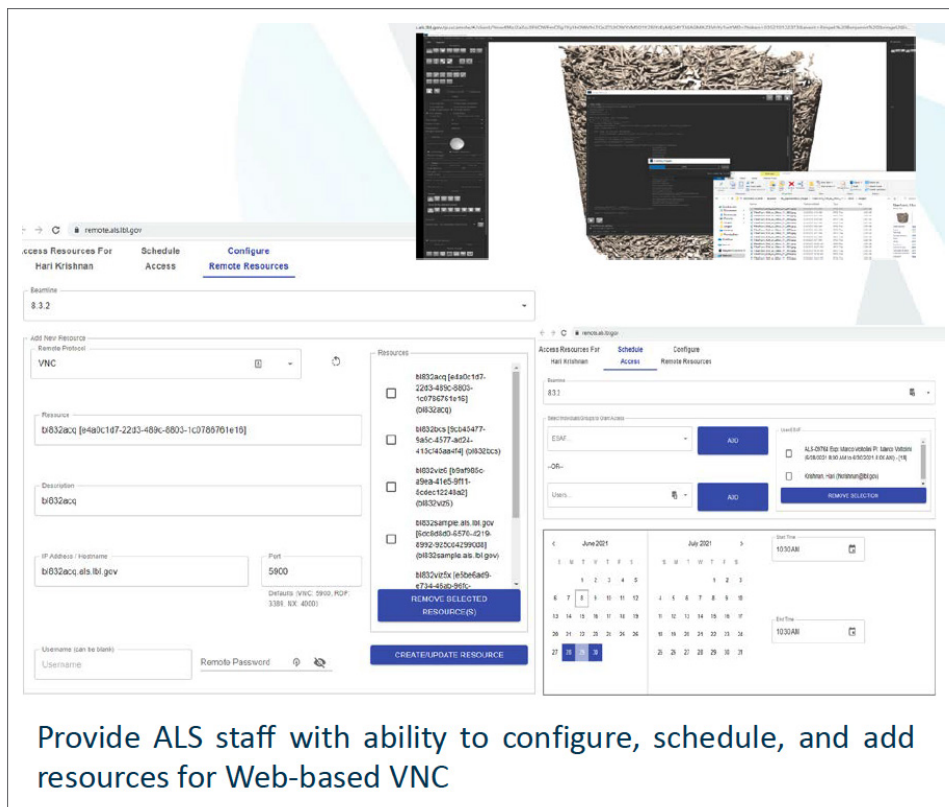
Figure 5.1.15: ALS Scheduling System

Data Portal

The ALS Data Portal, implemented using the cross-facility open-source SciCat scientific catalog project, has been implemented to help beamline staff and users to catalog, label, and find data sets taken at the ALS. With this database in place, users can find and download data sets. Metadata from beamline scans is "ingested" into the data portal automatically as part of the previously mentioned data movement framework.

The tomography beamline is the first to have its data sets ingested into the data portal. The SAXS/WAXS beamline will come online very soon. Over the next two years, as beamlines come online with the data movement framework, they will also come online with the data portal. With the data portal in place, analysis workflows will be developed that take data sets known to the data portal and run processing routines on them. The results of those analysis routines will also be ingested into the data portal and associated with the raw data sets from which they were calculated. The data portal software allows users to download data sets over HTTP. In the near future, we plan to provide users with a convenient way to invoke Globus transfers from the data portal application.
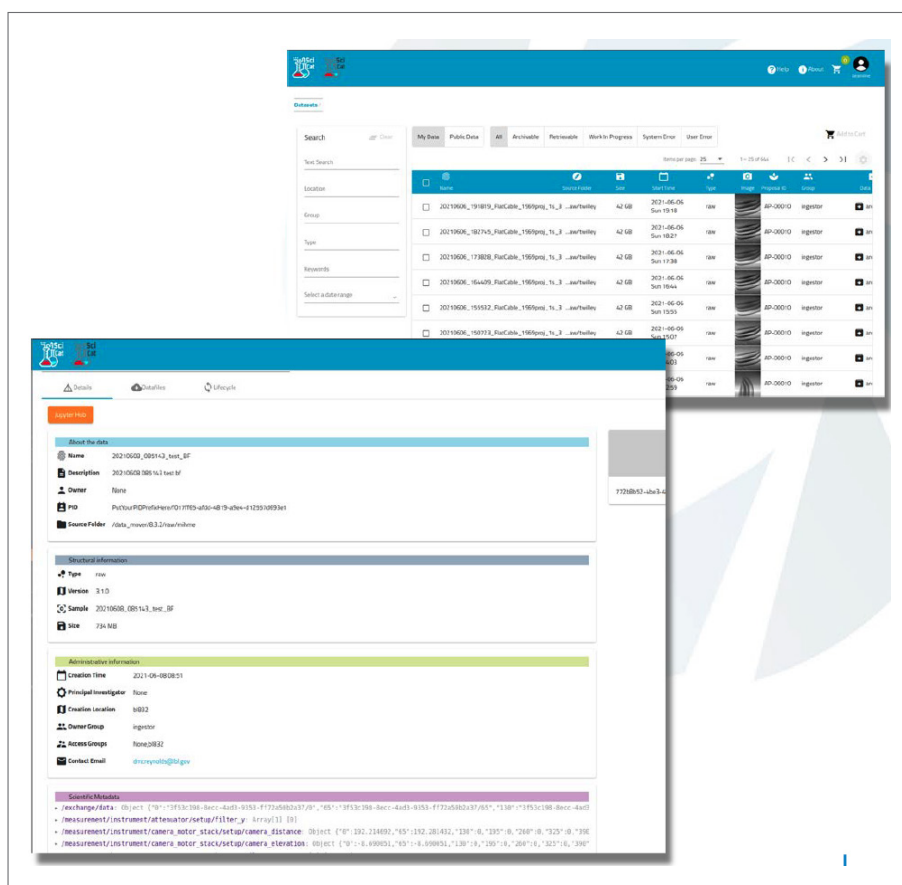
Figure 5.1.16: ALS Data Portal

5.1.2.6 Software Infrastructure

We break the software infrastructure into the following categories:

- Beamline/Endstation controls
- Data Movement
- Data Analysis
- AI/ML

Beamline/Endstation Controls

Beamline controls are using the Beamline Control System, based on LabView. A few end stations are using the Experimental Physics and Industrial Control System (EPICS)/Bluesky platforms.

Present–Two Years

Beamline and Endstation Controls use a variety of software tools:

- LabView is used to control all beamlines and a majority of endstations. LabView provides system controls and a user interface.
- Several end stations have adopted Bluesky[1] on top of EPICS and an EPICS/LabView hybrid.

Next Two to Five Years

- The Bluesky framework will be implemented at an increasing number of endstations.
- Autonomous data collection frameworks (e.g., gpCam[2]) will be implemented at some beamlines to help provide more intelligent scans.
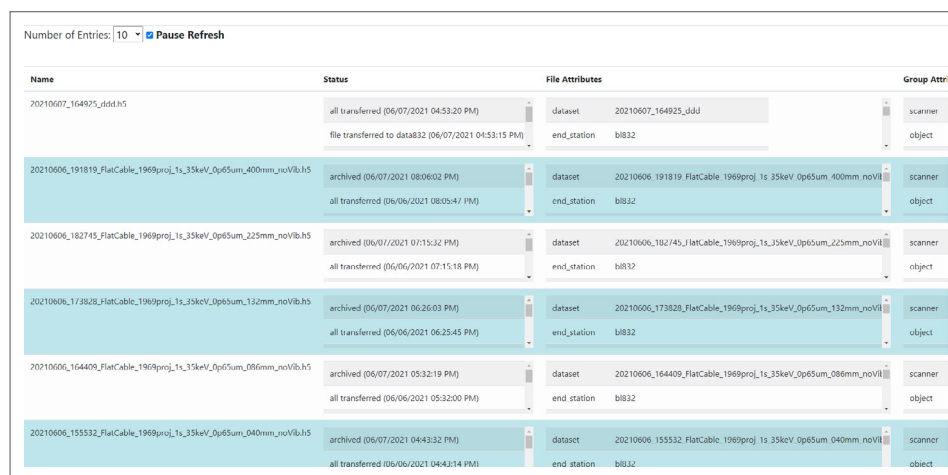
Beyond Five Years

- AI-driven data reduction techniques will be a critical part of the ALS infrastructure, applied at many points along the collection, movement and analysis pipelines

Data Movement

The ALS is using a wide variety of data analysis tools and software. To streamline the process, the ALS is building out the framework described earlier in this document.

We do anticipate that ML will play a much greater role in the future. Figure 5.1.17 shows the graphical user interface (GUI) to track data movement from the tomography beamline to NERSC. It is a file-based approach for now and uses Globus.

Figure 5.1.17: Data management portal that provides users with current state of transfer using Globus. The portal includes document-based support for supplying visual metadata content to users. Additionally, services in the portal enable users to query state of operations.

---

[1] https://blueskyproject.io/
[2] https://gpcam.lbl.gov/

Present–Two Years

- Globus SDK
  - internal transfers from machine to machine
  - external transfers (to NERSC)
- Custom movement code bases for tracking and invoking data transfers and metadata ingestion into the SciCat database
  - Python, PHP, FastAPI, MongoDB

Next Two to Five Years

- Similar software stacks as above, with an increased number of beamlines added to the framework

Beyond Five Years

- Similar software stacks as above, with an increased number of beamlines added to the framework

Data Portal

Once the data reaches NERSC, the data and metadata are ingested into SciCAT as described above. Local and remote users have access to their data and the opportunity to search for and annotate their data.
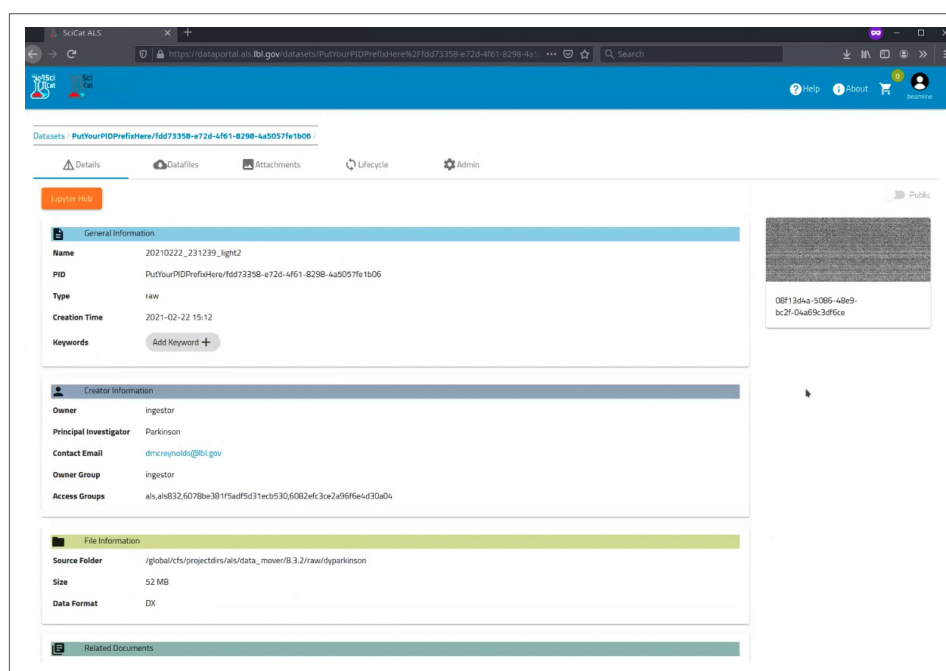
Figure 5.1.18: SciCAT

Present–Two Years

- SciCat Scientific Catalog
- Tracking, searching, downloading data sets
- Hosted on NERSC Spin service

Next Two to Five Years

- Similar software stacks as above, with an increased number of beamlines added to the framework

Beyond Five Years

- Similar software stacks as above, with an increased number of beamlines added to the framework

Data Analysis

For further data processing, the ALS is developing a framework to connect the data residing in SciCAT to a flexible analysis framework as described above. In this case, the data is exposed in Jupyter Lab running at NERSC. In addition to Jupyter, the ALS is developing a variety of GUI based analysis interfaces written in Dash. The availability of data and analysis resources in a web-based environment allows for remote processing of data as described above.
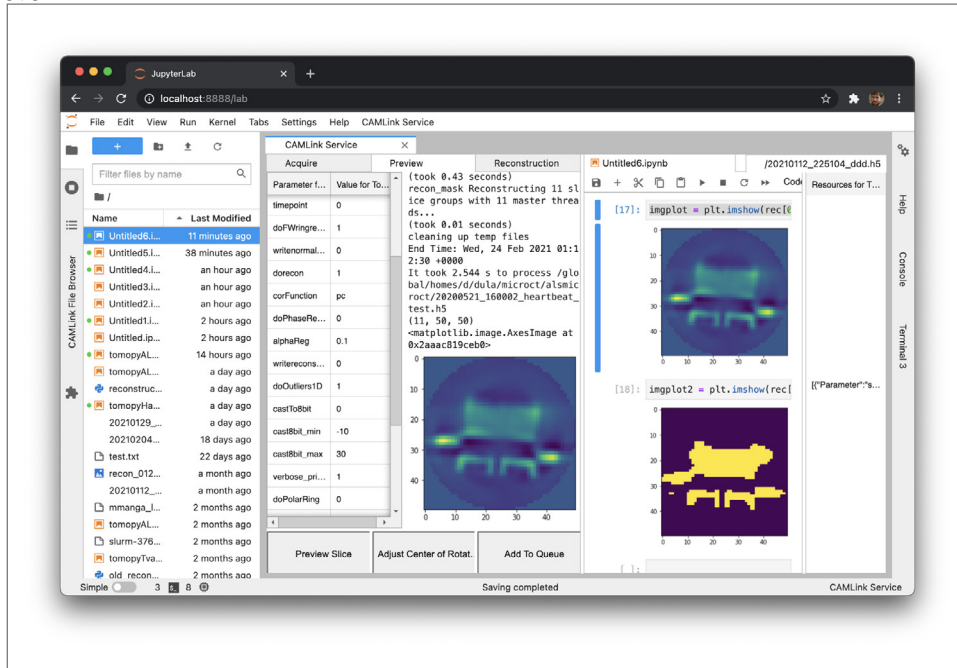
Figure 5.1.19: Main View: Showcases the JupyterLab extension utilized as custom user interface for performing Tomography pipeline through a tailored sequence of steps. (Inset image in orange): shows how the visual interface can seamlessly switch to Jupyter cells

Present–Two Years

- Technique-specific data analysis pipelines being developed, in use at the tomography beamline.
- Framework to run processing pipelines on different processors (e.g., ALS local cluster, NERSC).

Next Two to Five Years

- Increased number of techniques and beamlines supported by platform
- Integration with multiple HPC facilities (e.g., ALCF, OLCF)

Beyond Five Years

- Additional real-time and streaming capabilities

AI/ML

The ALS will employ ML to accelerate the processing of data at various stages of the data lifecycle, from acquisition through analysis. The ALS is developing the MLExchange project to promote the use of AI/ML techniques and lower the burden of using them for beamline scientists and beamline users. Models, data sets and analysis tools are available in a single deployment framework. Users can choose a variety of processing locations to work with models. This project uses a variety of tools including:

- version control systems like git and DVC
- container frameworks like Docker and Shifter
- user interface tools like Dash
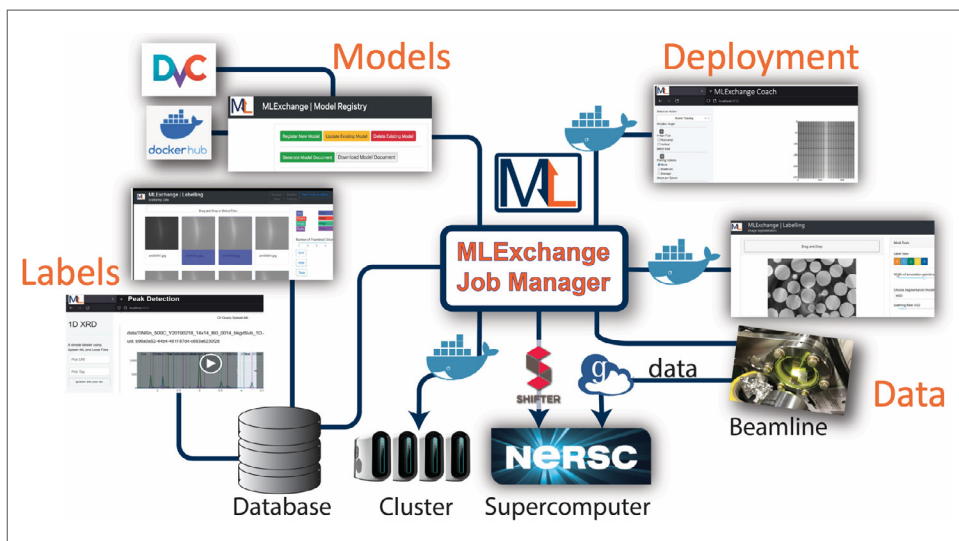
Figure 5.1.20: MLExchange project

Present–Two Years

- Demonstration and proof of concept
- Increasing user base as more modeling techniques are supported

Next Two to Five Years

- Scaling out to multiple facilities

Beyond Five Years

- Scaling out to multiple facilities

5.1.2.7 Network and Data Architecture

There are several Local Area Networks in use at the ALS. These networks are split between those used by the back-end control system and those used by the experimental beamlines. Most of the ALS LANs run at 1 Gbps with a few running at 10 Gbps. The ALS has two general purpose routers, one dedicated to the control networks and the other dedicated to beamline networks. Both routers are connected to the LBNLnet routers through dual 10 Gbps uplinks. The ALS also has a ScienceDMZ router which is connected to the ESnet routers through dual 10 Gbps uplinks. perfSONAR nodes exist at the most critical network locations, including in the ScienceDMZ.

In the short term (two-year budget horizon), the ALS is focusing on replacing the two general purpose routers, which are reaching their end of life in 2023. Two new routers are on order, and their installation is targeted for early 2023. These routers will be capable of 40 or 100 Gbps uplinks, with 10 or 40 Gbps links to the ALS networks. In addition to the router replacements, the ALS will be refreshing approximately 25 control network switches that will also be reaching end of life in 2023. Most of these switches will have 1 Gbps interfaces with 10 Gbps uplinks. There may be some delay in receiving this equipment related to COVID-19 supply-chain issues. The ALS ScienceDMZ router will be reaching end of life in 2024. LBNLnet is upgrading the LBNL ScienceDMZ network to support dual 100 Gbps paths all the way from user devices through the network border to ESnet and can provide very high data transfer rates to LBNL systems. Rather than replace the ALS ScienceDMZ router, the ALS may choose to take advantage of the LBNLnet ScienceDMZ router, possibly keeping the slower ALS ScienceDMZ router available as a backup path. perfSONAR nodes will be deployed to more network locations, including the new ScienceDMZ, covering all important distribution points.

Longer term (two to five years and beyond), the ALS has plans to upgrade all beamline network switches to 10 Gbps, and in some cases up to 40 Gbps when especially high data transfer rates are needed. This will require upgrading switch hardware and cabling at the beamlines, and from the beamlines to the ALS server room and routers. In addition, large portions of the ALS control networks will be replaced as part of the ALS-U upgrade project, currently scheduled for 2025–2026. This will include replacement of approximately 120 switches and network cabling, as well as adding interface modules to the ALS routers. LBNL will be upgrading its ESnet connection from 2x 100 Gbps to more than 1x 400 Gbps during this time period.
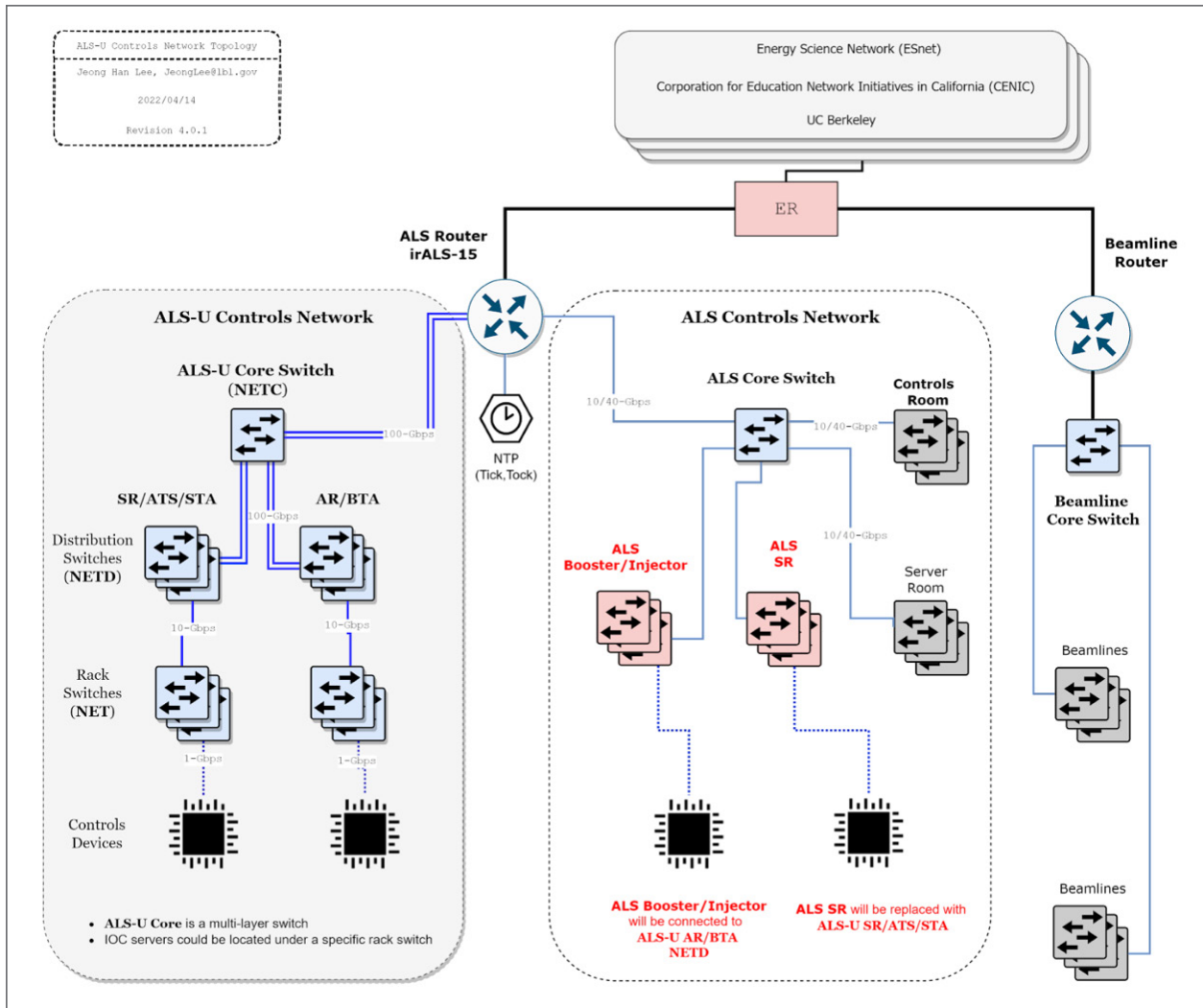
Figure 5.1.21: ALS Network Architecture
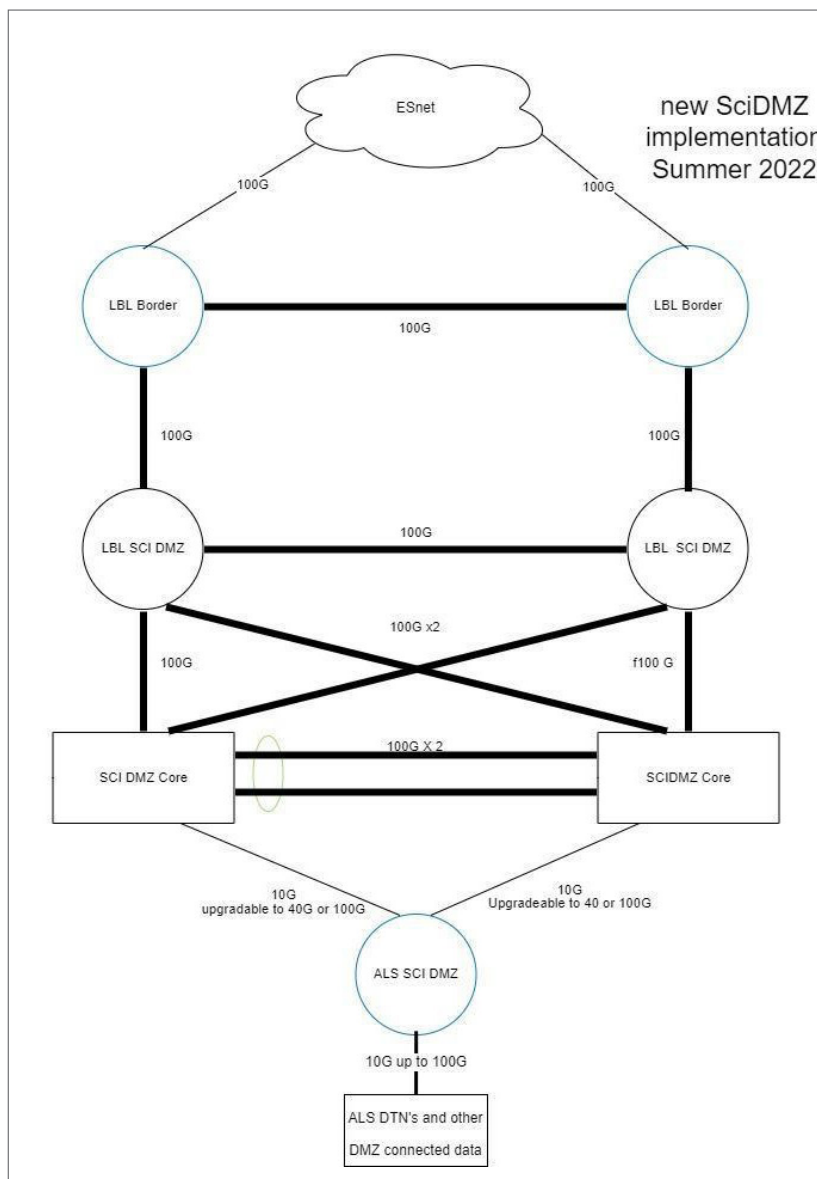
Figure 5.1.22: LBNL Network Architecture

5.1.2.8 Cloud Services

The ALS is looking into using commercial cloud computing but nothing has been developed or decided. The ALS makes use of the NERSC Spin service for hosting containers providing web and other services to its users.

5.1.2.9 Data-Related Resource Constraints

Data management and workflow tools are needed that integrate beamline instruments with computing and storage resources, for use during experiment, as well as facility user access for post-experiment analysis. Real-time data processing capabilities are required to significantly reduce data volumes and provide feedback during experiments to improve data quality and to drive the direction of ongoing measurements; the application of advanced mathematical algorithms, ML, and the integration of simulations and model-based approaches will allow automated steering of data collection. On-demand utilization of computing environments is required to enable near real-time data processing. Sufficient data storage and archival resources to house the continually increasing amounts of valuable scientific data produced by the ALS and ALS-U is required.

5.1.2.10 Outstanding Issues

None to report at this time.

5.1.2.11 Facility Profile Contributors

ALS Representation

- **Andreas Scholl**, *LBNL*, a_scholl@lbl.gov
- **Alexander Hexemer**, *LBNL*, ahexemer@lbl.gov
- **Dylan McReynolds**, *LBNL*, dmcreynolds@lbl.gov
- **Tanny Chavez**, *LBNL*, tanchavez@lbl.gov
- **Jason Jed**, *LBNL*, jjed@lbl.gov
- **Chris Harwood**, *LBNL*, harwood@lbl.gov
- **Nicholas Schwarz**, *ANL*, nschwarz@anl.gov
- **Stuart Campbell**, *BNL*, scampbell@bnl.gov
- **Apurva Mehta**, *SLAC National Accelerator Laboratory*, mehta@slac.stanford.edu
- **Vivek Thampy**, *SLAC National Accelerator Laboratory*, vthampy@slac.stanford.edu
- **Jana Thayer**, *SLAC National Accelerator Laboratory*, jana@slac.stanford.edu

ESCC Representation

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

5.2 Advanced Photon Source (APS)

The APS (see Figure 5.2.1), located at ANL, is a synchrotron light source, funded by the US DOE, Office of Science-BES to produce high-energy, high-brightness X-ray beams. The source is optimized to put large quantities of high-energy photons into a very small area in a very short time. Scientists from around the world utilize the APS to conduct forefront basic and applied research in the fields of materials science; biological and life science; physics; chemistry; environmental, geophysical, and planetary science; and innovative X-ray instrumentation.

Figure 5.2.1: Aerial view of the APS

5.2.1 Discussion Summary

- The APS is a scientific user facility that serves approximately 6,000 unique users per year.

- The APS-U Project will replace the entire APS storage ring. A one-year shutdown is required, and is scheduled to begin in April 2023.

- The APS currently operates 68 beamlines. As part of the APS-U Project, feature beamlines were approved for construction as either new, or replacement of existing beamlines. In addition, several beamlines were selected for major enhancements.

- The feature and enhanced beamlines will generate substantially more data than their contemporary equivalents, requiring step-change improvements in networking, controls and data acquisition, computing, workflow, data reduction, and analysis tools to operate effectively.

- APS-U data generation rates may be too large for traditional file-based workflows, which will facilitate a move to streaming-based workflows directly to computer system memory requiring robust facility and Laboratory networking.

- During APS-U, it will be impractical and unreasonable to support the scale of computing required with only local APS resources. The APS and ALCF have partnered to deliver a new model of computing, tightly coupling APS experiment instruments with ALCF supercomputers to accelerate scientific discovery.

- APS users, and those at other US DOE-funded facilities, often leverage the complete ecosystem of DOE experimental facilities. The process of science often involves combining and analyzing data from multiple sources. As data rates and complexity continue to increase, sufficient networking connectivity, bandwidth, and reliability is required to connect measurement facilities, computing facility, and user home institutions to enable effective data management and analysis.

- The deployment of remote access tools enabled experimenters to run measurements remotely during the COVID pandemic, and led to an increase in remote users. The APS anticipates that remote access will continue, and thus a critical infrastructure component.

- The meaning and size of the data, the processing steps, and the analysis and interpretation will vary depending on the beamline, the measurement technique, the detector(s) utilized, and the scientific goal.

- Generally, data is first stored on a computer attached to the acquisition detector or some other local beamline storage system.

- The final analysis and interpretation of the data for publication is generally carried out by the experiment team at their home institutions, is very experiment specific, and may take months or even years to perform.

- The APS Data Management System, a facility-wide software and hardware infrastructure for managing data and workflows, provides a data management, workflow, and storage systems. This framework integrates with 50 of the 68 APS beamlines, and facilitates the automated transfer of data between acquisition devices, computing resources, and data storage systems. The system utilizes Globus, GridFTP, remote sync, and secure copy protocol for internal data transfer within and external to the facility.

- Use of multimodal data requires more sophisticated data processing, and requires increases in computing capabilities which can include training AI/ML models as well as real-time analysis and feedback to enable autonomous experiment steering. The APS is exploring the utilization of edge-computing resources, coupled closely to detectors and instruments, to facilitate AI/ML data reduction algorithms.

- The APS adopts a graded approach to resource utilization:

- multicore processors and GPU units are available local to beamlines

- APS maintains an on-site computing cluster

- ANL maintains computing resources as a part of the Laboratory Computing Resource Center

- DOE HPC facilities (e.g., the ALCF, NERSC, and the OLCF) are available via allocation procedure

- Data is transferred to a local beamline workstation, the local APS distributed-memory computing cluster, or a computing center located on the Argonne campus, such as the ALCF or the Argonne Laboratory Computing Resource Center, for processing.

- The APS provides approximately 3.6 petabytes (PB) of central disk storage for short- and medium-term data retention, and several DTNs for reliable, high-speed data movement as a part of the APS Data Management System.

- The ALCF provides approximately 10 PB of tape storage for longer-term data retention. The ALCF also has a 100-PB CFS and a 100-PB project file system, along with tape storage that is available for APS use.

- APS X-ray Science Division operates beamlines that collect approximately 6 PB to 7 PB of raw data per year. APS-U estimates that the volume will increase by at least two orders of magnitude to 100s of PBs of raw data per year.

- The computing resources required by the APS are anticipated to grow by at least two orders of magnitude to keep up with data generation rates. Prior to the APS-U, most data processing can be performed within the range of teraflops per second of computing resources. In the APS-U Era, first-pass data processing at the APS will require on-demand access to tens of petaflops per second (PFLOPps) of computing resources.

- Raw and processed data is often stored on a large disk system at the APS using the APS Data Management System from which users may retrieve the data or share the data with collaborators.

- Currently, the APS only guarantees storage space for a minimum of three months; the long-term management of data is the responsibility of the user group that collected the data.

- In some cases, experimenters transfer the raw data to computing resources at their home institutions for all further processing.

- Demands for increased data processing capabilities in the APS-U Era are driven by new scientific opportunities enabled by the upgraded facility.

- The increase in brightness and advances in detector data rates will generate multiple orders of magnitude more data than are generated today; this increase in data volume necessitates an increase in processing power to keep pace.

- The APS is working to develop a computational data fabric for end-to-end data lifecycle management. These tools will be applied at more beamlines in the years after the APS-U.

- The APS does not presently heavily utilize cloud services for its scientific data infrastructure. The APS does not have any immediate plans to further utilize cloud services in this area.

- The current ANL network supports 10 Gbps, 40 Gbps, 100 Gbps, and 400 Gbps for WAN and LAN uses, and features a Science DMZ to support ANL facility use.

- As data rates and volumes continue to grow, greater demands will be placed on the APS network. This is especially true for the APS-U feature and enhanced beamlines. The APS is updating its network architecture and infrastructure to better serve the beamlines as it enters the APS-U Era.

- The APS network consists of a pair of core switches that are connected via 2x40 Gbps to the APS Tier 2 firewall, which in turn connects to the Argonne Tier 1 firewall with 2x100 Gbps capacity.

- The APS network will refresh hardware in the two- to five-year timeframe to support 400 Gbps in support of the APS-U upgrade, and support connectivity to the ANL Science DMZ for certain use cases.

5.2.2 APS Facility Profile

Each of the 68 beamlines at the APS offers a unique combination of capabilities. Some of the main considerations are energy range and tunability, special sample environments, time structures, and beam size (see Figure 5.2.2). The energies used range from relatively "soft" X-rays (as low as 250 eV) to "tender" X-rays (3-5 keV) to "hard" X-rays at 100 keV and sometimes higher. At many beamlines, the energy can be tuned with relative ease. Samples can be examined under extreme conditions of temperature and pressure, and several facilities are available for samples requiring special handling (for example, biohazards and radioactive samples). Many experiments involve timing through correlation with a pulsed laser or with the time structure of the X-ray pulses, for example. In the

typical operating mode, the X-rays come in evenly spaced bunches or pulses. Some beamlines employ additional optics to narrow the already tight beam into even smaller spots, offering spatial resolution into the 20-nm range.

5.2.2.1 Science Background

The ongoing APS-U Project will replace the entire APS storage ring with a ring based on an MBA lattice design including reverse bends. The new storage ring will increase the APS brilliance by factors of up to 500 depending on X-ray energy and make the APS the brightest hard x-ray synchrotron light source in the world. Moreover, because of both ongoing developments at the APS in superconducting undulators and the fact that the APS is the highest energy storage ring light source in the Western Hemisphere, the APS will continue to be world-leading in high-energy X-ray capabilities. A one-year shutdown is required for removal and replacement of the storage ring followed by time for instrument commissioning; the shutdown period is scheduled to begin in April 2023.

As part of the APS-U project, feature beamlines were selected for new installation and/or complete replacement based on their promise to best exploit the capabilities of the new source, namely brightness, coherence, and high-energy X-rays. In addition, based on similar criteria, several beamlines were selected for major enhancements. Because of the greatly enhanced brightness, coherence, and signal at high X-ray energies along with new, state-of-the-art, highly pixelated commercial detectors that are part of these projects and amplify these gains, the feature and enhanced beamlines require significant improvements in networking, controls and data acquisition, computing, workflow, data reduction, and analysis tools to operate effectively. Detailed information about computing and data needs, plans, and gaps may be found in the "APS Scientific Computing Strategy" document[3].

The meaning of data, the size of data, the processing steps applied to data, and the analysis and interpretation of data vary depending on the beamline, the measurement technique, the detector(s) utilized, and the scientific goal. Raw data is generated primarily by two-dimensional (area), one-dimensional (strip), or point detectors as a part of scattering, imaging, or spectroscopy measurements. Raw data may represent a scattering pattern, a transmission image, or spectra, for example. The size of the data may vary from a few megabytes to hundreds of terabytes per allocated beam time. Some form of data processing or reduction is usually performed after data is collected to transform the data from a technique or detector representation to the physical space of interest in an analyzable representation, such as a series of sinograms into a three-dimensional world-space volume, or spectra into elemental concentrations.

Generally, data is first stored on a computer attached to the acquisition detector or some other local beamline storage system. In many cases, the data is transferred to a local beamline workstation, the local APS distributed-memory computing cluster, or a computing center located on the Argonne campus, such as the ALCF or the Argonne Laboratory Computing Resource Center, for processing. Raw and processed data is often stored on a large disk system at the APS using the APS Data Management System (described in detail later in this document) from which users may retrieve the data or share the data with collaborators. Currently, the APS only guarantees storage space for a minimum of three months; the long-term management of data is the responsibility of the user group that collected the data[4]. In some cases, experimenters transfer the raw data to computing resources at their home institutions for all further processing. The final analysis and interpretation of the data for publication is generally carried out by the experiment team at their home institutions, is very experiment specific, and may take months or even years to perform.

---

[3] "The APS Scientific Computing Strategy," www.aps.anl.gov/files/APS-Uploads/XSD/XSD-Strategic-Plans/APSScientificComputingStrategy-2021-09-24-FINAL.pdf.
[4] "The Advanced Photon Source Data Management and Retrieval Practices," www.aps.anl.gov/Users-Information/Help-Reference/Data-Management-Retrieval-Practices
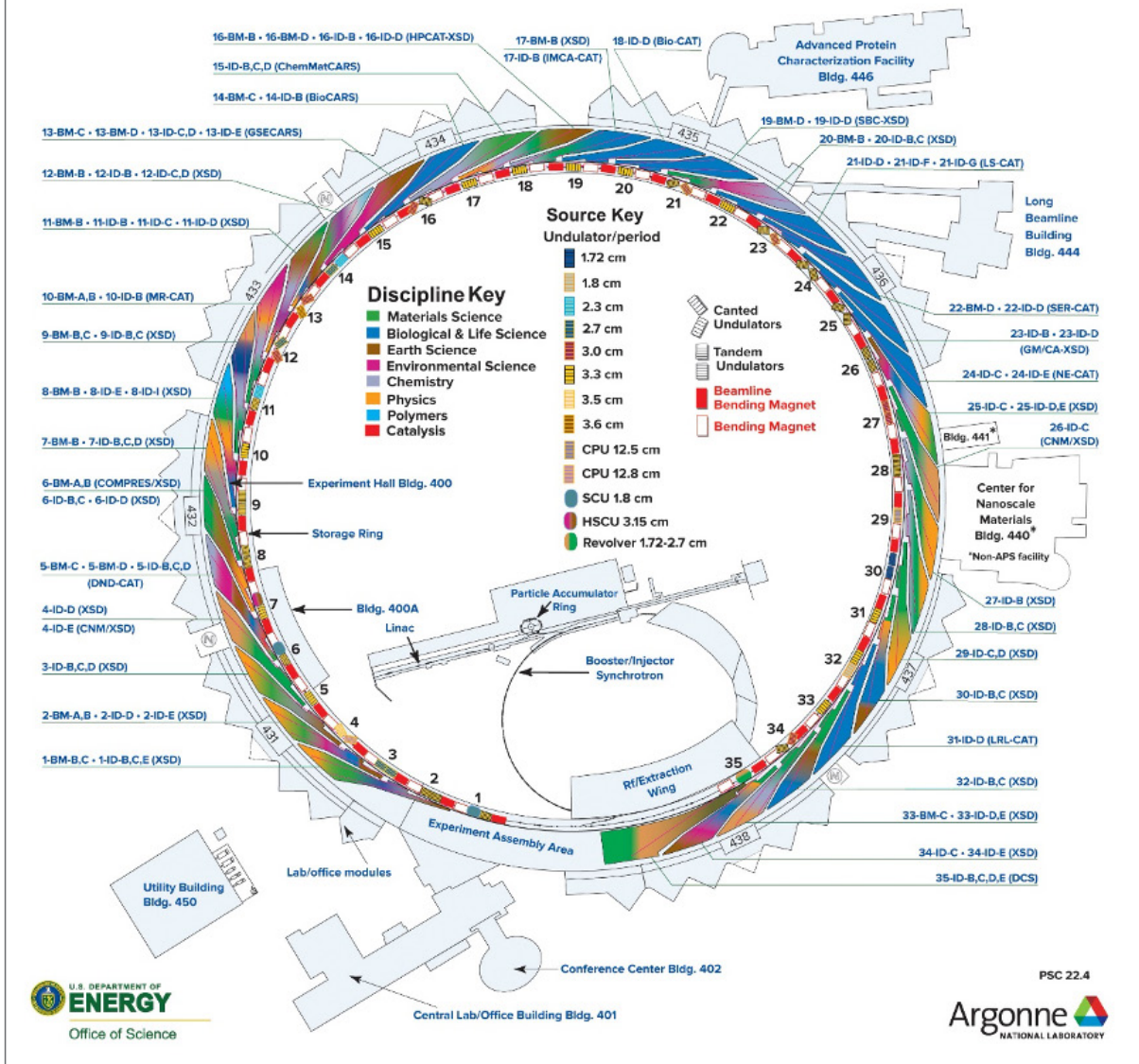
Figure 5.2.2: Diagram of APS beamlines, systems, and facilities

5.2.2.2 Collaborators

The APS is a scientific user facility with a large and diverse user community. The number of both on-site and remote users per year has steadily increased, peaking to almost 6,000 unique users per year (see Figure 5.2.3), resulting in approximately, at its peak, 2,500 publications per year (see Figure 5.2.4). A recent decrease in the number of users per year is due partly to a change in the way remote users for mail-in experiments are counted, but more so to restrictions imposed during COVID-19. In response, the deployment of remote access tools enabled experimenters to run measurements remotely and led to an increase in remote users as a fraction of total users in fiscal year (FY) 2021.

APS users perform experiments in many fields, including the biological and life sciences, the materials sciences, chemistry, and physics (see Figure 5.2.5). APS users are funded by a diverse group of sources, including the US DOE Office of Science and the National Nuclear Security Administration, the US National Science Foundation (NSF), the US National Institutes of Health (NIH), and industry, amongst others (see Figure 5.2.6). Facility users come from throughout the US (see Table 5.2.2) as well as from many countries (see Table 5.2.3).

Approximately one-half of the beamlines at the APS are managed directly by the APS X-ray Science Division. Others are managed by collaborative access teams (CATs), which comprise scientists from universities, industry, and federal and private research laboratories. Beam time is allocated to experiments in one of three ways: (1) the general user mode is the primary means of access for external users in which applications for beam time are peer reviewed; (2) the partner user mode provides access for projects that require guaranteed beam time over multiple cycles and will ultimately benefit the general user community; and (3) CATs provide beam time directly to CAT members. The APS operates for three three-month runs (or cycles) each year, with a one-month shutdown between runs. The runs typically span February to April, June to August, and October to December. Depending on the beamline and type of experiment, beam time may range from a few hours to a few weeks in duration.
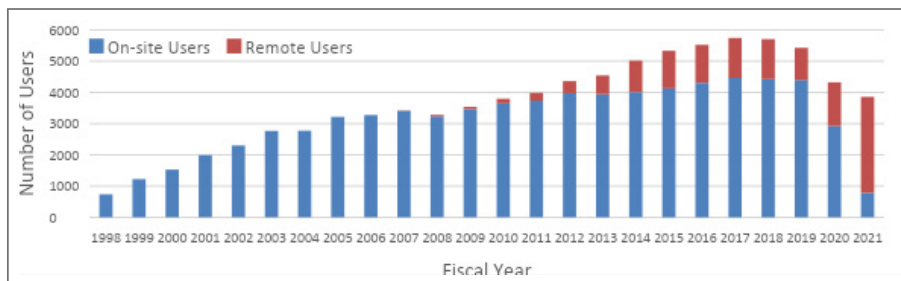
Figure 5.2.3: Total number of APS users per FY. Note: (a) Prior to FY14, mail-in users were not included in the Remote category; (b) In FY20, new BES user counting policy has been applied so that only 1 unique user is associated with mail-in experiments and the user is only counted once in the whole population.

User/Collab-orator 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
US UNIVERSITY-BASED PIS	Both	data portal, data transfer (Globus, sftp, cloud storage), portable hard drive	1 MB–10 TB	ad hoc	N	N/A
US NATIONAL LABS BASED PIS	Both	data portal, data transfer (Globus, sftp, cloud storage), portable hard drive	1 MB–10 TB	ad hoc	N	N/A
INTERNATIONAL PIS	Both	data portal, data transfer (Globus, sftp, cloud storage), portable hard drive	1 MB–10 TB	ad hoc	N	N/A

Table 5.3.1: APS Collaboration Space
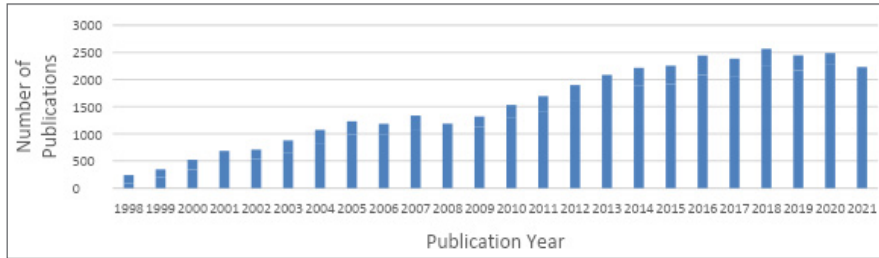
Figure 5.2.4: Number of publications attributed to the APS per publication year
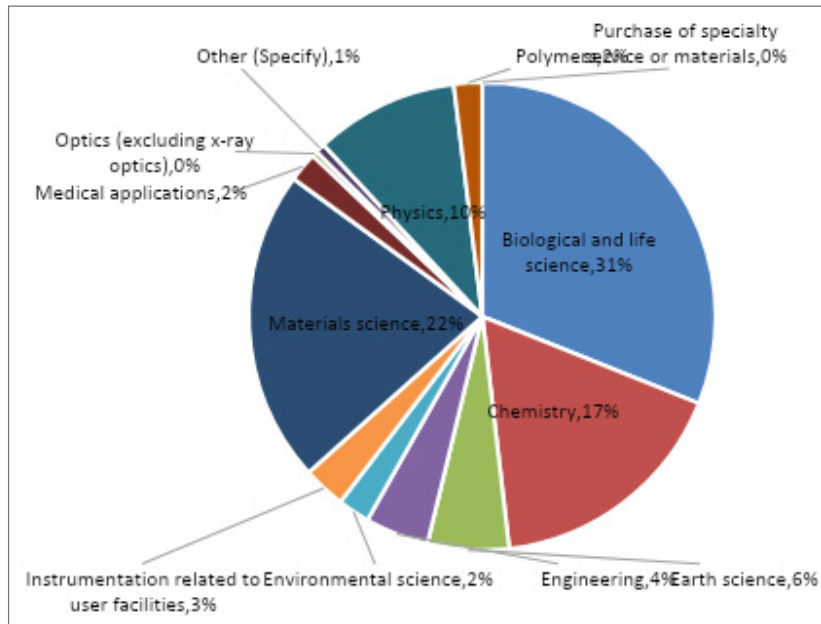
Figure 5.2.5: Percentage of APS users by experiment subject for FY 2021
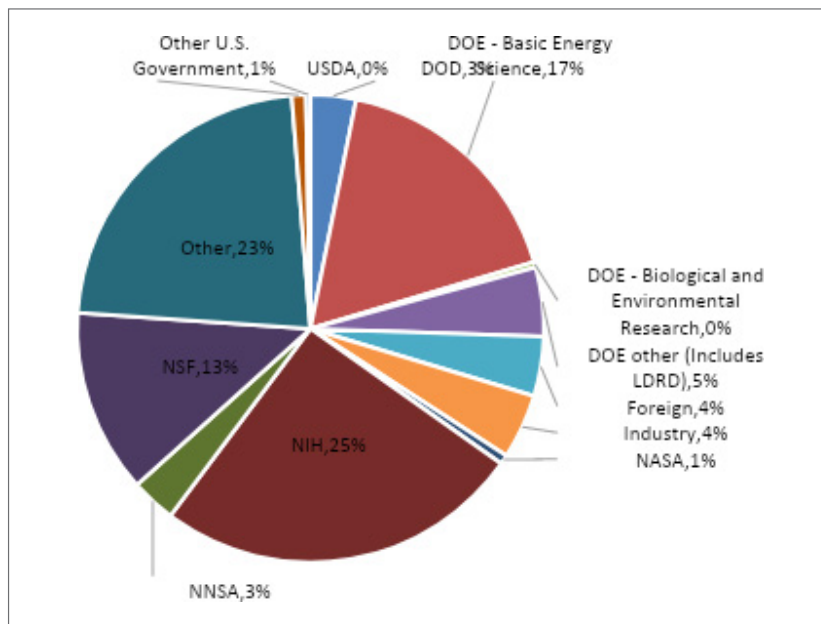
Figure 5.2.6: Percentage of APS users by source of support for FY 2021

State Description	Number of Users	State Description	Number of Users
ALABAMA	24	NEBRASKA	17
ARIZONA	23	NEVADA	10
ARKANSAS	2	NEW HAMPSHIRE	3
CALIFORNIA	369.37	NEW JERSEY	44
COLORADO	54	NEW MEXICO	43.5
CONNECTICUT	45	NEW YORK	243
DELAWARE	11	NORTH CAROLINA	89
DISTRICT OF COLUMBIA	19	NORTH DAKOTA	9
FLORIDA	62.1	OHIO	73.4
GEORGIA	85	OKLAHOMA	5
HAWAII	8	OREGON	12
IDAHO	1	PENNSYLVANIA	98
ILLINOIS	780.84	PUERTO RICO	4
INDIANA	98.05	RHODE ISLAND	11
IOWA	45	SOUTH CAROLINA	21
KANSAS	4	SOUTH DAKOTA	1
KENTUCKY	13	TENNESSEE	99
LOUISIANA	11	TEXAS	155
MAINE	5	UTAH	18.5
MARYLAND	151	VERMONT	2
MASSACHUSETTS	157.22	VIRGINIA	37.67
MICHIGAN	134	WASHINGTON	67.33
MINNESOTA	55	WEST VIRGINIA	6
MISSISSIPPI	4	WISCONSIN	39.95
MISSOURI	46	WYOMING	5
MONTANA	1	Sum:	3322.93

Table 5.2.2: APS users by state for FY 2021. Fractional numbers account for users who relocate to different states during the reporting period.

Institution Country	Number of Users
AUSTRALIA	8
BRAZIL	5
BULGARIA	1
CANADA	117
CHINA, PEOPLES REPUBLIC OF	51.93
DENMARK	17
FRANCE	29
GERMANY	28
HONG KONG	1
INDIA	2
IRELAND	2
ITALY	6
JAPAN	5
KOREA, SOUTH	34
MEXICO	1
NETHERLANDS, THE	2
PORTUGAL	1
SINGAPORE	3
SPAIN	2
SWEDEN	6.14
SWITZERLAND	2
TAIWAN	8
THAILAND	1
TURKEY	2
UNITED KINGDOM	28
USA	3322.93
Total countries: 26	3686

Table 5.2.3: APS users by home institution country for FY 2021. Fractional numbers account for users who change host institution to one that is in a different country during the reporting period.

5.2.2.3 Instruments and Facilities

The APS currently operates 68 beamlines, each with its own unique data generation characteristics. As part of the APS-U Project, feature beamlines were approved for construction as either new, or replacement of existing beamlines. In addition, several beamlines were selected for major enhancements. Because of the greatly enhanced brightness, coherence, and signal at high X-ray energies along with new state-of-the-art, high-bandwidth commercial detectors that are part of these projects and amplify these gains, the feature and enhanced beamlines will generate substantially more data than their contemporary equivalents, requiring step-change improvements in networking, controls and data acquisition, computing, workflow, data reduction, and analysis tools to operate effectively. Detailed information about computing and data needs, plans, and gaps may be found in the "APS Scientific Computing Strategy" document.

APS beamlines often have dozens of motors, readout electronics, and X-ray detectors that are controlled by real-time hardware devices. Many X-ray detectors have specific manufacturer-supplied interfaces. FPGA or ARM devices are often used for real-time coordination of devices used during experiments.

As data rates and volumes continue to grow, greater demands will be placed on the APS network. This is especially true for the APS-U feature and enhanced beamlines. The APS is updating its network architecture and infrastructure to better serve the beamlines as it enters the APS-U Era. Figure 5.2.7 depicts the APS-U Era network architecture and infrastructure plan. Refer to Section 5.2.2.7, Network and Data Architecture, for a detailed description of the APS network and planned upgrades.

Demands for increased data processing capabilities in the APS-U Era are driven by new scientific opportunities enabled by the upgraded facility. The increase in brightness and advances in detector data rates will generate multiple orders of magnitude more data than are generated today; this increase in data volume necessitates an increase in processing power to keep pace. The utilization of multimodal data to answer new questions requires more complex and sophisticated data processing algorithms requiring increases in computing capabilities. Increases in computing power are needed by advanced algorithms for existing techniques that, for example, provide higher-fidelity results, and to train AI/ML models. The need for real-time analysis and feedback to make crucial experiment decisions and enable autonomous experiment steering also requires more computing cycles than have been traditionally utilized.

The computing resources required by the APS are anticipated to grow by at least two orders of magnitude to keep up with data generation rates. Prior to the APS-U, most data processing can be performed within the range of teraflops per second of computing resources. In the APS-U Era, first-pass data processing at the APS will require on-demand access to tens of petaflops per second (PFLOPps) of computing resources. There is wide variability in the computational requirements among techniques and processing approaches, with those instruments and techniques that benefit most from high-energy, high-brightness, and coherent X-rays driving most requirements[5].

---

5 Schwarz N., Campbell S., Hexemer A., Mehta A., Thayer J. (2020) Enabling Scientific Discovery at Next-Generation Light Sources with Advanced AI and HPC. In: Nichols, J., Verastegui, B., Maccabe, A., Hernandez, O., Parete-Koon, S., Ahearn, T. (eds.) "Driving Scientific and Engineering Discoveries Through the Convergence of HPC, Big Data and AI." SMC 2020. Communications in Computer and Information Science, vol 1315. Springer, Cham.
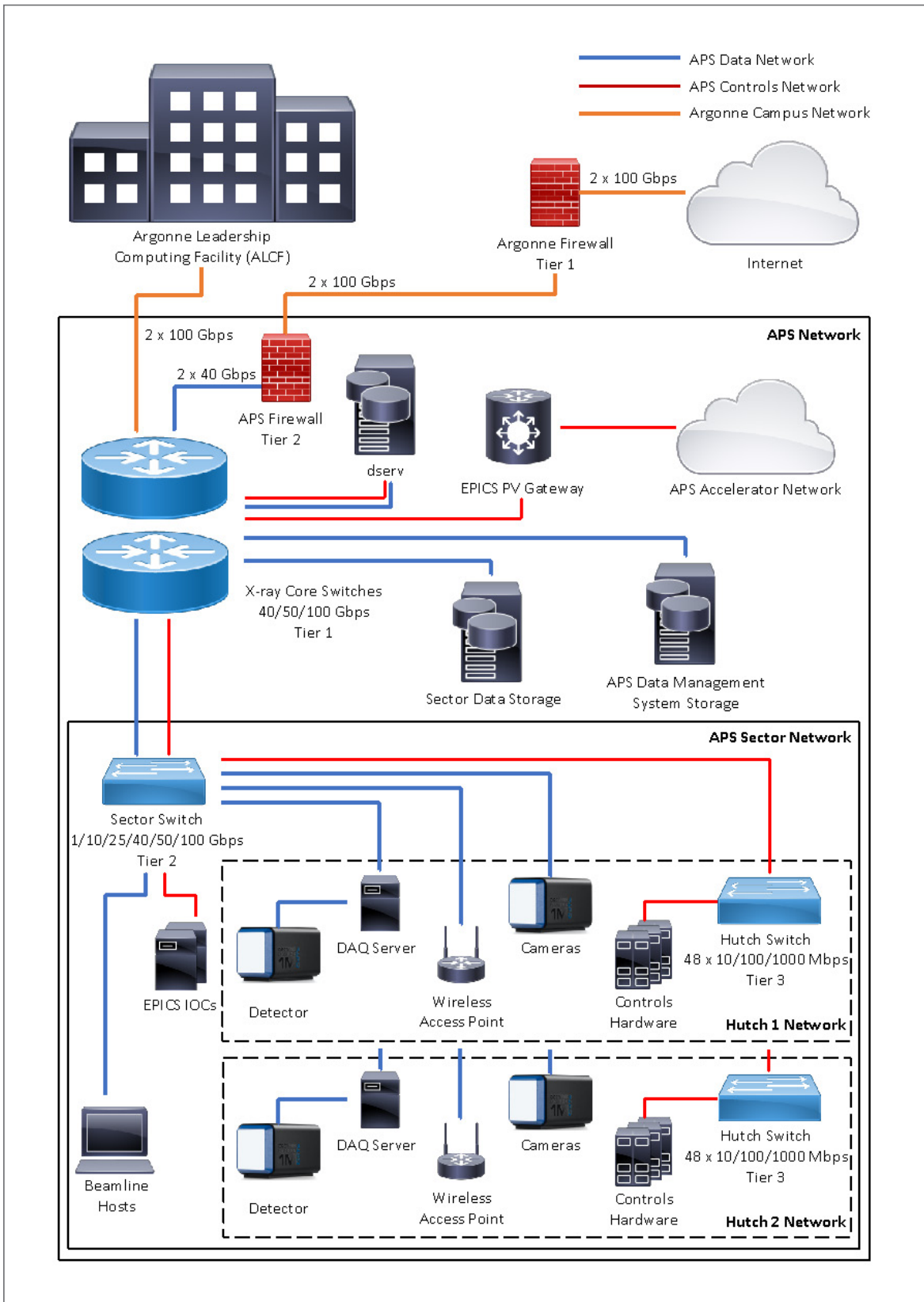
Figure 5.2.7: The APS beamline network and architecture today. See Section 5.2.2.7 for upgrades that are planned over the coming years.

Figure 5.2.8: Computing resources and respective specifications and performance available for use by the APS at Argonne

To satisfy these needs, the APS adopts a graded approach to resource utilization. Small-scale resources, such as multicore processors and GPUs local to beamlines, will be used when sufficient. For moderate computational needs, the APS maintains an on-site computing cluster, and Argonne maintains computing resources as a part of the Laboratory Computing Resource Center. For the most demanding computational problems, large-scale computing facilities must be used, including the ALCF, NERSC, and the OLCF. To mitigate challenges surrounding processing and storing such large, anticipated data volumes, the APS is exploring the utilization of edge-computing resources coupled closely to detectors and instruments, to run AI/ML data reduction algorithms. See Figure 5.2.8 for a list of computing resources available at Argonne.

Edge computing offers the ability to process data quickly on or near detectors and experiment instrumentation without the need to first transfer all data to high-end computing resources. This is particularly promising for handling large data when coupled with machine-learning methods. Using only a subset of data, machine-learning models may be trained on supercomputers. The trained model is then run using edge-computing devices to process newly acquired data, providing fast feedback for experiment steering.

Currently, the APS provides approximately 3.6 PB of central disk storage (easily expandable to 15 PB) for short- and medium-term data retention, and several DTNs for reliable, high-speed data movement internally and externally (see Figure 5.2.9) as a part of the APS Data Management System. The APS plans to double the size of available storage to approximately 8 PB within the next year. The ALCF currently provides approximately 10 PB of tape storage (easily expandable to meet future APS needs) for longer-term data retention. The ALCF has recently deployed a 100-PB CFS (Eagle) and a 100-PB project file system (Grand) along with additional tape storage that is available for APS use.
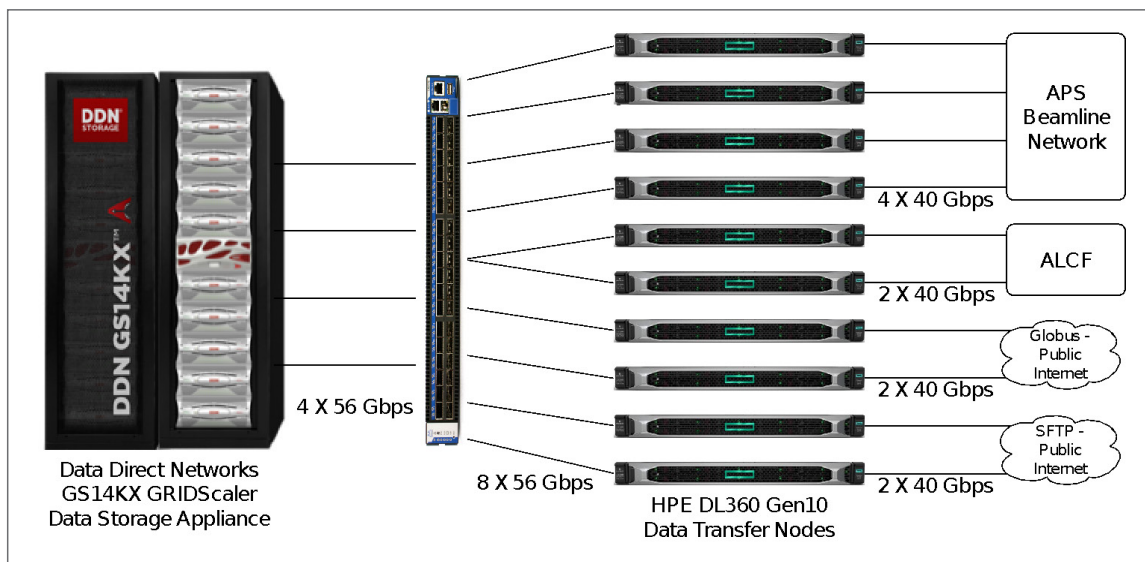
Figure 5.2.9: The APS Data Management System storage infrastructure. A Data Direct Networks storage appliance connects to an InfiniBand switch. Four DTNs link the storage to the beamline network, two DTNs connect the storage to the ALCF, two DTNs provide links to the public Internet for Globus transfers, and two DTNs provide links to the public Internet for Secure File Transfer Protocol (SFTP) transfers.

Today, the APS X-ray Science Division operates beamlines that collect approximately 6 PB to 7 PB of raw data per year. In the coming years, once the upgraded APS storage ring and beamlines are commissioned and brought into operation, it is estimated that the volume of data generated at the APS will increase by at least two orders of magnitude to 100s of PBs of raw data per year (see Figure 5.2.10).
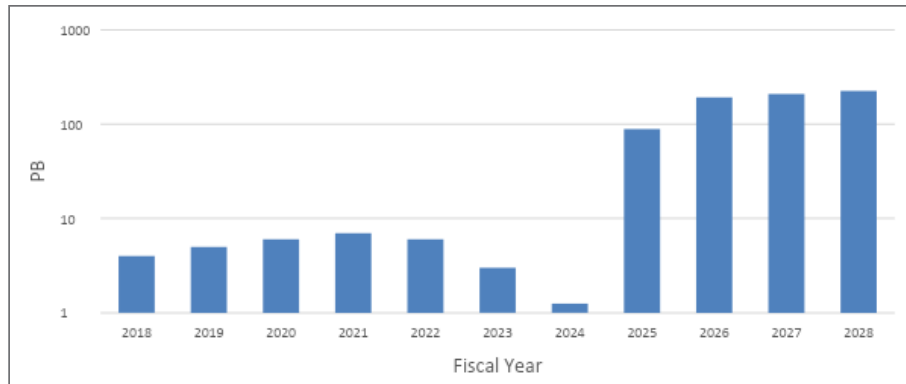
Figure 5.2.10: Log Scale: Estimated data generation volumes per year at the APS. Data generation will drop temporarily due to the installation of storage ring and beamline upgrades.

Detailed information about the data volumes that are anticipated to be generated by the feature beamlines may be found in the "APS Scientific Computing Strategy" document [1]. The APS is in the process of collecting similar information for the beamlines that will receive major enhancements. Abridged information on anticipated data volumes for the feature beamlines and a comparison to current instruments, where available, may be found in the tables below.

Today	Technique	Data Set Size (GB)	Data Per Day (TB)	Data Per Year (TB)
TODAY	Bragg coherent diffraction imaging (CDI)	0.06	0.0033	0.65
APS-U ERA	Fast Bragg CDI	0.49	1.32	220
	High-Resolution CDI	> 5.86	0.81	35

Table 5.2.4: Data generation rates today at the beamline 34-ID-C Bragg coherent diffraction imaging (CDI) instrument (for comparison) and estimated data generation rates at the ATOMIC APS-U feature beamline. For brevity, these tables represent only a fraction of the total 68 beamlines at the APS.

Today	Technique	Data Set Size (GB)	Data Per Day (TB)	Data Per Year (TB)
	GIWAXS	1.3	3.5	276.48
	GIWAXS-XPCS	197.8	69.5	5550.08
	GISAXS	2.4	6.3	501.76
APS-U ERA	GISAXS-XPCS	201.2	70.7	5642.24
	Fast GIXS	988.8	38.9	655.36
		1788.1	70.4	1187.84
	CSSI	357.6	40.2	3891.20

Table 5.2.5: Estimated data generation rates at the Coherent Surface Scattering Imaging (CSSI) APS-U feature beamline. For brevity, these tables represent only a fraction of the total 68 beamlines at the APS. (GIWAXS: Grazing-incidence wide-angle scattering; GISAXS: Grazing-incidence small-angle scattering.)

Today	Technique	Data Set Size (GB)	Data Per Day (TB)	Data Per Year (TB)
TODAY	Near-Field Diffraction	11	1	62
	Far-Field Diffraction	22	12	647
	Far-Field Diffraction	11	3	153
	Far-Field Diffraction	45	12	366
	Far-Field Diffraction	10	92	2,896
	Diffraction Tomography	281	1	10
	Diffraction Tomography	1,125	4	40
	Diffraction Tomography	2,250	3	35
	Diffraction Tomography	9,000	13	140
	Imaging Tomography	15	1	41
APS-U ERA	Near-Field Diffraction	13	17	894
	Far-Field Diffraction	25	116	9,706
	Diffraction Tomography	249	21	441
	Diffraction Tomography	1,825	154	3,233
	Diffraction Tomography	26.37	2.22	47
	Diffraction Tomography	1,993	105	1,103
	Diffraction Tomography	14,596	410	4,300
	Diffraction Tomography	211	45	467
	Imaging Tomography	34	4	134
	Fast Imaging Tomography	34	10	99

Table 5.2.6: Data generation rates today at the 1-ID High-Energy Diffraction Microscopy (HEDM) instrument (for comparison) and estimated data generation rates at the HEXM APS-U feature beamline. For brevity, these tables represent only a fraction of the total 68 beamlines at the APS.

Today	Technique	Data Set Size (GB)	Data Per Day (TB)	Data Per Year (TB)
TODAY	2-ID-D Ptychography	986	8.32	87
	2-ID-D Diffraction	0.20	0.21	13
	2-ID-E XRF	1.91	0.01	2.16
	BNP XRF	0.69	0.01	2.16
APS-U ERA	ISN XRF	1.30	1.98	104
	ISN Ptychography	204,322	181.02	7,603
	ISN Diffraction	40.86	51.72	3,528

Table 5.2.7: Data generation rates today at the 2-ID-D ptychography and diffraction, 2-ID-E x-ray fluorescence (XRF), and Bio Nano-Probe (BNP) XRF instruments (for comparison) and estimated data generation rates at the In Situ Nanoprobe (ISN) APS-U feature beamline. For brevity, these tables represent only a fraction of the total 68 beamlines at the APS.

Today	Technique	Data Set Size (GB)	Data Per Year (TB)
TODAY	4-ID-D XAS / XMCD	0.53	0.000986
	4-ID-D XAS / XMCD (mapping 10 um, high-pressure 1 Mbar)	0.53	0.002129
	4-ID-D Hard Resonant Magnetic Scattering (mapping 100 nm)	0.0004	0.000001
APS-U ERA	4-ID-G Hard Resonant Magnetic Scattering (mapping 100 nm)	9.20	834.13
	4-ID-G Hard Resonant Magnetic Scattering – Polarization Modulated (mapping 100 nm)	229.86	41,706.38
	4-ID-G Hard Resonant X-ray Ptychography	25.54	834.13
	4-ID-G Hard Resonant X-ray Ptychography – Polarization Modulated	638.51	41,706.38
	4-ID-G Bragg CDI Magnetic Contrast	638.51	417.06
	4-ID-G Tomographic CDI	10.22	834.13
	4-ID-H XAS / XMCD / XMLD (mapping 300 nm, high-pressure 7 Mbar)	0.26	0.002129

Table 5.2.8: Data generation rates today at the 4-ID beamline instruments (for comparison) and estimated data generation rates at the Polarization Modulation Spectroscopy (Polar) APS-U feature beamline. For brevity, these tables represent only a fraction of the total 68 beamlines at the APS. (XMCD: X-ray magnetic circular dichroism; XMLD: X-ray magnetic linear dichroism.)

Today	Technique	Data Set Size (GB)	Data Per Day (TB)	Data Per Year (TB)
TODAY	2-ID-D Ptychography	986	8.32	87
	2-ID-D Diffraction	0.20	0.21	13
	2-ID-E XRF	1.91	0.01	2.16
	BNP XRF	0.69	0.01	2.16
APS-U ERA	PtychoProbe XRF	0.65	0.66	35
	PtychoProbe Ptychography	2,043	206.88	8,689
	PtychoProbe Diffraction	40.86	129.30	8,146

Table 5.2.9: Data generation rates today at the 2-ID-D ptychography, 2-ID-E XRF, and BNP XRF instruments (for comparison) and estimated data generation rates at the PtychoProbe APS-U feature beamline. For brevity, these tables represent only a fraction of the total 68 beamlines at the APS.

Today	Technique	Data Set Size (GB)	Data Per Day (TB)	Data Per Year (TB)
TODAY	XPCS	18.61	1.31	102
APS-U ERA	XPCS — Fast	2,235	75	3,072
	XPCS — Fast	2,980	377	15,360
	XPCS — Average	670	141	6,144
	XPCS — Average	1,788	50	2,048

Table 5.2.10: Data generation rates today at the 8-ID instruments (for comparison) and estimated data generation rates at the XPCS APS-U feature beamline. For brevity, these tables represent only a fraction of the total 68 beamlines at the APS.

Today	Technique	Data Set Size (GB)	Data Per Day (TB)	Data Per Year (TB)
TODAY	Wire Scan	3.91	1.32	125
		5.86	1.98	21
		3.91	2.64	249
		5.86	3.96	42
APS-U ERA	Wire Scan — Fast	23.44	28.25	1,187
		39.06	47.08	494
	Wire Scan — Average	23.44	7.06	1,038
		39.06	11.77	124
	Wire Scan — Fast	4.69	28.25	1,187
		7.81	47.08	494
	Wire Scan — Average	4.69	7.06	1,038
		7.81	11.77	124

Table 5.2.11: Data generation rates today at the 34-ID-E Laue diffraction instrument (for comparison) and estimated data generation rates at the 3D Micro and Nano Diffraction (3DMN) APS-U feature beamline. For brevity, these tables represent only a fraction of the total 68 beamlines at the APS.

5.2.2.4 Generalized Process of Science

There is great diversity in the process of science at the APS due to the large and diverse user community. Multiple processes and workflows exist for each beamline, instrument, and technique, and new processes and workflows are continually developed and refined. Refer to the "APS Scientific Computing Strategy" document for detailed workflows for individual beamlines. Many of the processes, however, may be categorized into three main themes:

Processing/Reducing/Analyzing Large Amounts of Data from Light Source Instruments

As described earlier in this document, data is first stored on a computer attached to the acquisition detector or some other local beamline storage system. In many cases, the data is transferred to a local beamline workstation, the local APS distributed-memory computing cluster, or a computing center located on the Argonne campus, such as the ALCF, for processing. Raw and processed data is often stored on a large disk system at the APS using the APS Data Management System from which users may retrieve the data or share the data with collaborators. Researchers may refine results and combine data collected at the APS with data from other sources, including simulations; this is usually performed outside of the facility-based workflow. Over the coming few years, due to increases in data volumes and complexity, and aligning with the completion of the upgrade project, the APS anticipates leveraging supercomputers at the ALCF for much of its data processing needs. See Figure 5.2.11.
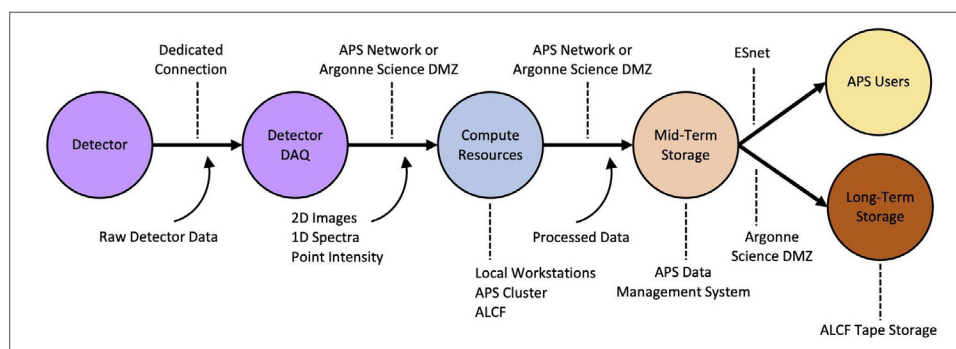
Figure 5.2.11: Generalized APS data processing/reduction/analysis workflow process
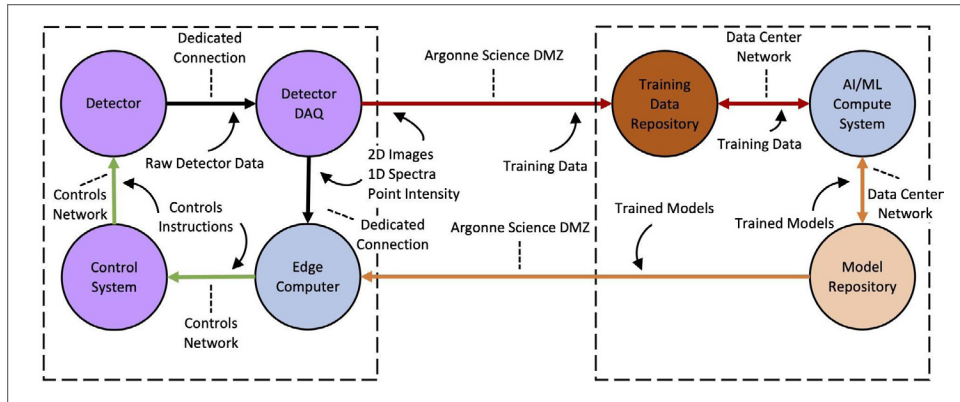
Figure 5.2.12: Data process for an autonomous experiment steering workflow using AI/ML

Today, users analyze much of their APS-generated data at their home institutions. The amount of computing resources needed to analyze APS data is anticipated to greatly increase in the coming years due to the estimated increase in APS data generation rates. This amount of computing resources may likely be out of reach for individual user groups, necessitating the use of APS-provided computing resources.

Increased data generation rates in the APS-U Era will likely be too great for traditional file-based workflows. The APS will utilize streaming-based workflows in which data is transferred from detector system memory directly to computer system memory, bypassing intermediate file systems, requiring robust facility and Laboratory networking. Event mode data collection will also be utilized in the future, requiring the development of new methods for identifying events and analyzing event-based data streams.

Adaptive Feedback and Autonomous Experiment Steering Using Advanced Computational Techniques, Including AI/ML Methods

An area of active development that will soon begin showing benefits is the use of advanced computational techniques, especially AI/ML, for adaptive feedback and autonomous experiment steering[678]. In this scenario, data collected from a detector acquisition system is considered training data for AI/ML models. This data is sent to a large-scale computer where AI/ML models are trained. The training data is stored in a repository for later use. The trained model is also stored in a repository for later use and sent to an edge-computing device near the detector. Subsequent data sets are processed on the edge device near the instrument generating control instructions sent to the instrument. See Figure 5.2.12.

---

[6] Cherukara, M., Nashed, Y., Harder, R.J., "Real-time coherent diffraction inversion using deep generative networks," Sci. Rep. 8.1, 1-8 (2018).

[7] Liu, Z., Sharma, H., Park, J. S., Kenesei, P., Almer, J., Kettimuthu, R., Foster, I. T., "BraggNN: Fast X-ray Bragg Peak Analysis Using Deep Learning," arxiv.org:2008.08198 (2020).

[8] Liu, Z., Ali, A., Kenesei, P., Miceli, A., Sharma, H., Schwarz, N., Trujillo, D., Yoo, H., Coffee, R., Layad, N., Thayer, J., Herbst, R., Yoon, C., Foster, I., "Bridging Data Center AI Systems with Edge Computing for Actionable Information Retrieval," Proceedings of the 3rd Annual Workshop on Extreme-Scale Experiment-in-the-Loop Computing (XLOOP 2021), held in conjunction with SC'21, November 19, 2021.

Coupling Large-scale Simulations, Digital Twins, Surrogate Models, and AI/ML With Experiment Data in Real-time to Drive Experiment Design and Experiment Decisions

Building on the previous process, the coupling of large-scale simulations, digital twins, and surrogate models with AI/ML is an active area of exploration. This mode of science promises to help accelerate scientific discovery, for example, by enabling faster and more complex materials synthesis. In this scenario, a supercomputing resource is used to incorporate experimentally derived data into simulations, the results of which are used to guide experiments in real time and even to plan experiment campaigns and sample compositions (see Figure 5.2.13).

Note: In these cases, users may access the APS and computing facilities remotely.

The U.S. DOE-funded light sources have developed a common vision for computing across the facilities, the DISCUS, and a decade-long roadmap to achieve the vision (see Figure 5.1.12). This vision proposes a transformative computational fabric that covers the full lifecycle of data generated at the BES light sources to accelerate discovery and insight. This vision proposes connecting the 200+ instruments at the light sources to a multitiered computing landscape, including edge, local, campus, and ASCR program compute resources, and discoverable data repositories using high-performance, robust feature-rich networks. This fabric would facilitate the full lifecycle of data across the BES complex, including theory/modeling and simulation, experiment design, data generation at scientific instruments within the light sources, data reduction and processing, analysis and interpretation, and publication and dissemination, serving the 10,000+ light sources users per year.
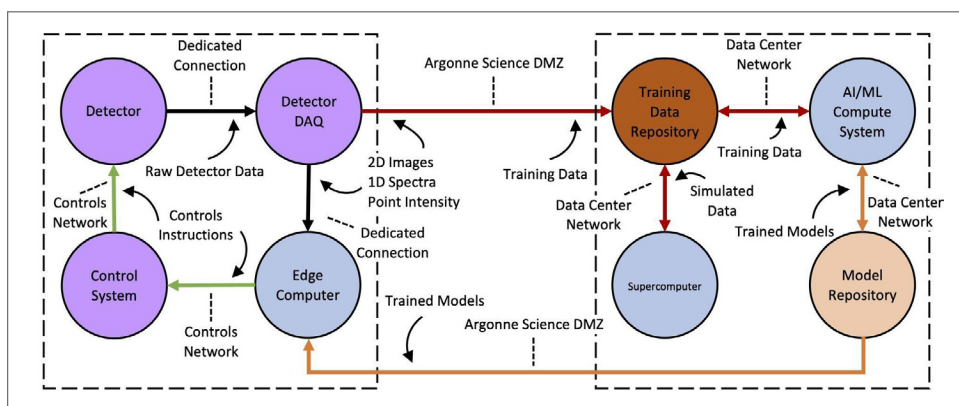
Figure 5.2.13: Data process for coupling large-scale simulations, digital twins, surrogate models, and AI/ML with experimentally derived data in real time

5.2.2.5 Remote Science Activities

The majority of studies carried out at the APS have generally been performed by on-site experimenter-led teams. But there is a growing proportion of these investigations that is performed by remote researchers. These remote measurements are performed either through mail-in programs, where users send arrays of samples to be measured in an automated or semi-automated fashion with facility staff providing oversight of the measurements, or through remote access mechanisms whereby all or a part of the experiment team is geographically remote and remote access tools, particularly NX NoMachine, are used to directly access and manipulate computers running data acquisition applications. Due to COVID-19, the portion of remote experimenters at the BES light sources increased greatly (see Figure 5.2.3 for the APS example). The APS anticipates that remote access will continue to be a substantial mode of operation, and thus a critical infrastructure component for the facility going forward. Remote access experiments have the advantages of increasing the user base and scientific output of the facility and increasing accessibility to the facility for underserved communities.

In the APS-U Era, it will be impractical and unreasonable to support the scale of computing required with only local APS resources. The colocation of the APS and world-leading supercomputing infrastructure at the ALCF on the Argonne campus provides an unprecedented opportunity for collaboration. The APS and ALCF have partnered to deliver a new model of computing, tightly coupling APS experiment instruments with ALCF supercomputers to accelerate scientific discovery (see Figure 5.1.12).

The ALCF deployed a new computing system, Polaris, in 2022. Polaris is a combination commodity CPU/GPU -based system with performance of approximately 44 PFLOPps. This system follows a new model for supercomputing systems, Instrument to Edge (I2E), to better enable use by experimental facilities. Up to 4 PFLOPps of computing will be prioritized to explore on-demand use of high-end computing resources by experimental and observational facilities, including the APS. The APS is currently testing workflows for ptychography, XPCS, and HEDM instruments on the Polaris testing platform, Edith.

Work is underway to test preemptive scheduling queues to provide immediate, on-demand access for APS jobs. Gateway nodes on this system will provide the ability for the APS to stream data directly to Polaris from detectors, avoiding local file input/output. The APS is working with Argonne's Data Science and Learning Division and the Globus Services team to develop a computational data fabric for end-to-end data lifecycle management, Gladier[9].

Combining these new capabilities will provide the necessary coupling between the APS and the ALCF to more seamlessly utilize large computing resources to enable the data processing needed in the APS-U Era. This model, once refined using Polaris, will be deployed on more computing resources at the ALCF. These capabilities will be employed for many other APS techniques and beamlines for data processing during beam time and for postprocessing by APS users after allocated experiment time is over.

Designed in collaboration with Intel and Cray, the 11 PFLOPps Theta supercomputer serves as a stepping stone to the next leadership-class ALCF supercomputer, the Aurora exascale supercomputer, which is scheduled to become available in 2023. Aurora is designed to support numerical simulation, data analysis, and deep learning applications. To this end, it is architected with a mix of Intel central processing units and GPUs to deliver sustained performance of greater than one exaflopps 1018 full-precision floating-point operations per second, and substantially higher compute rates at reduced precision. It will have aggregate system memory of more than 10 PBs. The APS will utilize this new class of supercomputer to couple the results of simulations and modeling with experiment data and train ML models in real time.

---

[9] Vescovi, R., Chard, R., Saint, N., Blaiszik, B., Pruyne, J., Bicer, T., Lavens, A., Liu, Z., Papka, M., Narayanan, S., Schwarz, N., Chard, K., Foster, I., "Linking Scientific Instruments and HPC: Patterns, Technologies, Experiences," https://arxiv.org/abs/2204.05128.

For data storage, the ALCF currently provides approximately 10 PB of tape storage (easily expandable to meet future APS needs) for longer-term data retention. The ALCF has recently deployed a 100-PB CFS (Eagle) and a 100-PB project file system (Grand) along with additional tape storage that is available for APS use. The APS will continue to work closely with the ALCF to help address data storage challenges.

APS users not only perform experiments using the APS, but also at other US DOE-funded facilities, including light sources, neutron sources, and nanoscience research centers, and at their home institutions. The process of science often involves combining and analyzing data from multiple sources. As data rates and complexity continue to increase, sufficient networking connectivity, bandwidth, and reliability is required to connect measurement facilities, computing facility, and user home institutions to enable effective data management and analysis.
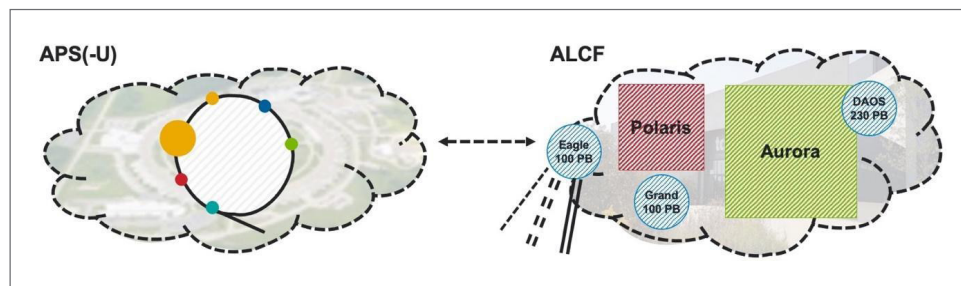
Figure 5.2.14: The APS and ALCF have partnered to deliver a new model of computing, tightly coupling APS experiment instruments with ALCF supercomputers and storage infrastructure, to accelerate scientific discovery

5.2.2.6 Software Infrastructure

The APS accelerator systems and beamline instruments rely primarily on the EPICS for low-level device control. The APS will upgrade EPICS to version 7 in the APS-U Era. The Bluesky[10] suite of tools will be adopted for high-level experiment control as beamlines and instruments are upgraded. As mentioned in the previous section, remote access tools, such as NX NoMachine, will continue to be utilized to enable remote access for APS experiments.

The APS Data Management System, the facility-wide software and hardware infrastructure for managing data and workflows, provides a data management, workflow, and storage system for managing data resources. The APS Data Management System integrates with beamline data workflows and large data storage systems. These tools automate the transfer of data between acquisition devices, computing resources, and data storage systems. Ownership and access permissions are granted to the users signed up to perform a particular experiment.

APS Data Management System[11][12] users can download data at their home institutions by using Globus or SFTP. The APS Data Management System utilizes Globus, GridFTP, remote sync, and secure copy protocol for internal data transfer within the facility. Some industrial partner beamlines at the APS utilize commercial systems, such as Aspera, or other proprietary tools for data transfer.

Since the system's inception in 2013, the APS Data Management System has been adopted at approximately 50 of 68 APS beamlines. The system is in use at both APS-managed and CAT-managed beamlines. Use of the system increased by 10 additional beamlines over a two-year timeframe (2020-2022) to help facilitate remote experiments due to COVID-19. The APS will continue to deliver a multitiered data management and distribution system for all current and future APS beamlines.

---

[10] Allan, D., Caswell, T., Campbell, S., Rakitin, M., "Bluesky's ahead: a multi-facility collaboration for an a la carte software project for data acquisition and management," Synch. Radiat. News 32(3), 19–22 (2019).

[11] Veseli, S., Schwarz, N., Schmitz, C., "APS Data Management System," J. Synchrotron Rad. 25, 1574-1580 (2018).

[12] Schwarz, N., Veseli, S., Jarosz, D., "Data Management at the Advanced Photon Source," Synch. Radiat. News 32:3, 13-18 (2019).

The APS is working with the Argonne Data Science and Learning Division and the Globus Services team to develop a computational data fabric for end-to-end data lifecycle management. This fabric, Gladier, connects and automates many stages of the data lifecycle from acquisition to processing to publication. Science web portals will allow APS users to view and download their data and reprocess their data on large-scale computing resources using Globus Automate and FuncX. The APS and the Globus team have prototyped this computational fabric for XPCS (see Figure 5.2.15) and serial crystallography, and are working to develop such workflows for ptychography, HEDM, and Bragg coherent diffraction imaging over the next two years to be ready for use at the newly upgraded APS facility.
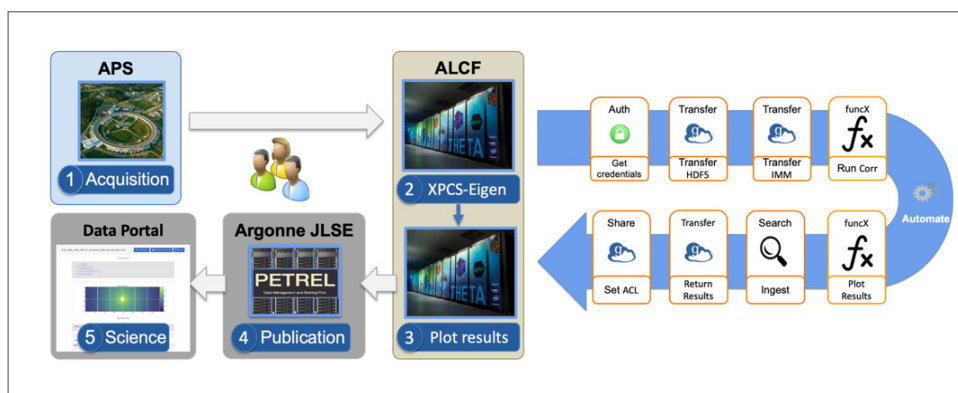
Figure 5.2.15: Automation used to perform on-demand analysis of XPCS data using computing resources at the ALCF. Data is transferred to ALCF where compute nodes are provisioned to perform analysis, extract metadata, and plot results, which are published to a Globus data portal for user analysis.

These tools will be applied at more beamlines in the years after the APS-U and will serve as the basis for enabling searchable data catalogs and adopting FAIR data practices. The Materials Data Facility and the DOE Office of Scientific and Technical Information will serve as digital object identifier-generating services for APS data sets. In the long term, the APS will adopt or develop tools that provide metadata catalogs, electronic logbooks, and sample tracking systems.

The APS is focusing data analysis algorithm and software development in the areas needed to answer the novel scientific inquiries enabled by the future APS. These areas are techniques driven by coherence, imaging, and high-energy, and multimodal techniques. Algorithms and software are being developed to analyze and reconstruct massive data volumes, bridge across length and time scales, identify and classify features and patterns, and provide feedback to experiments dynamically using real-time reduction and novel AI/ML approaches.

Coherence, imaging, high-energy, and multimodal techniques are already the most computationally intensive techniques performed at the APS, and throughput demands are expected to grow by as much as multiple orders of magnitude due to improved detectors and the upgraded source. Data reduction and analysis will rely heavily on the use of HPC, utilizing appropriate technologies such as multithreading, GPUs, edge devices, and distributed computing environments to obtain results with near real-time completion so that results enable user-driven or even automated steering of experiments.

Key software developments have been made in this area over the past years. Applications will continue to be developed for improved performance and algorithms. A complete list of software produced at the APS can be found on the "APS Catalog of Data Analysis Software" web site[13].

---

[13] "The Advanced Photon Source Catalog of Data Analysis Software," www.aps.anl.gov/Science/Scientific-Software.

5.2.2.7 Network and Data Architecture

The Argonne campus network connects 100 buildings and 12 data centers in a multivendor switching and routing network. The campus has a new robust single-mode fiber cable plant that has plenty of capacity and ability to be expanded.

External network connectivity is provided by ESnet, Metropolitan Research and Education Network (MREN), Internet2, and Internet2 Peering Exchange (I2PX). ESnet recently migrated Argonne to the new ESnet6 network and has redundant optical transport and routers deployed on the Argonne campus in a redundant configuration supporting up to 400GE connections.

There are two primary networking node locations distributed between the north (Building 221) and south (Building 541B) sides of the campus to provide network redundancy. Generally, all buildings and data centers have connections to each of these networking nodes. The fiber used to connect to off-site providers is diverse and exits on the east and west sides of the campus.

The Argonne networks supports many connection speeds. The WAN supports 10GE, 40GE, 100GE, and 400GE. The campus core network support speeds of 1GE, 10GE, 40GE, and 100GE. Building connections are generally multiple 10GE with some 40GE. Data centers 10GE, 40GE, and 100GE. LAN connections within buildings are generally 1GE and 10GE.

See Figure 5.2.16 and Figure 5.2.17 for diagrams of the current and anticipated APS network connections to the Argonne campus network, the ALCF, MREN, Internet2, and ESnet. See Table 5.2.12 for planned network upgrades.

ANL WAN

The Argonne WAN network connectivity is provided by ESnet and MREN (Internet2 gateway, Peer Exchange (I2PX)) ESnet has optical and routing gear on the Argonne campus. Argonne has two Juniper MX960 border routers that are split between buildings 221 and 541B for redundancy and diversity.

Argonne uses ESnet OSCARS virtual circuits. (On-Demand Secure Circuits and Advance Reservation System)

The Argonne connections to ESnet are 2x100GE. The connections to MREN are 3x10GE. The border routers are connected to each other with 2x100GE connections. The connections to the campus Core network are 2x40GE and 40GE to the perimeter firewall.

Argonne has one perfSONAR connected to the border at 40GE.

Argonne deploys cyber security at the border with blackhole routing and network traffic capture taps for traffic collection and intrusion detection.

ANL Science DMZ

The Argonne ScienceDMZ is composed of Juniper QFX series equipment that is connected directly to the Argonne border routers with redundant 2x100GE links. The ScienceDMZ provides high-speed connectivity between scientific organizations for the exchange of data without taking the traditional path through a firewall.

Connections in to the Science DMZ are a minimum of 100GE with most facilities connecting at 2x100G. This allows collaborators to exchange data in a high-speed environment that does not affect commodity network connectivity.

ANL Core

The Argonne core network is provided by 2 Cisco Nexus 7710s spread across buildings 221 and 541B for redundancy. These switches offer network speeds at 1GE, 10GE, 40GE, and 100GE. The Core provide connections to the buildings and many of the data centers on the Argonne campus.

ANL LAN

The Argonne LAN consists of the building access and aggregation switches. The Laboratory has standardized on Cisco and Aruba switches in the LAN. There is a combination of 1GE, 10GE, and 40GE network speeds in the LAN. Desktops are connected at 1GE.

ANL Data Centers

The primary data center at Argonne consists of Cisco Nexus equipment in an application centric infrastructure (ACI) fabric. It connects to the Argonne core network at 4x40GE and offers 2x40GE connectivity to every top of rack switch. There is a combination of 1GE, 10GE, 25GE, and 40GE to host in this space.

APS

Refer to Figure 5.2.7 for a diagram of the APS beamline network and architecture. The center of the APS beamline network consists of a pair of core switches (HPE Aruba 6410) located in the APS data center. These Tier 1 switches provide all routing to beamline subnets and to other parts of the APS, Argonne, and the Internet via ESnet. The core switches are configured in a redundant active/active configuration. The core switches provide multiple 40/50/100-Gbps ports. These core switches are connected via 2 x 40-Gbps uplinks to the APS Tier 2 firewall, which in turn connects to the Argonne Tier 1 firewall with 2 x 100-Gbps uplinks. The Tier 1 Argonne firewall connects to the Internet via ESnet using 2 x 100-Gbps uplinks. The APS core switches also connect directly to the ALCF via 2 x 100-Gbps uplinks. The same core switches connect to the storage systems for the APS Data Management System, sector data storage systems, and the data servers that host beamline control system configurations and software to interface with the APS accelerator network.

Each sector at the APS has a Tier 2 switch (HPE Aruba 6410) that serves to connect beamline devices and to connect the beamline to the core APS switches. The Tier 2 switches connect to beamline computers, control system EPICS IOCs, detectors and data acquisition servers, wireless access points, cameras, and controls hardware. Each Tier 2 beamline network switch will provide line rate 10/100/1000 megabits (Mbps) ports for the majority of devices at the beamline, as well as high-speed line rate 10/25/40/50/100-Gbps ports for data acquisition where needed. Uplinks to the APS Tier 1 core switches will be sized appropriately based on beamline needs.

A Tier 3 managed switch with 48 x 10/100/1000-Mbps ports may be deployed at each experiment hutch f or controls hardware stations to provide a dynamic cabling environment and to isolate beamline controls hardware traffic.

The APS will adopt a supervisory control and data acquisition architecture for the APS-U beamline control system network. Controls and data analysis network traffic will be separated and isolated from outside networks for maximum performance and security. Wireless access points can also be provided inside the hutch to support, for instance, advanced sensors or augmented-reality headsets.

An additional 96 pairs of single-mode fiber were installed from the APS data center to each of the laboratory office module network closets (768 pairs in total). This additional fiber infrastructure will provide sufficient network bandwidth from the beamlines to the data center for the next decade.
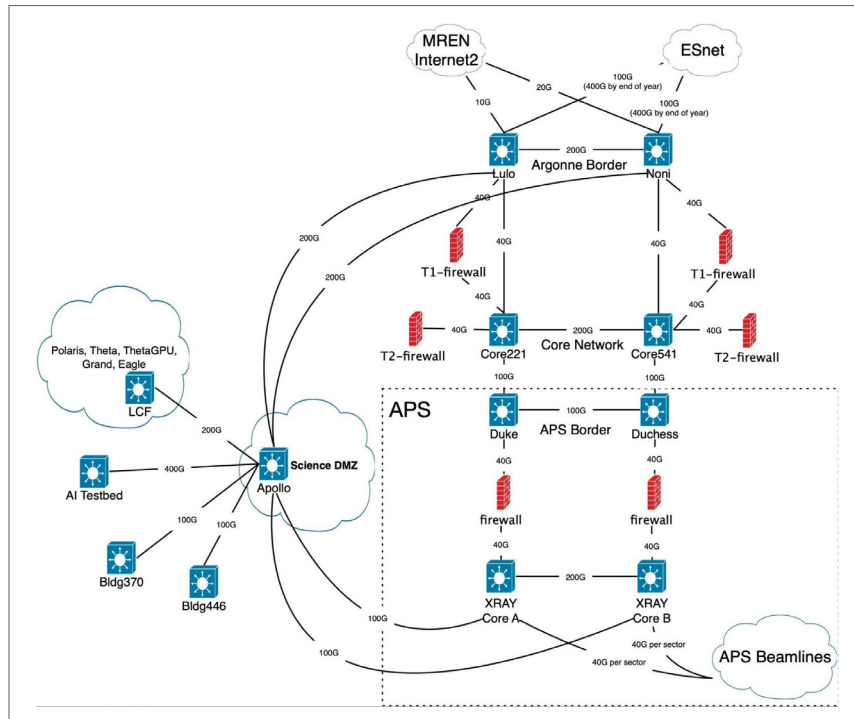
Figure 5.2.16: Diagram of the present APS network connections to the Argonne campus network, the ALCF, MREN, Internet2, and ESnet
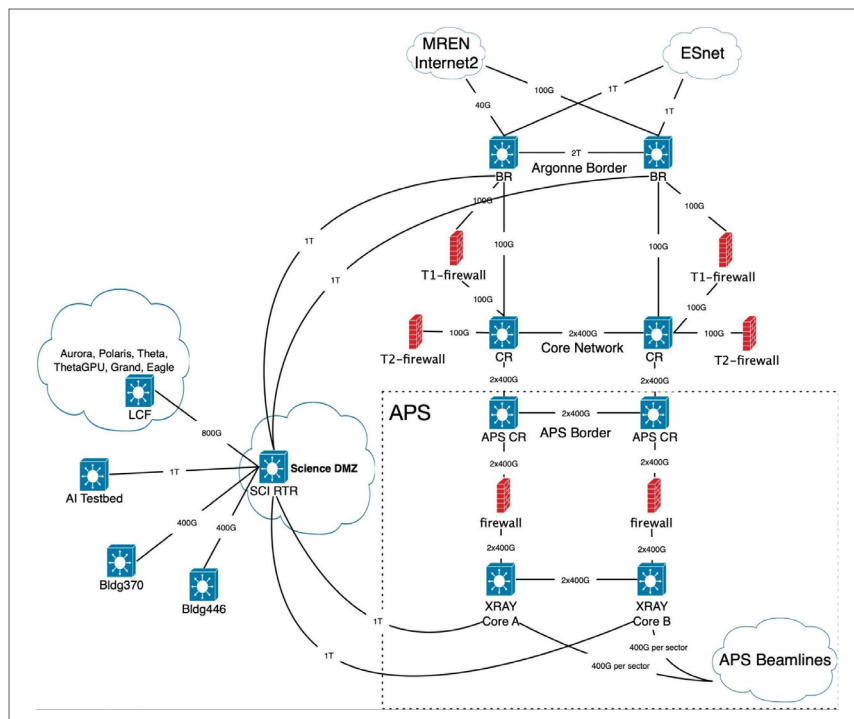
Figure 5.2.17: Diagram of the anticipated APS network connections to the Argonne campus network, the ALCF, MREN, Internet2, and ESnet in five years

	Present – 2 Years	2 – 5 Years	5+ Years
APS	• Complete individual beamline network switch upgrades • Increase bandwidth for each APS sector to dual 100-Gbps or dual 200-Gbps links as needed	Continue to increase bandwidth for each APS sector to dual 100-Gbps or dual 200-Gbps links as needed	• Networking hardware refresh • Terabit connectivity to the Argonne ScienceDMZ • Dual 400-Gbps uplinks from each APS sector • Dual 400-Gbps uplinks to the regular campus network
Data Centers	Hardware refresh for top of rack switches	Spine switch replacement	Terabit connectivity
LAN	• Continual refresh of switches older than 5 to 7 years • Wireless access point upgrades	• Consider an overlay network topology • Ongoing switch refreshes • Zero-trust networking	• Upgrade to next-generation devices • 10-Gbps to desktops
Core	No changes	Upgrade to next-generation devices	Terabit connectivity
Science DMZ	400 Gbps connectivity	Router upgrade	Terabit connectivity
WAN	400GE upgrade	Upgrade border routers Multiple 400-Gbps links	Terabit WAN connections

Table 5.2.12: Planned Argonne network upgrades

5.2.2.8 Cloud Services

The APS does not presently heavily utilize cloud services for its scientific data infrastructure. The APS does not have any immediate plans to further utilize cloud services in this area. The APS does rely on Globus tools for distributing data to its users and for data transfer to computing centers, and is prototyping workflows using Globus Automate, Globus FuncX, and Globus Gladier. The APS has prototyped the limited use of Microsoft Azure for one scientific computing use case and has prototyped the use of containers on the Argonne Laboratory Computing Resource Center. In the past, the APS utilized the Argonne Magellan platform for XPCS data reduction. The APS does utilize a number of software-as-a-service products for communication with its users and to enable remote experiments, including Box, GitHub, Slack, and Zoom. APS users may utilize cloud services outside of the APS infrastructure as a part of their data management and data analysis efforts.

The APS, along with the other US DOE-funded light sources, continues to evaluate the capabilities and cost-effectiveness of using cloud services.

5.2.2.9 Data-Related Resource Constraints

The APS is facing a multiple order-of-magnitude increase in demand for computing resources over the next decade. Data management and workflow tools are needed that integrate beamline instruments with computing and storage resources for use during experiments, as well as to facilitate user access. Real-time data analysis capabilities are required to significantly reduce data volumes and provide feedback during experiments to improve data quality and to drive the direction of ongoing research. The application of advanced mathematical algorithms, ML, and the integration of simulations and model-based approaches will allow automated steering of data collection. On-demand utilization of computing environments is required to enable near real-time data processing. Sufficient data storage and archival resources to house the continually increasing amounts of valuable

scientific data produced by the APS is required. Advances in networking serve as the basis to realize these critical capabilities.

5.2.2.10 Outstanding Issues

None to report at this time.

5.2.2.11 Facility Profile Contributors

APS Representation

- **Laurent Chapon**, *ANL*, lchapon@anl.gov
- **Nicholas Schwarz**, *ANL*, nschwarz@anl.gov
- **Richard Fenner**, *ANL*, fenner@anl.gov
- **David Leibfritz**, *ANL*, leibfritz@anl.gov
- **Alec Sandy**, *ANL*, asandy@anl.gov
- **Brandon Siegel**, *ANL*, bsiegel@anl.gov
- **Stefan Vogt**, *ANL*, svogt@anl.gov
- **Stuart Campbell**, *BNL*, scampbell@bnl.gov
- **Alexander Hexemer**, *LBNL*, ahexemer@lbl.gov
- **Apurva Mehta**, *SLAC National Accelerator Laboratory*, mehta@slac.stanford.edu
- **Vivek Thampy**, *SLAC National Accelerator Laboratory*, vthampy@slac.stanford.edu
- **Jana Thayer**, *SLAC National Accelerator Laboratory*, jana@slac.stanford.edu

ESCC Representation

- **Linda Winkler**, *ANL*, winkler@mcs.anl.gov
- **Corey Hall**, *ANL*, chall@anl.gov

5.3 National Synchrotron Light Source II (NSLS-II)

The National Synchrotron Light Source II (NSLS-II) is a state-of-the-art synchrotron facility that provides extremely stable and bright photon beams, from infrared to hard X-rays, and data infrastructure to enable multiscale, multimodal, high-resolution studies on diverse systems of materials.

5.3.1 Discussion Summary

- NSLS-II began user operations in 2015.

- Approximately 1,800 distinct researchers used NSLS-II beamlines for their research prior to the pandemic, with approximately 1,400 seen during the pandemic years.

- The NSLS-II currently operates 28 beamlines, each with its own unique data generation characteristics. Depending on the beamline and measurement, beam time may range from a few hours to more than a week in duration.

- There has been a trend to increasing remote access to synchrotrons, thanks to the ongoing advances in robotic sample handling, automated data acquisition and data management, technology improvements to light sources and detectors, and the COVID pandemic.

- The pandemic has heightened the needs for NSLS-II capabilities be operated and accessed remotely, both by staff and by the users. NSLS-II has aggressively ramped up the infrastructure and capabilities for remote access, along with associated cyber security measures. As the need for tools such as AI/ML to reduce human intervention increases, remote capabilities are of high strategic importance for the future health of the facility.

- The meaning of data, the size of data, the processing steps applied to data, and the analysis and interpretation of data vary depending on the beamline, the measurement technique, the detector(s) utilized, and the scientific goal. The size of the data may vary from a few megabytes to hundreds of terabytes per allocated beam time. Some form of data processing or reduction is usually performed after data is collected.

- NSLS-II has a common approach to control software. All beamlines run the Bluesky software suite for collecting, storing, managing, and partially analyzing data. Generally, data is first stored on a computer attached to the acquisition detector and then written to a shared central storage system. In addition to storing the raw data, processed data is also often stored on the central shared storage system.

- Use of multimodal data requires more sophisticated data processing, and requires increases in computing capabilities which can include training AI/ML models as well as real-time analysis and feedback to enable autonomous experiment steering. The APS is exploring the utilization of edge-computing resources, coupled closely to detectors and instruments, to facilitate AI/ML data reduction algorithms.

- Although the data generated by individual experiments varies, over the next decade the combined data generation rates for NSLS-II (and other DOE-funded light sources) will reach the exabyte per year range. Processing power of hundreds of PFLOPps is expected, which implies strong networking capabilities top link light source instruments to edge, local, campus, and centralized computing facilities reliability, and with low latency. Terabit per second networking and beyond will be required to handle the large amounts of data expected in the coming years.

- The computing resources required by the NSLS-II are anticipated to grow steadily with the overall increase of data generation rates. For the most demanding computational problems, large-scale computing facilities must be used, including the ALCF, NERSC, and the OLCF.

Sending data from a light source facility to the HPC facility requires a data path of sufficient bandwidth.

- Experimenters often transfer the raw data to computing resources at their home institutions for all further processing. The final analysis and interpretation of the data for publication is generally carried out by the experiment team at their home institutions, is very experiment and technique specific, and may take months to even years to perform.

- The NSLS-II only guarantees storage space for a minimum of 12 months; the long-term management of data is the responsibility of the user group that collected the data. The NSLS-II provides approximately 3.2 PB of central disk storage, which is expandable.

- Several DTNs are available for reliable, high-speed data movement. NSLS-II users can download data at their home institutions by using Globus (paid subscription) or sftp. Users often utilize their own "cloud storage" accounts (e.g., Google Drive, Dropbox, etc.) to transfer data.

- There is a high demand from the user community to integrate cloud services for both file transfer (e.g., Dropbox, Google Drive, etc.) and for communication (e.g., Slack) into all areas of the data lifecycle and compute workflows. We have seen this demand increase during the COVID-19 pandemic.

- BNL features a Tbps HTSN that serves as the primary network transport for all data intensive collaborations at BNL, and access to HPC and High-Throughput Computing (HTC) resources internal and external to the lab.

- The BNL network perimeter includes 3x100 Gbps connections to ESnet, and average 15–20 PB of data monthly.

- The NSLS-II facility is connected via a 400 Gbps to the BNL HTSN to meet the demands of ever-increasing data rates and volumes.

- BNL will upgrade network capabilities in the one- to two-year timeframe to support 400 Gbps connectivity, and beyond.

5.3.2 NSLS-II Facility Profile

As a US DOE Office of Science User Facility, the National Synchrotron Light Source II (NSLS-II) enables a collaborative, holistic approach to advance scientific endeavors by offering free access to highly advanced instruments and unique expertise. NSLS-II creates light beams 10 billion times brighter than the sun, directing them towards specialized experimental stations called beamlines (see **Figure 5.3.1**).

Figure 5.3.1: Aerial View of the National Synchrotron Light Source II (NSLS-II)
Currently, NSLS-II has 28 beamlines in operation, and 4 additional beamlines under construction. These beamlines (see Figure 5.3.2) offer unique, cutting-edge research tools, including higvh-throughput robot-driven sample processing, high spatial resolution imaging, and high-energy resolution. The storage ring is routinely operated at 400 mA and 30 pm-rad vertical emittance, and at 8 pm-rad during special running periods. NSLS-II has been implementing a multiyear development plan for the accelerator systems. Once this is completed, it will allow the NSLS-II storage ring to be reliably operated at its mature performance specifications of 500 mA and 8 pm-rad vertical emittance.

5.3.2.1 Science Background

As one of the newest, most advanced synchrotron light sources in the world, NSLS-II enables its growing research community to study materials with nanoscale resolution and exquisite sensitivity by providing cutting-edge capabilities. Together with visiting researchers from all around the world, interdisciplinary teams at NSLS-II uncover the atomic structure, elemental make-up, and electronic behavior of materials. By creating this new, deeper understanding of materials, these research teams advance our knowledge in a wide range of scientific disciplines such as life sciences, quantum materials, energy storage, advanced materials science, physics, chemistry, and biology.
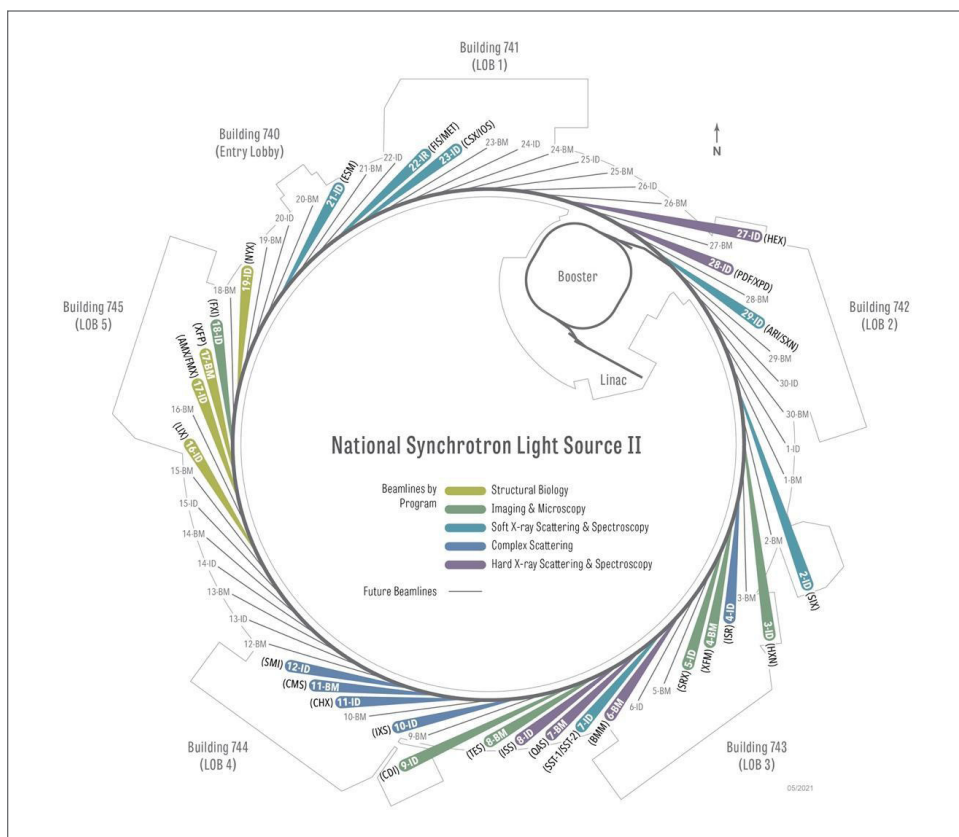
Figure 5.3.2: NSLS-II beamline diagram based on current funding assumptions

All beamlines at NSLS-II are organized into five photon science programs, based on the research capabilities they offer and staff expertise. Beyond the current existing portfolio of beamlines, NSLS-II has the capacity for additional 30 beamlines with four new beamlines already under development. To ensure all research tools are always cutting edge, we have a team of experts working on advancing detectors, X-ray optics, precision engineering, and positioning, as well as on theoretical simulations. Detailed information about computing and data needs, plans, remote access and gaps may be found in the NSLS-II Strategic Plan document[1].

Part of the holistic approach to research at NSLS-II is a common approach to control software for scientific instruments. All NSLS-II beamlines run the Bluesky software suite for collecting, storing, managing, and partially analyzing data. Generally, data is first stored on a computer attached to the acquisition detector and then written to a shared central storage system. In addition to storing the raw data, processed data is also often stored on the central shared storage system.

The meaning of data, the size of data, the processing steps applied to data, and the analysis and interpretation of data vary depending on the beamline, the measurement technique, the detector(s) utilized, and the scientific goal. Raw data is generated primarily by two-dimensional (area), one-dimensional (strip), or point detectors as a part of scattering, imaging, or spectroscopy measurements. The data from all these detectors either takes the form of an image or histogram, or as a stream of time stamped photon detection events. Raw data may be represented by a scattering pattern, a transmission image, or spectra, for example. The size of the data may vary from a few megabytes to hundreds of terabytes per allocated beam time. Some form of data processing or reduction is usually performed after data is collected to transform the data from a technique or detector representation to an analyzable representation, such as a series of sinograms into a three-dimensional world-space volume, or spectra into elemental concentrations.

---

1 The National Synchrotron Light Source II Strategic Plan, https://www.bnl.gov/nsls2/docs/pdf/nsls2-strategic-plan.pdf

Currently, the NSLS-II only guarantees storage space for a minimum of 12 months; the long-term management of data is the responsibility of the user group that collected the data[2]. In some cases, experimenters transfer the raw data to computing resources at their home institutions for all further processing. The final analysis and interpretation of the data for publication is generally carried out by the experiment team at their home institutions, is very experiment and technique specific, and may take months to even years to perform.

5.3.2.2 Collaborators

NSLS-II began user operations in 2015. In fiscal year 2019, 1755 distinct researchers used NSLS-II beamlines for their research. This number in fiscal year 2020 was 1355 due to the impact of the global COVID-19 pandemic. Despite this, NSLS-II saw 577 publications in calendar year 2020, with more than 40% of these publications in journals with impact factor great than 7.0, demonstrating the high impact of our relatively young facility.

NSLS-II operates in three 4-month cycles and all beam time is requested each cycle through the web-based Proposal Allocation, Safety, and Scheduling System (PASS) system. In this system, a beam time proposal describes the scientific experiments to be performed and identifies the experimental team. Proposals have different durations depending upon the type of proposal submitted. For any proposal that has a duration of more than one cycle, a beam time request (BTR) must be submitted for every cycle that a user requests beam time. Every BTR also requires a Safety Approval Form, which is also submitted through the PASS system once beam time is allocated.

There are several modes of access for beam time at NSLS-II, which are described in detail below. However, they all follow a common life cycle that involves the following steps: proposal submission, feasibility evaluation, peer review, allocation of beam time, scheduling, carrying out the work, and reporting the results.

- **General User (GU) Proposals** — Most proposals requesting beam time at NSLS-II are GU proposals. GU proposals are for scientists that require beam time on beamlines that routinely support the technique needed for their experiment. For GU experiments, users often bring only samples, but can also provide custom instrumentation for the duration of their experiments. GU proposals are valid for one year (3 beam time cycles). Up to 3 beamlines and 2 CFN instruments may be requested on a GU proposal. If a proposal is allocated beam time, it will also receive time on the CFN instrument(s), subject to feasibility review and sufficient availability. Each proposal requests a lifetime number of 8-hour shifts to complete the work for each beamline requested. For each cycle that a user requests beam time, he/she must submit a BTR against their GU proposal. These proposals are peer reviewed and allocated by the NSLS-II Proposal Review Panel. All GU proposals are considered active until either: (a) all beam time allocated to the proposal for its lifetime has been used, (b) the proposal is withdrawn, or (c) one year has elapsed.

- **Rapid-Access (RA) Proposals** — RA proposals are specifically for rapid access to beam time f or "hot topics" or for straightforward experiments using routine techniques with a fast turnaround time. RA proposals are valid for one beam time cycle and typically request a very small amount of beam time on one beamline. These proposals are peer reviewed and allocated by the NSLS-II Proposal Review Panel with a 1–2-week turnaround time prior to running the experiments.

- **Joint Bio-SAXS / Bio-SANS Proposals** — A collaboration has been established between the Bio-SAXS (LiX) beamline at NSLS-II and the Bio-SANS beamline at HFIR (ORNL) to form a SAXS/SANS Joint Access Program. In this program, users only need to submit one beam time proposal for access to both facilities. In this proposal, the user must justify the need for both techniques.

---

2 The National Synchrotron Light Source II Data Management Policy, https://www.bnl.gov/nsls2/userguide/post-experiment.php#step3

- **Block Allocation Groups** (**BAGs**) — A mode of beam time access at NSLS-II intended for groups of researchers that want to combine their short beam time requests into a single proposal in order to permit greater flexibility in beam time allocation and scheduling. BAG proposals may be motivated by shared scientific interest, geographical location, affiliation, common experimental setup, or other synergistic reasons. Combining the beam time of individual groups permits greater flexibility in the choice of projects and samples during a given allocation period and offers the individuals in the BAGs the benefit of access to more regular allocation of beam time. The term of a BAG proposal is two years (six beam time cycles). Up to five beamlines may be requested for a BAG proposal.

- **Partner User** (**PU**) **Proposals** — Partner Users (PU) proposals are for individuals or groups who need regular access to beam time on NSLS-II beamlines to carry out their work and who also wish to partner with the facility in making contributions that benefit other facility users by enhancing the utilization or capabilities of the facility or contributing to its operation. Possible examples include, but are not limited to, creating or expanding a user community, contributing a sophisticated endstation, contributing staff and/or equipment to provide user support for a given program, or the design, construction, or operation of endstation equipment, or even a whole beamline. NSLS-II staff may be PU members or PIs with the approval of the NSLS-II Director. The lifetime of a PU proposal will typically be up to three years, although it may be up to five years in special circumstances. Each beamline at NSLS-II has a maximum 40% of the available beamtime available for all PUs on that beamline.

- **Proprietary Proposals** — Proprietary research is work conducted under a Class Waiver for Proprietary Users of Energy Research Designated User Facilities. Private individuals, representatives from educational institutions, nonprofit organizations, or industry, may conduct such research. Under the terms of the DOE Class Waiver, the user is obligated to pay the full-cost recovery rate for use of NSLS-II. In return, the user has the option to take title to any inventions made during the proprietary research program and to treat as proprietary all technical data generated during the proprietary research program. The terms and conditions under which proprietary research may be conducted at the NSLS-II are set forth in the Proprietary User's Agreement, which must be in place before any experiment can commence. Proprietary work requires the submission of a Proprietary Proposal, which must contain a functional nonproprietary description of the work. Proprietary proposals are RA proposals that are reviewed by NSLS-II management.

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
US UNIVERSITY-BASED PIS	Both	Data portal, data transfer (Globus, sftp, cloud storage), portable hard drive	1 MB–10 TB	ad hoc	N	N/A
US NATIONAL LABS BASED PIS	Both	Data portal, data transfer (Globus, sftp, cloud storage), portable hard drive	1 MB–10 TB	ad hoc	N	N/A
INTERNATIONAL PIS	Both	Data portal, data transfer (Globus, sftp, cloud storage), portable hard drive	1 MB–10 TB	ad hoc	N	N/A

Table 5.3.1: NSLS2 Collaboration Space

Depending on the beamline and measurement, beam time may range from a few hours to more than a week in duration. As previously stated, NSLS-II users perform experiments covering many scientific fields, such as life sciences, quantum materials, energy storage, advanced materials science, physics, chemistry, and biology. In response to the pandemic, the deployment of remote access tools has enabled experimenters to run measurements remotely and has led to an increase in remote users as a fraction of total users in FY 2021. NSLS-II caters for users that come from both throughout the United States and overseas.

5.3.2.3 Instruments and Facilities

The NSLS-II currently operates 28 beamlines, each with its own unique data generation characteristics. NSLS-II continues to develop new beamlines to meet the research needs of the scientific community. Currently NSLS-II is constructing a high-energy X-ray scattering and imaging beamline (HEX), funded by New York State, to be completed in 2022. In addition, three undulator beamlines, Coherent Diffraction Imaging (CDI), ARPES and RIXS Imaging, and Soft X-ray Nanoprobe, are under construction in a DOE major items of equipment project, NEXT-II, to be completed in 2026-28 timeframe. At that time, NSLS-II will have 32 beamlines in operations, out of its capacity of ~60 beamlines.

NSLS-II beamlines often have many dozens of motors, readout electronics, and X-ray detectors that are controlled by real-time hardware devices. Many X-ray detectors have specific manufacturer-supplied interfaces. FPGA or ARM devices are often used for real-time coordination of devices used during measurements.

To meet the demands of ever-increasing data rates and volumes, there have been recent upgrades to the NSLS-II networking. The NSLS-II facility is now connected via a 400 Gbps to the BNL HTSN. Refer to Section 5.3.2.7, Network and Data Architecture, for a detailed description of the BNL HTSN network.

The utilization of multimodal data to answer new questions requires more complex and sophisticated data processing algorithms requiring increases in computing capabilities. Increases in computing power are needed by advanced algorithms for existing techniques that, for example, provide higher-fidelity results, and to train AI/ML models. The need for real-time analysis and feedback to make crucial experiment decisions and enable autonomous experiment steering also requires more computing cycles than have been traditionally utilized.

The computing resources required by the NSLS-II are anticipated to grow steadily as beamlines mature, new beamlines are constructed, and detector developments all contribute to increase the overall data generation rates. There is wide variability in the computational requirements among techniques and processing approaches

with those instruments and techniques that benefit most from high-energy, high-brightness, and coherent X-rays driving most requirements [3].

Edge computing offers the ability to process data quickly on or near detectors and experiment instrumentation without the need to first transfer all data to high-end computing resources. This is particularly promising for handling large data when coupled with machine-learning methods. Using only a subset of data, machine-learning models may be trained on supercomputers. The trained model is then run using edge-computing devices to process newly acquired data, providing fast feedback for experiment steering.

The pandemic has heightened the needs for advanced, readily available capabilities at NSLS-II, as well as the need for improved experiment infrastructure and processes to allow these capabilities be operated and accessed remotely, both by staff and by the users. NSLS-II has been aggressively ramping up the infrastructure and capabilities for remote access, along with associated cyber security measures. This is both a short-term need driven by the current pandemic, and a long-term need to take full advantage of the high-brightness and high-throughput capabilities that NSLS-II provides. As the time for a single measurement decreases, the need for tools such as AI/ML to reduce human intervention increases. Thus the remote experiments capabilities are of high strategic importance for the future health of the facility.

Currently, the NSLS-II provides approximately 3.2 PB of central disk storage (which is easily expandable, and is scheduled to be expanded to 6.4 PB before the end of 2022) for short- and medium-term data retention, and several DTNs for reliable, high-speed data movement. The NSLS-II is in the process of collecting revised detailed data rate information for all its beamlines to include all the improvements and upgrades that have been implemented (and also those that are planned) during the initial years of operations.

For the most demanding computational problems, large-scale computing facilities must be used, including the ALCF, NERSC, and the Oak Ridge Leadership Computing Facility (OLCF).

5.3.2.4 Generalized Process of Science

There is great diversity in the process of science at the NSLS-II due to the large and diverse user community. Multiple processes and workflows exist for each beamline, instrument, and technique, and new processes and workflows are continually developed and refined. Many of the processes, however, may be categorized into three main themes:

Processing/Reducing/Analyzing Large Amounts of Data from Light Source Instruments

Experiments that generate complex and large data volumes are challenging to execute and often require fast turnaround between data acquisition and data analysis to enable experimenters to make informed decisions when driving experiments to make use of the limited resources of X-ray beam time and samples. Data collection rates are growing exponentially due to light source, instrument and detector upgrades, and computational requirements are also growing in proportion. To analyze data on experiment timescales, it is necessary to send data to remote HPC facilities such as NERSC, the ALCF, or the OLCF. Sending data from a light source facility to the HPC facility requires a data path of sufficient bandwidth. Raw and processed data is often stored on a large disk system at the NSLS-II from which users may retrieve the data or share the data with collaborators. Researchers may refine results and combine data collected at the NSLS-II with data from other sources, including simulations; this is usually performed outside of the facility-based workflow. Over the coming few years, due to increases in data volumes and complexity, the NSLS-II anticipates leveraging computational resources outside of BNL.

---

3 Schwarz N., Campbell S., Hexemer A., Mehta A., Thayer J. (2020) Enabling Scientific Discovery at Next-Generation Light Sources with Advanced AI and HPC. In: Nichols J., Verastegui B., Maccabe A., Hernandez O., Parete-Koon S., Ahearn T. (eds) Driving Scientific and Engineering Discoveries Through the Convergence of HPC, Big Data and AI. SMC 2020. Communications in Computer and Information Science, vol 1315. Springer, Cham.

Adaptive Feedback and Autonomous Experiment Steering Using Advanced Computational Techniques, Including AI/ML Methods

Autonomous experiment steering is an emerging mode of conducting science at the 5 US DOE-funded light sources. Within the light sources, this capability has the promise to provide automated setup of the source and sample alignment, intelligent data collection, quality verification, data reduction, and coupling of experimentally derived data with information derived from theory, models, and simulations[456]. Autonomous experiment steering has the potential to unlock new materials science knowledge to, for example, better understand failure modes in materials, enable the synthesis of new materials, aid in the creation of purpose-built designer materials, and assist in additive manufacturing processes. Feedback must often be obtained on timescales too short for humans to react to plan and steer experiments to, for example, catch rare events or see fast processes. Due to the intrinsic capabilities of the current and soon to be upgraded sources, coupled with high data rate detectors that generate large volumes of data at increasingly higher rates, advanced computational techniques, including AI/ML, must be employed to realize autonomous steering of light source experiments. These methods require the utilization of considerable supercomputing power to process data and train AI/ML models that may then be used to make real-time decisions using edge or local computing systems

An area of active development that will soon begin showing benefits is the use of advanced computational techniques, especially AI/ML, for adaptive feedback and autonomous experiment steering. In this scenario, data collected from a detector acquisition system is considered training data for AI/ML models. This data is sent to a large-scale computer where AI/ML models are trained. The training data is stored in a repository of training data for later use. The trained model is also stored in a repository for later use and sent to an edge-computing device near the detector. Subsequent data sets are processed on the edge device near the instrument generating control instructions sent to the instrument.

Although the data generated by individual experiments may vary greatly, over the next decade the combined data generation rates of the 5 US DOE-funded light sources is expected to approach the exabyte per year range (see **Figure 5.3.3**). In order to process this data, the light sources are expected to require tens to many hundreds of PFLOPps of on-demand processing capacity (see **Figure 4**). Networking must be able to connect light source instruments to edge, local, campus, and centralized computing facilities reliability, and with low latency. Terabit per second networking and beyond will be required to handle the large amounts of data expected in the coming years.

Autonomous experiment steering will become an increasingly relied upon capability at the light sources as the decade progresses. Due to the high computational cost associated with data processing, reduction, and analysis, and of model training on such large data sets, the light sources will need to employ the use of geographically distributed computing facilities and data repositories. This necessitates high bandwidth, low latency network connections between the light sources and remote computing and data storage centers.

4 Phillip M Maffettone et al 2021 Mach. Learn.: Sci. Technol. 2 025025

5 Campbell SI, Allan DB, Barbour AM, et al (2021) Outlook for artificial intelligence and machine learning at the NSLS-II. Machine Learning: Science and Technology 2:013001. https://doi.org/10.1088/2632-2153/abbd4e

6 D. Olds, D. B. Allan, T. A. Caswell, J. Lynch, P. M. Maffettone and S. I. Campbell, "Optimizing High-Throughput Capabilities by Leveraging Reinforcement Learning Methods with the Bluesky Suite," 2021 3rd Annual Workshop on Extreme-scale Experiment-in-the-Loop Computing (XLOOP), 2021, pp. 36-42, doi: 10.1109/XLOOP54565.2021.00011, https://ieeexplore.ieee.org/document/9652809.
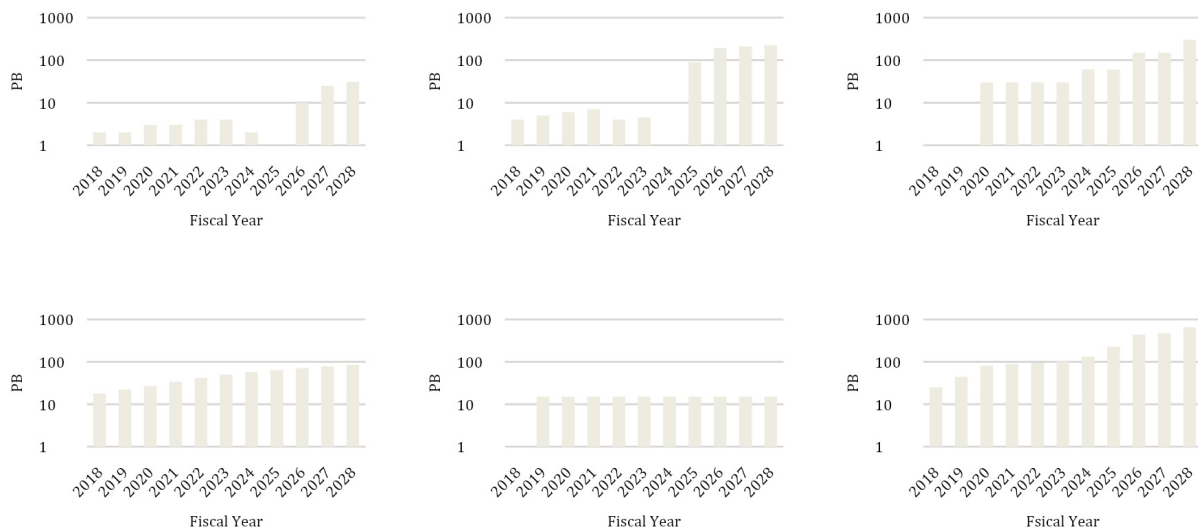
Figure 5.3.3: Log Scale: Estimated data generation rates per year at the BES light sources. At the ALS and APS, data generation will stop during 2025 and 2024, respectively, due to installations of new storage rings. Aggregate data generation across the BES light sources will approach the exabyte (EB) range per year by 2028. The differences in data generation rates across the facilities depend on the number, rate, and resolution of the detectors at each instrument which in turn depend on factors like the brightness of the source and the actual requirements of the experiment technique specific to that particular instrument.

YEAR	Facility				
	ALS	APS	LCLS/LCLS-II	NSLS-II	SSRL
2022	0.1 PFLOPS	4 PFLOPS	1–100 PFLOPS	2.5 PFLOPS	< 1 PFLOPS
2028	30 PFLOPS	50 PFLOPS	50–1000 PFLOPS	45 PFLOPS	< 1 PFLOPS

Table 5.3.2: Estimated PFLOPps of on-demand computing resources required by each of the BES light sources by 2022 and 2028

Coupling large-scale simulations, digital twins, surrogate models, and AI/ML with experiment data in real-time to drive experiment design and experiment decisions: Coupling of simulations and development of "digital twins" to light source experiments has the potential to unlock new materials science knowledge to, for example, better understand failure modes in materials, enable the synthesis of new materials, aid in the creation of purpose-built designer materials, and assist in additive manufacturing processes. It will also allow for a more efficient and optimum use of beamtime at the light sources. For a high-level flow chart of combining experiments with digital twins, see **Figure 5.3.4**.

The coupling of simulations with light source experiments can be split up into three main areas in the experimental life cycle. Firstly, before the experiment, simulations can be used to help prepare, plan, and determine if the experiment is even feasible. Secondly, during the allocated beam time, simulations can help guide and inform the strategy and guide the experiment. Finally, simulations can be used to aid in the data analysis in order to extract the maximum scientific information from the data.
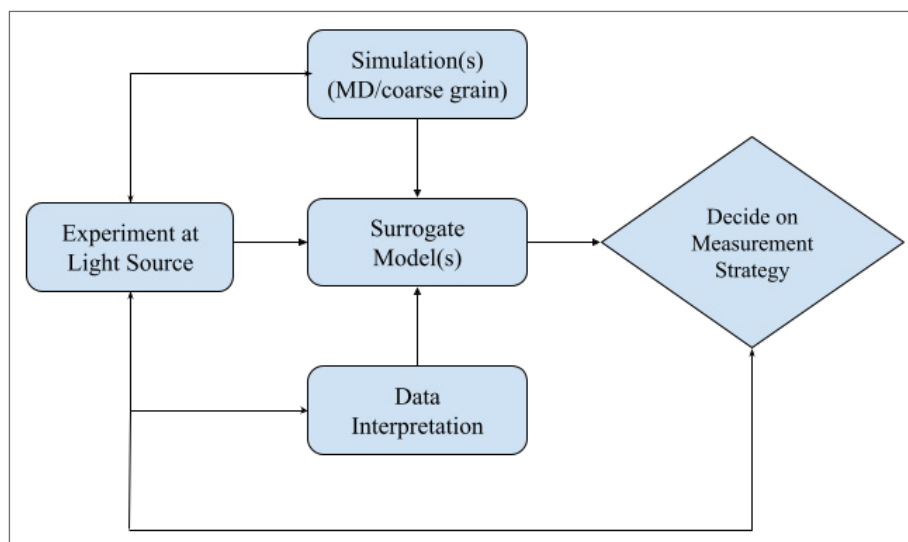
Figure 5.3.4: Flow chart of using a digital twin with light source experiments

This mode of science promises to help accelerate scientific discovery, for example, by enabling faster and more complex materials synthesis. In this scenario a supercomputing resource is used to incorporate experimentally derived data into simulations, the results of which are used to guide experiments in real-time and to even plan experiment campaigns and sample compositions.

Coupling simulations experiment steering will become an increasingly relied upon capability at the light sources as the decade progresses. Due to the high computational cost associated with data processing, reduction, and analysis, and of model training on such large data sets, the light sources will need to employ the use of geographically distributed computing facilities and data repositories. This necessitates high bandwidth, low latency network connections between the light sources and remote computing and data storage centers.

Unified solutions across the light sources are required in order to leverage efficiencies of scale, and to provide facility users with the ability to easily and transparently manipulate data across the light sources. A shared computational fabric for the complex should be developed that connects light source instruments (and other scientific user facilities) to a multitiered, distributed computing landscape, including edge, local, campus, and supercomputing centers, data repositories and archives, and facility user institutions in a seamless and transparent manner. This necessitates the development of advanced networking capabilities and increased networking bandwidth, sustainable and discoverable data repositories, on-demand real-time supercomputing access, and workflow and orchestration tools. Also, as the majority of light source users are not experts in using computational facilities, there needs to be easy access and both in terms of obtaining the resources and performing the calculations.

Note: In these cases, users may access the NSLS-II and computing facilities remotely.

NSLS-II has been working closely with the other four BES light source facilities, coordinated through the 5-way LSDCSC. This group produced a unified vision for the distributed data infrastructure to enable user science, the DISCUS, and a decade long roadmap to achieve the vision (see **Figure 5.1.12**). This vision proposes a transformative computational fabric that covers the full lifecycle of data generated at the BES light sources to accelerate discovery and insight. This vision proposed connecting the 200+ instruments at the light sources to a multitiered computing landscape, including edge, local, campus, and ASCR compute resources, and discoverable data repositories using high-performance, robust feature-rich networks. This fabric would facilitate the full lifecycle of data across the complex, including theory/modeling and simulation, experiment design, data generation at scientific instruments within the light sources, data reduction and processing, analysis and interpretation, and publication and dissemination, serving the 10,000+ light sources users per year.

5.3.2.5 Remote Science Activities

There has been a trend to increasing remote access to synchrotrons in recent years, thanks to the ongoing advances in robotic sample handling, and in automated data acquisition and data management. The current COVID-19 restrictions have accelerated such demands. Furthermore, increasingly brighter synchrotron sources and faster detectors have dramatically shortened the data collection times, to the point that data collection is the least time-consuming activity in an increasing number of experiments.

Working with other BES light source facilities, we have formed a remote experiments task force to work on implementing remote access and tele-experiments at NSLS-II, and to look at access policies, procedures, and user experience. We see remote experiments as essential for NSLS-II to maintain leadership and attract the best science in an increasingly connected world.

5.3.2.6 Software Infrastructure

The NSLS-II accelerator systems and beamline instruments rely primarily on the EPICS for low-level device control. Bluesky[7] is used for high-level experiment control and orchestration. Remote access tools, such as Apache Guacamole, will continue to be utilized to enable remote access for NSLS-II experiments.

NSLS-II users can download data at their home institutions by using Globus (paid subscription) or sftp. Users often utilize their own "cloud storage" accounts (e.g., Google Drive, Dropbox, etc.…) to transfer data.

The initial data management layer for experimental data is the Bluesky databroker/tiled software[8] which is tightly integrated into the data acquisition system. This is developed as a collaborative open-source project with contributors from other facilities, both within the DOE complex and internationally. This system aims to provide a consistent "data API" rather than prescribe a given on-disk data format.

These tools will serve as the basis for enabling searchable data catalogs and adopting FAIR data practices. The Materials Data Facility (MDF) and the DOE Office of Scientific and Technical Information will serve as a DOI generating service for NSLS-II data sets. In the long term, the NSLS-II will adopt or develop tools that provide metadata catalogs, electronic logbooks, and sample tracking systems.

There is a high demand from the user community to integrate cloud services for both file transfer (e.g. Dropbox, Google Drive, etc.) and for communication (e.g. Slack) into all areas of the data lifecycle and compute workflows. We have seen this demand increase during the COVID-19 pandemic.

Coherence, imaging, high-energy, and multimodal techniques are generally the most computationally intensive techniques performed at the NSLS-II. Data reduction and analysis will rely heavily on the use of HPC, utilizing appropriate technologies such as multithreading, General Purpose GPUs, edge devices, and distributed computing environments to obtain results with near real-time completion, so that results enable user-driven or even automated steering of experiments.

5.3.2.7 Network and Data Architecture

BNL High-Throughput Science Network

BNL has implemented a vendor agnostic, resilient, scalable and modular terabit per second (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.

---

7 Allan, D., Caswell, T., Campbell, S., Rakitin, M. (2019) Bluesky's ahead: a multi-facility collaboration for an a la carte software project for data acquisition and management. Synch. Radiat. News 32(3), 19–22.
8 Tiled website — https://blueskyproject.io/tiled/

The HTSN has five primary components:

- **Network perimeter**
    - Three diverse 100 Gbps circuits that peer with ESnet. 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 15–20 PB of data monthly.
- **Science DMZ**
    - Supports open, high-speed WAN/Internet) access for all scientific collaborations throughout the BNL campus.
- **Science Core**
    - A Tbps Data Center Interconnect (DCI) for data intensive collaborations at BNL. This network interconnect enables high-speed connectivity between collaborations such as ATLAS, STAR, PHENIX, CAD, CFN, NSLS-II, HPC Clusters and the Scientific Data and Computing Center (SDCC).
    - Intelligence and routing policies are applied within the Science Core to restrict or grant access to specific resources within the SDCC.
- **Spine**
    - 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.
    - The responsibility of the Spine is fast packet forwarding and flexibility, not policy insertion or server termination.
    - External Border Gateway Protocol (eBGP) is utilized throughout the HTSN. eBGP was chosen for its ability to immensely scale and to create modularity and fault domain isolation down to the rack level. Each Spine group shares the same Autonomous System Number (ASN) but does not have Internal BGP (iBGP) peering's between them. Each Leaf or pairs of Leaves will require their own ASN.
- **Storage Core**
    - A redundant terabit per second switching block that aggregates high-performance storage services

BNL Next-Generation Network Perimeter

The network perimeter at BNL is a high-speed and fault-tolerant network infrastructure that provides the BNL site connectivity to the Internet and various scientific wide-area networks. It supports numerous data intensive collaborations such as BES, Biological and Environmental Research, HEP, and Nuclear Physics. It also supports critical campus services such as workstations, phone service, security, safety and monitoring and enterprise and cloud computing. Since being placed into production in 2013, the network perimeter has transmitted over 100+ PBs of data per year to numerous scientific collaborations worldwide.

In 2013 the BNL network perimeter was bleeding edge 100 GbE technology. Now, the hardware has reached 8+ years in age and it is no longer cost-effective to purchase additional hardware for these platforms. Newer platforms today support much greater 100 GbE interface densities along with supporting 400 GbE which will allow BNL to support all its data intensive collaborations well into the future. With these factors in mind, BNL will possibly procure a next-generation network perimeter within the next one to two years. This will prepare the laboratory to meet the missions needs of all the data intensive collaborations at BNL.

5.3.2.8 Cloud Services

Currently, we make limited use of Amazon AWS and Microsoft Azure cloud platforms in the areas of prototyping, testing and standing up computational/storage resources to support training and demonstrations. There are a number of Software as a Services in production use, including Slack, Dropbox, GitHub Enterprise Cloud and Office 365.

The NSLS-II, along with the other US DOE-funded light sources, continues to evaluate the capabilities and cost-effectiveness of using cloud services.

5.3.2.9 Data-Related Resource Constraints

Data management and workflow tools are needed that integrate beamline instruments with computing and storage resources, for use during experiment, as well as facility user access for post-experiment analysis. Real-time data processing capabilities are required to significantly reduce data volumes and provide feedback during experiments to improve data quality and to drive the direction of ongoing measurements; the application of advanced mathematical algorithms, ML, and the integration of simulations and model-based approaches will allow automated steering of data collection. On-demand utilization of computing environments is required to enable near real-time data processing. Sufficient data storage and archival resources to house the continually increasing amounts of valuable scientific data produced by the NSLS-II is required.

5.3.2.10 Outstanding Issues

None to report at this time.

5.3.2.11 Facility Profile Contributors

NSLS-II Representation

- John Hill, BNL, hill@bnl.gov
- Stuart Campbell, BNL, scampbell@bnl.gov
- Paul Cianciulli, BNL, pcianciulli@bnl.gov
- Nicholas Talbot, BNL, ntalbot@bnl.gov
- Nicholas Schwarz, ANL, nschwarz@anl.gov
- Jana Thayer, SLAC National Accelerator Laboratory, janajana@slac.stanford.edu
- Alex Hexemer, LBNL, ahexemer@lbl.gov
- Apura Mehta, SLAC National Accelerator Laboratory, jana mehta@slac.stanford.edu
- Vivek Thampy, SLAC National Accelerator Laboratory, janavthampy@slac.stanford.edu

ESCC Representation

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

5.4 Linac Coherent Light Source (LCLS)

The LCLS, located at the SLAC National Accelerator Laboratory was the world's first free electron laser operating in the "Hard X-Ray" regime, funded by the US DOE, Office of Science, Office of BES. Operational since 2009, LCLS has had a dramatic effect on a broad cross-section of scientific fields ranging from atomic and molecular science, ultrafast chemistry and catalysis, quantum materials, structural biology, high energy-density science, and nonlinear photon science. LCLS produces X-rays that are used to image physical and chemical systems with atomic resolution, femtosecond precision, and chemical specificity.

5.4.1 Discussion Summary

- The LCLS, located at the SLAC National Accelerator Laboratory, features10 specialized instruments.

- LCLS has had over 13,000 scientific user visits, and more than 3000 unique users over the course of operation. On average, LCLS has around 1000 users, and 700 unique users, every year.

- Unlike at a synchrotron where X-rays can be delivered to multiple instruments at once, the linear accelerator can only deliver X-rays to one instrument at a time, although it is possible for other experiments to run parasitically using the X-rays that pass through the first instrument. As a result, there are usually 1–2 experiments during a shift.

- Operating experiments is the shared responsibility of beamline scientists, local staff, and on-site users. LCLS has web services, a data management system, and a data analysis infrastructure that allows for virtual access to enable on and off-site analysis. Because of the uniqueness and changeability of LCLS instruments, LCLS does not envision remote operation of the instruments.

- The LCLS-II upgrade represents an advance in X-ray laser technology and will be a transformative tool for energy science. The LCLS-II upgrade is scheduled to come online in 2023.

- LCLS-II will feature a significant increase in the data throughput from today's 1–5 GBps to 200 GBps. Future planned upgrades are expected to increase the throughput to multiple TBps by 2028.

- LCLS uses the same data system at all instruments. The data system provides data acquisition, online data reduction, real-time monitoring, fast feedback for data quality monitoring, data storage and management, data archiving, a local compute facility for offline analysis and an offline analysis framework for data access.

- The LCLS data system handles the transparent data movement within several layers of computing in the pipeline from the Detector Edge through the data reduction compute layer, to the data cache where it is accessible to users for fast feedback analysis.

- Key elements of a future data management strategy for LCLS include a common API for accessing network and computing resources, parallel data transfer tools, high-fidelity data transfer, network performance monitoring, reservations, and dynamic network provisioning.

- Analysis is performed by user teams using available computing resources at SLAC, DOE HPC facilities, or the users' home institutions. Analysis is typically begun while the experiment is running, but can be refined and repeated with different parameters many times after execution. Real-time feedback is critical to drive experiments and ensure experiment success.

- Computing systems are important to facility operation, data interpretation, and overall scientific productivity. There are a number of sources and sinks for LCLS data, depending on the scientific workflow and the needs of the experimenters. The data paths can be summarized as follows:

    — LCLS facility to ASCR compute facility

    — LCLS facility to light source or national lab

    — LCLS facility to university/home institution

    — LCLS facility to cloud computing/storage such as Google, Amazon AWS, or Microsoft Azure.

- LCLS facility to ASCR facility is the most heavily used, and is used:

    — to analyze raw data or do post-experiment analysis

    — archive LCLS data sets (currently between LCLS and NERSC)

    — train/retrain AI/ML models

    — transmit simulation data that can be used during the experiment

- The LCLS to Lab data path is used when multimodal analysis is desired, or to support other remote computational resources.

- The LCLS facility to university/home institution is used for the transfer of data sets post experiment, or for the use of remote computational resources.

- The LCLS facility to cloud computing/storage is typically used by users, and not the facility, when they are attempting to share or back-up data, or perform some varieties of analysis.

- In order to analyze data on experiment timescales, LCLS is sending data to remote HPC facilities such as NERSC, the ALCF, or the OLCF. Sending data from a light source facility to the HPC facility requires a data path of sufficient bandwidth.

- Experiment data and metadata collected at LCLS may be stored at and retrieved from the tape for a period of 10 years. A copy of the data remains at SLAC and another copy is made at a remote site, such as NERSC. LCLS provides space for all experimental data at no cost to the user.

- Local SLAC compute resources are sufficient to provide quasi-real-time analysis and feedback for roughly 80% of LCLS experiments, but for the remaining 20% of experiments, LCLS computational capabilities are not sufficient. LCLS intends to stream data to a DOE HPC (NERSC, OLCF, ALCF) for analysis.

- During LCLS-II operation, it is expected that local computing capabilities will still account for 80% of experiments and will require processing resources on the order of 5 PFLOPS and storage capacities above 100 PB by 2027. LCLS-II will continue to use DOE HPC facilities for the other 20% of experiments with the largest computing demands, and will heavily leverage ESnet network resources to facilitate data mobility.

- Most users complete their offline analysis using SLAC-provided computing resources in the 4 months following their experiment. On average, a typical user will rerun over their entire data set up to 10 times.

- LCLS produces data sets between 1–10 TB in size. LCLS-II and LCLS-II-HE has an expectation of generating data sets of about 1 PB per shift or several PBs per experiment.

- Approximately a tenth of users copy data to their home institutions, usually using Globus.

- The further development and application of autonomous experiment steering requires that a number of gaps in capabilities and infrastructure be addressed:
  - sufficient and reliable bandwidth.
  - sufficient and sustainable data storage resources.
  - on-demand access to large-scale computing systems.
  - transparent access to facilities and systems through federated identity and shared data protocols.
  - software tools and infrastructure to facilitate the development of scientific data workflows.
- Networking must be able to robustly connect LCLS instruments to remote computing facilities with low latency. Terabit per second networking and beyond will be required to handle the large amounts of data expected towards the end of the decade. There is a need to upgrade the SLAC/LCLS ESnet connections to keep pace with the expected data rates of 200 Gbps in 2023 and 1 Tbps by 2028 enabling streaming data transfer from LCLS to ASCR computing facilities.
- Autonomous experiment steering will be crucial to experiment success, and requires the use of geographically distributed computing facilities and data repositories, which in turn necessitates high bandwidth, low latency network connections.

5.4.2 LCLS Facility Profile

LCLS provides ultrashort pulses (from 200 attoseconds to 200 femtoseconds), with peak brightness a billion times higher than synchrotron sources, over an energy range from 250 to 25,000 eV, at 120 pulses per second. LCLS is currently undergoing major upgrades, LCLS-II and LCLS-II-HE, to provide a million pulses per second, delivering orders of magnitude enhancement of the average power and coherent flux compared to any other source. LCLS is one of five national user light source facilities available to users from academia, industry, government agencies, and research institutions worldwide.

5.4.2.1 Science Background

The overall layout of the LCLS instruments is shown in **Figure 5.4.1**. LCLS features 10 specialized instruments, each with a dedicated team of scientists and support staff to conduct pioneering research and provide the technical expertise required to operate the cutting-edge instrumentation. In order to detect the interaction of photons with matter, each instrument is surrounded by a variety of detectors and diagnostics such as ion time-of-flight chambers, gas bottles, diodes, and area detectors (cameras). This instrumentation captures precisely time-stamped information about the beam, such as energy and profile, the environment, the sample, and the location of the photons that scattered off the sample. Detector data for each independent X-ray shot is combined with information about the sample environment and other diagnostics to infer the properties of the system under study. In order to make sense of this information, science-domain-specific techniques are used to extract the physically meaningful information about the structure and dynamics of the target under study. The analysis is typically performed by the user team using available computing resources at SLAC, ASCR High End Computing (HEC) facilities, or the users' home institutions. The complex analysis that is used to answer a specific scientific question is typically begun while the experiment is running but the analysis is refined and repeated with different parameters many times after the experiment has concluded. Due to the dynamic nature of these experiments, real-time feedback is critical to drive experiments and ensure experiment success.

The LCLS-II upgrade represents an advance in X-ray laser technology and will be a transformative tool for energy science. It will qualitatively change the way in which X-ray scattering, spectroscopy and imaging will be used in the future, to observe in ways never before possible, how natural and artificial systems function, spanning multiple decades of time scales (down to the attosecond regime) and multiple spatial scales (down to the atomic regime). LCLS-II will further enable powerful new ways to capture rare chemical events, characterize

fluctuating heterogeneous complexes, and reveal underlying quantum phenomena in matter using nonlinear, multidimensional, and coherent X-ray techniques that are only possible with a true X-ray laser.
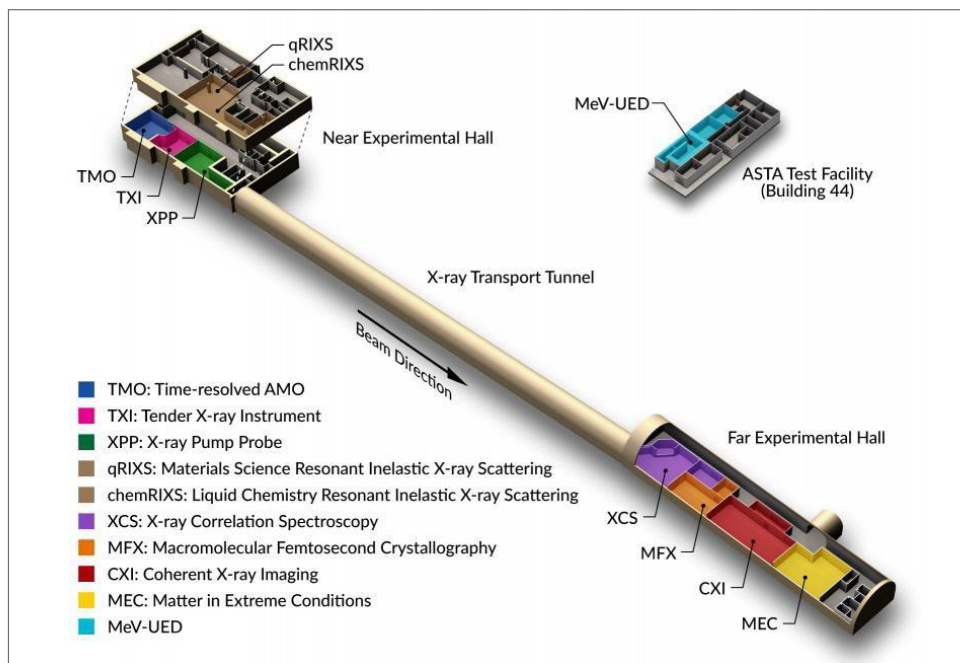
Figure 5.4.1: LCLS features 10 specialized instruments, each with a dedicated team of scientists and support staff, to conduct pioneering research and assist users with experiments. Each hutch is equipped with a suite of instruments to assist in gathering a wide range of data using various specialized techniques, from telltale signatures of electrons and ions to the intricate patterns left by crystallized samples struck by the X-ray laser.

5.4.2.2 Collaborators

LCLS is a scientific user facility with a large and diverse user community. With over 13,000 scientific user visits in its first 10 years of operation, researchers from around the world have conducted groundbreaking experiments in a variety of fields. LCLS has 3000+ unique users and has conducted 900+ experiments on its 10 beamlines. On average, LCLS has around 1000 users and 700 unique users every year.

Unlike at a synchrotron where X-rays can be delivered to multiple instruments at once, the linear accelerator can only deliver X-rays to one instrument at a time, although it is possible for other experiments to run parasitically using the X-rays that pass through the first instrument. As a result, there are usually 1–2 experiments during the day shift and 1–2 experiments on the night shift, with the maximum possible number of experiments running at any given time to maximize the use of the facility by users. Multiplexing, in which multiple instruments run simultaneously, is nevertheless common and at least 2–4 experiments at different instruments may be operating in any given week. Running each experiment is the shared responsibility of beamline scientists and users. LCLS provides data management tools, such as a file manager and web services to configure and launch analysis workflows, and a data analysis infrastructure that allows for virtual access to all of the resources necessary to analyze the data. In most cases, all the same resources are available to users' doing analysis at home as on-site at SLAC.
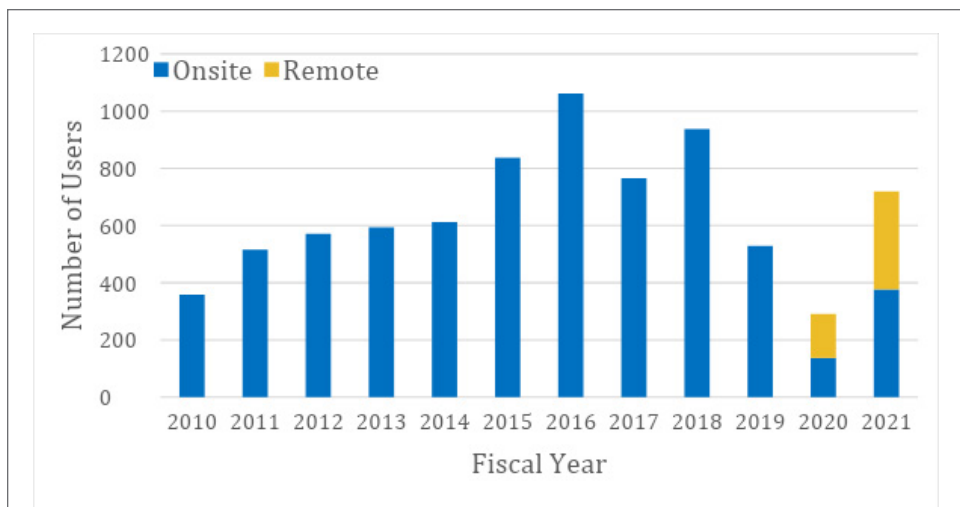
Figure 5.4.2: Total number of LCLS users per fiscal year. During the COVID pandemic from 2020-2021, many users were remote and LCLS staff together with a limited number of on-site users managed the experiments.

LCLS users perform experiments across a broad cross-section of scientific fields ranging from atomic and molecular science, ultrafast chemistry and catalysis, quantum materials, structural biology, high energy-density science, and nonlinear photon science.
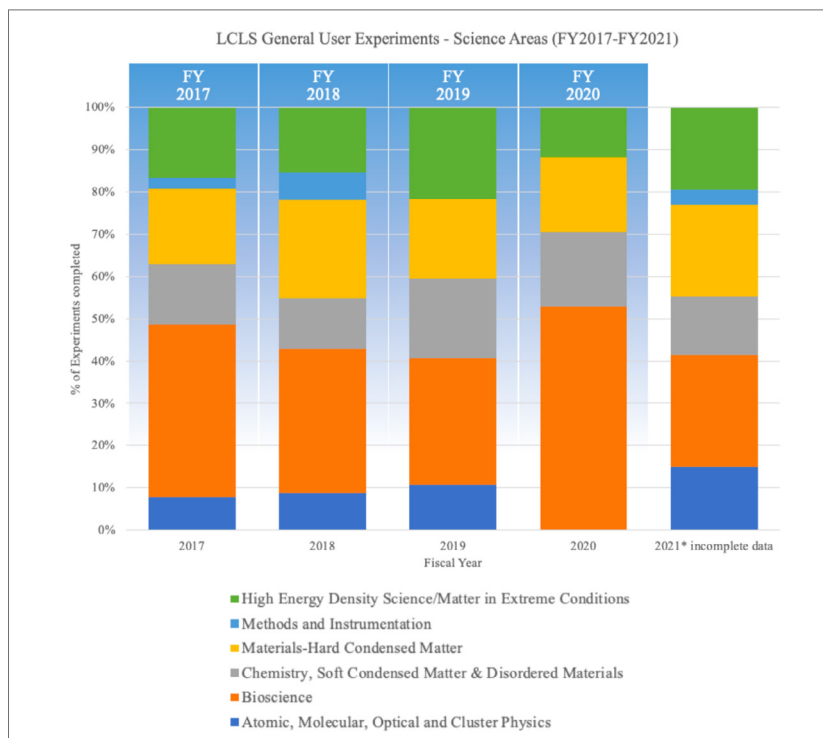
Figure 5.4.3: LCLS Users and Science Areas by Fiscal Year, including MeV-UED (noting that the X-ray facility ran for less than half of FY17, only 10 weeks in FY19, and only 6 weeks in FY20).

LCLS users come from academia, industry, government agencies, and research institutions worldwide.

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
LCLS FACILITY TO ASCR FACILITY — STREAMING OR POST-EXPERIMENT ANALYSIS	Primary	Automated data transfer using bbcp, xrootd. Other: LCLS facility handles data transfer from LCLS POSIX file system to remote facility	1 PB; range from 1 TB to tens of PB. Maximum rate is of order 100 GBps	Every 2 months (depends on LCLS experiment schedule) intense data transfer to ASCR facility for a period of ~1 week	Every 2 months for 1 week at a rate of < 1 GBps	For streaming data, need low latency and high bandwidth during the experiment; reservation or guarantee of bandwidth end-to-end might be necessary
LCLS FACILITY TO ASCR FACILITY – LCLS DATA ARCHIVAL	Secondary	Automated data transfer using rucio, xrootd, high performance storage system (HPSS)	1 PB; range from 1 TB to tens of PB. In aggregate, we expect of order 300 PB/year	Continuous	Data restore from tape archive (rare, because this is a second copy of the data)	N/A
LCLS FACILITY TO ASCR FACILITY – (RE)TRAINING AI/ML MODELS	Primary	Automated data transfer using bbcp, xrootd. Other: LCLS facility handles data transfer from LCLS, POSIX file system to remote facility	1 TB; range from 1 GB to tens of TB. Maximum rate is of order 1–10 GBps	Continuous	Yes, at a rate of < 1 GBps	Need low latency during experiments
LCLS FACILITY TO OTHER LIGHT SOURCE/LAB/ UNIVERSITY — MULTIMODAL ANALYSIS, POST-EXPERIMENT ANALYSIS, ARCHIVAL	Secondary	Globus, scp; these transfers are done by the user and we do not monitor the traffic or how the transfers are accomplished.	10% or fewer experiments may copy data to a remote site post-experiment; as data sets grow, it becomes less likely that users will copy data to home institutions. HDF5 files are most likely to be copied. These files are likely to range from 1 TB to tens of PB. Most likely value is ~ 1 PB per experiment	Ad hoc	No	N/A

Table 5.4.1: LCLS Collaboration Space
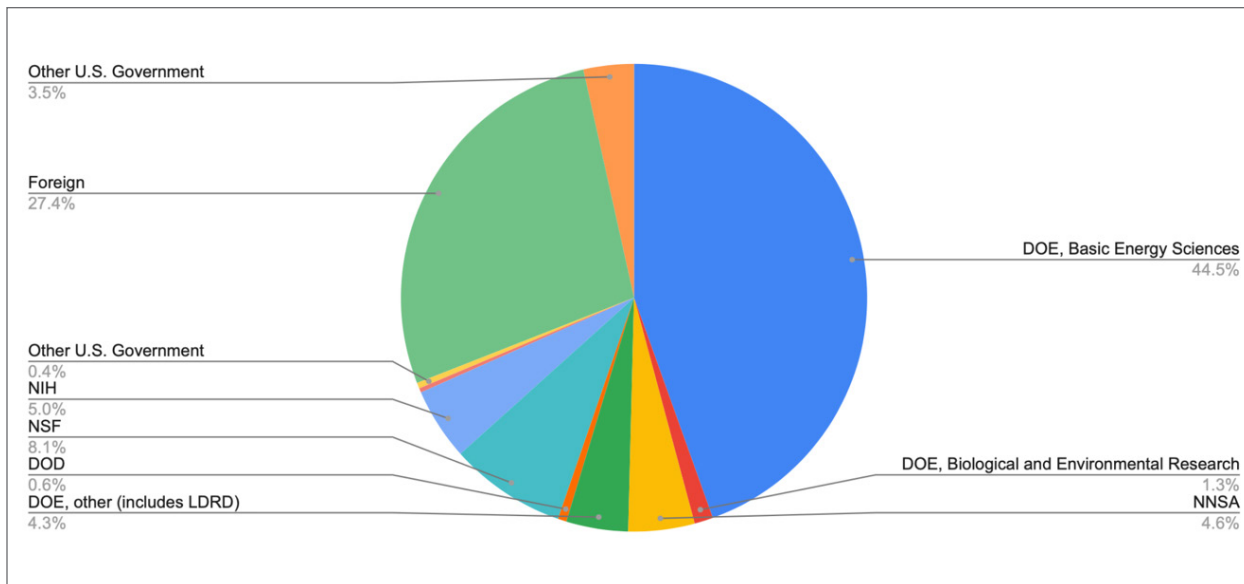
Figure 5.4.4: Percentage of LCLS users by source of support for FY2021

State	Number of Users	State	Number of Users
Arizona	44	Nebraska	9
California	297	Nevada	3
Colorado	2	New Jersey	7
Connecticut	5	New Mexico	17
Illinois	24	New York	33
Indiana	2	Ohio	5
Kansas	10	Pennsylvania	5
Louisiana	1	Rhode Island	9
Maryland	3	Texas	3
Massachusetts	12	Utah	2
Michigan	8	Washington	5
Minnesota	3	Wisconsin	6
Missouri	4	District of Columbia	3

Table 5.4.2: LCLS users by state for FY2021

Advanced computing systems are playing an increasingly important role in facility operation, data interpretation, and overall scientific productivity. There are a number of sources and sinks for LCLS data, depending on the scientific workflow and the needs of the experimenters. Most LCLS data is generated by the LCLS facility during experiments. Some LCLS data products are generated by users during the analysis of the LCLS data sets acquired during an experiment. This data may be generated locally at SLAC or it may be generated at a remote compute resource, such as an ASCR facility. Users may also work with theorists to generate simulations that are relevant to an experiment; this data may be generated at the ASCR compute facilities or at a university or the users' home institutions. Users may also compare data taken at LCLS to data taken at one of the other light sources. During an experiment or during the post-experiment analysis phase, users may generate more results data and they may need to compare the simulated results to experimental results from any of the light sources. The data paths can be summarized as follows: 1) LCLS facility to ASCR compute facility, 2) LCLS facility to light source or national lab, 3) LCLS facility to university/home institution, and 4) LCLS facility to cloud computing/ storage such as Google, Amazon AWS, or Microsoft Azure.

The first, LCLS facility to ASCR facility is the most heavily used. It is used to do streaming analysis using ASCR computing to analyze raw data or do post-experiment analysis. A primary copy of the data set is always kept at LCLS and a secondary copy is kept at NERSC. The LCLS facility to ASCR facility workflow is used to archive LCLS data sets at NERSC and to train/retrain AI/ML models that are being used by software at a beamline during experiments or to analyze data post experiment. An emerging use case is the use of simulations, performed at the ASCR facility, and then used during the experiment to inform data analysis.

The second data path, lab to lab, is used when multimodal analysis is desired, or if a lab/university has a specialized local computing resource that is used for analysis.

The third data path, LCLS facility to university/home institution is primarily used for the slow transfer of data sets post experiment to a users' local storage/computing resources.

Finally, the LCLS facility to cloud computing/storage is not used by the facility but may be employed by some users. The LCLS facility has investigated the possibility of using cloud resources but found that the costs associated with I/O were unsustainably high. However, LCLS does not prevent users from making use of these resources if they wish, and if multimodal analysis is desired, cloud storage and computing becomes a more attractive option. In some cases, such as streaming real-time analysis during an experiment, there may be strong real-time constraints on the flow of data between the facility and a computing resource which may limit the use of commercial cloud resources.

5.4.2.3 Instruments and Facilities

LCLS serves over 3000 users with 10 instruments and uses the same data system at all instruments, a key distinction between LCLS and the other light sources. The data system provides data acquisition, online data reduction, real-time monitoring, fast feedback for data quality monitoring, data storage and management, data archiving, a local compute facility for offline analysis and an offline analysis framework for data access. The facility operates 24/7 with brief shutdowns for maintenance about twice a year.

Day in the Life of an LCLS User — The Data Lifecycle

The majority of users gain access to BES user facilities via a merit-based peer-review process in which scientific proposals for access to facility's cutting-edge capabilities are evaluated by an independent proposal review panel. Calls for these proposals typically take place twice a year. Top-ranking proposals are awarded beam time, and the LCLS facility begins discussions with users about their experiment's needs as soon as beam time is awarded. Users may return to the facility with each new run or they may not return to the facility for years at a time. The LCLS analysis framework and data management framework must be capable of retrieving old data and enable users to access and analyze older data together with newer data.

Individual users' experiments are typically allocated ~5 shifts of 12 hours each. Users may begin generating simulations, growing samples, or building instrumentation several months in advance of the beam time. Experiments are either on a day shift, 9 AM–9 PM or a night shift, 9 PM–9 AM and multiple experiments may run simultaneously in any given week. Experiment configurations at beamlines are not static and change weekly when the previous user group rotates out and the next user group rotates in. Configurations at beamlines may also change during an experiment in response to an experiment's needs: detectors and diagnostics may be added or removed, the samples under study are replenished, renewed, or replaced with different samples, and other aspects of the instrument such as laser timing, mirror alignment, etc. may be adjusted for optimal performance. Some experiments require coordinated excitation of the sample followed by data-taking. For example, a laser pulse may excite a material which is followed up by a precisely timed series of X-ray pulses to image the system at different time intervals following the stimulus. In this case, the data system coordinates this sequence of events and collects metadata that is stored alongside the instrument data to describe the conditions under which the experiment was done.

Prior to an experiment, users may generate simulations in collaboration with theorists. The resulting data is privately managed by the users. These simulations are used to inform the parameters of an experiment and may be compared to experimental results during an experiment. Users arrive on-site at the facility several days prior to their beamtime to assess the beamline, set up specialized equipment for the experiment, or test analysis workflows, a task which may require some access to computing resources. The number of collaborators associated with an experiment ranges from 5 to 80, some of which are on-site and driving the experiment during the beam time while the others are analyzing data from their home institutions and participating with on-site colleagues via telecommunication software.

During the experiment, beamline scientists work closely with experimenters to optimize the data acquisition (calibrate detectors, detector to IP distances, sample delivery, hit rate) and optimize optics/laser timing. Once those are setup, the interactive process of data acquisition begins. Data is acquired for about 10 minutes at a time, and analyses are run "live" during acquisition with the goal of answering a scientific question within a few minutes of stopping acquisition. This feedback allows experimenters to decide how best to steer the experiment. Data quality is also assessed as part of this process: is the calibration/timing/alignment good? Is the data reduction performing optimally? Is the signal to noise as expected? In addition to data collection for scientific inquiry, data is collected to calibrate the detectors, align the beamline, adjust laser timing, and tune sample delivery. By the last few days of beamtime, most of the beamline parameters have been optimized to extract the most science information. As a result, the average rates of useful data written to disk usually increase during the last days of an experiment.

Real-time data analysis is desired in order to optimize the use of limited beamtime and sample. Users need to analyze data at the rate of acquisition in order to obtain analysis results within minutes of ending a period of data acquisition. Local SLAC compute resources are sufficient to provide quasi-real-time analysis and feedback for roughly 80% of LCLS experiments, but for the remaining 20% of experiments, LCLS computational capabilities are not sufficient. LCLS intends to stream data to a DOE Leadership Class Facility (NERSC, OLCF, ALCF) for analysis. After data is copied to the spindle-based offline storage, users may copy data to their home institution for analysis. Approximately a tenth of users copy data to their home institutions, usually using Globus. Most users complete their offline analysis using SLAC-provided computing resources in the 4 months following their experiment. On average, a typical user will rerun over their entire data set up to 10 times. In some cases, users will also reanalyze data taken during a previous beamtime (up to 10 years ago) or combine results with data taken at other light sources to include (but not limited to) ALS, APS, NSLS-II, SSRL, SACLA, PAL, EuXFEL.

Requirements — Throughput, Storage, Computing

The LCLS-II upgrades, scheduled to come online in 2023, will increase the repetition rate from 120 Hz to 1 MHz Coupled with the adoption of ultrahigh repetition rate imaging detectors, there will be a significant increase in the data throughput from today's 1–5 GBps to 200 GBps in 2023. Future planned upgrades are expected to

increase the throughput to multiple TBps by 2028. The throughput will be handled internally by the LCLS facility, but the reduced data that is written to disk, represented by the blue line in **Figure 5.4.5** below, will be archived and potentially copied or moved to other facilities for analysis.
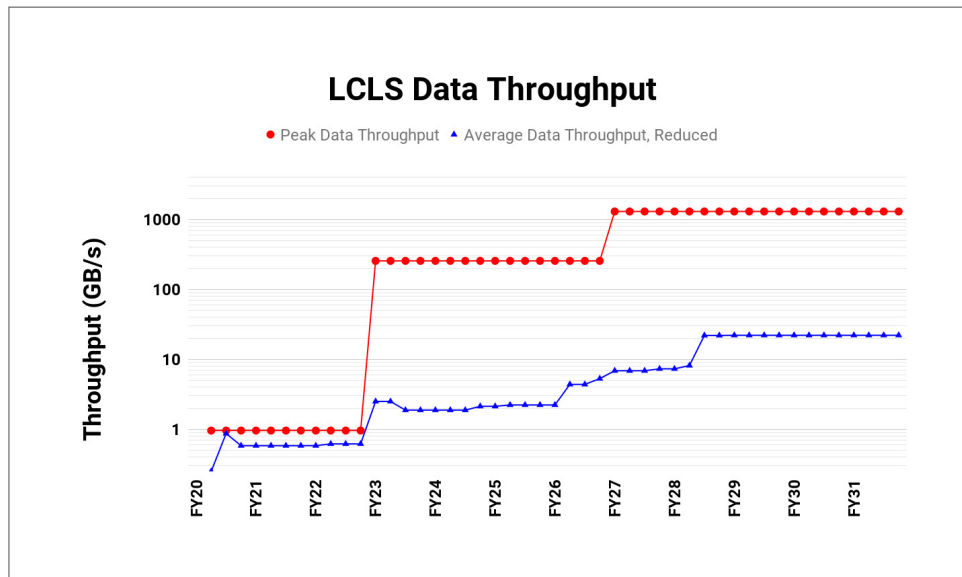
Figure 5.4.5: Peak data throughput (red), the amount of science data produced by the sensors of an instrument every second. LCLS-I throughput, with a repetition rate of 120 Hz, is about 1–5 GBps. LCLS-II throughput, operating at a repetition rate of 1 MHz, beginning in 2023 will have a peak throughput of 200 GBps. In 2027, when LCLS-II-HE instruments come online at 1 MHz, the throughput is expected to be ~1 TBps. When the effects of the data reduction pipeline and the duty cycle of the experiment are taken into account, the average data throughput (blue) is about 20 GBps.

In aggregate, data sets produce of order 1–10 TB of data per experiment in the 120 Hz mode of LCLS. With LCLS-II and LCLS-II-HE running at 1 MHz, the expectation is that a typical experiment will generate about 1 PB per shift or several PBs per experiment, already reduced prior to being written to disk via the real-time data reduction pipeline.
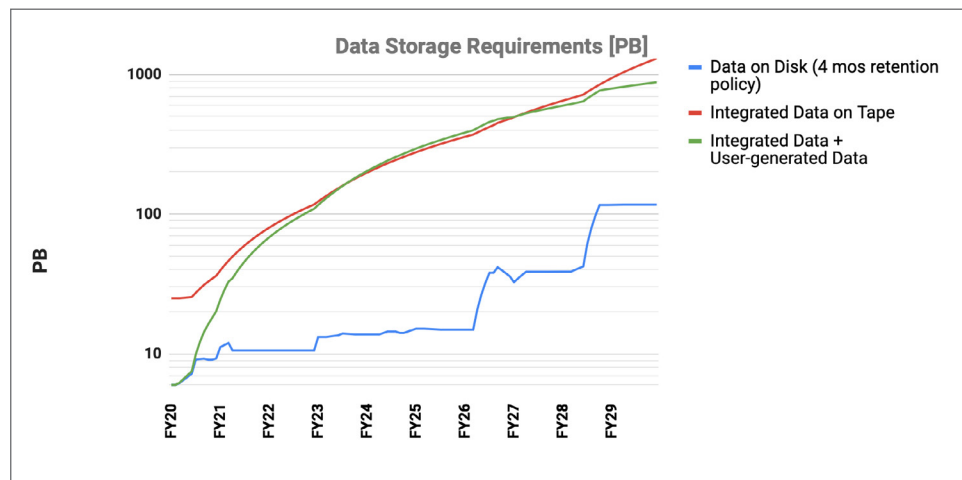
Figure 5.4.6: This figure shows the integrated data as a function of time in PB assuming a data retention policy of 4 months on disk and 10 years on tape. The blue curve shows local disk storage requirements as a function of time, requiring 100 PB of disk storage by the end of the decade. The red curve shows tape storage requirements as a function of time.

By 2028, data aggregated at rates of multiple terabits per second (100–300 Gbps) within three years and exceeding 1 Tbps in five years may flow via ESnet from any of the light sources (ALS, APS, LCLS, NSLS-II, SSRL) to any of the ASCR compute facilities (ALCF, NERSC, OLCF).
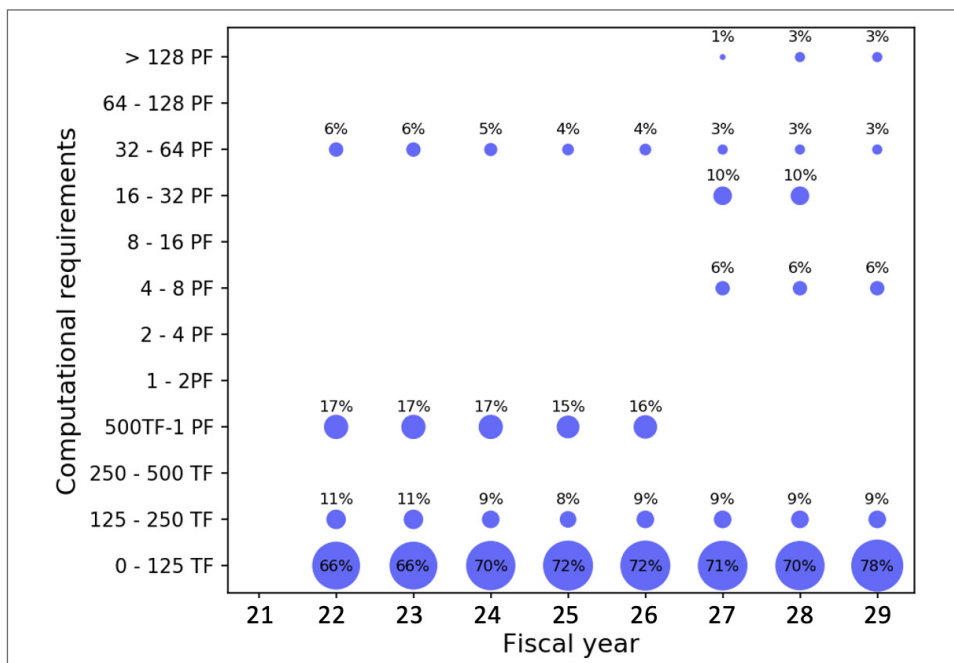
Figure 5.4.7: Computational requirements for all experiments as a function of time. In 2023, 66% of experiments will require between 0 and 125 TFLOPs of computing. When LCLS-II turns on, most experiments will still require < 1 PFLOPs of computing, but a larger fraction will require of order 5 PFLOPs. Experiments requiring < 5 PFLOPs will be analyzed locally. Any experiment requiring > 5 PFLOPs will be streamed to HEC for remote processing.

LCLS-II Data System

Computing demands at LCLS are driven by the repetition rate of the source, advances in detector technology, advances in data analysis algorithms, and the requirement to provide flexible and easy-to-use fast feedback to users in real time, a challenge given the weekly turnaround of experiments and the number of new user groups. The LCLS-II data system, the main components of which are described in **Figure 5.4.8** below, provides core hardware and software infrastructure for scalable data acquisition, online monitoring, offline analysis, and data management enabling scientists to efficiently go from measurement to scientific insight. The LCLS-II data system has some challenging characteristics that may impact networking requirements:

- Ability to provide fast feedback to the experimenters on the timescale of seconds and, for more complex analyses, minutes — this is required in order to reduce the time to complete the experiment, improve the quality of the data and increase the experiment success rate.

- 24/7 availability.

- The analysis pattern is characterized by bursts of short jobs, requiring very short startup time.

- Storage represents a significant fraction of the overall system, both in terms of cost and complexity.

- Throughput between the front-end electronics and the storage layers and between storage and processing is a critical element of the system.

- Speed and flexibility of the development cycle are critical due to the wide variety of experiments, the rapid turnaround required, and the need to modify data analysis during experiments.

The LCLS facility handles the challenges of throughput increases, larger data sets, complex, computationally demanding analyses and shot-by-shot analysis requirements. Users then provide the last mile and develop their analysis on top of this stack.
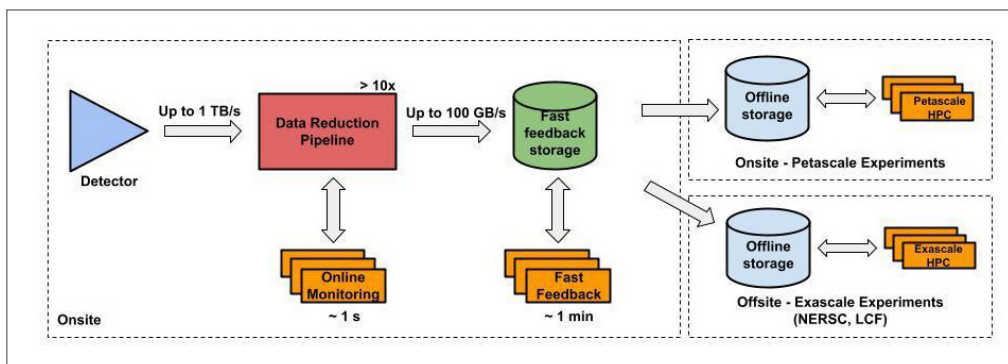
Figure 5.4.8: Main components and data flow of the LCLS Data System. For each of the LCLS instruments, there is a dedicated infrastructure for reading out detectors and a shared infrastructure for data reduction and fast feedback. The data center is also a shared resource and may be supplied by local resources at SLAC or by remote HPC facilities.

The LCLS data system handles the transparent data movement within several layers of computing in the pipeline from the Detector Edge (FPGA/ASIC) through the data reduction compute layer (CPU, FPGA, GPU), to the data cache (NVMe-based disk storage) where it is accessible to users for fast feedback analysis for approximately 1–5 shifts. From the data cache, the first persistent storage layer, data is automatically transferred by the data system to offline storage using bbcp and xrootd, where the data will remain on disk for 4 months post experiment. Data is also automatically archived to tape where data remains for 10 years. A second copy of the data is held on tape at NERSC.

The types of data that are recorded vary from instrument to instrument and from experiment to experiment and include large imaging detectors, digitizers that produce waveforms, and point values such as temperatures, pressures, motor positions, etc. The dominant contribution comes from the large imaging detectors. Data is stored in the LCLS data format: xtc for LCLS-I and xtc2 for LCLS-II. This data format has the merit of being the same on the wire as it is on disk, and in the case of LCLS-II is self-describing, which means there are no additional encode/decode steps as the data pass through the computing layers of the data system. LCLS-II must accommodate variable length data that can be reduced in one of approximately 10 ways (lossless compression, binning, SZ compression, region of interest, peak finding, projections onto an axis, calculation/feature extraction, MPEG-style compression, and integration).

Data Retention Policies

LCLS is committed to providing its users with their data in a timely and convenient fashion. Experiment data and metadata collected at LCLS may be stored at and retrieved from the tape for a period of 10 years. A copy of the data remains at SLAC and another copy is made at a remote site, such as NERSC. LCLS provides space for all experimental data at no cost to the user. This includes the data from the detectors as well as the data derived from user analysis. Raw data is available as xtc or xtc2 files. Users may translate the data from xtc/xtc2 to HDF5 format using a utility provided by the LCLS facility. The translation is not performed automatically, but is done on demand at the user's request. After 4 months, data is purged from disk but is still available on tape and may be retrieved by the user on demand for a period of 1 month. The time to restore files varies from a few tens of

minutes to a day or longer as restore requests are run with lower priority than data transfer and archival for operating experiments.

Compute Resources

The computing system for LCLS-II-HE will require both access to the DOE HEC facilities for the highest demand experiments in the floating point operations per second (FLOPS) exascale and dedicated, local capabilities for storage and analysis for standard experiments. Local capabilities, used for 80% of experiments, will require processing resources on the order of 5 PFLOPS and storage capacities above 100 PB by 2027 (these numbers assume the current data retention policy and the expectation that we will reduce the data volumes on the fly by at least a factor of 10).
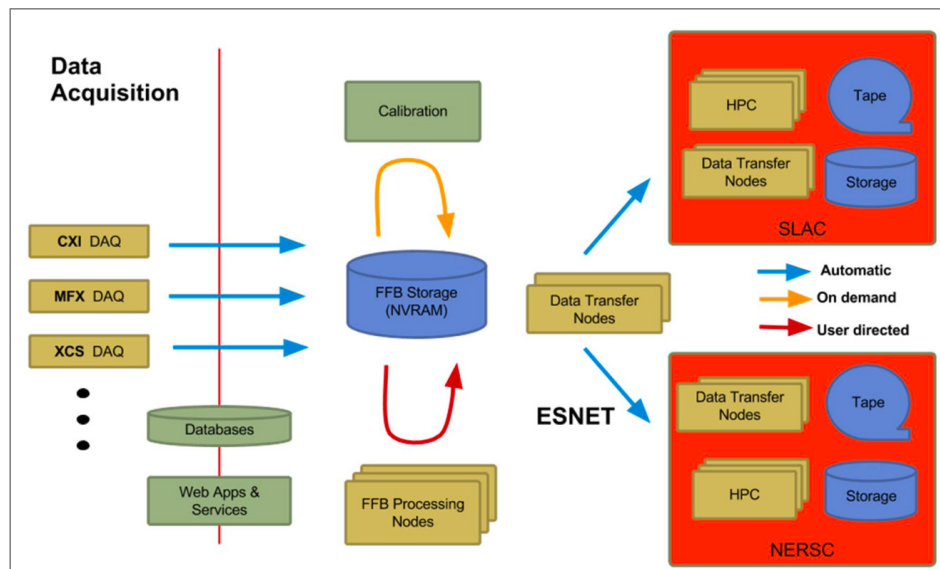
Figure 5.4.9 The LCLS-II data system automatically transfers data from the DAQ to the fast feedback storage layer (data cache) and to local or remote sites. Access to remote sites, such as NERSC, is via ESnet. Routine data taking and archival activities are automatic, but certain activities such as restoration of data from tape to disk and analysis workflows may be user directed.

HEC facilities are used for 20% of experiments with the largest computing demands. Access to remote sites is via ESnet. Local resources are preferred for those experiments that do not require a supercomputer due to the challenge of coordinating planned outages with HEC facilities across the entire year and risks related to HEC unplanned outages and reduce some of the complexities associated with running on a supercomputer (resource orchestration, user accounts and privileges, priority management, container requirements, etc.).
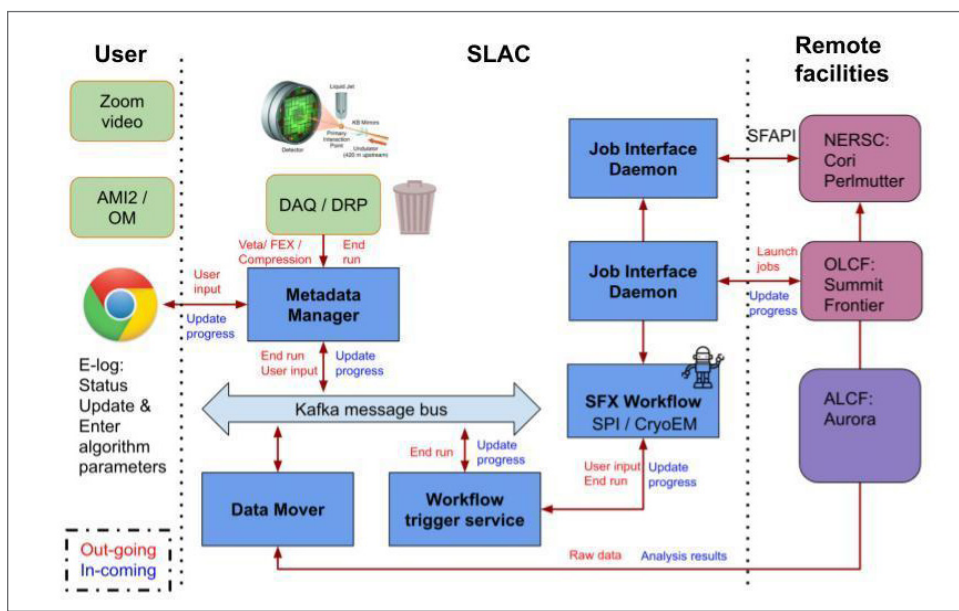
Figure 5.4.10: The LCLS-II data systems provides the capability to trigger automatic workflows at remote sites. Communication between the LCLS-II data system and the remote facility proceeds along two paths, a data path and a control path. Reservations are made at the remote facility in advance of the experiment. Data is transferred to the remote facility over ESnet and results back to SLAC; the data transfers happen automatically once the reservations are in place and connections enabled. There is a control path for triggering workflows and monitoring job status.

Use Case: Remote Quasi-real-time Analysis of Large Streaming Data Sets

X-ray scattering experiments are a powerful tool to determine the molecular structure and function of unknown samples, such as COVID-19 viral proteins. In crystallography experiments, molecular structure is determined by merging the X-ray diffraction patterns from millions to billions of protein crystals exposed in random orientations. The near real-time interpretation of structure revealed by X-ray diffraction requires significant computational resources. Similarly, techniques such as ptychography, X-ray tomography, XPCS, and X-ray fluorescence generate data that requires significant computational resources to process and analyze.

The high repetition rate and uniform/programmable time structure of LCLS-II and LCLS-II-HE will provide a transformational capability to collect 108–1010 scattering patterns (or spectra) per day with sample replacement between pulses. By exploiting revolutionary advances in data science, developing and applying advanced computational algorithms to massive data sets (e.g. kinetic inference, Bayesian analysis, pattern recognition, manifold maps, and ML algorithms) it will become possible for the first time to characterize heterogeneous ensembles of particles, map stochastic dynamics, or extract new information about rare transient events from comprehensive data sets of X-ray scattering patterns and/or spectra. These capabilities promise to revolutionize the areas of chemistry and materials science. Such "big data" experiments are challenging to execute and often require fast turnaround between data acquisition and data analysis to enable experimenters to make informed decisions when driving experiments in order to make use of the limited resources of X-ray beam time and samples. Data collection rates are growing exponentially due to light source and detector upgrades; computational requirements are also growing in proportion. In order to analyze data on experiment timescales, LCLS is sending data to remote HPC facilities such as NERSC, the ALCF, or the OLCF. Sending data from a light source facility to the HPC facility requires a data path of sufficient bandwidth.

Use Case #2: Autonomous Experiment Steering for Scientific Discovery at LCLS

Autonomous experiment steering is an emerging mode of conducting science at LCLS.

This capability has the promise to provide automated setup of the source and sample alignment, intelligent data collection, quality verification, data reduction, and coupling of experimentally derived data with information derived from theory, models, and simulations. Autonomous experiment steering has the potential to unlock new materials science knowledge to, for example, better understand failure modes in materials, enable the synthesis of new materials, aid in the creation of purpose-built designer materials, and assist in additive manufacturing processes.

Feedback must often be obtained on timescales too short for humans to react in order to plan and steer experiments to, for example, catch rare events or see fast processes. Due to the intrinsic capabilities of the current and soon to be upgraded sources, coupled with high data rate detectors that generate large volumes of data at increasingly higher rates, advanced computational techniques, including AI/ML, must be employed to realize autonomous steering of light source experiments. These methods require the utilization of considerable supercomputing power to process data and train AI/ML models that may then be used to make real-time decisions using edge or local computing systems (see **Figure 5.4.11**).

Figure 5.4.11: Prototypical autonomous experiment ML data workflow. Data generated at a light source instrument (in this case at LCLS) is streamed directly to a supercomputing center (in this case ALCF). An ML model is trained on the fly. The trained model is deployed on an edge-computing device at the light source instrument to make experiment decisions in real time.

Use Case #3: Autonomous Experiment Steering for Scientific Discovery at LCLS

Coupling of simulations and development of "digital twins" to LCLS experiments has the potential to unlock new materials science knowledge to, for example, better understand failure modes in materials, enable the synthesis of new materials, aid in the creation of purpose-built designer materials, and assist in additive manufacturing processes. It will also allow for a more efficient and optimum use of beamtime at the light sources.

The coupling of simulations with light source experiments can be split up into three main areas in the experimental life cycle. Firstly, before the experiment, simulations can be used to help prepare, plan, and determine if the experiment is even feasible. The results of these simulations may be generated and stored at a HEC facility. Secondly, during the allocated beam time, simulations can help guide and inform the strategy and guide the experiment. Finally, simulations can be used to aid in the data analysis in order to extract the maximum scientific information from the data. Below is a diagram that shows how a digital twin interacts with LCLS experiments.
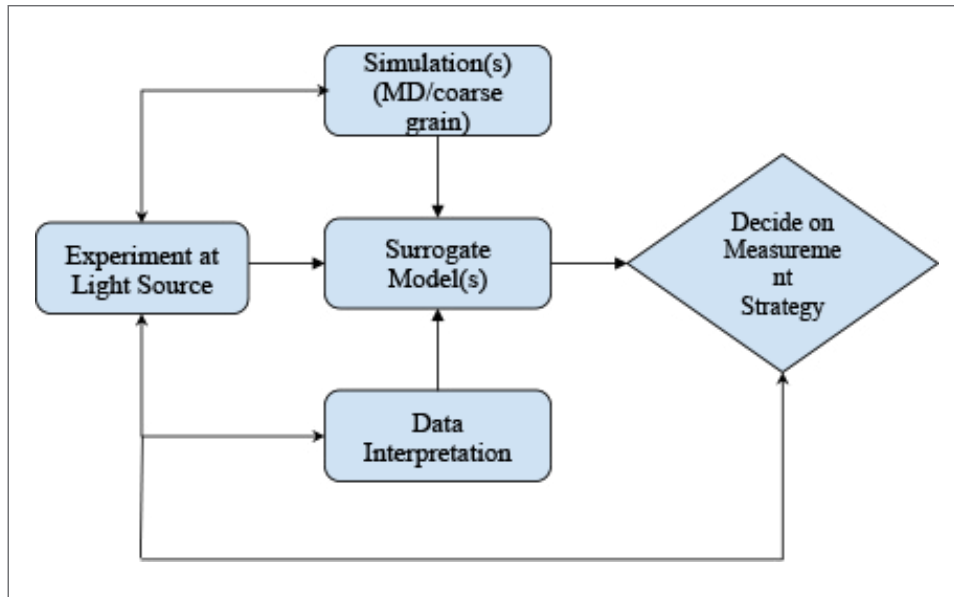
Figure 5.4.12: Experimental data is generated at a Light Source facility, such as LCLS (at left). While the data is being analyzed/interpreted and interpretation results are being produced, a series of simulations are spawned elsewhere using parameters that match current experimental conditions. These results are similarly analyzed and results are produced. The results of the simulation and the experimental data interpretation are used to derive the experiment to its next, optimal operating point. Both the experimental data analysis and the simulations may be compute intensive. The simulations may be launched at a different HEC site than the data interpretation jobs.

Data Management Gaps and/or Network Bottlenecks

Networking must be able to robustly connect LCLS instruments to remote computing facilities with low latency. Terabit per second networking and beyond will be required to handle the large amounts of data expected towards the end of the decade. There is a need to upgrade the SLAC/LCLS ESnet connections to keep pace with the expected data rates of 200 Gbps in 2023 and 1 Tbps by 2028 enabling streaming data transfer from LCLS to ASCR computing facilities.

The large amounts of data, large number of files, heterogeneous storage systems, nationally distributed data centers, and a desire to organize, manage, and access data at scale from any light source or HPC facility will require a data management system that can transparently, and in real time, move data between storage and computing, and provide for open public data repositories that will enable AI/ML technologies. Methods and tools for moving data with high fidelity across networks and methods for parallelizing and prioritizing data movement to enable HPC jobs to start quickly after data acquisition are needed. Optimized methods for streaming data to allow parallel execution on HPC resources need to be developed.

Key elements of a future data management strategy for LCLS include a common API for accessing network and computing resources, parallel data transfer tools, high-fidelity data transfer, network performance monitoring, reservations, and dynamic network provisioning. Some thought should be given to defining a methodology for handling simulated and other data to enable multimodal analysis and for reporting meaningful intermediate analysis results back to the user without abusing the network.

The further development and application of autonomous experiment steering at LCLS requires that a number of gaps in capabilities and infrastructure across the light sources, Laboratories, and computing centers be filled, including sufficient, reliable bandwidth, sufficient and sustainable data storage resources, on-demand access to large-scale computing systems, transparent access to facilities and systems across the complex, including

federated identity and shared data protocols, and software tools and infrastructure to facilitate the development of scientific data workflows that operate across the DOE's distributed resource landscape.

As the decade progresses, autonomous experiment steering will be crucial to experiment success, and this will require the use of geographically distributed computing facilities and data repositories, which in turn necessitates high bandwidth, low latency network connections between the LCLS and remote computing and data storage centers.

Coupling simulations experiment steering will become an increasingly relied upon capability at LCLS as the decade progresses. Since the simulations are generated at a different facility than the experimental data, there will be some data movement, potentially real-time if digital twins are spawned during an operating experiment, to collect the experimental and theoretical results for comparison. It is expected that this data transfer will be small, of order 1 GB/s, since the actual simulation products and raw experimental data will not be copied, only the results.

5.4.2.4 Generalized Process of Science

The process of science is different for each instrument and experiment. Nevertheless, there is overlap in the methodologies of LCLS experiments. We have identified roughly 20 unique workflows at LCLS and will describe the use cases that require the most network and compute resources. All other LCLS use cases should fit within this envelope.

Use Case: Serial Femtosecond X-ray Crystallography

Serial Femtosecond X-ray Crystallography (SFX) reveals the reaction mechanism by providing separate atomic structures for each metastable state, and several time points in between. SFX experiments offer huge benefits to the study of macromolecules, including the availability of femtosecond time resolution and the avoidance of radiation damage under operando conditions. SFX techniques will be instrumental in many experiments, ranging from the determination of macromolecules structure and dynamics, to the understanding and controlling of materials nucleation pathways, to the study of oxygenic photosynthesis.

In SFX, the structural information is derived from the diffraction data collected from a stream of individual crystals, with the primary feature extraction step consisting of measuring the Bragg spot intensities on each diffraction pattern. The main steps in the SFX algorithm are (1) identifying the Bragg diffraction spots, (2) deducing the geometry of the lattice repeat, (3) refining the model again and (4) summing the X-ray signal in each spot for further analysis. See **Figure 5.4.13** for a pictorial representation of the SFX pipeline.
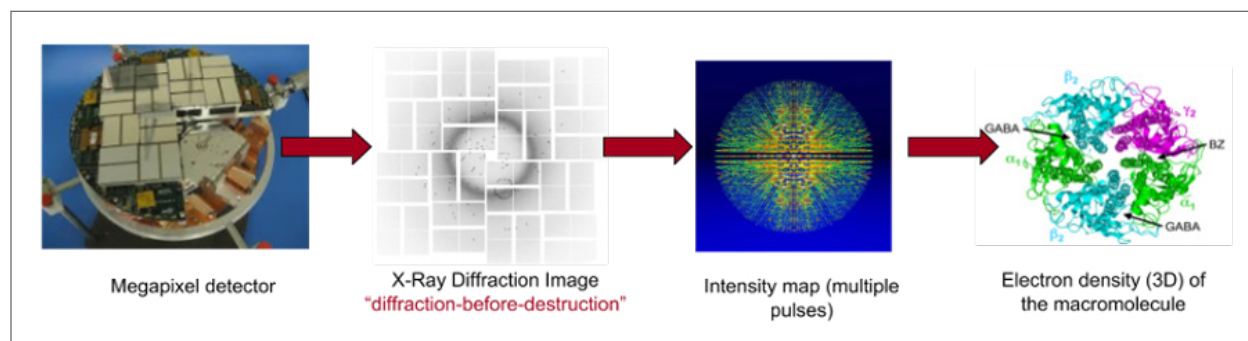
Figure 5.4.13: SFX Pipeline

Based on a 10% hit rate and on the CCTBX workflow compute requirements, we project that the rate of useful SFX events will approach 5 kHz and the computing resources needed to reconstruct SFX data in LCLS-II will run routinely in the few tens of PFLOPS and some of the more demanding algorithms, like IOTA (Integration Optimization Triage and Analysis), will require several tens of PFLOPS.

Use Case: Heterogeneous Catalysis

A second illustrative example is in the area of heterogeneous catalysis, where functional systems are neither homogeneous, nor static. The evolution of atomic and electronic structure, the making and breaking of chemical bonds, and the exchange of vibrational energy through intermediate states ultimately determine functionality. These interactions further lead to dynamic restructuring of catalyst materials during reaction. Knowing the time evolution of the atomic and electronic structure of molecules and substrates, particularly near elusive transition states, is critical to developing a predictive understanding for catalysts design. Today we are unable to develop a complete picture of this structural evolution — neither on the ultrafast timescales relevant to atomic motion, nor the nsec to msec timescales characteristic of diffusion and materials evolution. LCLS has enabled the study of simple surface reactions, and reported the first observation of a surface transition state — based on studies of ideal crystals, with reactants prepared at high concentrations in vacuum. However, these studies do not address "working" catalysts which are typically of low dimensionality (e.g. nanoclusters of metals on an oxidic support), where coupled fluctuations of the electronic and atomic structure become increasingly important.

LCLS-II-HE, coupled with advanced computational approaches applied to massive data sets, will enable completely new approaches for simultaneously following both the atomic and electronic structure of heterogeneous catalysts in operation. In one such proposed approach, nanocatalysts are prepared (either with or without preadsorbed reactants) and interrogated pulse by pulse using a gas-phase jet, liquid particle injector, or translating tape. Multimodal characterization incorporates ultrafast X-ray spectroscopy (e.g. EXAFS, XES, or RIXS) or photoelectron spectroscopy to probes the electronic structure and chemical environment, while coherent X-ray scattering from the same hard X-ray pulse probes the atomic structure. Demonstration experiments in the soft X-ray range at FLASH and at LCLS highlight the promise of this approach for characterizing small heterogeneous ensembles of nanoparticles at the atomic scale using hard X-rays.

Use Case: Single-Particle Imaging (SPI)

SPI is a promising emerging technique for such applications, Here, diffraction images are collected from individual particles, and are used to determine molecular (or atomic) structure, even from multiple conformational states (or nonidentical particles) in operating conditions that are inaccessible through other methods. These techniques will be instrumental in many areas of science, ranging from understanding and controlling nanomaterials self-assembly, to the chemistry and morphology of combustion aerosols to the coupled electronic and nuclear dynamics in heterogeneous (nano) catalysis.

However, determining structure from these experiments is challenging, since orientations and states of imaged particles are unknown and images are highly contaminated with noise. Furthermore, the number of useful images is often limited by achievable single-particle hit rates. Advanced algorithms currently under development (e.g. M-TIP by the CAMERA group) introduce an iterative projection framework to simultaneously determine orientations, states, and molecular structure from limited single-particle data by leveraging structural constraints throughout the reconstruction and offer a potential pathway to increasing the amount of information that can be extracted from single-particle diffraction.

A fundamental challenge in SPI is that the orientation of each imaged particle is unknown and must be recovered to determine structural information. Additionally, many samples display conformational flexibility and may exist in one of many possible structural states. To account for varying structural states and avoid a loss of resolution due to averaging of states, the diffraction patterns may need to be classified to the correct state. Furthermore, single particles scatter very few photons; hence the images are heavily contaminated by shot noise, often with less than a photon per Shannon pixel at high scattering angles.

5.4.2.5 Remote Science Activities

Operating experiments is the shared responsibility of beamline scientists, local staff, and on-site users. LCLS has web services, a data management system, and a data analysis infrastructure that allows for virtual access to all of the resources necessary to analyze the data taken at LCLS from a remote location. In most cases, all the same resources are available to users' doing analysis at home as at SLAC. In most cases, the data remains in LCLS local storage and LCLS offline processing resources are used to analyze. Because of the uniqueness and changeability of LCLS instruments, LCLS does not envision remote operation of the instruments.

The exascale workflow for LCLS-II-HE, illustrated in **Figure 5.4.14**, has the following key network/bandwidth requirements:

- Ability to fully utilize terabit per second links between SLAC and HEC via ESnet in order to achieve full streaming bandwidth from LCLS-II-HE.

- Ability to organize data sets (images) within a deeper storage hierarchy to allow the placement of images on burst buffers within the supercomputer system. For some architectures this will require the ability to control placement based on the exascale system`s network topology and proximity of burst buffers to computational elements.

- Ability to control resource allocation within the exascale system to enable on-demand job scheduling, which may include allocation of different resources at different times, such as burst buffer allocation prior to compute node allocation.

- Ability to orchestrate resource allocation, data movement, and data analysis execution with predictable performance and resiliency.
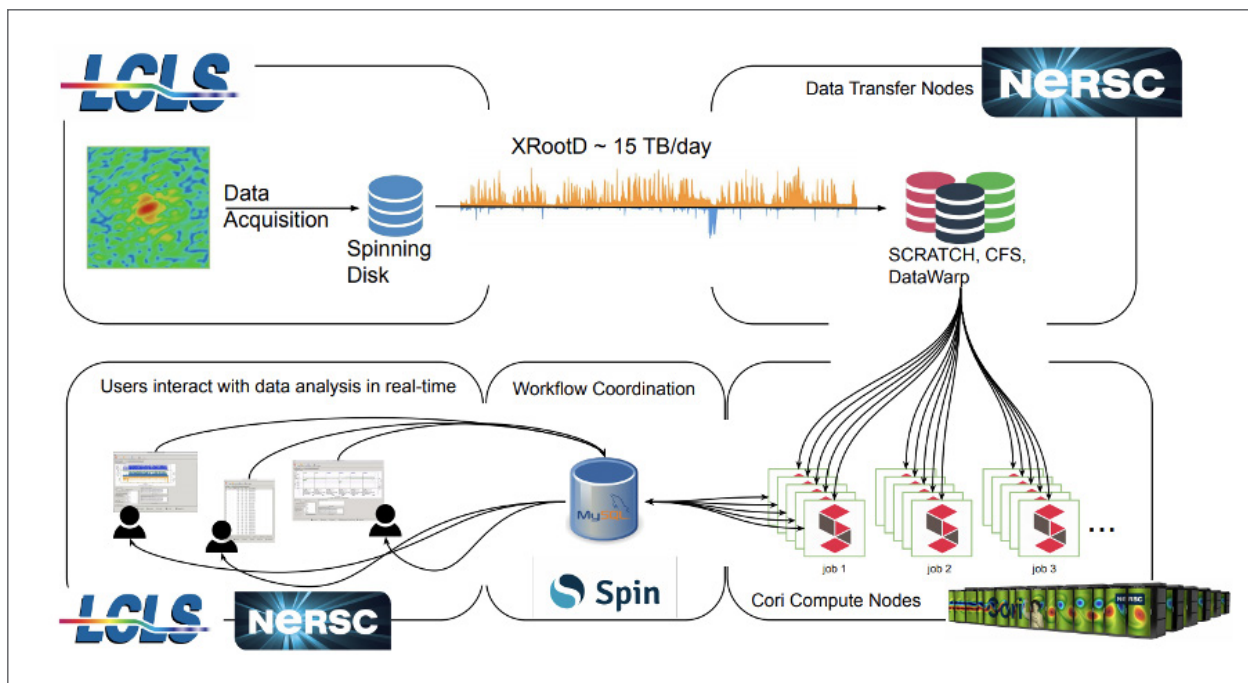
Figure 5.4.14: LCLS SFX/SPI example is a representative information extraction pipeline. In this example, raw data is acquired at the light source facility (LCLS), automatically transferred to an ASCR facility (NERSC) via ESnet, and analyzed using a scalable software library appropriate to the science domain. It is possible to achieve a turnaround time from data acquisition to full molecular reconstruction within 10 minutes. (ref arXiv.2106.11469v2). The analysis pattern is characterized by bursts of short jobs, requiring very short startup times. Current data collection rates are about 15 TB/day for high compute experiments. Within three to five years, it is expected that this rate will increase to 500 TB/day and in 5+ years to > 1 PB/day. During analysis execution at a remote HPC site, updates on job progress, intermediate results, and logfiles may be transferred back to the originating facility. After analysis jobs complete, summary results are transferred and made available to users via a web browser at LCLS. In total, this data stream from NERSC back to LCLS is about 1 Gb/s, but is likely to increase by at least a factor of 10x within three years as more sophisticated methods of visualizing intermediate results are implemented. In addition to transferring data for running experiments to NERSC for fast-feedback processing, LCLS also uses ESnet to archive data at NERSC. Typical transfers are 8–24 Gbps but are expected to increase by a factor of at least 10 within a year and a factor of 100 within three years.

As an end-to-end challenge problem, XFEL applications require significant orchestration of compute, network, and storage resources and present a model use case for ESnet R&D into network operating systems [ENOS].

LCLS has developed a powerful data management system that handles both the automatic workflows of the data through the various storage layers, e.g. long-term data archival, and the users' requests through a web portal, e.g. restoring data from tape.

Some aspects of the current system, such as checksum calculations, HPSS interface, and lack of prioritization, will become limitations at higher data volumes and will need to be upgraded for LCLS-II-HE. The LCLS data management system is capable of offloading the offline processing and storage capabilities to multiple HEC facilities via ESnet. Users can track and manage the experimental data transparently across the different systems.

5.4.2.6 Software Infrastructure

The LCLS accelerator systems and beamline instruments use the EPICS for low-level device control. The Bluesky suite of tools has been adopted for high-level experiment control. As mentioned in the previous section, remote access tools, such as ssh, NX NoMachine, will continue to be utilized to enable remote access for LCLS experiments.

The basic software infrastructure required to support the following functions of the LCLS-II data management system are as follows:

1. Security — Authentication/Authorization (currently WebAuth using SLAC Lightweight Directory Access Protocol (LDAP), Stanford SSO)

2. Scan/run control (Bluesky, Python, EPICS, Kafka message bus, custom data acquisition software)

3. File catalog (Rucio)

4. Experiment metadata (mongoDB, Kafka)

5. Electronic Logbook (custom web application)

6. Samples (mongoDB)

7. Workflow (Kafka, job submission via LSF/SLURM, AirFlow, APIs provide feedback)

8. Feedback and reporting (custom web form)

9. Instrument operator portals (MongoDB)

In addition, LCLS-II software uses tools like Kafka, MongoDB, and Python to create a development environment that is scalable, reliable, and easy to modify and build on:

1. Composability/site specificity (Kafka/Websocket, custom web service)

2. Scalability/Reliability (MongoDB, Kafka)

3. Backup/restore (MongoDB: mongodump, mongorestore)

4. Scriptability (Python, Flask, Kerberos endpoint, PAM endpoint)

5. Ease of deployment/upgrades (Python+javascript application)

6. Ease of development (Conda, Kubernetes, document-based store with Python_Javascript)

7. Code repositories (git)

8. Continuous integration (Travis)

9. Code deployment via containers (Docker, Singularity)

10. Interactive analysis (Jupyter)

The LCLS-II data management system uses a message-oriented architecture using Kafka as a message bus where all changes of interest are published on a message bus. This enables the composability of applications. The ability to separate out site-specific features into serverless components that react to these messages and shape the business data to suit the needs of the site is also useful.

Security — Authentication/Authorization

The LCLS facility has multiple instruments with multiple end stations; each end station is used by teams of users (an experiment) in a serially scheduled fashion to gather scientific data. Thus, in all the facilities, the data management system is a multitenancy environment. With some operational exceptions, only the collaborators of an experiment are allowed to read/write data belonging to the experiment.

To use the data management system, users have to authenticate themselves. To simplify account management at the facilities, integration with an external user directory (LDAP) is essential. Access to scientific data must be restricted to collaborators of the experiments where experiments are the units of authorization. There is a mechanism for the PI and others that the PI nominates to control access to the science data by adding or removing collaborators from an experiment via the electronic logbook, or eLog.

Scan/Run Control

Scientific data is gathered by the LCLS Data Acquisition System (DAQ) in units called runs/scans. Each instrument has its own DAQ capable of starting a run, gathering scientific data and writing it into files, and ending a run to indicate closure of the data gathering process. A run is also the unit of analysis; metadata, files and other data management objects are associated with a unique (to the experiment) run number. Kafka messages are published at each "start run" and "end run" transition; previously configured workflows can be launched based on arrival of these triggers.

File Catalog

Scientific data gathered by the DAQ is written into multiple files that are then registered with the data management system which associates them with the current run and places them in a file catalog. The file catalog is subsequently used as a reference/index for the rest of the experiment life cycle. To facilitate error detection, checksums and sizes are also stored with each file.

Files with scientific data are often moved/copied/synced between locations, for example, NERSC, to allow for analysis of larger data sets using HPC. This location information is also kept in the file catalog. Kafka messages are published on file registration/file movement.

Experiment Metadata

For each run/scan, the DAQ gathers metadata from the control system and other sources and stores it in the database using a web service API call. In addition, analysis of the science data can also generate new metadata, for example, hit counts, that can also be written to the database using the API.

Electronic Logbook (eLog)

The application also supports an electronic logbook for each experiment where notes for the experiment can be maintained. Logbook entries can have multiple attachments and can also be formatted HTML content. Rich text search of the elog entries is supported in addition to infinite scroll for faster page rendering. Since elog entries are the principal means of keeping track of the progress of an experiment, automatic email notifications are supported. The application supports the notion of grouping runs together and associating them with a particular sample to make it easier for the user to analyze similar data in the same context.

Within the eLog, the user can configure a workflow in which the arrival of data (or a begin run or an end run) triggers automated analysis of the scientific data at a local or remote site. In the data manager, workflows are triggered from Kafka messages. For example, automated data quality workflows are triggered using "end run" Kafka messages. Site-specific serverless components can consolidate Kafka messages into higher level messages that can trigger more complex workflows, for example, to trigger analysis on CORI when all files in a run are transferred to NERSC. Workflow definitions are scripts that are executed using the privileges of the user defining the workflow. The script typically uses a batch processing tool like LSF/SLURM to submit the analysis job; the batch jobs can then update the state of the workflow execution using web service APIs to provide feedback (for example, progress bars). AirFlow is used to support DAG workflows, and LCLS is exploring enhancements to this feature of the application.

Finally, all user facilities have instrument support personnel responsible for the operations of the facility. The data manager application has an instrument operator portal with support for creating/editing experiments/instruments. For each hutch, there is a "current" experiment. The "current" experiment feature is also used to support operator accounts; these are shared user accounts specific to an instrument. Access is granted to an instrument's operator account for the "current" experiment for that endstation. The instrument ops portal also has some basic OLAP reporting that are computed periodically using MongoDB aggregates and cached in the database.

Scalability/Reliability

All participants in the data manager architecture are clustered with support for partitioning/replication. The backend uses MongoDB which has excellent support for sharding (partitioning) and replication. To facilitate multitenancy, the data manager uses a MongoDB database per experiment. This restricts most operations to function within the context of an experiment. In addition to security, this also reduces the query space for most operations which improves latency.

Scriptability

The data manager application exposes a web service API and the UI uses this API solely for its operation. Thus, everything that can be done from the UI can be done from within Python/other scripting languages. To support automated scripting, we offer three authentication endpoints.

1. As we mentioned before, an operator account that applies only to the currently active experiment at an end station and is restricted at the web server to machines belonging to the end station.

2. A Kerberos endpoint that lets the user use their own credentials by getting a Kerberos token and then using this token to create a header for web service API calls.

3. A PAM endpoint that lets the user use their userid/password.

The DAQ uses the same API. Other Python clients, for example, a tool that grabs screenshots, also use the same API to push those screenshots into the eLog with a comment and tags supplied by the user.

Deployment

The data manager application is largely a Python+Javascript application. LCLS uses Conda for both the Python and Javascript components; for the Javascript libraries, node is installed as part of the Conda environment with the required libraries. The application then has interceptors to serve the Javascript files from within the Conda environment. CryoEM uses Kubernetes mounting storage from the IBM General Parallel File System (GPFS).

Future Work

1. Federated logins: Support for users using credentials from their home institutions.

2. Workflows: Support for more complex workflows.

5.4.2.7 Network and Data Architecture

The LCLS-II data management system and strategy to use local SLAC compute for the majority of experiments and DOE HEC facilities (OLCF, ALCF, and NERSC) for the most computationally demanding LCLS experiments has already been described above.

Figure 5.5.15 shows a different view of the LCLS-II data system, one that includes the geographic locations of each of the elements of the data system. The instrumentation and sources of data are located in one of the 10instruments in the LCLS Facility Experimental Halls. Data is transferred via a point-to-point connection from each instrument to an optical switch in the Near Experimental Hall Telecon room. The switch connects the data to the data reduction pipeline and downstream elements. After the data is reduced in volume by the data reduction pipeline, it is written to NVMe in the fast feedback storage layer. From there, it is transferred to spindle-based offline storage or transferred to NERSC via ESnet using up to 10 DTNs, each with 100 Gbps of bandwidth.
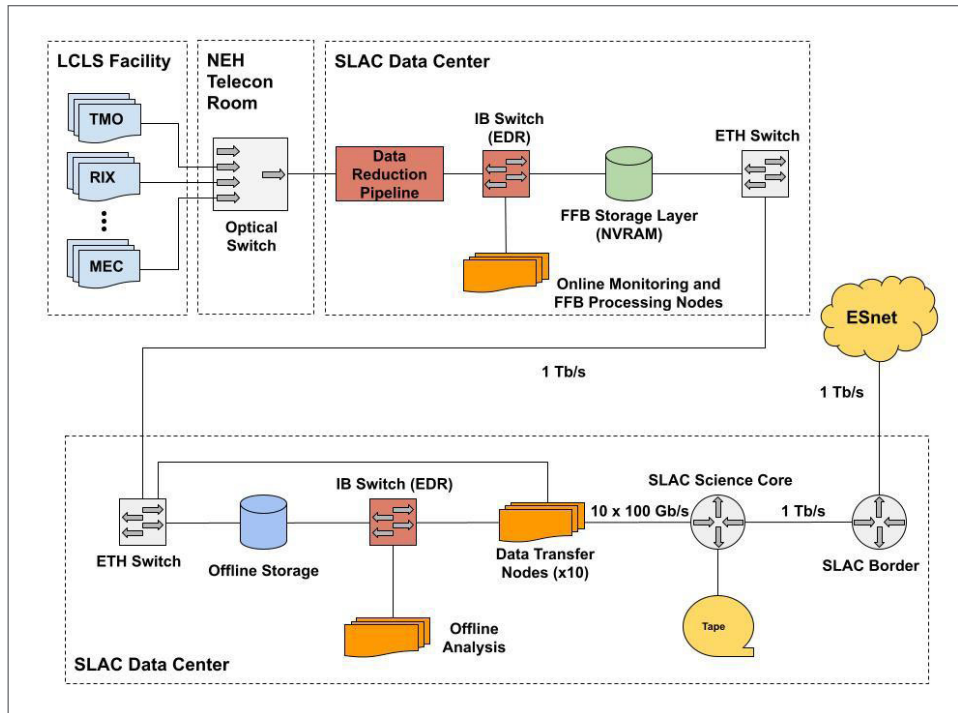
Figure 5.5.15: SLAC Network Diagram

The plot below shows SLAC's plans (yellow/orange line) for upgrading the border network to 1 Tbps. If LCLS is capable of producing data at 1 Tbps, then the HEC facility on the opposite end, such as NERSC, and ESnet in between will also need to be capable of handling 1 Tbps.
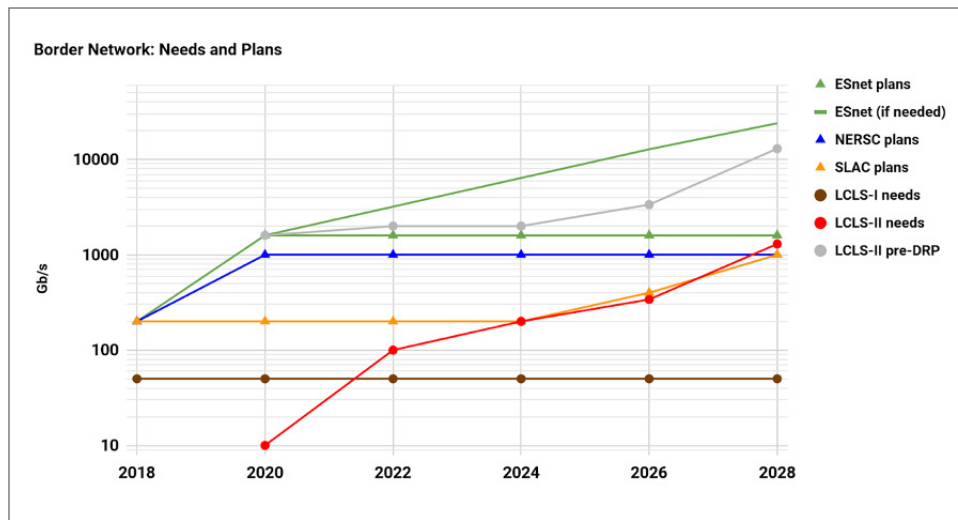
Figure 5.5.16: This plot shows the needs of the data source (LCLS), the plans for the data sink (NERSC) and the current projected plans for ESnet. The gray line shows what LCLS needs would be if there were no data reduction. LCLS-I, at 120 Hz, is the brown line. LCLS-II needs with the planned level of data reduction are shown in red. NERSC plans are shown in blue and anticipated ESnet plans are shown with the green triangle. The green line (no triangle) is the hypothetical projection of what it would take to satisfy LCLS-II needs if there were no data reduction pipeline . Note that the ESnet plans were communicated to LCLS some years ago, but are subject to change at any time. The green triangle projection shown here is not binding, but meant to show what would satisfy most of the anticipated LCLS-II requirements.

5.4.2.8 Cloud Services

None to report at this time.

5.4.2.9 Data-Related Resource Constraints

None to report at this time.

5.4.2.10 Outstanding Issues

None to report at this time.

5.4.2.11 Facility Profile Contributors

LCLS Representation

- **Jana Thayer,** *SLAC National Accelerator Laboratory*, jana@slac.stanford.edu
- **Apurva Mehta,** *SLAC National Accelerator Laboratory*, mehta@slac.stanford.edu
- **Leilani Conradson,** *SLAC National Accelerator Laboratory*, leilani@slac.stanford.edu
- **Wilko Kroeger,** *SLAC National Accelerator Laboratory*, wilko@slac.stanford.edu
- **Vivek Thampy,** *SLAC National Accelerator Laboratory*, vthampy@slac.stanford.edu
- **Stuart Campbell,** *BNL*, scampbell@bnl.gov
- **Nicholas Schwarz,** *ANL*, nschwarz@anl.gov
- **Alex Hexemer,** *LBNL*, ahexemer@lbl.gov

ESCC Representation

- **Mark Foster,** *SLAC National Accelerator Laboratory*, fosterm@slac.stanford.edu

5.5 Stanford Synchrotron Radiation Light (SSRL)

The SSRL is a national user facility dedicated to providing synchrotron radiation for scientific and technological research in a wide range of areas in support of users from academia, national laboratories, and industry.

5.5.1 Discussion Summary

- The SSRL has 32 operational beamline stations on SPEAR3, of which 26 can operate simultaneously, and with four of them being PRT/partner stations.

- SSRL provides facilities to approximately 1700 unique scientists per year (pre-pandemic).

- SSRL operates as a dedicated synchrotron radiation light source for approximately 9 months each year (mid-October through early August).

- Depending on the beamline and measurement, beam time may range from a few hours to more than a week in duration.

- The meaning, size, processing steps, and analysis and interpretation of the scientific data vary depending on the beamline, the measurement technique, the detector(s) utilized.

- SSRL data is first stored on a computer attached to the acquisition detector or beamline control system. Data can be transferred via shared filesystem (NFS) where it can be downloaded to an analysis machine.

- SSRL has been working closely with the other four BES light source facilities, coordinated through the 5-way LSDCSC. This group produced a unified vision for the distributed data infrastructure to enable user science, the DISCUS.

- SSRL plans to implement the Bluesky databroker/tiled software for the initial data management layer to enable a data acquisition system consistent with the other light sources. This has been developed as a collaborative open-source project. This system aims to provide a consistent "data API" rather than prescribe a given on-disk data format.

- Imaging, high-energy, high-speed operando and multimodal techniques are generally the most computationally intensive techniques performed at the SSRL. Data reduction and analysis rely heavily on the use of HPC, GPUs, edge devices, and distributed computing environments to obtain results with near real-time completion.

- The computing resources required by the SSRL continue to grow, and there is currently a wide variability in the computational requirements, use of SLAC computing resources (e.g., local, institutional, or edge-computing resources) offers the ability to process data quickly without the need for large-scale data movement.

- To analyze data on experiment timescales, we anticipate that it will soon become necessary to send data to remote HPC facilities such as the Stanford Research Computing Facility (SRCF), SLAC Computing Services (SCS), or NERSC. Sending data from a light source facility to the HPC facility requires a data path of sufficient bandwidth.

- SSRL guarantees storage space for a minimum of 24 months; the long-term management of data is the responsibility of the user group that collected the data.

- In some cases, experimenters transfer the raw data to computing resources external to SLAC; either at their home institutions or dedicated HPC facilities.

- SSRL users can download data at their home institutions or on their personal computers using FTP servers or by direct download onto portable hard drives. Users also often utilize their own "cloud storage" accounts (e.g., Google Drive, Dropbox, etc.) to transfer data.

- There is a high demand from the user community to integrate cloud services for both file transfer and sharing, as well as and for communication into all areas of the data lifecycle and compute workflows. We have seen this demand increase during the COVID-19 pandemic.

- SSRL uses Amazon AWS and Microsoft Azure cloud platforms in the areas of prototyping, testing, and standing up computational/storage resources to support training and demonstrations.

- SSRL, and other BES facilities, continue to evaluate the capabilities and cost-effectiveness of using cloud services.

- External network connectivity to SSRL consist of 2x10 Gbps uplinks provided by SLAC networking, which obtains its external network connectivity via 100Gb connectivity from ESnet and Stanford University to Internet2.

5.5.2 SSRL Facility Profile

The goal of SSRL is to enable outstanding scientific research for a broad GU community and provide excellent facilities and user support. The facility is well-known and recognized for pioneering contributions in new synchrotron methods and instrumentation and their use in enabling new science. The accelerator was completely replaced in 2004 by a 3rd generation, low-emittance storage ring, SPEAR3, jointly funded by the DOE and the NIH. SPEAR3 operates in top-off injection mode at 3 GeV and 500 mA, and at an emittance of 7 nm with a high degree of reliability.
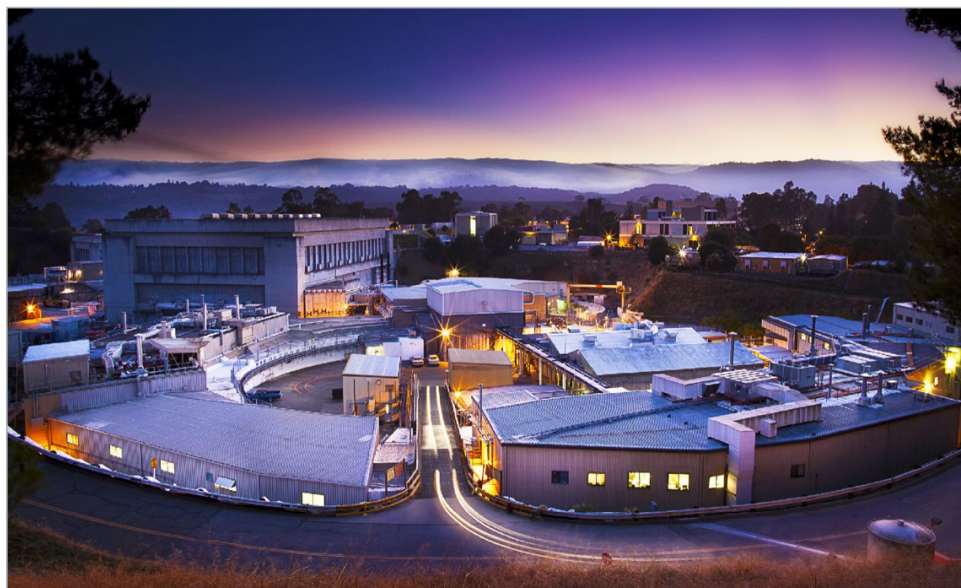
Figure 5.5.1: Aerial View of the SSRL

5.5.2.1 Science Background

There are currently 32 operational beamline stations on SPEAR3, of which 26 can operate simultaneously, and with four of them being PRT/partner stations[1]. Of the 32, 23 have insertion device sources (undulator, wiggler) and 9 have bending magnet sources. An additional two beamline stations, on two beamlines, are in the commissioning stage. The majority of beam time is allocated via a GU proposal approach and the proposal review and beam time allocation is managed centrally for all beamlines.

---

[1] Stanford Synchrotron Radiation Lightsource Strategic Plan,
https://www-ssrl.slac.stanford.edu/content/sites/default/files/documents/ssrl_strategic_plan_2021-2025.pdf

SSRL provides essential scientific facilities to ~1700 unique scientists per year (pre-pandemic operations), resulting in ~700 publications per year. The strong commitment to science education and mentoring is demonstrated by the more than ~120 theses per year citing support for work partly or fully supported at SSRL and enhanced by a significant training and outreach activity. Operation of SSRL is funded by the US Department of Energy, Office of BES, with additional support for SSRL's structural molecular biology program from the NIH and the DOE Office of Biological and Environmental Research.
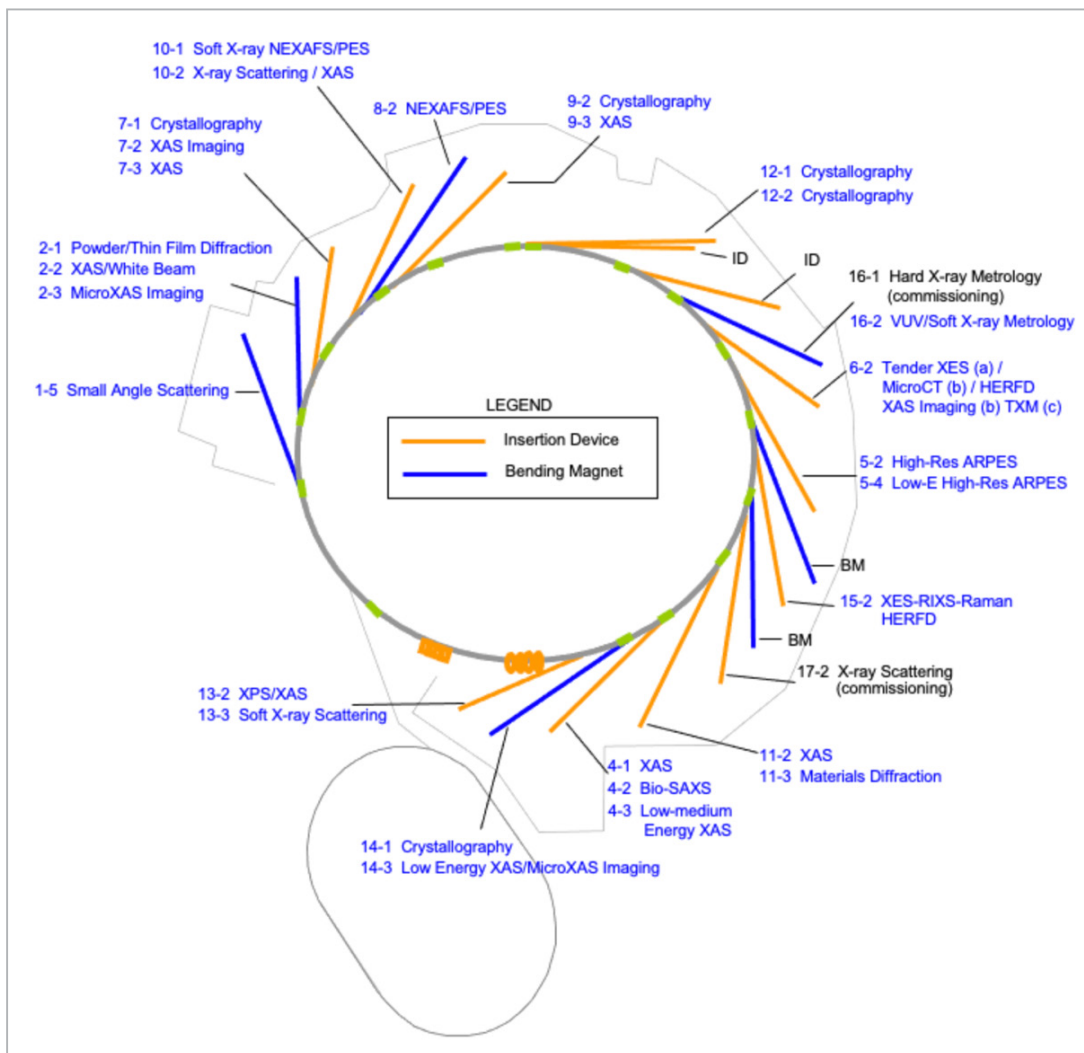
Figure 5.5.2: SSRL Beamline Map (2022)

The five main research areas at SSRL (Materials Science, Quantum Materials, Chemistry and Catalysis, Structural Molecular Biology, and Environmental/Geological Science) address three themes: Accelerating Materials Design, Understanding Catalytic Function and Interfacial Reactions with Atomic Precision, and Identifying how Collective Function Emerges from Constituent Interactions. The range of scientific research areas along with the varied experimental techniques used at SSRL results in huge volumes of experimental data of considerable complexity.

The meaning of data, the size of data, the processing steps applied to data, and the analysis and interpretation of data vary depending on the beamline, the measurement technique, the detector(s) utilized, and the scientific goal. Raw data is generated primarily by two-dimensional (area), one-dimensional (strip), or point detectors as

a part of scattering, imaging, or spectroscopy measurements. Raw data may represent, a scattering pattern, a transmission image, or spectra, for example. The size of the data may vary from a few megabytes to hundreds of terabytes per allocated beam time. Some form of data processing or reduction is usually performed after data is collected to transform the data from a technique or detector representation to the physical space of interest in an analyzable representation, such as a series of sinograms into a three-dimensional world-space volume, or spectra into elemental concentrations.

Generally, data is first stored on a computer attached to the acquisition detector or some other local beamline storage system. In many cases, the data is transferred to a shared network drive (NFS), from where it can be downloaded to an analysis machine. Currently, SSRL guarantees storage space for a minimum of 24 months; the long-term management of data is the responsibility of the user group that collected the data. In some cases, experimenters transfer the raw data to computing resources at their home institutions for all further processing. The final analysis and interpretation of the data for publication is generally carried out by the experiment team at their home institutions, is very experiment and technique specific, and may take months to even years to perform.

5.5.2.2 Collaborators

SSRL operates as a dedicated synchrotron radiation light source for approximately 9 months each year (usually from mid-October through early August). The SSRL User Facility Access Policy, available on the SSRL website[2], provides a concise overview of the access policy for SSRL's beamlines and instrumentation. SSRL has established a scientific peer-review process focused on transparency, equity, and efficiency. It recognizes the needs and contributions of users and staff to maintain and strengthen SSRL's position as an internationally leading photon science user facility. Policies that impact the user community are developed in consultation and coordination with external advisory committees, including the SSRL Proposal Review Panel (PRP), SSRL Users' Executive Committee, and the SSRL Scientific Advisory Committee.

Access to SSRL is mainly granted through proposals that are peer reviewed and rated by the PRP. The PRP consists of scientific experts, all of whom are external to the SSRL facility. The PRP is organized into five subpanels by scientific disciplines and techniques that cover a broad range of basic and applied science as well as method and instrumentation development. Since submitting a proposal is the first step to access beam time at SSRL. New users are encouraged to contact SSRL scientists, who are available to advise and answer questions about the SSRL SPEAR3 accelerator, beamlines, instrumentation, capabilities, science or techniques in order to help users plan their experiments and proposals. Modes of accessing beam time at SSRL include GU and PU proposals. A small percentage of beam time is available for SSRL staff access and through the Director's discretionary access. In addition, GU or PU proposals can be submitted for nonproprietary or proprietary research. User access types include:

- GU Access
  - Standard GU Proposals
  - RA GU Proposals
  - Letter of Intent/Director's Discretionary Access (LOI)
- PU Access
  - Collaborative Access Proposals (CAPs)
  - Participating Research Teams (PRTs)

Depending on the beamline and measurement, beam time may range from a few hours to more than a week in duration. As previously stated, SSRL users perform experiments covering many scientific fields, such as life sciences, quantum materials, energy storage, advanced materials science, physics, chemistry, and biology (Figure 5.5.3). These users also span a range of expertise levels (Figure 5.5.4) and come from varied institutions

---

[2] http://www-ssrl.slac.stanford.edu/content/user-resources/user-facility-access-policy

(Figure 5.5.5). In response to the pandemic, the deployment of remote access tools has enabled experimenters to run measurements remotely and has led to an increase in remote users as a fraction of total users in FY 2021. SSRL caters for users that come from both throughout the United States and overseas. 5.5.6 shows the total number of publications, theses, books and conference proceedings increasing steadily at peaking at over ~700 per year.
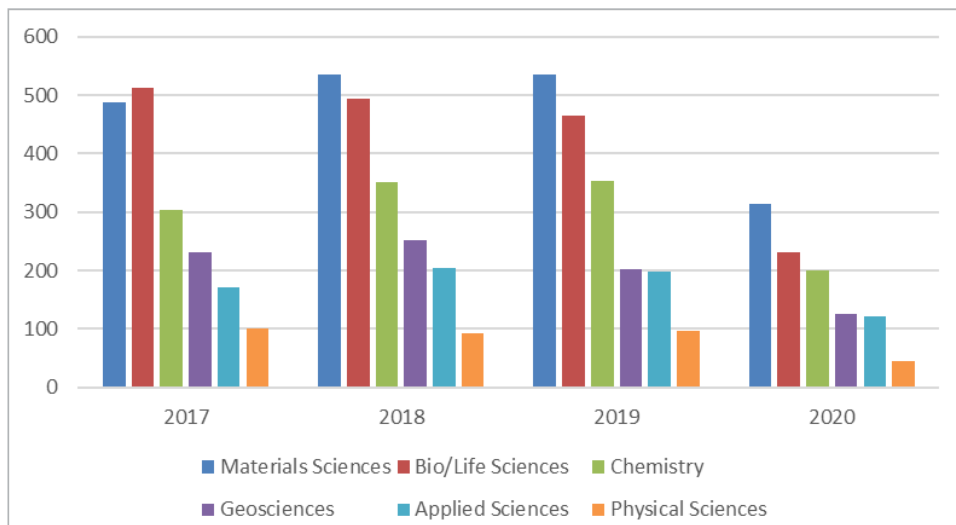
Figure 5.5.3: Breakout of requested and allocated experiments by scientific disciplines during FY-2017FY2020, summed over all the beamlines
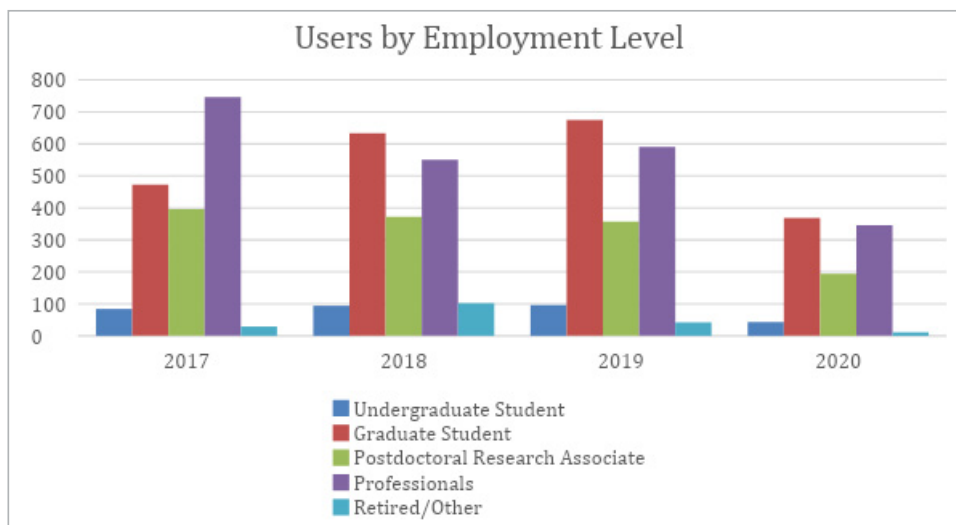
Figure 5.5.4: Proportions of different categories of users
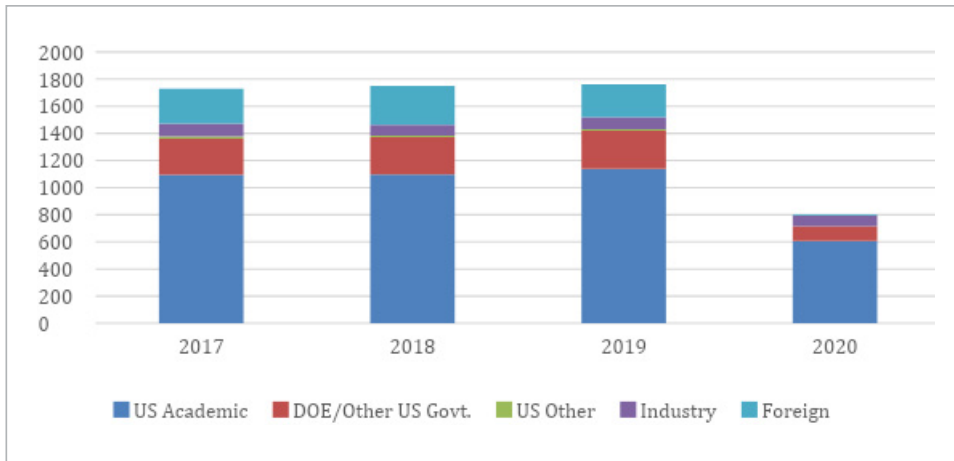
Figure 5.5.5: SSRL usage by institution type
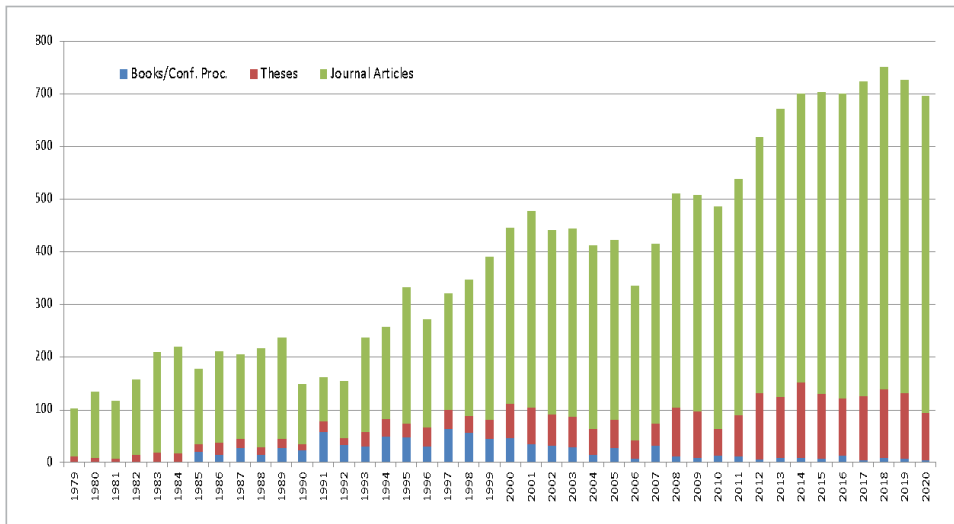
Figure 5.5.6: Publications, Doctoral Theses and Books/Conference Proceedings attributed to SSRL

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 tweo the source and method	Any known issues with data sharing
US UNIVERSITY-BASED PIS	Both	data portal, data transfer (sftp, cloud storage), portable hard drive	1 MB–10 TB	Ad hoc	N	N/A
US NATIONAL LABS BASED PIS	Both	data portal, data transfer (sftp, cloud storage), portable hard drive	1 MB–10 TB	Ad hoc	N	N/A
INTERNATIONAL PIS	Both	data portal, data transfer (sftp, cloud storage), portable hard drive	1 MB–10 TB	Ad hoc	N	N/A

Table 5.5.1: SSRL Collaboration Space

5.5.2.3 Instruments and Facilities

SSRL currently has 32 operational beamline stations on SPEAR3, of which 26 can operate simultaneously, and with four of them being PRT/partner stations. Of the 32, 23 have insertion device sources (undulator, wiggler) and 9 have bending magnet sources. An additional two beamline stations, on two beamlines (BL 15-2 and BL 17-2), are in the commissioning stage.

Summarized below are the experimental instruments, techniques and scientific areas for the different beamlines.

Station	Technique	Field of Science	Instruments
BL2-1	Powder/Thin Film Diffraction and Reflectivity	Materials / Applied Energy	Diffractometer, Scintillator and Si Drift Detector, PILATUS 100K, Stages, Cryostat, Furnace
BL2-2	XAS / In-situ Catalysis XAS	Catalysis / Chemistry / Energy / Materials	Si Drift Detector; Detector Arrays, Reactor Cells, Gas Feeds, Mass Flow Control System
BL2-3	MicroXAS Imaging	Environmental / Geoscience / Biology / Medical / Materials / Chemistry	SIGRAY Microfocus System; Si Drift Detector

Table 5.5.2: Scientific Activity and Instrumentation for the Stations on Beam Line 2

Station	Technique	Field of Science	Instruments
BL4-1	X-ray Absorption Spectroscopy	Environmental / Geoscience / Chemistry / Materials	Varies: Rails, 30-element Ge Detector Array, Ionization Chamber Detector, LHe and LN Cryostats, Sample Magazines and Automated Sample Changer
BL4-2	Small and Wide-Angle X-ray Scattering	Biology	Small-Angle X-ray Scattering Camera w/CCD & PAD Detectors, Wide-Angle X-ray Scattering Instrumentation, Stopped-Flow Mixer, HPLC
BL4-3	X-ray Absorption Spectroscopy	Chemistry / Catalysis / Environmental / Biology / Materials	Varies: Rails, Ge Detector Array, Ionization Chamber, PIPS and Si Drift Detectors, LHe Cryostat, Tender X-ray Instrumentation, Sample Magazines and Automated Sample Changer

Table 5.5.3: Scientific Activity and Instrumentation for the Stations on Beam Line 4

Station	Technique	Field of Science	Instruments
BL5-2	Ultrahigh Energy Resolution Photoemission, ARPES	Correlated Materials	SCIENTA DA30 Analyzer, MBE Deposition System
BL5-4	Low-Energy, High-Resolution Photoemission, ARPES	Correlated Materials	SCIENTA R4000 Analyzer, Low Energy Electron Diffraction

Table 5.5.4: Scientific Activity and Instrumentation for the Stations on Beam Line 5

Station	Technique	Field of Science	Instruments
BL6-2A (FRONT)	Tender X-ray XES	Chemistry / Materials / Biology	Tender XES Instrument
BL6-2B (MIDDLE)	HERFD MicroXAS Imaging / MicroCT	Geology / Forensics / Materials / Biology / Chemistry / Catalysis	HERFD XAS Imaging Instrument / Micro Computed Tomography Instrument
BL6-2C (REAR)	Hard X-ray Transmission Microscopy	Catalysis / Materials / Chemistry / Geochemistry / Forensics	Zeiss Xradia Microscope with CCD detector

Table 5.5.5: Scientific Activity and Instrumentation for the Stations on Beam Line 6

Station	Technique	Field of Science	Instruments
BL7-1	Macromolecular Crystallography	Structural Molecular Biology	ADSC Q315R CCD Detector, Goniometer, Cryostat, Robotic Sample Changer, Rapid Freeze Quencher
BL7-2	MicroXAS Imaging	Biology / Medical / Geochemistry / Chemistry / Materials	Capillary Optics and Variable Apertures, Imaging Instrumentation, Vortex Detectors
BL7-3	X-ray Absorption Spectroscopy	Structural Molecular Biology	Rails, Ge Detector Array, LHe Cryostat, PIPS Detector

Table 5.5.6: Scientific Activity and Instrumentation for the Stations on Beam Line 7

Station	Technique	Field of Science	Instruments
BL8-2	Photoemission; NEXAFS; Circular Dichroism	Material Science, Physics, Surface Science, Chemistry	100-1300 eV SGM. Chamber 1: CMA Chamber 2: Janis cryostat (4-350K); "ambient" pressure ($10^{-4} \sim 10^{-11}$ Torr)

Table 5.5.7: Scientific Activity and Instrumentation for the Stations on Beam Line 8

Station	Technique	Field of Science	Instruments
BL9-2	Macromolecular Crystallography	Structural Molecular Biology	PILATUS 6M PAD, SSRL Single-axis micro-diffractometer, Cryostat, Robotic Sample Changer, UV-VIS micro-spectrophotometer
BL9-3	X-ray Absorption Spectroscopy	Structural Molecular Biology	Varies: Rails, 100-Pixel Monolithic Ge Detector, PIPS Detector, LHe Cryostat, Single Crystal XAS

Table 5.5.8: Scientific Activity and Instrumentation for the Stations on Beam Line 9

Station	Technique	Field of Science	Instruments
BL10-1	Photoemission; NEXAFS	Materials / Chemistry	TES Spectrometer, Scienta SES-100, Silicon Diode AXUV100, Channeltron, Manipulator (~80-1100K)
BL10-2A (FRONT)	X-ray Diffraction/ Scattering	Materials	Two-circle goniometer, Specialized Experimental Specific Instrumentation, EIGER1M Detector
BL10-2B (REAR)	In situ X-ray Absorption Spectroscopy	Catalysis / Materials	Dual Sample Stages, Ge Detector Array, Infrared Spectrometer, Compressed Gas and Liquid Feeds, Mass Spectrometer, Gas Chromatograph

Table 5.5.9: Scientific Activity and Instrumentation for the Stations on Beam Line 10

Station	Technique	Field of Science	Instruments
BL11-2	X-ray Absorption Spectroscopy	Biogeochemical and Interfacial / Environmental / Chemistry / Catalysis	Varies: Rails, 100-Pixel Monolithic Ge Detector, Ionization Chamber Detector, LHe and LN Cryostats, GIXAS, Environmental Chamber for Radioactive Material Handling, Sample Magazines and Automated Sample Changer
BL11-3	X-ray Scattering; Thin Film Diffraction	Materials / Environmental / Polymers / Catalysis	Rayonix 225 CCD Detector, Vortex, Kappa Goniometer

Table 5.5.10: Scientific Activity and Instrumentation for the Stations on Beam Line 11

Station	Technique	Field of Science	Instruments
BL12-1	Macromolecular Crystallography	Structural Molecular Biology	Dectris EIGER X 16M PAD Detector, Microdiffractometer, Cryostat, Robotic Sample Changer for Cryogenic and RT Samples
BL12-2	Macromolecular Crystallography	Structural Molecular Biology	Dectris PILATUS 6M PAD Detector, Microdiffractometer, Microcollimators, Cryostat, Robotic Sample Changer for Cryogenic Samples

Table 5.5.11: Scientific Activity and Instrumentation for the Stations on Beam Line 12

Station	Technique	Field of Science	Instruments
BL13-2	Soft X-ray Absorption Spectroscopy, Photoemission	Catalysis / Surface Chemistry / Materials	Electron Analyzer X-ray Photoemission System; Ambient Pressure Photoemission System
BL13-3	Resonant Soft X-ray Scattering Resonant Inelastic X-ray Scattering Soft X-ray Absorption Spectroscopy X-ray Magnetic Circular Dichroism	Materials Science / Quantum Materials / Energy Science / Heterostructures	Scattering Chamber incl. in-Vacuum Diffractometer, X-ray Spectrometer (Transition-Edge-Sensor, TES), In-Vacuum CCD Area Detector, Photodiodes, Cryostat, Sample Holder with Magnet

Table 5.5.12: Scientific Activity and Instrumentation for the Stations on Beam Line 13

Station	Technique	Field of Science	Instruments
BL14-1	Macromolecular Crystallography	Structural Molecular Biology	MAR325 CCD Detector, SSRL Single-axis microdiffractometer, LN2 Cooler, Robotic Sample Changer
BL14-3A	X-ray Absorption Spectroscopy	Biology / Environmental / Chemistry / Materials	Vortex Si Drift Detector, PIPS Detector, Inert He Atmosphere Sample Environment, Multisample Delivery
BL14-3B	MicroXAS Imaging	Biology / Environmental / Biogeochemistry / Chemistry / Materials	SIGRAY Microfocus System; Si Drift Detector; XAS Imaging Instrumentation

Table 5.5.13: Scientific Activity and Instrumentation for the Stations on Beam Line 14

Station	Technique	Field of Science	Instruments
BL15-2	Inelastic Scattering / Advanced Spectroscopy / HERFD Spectroscopy / Time-resolved Spectroscopy	Chemistry / Materials / Biology	Advanced Spectroscopy Instrumentation Ultrafast MHz Fiber Laser Fast Detectors

Table 5.5.14: Scientific Activity and Instrumentation for the Stations on Beam Line 15

Station	Technique	Field of Science	Instruments
BL16-1	X-ray absorption, reflectivity	Metrology	Hutch configured for custom instrumentation
BL16-2	X-ray absorption, reflectivity	Metrology	Reflectometer, absorption spectrometer

Table 5.5.15: Scientific Activity and Instrumentation for the Stations on Beam Line 16

Station	Technique	Field of Science	Instruments
BL17-2	X-ray diffraction, scattering	Materials / Energy	Materials diffraction and scattering instrumentation, multiple detector systems, operando sample environments

Table 5.5.16: Scientific Activity and Instrumentation for the Stations on Beam Line 17

Each of these beamlines has its own unique data generation characteristics. The existing beamlines continue to be upgraded and new beamlines developed to meet the research needs of the scientific community. SSRL beamlines often have many dozens of motors, readout electronics, and X-ray detectors that are controlled by real-time hardware devices. Many X-ray detectors have specific manufacturer-supplied interfaces. FPGA or ARM devices are often used for real-time coordination of devices used during measurements. To meet the demands of ever-increasing data rates and volumes, there have been recent upgrades to the SSRL networking. For details, please see Section 5.6.2.7.

The utilization of multimodal data to answer new questions requires more complex and sophisticated data processing algorithms requiring increases in computing capabilities. Increases in computing power are needed by advanced algorithms for existing techniques that, for example, provide higher-fidelity results, and to train AI/ML models. The need for real-time analysis and feedback to make crucial experiment decisions and enable autonomous experiment steering also requires more computing cycles than have been traditionally utilized.

The computing resources required by the SSRL continue to grow as new beamlines are developed and existing instrumentation is upgraded to use higher speed area detectors, especially area detectors. There is wide variability in the computational requirements among techniques and processing approaches with those instruments and techniques that benefit most from high-energy, high-brightness, and cutting-edge operando/in-situ experiments driving most requirements.

Edge computing offers the ability to process data quickly on or near detectors and experiment instrumentation without the need to first transfer all data to high-end computing resources. This is particularly promising for handling large data when coupled with machine-learning methods. Using only a subset of data, machine-learning models may be trained on supercomputers. The trained model is then run using edge-computing devices to process newly acquired data, providing fast feedback for experiment steering.

The pandemic has heightened the needs for advanced, readily available capabilities at SSRL, as well as the need for improved experiment infrastructure and processes to allow these capabilities to be operated and accessed remotely, both by staff and by the users. SSRL has been aggressively ramping up the infrastructure and capabilities for remote access. As the time for a single measurement decreases, the need for tools such as AI/ML to reduce human intervention increases. Thus, improving remote experiments capabilities are of high strategic importance for the future health of the facility.

5.5.2.4 Generalized Process of Science

There is great diversity in the process of science at SSRL due to the large and diverse user community. Multiple processes and workflows exist for each beamline, instrument, and technique, and new processes and workflows are continually developed and refined. Many of the processes, however, may be categorized into three main themes:

Processing/Reducing/Analyzing Large Amounts of Data from Light Source Instruments

Experiments that generate complex and large data volumes are challenging to execute and often require fast turnaround between data acquisition and data analysis to enable experimenters to make informed decisions when driving experiments to make use of the limited resources of X-ray beam time and samples. Data collection rates are growing exponentially due to light source, instrument and detector upgrades, and computational requirements are also growing in proportion. To analyze data on experiment timescales, we anticipate that it will soon become necessary to send data to remote HPC facilities such as the SRCF, SCS, or NERSC. Sending data from a light source facility to the HPC facility requires a data path of sufficient bandwidth. Raw and processed data is often stored on a large disk system at SSRL from which users may retrieve the data or share the data with collaborators. Researchers may refine results and combine data collected at the SSRL with data from other sources, including simulations; this is usually performed outside of the facility-based workflow. Over the coming few years, due to increases in data volumes and complexity, SSRL anticipates leveraging computational resources outside of SLAC.

Adaptive Feedback and Autonomous Experiment Steering Using Advanced Computational Techniques, Including AI/ML Methods

Autonomous experiment steering is an emerging mode of conducting science at the 5 US DOE-funded light sources. Within the light sources, this capability has the promise to provide automated setup of the source and sample alignment, intelligent data collection, quality verification, data reduction, and coupling of experimentally derived data with information derived from theory, models, and simulations[345]. Autonomous experiment steering has the potential to unlock new materials science knowledge to, for example, better understand failure modes in materials, enable the synthesis of new materials, aid in the creation of purpose-built designer materials, and assist in additive manufacturing processes. Feedback must often be obtained on timescales too short for humans to react to plan and steer experiments to, for example, catch rare events or see fast processes. Due to the intrinsic capabilities of the current and soon to be upgraded sources, coupled with high data rate detectors that generate large volumes of data at increasingly higher rates, advanced computational techniques, including AI/ML, must be employed to realize autonomous steering of light source experiments. These methods require the utilization of considerable supercomputing power to process data and train AI/ML models that may then be used to make real-time decisions using edge or local computing systems

An area of active development that will soon begin showing benefits is the use of advanced computational techniques, especially AI/ML, for adaptive feedback and autonomous experiment steering. In this scenario, data collected from a detector acquisition system is considered training data for AI/ML models. This data is sent to a large-scale computer where AI/ML models are trained. The training data is stored in a repository of training data for later use. The trained model is also stored in a repository for later use and sent to an edge-computing device near the detector. Subsequent data sets are processed on the edge device near the instrument generating control instructions sent to the instrument.

Although the data generated by individual experiments may vary greatly, over the next decade the combined data generation rates of the 5 US DOE-funded light sources is expected to approach the exabyte per year range. In order to process this data, the light sources are expected to require tens to many hundreds of PFLOPps of on-demand processing capacity. Networking must be able to connect light source instruments to edge, local, campus, and centralized computing facilities reliability, and with low latency. Terabit per second networking and beyond will be required to handle the large amounts of data expected in the coming years.

3 Phillip M Maffettone et al 2021 Mach. Learn.: Sci. Technol. 2 025025

4 Campbell SI, Allan DB, Barbour AM, et al (2021) Outlook for artificial intelligence and machine learning at the NSLS-II. Machine Learning: Science and Technology 2:013001. https://doi.org/10.1088/2632-2153/abbd4e

5 D. Olds, D. B. Allan, T. A. Caswell, J. Lynch, P. M. Maffettone and S. I. Campbell, "Optimizing High- Throughput Capabilities by Leveraging Reinforcement Learning Methods with the Bluesky Suite," 2021 3rd Annual Workshop on Extreme-scale Experiment-in-the-Loop Computing (XLOOP), 2021, pp. 36-42, doi: 10.1109/XLOOP54565.2021.00011

Autonomous experiment steering will become an increasingly relied upon capability at the light sources as the decade progresses. Due to the high computational cost associated with data processing, reduction, and analysis, and of model training on such large data sets, the light sources will need to employ the use of geographically distributed computing facilities and data repositories. This necessitates high bandwidth, low latency network connections between the light sources and remote computing and data storage centers.

Coupling Large-scale Simulations, Digital Twins, Surrogate Models, and AI/ML With Experiment Data in Real Time to Drive Experiment Design and Experiment Decisions

Coupling of simulations and development of "digital twins" to light source experiments has the potential to unlock new materials science knowledge to, for example, better understand failure modes in materials, enable the synthesis of new materials, aid in the creation of purpose-built designer materials, and assist in additive manufacturing processes. It will also allow for a more efficient and optimum use of beamtime at the light sources. For a high-level flow chart of combining experiments with digital twins, see **Figure 5.5.5**.

The coupling of simulations with light source experiments can be split up into three main areas in the experimental life cycle. Firstly, before the experiment, simulations can be used to help prepare, plan, and determine if the experiment is even feasible. Secondly, during the allocated beam time, simulations can help guide and inform the strategy and guide the experiment. Finally, simulations can be used to aid in the data analysis to extract the maximum scientific information from the data.
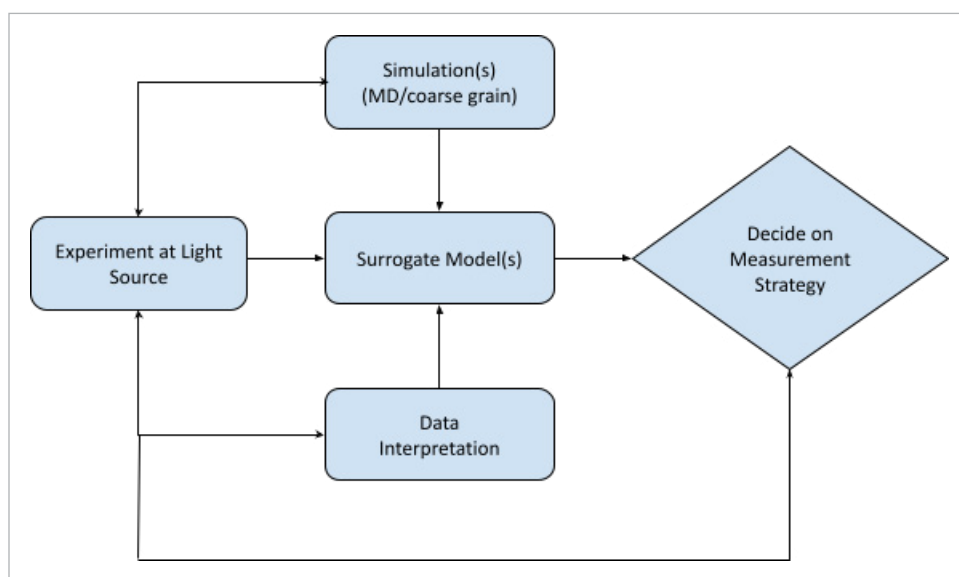
Figure 5.5.7: Flow chart of using a digital twin with light source experiments

This mode of science promises to help accelerate scientific discovery, for example, by enabling faster and more complex materials synthesis. In this scenario a supercomputing resource is used to incorporate experimentally derived data into simulations, the results of which are used to guide experiments in real time and to even plan experiment campaigns and sample compositions.

Coupling simulations experiment steering will become an increasingly relied upon capability at the light sources as the decade progresses. Due to the high computational cost associated with data processing, reduction, and analysis, and of model training on such large data sets, the light sources will need to employ the use of geographically distributed computing facilities and data repositories. This necessitates high bandwidth, low latency network connections between the light sources and remote computing and data storage centers.

Unified solutions across the light sources are required in order to leverage efficiencies of scale, and to provide facility users with the ability to manipulate data easily and transparently across the light sources. A shared computational fabric for the complex should be developed that connects light source instruments (and other scientific user facilities) to a multitiered, distributed computing landscape, including edge, local, campus, and supercomputing centers, data repositories and archives, and facility user institutions in a seamless and transparent manner. This necessitates the development of advanced networking capabilities and increased networking bandwidth, sustainable and discoverable data repositories, on-demand real-time supercomputing access, and workflow and orchestration tools. Also, as most light source users are not experts in using computational facilities, there needs to be easy access and both in terms of obtaining the resources and performing the calculations.

SSRL has been working closely with the other four BES light source facilities, coordinated through the 5-way LSDCSC. This group produced a unified vision for the distributed data infrastructure to enable user science, the DISCUS, and a decade long roadmap to achieve the vision (see **Figure 5.1.12**). This vision proposes a transformative computational fabric that covers the full lifecycle of data generated at the BES light sources to accelerate discovery and insight. This vision proposed connecting the 200+ instruments at the light sources to a multitiered computing landscape, including edge, local, campus, and ASCR compute resources, and discoverable data repositories using high-performance, robust feature-rich networks. This fabric would facilitate the full lifecycle of data across the complex, including theory/modeling and simulation, experiment design, data generation at scientific instruments within the light sources, data reduction and processing, analysis and interpretation, and publication and dissemination, serving the 10,000+ light sources users per year.

5.5.2.5 Remote Science Activities

There has been a trend to increase remote access to synchrotrons in recent years, thanks to the ongoing advances in robotic sample handling, and in automated data acquisition and data management. The current COVID-19 restrictions have accelerated such demands. Furthermore, increasingly brighter synchrotron sources and faster detectors have dramatically shortened the data collection times, to the point that data collection is the least time-consuming activity in an increasing number of experiments.

Working with other BES light source facilities, we have formed a remote experiments task force to work on implementing remote access and tele-experiments at SSRL, and to look at access policies, procedures, and user experience. We see remote experiments as essential for SSRL to maintain leadership and attract the best science in an increasingly connected world.

5.5.2.6 Software Infrastructure

SSRL accelerator systems rely primarily on the EPICS for low-level device control while the beamline instruments rely partly on EPICS and partly on SPEC and other custom software for low-level device control. High-level experiment control and orchestration is implemented using SPEC and other custom control software. The Bluesky[6] suite of tools has been used for certain high-throughput applications and is being adopted for high-level experiment control as beamlines and instruments are upgraded. Remote access is primarily through tools such as Windows Remote Desktop and NoMachine.

SSRL users can download data at their home institutions or on their personal computers using FTP servers or by direct download onto portable hard drives. Users also often utilize their own "cloud storage" accounts (e.g., Google Drive, Dropbox, etc.) to transfer data.

Going forward, the plan is to implement the Bluesky databroker/tiled software[7] for the initial data management layer to enable a data acquisition system consistent with the other light sources. This has been developed as a

---

[6] Allan, D., Caswell, T., Campbell, S., Rakitin, M. (2019) Bluesky's ahead: a multi-facility collaboration for an a la carte software project for data acquisition and management. Synch. Radiat. News 32(3), 19–22.

[7] https://blueskyproject.io/tiled/

collaborative open-source project with contributors mainly from other facilities, both within the DOE complex and internationally. This system aims to provide a consistent "data API" rather than prescribe a given on-disk data format.

These tools will serve as the basis for enabling searchable data catalogs and adopting FAIR data practices. The MDF and the DOE Office of Scientific and Technical Information will serve as a DOI generating service for SSRL data sets. In the long term, SSRL will adopt or develop tools that provide metadata catalogs, electronic logbooks, and sample tracking systems.

There is a high demand from the user community to integrate cloud services for both file transfer (e.g., Dropbox, Google Drive, etc.) and for communication (e.g., Slack) into all areas of the data lifecycle and compute workflows. We have seen this demand increase during the COVID-19 pandemic.

Imaging, high-energy, high-speed operando and multimodal techniques are generally the most computationally intensive techniques performed at the SSRL. Data reduction and analysis will eventually rely heavily on the use of HPC, utilizing appropriate technologies such as multithreading, General Purpose GPUs, edge devices, and distributed computing environments to obtain results with near real-time completion, so that results enable user-driven or even automated steering of experiments.

SSRL is focusing data analysis algorithm and software development in the areas needed to answer the novel scientific inquiries enabled by high-speed operando and multimodal experiments as SSRL instruments are being upgraded to allow these capabilities. Algorithms and software are being developed to analyze and reconstruct massive data volumes, bridge across length and time scales, identify and classify features and patterns, and provide feedback to experiments dynamically using real-time reduction and novel AI/ML approaches. Third-party vendors such as Modelyst[8] are being used to integrate multimodal data into a PostgreSQL database to facilitate querying, visualization and analysis of data collected using different modalities. The goal is to eventually integrate the methods developed here into the wider Bluesky suite to provide a consistent toolset to visualize and analyze data.

5.5.2.7 Network and Data Architecture

The SSRL network is composed of one data center, and it spans 20 buildings on the SLAC campus which are connected via a multimode and single-mode cable plant. The core and leaf infrastructure consist of mainly Cisco equipment.

External network connectivity to SSRL consist of 2x10 Gbps uplinks provided by SLAC networking, which obtains its external network connectivity via 100Gb connectivity from ESnet and Stanford University/Internet2.

---

[8] https://www.modelyst.com
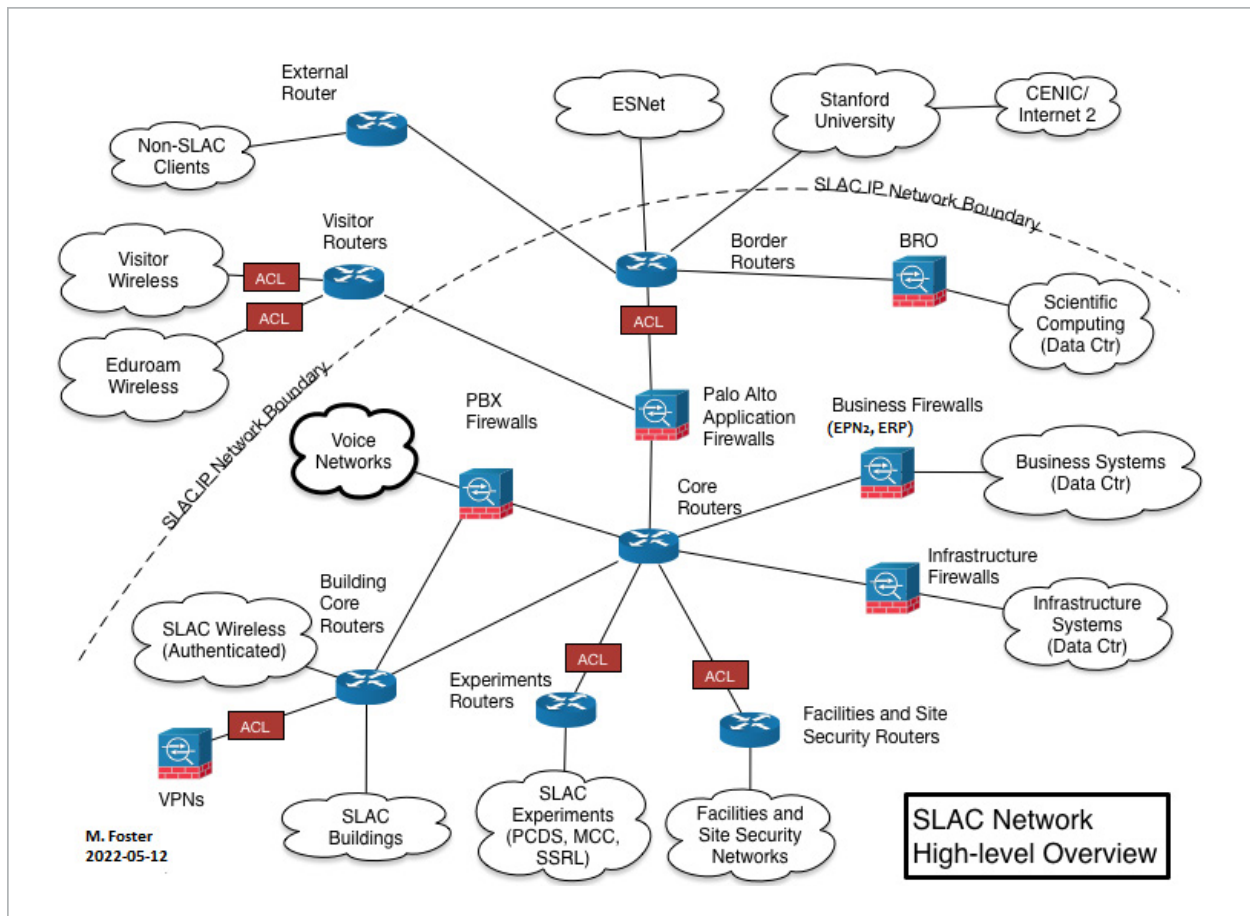
Figure 5.5.8: High-level Overview of SLAC Network

Primary network connectivity to SSRL is distributed between SLAC's B050 data center and SSRL, B120. Two 2x10 Gbps uplinks are provided over the same fiber optic trunk cable. Fiber redundancy for this interconnectivity between SLAC and SSRL currently does not exist. The fiber optic cable plant for SSRL starts in B120 and traverses ~20 buildings. Network connectivity to the SSRL enterprise and mission critical environments are segmented via dedicated network hardware. The network boundary for SSRL the enterprise network and mission critical environments consist of a Fortinet firewall. The firewall provides further segmentation for SSRL mission critical environments.

SSRL has a variety of distribution layer connection speeds that support the SPEAR3 accelerator, SSRL beamline data acquisition, and enterprise environments. The bandwidth ranges are 1Gbps, 10 Gbps, and 100 Gbps. Bandwidth is provided for each application accordingly.

SLAC WAN

The SSRL WAN network connectivity is provided by SLAC. SLAC has redundant 100 Gbps connections to its core routers from ESnet and Stanford University/Internet2.
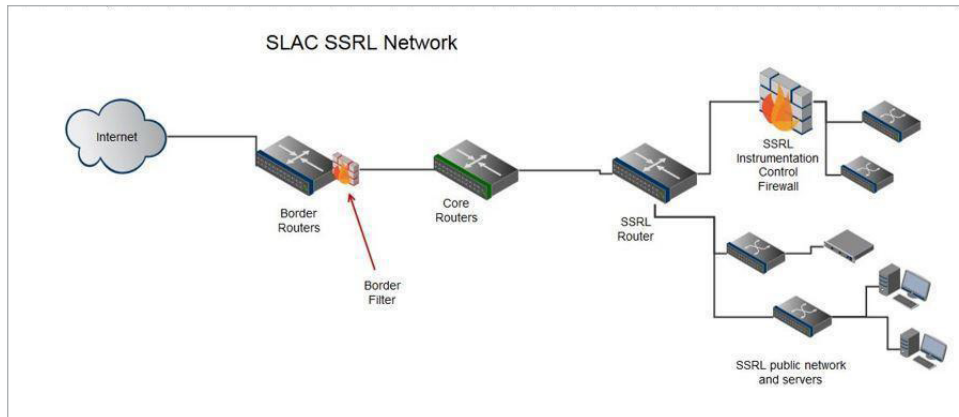
Figure 5.5.9: High-level Overview of SLAC Network

SLAC Core

SSRL core enterprise network is provided by 1 Cisco 6509. The 6509 provides network connectivity for the SSRL enterprise environment and 1x10 Gbps downstream link to the SSRL firewall for connectivity to the SPEAR accelerator and beam line data acquisition environment. The Core for the SSRL mission critical environments consists of a Fortinet firewall for routing and a variety of a Cisco Nexus 7009 and Cisco Nexus 5548 for LAN connectivity.

- **Present to two years:**
  - — Upgrade Core for SSRL enterprise networks from 1xCiso 6509 to 2xCisco Nexus9000 C9336C-FX2 Chassis
- **Two to five years:**
  - — Consider new modular data center for SSRL
  - — Rearchitect SSRL fiber optic plant
  - — Rearchitect SSRL core and switch infrastructure
- **Five+ years:**
  - — Replace core devices and possibly increase to terabit connectivity

SLAC LAN

SSRL LAN connectivity for the SSRL enterprise network is provided by a Cisco 6509 which consists of switch uplinks with speeds that vary between 2x1Gbps and 2x10 Gbps. SSRL LAN connectivity for the mission critical environment consists of speeds that vary between 2x10 Gbps and 2x100 Gbps for switch and server connectivity.

- **Present to two years:**
  - — Continual refresh of switches older than five to seven years
  - — Wireless access point upgrades
- **Two to five years:**
  - — Continual refresh of switches older than five to seven years
  - — Consider new modular data center for SSRL
  - — Rearchitect SSRL fiber optic plant
  - — Rearchitect SSRL core and switch infrastructure

- **Five+ years:**
  - — Deploy NextGen switching hardware
  - — Standardize 10 Gbps environment for desktops

SLAC Data Center

SSRL has one primary data center but has multiple server rooms spanned across multiple SSRL buildings. Cisco Nexus core distributions switches resides within these locations and provide connectivity from 10 Gbps to 100 Gbps for required services via ToR switches.

- **Present to two years:**
  - — Continual refresh for ToR switches older than five to seven years
- **Two to five years:**
  - — Consider new modular data center for SSRL
  - — Upgrade fiber optic cable plant
- **Five+ years:**
  - — Deploy NextGen routing and switching hardware

SSRL Modular Data Center Justification

SSRL needs a modern centralized data center to house computing and networking infrastructure. Currently SSRL has one ~720 sq. ft. official data center. In addition, three separate rooms are used to house additional computing and networking equipment. This configuration poses a management challenge due to the following factors: cooling, power, capacity, and redundancy. Each of these locations are at 95% capacity and limit large-scale growth. These limitations force divisions to find one-off solutions as needed. A data center ~5,000 sq. ft. would suffice for current and future needs at SSRL.

5.5.2.8 Cloud Services

Currently, we make limited use of Amazon AWS and Microsoft Azure cloud platforms in the areas of prototyping, testing, and standing up computational/storage resources to support training and demonstrations. There are several Software as a Services in production use, including Slack, Dropbox, GitHub Enterprise Cloud and Office 365.

SSRL, along with the other US DOE-funded light sources, continues to evaluate the capabilities and cost-effectiveness of using cloud services.

5.5.2.9 Data-Related Resource Constraints

SSRL is facing a multiple order-of-magnitude increase in demand for computing resources over the next decade. Data management and workflow tools are needed that integrate beamline instruments with computing and storage resources, for use during experiment, as well as facile user access for post experiment analysis. Real-time data analysis capabilities are required to significantly reduce data volumes and provide feedback during experiments to improve data quality and to drive the direction of ongoing measurements; the application of advanced mathematical algorithms, ML, and the integration of simulations and model-based approaches will allow automated steering of data collection. On-demand utilization of computing environments is required to enable near real-time data processing. Sufficient data storage and archival resources to house the continually increasing amounts of valuable scientific data produced by SSRL is required. Advances in networking serve as the basis to realize these critical capabilities.

5.5.2.10 Outstanding Issues

None to report at this time.

5.5.2.11 Facility Profile Contributors

SSRL Representation

- **Paul McIntyre,** *SLAC National Accelerator Laboratory,* pcm1@slac.stanford.edu
- **Vivek Thampy,** *SLAC National Accelerator Laboratory*, vthampy@slac.stanford.edu
- **Christopher S. Ramirez,** *SLAC National Accelerator Laboratory*, ramirez@slac.stanford.edu
- **Cathy J. Knotts,** *SLAC National Accelerator Laboratory*, knotts@slac.stanford.edu
- **Jana Thayer,** *SLAC National Accelerator Laboratory*, jana@slac.stanford.edu
- **Stuart Campbell,** *BNL,* scampbell@bnl.gov
- **Nicholas Schwarz,** *ANL,* nschwarz@anl.gov
- **Alex Hexemer,** *LBNL,* ahexemer@lbl.gov

ESCC Representation

- **Mark Foster,** *SLAC National Accelerator Laboratory,* fosterm@slac.stanford.edu

5.6 High Flux Isotope Reactor (HFIR) and Spallation Neutron Source (SNS)

The SNS and the HFIR are both DOE BES Scientific User Facilities at ORNL. These facilities are operated to support research using neutrons in a variety of science areas including structural biology, crystalline materials, polymers, magnetism and superconductivity, chemistry, and disordered materials.

5.6.1 Discussion Summary

- The SNS and the HFIR are both DOE BES Scientific User Facilities at ORNL.

- Between the two user facilities there are 2,000 people who wish to access data remotely per year, and typical beam time requests range from 2–16 days.

- HFIR and SNS provide researchers with two suites of neutron scattering instruments and beam lines. The instruments are supported by a variety of sample environments and data analysis and visualization capabilities. Each instrument uses a common network and computing architecture.

- A remote login infrastructure provides remote users access to both data and computing resources. During the pandemic, remote user capability was added to the instrument systems. This capability has been deployed to 24 of 32 instruments and gives remote users the ability to control the instrument as if they were physically located at the instrument hutch.

- Remote users login through the analysis cluster and then if they meet certain criteria, they can remote into beamline control computers. In order to do so users must be on the current experiment proposal, trained for remote experiments on the instrument, and remote experiments must be enabled for that beamline.

- Data acquisition collects all aspects and metadata and stored in an HDF-5 data file using the NeXus data schema. A single measurement is called a "run", and experiments are composed of many runs. Collections of runs may be grouped into a "reduced" data set. Stored procedures are executed on computing clusters at the end of every "run" such that reduced data can be provided to the users as soon as possible following the end of data collection.

- SNS data sets tend to be larger than those produced at HFIR; for the SNS, "runs" range from tens of MBs to 4 GBs. The highest data rate SNS instruments are capable of producing data sets approaching 1 TB. The imaging and tomography beam lines have potential to produce even larger data sets, particularly at the new instrument under construction at the SNS.

- The maximum data rate for SNS ranges from 33 KBps to 157 MBps across the instruments. The data flows through the cached storage and into the shared analysis file system according to the user's choices on utilizing the instrument.

- The SNS-FTS is constructing the VENUS imaging instrument. It will require 10Gb networking to support the expected data volumes; initial estimates are to generate 20–30 TB of raw data per day. Reduced data will be a fraction of that amount, but the intention is to keep all raw data. The next-generation detector in development (5+ years) will have approximately 16 times the data requirements.

- Recent upgrades to the NOMAD instrument established 10Gb networking and streamlined data processing software to handle the increased data rates of 200 MBps.

- The SNS STS will add a new suite of instruments to the SNS during the next 10 years. The data rates and storage needs are not well defined at present, but it is safe to assume that the STS will add another 30% to networking and storage requirements over the current FTS plus another VENUS-like instrument.

- The facility data network is internal to ORNL. Data flows from the instruments to a centralized data management system and is processed either locally on facility computers or on other ORNL computers.

- SNS provides shared analysis computing resources for its users via a cluster of computing resources and software. These computing resources are available while the user is on-site conducting the experiment and remain available for remote login afterward.

- Software systems enable the use of the on-premise edge computing, the Compute and Data Environment for Science (CADES) and the computing facilities provided by the OLCF.

- Data produced during the approved beam time for an experiment is held for an indeterminate time at the facility storage system. Currently, all data produced since the SNS was constructed has been stored. For HFIR, all data produced on instruments since approximately 2012 is stored.

- Users may use scp or sftp to transfer their experiment data to their own computers for analysis at their convenience.

- The facilities are considering cloud resources in lieu of the tape storage system, but have not decided on this yet. Users may use cloud servers in their analysis workflows, but do so outside of the standard workflow pattern.

- ORNL connects to ESnet via redundant border routers at 100 Gbps. The expectation is that these connections will soon be upgraded to 400G connections.

- ORNL utilizes a Science DMZ architecture for high-performance data transfer. Globus is the approved transfer method. A border perfSONAR node is connected to the border router and participates in ESnet perfSONAR testing.

- To date, Science DMZ, perfSONAR and Globus have not been incorporated into the network and data architecture of HFIR and SNS and, because the use of CADES and OLCF machines are not currently part of the analysis workflow.

5.6.2 HFIR and SNS Facility Profiles

Each facility consists of two parts: a neutron production part and an instrument part. Neutron production at HFIR is performed by a uranium nuclear reaction whose reactor vessel is designed for fast refueling and currently operates 7 fuel cycles per year. Neutron production at the SNS is performed by producing a burst of protons that are guided to a flowing liquid mercury target at 60Hz repetition rate; the SNS operates two cycles per year, allowing for maintenance activities of the neutron production part. The pulsed nature of the SNS allows for time-resolved measurements that provide visibility into dynamic processes within materials. The SNS contains 18 instruments in the user program plus a fundamental physics instrument operated by university collaborators. HFIR contains 12 instruments in the user program plus a development instrument and a sample alignment station.

5.6.2.1 Science Background

Each instrument is constructed with a unique topology that optimizes for some aspect of the physical process that is to be investigated by the researchers operating it. There are spectrometers which measure composition and states of matter in materials, diffractometers that obtain information of the structure, spacing and organization of crystalline materials, and reflectometers that probe the interfaces of solid/solid, solid/liquid or free liquid layered materials. Experiments are typically performed by placing a representative sample of the material of interest within the direct neutron beam of an instrument. The neutrons interact with the sample material, and a portion of these neutrons is scattered and detected by the instrument detector systems. These scatter patterns are analyzed with an understanding of the specific geometry of the instrument, the position of the sample w/r/t the neutron beam and the knowledge of sample environment and the state of the instrument (slits, collimators and neutron

beam characteristics). The instruments are designed and optimized to statistically measure a range of size scales from atomic through molecular to macromolecular structures.

Data acquisition collects event data showing the position of the neutron on the detector array and its time of flight (energy), along with metadata for the sample's environment (temperature, pressure, electric or magnetic field, etc.) and instrument configuration (neutron choppers, motion controls, etc.). The data is processed to produce an HDF-5 data file using the NeXus data schema[9]. A single measurement is called a "run," and experiments are composed of many runs. The data acquisition system ensures that the data file for a run is collected concurrently with the experiment. Collections of runs may be grouped into a "reduced" data set. Stored procedures are executed on computing clusters at the end of every run such that reduced data can be provided to the users as soon as possible following the end of data collection. Experiment automation systems ensure that the instruments are configured for the next data collection state so as to maximize the active data collection time during an experiment. Data collected prior to reduction is cached locally on the instrument data systems to allow rereduction of runs if certain administrative conditions (wrong experiment or sample data entered by users, exceptions raised in the stored procedures for reduction or other errors) occur. SNS data sets tend to be larger than those produced at HFIR; for the SNS, runs range from tens of MBs to 4 GBs. The highest data rate SNS instruments are capable of producing data sets approaching 1 TB. The imaging and tomography beam lines have potential to produce even larger data sets, particularly at the new instrument under construction at the SNS.

The facility data network is internal to ORNL. Data flows from the instruments to a centralized data management system and is processed either locally on SNS computers or on other ORNL computers. SNS provides a cluster of computing resources and software for users to process and analyze their data. These computing resources are available while the user is on-site conducting the experiment and remain available for remote login afterward. This is because the instrument scientists who serve the users collaborate with them to design the analysis workflows as they are highly instrument specific and dependent on the experiment that is performed.

5.6.2.2 Collaborators

Between the two user facilities there are 2,000 people who wish to access data remotely per year. That is 2,000 unique logins to long-term data storage and/or the analysis cluster. There are approximately 16,000 user accounts in total. At any point in time there are approximately 250 users logged in. They come from national and international research facilities and they have varied computational skill sets and resources at their home institution. Use of the instruments is proposal-driven and governed by a Science Review Committee which reviews proposals in subcommittees to discuss and rank the proposals with many high-quality proposals not making the cutoff due to oversubscription. The figure below shows the global distribution of users. Beam time requests range from 2–16 days.

---

[9] https://www.nexusformat.org
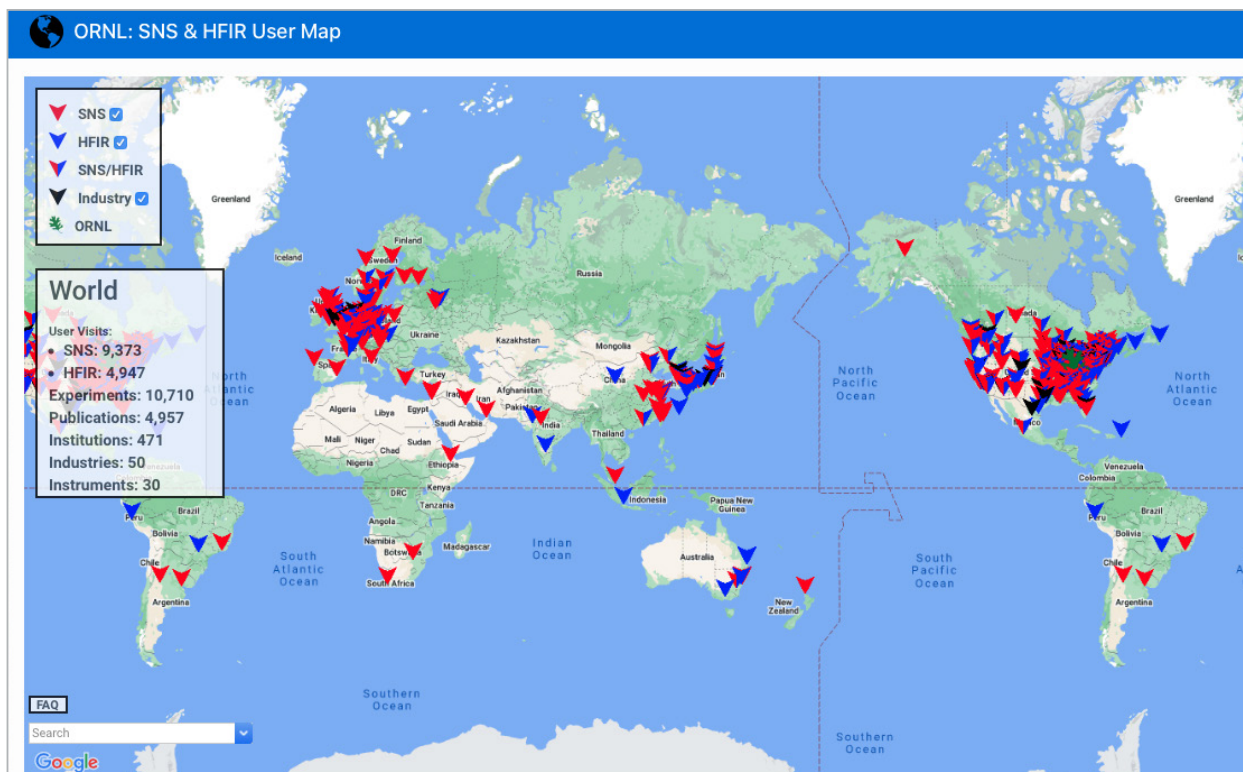
Figure 5.6.1: SNS and HFIR User Map Provided by the NScD User Office

Data produced during the approved beam time for an experiment is held for an indeterminate time at the facility storage system. Currently, all data produced since the SNS was constructed has been stored. For HFIR, all data produced on instruments since approximately 2012 is stored. Users may use scp or sftp to transfer their experiment data to their own computers for analysis at their convenience.Each instrument uses a common

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
HFIR USERS	secondary	data transfer	22GB	ad hoc	N	N/A
SNS USERS	secondary	data transfer	1TB	ad hoc	N	N/A

Table 5.6.1: HFIR and SNS Collaboration Space

5.6.2.3 Instruments and Facilities

HFIR and SNS provide researchers with two complementary world-class suites of neutron scattering instruments and beam lines. The instruments are supported by a variety of sample environments and data analysis and visualization capabilities[10].

---

[10] https://neutrons.ornl.gov/instruments
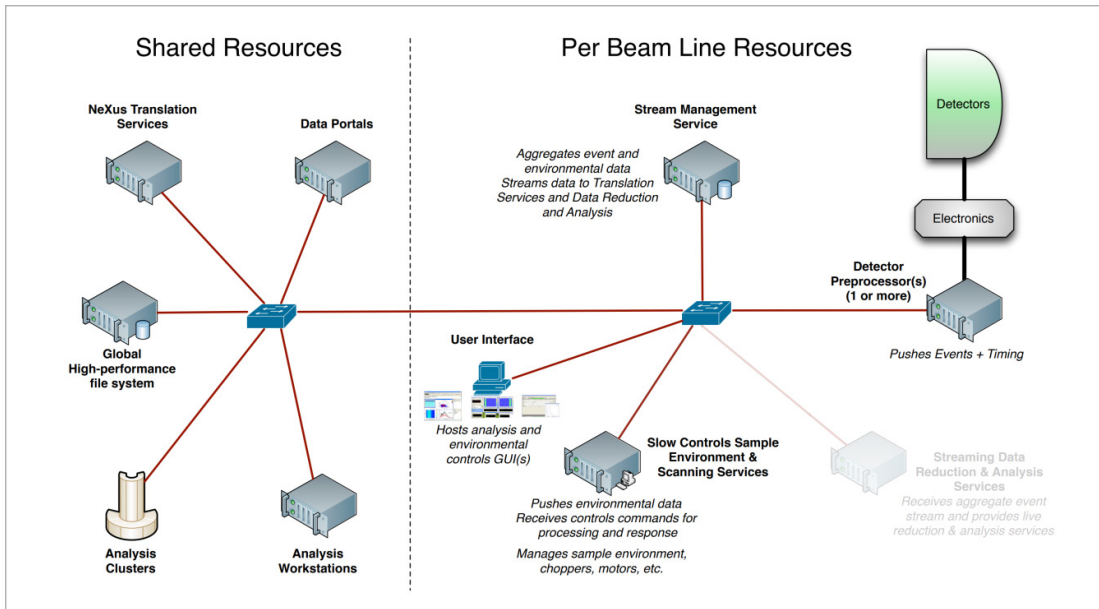
Figure 5.6.2: General Per-Beamline Computing Resource Diagram

Each instrument uses a common network and computing architecture. The network address space is ordered with each instrument occupying one routed class-C address space accessible from the facility network and one private class-C address space containing sample environment and control equipment accessible only by the per-beamline resources (which are further described in Section 5.6.2.7). As shown in the diagram above, data produced by the detectors is fused and time-ordered with the instrument slow-controls process values taken during the experiment. This data is stored on the Streaming Management Service computer which holds between 20–80 TB of cached storage of prereduced data as described in Section 5.6.2.1.

Display Name	Sample Co...	Mean	Median	Standard Deviation	Min Value	Max Value
🟥 RTBT:Diag:BCM25:Power1_KW	9756	1265.6926629199183	1400.635559510687	389.49370332254784	0.0	1413.580582947173
🟦 BL1B:Det:N1:Det1:EventRate_RBV	7113	674756.6950778328	473754.5984848485	1269047.1916842775	0.0	1.9630226244094487E7
🟩 BL3:Det:N1:Det1:EventRate_RBV	8000	33868.87270623871	17010.438823630284	68051.99823428123	0.0	452247.037037037
🟧 BL5:Det:N1:Det1:EventRate_RBV	7318	6226.156299908173	2673.3330962541486	11944.072940796792	0.0	88831.29629629629
🟦 BL6:Det:N1:Det1:EventRate_RBV	9130	26826.530106608754	1478.9212962962963	63576.553130718676	0.0	801043.0825688073
🟦 BL7:Det:N1:Det1:EventRate_RBV	6956	4335.031140996885	2356.1903021442495	4390.915509981236	0.0	43919.3
🟦 BL9:Det:N1:Det1:EventRate_RBV	4922	104673.648455686	44149.6867816092	200162.77895254767	0.0	892141.5617977529
🟦 BL11A:Det:N1:Det1:EventRate_RBV	8528	2873.3126257963936	1829.2153996101365	3075.0367197756686	0.0	16975.40740740741
🟦 BL11B:Det:N1:Det1:EventRate_RBV	8919	694.5418996719058	504.3611111111111	2092.948480596513	0.0	39888.398148148146
🟨 BL12:Det:N1:Det1:EventRate_RBV	8709	43731.31255267443	31060.99074074074	54556.8060036022	0.0	273461.27777777775
🟦 BL14B:Det:N1:Det1:EventRate_RBV	8808	118.29898194726606	20.645507893214315	259.6482243945973	0.0	4154.958333333333
🟦 BL16B:Det:N1:Det1:EventRate_RBV	5481	4657.6448041197145	1740.1781609195402	7084.0689386629465	0.0	64837.794444444444
🟦 BL17:Det:N1:Det1:EventRate_RBV	5639	10539.244605744732	1434.0114942528735	37996.407390068176	0.0	1599761.0416666667
🟩 BL18:Det:N1:Det1:EventRate_RBV	5584	11094.537271317944	3439.7018199233717	128109.45325221999	0.0	2709649.5

Table 5.6.2: SNS Event Data Rate for the FY22-A RunCycle (events per sec–8 Bytes per event)

Data rates of the SNS instruments are characterized in the Table above. The maximum data rate ranges from 33 KBps to 157 MBps across the instruments. The data flows through the cached storage and into the shared analysis file system according to the user's choices on utilizing the instrument. As described in Section 5.6.2.1, the conversion to the NeXus format and transfer to the analysis file system occurs concurrently with the data collection according to the workflow described in Section 5.6.2.4.

The SNS-FTS is constructing the VENUS imaging instrument. It will require 10Gb networking and will dwarf current experiment data sizes. In the initial build, it will generate 20–30 TB of raw data per day on a Multichannel Plate (MCP) detector comprising a 2x2 array of "Timepix3" sensors. Reduced data will be a fraction of that amount, but the intention is to keep all raw data initially until neutron clustering and AI reduction algorithms are vetted. The next-generation detector in development (in five years) will have approximately 16 times the data requirements, as it anticipates using an 8x8 array of Timepix4 MCP sensors. Recent upgrades to the NOMAD instrument which fully built out the detector array have also pushed bandwidth and data processing capabilities. The recent upgrade to the instrument established 10 Gb networking and streamlined data processing software to handle the increased data rates generated by NOMAD. The desired data rate at NOMAD is sustained 25M neutron events per second or 200 MBps. The proton power upgrade at the SNS will increase neutron event rate roughly 40% at all instruments and storage requirements will increase correspondingly.

The SNS STS will add a new suite of instruments to the SNS during the next 10 years. The data rates and storage needs are not well defined at present. The STS will receive 25% of a 2.8MW beam or roughly 0.7MW, half of what current instruments receive now on SNS instruments. The numbers of instruments to be built at STS are in the 10–12 range with an imaging instrument in the mix. It is safe to assume that the STS will add another 30% (half of 60%) to networking and storage requirements over the current FTS plus another VENUS-like instrument.

HFIR plans to change out the beryllium jacket on the reactor within five years and will be refreshing the instrument suite with minor upgrades during this time. The instruments in the cold guide hall (See the HFIR Instrument Layout figure below) will be reconstructed as the hall will be expanded and made larger. A neutron spin-echo instrument is planned to be added. The IMAGING beamline will receive the same Timepix3 MCP detector as VENUS. All other instrument data rates and storage needs are well-known, our current capabilities meet these requirements.

SNS provides shared analysis computing resources for its users. A typical analysis cluster computer has 32 cores (3.4GHz) and 1 TB of RAM. A computing cluster is available for processing large data sets, we have 8 Analysis machines in the cluster. A typical analysis cluster computer has 32 cores (3.4GHz) and 1 TB of RAM. Some beamlines have their own dedicated analysis machine physically located at the beamlines. The centralized data storage is a Quantum Stornext File system with 2.2 PBs of data storage and a backup tape file system with 4.4 PBs of data storage. The tape system is designed to be big enough for 2 copies of all data, which can be maintained at separate geographic locations. A remote login infrastructure provides remote users access to both data and computing resources. During the pandemic, remote user capability was added to the instrument systems. This capability has been deployed to 24 of 32 instruments and gives remote users the ability to control the instrument as if they were physically located at the instrument hutch.

SNS	Average File Size	Number of Experiments	Number of Directories	Total Data set Size (GB)	Anticipates File Size Increase 2–5 years roughly 40% increase in file size & bandwidth due to 1.4MW– 2.0MW Proton Power Upgrade	Strategic Planning (Beyond 5 years)
ARCS	65 MB	119792	316	8217.3	40%	Same as 2–5 years
BASIS	43 MB	73769	182	3372.8	40%	Same as 2–5 years
CNCS	32 MB	198795	334	6745.4	40%	Same as 2–5 years
CORELLI	232 MB	243629	522	59338.9	40%	Same as 2–5 years
EQSANS	320 MB	49626	548	16663.7	40%	Same as 2–5 years
FNPB	*				40%	Same as 2–5 years
HYSPEC	49 MB	291353	470	15084.1	40%	Same as 2–5 years
LIQREF	13 MB	56247	270	794.9	40%	Same as 2–5 years
MAGREF	38 MB	11206	158	453.1	40%	Same as 2–5 years
MANDI	1126 MB	2708	156	3199.8	40%	Same as 2–5 years
NOMAD	3178 MB	94999	1376	316640.5	40%	Same as 2–5 years
NSE	**				40%	Same as 2–5 years
POWGEN	338 MB	18121	1128	6423.7	40%	Same as 2–5 years
SEQUOIA	136 MB	209078	672	30021.2	40%	Same as 2–5 years
SNS NEUTRONS AND PRESSURE (SNAP)	1200 MB	14937	250	18803.5	40%	Same as 2–5 years
TOPAZ	1390 MB	13816	252	20146.1	40%	Same as 2–5 years
USANS	59 MB	38009	220	2360.8	40%	Same as 2–5 years
VENUS	Currently being built	N/A			25 TB/Day	200 TB/Day Eventually
VISION	4102 MB	52824	704	227244.0	40%	Same as 2–5 years
VULCAN	109 MB	67908	450	7778.0	40%	Same as 2–5 years

*   Fundamental Physics Beamline is not managed by Neutron Sciences
** Just recently transitioned to NScD ORNL management, Data is stored separately

Table 5.6.3: SNS Instrument Run Sizes

SNS	Average File Size	Number of Experiments	Number of Directories	Total Data set Size (GB)	Anticipates File Size Increase 2–5 years roughly 40% increase in file size & bandwidth due to 1.4MW–2.0MW Proton Power Upgrade	Strategic Planning (Beyond 5 years)
BIO-SANS	427 MB	51550	263	4499.2	No major increases anticipated in 2–5 years	Same as 2–5 years
CTAX	16 KB	31920	459	0.6	No major increases anticipated in 2–5 years	Same as 2–5 years
DEMAND	14 KB	105960	878	1.6	No major increases anticipated in 2–5 years	Same as 2–5 years
DEV BEAM	*			6644.0	No major increases anticipated in 2–5 years	Same as 2–5 years
FIE-TAX	13 KB	51678	670	0.7	No major increases anticipated in 2–5 years	Same as 2–5 years
GP-SANS	85 MB	51550	263	4499.2	No major increases anticipated in 2–5 years	Same as 2–5 years
HIDRA	117 MB	3181	124	384.1	No major increases anticipated in 2–5 years	Same as 2–5 years
IMAGINE	**			63.0	No major increases anticipated in 2–5 years	Same as 2–5 years
IMAGING				60421.0	25 TB/Day	Same as 2–5 years
POWDER	61 KB	16769	1028	1.1	No major increases anticipated in 2–5 years	Same as 2–5 years
PTAX	15 KB	49563	556	0.8	No major increases anticipated in 2–5 years	Same as 2–5 years
TAX	13 KB	43014	559	0.6	No major increases anticipated in 2–5 years	Same as 2–5 years
WAND²	12 MB	875234	446	11114.8	No major increases anticipated in 2–5 years	Same as 2–5 years

* Files maintained by development scientists
** Files maintained separately, unique control system

Table 5.6.4: HFIR Instrument Run Sizes
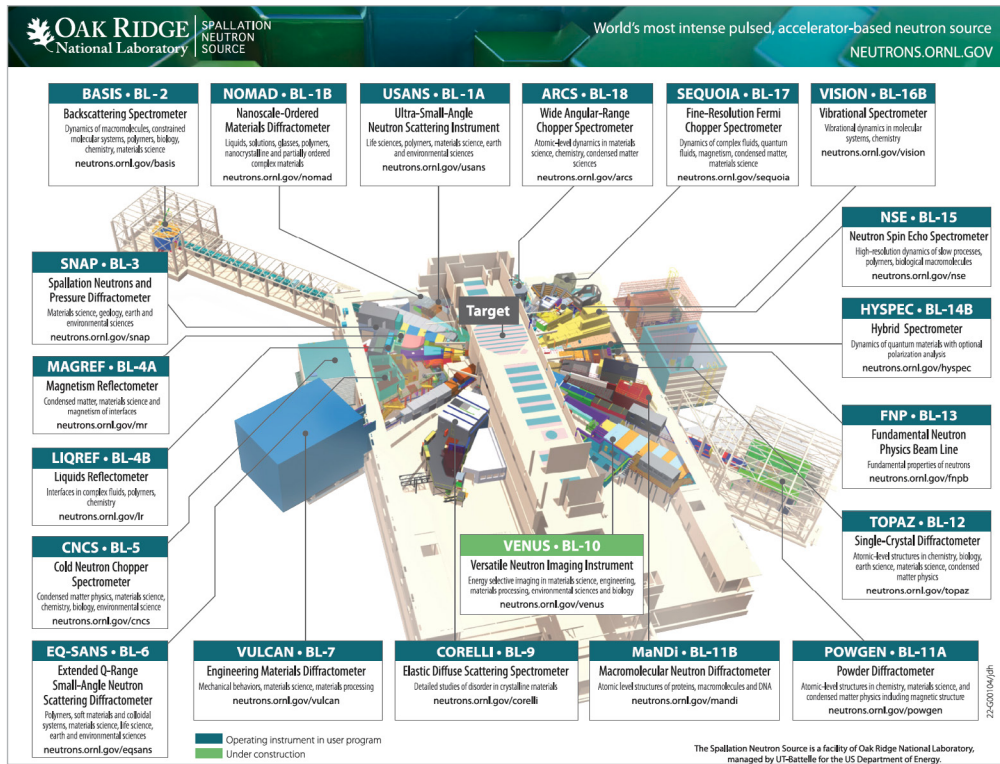
Figure 5.6.3: SNS Instrument Layout
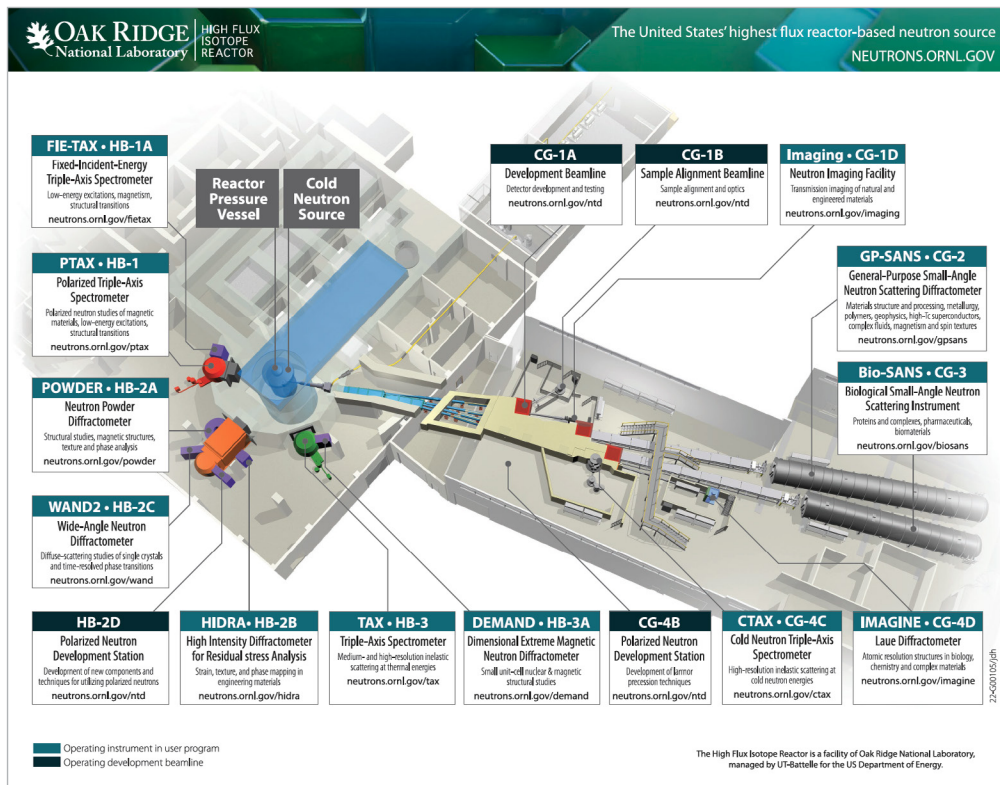
Figure 5.6.4: HFIR Instrument Layout

5.6.2.4 Generalized Process of Science

As referenced in Section 5.6.2.3, the workflow for instrument data acquisition is implemented by a software system named Accelerating Data Acquisition and Analysis (ADARA[11]) whose open-source code is distributed via its github portal[12].

ADARA Architecture & Components - Overview
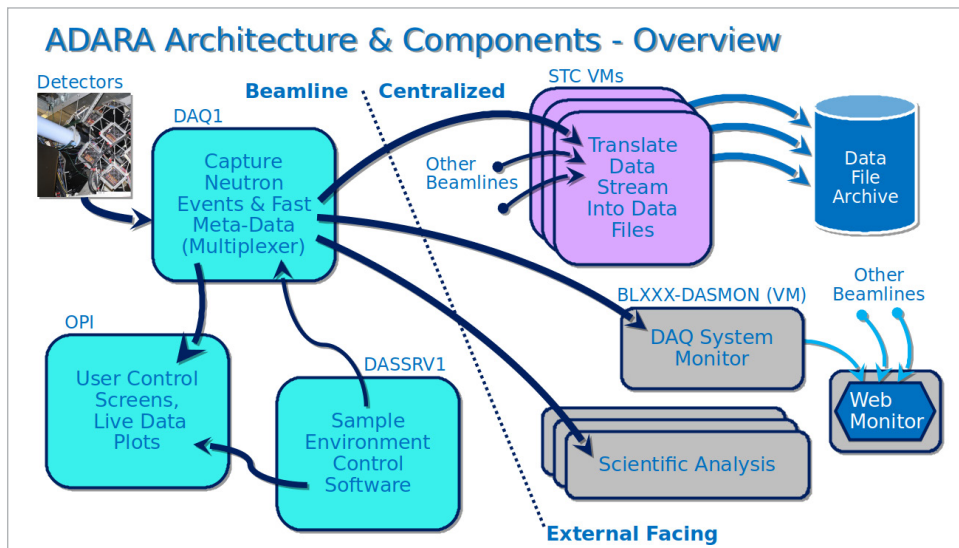
Figure 5.6.5: General Data Workflow Diagram

This workflow aggregates the detector electronics producers, fast metadata and sample-environment control data. Detector electronics produce tuples of (position, time) for all neutron counts. Fast metadata represents various processes that are sampled, time resolved, and collated into waveforms that can be understood at the same time scale as neutron data. Examples of this are: data from beam monitors that provide a measure of incident neutron flux on a sample, data from top-dead center sensors of neutron beam choppers that control the bandwidth and incident energy of the neutron beam, and experiment-specific signals collected by various systems such as potentiostat, rheometers, pulsed-magnet controllers. The experiment-specific signals allow subsequent analysis of the data collected to be interpreted according to the electrical, physical or magnetic conditions affecting the sample in a dynamic way.

In order to meaningfully use the data collected, neutron counts must be translated from instrument coordinates (two-dimensional detector positions) into physical coordinates (three-dimensional positions relative to the sample centroid) and subsequently into scientific coordinates (momentum transfer space or lattice parameters for example). The translation into physical coordinates is performed automatically as part of the NeXus file generation using Instrument Definition Files (IDF) that are produced by careful surveys of the precise physical location of the detector and sample positions upon instrument configuration changes. Each instrument maintains a revision-controlled version of the IDF files which allows retranslations to occur from the data collected at any time since the instrument operation started. The conversion into scientific coordinate systems and subsequent modeling is highly instrument and experiment specific. These are performed by users as guided by the instrument scientists using the analysis computational resources, taking as their inputs the experiment data collected. On certain instruments a framework called MANTID Workbench[13] is used to provide access to common algorithms used at several neutron scattering facilities worldwide. This framework is used both for

---

[11] https://ieeexplore.ieee.org/document/6972268
[12] https://github.com/ADARA-Neutrons/adara
[13] https://www.mantidproject.org

auto-reduction processes which execute automatically following the completion of data collection on the NeXus files as well as by users and instrument scientists to visualize and interpret the experiment data collected before performing further modeling and experiment-specific analysis.

Users will run model refinement codes iteratively to validate their models against experimental results. This work may occur on the facility analysis cluster close to the stored data sets, or it may be performed on users' computational resources either on their own portable computers or at their home facilities.
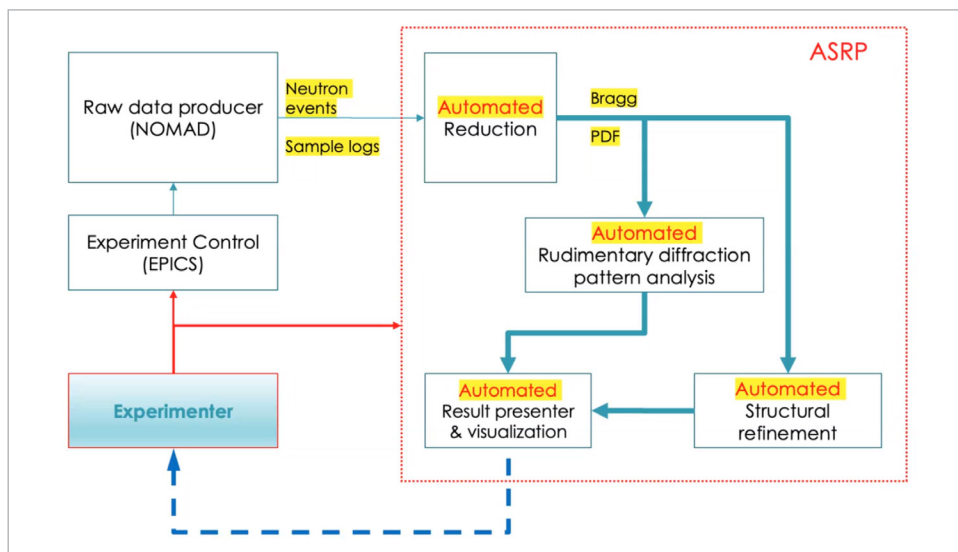
Figure 5.6.6: Future enhanced analysis workflows

There are current projects ongoing to study and develop enhanced automated analysis workflows. These projects seek to take advantage of the diverse networking and computational capabilities available at the ORNL site. The basic components of this effort are shown in the below diagram.
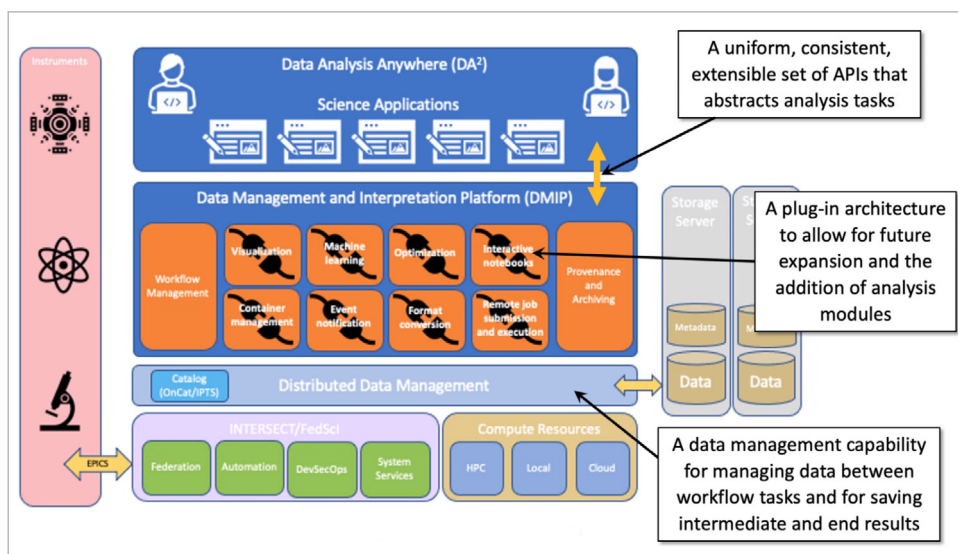
Figure 5.6.7: Vision for a Neutron Data Analysis and Interpretation Capability

The basic principle is to chain data artifacts from one stage of the workflow into the next. Each stage of the workflow is a plug-in containerized instance with provenance and tracking, that is ideally able to be regenerated as needed from software specifications (think: dockerfiles, singularity definition files, etc.). These workflows envision invocations of commercial software alongside open-source, facility- and user-written codes. When implemented in the next two to five years, these software systems will enable the use of the on-premise edge computing, the CADES[14]) and the computing facilities provided by the OLCF[15]).

Beyond the five-year timeframe, the next frontier is the integration of hyperconverged infrastructure electronics components that will allow for real-time analysis of experiments with the simultaneous comparison to the results of science modeling codes that are currently performed over several months after experiment data collection. These edge-computing resources move closer to the data sources and so have an opportunity to transform the nature of experiments conducted at the facilities by providing immediate visualization of experiments in the scientific coordinate systems in real-time as data is being collected while simultaneously rendering the results of the users' computational models using the parameters extracted from the immediate analysis of the experimental data being collected.
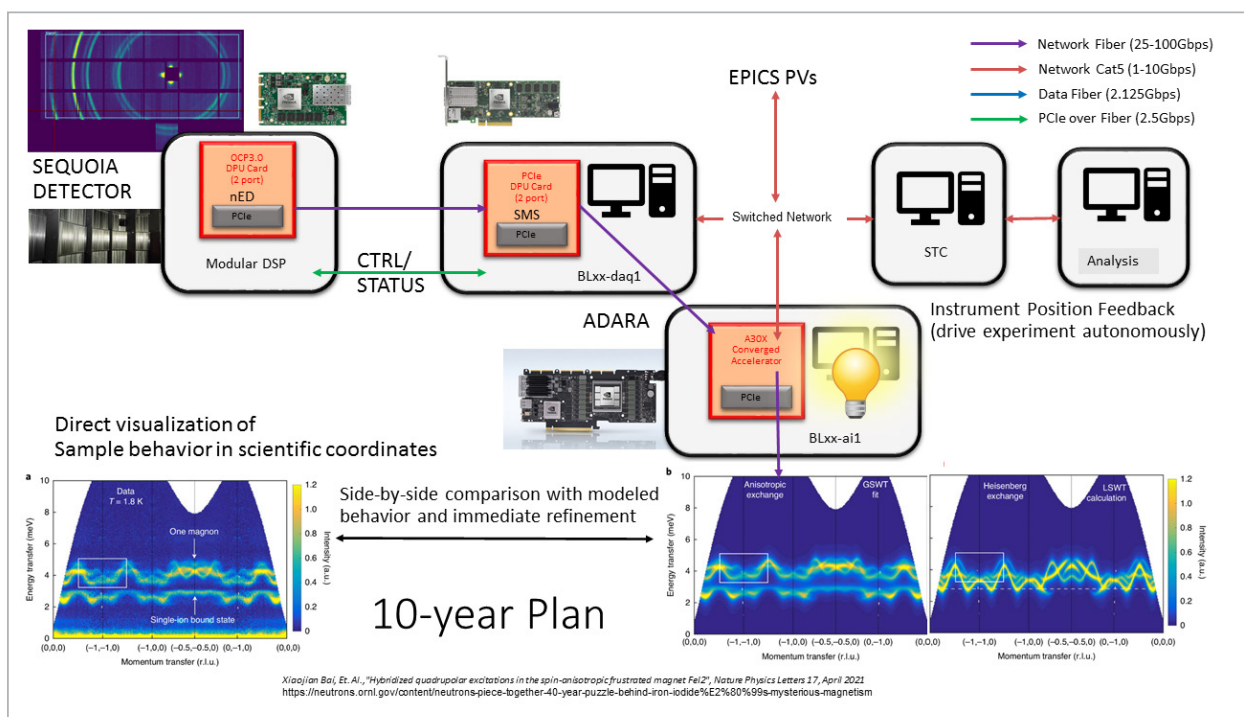
Figure 5.6.8: Vision for a Real-Time Hyperconverged Infrastructure for Knowledge Discovery

5.6.2.5 Remote Science Activities

All users can login to the analysis cluster remotely to analyze their data. Remote experiment capability is available on 24 of 32 instruments. This gives the user the ability to control the instrument as if they were at the instrument hutch from anywhere in the world. Remote users login through the analysis cluster and then if they meet certain criteria, they can remote into beamline control computers. In order to do so users must be on the current experiment proposal, trained for remote experiments on the instrument, and remote experiments must be enabled for that beamline. Multifactor authentication is being rolled out this summer.

The basic principle is the use of a browser based remote desktop software (ThinLinc[16]). This software is integrated with the LDAP[17]) used to authenticate users both internal and external. This software allows a remote user to gain access to a remote operator interface (OPI) machine. This OPI operates in parallel to the beamline OPI so that remote and local users see the same view of the instrument. An OPI is merely a user interface layer to a distributed process variable system that is described in Section 5.6.2.6.
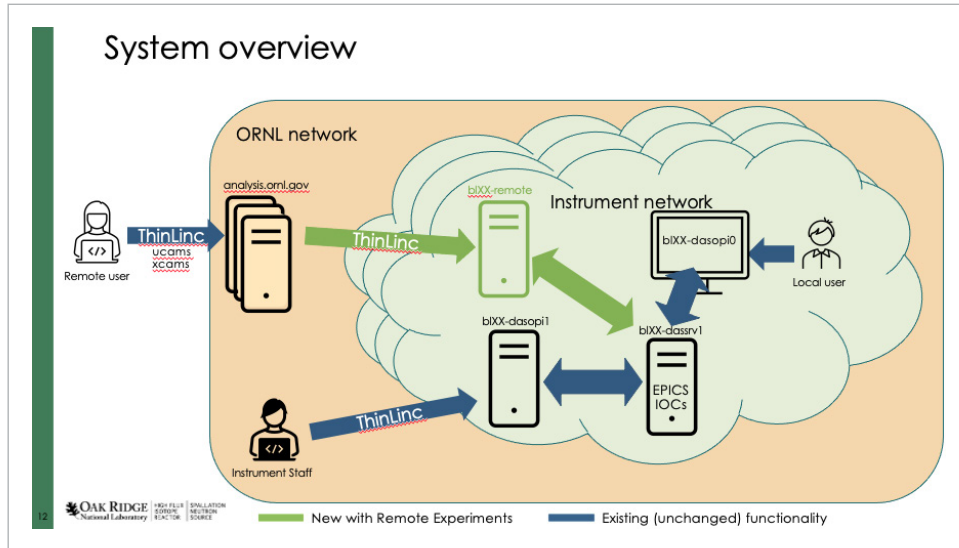
Figure 5.6.9: Overview of connections for Remote Experiments

5.6.2.6 Software Infrastructure

The EPICS) is used for instrument control on 24 of the 32 instruments at the SNS and HFIR combined. EPICS is a toolkit that enables the creation of distributed control systems over standard networking via a process variable distribution protocol interconnecting many Input/Output Controllers (IOC). The IOC is the fundamental unit of control of hardware whose parameters are exposed as process variables. The IOCs operate independently from each other, and through their exchange of process variables over the network and defined

There are 7 instruments at HFIR which run the Spectrometer Instrument Control Environment (SPICE), a LabVIEW-based control system. The SPICE instruments are scheduled to be converted to EPICS over the next several years.

The MANTID framework is used for analyzing the data, subject to the discussion and roadmap described in Section 5.6.2.4.

5.6.2.7 Network and Data Architecture

ORNL WAN Networking

ORNL connects to ESnet via redundant border routers, each of which currently connects to a diverse ESnet router at 100G. The expectation is that these connections will soon be upgraded to 400G connections. The ORNL border routers connect the ORNL Enterprise network, which includes SNS, CNMS and HFIR, with the OLCF and ESnet. This connectivity is depicted in Figure 5.6.10.

---

[16] https://www.cendio.com
[17] https://datatracker.ietf.org/doc/html/rfc4510
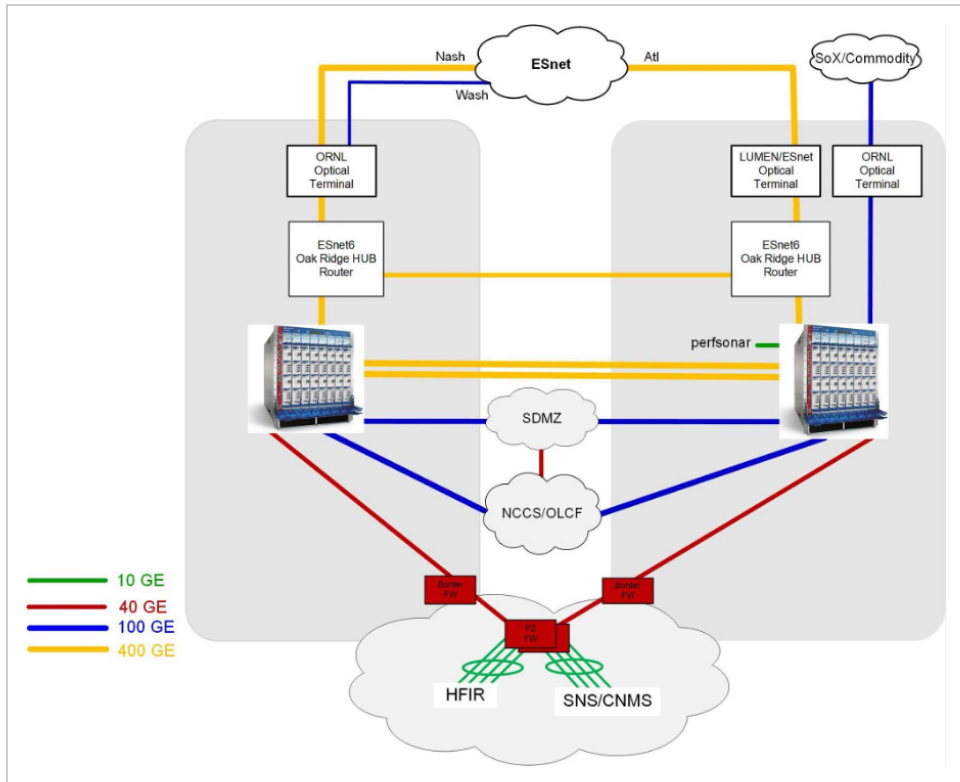
Figure 5.6.10: ORNL WAN Connectivity

The Enterprise firewalls are connected to the border routers at 40G currently. Connections to the Enterprise core from SNS/CNMS and HFIR are also 40G currently. Upgrades to these core switches are in progress and uplinks from SNS/ complex materials scattering (CMS) are expected to become 100G early in FY23.

ORNL does utilize a Science DMZ architecture for high-performance data transfer. This environment connects to the border routers with 10/40/100G DTN connections available. Globus is the approved transfer method. A border perfSONAR node is connected to the border router and participates in the ESnet grid.

EPICS Networking
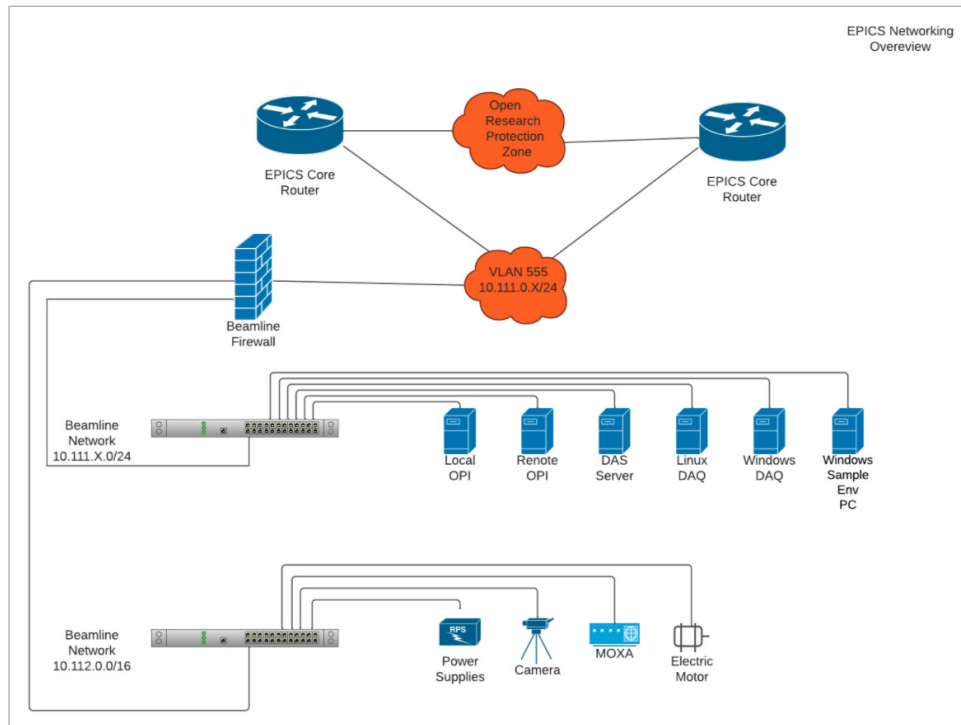
Figure 5.6.11: EPICS Beamline Network Diagram

As referenced in Section 5.6.2.3, the instrument control network consists of two segments: the server network that has full access to the open research network at ORNL through core routers. Each beamline network is isolated by a firewall to secure the beamline computers while allowing machines on the research network read-only access to the instrument state. This is convenient to allow any workstation in the facility to visualize instrument state and perform historical queries and check alarm conditions. The network links are commonly 1G links, with some beamlines having 10G links between the router and the facility networking.

To date, Science DMZ, perfSONAR and Globus have not been incorporated into the network and data architecture because the use of CADES and OLCF machines are not part of the analysis workflow at the moment. As described in Section 5.6.2.6, these resources are being developed for automated analysis, and so these components are likely to become used in the two-to five-year timeframe.

5.6.2.8 Cloud Services

We are considering cloud resources in lieu of the tape storage system, but have not decided on this yet.
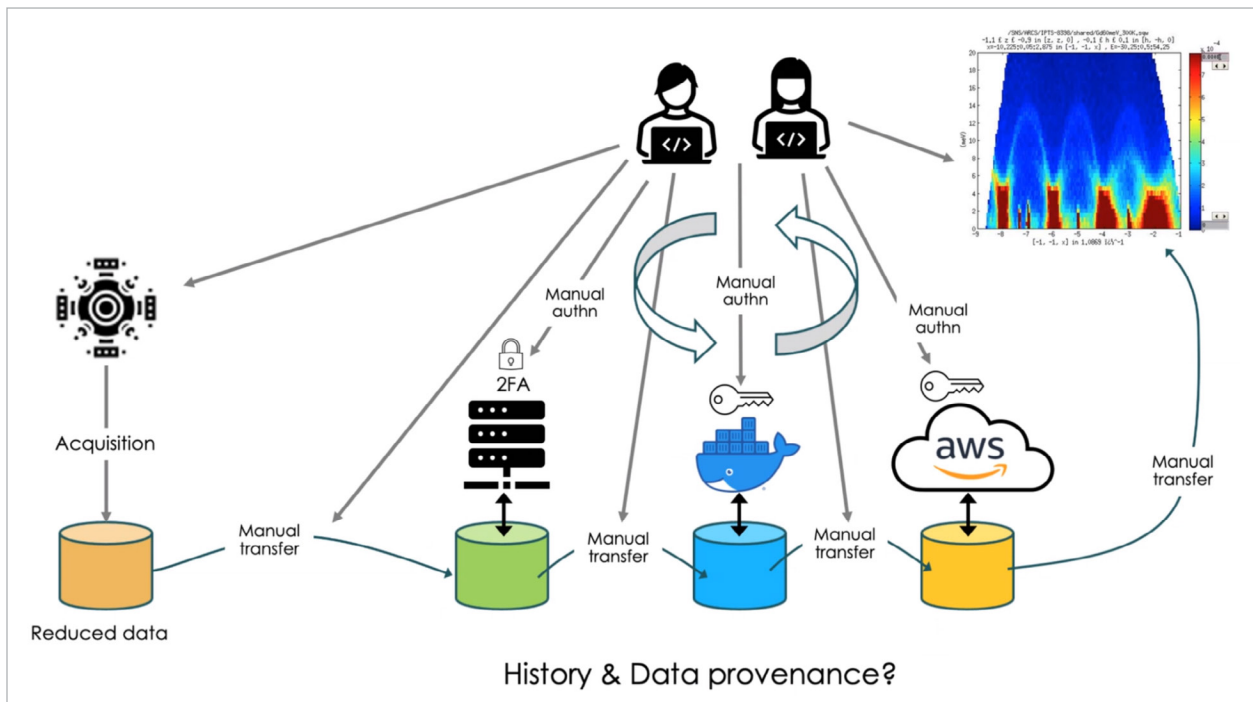
Figure 5.6.11: EPICS Beamline Network Diagram

Users, on the other hand may use cloud servers in their analysis workflows as described in the above figure. This highlights the need for the automated and federated workflows described in Section 5.6.2.4. As was mentioned there, the first implementation of these workflows will use the internal "cloud" resources that have been developed at ORNL over the last several years before external clouds are considered.

The two-factor authentication implementation alluded to in Section 5.6.2.5, is planned to use the Microsoft Authenticator over their Azure cloud.

5.6.2.9 Data-Related Resource Constraints

The IMAGING beamlines, VENUS in particular which is being built now will stress our storage and bandwidth capabilities currently. The storage and bandwidth requirements for this were described in Section 5.6.2.3. Eventually all three facilities: SNS-FTS, HFIR and SNS-STS will have an imaging beamline producing data of the magnitude of VENUS so it will be a major consideration.

5.6.2.10 Outstanding Issues

None to report at this time.

5.6.2.11 Facility Profile Contributors

HFIR and SNS Representation

- **Rob Knudson**, *ORNL*, knudsoniroiv@ornl.gov
- **Bogdan Vacaliuc**, *ORNL*, vacaliucb@ornl.gov
- **Rich Crompton**, *ORNL*, cromptonra@ornl.gov
- **Matt Bedynek**, *ORNL*, bedynekmj@ornl.gov
- **Jeeem Kohl**, *ORNL*, kohlja@ornl.gov
- **Steven Hartman**, *ORNL*, hartmansm@ornl.gov
- **Greg Watson**, *ORNL*, watsongr@ornl.gov
- **Jon Fortnoy**, *ORNL*, fortneyjm@ornl.gov
- **Thomas Proffen**, *ORNL*, tproffen@ornl.gov

ESCC Representation

- **Susan Hicks**, *ORNL*, hicksse@ornl.gov

5.7 Center for Functional Nanomaterials (CFN)

The CFN is a NSRC operated for the US DOE at BNL. As a national scientific user facility, the CFN provides users a research experience supported by top-caliber scientists and with access to state-of-the-art instrumentation.

5.7.1 Discussion Summary

- The CFN is an NSRC operated for the US DOE at BNL. Scientific projects at CFN fall under three nanoscience themes: nanomaterial synthesis by assembly, accelerated nanomaterial discovery, and nanomaterials in operando conditions.

- CFN is not a single collaboration. The CFN facility includes advanced instrumentation in nanolithography, materials preparation, electron and photon probes, and computational resources. They all operate independently.

- For experimental facilities, CFN is also rapidly expanding its ability to support users who are not physically present at the facility. This program builds on existing support for remote user access to computing resources. It has facilitated remote engagement of users with CFN staff, remote operation of instruments during experiments, a mail-in process for user sample analysis, and off-site access to experimental data and CFN-supported analytics software.

- CFN's capabilities require effective utilization of high-rate data streams and real-time management of experimental conditions; this begins with raw data management (e.g., high rates and volumes from instruments), but also developing a suite of tools to optimize facility usage. CFN goal for continued growth in this area is to offer integrated data management, analytics, and simulation tightly coupled to target experimental facilities.

- Many instruments at CFN are operated stand-alone; in some cases, the instruments do not have sufficient storage needs and are being incorporated into the SDCC at BNL. This has revealed some significant networking challenges in ensuring ample capacity and low latency.

- The primary example for storage needs is the CFN TEM facilities which can generate 3 GBps data flows. To support this requirement, CFN installed fiber optic networking and purchased 1 PB of GPFS storage integrated into SDCC operated facilities. This allowed secondary transfer, storage, and processing of those data bursts.

- A planned collaboration with LBNL to operate a high frame rate camera, capable of generating data rates of 50 GB/s, suggests that about 1 PB per year will be generated.

- In the next two- to five-year timeframe, additional direct detection capabilities will likely be added with acquisition of new STEMs which could drive the data generation rate up to as much as 4 PB per year.

- Beyond the five years' timeframe CFN will likely replace all its five (S)TEMs currently in operation. There will be data transfer involved between national laboratories to take advantage of supercomputer capabilities and data analysis tools available at any particular Laboratory.

- CFN is investigating whether software tools developed for large-scale scientific facilities can be used. CFN is leveraging data-transfer and storage technologies developed for high-energy physics, and will investigate whether data acquisition and management tools developed for synchrotron facilities can be adapted to lab-based tools. In particular, the Bluesky framework may provide a robust way to track data and metadata in a schema-free manner.

- CFN operates a midrange HPC facility, and it is managed by the SDCC, part of the Computer Science Initiative (CSI) at BNL.

- Direct data transfer between the beamline workstations and the HPC at SDCC is not made available for security concerns. Instead, a workaround is designed to use the Amazon S3 as a communication intermediate, where commands are passed back and forth through the file system.

- The CFN's computing facility is accessed fully remotely through SSH gateways. Data transfer is commonly done with Globus.

- The physical storage and data management needs of CFN facilities located at NSLS-II will follow policies implemented at that facility. In the one- to three-year timeframe, the most data intensive facilities are expected to be X-ray scattering, with steady-state data rates approaching 10 PB/year, and peak data rates during burst acquisition of >1 TB/hour.

- BNL features a Tbps HTSN that serves as the primary network transport for all data intensive collaborations at BNL, and access to HPC and HTC resources internal and external to the lab.

- The BNL network perimeter includes 3x100 Gbps connections to ESnet, and average 15–20 PBs of data monthly.

- BNL will upgrade network capabilities in the one-to two-year timeframe to support 400 Gbps connectivity, and beyond.

5.7.2 CFN Facility Profile

The CFN mission is advancing nanoscience, by being an essential resource for the worldwide scientific community and by carrying out transformative nanoscience research to support the energy, economic, and national security of the United States. Strategic partnerships are crucial to CFN mission success, including the strong synergy with the National Synchrotron Light Source II (NSLS-II), also located at BNL.

5.7.2.1 Science Background

Scientific projects at CFN fall under three nanoscience themes: nanomaterial synthesis by assembly, accelerated nanomaterial discovery, and nanomaterials in operando conditions. These themes reflect the technical expertise of the staff and guide the development of cutting-edge facilities. The CFN facility is envisioned with the entire process of materials research in mind (synthesis and fabrication, advanced characterization, and understanding), such that users access an integrated set of tools for a complete research experience under one roof. The CFN operates advanced instrumentation in nanolithography, materials preparation, electron and photon probes, including those located at NSLS-II, and computational resources with diverse software tools for nanoscience theory, simulation, and data analytics.

CFN's in situ and operando characterization capabilities exemplify the needs of effective utilization of high-rate data streams and real-time management of experimental conditions. Challenges begin with the acquisition and management of raw data, owing to the high rates and large volumes of data generated by these leading scientific instruments. However, the larger challenge is to develop the suite of tools needed to optimize the usage of precious facility time. In practice, this means gathering the most meaningful data sets within the limited experimental duration that can be allocated, in service of the ultimate goals of research productivity and scientific impact. This brings the focus to developing not just a baseline set of tools for experimental operations and raw data management, but to workflow software tools that encompass database access, in situ data analytics including application to multimodal data, and access to on-demand computing to meet heavy data analysis needs and for physical simulations to be carried out during the timeframe of the experimental work at the facility.

5.7.2.2 Collaborators

CFN is not a single collaboration. The CFN facility includes advanced instrumentation in nanolithography, materials preparation, electron and photon probes, and computational resources. They all operate independently. CFN staff engage in many different collaborations with Users from around the world who come to CFN for a

particular instrument or a combination of instruments. CFN scientists also collaborate with colleagues from other DOE facilities for development of new instruments.

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
BNL USERS/ COLLABORATORS	Primary	Data Transfer	Instrument dependent	Ad hoc	N	N/A
USERS/ COLLABORATORS FROM OTHER DOE LABS	Primary	Data Transfer	Instrument dependent	Ad hoc	N	N/A
US UNIVERSITY-BASED PIS	Primary	Data Transfer	Instrument dependent	Ad hoc	N	N/A
INTERNATIONAL PIS	Primary	Data Transfer	Instrument dependent	Ad hoc	N	N/A

Table 5.7.1: CFN Collaboration Space

5.7.2.3 Instruments and Facilities

Within the CFN facility set, the expanding scope of capabilities driven by in situ and operando characterization needs are the primary target for integrated data management and analysis. These include data acquisition modes in transmission electron microscopy that represent the largest raw data challenge within the CFN, several end stations at National Synchrotron Light Source II (NSLS-II) where CFN has partnerships, and a suite of scanning probe instruments. In parallel, CFN has operated a midrange HPC facility since its inception. That facility has been migrated to a paid-in, stakeholder share of joint facilities managed by the SDCC, part of the CSI at BNL. CFN internal and user science projects routinely involve collaboration between experiment and physical theory supported by computation. The computational demands in support of that physical theory will continue to grow with the complexity of the systems and measurements done in CFN facilities. Furthermore, the CFN goal for continued growth in this area is to offer integrated data management, analytics, and simulation tightly coupled to target experimental facilities. Such an offering would empower the nanoscience user community to tackle materials discovery problems of previously unattainable complexity.

5.7.2.4 Generalized Process of Science

Many instruments at CFN are operated stand-alone. However, in a few cases, the computers attached to the instruments cannot meet the computing and/or storage needs, and we have been working to incorporate the SDCC at BNL, which also manages our HPC, into the workflow. Here we have met significant networking challenges.

The primary example for storage needs is the CFN TEM facilities. The presently installed Gatan K2-IS camera generates ∼ 3 GB of data per second. At the peak data rate, data is currently streamed for storage on an intermediate solid-state disk array in a vendor proprietary format that limits uninterrupted video streaming to 15 minutes. CFN installed fiber optic networking and purchased 1 PB of GPFS storage integrated into SDCC operated facilities. This allowed secondary transfer, storage, and processing of those data bursts. However, the current configuration does not allow for the uninterrupted streaming of data from the camera to large-scale storage that would enable longer experimental runs. A planned collaboration with LBNL to install a ∼ 90,000 frame rate camera will generate ∼ 50 GB of data per second. Once it is achieved, an estimate that includes a

realistic assessment of user project needs and workflows suggests that about 1 PB per year will be generated. In the next two- to five-year timeframe, additional direct detection capabilities will likely be added with acquisition of new STEMs which could drive the data generation rate up to as much as 4 PB per year. Beyond the five years' timeframe CFN will likely replace all its five (S)TEMs currently in operation. All newly acquired (S)TEMs will have at least one fast direct electron detector installed with increased number of active pixels and higher frame rate in comparison with current state-of-the-art solutions. For the purpose of data analysis there will be a potential data transfer involved between national laboratories to take advantage of supercomputer capabilities and data analysis tools available at any particular Laboratory.

As an example of computing needs by an instrument, the CFN is leading research on autonomous experimentation at the CMS beamline of NSLS-II. Sophisticated machine-learning models are developed to control the experiments. However, these compute-demanding models run too slow on local workstations and would cause significant waste of precious beamline time. We have been working to include the SDCC HPC cluster into the workflow. However, direct data transfer between the beamline workstations and the HPC at SDCC is not made available for security concerns. Instead, a workaround is designed to use the Amazon S3 as a communication intermediate, where commands are passed back and forth through the file system. Our initial tests were successful. We are able to control the beamline based on a model computed on the HPC. However, because the HPC is a shared resource managed with SLURM, we have to use 'reservation' to guarantee compute for the experiment, which requires good planning. It is still a challenge to get real-time computing instantly as needed.

5.7.2.5 Remote Science Activities

The CFN's computing facility is accessed fully remotely through SSH gateways. Data transfer is commonly done with Globus. For experimental facilities, CFN is also rapidly expanding its ability to support users who are not physically present at the facility. This program builds on existing support for remote user access to computing resources. It has facilitated remote engagement of users with CFN staff, remote operation of instruments during experiments, a mail-in process for user sample analysis, and off-site access to experimental data and CFN-supported analytics software.

5.7.2.6 Software Infrastructure

CFN is investigating whether software tools developed for large-scale scientific facilities can be used for smaller-scale, but high data-rate, lab tools. CFN is already leveraging data-transfer and storage technologies developed for high-energy physics, and will investigate whether data acquisition and management tools developed for synchrotron facilities can be adapted to lab-based tools. In particular, the Bluesky framework in use at NSLS-II and being deployed at other synchrotron facilities provides a robust way to track data and metadata in a schema-free manner.

Projects and collaborations directed to data analytics and autonomous experimentation will continue, including testing with users, with planned internal resources for selected goals. For example, XANES is a widely used technique with broad applicability. Improving methodologies for interpretation of the spectra can have an impact across many materials and chemical science use cases. Improving data analytics includes linking domain-specific tools for direct simulation of spectra to applied math tools for automated spectral comparison and ML to link structure descriptors to spectral descriptors. Furthermore, many user projects take advantage of multiple facilities to generate complementary data sets. Advanced data analytics for multimodal data sets, leveraging emerging applied math techniques, will also be essential. Such advances, implemented in validated software tools and combined with curated and extensible databases of both experimental and computational results, can significantly enhance scientific impact from target facilities in a three- to five-year timeframe. However, CFN does not have the software development staff resources needed to harden tools for robust user operations.

5.7.2.7 Network and Data Architecture

BNL High-Throughput Science Network

BNL has implemented a vendor agnostic, resilient, scalable and modular terabit per second (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 primary components:

- **Network perimeter**
  - Three diverse 100 Gbps circuits that peer with ESnet. 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 15–20 PBs of data monthly.
- **Science DMZ**
  - Supports open, high-speed WAN/Internet) access for all scientific collaborations throughout the BNL campus.
- **Science Core**
  - A Tbps DCI for data intensive collaborations at BNL. This network 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.
- **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.
  - 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 ASN but does not have Internal BGP (iBGP) peering's between them. Each Leaf or pairs of Leaves will require their own ASN.
- **Storage Core**
  - A redundant terabit per second switching block that aggregates high-performance storage services.

BNL Next-Generation Network Perimeter

The network perimeter at BNL is a high-speed and fault-tolerant network infrastructure that provides the BNL site connectivity to the Internet and various scientific wide area networks. It supports numerous data intensive collaborations such as BES, Biological and Environmental Research, HEP, and Nuclear Physics. It also supports critical campus services such as workstations, phone service, security, safety and monitoring and enterprise and cloud computing. Since being placed into production in 2013, the network perimeter has transmitted over 100+ PBs of data per year to numerous scientific collaborations worldwide.

In 2013 the BNL network perimeter was bleeding edge 100 GbE technology. Now, the hardware has reached 8+ years in age and it is no longer cost-effective to purchase additional hardware for these platforms. Newer platforms today support much greater 100 GbE interface densities along with supporting 400 GbE which will allow BNL to support all its data intensive collaborations well into the future. With these factors in mind, BNL will possibly procure a next-generation network perimeter within the next one to two years. This will prepare the laboratory to meet the missions needs of all the data intensive collaborations at BNL.

5.7.2.8 Cloud Services

We have limited use cases of DropBox and OneDrive.

5.7.2.9 Data-Related Resource Constraints

The physical storage and data management needs of CFN facilities located at NSLS-II will follow policies implemented at that facility. In the one- to three-year timeframe, the most data intensive facilities are expected to be X-ray scattering, with steady-state data rates approaching 10 PB/year, and peak data rates during burst acquisition of >1 TB/hour.

The CFN TEM facilities are the primary challenge in data-related resource constraints. The presently installed Gatan K2-IS camera can generate ~ 3 GB of data per second. At the peak data rate, data is currently streamed for storage on an intermediate solid-state disk array in a vendor proprietary format that limits uninterrupted video streaming to 15 minutes. CFN installed fiber optic networking and purchased 0.6 PB of GPFS storage, now integrated into SDCC operated facilities to replace the 32 TB of SAN storage originally provided with the instrument. This allowed secondary transfer, storage, and processing of those data bursts. However, the current configuration does not allow for the uninterrupted streaming of data from the camera to large-scale storage that would enable longer experimental runs. Achieving this represents an internal project in the planning phase in collaboration with LBNL to install a ~ 90,000 frame rate camera which will generate ~ 50 GB of data per second. Once it is achieved, an estimate that includes a realistic assessment of user project needs and workflows suggests that about 1 PB per year will be generated. In the next two- to five-year timeframe, additional direct detection capabilities will likely be added with acquisition of new STEMs which could drive the data generation rate up to as much as 4 PB per year. Beyond the five years' timeframe CFN will likely replace all its five (S) TEMs currently in operation. All newly acquired (S)TEMs will have at least one fast direct electron detector installed with increased number of active pixels and higher frame rate in comparison with current state-of-the-art solutions. For the purpose of data analysis there will be a potential data transfer involved between national laboratories to take advantage of supercomputer capabilities and data analysis tools available at any particular Laboratory.

5.7.2.10 Outstanding Issues

None to report at this time.

5.7.2.11 Facility Profile Contributors

CFN Representation

- **Charles Black,** *BNL,* ctblack@bnl.gov
- **Qin Wu,** *BNL,* qinwu@bnl.gov
- **Dmitri Zakharov,** *BNL,* dzakharov@bnl.gov

ESCC Representation

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

5.8 Center for Integrated Nanotechnologies (CINT)

The Center for Integrated Nanotechnologies (CINT) is a DOE Office of Science NSRC operating as a national user facility. As a vibrant partnership between LANL and SNL, CINT leverages the unmatched scientific and engineering expertise, as well as special capabilities, of our host DOE laboratories, creating a unique user facility environment among the NSRCs.

5.8.1 Discussion Summary

- The Center for Integrated Nanotechnologies (CINT) is a DOE Office of Science NSRC operating as a national user facility. As a vibrant partnership between LANL and SNL.

- CINT has users from most US states (42) and several (26) foreign countries. In a typical pre-COVID year, CINT would host approximately 800 users both on-site and remotely.

- CINT is largely an experimental facility with significant theoretical expertise, where theory and experiment often work together collaboratively. CINT is moving towards AI/ML with a focus on creating and understanding nanomaterials.

- Collaborating institutions (user home institutions) are generally academic, other national laboratory, and some industry. CINT also has strong relationships with other user facilities including the four sister NSRCs, the National High Magnetic Field Laboratory, and various Office of Science neutron and X-ray sources.

- Data set sizes, file types, file numbers etc. vary widely, as many capabilities are vendor based and proprietary. Other data sets may be generated by in-house software, e.g. LabView. Data sets would typically be less than 1GB.

- The generalized workflow for CINT is as follows to support experimental science:

  - Small-scale experiments involving synthesis or analysis are planned.

  - When instruments are involved, data is collected on the associated instrument computer.

  - Data is typically analyzed on the associated computer using proprietary software, if the instrument is proprietary.

  - Data may be analyzed on a separate workstation if the data is nonproprietary (Igor Pro, Origin etc.)

  - Transfer of data between computers or off-site takes place using encrypted USB or general network services. Methods for moving data are dictated by host laboratory IT security policies.

- For large-scale theory, modeling and simulation: CINT users are granted remote access to institutional computing resources in support of CINT-approved user projects. CINT supports these projects from the scientific perspective but does not support them with respect to infrastructure.

- With the exception of large-scale computing that takes place on host laboratory institutional computers, most data collected at CINT is collected on desktop-type workstations connected to a large variety scientific instruments. These instruments range from vendor-provided (electron microscopy, s-ray diffraction, various spectrometers) to in-house built (ultrafast laser spectroscopy) running on LabView or similar.

- Data is collected by users and staff through either through network transfer (email, ftp) or encrypted USB drive as required by host lab security policies. Many instrument computers are not networked due to patching and reboot requirements, so data must be transferred via USB drive and are of size GB or less.

- Some collaboration and data sharing are allowed through cloud resources, primarily Google Drive.

- As data movement is not a major inhibitor of our work, we have no long-term (5-year) plans to adjust the data collection and distribution approach. Further, CINT is restricted by host laboratory security policies that limit our connectedness to outside entities.

- Due to host laboratory IT security and site-security policies, remote experimental work is typically limited to what can be accomplished vie teleconference platforms such as Zoom. Collaborative writing can be performed through Google Suite. We do not allow users to participate remotely using telepresence tools such as augmented-reality glasses (HoloLens or similar) or telepresence robots.

- Data tends to be managed locally on instrument computers, or individual laptops and desktops. Movement of data may occur via email, FTP, or Google Drive.

- CINT host laboratories limit the use of off-site cloud-based computing resources per IT security policies. The only currently approved cloud resource is Google Suite, through LANL contract.

- Forward looking, it would be helpful to have an accessible network for the transfer of data within a facility, or between the two halves of CINT, that was independent of host laboratory networks and could therefore have a less restrictive security policy.

- The local network architecture for the LANL CINT facility is LANL's standard enterprise deployment. It features 1Gbps facility LAN capabilities, and 10 Gbps connectivity to the laboratory network. The laboratories are connected at 100 Gbps to ESnet.

- The laboratory networking infrastructure features the ability to access institutional computing, as well as DTNs to facilitate mobility for approved users to external sites.

- CINT's most pressing network constraint originates in host laboratory security policy. Currently, CINT utilizes the host laboratory networks for connectivity of instruments and office computers. If an additional layer could be added which would satisfy host laboratory IT requirements (or not be subject to them) and allow easier movement of small data sets between the CINT facilities at the two host laboratories (or remote access to certain experimental tools to off-site, nonbadged users) then that would be a benefit to CINT.

5.8.2 CINT Facility Profile

Our users and staff conduct research projects within and across the Core Facility in Albuquerque, NM, and the Gateway Facility in Los Alamos, NM (**Figure 5.8.1**). By creating a collaborative community of diverse users matched to expert facility scientists with advanced capabilities, CINT fosters high-impact nanoscience discoveries, leads next-generation technique development, and advances the frontiers of knowledge beyond what is achievable by individual researchers or any single institution.

Figure 5.8.1: CINT Core Facility (left) and Gateway Facility (right)

5.8.2.1 Science Background

The overarching goal of CINT is to be the national resource for research expertise and advanced capabilities to synthesize, fabricate, characterize, understand, and integrate nanostructured materials. By developing innovative systems with unprecedented functionality, we aim to inspire revolutionary nano-enabled technologies.

We have world-leading scientific expertise and specialized capabilities to synthesize, fabricate, characterize, and understand nanomaterials in increasingly complex integrated environments. Our expertise is organized in four scientific thrust areas:

- Quantum Materials Systems : Understanding and designing nanomaterials to create new functionalities based on quantum effects that span multiple length scales (from nm to mm).

- Nanophotonics and Optical Nanomaterials: Discovery, synthesis, and integration of optical nanomaterials; exploitation and characterization of emergent or collective electromagnetic and quantum optical phenomena, from nanophotonics and metamaterials to quantum coherence.

- In-situ Characterization and Nanomechanics (ICNM): Developing and implementing world-leading capabilities to study the dynamic response of materials and nanosystems to mechanical, electrical, radiation, or other stimuli.

- Soft, Biological, and Composite Nanomaterials: Solution-based materials synthesis and assembly of soft, composite, and artificial biomimetic nanosystems.

As a national user facility, CINT has users from most US states (42) and several (26) foreign countries. In a typical pre-COVID year, CINT would host approximately 800 users both on-site and remotely, who collectively have a few hundred active user projects. As described above, CINT is largely an experimental facility with significant theoretical expertise, where theory and experiment often work together collaboratively.

CINT theoretical efforts generally utilize institutional computing resources at the two host laboratories, running custom codes developed by CINT-supported theorists with a focus on nanomaterials science.

Collaborating institutions (user home institutions) are generally academic, other national laboratory, and some industry. CINT also has strong relationships with other user facilities including the four sister NSRCs, the National High Magnetic Field Laboratory, and various Office of Science neutron and X-ray sources.

While CINT is largely an experimental facility focused on smaller-scale synthesis, fabrication, analysis and modeling efforts, CINT is moving towards AI/ML with a focus on creating and understanding nanomaterials.

AI/ML Science Opportunity

ML has been around since the dawn of digital computing. But recent advances in computing hardware and algorithm development, along with an exponential acceleration in the quantity of scientific data being collected, have triggered the proliferation of ML throughout all fields of science. In the rush to utilize these new tools, there is still much to learn regarding how to apply these methods in a way that advances our fundamental understanding of the physical world, and in our specific case the understanding of nanoscale phenomena. The most compelling applications of ML will be toward incredibly difficult problems and where the outcomes are not known a priori, such as materials subjected to extreme environments, autonomous optimization of synthetic routes for new materials discovery, and intelligent/autonomous data collection. The largest impacts will be from data analysis that is not tenable by human processing, either because of the needed response time or the enormous quantity of data.

Positional Science and Capabilities

Our current ML expertise and capabilities include automated qubit optimization, prediction of new synthesis routes (physical vapor deposition) for novel metastable nanostructured alloys, virtual microscopy, and acceleration of quantum simulations. Building upon these capabilities will allow us to make scientific advances in areas of

fundamental importance to CINT, such as quantum materials, materials response to extreme environments, and nanomaterials discovery.

Automated Qubit Optimization

Semiconductor qubit devices are becoming more complex as more qubits are added to 1D arrays and further extended into 2D qubit arrays. As the number of qubits increases, so do the number of controls and cross-correlations between controls. Configuring these complex devices for qubit operation quickly becomes intractable for human experimenters. CINT has developed the capability to autonomously tune qubits using ML and image analysis techniques. This capability, which will continue to evolve, will accelerate scientific discovery through full, autonomous, device configuration.

Predicted Synthesis of Metastable Phases

CINT has developed a reduced-order model for accelerated microstructure evolution predictions by utilizing time-series multivariate regression splines or long short-term memory deep learning algorithms within a phase field simulation. We demonstrated this capability to predict microstructure evolution during spinodal decomposition and also are applying it to mesoscale models for corrosion (atmospheric corrosion and molten salt corrosion). This same framework will be used to predict synthesis routes (physical vapor deposition) for novel metastable nanostructure alloys with properties desirable for combined extreme environment applications.

Accelerating Quantum Simulations

Quantum chemical computation is a foundation of materials science but is limited in application due to well-known issues of high computational cost, small system sizes, and short timescales. Emerging data science approaches promise to break the existing scaling barriers and enable training of neural networks capable of quantitative predictions for much larger materials system as well as reducing the cost of these calculations to a level similar to classical force fields. CINT has already demonstrated this capability using a transfer learning algorithm, trained on 5 million highly accurate density functional theory calculations, to generate a coupled cluster model that retains quantitative accuracy. The resulting models will encode chemical and physical information that is extensible to much larger systems, and transferable to new types of processes.

5.8.2.2 Collaborators

Generally, CINT users are small-scale research groups from academia, national laboratories, and some industry from around the US and world. With the exception of large-scale computing that takes place on host laboratory institutional computers, most data collected at CINT is collected on desktop-type workstations connected to a large variety scientific instruments. These instruments range from vendor-provided (electron microscopy, s-ray diffraction, various spectrometers) to in-house built (ultrafast laser spectroscopy) running on LabView or similar.

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
WORLDWIDE	Users typically take data with them, or we transfer via network services	USB drive, FTP, Google Drive	Gigabytes or less	Daily	Metadata — yes during manuscript drafting. Original — no	No

Table 5.8.1: CINT Collaboration Space

Data is collected by users and staff through either through network transfer (email, ftp) or encrypted USB drive as required by host lab security policies. Many instrument computers are not networked due to patching and reboot requirements, so data must be transferred via USB drive.

Some collaboration and data sharing are allowed through cloud resources, primarily Google Drive.

5.8.2.3 Instruments and Facilities

CINT hosts a collection of small-scale capabilities that include numerous materials science analytical tools, chemical spectroscopy, ultrafast laser spectroscopy, diffraction, imaging, micromechanical testing, as well as comprehensive organic synthesis, materials synthesis, and device fabrication. We also access a number of theory, modeling and simulation capabilities that are hosted on institutional computing resources.

Outside of institutional computing resources, CINT capabilities generally run on small-scale desktop computers connected to scientific instruments. These desktop computers may or may not be networked.

Data set sizes, file types, file numbers etc. vary widely, as many capabilities are vendor based and proprietary. Other data sets may be generated by in-house software, e.g. LabView. Data sets would typically be less than 1GB.

Because we are an experimental materials science facility, and data movement is not a major inhibitor of our work, we have no long-term (5-year) plans to adjust the data collection and distribution approach. Further, CINT is restricted by host laboratory security policies that limit our connectedness to outside entities.

5.8.2.4 Generalized Process of Science

The generalized workflow for CINT experimental science is as follows:

1. Small-scale experiments involving synthesis or analysis are planned.
2. When instruments are involved, data is collected on the associated instrument computer.
3. Data is typically analyzed on the associated computer using proprietary software, if the instrument is proprietary.
4. Data may be analyzed on a separate workstation if the data is nonproprietary (Igor Pro, Origin etc.).
5. Transfer of data between computers or off-site takes place using encrypted USB or general network services. Methods for moving data are dictated by host laboratory IT security policies.

Forward looking, it would be helpful to have an accessible network for the transfer of data within a facility, or between the two halves of CINT, that was independent of host laboratory networks and could therefore have a less restrictive security policy.

5.8.2.5 Remote Science Activities

For CINT, remote experimental work is typically limited to what can be accomplished vie teleconference platforms such as Zoom due to host laboratory IT security and site-security policies. Collaborative writing can be performed through Google Suite. We do not allow users to participate remotely using telepresence tools such as augmented-reality glasses (HoloLens or similar) or telepresence robots. During the pandemic, Because CINT approaches are host-lab policy-based, we do not anticipate changing in the near term.

For large-scale theory, modeling and simulation: CINT users are granted remote access to institutional computing resources in support of CINT-approved user projects. CINT supports these projects from the scientific perspective but does not support them with respect to infrastructure.

5.8.2.6 Software Infrastructure

Experimental: Data tends to be managed locally on instrument computers, or individual laptops and desktops. Movement of data may occur via email, FTP, or Google Drive.

Large-scale Computing: CINT Staff Have Developed and Are Developing a Number of Codes to Use to Support Nanomaterials Science: LLAMPS, MEXMD, LATTE, TBM3. For Applications, See Introduction Above.

5.8.2.7 Network and Data Architecture

The following information is for the CINT Gateway at LANL. CINT Core facilities at Sandia are similar.

The local network architecture for the LANL CINT facility is LANL's standard enterprise deployment. It features the following key components:

- Gigabit Copper Ethernet Interface
- Ten Gigabit Uplinks from facility
- Shared 100 Gigabit connectivity to ESnet
- Enterprise storage networks

This local enterprise network has direct connections to LANL's large-scale computing capabilities. The institutional computing program provides the following resources:

- High-speed DTN services
- 900 million CPU hours per year
- Conventional modeling and simulation
- ML
- Data analysis
- Quantum computing

5.8.2.8 Cloud Services

CINT host laboratories limit the use of off-site cloud-based computing resources per IT security policies. The only currently approved cloud resource is Google Suite, through LANL contract.

5.8.2.9 Data-Related Resource Constraints

CINT's most pressing network constraint originates in host laboratory security policy. Currently, CINT utilizes the host laboratory networks for connectivity of instruments and office computers. If an additional layer could be added which would satisfy host laboratory IT requirements (or not be subject to them) and allow easier movement of small data sets between the CINT facilities at the two host laboratories (or remote access to certain experimental tools to off-site, nonbadged users) then that would be a benefit to CINT.

5.8.2.10 Outstanding Issues

None to report at this time.

5.8.2.11 Facility Profile Contributors

CINT Representation

- **Adam Rondinone,** *LANL,* rondinoneaj@lanl.gov

ESCC Representation

- **Dylan Merrigan,** *LANL,* dylan@lanl.gov

5.9 Center for Nanoscale Materials (CNM)

The CNM at ANL is one of five NSRCs operated by the DOE Office of Science as a user facility providing expertise, instrumentation, and infrastructure for interdisciplinary nanoscience and nanotechnology research. Academic, industrial, and international researchers can access the center through its user program for both nonproprietary and proprietary research.

5.9.1 Discussion Summary

- The CNM at ANL is one of five Nanoscale Science Research Centers (NSRCs) operated by the DOE Office of Science.

- During FY21, the CNM hosted 349 on-site and 353 remote user researchers from academia, national laboratories and industry.

- CNM research activities span the fields of optical and electron microscopy, optical microspectroscopy, laboratory source X-ray diffraction, scanning tunneling microscopy and spectroscopy, high-magnetic-field optical spectroscopy and microscopy, as well as the production of simulations.

- The CNM also houses a comprehensive suite of theory and modeling tools, an HPC cluster, and various software and modeling/simulation tools.

- The CNM HPC Cluster (Carbon) provides GPUs and edge-computing capabilities.

- The experimental and computational data created on-site is stored in the CNM data repository for two to five years as space permits. User generated data has the user as the primary custodian with the laboratory as secondary custodian following the institutional user agreements in place for site access.

- The experimental and computational data created on-site is transferred to external users through the Globus platform, physical hard drives, ftp/scp, or the ANL BOX system, and further analysis is typically performed locally post transfer.

- Remote science is enabled via telepresence engagement of the staff researcher or remote operation directly by the user through network access to control screens, or through remote access to Carbon cluster. Experimental work is remotely driven through a Bomgar Remote Access Server providing secure visualization of process control screens and on-site analytical tools to remote clients, data transfer using Globus is typically done at the conclusion of the experiment or synchronously throughout the experimental time, allowing for iteration or externally hosted analysis.

- Remote analytical tools for simulation such as ALCF resources for AI/ML training (Cerebrus/Polaris) are also used. Computations are routinely performed remotely by the CNM staff across different computing clusters and facilities within Argonne (Bebop, LCRC, Blues and Theta) as well as other DOE supercomputing facilities (NERSC).

- In the next two to five years, CNM plans to release custom research data management systems (RDMS) to aid in the data curation, analysis, and publication of scientific data produce at CNM.

- There is a need for common standards and shared workflows for data across the NSRCs which will not only provide an effective data solution, but will also enable cross-center data sharing, augmentation, and manipulation. NSRCs have different schemes and different levels of implementation for acquiring, labeling, storing, and providing access to the heterogeneous data generated. The lack of data pipeline standards causes issues, such as differences in metadata associated with different data types and different levels of data manipulation. Top address this, many levels of technical details need to be worked out:

    — data formats.

    — software development framework.

    — sample tracking.

    — metadata capturing.

    — labeling.

    — data sets from correlated measurements, and the corresponding simulations, need to be handled in a coordinated manner to extract synergistic information.

- CNM data sets range in size from MB to GB per entry, and up to 1000s of entries per year; requiring data storage to support TB scale data over time. Emerging and planned instruments can stream data at higher rates (e.g., TB scale) and will require faster and more plentiful storage options provided by ANL or through collaboration with ALCF.

- During the experimental time, characterization, synthesis, or fabrication data is locally analyzed as it is being acquired at sufficient fidelity to enable on the fly decision-making on experimental process.

- The ANL Self Driving Lab Initiative creates a local control of experimental process based on data recognition which requires transfer of data to HPC resources for external training of algorithmic approaches that are then locally deployed.

- In the next two to five years, we are working towards integrating experimental synthesis, microscopy and spectroscopy characterization, and AI inference from experimental data on the same compute and data platform.

- Our vision beyond five years is to realize fully automated workflows that will allow for seamless sharing of experimental and simulation data sets amongst the users and allow them to perform multiple types of analysis (including AI/ML).

- We also envision setting up of digital twins in the 5–10 year timeframe, that will mimic our experimental synthesis and characterization platforms and facilitate virtual experiments allowing users to map experimental controls to the intended outputs and exhaustively perform synthetic experiments before the actual ones.

- The CNM does not presently utilize cloud services for its scientific data infrastructure or for computations/data analysis. Cost-effectiveness was a major factor hindering the use of cloud computing services for CNM users. We do not have any immediate plans to utilize cloud services for our need.

- The Argonne WAN network connectivity is provided by ESnet and MREN (Internet2 gateway, Peer Exchange (I2PX)) ESnet has optical and routing gear on the Argonne campus.

- The current ANL network supports 10 Gbps, 40 Gbps, 100 Gbps, and 400 Gbps for WAN and LAN uses, and features a Science DMZ to support ANL facility use.

- The Argonne Science DMZ is connected directly to the Argonne border routers with redundant 2x100GE links. The Science DMZ provides high-speed connectivity between scientific organizations for the exchange of data without taking the traditional path through a firewall.

- Connections in to the Science DMZ are a minimum of 100GE with most facilities connecting at 2x100G. This allows collaborators to exchange data in a high-speed environment that does not affect commodity network connectivity.

5.9.2 CNM Facility Profile

As a DOE-funded research center, the CNM is at the forefront of discovery science that addresses national grand challenges encompassing the topics of energy, information, materials and the environment. The scientific strategy of the CNM is consolidated under the following three crosscutting and interdependent scientific themes: (a) Quantum materials and sensing; (b) Manipulating nanoscale interactions; and (c) Nanoscale dynamics. Collectively, they aim at the discovery and hierarchical integration of materials across different length scales, and at the extremes of temporal, spatial, and energy resolutions.

5.9.2.1 Science Background

The CNM provides world-leading expertise and tools to its users. Some of the distinguishing capabilities available at the CNM include the Ultrafast Electron Microscope (UEM), the Hard X-ray Nanoprobe operated in partnership with the APS, novel scanning tunneling microscopes, ultrafast optical spectroscopy techniques over a broad ultraviolet to terahertz spectral range, quantum optics characterization capabilities with single nanoparticle magneto-photoluminescence spectroscopy, nanomechanical and plasmonic structure fabrication, an ultralow temperature (10 mK) quantum research laboratory, a superlubricity laboratory, a chromatic aberration corrected TEM, nanofabrication at the largest research clean room in the Midwest, and the Carbon HPC cluster. The CNM currently employs 69 employees, who contribute to world-leading scientific programs in addition to supporting the users of the facility. During FY21, the CNM hosted 349 on-site and 353 remote user researchers from academia, national laboratories and industry. The experimental and computational data created on-site is transferred to external users through the Globus platform, physical hard drives, ftp/scp, or the ANL BOX system, and further analysis is typically performed locally post transfer. This data is additionally stored in the CNM data repository or in the case of beamline data using the APS Nautilus server for two to five years as space permits. User generated data has the user as the primary custodian with the laboratory as secondary custodian following the institutional user agreements in place for site access.

5.9.2.2 Collaborators

The CNM user base is distributed nationwide with regional collaborators spanning the Midwest, California and the east coast of the United States. We also cater to international users (13% of our users located internationally) from European, Pacific and South American origin. The primary need for data transfer is from the CNM facility to researchers in US academia and other US national laboratories. Data directionality is primarily outward facing, with data generated at the CNM being transferred to user groups and external collaborators. We partner with several Energy Frontier Research Centers, Multi-University Research Institutes and NSF research centers through PU agreements, most recently with the Q-Next National Quantum Information Science Research Center led by ANL and SLAC which will ramp up research operations over the next five years.

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 PSE, CELS, QNEXT, JCESR	Secondary	GLOBUS, ANL BOX	~100GB	Ad hoc (per experiment/ simulation)	N	Access to proprietary analytical software
US AND INTERNATIONAL ACADEMIC INSTITUTIONS	Secondary	GLOBUS, portable hard drives, scp	~100GB	Ad hoc (per experiment/ simulation)	N	Coordination of posttransfer analysis, transfer speed, data synchronization across different computational clusters
INDUSTRIAL USERS	Secondary	GLOBUS, BOX, Hard Drives	~50GB	Ad hoc (per experiment/ simulation)	N	Coordination of posttransfer analysis, transfer speed, data synchronization across different computational clusters

Table 5.9.1: CNM Collaboration Space
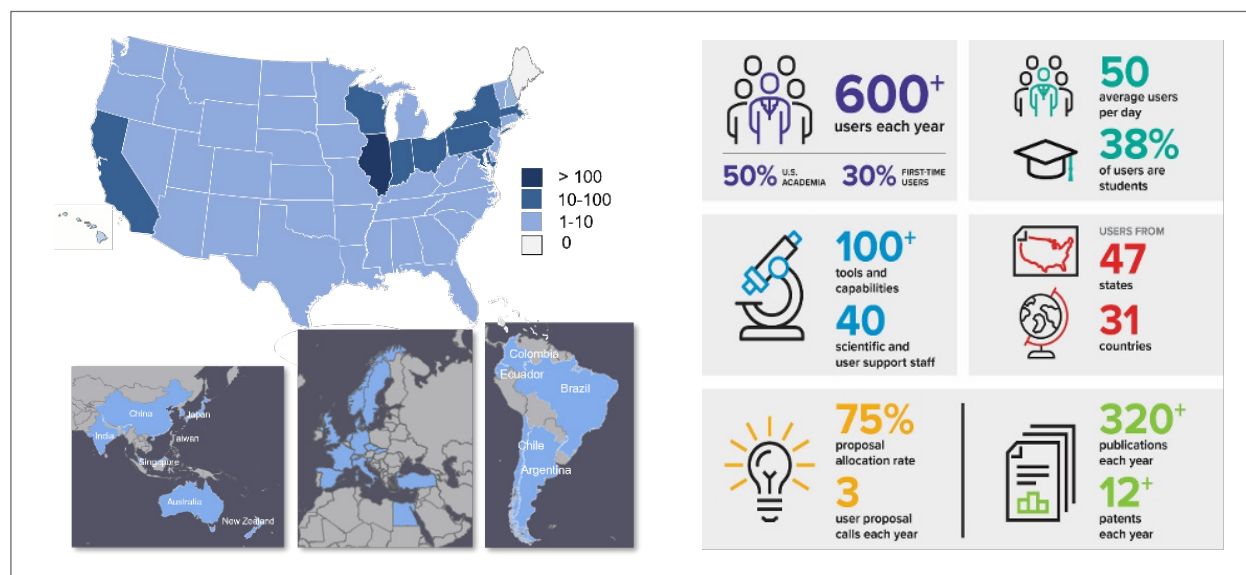
Figure 5.9.1: Geographical distribution and statistics of CNM user program

5.9.2.3 Instruments and Facilities

Argonne's CNM offers a wide range of capabilities designed to enhance the study and creation of nanomaterials. From X-ray microscopy that uses the power of Argonne's APS to clean room–based nanofabrication techniques, the CNM provides its staff and users with a broad combination of scientific resources involving data entries that vary in size from kilobytes to terabytes at collection speeds spanning from weeks to microseconds.
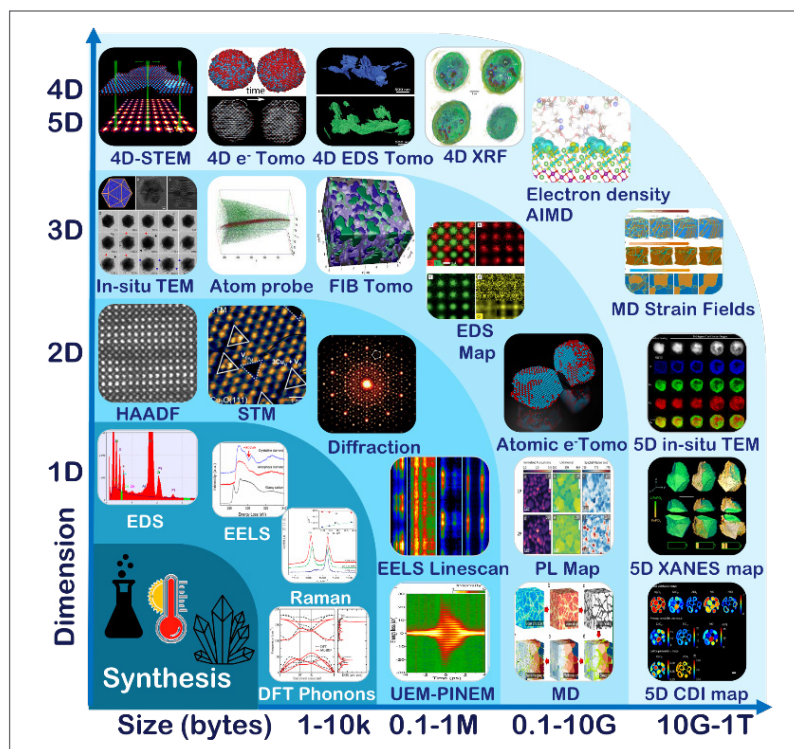
Figure 5.9.2: Examples of data types generated at CNM

We pioneer new clean room–based nanofabrication techniques to integrate novel materials into unique structures and devices to realize nanofabricated structures for the transduction of mechanical, optical, and electrical signals. Data generated from these activities is primarily surface optical and electron microscopy, optical microspectroscopy and laboratory source X-ray diffraction from the nanoscale to wafer-scale characterization. Our low temperature Quantum Laboratory characterizes superconducting quantum devices with DC transport conductance measurements, RF/microwave transmission and reflection spectra (typically ~10 MB per entry, ~1000 entries per yr).

Our toolbox of synthesis techniques includes chemical and nanoparticle synthesis methods, various gas- and liquid-phase deposition methods, general wet-lab and air-free chemistry techniques, and methods for the preparation of samples as films and surfaces. Nano-bio hybrid materials and biological assembly efforts center on nature-inspired materials that potentially can support energy conversion and energy transport, as well as understanding biosensing mechanisms in cell-like environments that are functionalized with engineered nanomaterials. We are able to hybridize metals, organics, semiconductors, and dielectrics with biomaterials to create nanobio metamaterials with unique properties. Data generated from these activities is primarily optical microscopy, electrical characterization and conventional and ultrafast optical spectroscopy (typically ~1GB/entry ~10entries/year). Our recent development of autonomous synthesis within the ANL Self Driving Lab Initiative creates a local control of experimental process based on data recognition which requires transfer of data to HPC resources for external training of algorithmic approaches that are then locally deployed.

In addition, we seek to engineer defects with atomic precision in nanoparticles and two-dimensional nanomaterials with a view to understanding their role in charge and energy transfer as well as their prospects for quantum science. Because defects play a critical role in the properties and behavior of nanostructures due to their small size, a detailed understanding of their influence is critical to nanoscience and developing nanotechnologies. Data generated from these activities is primarily scanning tunneling microscopy and spectroscopy as well as high-magnetic-field optical spectroscopy and microscopy (typically ~30 MB per entry, ~200 entries per yr).

We provide a wide range of capabilities and expertise to help elucidate structural, structure-property and functional information from nanoscale particles, assemblies, devices and systems. The methods span numerous aspects of magnetic/electrical measurements, metrology, microscopy, spectroscopy, quantum science and wear/friction measurements. Instruments range from those commonly found in research laboratories, such as scanning electron microscopy and powder diffraction, to specialized and unique capabilities such as UEM and the Hard X-Ray Nanoprobe at the APS. Our electron microscopy data flow includes structure and spectroscopy of materials using imaging, diffraction and energy loss spectroscopy (typically ~20 MB/entry, ~500k entries per yr). Recently installed high-speed electron microscopy leveraging a K2 detector (~103 frames per second) generates significantly higher data volume and velocity, at 5 MB/entry and up to 720k entries/hour. Our X-ray microscopy capabilities include nano-focused coherent Bragg diffraction imaging and X-ray fluorescence spectroscopy (~0.1TB per entry, ~36 entries per yr) — these are expected to have 100x data rate increase due to the APS diffraction-limited source upgrade with first light in 2024 making full use of our EIGER2-X1M (1028x1062x32bit 2000fps) collection bandwidth — the planned approach for data collection depends on high-throughput ptychography recently demonstrated with locally deployed edge computing delivering real-time imaging that is continuously improved with HPC live-training algorithmic recognition (fig 3). Our electron microscopy capabilities are also planned for significant upgrades with installation of two new microscopes with high dynamic range electron microscope pixel array detectors (128x128x32bit 1100fps) demanding high throughput (up to 7 million entries/hour) data management compatible with external transfer, inference with AI workflow, and additional postanalysis.
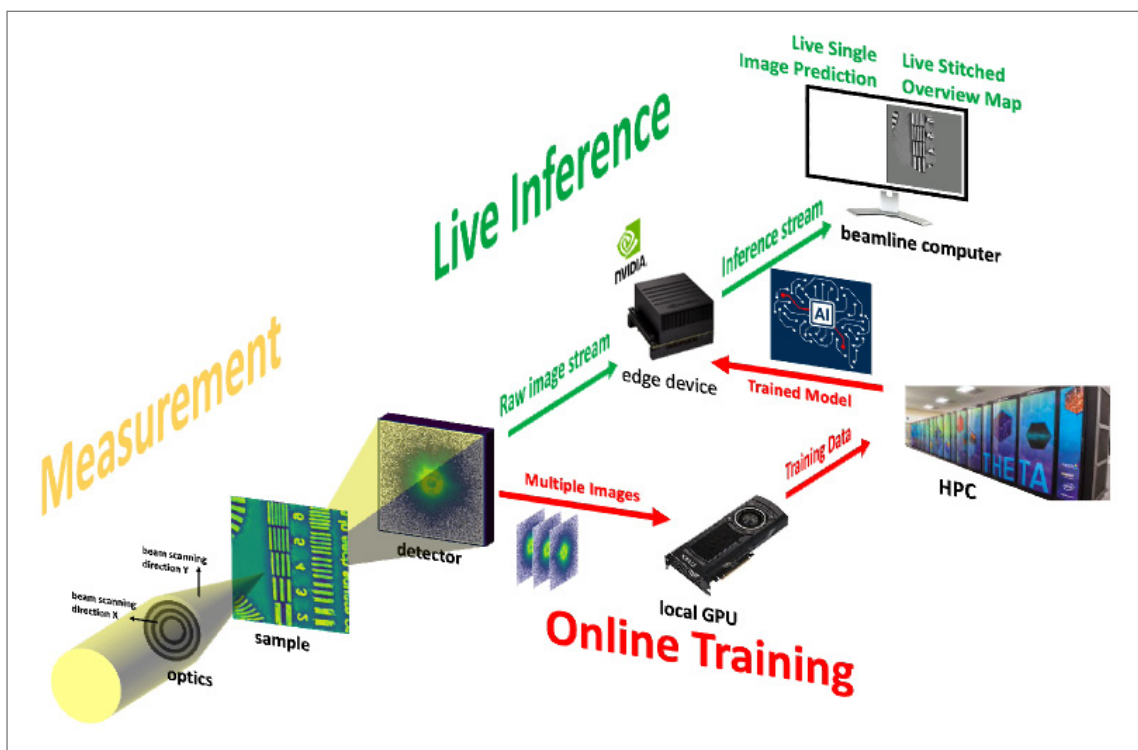
Figure 5.9.3: Demonstration of edge computing in development for high bandwidth microscopy experiments

The CNM also houses a comprehensive suite of theory and modeling tools, an HPC cluster, and various software and modeling/simulation tools. The theory, modeling, and simulation methods include numerous density functional theory packages, molecular dynamics, Monte Carlo, genetic algorithm, finite difference time domain, and AI/ML. The CNM HPC Cluster (Carbon) provides GPUs and edge-computing capabilities. A recent upgrade of 22 nodes each with 2 NVIDIA TESLA V100 GPUs is geared towards AI workflows, including training of

neural network models for image analysis and performing inference from high-speed electron microscopy data. Software and modeling/simulation tools include BLAST (Bridging Length/Timescales via Atomistic Simulation Toolkit), FANTASTX (Fully Automated Nanoscale to Atomistic Structure from Theory and eXperiment), computational electrodynamics (e.g., Lumerical, MEEP), and other specialized analysis software and modeling expertise.

In the next two to five years, we are working towards integrating experimental synthesis, microscopy and spectroscopy characterization, and AI inference from experimental data on the same compute and data platform. Our vision beyond five years is to realize fully automated workflows that will allow for seamless sharing of experimental and simulation data sets amongst the users and allow them to perform multiple types of analysis (including AI/ML). We also envision setting up of digital twins in the 5–10 year timeframe, that will mimic our experimental synthesis and characterization platforms and facilitate virtual experiments allowing users to map experimental controls to the intended outputs and exhaustively perform synthetic experiments before the actual ones.

5.9.2.4 Generalized Process of Science

The mission of the NSRCs is twofold: to enable the external scientific community to carry out high-impact nanoscience projects through an open, peer-reviewed user program, and to conduct in-house research to discover, understand, and exploit functional nanomaterials for society's benefit. We conduct three calls for user proposals a year — these remain active for one year with possible extension to two, resulting in approximately 250 active proposals per year. During the experimental time, characterization, synthesis, or fabrication data is locally analyzed as it is being acquired at sufficient fidelity to enable on the fly decision-making on experimental process. This analysis is typically improved upon post acquisition either through data transfer to remote institutions or remote access to on-site analytical tools, and then integrated with simulation data prior to publication and contributed to external data repositories.
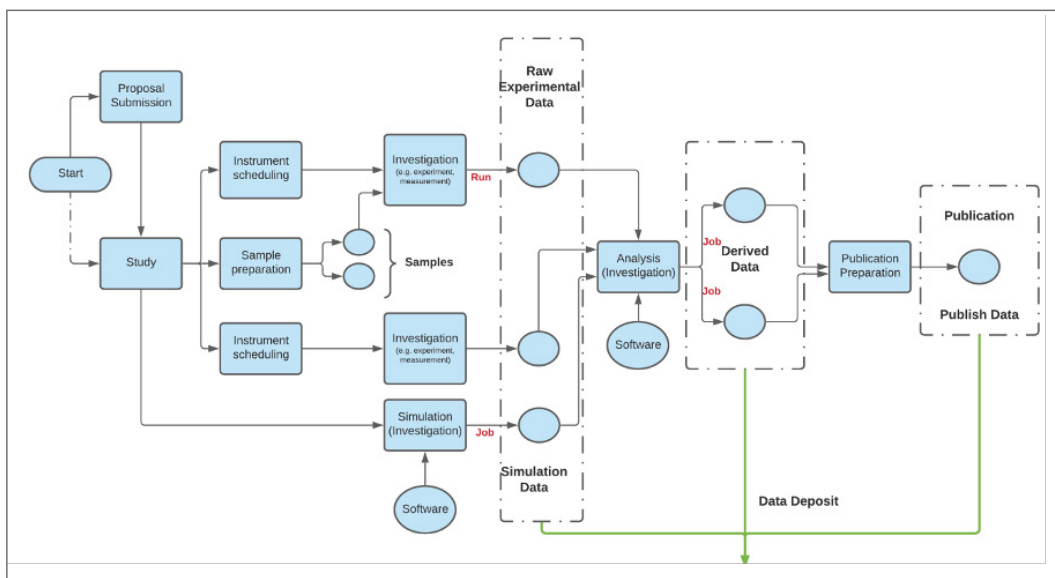
Figure 5.9.4: Example of data workflow at CNM

It is a two- to five-year developmental goal of the CNM to expand and integrate this process into a coordinated data platform to allow the development of common data schemas as well as interconnected software tools and workflows to allow users and staff to acquire, annotate, access and analyze or apply AI/ML tools to complex and often correlated scientific data produced.

5.9.2.5 Remote Science Activities

The CNM has over 30 capabilities currently available remotely spanning synthesis, lithography, spectroscopy, electron / X-ray microscopy, and modeling and simulations. Remote science is enabled via telepresence engagement of the staff researcher or remote operation directly by the user through network access to control screens, or through remote access to Carbon cluster. Experimental work is remotely driven through a Bomgar Remote Access Server providing secure visualization of process control screens and on-site analytical tools to remote clients, data transfer using Globus is typically done at the conclusion of the experiment or synchronously throughout the experimental time, allowing for iteration or externally hosted analysis. Remote analytical tools for simulation such as ALCF resources for AI/ML training (Cerebrus/Polaris) are also used. Computations are routinely performed remotely by the CNM staff across different computing clusters and facilities within Argonne (Bebop, LCRC, Blues and Theta) as well as other DOE supercomputing facilities (NERSC).

5.9.2.6 Software Infrastructure

The CNM primarily utilizes two storage technologies, block and object storage. The software solution for block storage access is Luster, and DataCore Swarm for object storage. Both storage solutions utilize the Globus data platform services to share and transfer data with remote users or facilities. In addition, DataCore Swarm provides an S3 API to access, transfer, and share data. In the next two to five years, CNM plans to release custom RDMS to aid in the data curation, analysis, and publication of scientific data produce at CNM.

5.9.2.7 Network and Data Architecture

The Argonne campus network connects 100 buildings and 12 data centers in a multivendor switching and routing network. The campus has a new robust single-mode fiber cable plant that has plenty of capacity and ability to be expanded.

External network connectivity is provided by ESnet, MREN, Internet2, and Internet2 Peering Exchange (I2PX). ESnet recently migrated Argonne to the new ESnet6 network and has redundant optical transport and routers deployed on the Argonne campus in a redundant configuration supporting up to 400GE connections.

There are two primary networking node locations distributed between the north (Building 221) and south (Building 541B) sides of the campus to provide network redundancy. Generally, all buildings and data centers have connections to each of these networking nodes. The fiber used to connect to off-site providers is diverse and exits on the east and west sides of the campus.

The Argonne networks supports many connection speeds. The WAN supports 10GE, 40GE, 100GE, and 400GE. The campus core network support speeds of 1GE, 10GE, 40GE, and 100GE. Building connections are generally multiple 10GE with some 40GE. Data centers 10GE, 40GE, and 100GE. LAN connections within buildings are generally 1GE and 10GE.

ANL WAN

The Argonne WAN network connectivity is provided by ESnet and MREN (Internet2 gateway, Peer Exchange (I2PX)) ESnet has optical and routing gear on the Argonne campus. Argonne has two Juniper MX960 border routers that are split between buildings 221 and 541B for redundancy and diversity.

Argonne uses ESnet OSCARS virtual circuits. (On-Demand Secure Circuits and Advance Reservation System)

The Argonne connections to ESnet are 2x100GE. The connections to MREN are 3x10GE. The border routers are connected to each other with 2x100GE connections. The connections to the campus Core network are 2x40GE and 40GE to the perimeter firewall.

Argonne has one perfSONAR connected to the border at 40GE.

Argonne deploys cyber security at the border with blackhole routing and network traffic capture taps for traffic collection and intrusion detection.

Present to two years

- 400 GE upgrade expected in FY22

Two to five years

- Upgrade border routers
- Multiple 400 GE

Five + years

- Terabit WAN connections

ANL Science DMZ

The Argonne ScienceDMZ is composed of Juniper QFX series equipment that is connected directly to the Argonne border routers with redundant 2x100 GE links. The ScienceDMZ provides high-speed connectivity between scientific organizations for the exchange of data without taking the traditional path through a firewall.

Connections in to the Science DMZ are a minimum of 100 GE with most facilities connecting at 2x100 G. This allows collaborators to exchange data in a high-speed environment that does not affect commodity network connectivity.

Present to two years

- 400GE connectivity

Two to five years

- Hardware upgrade

Five+ years

- Terabit connectivity

ANL Core

The Argonne core network is provided by 2 Cisco Nexus 7710s spread across buildings 221 and 541B for redundancy. These switches offer network speeds at 1GE, 10GE, 40GE, and 100GE. The Core provide connections to the buildings and many of the data centers on the Argonne campus.

Present to two years

- No changes

Two to five years

- Replace core devices to NextGen

Five+ years

- Terabit connectivity

ANL LAN

The Argonne LAN consists of the building access and aggregation switches. The Laboratory has standardized on Cisco and Aruba switches in the LAN. There is a combination of 1GE, 10GE, and 40GE network speeds in the LAN. Desktops are connected at 1GE.

Present to two years

- Continual refresh of switches older than five to seven years
- Wireless access point upgrades

Two to five years

- Consider an overlay network topology

- Ongoing switch refreshes

- Zero-trust networking

Five+ years

- Next-Gen

- 10GE to the desktop

ANL Data Centers

The primary data center at Argonne consists of Cisco Nexus equipment in an ACI fabric. It connects to the Argonne core network at 4x40GE and offers 2x40GE connectivity to every ToR switch. There is a combination of 1GE, 10GE, 25GE, and 40GE to host in this space.

Present to two years

- Hardware refresh for ToR switches

Two to five years

- Spine switch replacement

Five+ years

- Terabit connectivity
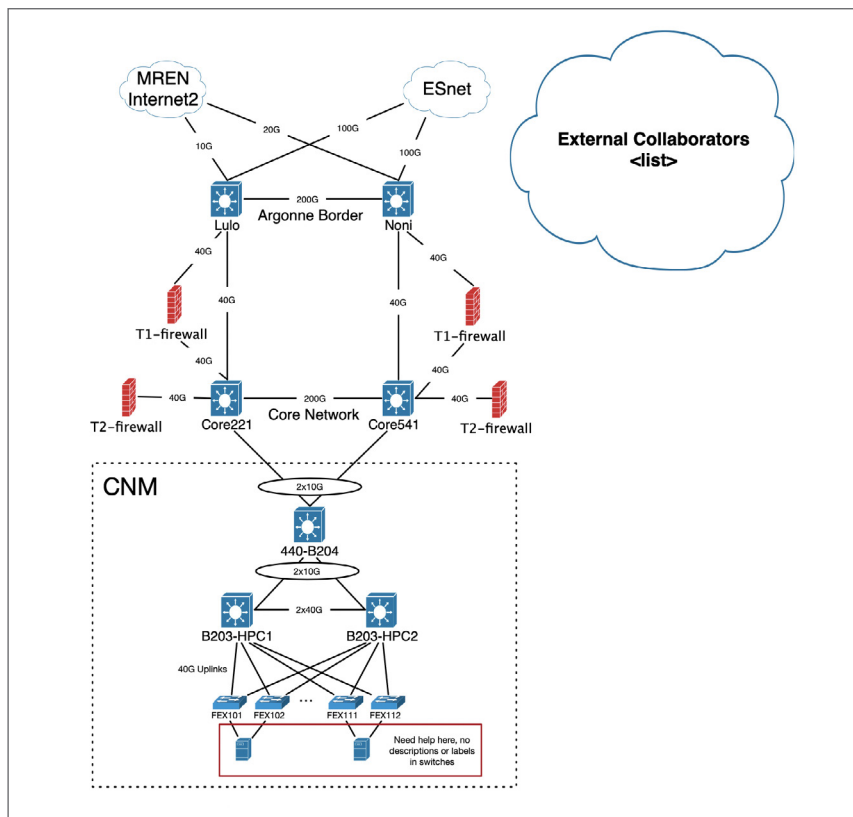
Figure 5.9.5: ANL Network Diagram

5.9.2.8 Cloud Services

The CNM does not presently utilize cloud services for its scientific data infrastructure or for computations/data analysis. Cost-effectiveness was a major factor hindering the use of cloud computing services for CNM users. We do not have any immediate plans to utilize cloud services for our need. CNM along with the other NSRC's will continue to evaluate the capabilities and cost-effectiveness of using cloud services in the near term.

5.9.2.9 Data-Related Resource Constraints

In addition to the need for high-bandwidth physical infrastructure at the present time the different NSRCs, and different facilities within each, have different schemes and different levels of implementation for acquiring, labeling, storing, and providing access to the heterogeneous data generated. The lack of data pipeline standards causes issues, such as differences in metadata associated with different data types and different levels of data manipulation. To capture and curate data from the NSRCs, many levels of technical details need to be worked out, such as data formats, software development framework, sample tracking, metadata capturing, and labeling. In addition, data sets from correlated measurements, and the corresponding simulations, need to be handled in a coordinated manner to extract synergistic information. These issues hamper the efficient use of NSRC data for scientific discovery. There is a need for common standards and shared workflows for data across the NSRCs which will not only provide an effective data solution, but will also enable cross-center data sharing, augmentation, and manipulation.

5.9.2.10 Outstanding Issues

None to report at this time.

5.9.2.11 Facility Profile Contributors

CNM Representation

- **Martin Holt,** *ANL,* mvholt@anl.gov
- **Anthony Avarca,** *ANL,* aavarca@anl.gov
- **Maria Chan,** *ANL,* mchan@anl.gov
- **Subramanian Sankaranarayanan,** *ANL,* skrssank@anl.gov

ESCC Representation

- **Linda Winkler,** *ANL,* winkler@mcs.anl.gov
- **Corey Hall,** *ANL,* chall@anl.gov

5.10 Center for Nanophase Materials Sciences (CNMS)

CNMS offers world-class capabilities and expertise for the synthesis, fabrication, characterization, and theoretical understanding of a wide variety of nanomaterials and nanostructures. Specific facilities at CNMS include laboratories for macromolecular synthesis including selective deuteration, laser synthesis of 2D materials, carbon nanostructures, layered materials, and a broad range of general-purpose laboratories. A 10,000 square foot clean room houses a complete suite of fabrication and lithography (optical and e-beam) instrumentation, as well as specialized approaches for direct-write nanofabrication using electron and ion beams as well as optical 3D patterning. Functional and structural characterization of materials is achieved by optical spectroscopies, nuclear magnetic resonance, X-ray diffraction; electrical, optoelectronic, and magnetic properties can be studied using a broad range of instruments. Quantitative functional measurements are also obtained in a spatially resolved matter using the broad range of imaging tools: a heavy emphasis is placed on cutting-edge scanning probe microscopies, state-of-the-art atomic-scale characterization using scanning transmission electron microscopy and electron microscopy using specific sample holders for in-situ and operando imaging, three-dimensional reconstruction using atom-probe tomography, and combinations of imaging and patterning using helium ion microscopy.

5.10.1 Discussion Summary

- The CNMS is a nanoscience user facility that provides services to users in areas of nanofabrication, synthesis, theory, and characterization.

- The CNMS serves over 600 unique users/year, In FY2021 70% of our users were nonlocal, with 120 US institutions from 36 states and territories, and 43 foreign institutions from 24 foreign countries.

- Users are assigned hours to work on the instruments, but no explicit computational or storage is provided. As such, users are given access to their data when on site, and encouraged to take it with them.

- Experimental data reduction is achieved at the source during the experiment with simple techniques such as signal averaging. Reduced experimental data is usually stored on local drives and then given to the user, through USB keys or other media. Most of the data resides in separate storage boxes that are unlinked. In five years, it is expected this will no longer be the case.

- The data from microscopy experiments is typically stored immediately on a network attached storage (NAS) device, and then transferred to dedicated storage for longer term (typically about five years). Data from simulations is typically handled differently, and stored locally on storage for a few years before going to tape storage for long-term archival.

- Substantial bottlenecks exist in terms of the analysis of the data streams emanating from scanning probe and electron microscopy instruments, as well as integrating feedback from theoretical/simulation insights. CNMS is working to directly couple the microscope data to simulations in real-time to assist the experimenter is of crucial importance in the quest towards automation and autonomous materials/physics discovery platforms.

- CNMS is currently trialing the use of DataFed, an ORNL-developed federated data management system for keeping track of both simulation and experimental data sets. DataFed is a software solution that links together geographically diverse file storage systems and makes all of them available through the same utility. DataFed can work with Globus and other tools.

- CNMS is trialing the use of a gateway device to transfer files from the instruments to the data storage system, with the files automatically ingested by DataFed so that the relevant metadata is also captured, thus enabling keyword searches by metadata fields. The data storage is local

- to ORNL. Any remote data sets can be accessed through collaborators through either Globus endpoints or through scp/ftp, or through the DataFed interface.

- CNMS requires more access to computation, which is particularly limited for the theory groups, and the changing nature of the HPC platforms which make older code quickly obsolete, requiring significant rewriting to accommodate new hardware upgrades.

- Networking is critical to achieve autonomous and "smart" characterization tools, and for physics discovery. Data captured at synthesis and characterization tools can be processed and the experimental conditions updated. A sufficient amount of computational and networking resources to keep up with experiment flow are required to achieve this vision.

- Most microscopy data set collections are in the 1–10GB range, with more for electron microscopy (can be >10GB for a single data set in that instance). The difficulty usually lies not in storage but in analysis, so that there is enough feedback for how to optimize measurement parameters before the user departs. For large files, data transfer can be an issue if the institutions involved do not possess Globus endpoints.

- STEM devices typically have data rates of approximately 100Mbps, and 4D-STEM can increase this to greater than 100Gbs. Data reduction using edge processing is critical to handle the volumes of the later.

- CNMS data is being transitioned to a large 2 PB storage server housed in the SNS. CNMS also runs a local server with about 100 TB of storage which is a windows file system that our staff can access to back up their data.

- The CNMS currently has 2 GPU systems for dedicated model training and deployment at the edge.

- For running simulations, current compute capacity is operated by CADES on dedicated CNMS resources.

- ORNL-provided virtual machines to run Jupyter notebooks on the CADES OpenStack cloud compute are available to facilitate computing. These resources are not always sufficient and some users opt out and process the data locally on their own laptops, or on more dedicated hardware available to them from their own institutions.

- Most of the data that is captured at the Center for user projects is passed back to the user in the form of hard drives or uploaded and then downloaded by the user onto their home institution devices. For the theory collaborations, Globus endpoints can be used for transfer of simulation data files.

- Significant computational needs beyond edge compute will be required, including GPUs for tasks such as model training, running molecular dynamics codes, or cluster compute for first principles approaches such as density functional theory.

- A dedicated pipeline with fast data-transfer rates connecting edge computers that are close to experiments, with such midcapacity and HPC infrastructure in national labs is required to drive specific simulations on the fly to both guide experimental investigation, as well as accelerate materials as well as fundamental physics discoveries.

- Most of the remote resources used are computational in nature, e.g., utilizing compute clusters at NERSC.

- For most of the facility, the data storage and transfer speeds are currently sufficient, and are not expected to be significant concerns in the next two years.

- Currently the CNMS makes limited use of cloud services. Our cloud service is a locally run OpenStack run by CADES, which provides small virtual machines to staff and users at the CNMS on an as-needed basis. These VMs can, and are, used for data processing and occasionally some more heavy parallel computation jobs. Many users utilize Colab, for access to a GPU for training, and demand for this will increase in the coming years, forcing investigation into other commercial or DOE-provided resources.

- Most of the instruments at CNMS are off-network. This presents significant challenges, and the lab has started an initiative termed INTERSECT that aims to create tools to better connect instruments to computational facilities.

- From the networking point of view, there are challenges in security data transfer rates, and physical cabling required for this to be achieved. This is because not all buildings have the same level of network speeds (ranges from 1 GBps to 10 Gbps).

- For transfer of large files, Globus is the preferred utility.

- ORNL connects to ESnet via redundant border routers at 100 Gbps. The expectation is that these connections will soon be upgraded to 400 G connections.

- ORNL utilizes a Science DMZ architecture for high-performance data transfer. Globus is the approved transfer method. A border perfSONAR node is connected to the border router and participates in ESnet perfSONAR testing.

5.10.2 CNMS Facility Profile

The CNMS is a nanoscience user facility that provides services to users in areas of nanofabrication, synthesis, theory, and characterization. Since the science is highly varied across the Center, we instead focus on one example here that links multiple groups together, from the materials synthesis to the characterization to computational and theory. Specifically, one major focus is on understanding and predicting defects in 2D and layered materials, with synthetic approaches to create and introduce defects, microscopy approaches to characterize and manipulate defects at the atomic scale, and theoretical approaches to model the observed behavior and use this knowledge for feedback to either synthesis, or to microscopy to direct atomic fabrication.

5.10.2.1 Science Background

Consider that scientific characterization tools, and in particular forms of scanning probe and electron microscopy, have been pivotal to increasing our understanding of nanoscience by enabling functional properties to me correlated with microstructural and atomic features of samples. Despite the proliferation of such tools, substantial bottlenecks exist in terms of the analysis of the data streams emanating from these tools, and the (much longer time) feedback from theoretical/simulation insights that can provide knowledge about the physical mechanisms underpinning the observed relationships and phenomena. Directly coupling the microscope data to simulations in real time to assist the experimenter is of crucial importance in the quest towards automation and autonomous materials/physics discovery platforms. A simple example of such a workflow is shown in Figure 5.10.1.

Figure 5.10.1: From Ghosh et al., npj Comp. Mater. 8, 74 (2022). (a) Deep learning models take input images and convert them into atomic coordinates. These are then used as input for simulations, which can then be run to provide insight into the physics and relevant chemistry, energetics of the different structures observed, and also provide guidance towards feedback to manipulate the structures observed towards predefined targets.

The data rates on traditional STEM are lower (~20 MBps), but the same can be applicable to 4D-STEM which can provide data rates of >100 Gbitps and which will need reduction at the edge before further processing can be accomplished. At the same time, significant computational needs beyond edge compute will be required, including GPUs for tasks such as model training (step a in Figure 5.10.1), and to run molecular dynamics codes, or cluster compute for first principles approaches such as density functional theory. These calculations can range from 1 to 10s of nodes, taking advantage of both midcapacity computing as well as DOEs HPC capabilities, including the hybrid CPU/GPU architecture available in the latter. Such calculations generate 0.5–5 GBs of data per calculation, that can quickly add up to few TBs of data with concurrent embarrassingly parallel high-throughput screening approaches. A dedicated pipeline with fast data-transfer rates connecting edge computers that are close to experiments, with such midcapacity and HPC infrastructure in national labs has the much needed, yet untapped potential, to drive specific simulations on the fly to both guide experimental investigation, as well as accelerate materials as well as fundamental physics discoveries.

The data from microscopy experiments is typically stored immediately on a NAS device, and then transferred to dedicated storage for longer term (typically about five years). Data from simulations is typically handled differently, and stored locally on storage for a few years before going to tape storage for long-term archival.

5.10.2.2 Collaborators

Collaborations for microscopy projects are typically those where the user/collaborator provide the sample to be investigated. Alternatively, the samples could be synthesized in house, and some characterization that is not available at CNMS may be performed externally at other facilities (for example, at a synchrotron such as APS for quantitative measurements of composition and structure of oxide films). Most of the data from such collaborations is first reduced before it is shared over cloud storage such as Dropbox, although in the case of large simulation files, Globus is often used to transfer files to local storage, if for example this is required to run analysis on HPC systems.

The CNMS serves over 600 unique users per year, and it would be impossible to list all collaborations here. That said, many of our users are local (coming from East TN, typically University of Tennessee, Knoxville), and then there are significant national and international collaborators from countries including Taiwan, UK, Germany, Spain, Australia, and many others. See Figures 5.10.2 and 5.10.3 for maps of the CNMS user community.

Figure 5.10.2: CNMS Users from North America
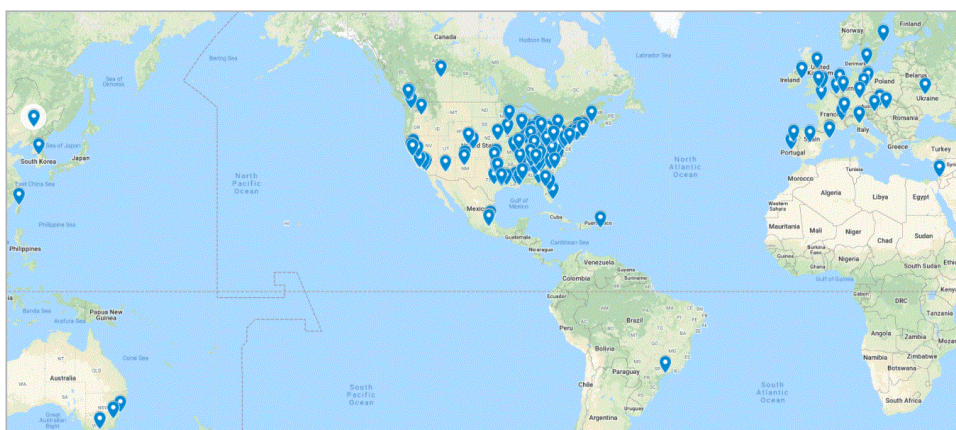
Figure 5.10.3: CNMS Users Worldwide

In FY2021 70% of our users were nonlocal, with 120 US institutions from 36 states and territories, and 43 foreign institutions from 24 foreign countries.

Most of the data that is captured at the Center for user projects is passed back to the user in the form of hard drives or uploaded and then downloaded by the user onto their home institution devices. For the theory collaborations, Globus endpoints can be used for transfer of simulation data files. We do not currently have any knowledge on the frequency of this process, or the average size of the data sets, given each type of instrument and experiment generates a different type of data set. However, in microscopy at least, most data set collections are in the 1–10GB range, with more for electron microscopy (can be >10GB for a single data set in that instance). The difficulty usually lies not in storage but in analysis, so that there is enough feedback for how to optimize measurement parameters before the user departs. For large files, data transfer can be an issue if the institutions involved do not possess Globus endpoints. In lieu of the table, we can produce the following information regarding our existing

While we do not expect significant increases in users in upcoming years (Center is essentially working at capacity), the volume of data is expected to increase.

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
DREXEL (EXAMPLE OF ADVANCED MICROSCOPY USER)	Secondary	Data transfer is through Globus to home institution	~100GB for one visit	Ad hoc	N	Getting data from instrument to the storage has been a challenge
HELMHOLTZ BERLIN (EXAMPLE OF TYPICAL MICROSCOPY USER)	Secondary	Dropbox	~20GB	Ad hoc	N	Can be slow
THEORY USER (TYPICAL) / FROM AROUND USA OR GLOBAL	Secondary	Globus from NERSC/OLCF or data-transfer nodes from ORNL-cluster to home institution,	~100s GB to ~1 TB	Ad hoc	N	Archival storage has limits for high-throughput DFT, or time-dependent dynamics simulations (e.g. 'many' large wave function files or refined molecular dynamics (MD) trajectories take up space) and archived data from tape can be slow to transfer or get corrupted over time.

Table 5.10.1: CNMS Collaboration Space

5.10.2.3 Instruments and Facilities

Theoretical research centers on understanding complex nanoscale materials spanning the full breadth of CNMS research and utilizes leadership computing as well as in-house multiteraflop compute clusters. Soon, cryo-electron microscopy capabilities will be serviced through 300kV Cryo-S/TEM (FEI Titan Krios G4) and a 120kV Cryo-TEM (FEI Talos L120C).

CNMS has a dedicated Nanophase server that is periodically backed up, but this is of limited size (10 TB). Additionally, CNMS has recently acquired a "Black pearl" storage system with 2 PB of storage and is housed in the SNS. The CNMS currently has 2 GPU systems (one Lambda server with 8 GPUs, and one DGX-2 with 16 GPUs) for dedicated model training and deployment at the edge. For running simulations, current compute capacity includes an 896 core AMD/Interlagos cluster on the OIC ORNL system and a Cray CS400 1216 core Intel Haswell cluster on the CADES Condo system. Soon, we expect an upgrade to a 24 node AMD system with ~3000 CPU cores.

For experiments, users are assigned hours to work on the instrument, but no explicit computational or storage is provided. As such, users are given access to their data when on-site, and encouraged to take it with them. For computing, we may also provide ORNL-provided virtual machines to run Jupyter notebooks on the CADES OpenStack cloud compute. These are small VMs that are typically 1 or 2 cores with 8GB RAM, although they can be expanded if needed. Given the small nature of the VM, most users still opt to process the data locally on their own laptops, or on more dedicated hardware available to them from their own institutions.

In the next two to five years, we expect that there will be more GPU-capable VMs that we can deploy for users (through services such as Jupyter Hub). Moreover, we expect more edge compute, in the form of FPGA devices, smaller-scale GPUs for low power inference (such as Jetson Nanos), and more clusters/cloud service offerings in terms of available memory, CPU core numbers, and storage. Beyond five years, our vision is more aligned with workflows for automated materials discovery, autonomous synthesis and characterization, and feedback to theory for physics discovery. Such workflows will tie together all levels of compute, require significant resource management, and are predicted to be largely autonomous.

5.10.2.4 Generalized Process of Science

Networking is critical to achieve autonomous and "smart" characterization tools, and for physics discovery, as shown in Figure 1. Essentially, data captured at synthesis and characterization tools can be processed and the experimental conditions updated, via suitable algorithms such as Bayesian optimization. However, such algorithms are compute hungry and therefore, data needs to be transferred from the instrument to the GPU cluster for computation, and then fed back to the instrument as directions for adjusting the instrument parameters to achieve optimal sampling and autonomous data capture.

Simulations that can guide and explain the observations encountered in the experiment are also critical. At this point, most simulations are performed either before or after the experiment is completed, and little to no feedback is available, either to experiment or theory. Recently we have begun exploring methods to launch jobs based on direct model inference at the characterization tool. This requires the necessary networking connections between HPC/cluster compute and the instruments themselves.

With respect to data analysis and reduction, in many instances, data reduction is achieved at the source during the experiment, i.e. with simple techniques such as signal averaging. This (reduced) data is usually stored on local drives and then given to the user, through USB keys or other media. Experimental and simulation data is not shared immediately and usually there is a considerable time lag; we do envision (as in Figure 1) that this paradigm is rapidly changing. Although standardized file structures are available for much of the simulation data, this is not universal, and the problem is exacerbated on the experimental side where every vendor chooses their own (often proprietary) standard. This makes data sharing more difficult, leading to the need to write 'converters' to a common file format, such as hdf5. Much of this work has been completed in the past few years. The future will revolve around (a) automating the pipelines for translation of the file to the standard selected, automating data processing based on the files generated, automating the theory-experiment matching, and then presenting visualization to the user so they can better understand the differences between the experiment and the simulated results, for immediate feedback to both domains. At this point, most of the data resides in separate storage boxes that are unlinked. In five years it is expected this will no longer be the case.

5.10.2.5 Remote Science Activities

Most of the remote resources used are computational in nature, e.g., utilizing compute clusters at NERSC. We are not aware of other (noncomputational) remote resources being utilized other than for storage (Google Drive and Dropbox).

5.10.2.6 Software Infrastructure

CNMS data is being transitioned to a large PB storage server. CNMS also runs a local server with about 100 TB of storage which is a windows file system that our staff can access to back up their data. CNMS is currently trialing the use of DataFed, an ORNL-developed federated data management system for keeping track of both simulation and experimental data sets. DataFed is a software solution that links together geographically diverse file storage systems and makes all of them available through the same utility. It offers considerable features relevant for scientific information flow data provenance tracking, metadata searches and includes a Python API for automating routines. DataFed can work with Globus and other tools.

We also publish data sets through ORNL's constellation application, which provides a quick interface for hosting the file, and providing a DOI. It is envisioned that this will be integrated into DataFed in the future.

For the experimental side, we are trialing the use of a gateway device to transfer files from the instruments to the data storage system, with the files automatically ingested by DataFed so that the relevant metadata is also captured, thus enabling keyword searches by metadata fields. The data storage is local to ORNL. Any remote data sets can be accessed through collaborators through either Globus endpoints or through scp/ftp, or through the DataFed interface (if the collaborator also utilizes DataFed).

For analysis of spectral and imaging data sets, the CNMS, with partners, developed the pycroscopy ecosystem of packages (github.com/pycroscopy). This software effort has amassed over 200k downloads in five years, and now consists of ~ a dozen important packages for different types of data analysis. All the packages are written in Python and are open-sourced. This ecosystem is divided primarily into three distinct areas, (1) Data input/output and utilities, (2) Generic imaging and spectroscopy analysis, and (3) More focused and domain-specific packages, as shown in Figure 1. The general feature of the ecosystem is a common 'currency' of data object (the sidpy data set object) that enables efficient processing, storage and visualization.
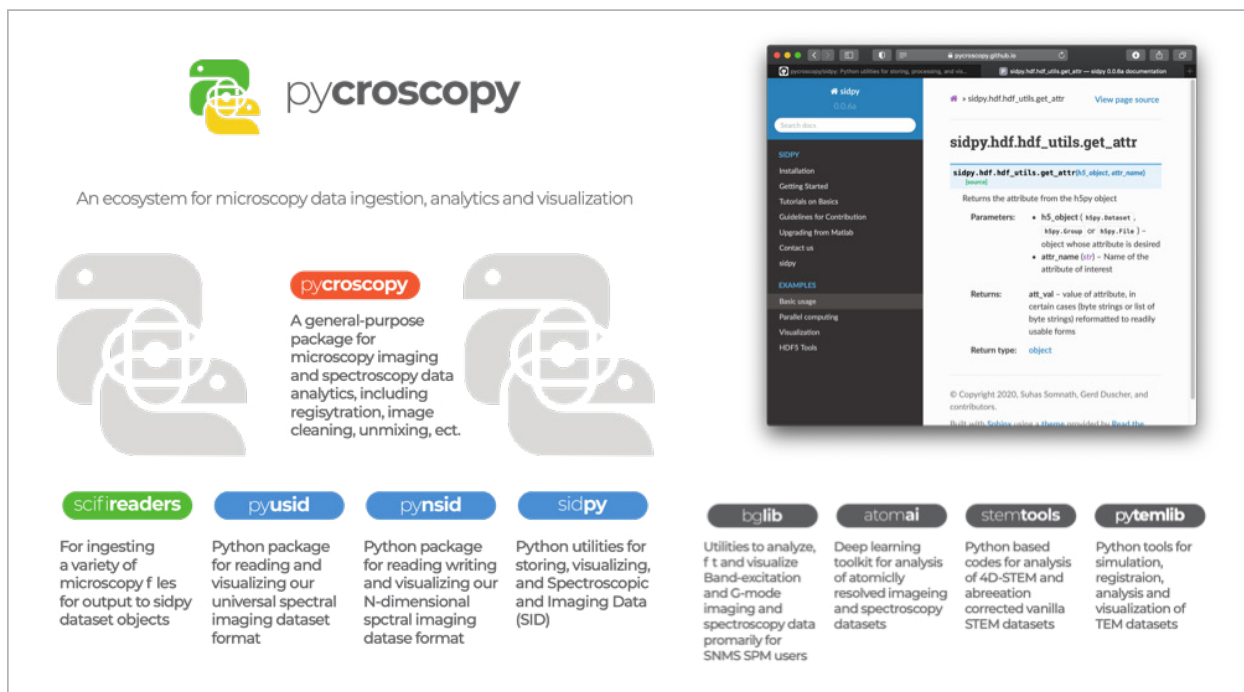
Figure 5.10.4: Flow chart of using a digital twin with light source experiments

5.10.2.7 Network and Data Architecture

ORNL connects to ESnet via redundant border routers, each of which currently connects to a diverse ESnet router at 100G. The expectation is that these connections will soon be upgraded to 400G connections. The ORNL border routers connect the ORNL Enterprise network, which includes SNS, CNMS and HFIR, with the OLCF and ESnet. This connectivity is depicted in Figure 5.10.4.
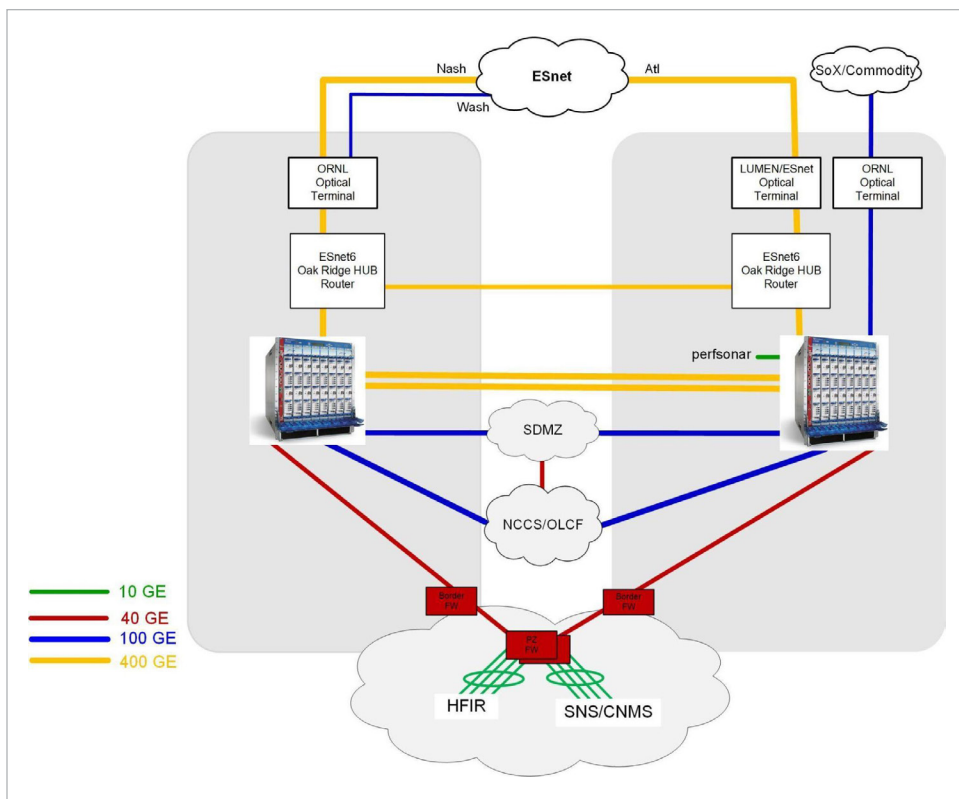
Figure 5.10.5: ORNL Network Diagram

The Enterprise firewalls are connected to the border routers at 40G currently. Connections to the Enterprise core from SNS/CNMS and HFIR are also 40G currently. Upgrades to these core switches are in progress and uplinks from SNS/CMS are expected to become 100G early FY23.

ORNL does utilize a Science DMZ architecture for high-performance data transfer. This environment connects to the border routers with 10/40/100G DTN connections available. Globus is the approved transfer method. A border perfSONAR node is connected to the border router and participates in the ESnet grid.

Most of the instruments at CNMS are off-network. This presents significant challenges, and the lab has started an initiative termed INTERSECT that aims to create tools to better connect instruments to computational facilities to enable "smart" instruments, labs of the future, etc. From the networking point of view, there are challenges in security data transfer rates (especially for advanced electron microscopy with fast camera detectors), and physical cabling required for this to be achieved. This is because not all buildings have the same level of network speeds

5.10.2.8 Cloud Services

Currently the CNMS makes limited use of cloud services, but this is expected to change. Our cloud service is a locally run OpenStack run by CADES, which provides small virtual machines to staff and users at the CNMS on an as-needed basis. These VMs can, and are, used for data processing and occasionally some more heavy parallel computation jobs (e.g., those requiring ~64 CPU cores), for special requests. Many of our users also utilize Colab, for access to a GPU for training. It is expected that as GPU usage continues to increase, either our local cloud compute capabilities will grow to accommodate this demand, or our users will shift to external cloud services such as AWS or Azure. This may run into the dozens of instances/year level. The CNMS is procuring two cloud infrastructure nodes FY22 to begin to address this need.

5.10.2.9 Data-Related Resource Constraints

For most of the facility, the data storage and transfer speeds are currently sufficient, and are not expected to be significant concerns in the next two years (perhaps with the exception of very particular experiments, such as 4D STEM). The main constraints are with the access to computation, which is particularly limited for the theory groups, and the changing nature of the HPC platforms which make older code quickly obsolete, requiring significant rewriting to accommodate new hardware upgrades. These are perhaps the most significant issues.

5.10.2.10 Outstanding Issues

None to report at this time.

5.10.2.11 Facility Profile Contributors

CNMS Representation

- **Rama Vasudevan,** *ORNL,* vasudevanrk@ornl.gov
- **Panchapakesan Ganesh,** *ORNL,* ganeshp@ornl.gov
- **Bobby Sumpter,** *ORNL,* sumpterbg@ornl.gov

ESCC Representation

- **Susan Hicks,** *ORNL,* hicksse@ornl.gov

5.11 The Molecular Foundry (TMF)

Supported by the DOE Office of BES through their NSRC program, TMF is one of five national User Facilities for nanoscale science that serves roughly 1000 academic, industrial and government scientists around the world each year. Users come to TMF to perform multidisciplinary research beyond the reach of their own laboratory. By taking advantage of TMF's broad spectrum of core capabilities and expertise, users increase the scope, technical depth, and impact of their research. Moreover, while at TMF, users' access LBNL's diverse scientific community that includes other user facilities.

5.11.1 Discussion Summary

- TMF is one of five national User Facilities for nanoscale science that serves roughly 1000 academic, industrial and government scientists around the world each year. Users come to TMF to perform multidisciplinary research beyond the reach of their own laboratory.

- TMF included 28 independent researchers, 18 technical staff.

- There are seven facilities at TMF:
  — Imaging and Manipulation of Nanostructures
  — Nanofabrication Facility
  — Theory of Nanostructured Materials Facility
  — Inorganic Nanostructures Facility
  — Biological Nanostructures Facility
  — Organic and Macromolecular Synthesis Facility
  — NCEM

- COVID 19 has greatly sped up our remote activities. Tools like VNC and Microsoft Remote Desktop are used to remotely log into the control PCs that power microscopes. Due to the inherent latency of remote use, this makes operating the microscope remotely challenging. Remote cameras and telepresence robots are also available remotely, and have been successfully used.

- TMF has users from around the world who come and generate data at our facility:
  — Most data is stored at the facility for least five years (and best effort beyond)
  — There are no facility-wide network drives or cloud-based systems to remotely access or back up data
  — Each user is encouraged to take a copy of their data; most do via the use of cloud storage or external media transfer
  — There is no general data access method: some use removable media, others store data at the facility and operate on it remotely, and some may use the local computing cluster
  — Data size can range from kilobyte text files to 700 GB raw binary files

- The data lifecycle heavily depends on the facility and equipment:
  ○ Facilities that focus on synthesis generate information about reactions and reactants.
  ○ Facilities such as NCEM and the Imaging floor generate many terabytes of data from detectors for imaging, diffraction and spectroscopy.

- NCEM has 9 electron microscopes of which four are large data generators:
  - TEAM I generates data sets that range in size from 10–50 GB over the course of 10 minutes, and several can be generated per day.
  - TEAM 0.5 generates data at 480 Gbit/s, with maximum data sets being 700 GB over a 15 seconds run. To support this use case the data must either be reduced locally, or transferred directly to NERSC (via dedicated 100 Gbps campus networking). It can produce 20 data sets (14 TBs) per day, but this is currently limited by the data reduction time more than the ability to use the instrument. Data generation will increase by 25% in the one-year timeframe, after an upgrade.
  - TitanX generates data in the 100 MB–2 GB range. Several of these can be acquired per day.
  - Themis generates data sets in the range of 10–50 GB in a few minutes. Several of these are acquired per day.
  - A new TEM, similar to TEAM I will arrive in one to two years.
  - A new TEM, similar to TEAM 0.5, will arrive in three+ years.
- Other parts of TMF generate GB to TB of data per year, and feature heterogeneous methods for data storage, sharing, and processing.
- Access to data generated by BES user facilities is typically governed by the user agreement, a legally binding document. In many (most) cases, the data is owned by the users, not the facility. In some instances, data is open and available, given there are no restrictions on the collected data beyond understanding how it was taken and managed.
- We utilize local computers with VNC or RDP access for some data analysis. Utilization of clusters and HPC is just starting and will become more important, but this is hampered by our ability to get data to these facilities off-site. It is also hampered by a lack of expertise in data processing in the user community.
- Data analysis happens using proprietary (graphical based) software and Jupyter. TMF provides access to proprietary software through dedicated computers, with remote access. Python analysis is done on the users' own computers or our computers using open-source software and algorithms developed at our facility or within the field of TEM.
- Remote data analysis is also a place for great improvement. Inspired by Jupyter@NERSC, TMF has deployed a couple of local machines for JupyterLab remote processing of data. This is very likely to be an important capability going forward as data analysis is an important bottleneck in our workflows.
- Simulations are sometimes done, but this is not common. TMF would like to make this easier and more common. Simulations are largely compute dependent with very small input values and are now GPU accelerated.
- NCEM microscopes are connected to local NAS with ~450 TB of capacity, using an internal 1Gbps LAN. Workstations use direct file mounts for data mobility. LAN network capacity is expandable to support 10 Gbps and 40 Gbps networking in the future.
- Some parts of TMF have semicentralized data storage (e.g., NAS for workstations, or institutional cloud resources at Google) which can be used temporarily. This allows the usage of TMF computational and software resources locally and remotely. Others may have access to "support PCs" that link instruments and have limited resources that can facilitate data taking, sharing, and temporary storage.

- TMF would like to provide facility-wide compute and storage resources, but this has been cost prohibitive both from a capital and operating expense perspective.

- TMF users sometimes have DOE HPC allocations. Data transfer between TMF's internal compute resources (managed by the High-Performance Computing Services (HPCS) group at LBNL) and DOE HPC centers, such as NERSC, is frequently required.

- Data sharing is extremely difficult. Many users come through the facility, and keeping track of whose data is where is far too time consuming. None of this has been automated, as it takes a dedicated team of professionals to keep such a system running.

- The center does not use any official cloud services. Users are known to use Google Drive (and other similar services) to store and move data.

- A Globus server with Science VM service that features 10 TB storage has been allocated to facilitate sharing data between TMF, ALS and JBEI.

- For large data sets, data storage and movement are bottlenecks in our current workflows.

- The local NAS at NCEM is ~450 TB and 90% full at all times with older data being manually deleted as needed.

- There is currently no standard back-up management system for TMF or for the data generated with our compute resources, nor IT staff with this as part of their remit.

- TMF is mostly networked with 1Gbps LAN networks, with some 10 Gbps connections to support point-to-point connections with larger data rates. There is a special 100 Gbps connection to NERSC to support the NCEM TEAM 0.5 instrument.

- In late 2022 TMF will commission a building-wide fiber network with 10, 40, 100 Gbit connections for each microscope and detector. The building has a 40 Gbit uplink that can be expanded. Data generation and movement will continue to grow exponentially as new microscopes and detectors with large data outputs are installed.

5.11.2 TMF Facility Profile

TMF features world-class scientists with expertise across a broad range of disciplines and state-of-the-art, often one-of-a-kind, instrumentation. TMF staff have developed new materials, tools, techniques, codes, and fundamental understanding across all fields of science, and their expertise is highly sought out by the scientific community. One strength is TMF's suite of custom automated synthesis robots, which are a pillar of TMF's efforts towards accelerated materials discovery, particularly in the synthesis of colloidal nanocrystals and bio-inspired peptoid polymers. TMF is also home to the TEAM microscopes, two of the most powerful electron microscopes in the world, which have been the platform for revolutionary advances in electron detectors, as well as other new developments and characterization techniques. Additionally, TMF has long been a leader in sub-10 nm fabrication, and the simulation and interpretation of X-ray spectroscopy experiments.

5.11.2.1 Science Background

TMF staff spend at least half their time working with outside users that are selected through an external, peer-review process. Staff then devote the remainder of their time to internal research activities, which can be augmented with postdoctoral fellows hired using internal or external grant support. Internal research programs advance the frontiers of nanoscale science by developing new capabilities that are made available to users. These advances are often inspired and supported by partnership with the user community at TMF and its fellow NSRCs and serve to attract new user groups once established: many new TMF capabilities arise out of synergistic projects with users.

Organized into seven interdependent research facilities that support crosscutting scientific themes, TMF provides access to state-of-the-art instrumentation, unique scientific expertise, and specialized techniques to help users address myriad challenges in nanoscience and nanotechnology. At the start of FY22, TMF included 28 independent researchers, 18 technical staff.

Imaging and Manipulation of Nanostructures

This facility develops and provides access to state-of-the-art characterization and manipulation of nanostructured materials — from "hard" to very "soft" matter — including electron, optical, and scanning probe microscopies.

Nanofabrication Facility

This facility focuses on understanding and applying advanced lithographies, thin film deposition, and characterization, emphasizing integration of inorganic, organic, and biological nanosystems with the potential for nanoelectronic, nanophotonic, and energy applications.

Theory of Nanostructured Materials Facility

This facility expands understanding of material systems and phenomena from molecules to nanoscale assemblies through the development and application of theories and computational methods that describe: energy and information management; pattern formation and nonequilibrium dynamics; and the interpretation and prediction of experimental observables in complex nanostructured systems.

Inorganic Nanostructures Facility

This facility is devoted to the science of semiconductor, carbon and hybrid nanostructures—including design, synthesis, and accelerated discovery of nanocrystals, nanowires, and nanotubes and their self-assembly into 3D mesoscale functional materials for use-inspired energy applications.

Biological Nanostructures Facility

This facility designs and synthesizes new materials based on the self-assembly of biopolymers and bio-inspired polymers, creates new nanocrystal probes for bioimaging, and develops synthetic biology techniques to reengineer organisms and create hybrid biomolecules to interface with a variety of applications.

Organic and Macromolecular Synthesis Facility

This facility studies "soft" materials, including the synthesis of organic molecules, macromolecules, polymers and their assemblies, with access to functional systems, photoactive materials, organic-inorganic hybrid structures, and porous materials.

NCEM

A world-renowned center for microscopy since 1983, and integrated into TMF in 2014, this facility features cutting-edge instrumentation, techniques and expertise required for exceptionally high-resolution imaging and analytical characterization of a broad array of materials.

The data lifecycle at TMF heavily depends on the facility and equipment that was used to acquire/generate the data. Facilities that focus on synthesis generate information about reactions and reactants which are being augmented with robots capable of running hundreds of reactions at one time. Other facilities such as NCEM and the Imaging floor generate many terabytes of data from detectors for imaging, diffraction and spectroscopy. Thus, we would have interest in the storage and retrieval of a wide range of types of data. No chain of custody or ownership is currently applied to these data sets. For example, at the NCEM facility all data is open to all users who can access the storage computers. This has not been a problem because without intimate knowledge of how

the data was taken and the sample it was taken from it is essentially useless. This points towards the ability to capture, store and catalog metadata as an extremely important part of the future of science at TMF.

TMF's Theory Facility has its own computing resources and pay per use resources managed by Berkeley Lab's IT Division and makes use of national resources such as NERSC. Users and Staff of the Theory Facility generate simulated data on these computing resources over a wide spectrum of rates/sizes — up to TBs per simulation. We make use of storage attached to our supercomputing resources and permit users to download their data to home institutions as they wish. User data is privately maintained according to UNIX permissions, with some users providing open access to their data to facilitate assistance with analysis by TMF Staff.

Compute Resources at TMF

TMF Compute Cluster resources include:

- ETNA: a 184-node cluster connected through a high-performance Mellanox 56 Gbps FDR Infiniband fabric. Etna nodes predominantly comprise 24-core (2x12) 2.3GHz Intel Xeon processors with 2.67GB per core, with an additional 3 Xeon Phi nodes, and 9 GPU nodes with 4 Tesla K80 GPUs each. The cluster nodes have access to a high speed 169.8 TB LUSTRE parallel file system. Etna has a theoretical performance of 158.7Tflops and was purchased by TMF using recapitalization funds in October 2016.

- VULCAN: a 278-node cluster with a QLogic 40 Gbps QDR Infiniband interconnect connected to a 41.3 TB LUSTRE parallel file system. Vulcan nodes come in two types: predominantly (242-nodes) those with two 2.4GHz Intel Xeon E5530 Quad-core Nehalem processors with 3 GB RAM per core; (4-nodes) 16-core Intel Xeon with 4 Tesla Nvidia GPUs per node; (24-nodes) 20-core Intel Xeon. Vulcan is also connected to an additional 57.0 TB BlueArc NFS file system. Theoretical performance is 18.1TFLOPS with 5.7 TB of total memory. Vulcan was provided to TMF by American Recovery and Reinvestment Act funds in Jan, 2010.

- NANO: 232-core Intel Xeon processor machine networked with high-speed, low-latency Infiniband interconnects; it has 455 GB of total memory and uses a 10.1 TB Panasas parallel file system. Nano was provided to TMF at its inception and expanded regularly during its first three years of operation.

Etna, Vulcan, and Nano are used exclusively by Theory Facility Staff and Users. Ideal for exploratory research, they provide the Theory Facility with the flexibility to address exciting new problems as they arise, permit fast turnaround for development projects, and are scalable for future expansion.

All of these computing resources are housed, installed, administered and maintained by the HPCS group within the IT Division at LBNL.

TMF users sometimes have additional NERSC energy research computing allocations process allocations for HPC tasks, which can also contribute to the aims and objectives of the facilities and the user communities. These allocations are handled by a separate peer-reviewed process by NERSC. Data transfer between TMF's internal compute resources (managed by the HPCS group at LBNL) and NERSC is frequently required.

5.11.2.2 Collaborators

TMF has users from around the world who come and generate data at our facility:

- In general, most data is stored at the facility with the understanding that we make the best effort to retain data for at least five years.

- We currently have no facility-wide network drives or cloud-based systems to remotely access or back up data, and we are not actively planning to implement such a system.

- We highly encourage each user to take a copy of their data and almost all do using Cloud storage or personal hard disk drives (HDDs).

- We have no general data access method and many different ways of doing it. Some take data home on HDDs, some store them at our facility to operate on remotely, some store data on TMF cluster. It is a wide range.

- Data size can range from kilobyte text files to 700 GB raw binary files. Many files are proprietary and generated by detectors with company software and formats.

- Data sharing is extremely difficult. We have many users come through the facility and keeping track of whose data is where is far too time consuming. None of this has been automated as it takes a dedicated team of professionals to keep such a system running.

The Theory Facility has centralized data resources defined by storage collocated with supercomputing resources, NAS for workstations, and (recently, provided by institutional subscription) Google Drive. Users and collaborators typically leave their raw data on these "local" resources for easier access to further calculation or TMF postprocessing tools. Generally, to complete publications, a minimal projection of this data is required for download to the user/collaborator home institution. The Theory Facility aims to facilitate and reinforce this model by providing Jupyter notebooks to access and process locally stored data and convert it to publishable figures.

Within NCEM, all data resides on "support PCs" which can be connected to the Internet and also on a local ~140 TB NAS administered by our staff. Most users take their data with them on hard drives but some leave very large data sets on the NAS to be analyzed later (locally or remotely). We would like to get away from this, but any solution considered and proposed has been significantly more expensive (5–10x) than having us locally install, maintain, and administer it.

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
INTERNATIONAL	Primary	ssh terminal, OpenOnDemand, Jupyter	1kB–1TB	very small data transfer weekly, very large rarely	not offten	collaboration on data limited due to UNIX permissions defaults

Table 5.11.1: TMF Collaboration Space

5.11.2.3 Instruments and Facilities

The equipment and requirements are different for each facility. Here is a short breakdown of the largest data generators.

NCEM

- 9 electron microscopes, of which 4 are large data generators.

- TEAM I has a detector that generates data sets on the scale of 10–50 GB which can be acquired in tens of minutes each. Several will be acquired in a day.

- TEAM 0.5 has a detector that generates data at 480 Gbitps. Maximum data sets are 700 GB acquired in 15 seconds. This data is reduced locally (10 minutes) or pulled to NERSC via dedicated 100 Gbit fiber (~3 GBps real world maximum) and reduced in < 5 minutes. The detector currently generates about 20 data sets (14 TBs) per day which is currently limited by the data reduction time (i.e. users would take more data if possible).

  — This detector will increase in data generation by 25% in the 1 year timeframe

- TitanX has a detector that generates data in and 100 MB–2 GB range. Several of these can be acquired per day.

- Themis has two detectors that can generate data sets in the range of 10–50 GB in a few minutes. Several of these are acquired per day.

- All microscopes are connected to a local NAS with ~450 TB of storage over internal 1 Gbit connections. We mount the drives using samba to copy data. 2 workstations are connected to the NAS by 10 Gbit links for local processing.

- A new TEM will soon be acquired (one- to two-year timeframe) which will have similar output to the TEAM I above.

- Internal network: all microscopes have recently been connected to a fiber network for internal data movement. This is still being built out but should provide > 10 Gbit links for all equipment with a >= 40 Gbit uplink to LBNLnet in the works.

- A second large data generating detector (> 480 Gbitps) is being built and should be operational in the coming three-year timeframe.

Imaging and Manipulation

- Diverse set of microscopes and other instruments from commercial and custom built with Windows based PC control.

- Locally generated data is saved to a network backup disk.

- Generation of tens of TB yearly, but expected to grow.

- Some data is proprietary but others produce HDF5.

Nanofabrication

- Many instruments. Very little data is generated or stored.

Biological Nanostructures

- 3 x optical microscopes and 3 x mass spectrometers which generate 1 Gigabyte.

- These generate Gigabytes per experiment which last about a day.

- Very similar to TEM data generation (single images and movies).

- Largest data generator is the superresolution microscope. One user has a 1 TB folder.

- A couple of computers connected by 10 Gbit for data movement and storage and analysis.

- Recently implemented a Globus server with Science VM service with 10 TB storage sharing data between TMF, ALS and JBEI.

Inorganic Nanostructures

- Five synthesis and characterization robots.

- >20 characterization tools, including spectrometers, diffractometers, microscopes, electrical testing setups, elemental analysis tools, fume hoods for synthesis.

- Data is stored locally or in online databases (Symyx lab execution and analysis (LEA) server, ESCALATE).

- Robots generate more data like spectra, X-ray diffraction patterns, images, recipes in XML files, etc.

- About 1–10 TB / year in total for robots.

Theory

- This facility maintains a cluster for simulations by staff and users. It is described previously and managed by IT.

Organic

- ~ 16 instruments in total running on Windows. 1/3 are Windows 7 or earlier
- All data stored locally on computers that generate the data

5.11.2.4 Generalized Process of Science

NCEM

In general, experiments have the goal of understanding the atomic structure, electronic structure, or composition of materials. We have two ways that users typically use our facility. Local long-term users and users who travel to the site from out of town/state/country. Every user brings their own samples which range very widely in their type. This makes it hard to generalize our workflows since each sample and experiment is different. Further, transmission electron microscopes are highly configurable which means data can be generated in many different ways such as imaging, diffraction and spectroscopy. Microscope experiments are generally booked on the half day or full day depending on the complexity.

Data analysis happens using proprietary (graphical based) software and Jupyter. We provide access to proprietary software through about 5 Windows computers now with remote access. Python analysis is done on the users' own computers or our computers using open-source software and algorithms developed at our facility or within the field of TEM.

Simulations are also sometimes done, but this is not common. We would like to make this easier and more common. Simulations are largely compute dependent with very small input values and are now GPU accelerated.

All data resides on "support PCs" which can be connected to the Internet and also on a local ~140 TB NAS administered by our staff. Most users take their data with them on hard drives but some leave very large data sets on the NAS to be analyzed later (locally or remotely). We would like to get away from this, but any solution considered and proposed has been significantly more expensive (5–10x) than having us locally install, maintain, and administer it.

Our recent network upgrade (> 10 Gbit) for all microscopes will improve our ability to move data around and out of the facility. This will likely push us to adopt newer technologies for data maintenance and lead to the generation of even more data.

Bio / Organic / Inorganic

Data is processed locally proprietary software on Windows computers, with final results including graphs and images and model fitting. Materials that are synthesized are documented in lab notebooks with the addition of some small analytical files like pictures or metadata. Most data that is generated is done during postanalysis. There is generally a need for centralized data storage linked to a user account and/or proposal.

Inorganic

The Inorganic Nanostructures Facility offers facilities for the manual and automated synthesis and characterization of novel inorganic materials, including colloidal nanoparticles, two-dimensional materials, and hybrid inorganic-organic frameworks.

Data generated in this facility includes:

1. synthetic recipes and process observations recorded in physical laboratory notebooks;
2. raw characterization data such as spectra, images generated by optical/Raman/and electron microscopes, X-ray diffraction patterns, and electrical measurement data.

Data is predominantly stored on local instrument computers; raw simulation data; and processed data stored on staff computers. In some cases, this data is stored in more centralized electronic repositories, including: electronic laboratory notebooks, e.g., Evernote and OneNote; personal and shared cloud folders, e.g., Google Drive; ScopeFoundry, as noted in the Imaging Facility plan, and local and remote servers for Python Jupyter notebooks.

In other instances, data is stored in centralized databases, such as ESCALATE and a commercial Laboratory Information Management System (LIMS) called Laboratory Execution and Analysis for generating, capturing, and storing data associated with two of its robotic synthesis platforms, HERMAN and WANDA. LEA has a modular architecture that stores data in a centralized LBNL Oracle database mediated by a TMF-based midlevel application server that mediates the transfer of data and metadata from client (robot) computers. Experimental designs are automatically stored in the central database, along with process logs and, in some cases, characterization data. A strength of this system is its Library Studio software that allows for complex chemical recipes to be designed and encoded (in XML) by users unfamiliar with automation. This robust platform is proprietary. Since this closed system does not have API's for open-source programming environments, it is difficult to integrate with ML algorithms and autonomous systems.

Centralization of the databases frees staff from having to maintain complex data architectures, but the disadvantage is that user data is at the mercy of policies outside of the control of scientific staff.

Theory

Very similar to other HPC cluster needs for materials physics simulations

Imaging

Every user brings their own samples which range very widely in their type. This makes it hard to generalize the workflows since each sample and experiment is different. Further, our microscopes are highly configurable which means data can be generated in many different ways such as imaging, spectroscopy, or hyperspectral images. Most microscopes use a custom ScopeFoundry instrument control code which outputs HDF5 data sets that include microscope metadata.

Data analysis happens using proprietary (graphical based) software and Jupyter. We provide access to proprietary software through about 5 Windows computers now with remote access. Python analysis is done on the users' own computers or our computers using open-source software.

5.11.2.5 Remote Science Activities

COVID 19 has greatly sped up our remote activities. We are using VNC type technologies to remotely log into microscopes. For Electron microscopes running a microscope remotely like this is very hard though. The most useful outcome is likely the ability to screen share an experiment session with remote collaborators (i.e., synthesis experts or data analysis experts or the project PI) to collaboratively work on an experiment together.

- For example, having more than 2 people in a microscope room is "too many cooks in the kitchen."
- For other microscopes (e.g., Optical spectroscopy) users often will setup their sample and do final alignments in person, then remotely control microscope for the duration of their experiment (1 hour– 3 days).

Microsoft Remote Desktop is also used to log into remote instruments to control experiments and analyze/extract data. Remote cameras and telepresence robots are used to monitor instruments. Remote data analysis is also a place for great improvement. Inspired by Jupyter@NERSC we have deployed a couple of local machines for JupyterLab remote processing of data. This is very likely to be an important capability going forward as data analysis is an important bottleneck in our workflows. Most users take their data home with them to analyze separately.

5.11.2.6 Software Infrastructure

- We commonly use proprietary software installed on Windows computers and allow for remote access to these machines (RDP and VNC). We are starting to use open-source software and algorithms based on Python, Numpy, and Jupyter.

- Theory mostly uses DFT and MD simulation software (both open-source and commercial) installed on the TMF clusters. Globus is often used to transfer large amounts of data.

5.11.2.7 Network and Data Architecture

TMF in building 67 is wired with 1 Gbit. Some dedicated 10 GBE lines have been run for specific point-to-point connections.

At NCEM in building 72: prior to 2022 almost all data movement has been at 1Gbit speeds. Exceptions are 1 microscope with a direct 10Gbit connection to a local NAS and 1 detector has a direct 100 Gbit connection to NERSC. In summer 2022 we will commission a building-wide fiber network with 10, 40, 100 Gbit connections for each microscope and detector. The building has a 40 Gbit uplink and this might also be expanded. We expect data generation and movement to continue to grow exponentially as new microscopes and detectors with large data outputs are installed.

5.11.2.8 Cloud Services

The center does not use any official cloud services. Users are known to use Google Drive (and other similar services) to store and move data. Druva is a common PC backup service used at the center.

5.11.2.9 Data-Related Resource Constraints

- In general users transport large amounts of data on personal HDDs is the current operational mode. We are interested in a more centralized location like a NAS with high-speed network connectivity.

- Science is also usually only captured in handwritten notebooks. A digital lab notebook would be an obvious next step. This will be hard to get users to adopt since they come from many different institutions and might have their own preferred software.

- For large data sets, data storage and movement are bottlenecks in our current workflows.

- We utilize local computers with VNC or RDP access for some data analysis. Utilization of clusters and HPC is just starting and will become more important, but this is hampered by our ability to get data to these facilities off-site. It is also hampered by a lack of expertise in data processing in the user community.

- The local NAS at NCEM is ~450 TB and 90% full at all times with older data being manually deleted as needed.

- There is currently no standard back-up management system for TMF or for the data generated with our compute resources, nor IT staff with this as part of their remit.

5.11.2.10 Outstanding Issues

None to report at this time.

5.11.2.11 Facility Profile Contributors

TMF Representation

- **Kristin Persson**, *LBNL*, kapersson@lbl.gov
- **Peter Ercius**, *LBNL*, percius@lbl.gov
- **Emory Chan**, *LBNL*, emchan@lbl.gov
- **David Prendergast**, *LBNL*, dgprendergast@lbl.gov
- **Edward Barnard**, *LBNL*, ESBarnard@lbl.gov
- **Sinead Griffin**, *LBNL*, SGriffin@lbl.gov

ESCC Representation

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

5.12 Autonomous Experiment Steering for BES Facilities

The autonomous experiment steering case study brings together BES community representatives from X-ray light sources, neutron scattering facilities, NSRCs, and independent researchers, to discuss this emerging area of scientific investigation. The case study profiles the overall concept, some early results, and a discussion about what the future will hold.

5.12.1 Discussion Summary

- Experimental facilities have a challenging task: equipment is advanced and acquisition rates are increasing, yet the complexity of scientific questions is also increasing meaning that it becomes less possible to address questions without parallel advances in experimental design.

- Modern scientific instruments, combined with the use of robotics and high-throughput workflows for sample preparation/loading, can acquire measurements at ever-increasing rates and resolutions. Autonomous experiment steering leverages these advantages by addressing the bottlenecks associated with data processing and automatic decision-making, thus enabling closed-loop experiments through efficient data analytics and advanced AI/ML approaches.

- Autonomous and automated experiments will become heavily adopted in the coming years, due to increases in efficiency, elimination of analysis and decision-making bottlenecks. This will enable physics discovery in materials design, synthesis, and characterization loops.

- In typical closed-loop experiments, raw data is collected, handled, and stored at the location they are generated using instruments maintained by the experimental teams at their home institutions. Depending on the computational power required for data analysis, preprocessed data is either analyzed locally or shared across divisions/facilities that might be physically located at great distances. Data from multiple sources might need to be gathered at a centralized location for multimodal analysis, and transferred to facilitate AI/ML based decision-making. This data is also archived for long-term access at database facilities, which ideally should provide data access control management.

- Along with the ongoing development of digital twins, it is expected that HPC systems will be one of the major data sources for future experimentation. Simulations are used in autonomous experiments for synthetic data generation to supplement experimental data. Distributed experiments have been proposed which leverage simulations using HPC infrastructure and parallel data acquisition at multiple facilities internationally.

- Autonomous intelligent decision-making (instead of automated decision-making) will be the one limiting factor of future self-driving labs. Most successful approaches are based on HPC-driven UQ, this will require readily accessible allocation of HPC resources and communication infrastructure.

5.12.2 Autonomous Experiment Steering for BES Facilities Case Study

Modern scientific instruments combined with the use of robotics and high-throughput workflows for sample preparation/loading can acquire measurements at ever-increasing rates and resolutions. Autonomous experiment steering leverages these advantages by addressing the bottlenecks associated with data processing and informed/automatic decision-making, thus enabling closed-loop experiments through efficient data analytics and advanced AI and ML approaches. Autonomous experiments will revolutionize traditional scientific methods and accelerate scientific discoveries orders of magnitude. When real-time data analysis can be performed, some intelligent decision can be made that shorten the overall experimental time by choosing the most impactful/important data to collect.

5.12.2.1 Science Background

In typical closed-loop experiments, raw data is collected using different scientific instruments maintained by the experimental teams at their home institutions. The rate and duration at which data is collected is highly dependent on the type of instruments and nature of the experiments, which can range from 10s to 10,000s of samples in hours to weeks. These experiments might be performed in parallel, or they might exhibit complex dependencies that require specific execution orders. The raw data is usually handled and stored locally at the location where they are generated. Depending on the computational power required for data analysis, preprocessed data is either analyzed locally or shared across divisions/facilities that might be physically located at great distances. Data from multiple sources might need to be gathered at a centralized location for multimodal analysis, and data might be transformed into various forms, as objective scores, images, trends/patterns, surrogate models to facilitate AI/ML based decision-making. Knowledge such as reaction mechanisms, materials properties, structure/morphologies are extracted from these experiments and are used for publication. This data is also archived for long-term access at database facilities, which ideally should provide data access control management.

Along with the ongoing development of digital twins, it is expected that HPC systems will be one of the data sources. Simulations are used in autonomous experiments for synthetic data generation to supplement experimental data. Experimental runs can also be used to tune and modify simulations to provide more meaningful results, and further the experimental design optimization process or shorten the experimental time by acquiring "intelligent" data. Distributed experiments have been proposed which leverage simulations using HPC infrastructure and parallel data acquisition at multiple facilities internationally. We expect that autonomous and automated experiments will become heavily adopted in the coming five years, due to increases in acquisition efficiency and quality, capability to perform new science that was previously infeasible due to analysis and decision-making bottlenecks, and enable physics discovery in materials design, synthesis and characterization loops.

Experimental facilities around the globe are facing a challenging task: while equipment is becoming increasingly advanced, leading to a steady rise in data acquisition rates, facilities are still outpaced by the increase in complexity of scientific questions and therefore unable to reliably answer them without parallel advances in experiment design and adequate computing infrastructure. The rise in complexity leads to high-dimensional parameter spaces, spanned by the parameters describing the sample, the instrument, and the data acquisition protocol, embracing the full range of synthesis, processing, and environmental parameters that describe the sample and its characterization throughout the design-experiment loop. These parameter spaces have to be explored efficiently in search of new scientific discoveries. Traditional methods address the rise in dimensionality by checking more and more possible configurations, counting on an increase in data-acquisition rates. This "brute-force" approach takes full advantage of advanced computing facilities for computation, storage, and novel detector technology with higher speed and dimensions (i.e., several millions of pixels). In a standard approach, a Cartesian grid is often defined, which is then used to automatically control and schedule measurements. This approach is often referred to as scanning or raster scanning. Since little information about the model is known beforehand, the grid is often defined to be very fine, which leads to long periods of data collection, and a significant amount of redundant data is often collected, processed, and stored. Another approach is to perform a coarse-grid scan first and then, based on the practitioner's input, focus on subregions of the parameter space using a fine-grid scan. This method needs human intervention and potentially leads to bias, lost information, and overly redundant data. As the dimensionality of the parameter space increases, grid-based approaches become increasingly impractical, since the number of grid points scales with the power of the dimension of the space — even a "simple" problem inhabiting a ten-dimensional parameter space is prohibitively expensive under such a brute-force approach. In the case of high-dimensional parameter spaces, practitioners often change their approach to an intuition-based technique, in which, after just a few measurements are acquired, the user attempts to make out patterns and trends in the data, which will then be used to steer the experiment. While data is collected more deliberately, this approach leads to highly trained scientists micromanaging the experiment, while choosing measurements suboptimally; after all, human brains are not well-equipped for pattern recognition

and decision-making in high-dimensional spaces. This results in the need for constant vigilance and attention from the experiment designer, as well as the expectation that the user will be able to interpret and integrate the current results on the basis of all the previous measurements. Additionally, user bias can creep in, in which the expectation that results should look a certain way, or that unexplored regions probably will not be interesting, can skew the investigation.

Autonomous intelligent decision-making (instead of automated decision-making) will be the one limiting factor of future self-driving laboratories and user facilities. Most successful approaches are based on HPC-driven UQ that can approximate functions in extremely data-sparse scenarios. This will require readily accessible allocation of HPC resources and communication infrastructure.

5.12.2.2 Collaborators

Complex scientific questions often require participation of researchers from several different institutions. This section highlights a few of such collaborative efforts.

The Center for Advanced Mathematics for Energy Research Application's (CAMERA) autonomous-experimentation algorithms are multi-institutional and facility endeavors with collaborators and user spread across the globe. The close collaborators consist of circa 25 groups in 5 countries. Raw data is rarely shared among these groups, it is much more common that algorithms are shared if the license permits it.

HYPERCT's autonomous hyperspectral (and multimodal) computed tomography is a project that spans across ORNL SNS and Brookhaven National Laboratory's (BNL)s National Synchrotron Light Source II (NSLS-II), utilizing both neutron and X-ray imaging beamlines to provide multimodal imaging capabilities for the GU community. The overarching goal of the scientific research associated with energy materials is to better understand electrochemical energy conversion, which is central to the performance and lifetime of energy storage devices. In a battery, irreversible processes happen across broad time and length scales, from the atomic level to the electrode level (cm), and from ms to days. Multiple imaging modalities are necessary to encompass the scales required to understand the changes in chemistry, crystal structure and electrode morphology of a battery during operation. These modalities often exist at more than one DOE BES Scientific User Facility (SUF), and the associated collection processes can vary widely by modality. These considerations necessitate careful planning for the efficient collection of informative data, data sharing across large-scale facilities, academia and/or industry, and state-of-the-art analysis and inference from the collected data.

Multimodal imaging is a technique that is increasingly data heavy due to advances in X-ray and neutron source brightness (leading to faster image acquisition) and improvements in detector technology such as larger detector areas, smaller pixel size, and the ability to measure events (leading to larger data sets). Advances in imaging algorithms make better use of all of this data but lead to additional computational expense. The computing power needed to reduce, process, analyze and visualize data, let alone combine different modalities together, often reside at national laboratories, again requiring close communication and coordination across facilities but are often distributed across various physical locations.

Data acquisition of complex system energy systems spawn across several DOE BES SUFs. In this case study, we utilize imaging capabilities at the SNS, at ORNL, and NSLS-II, at BNL. We combine this with image processing methods developed at Purdue University and ORNL and implemented on ORNL and BNL computational 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
LBNL	primary & secondary	cloud service, local database. Globus occasionally.	100 MB-100 GB per experiment or simulation. 1-10 TB/year	ad hoc, approx. daily	Y, through data mining, web	Access to proprietary software. Slow transfer of large data sets, esp. between data storage & computational resources. Lack of standardization. Scientific users not comfortable with databases. Compatibility with legacy equipment/software.
ANL/CNM (ALSO ANL/ LCRC, ANL/ ALCF)	primary	Data transfer using Globus, ANL BOX, NST server	GB range for experiments, GB-TB for simulations	ad hoc	N	Access to proprietary analytical software
BNL	primary (secondary at NSLS2)	Data transfer using Globus, S3, or cloud service. Users have access to NSLS2 data via Tiled and Jupyter Hub.	100 MB-10 Tb per experiment	ad hoc	Y, through messaging protocol	Network access routes are never simple. Lack of a single standard/convention used by all parties.
ORNL	primary	Data transfer with Globus, locally through gateway devices to brigde networks, data portals at neutron sources	MB-GB	variable	Y, through gateway devices	Networking from instruments can be challenging, data transfer workflow still in development at neutron sources
INTERNATIONAL UNIVERSITY LABORATORIES	primary	Kafka, Cloud services, e.g. AWS	100 MB-10 GB	Seconds to minutes for messages, weekly for large data transfer	Y	secure access

Figure 5.12.1: Autonomous Experiment Steering Collaboration Space

5.12.2.3 Use of Instruments and Facilities

ANL
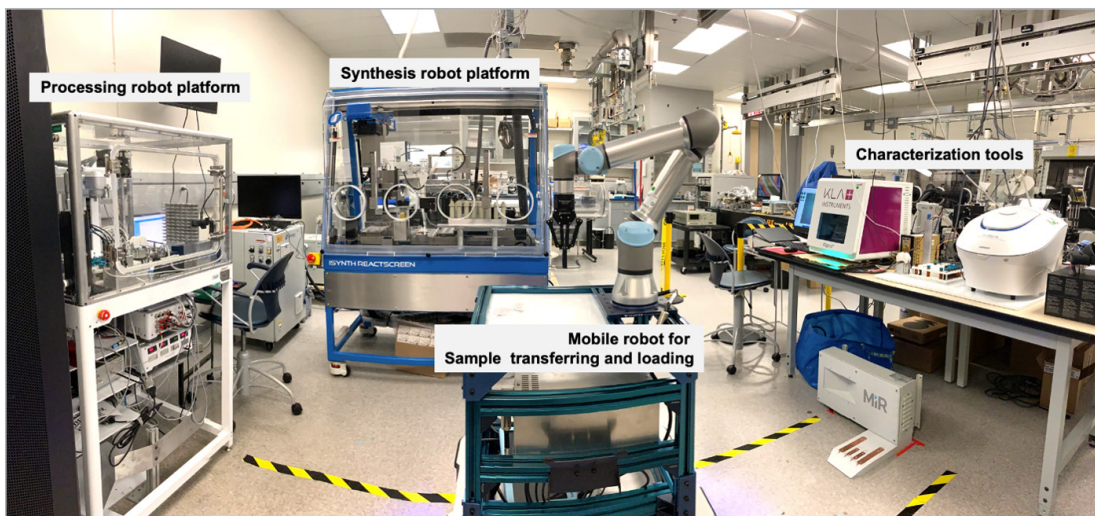
Figure 5.6.10: ORNL WAN Connectivity

- **Robotic solution-processing platform:** Pipetting system for solution transfer, liquid handling system, substrate handling system, blade coating station for film deposition, heating stage for film annealing, sample storage plate, and an enclosed frame for controlled experimental environment.

- **Chemspeed synthesis robot platform:** Liquid and powder handling system, reactor arrays, filtration system, chemical storage systems, and an enclosed frame for controlled experimental environment.

- **Mobile robot:** MIR200 wheel robot, UR5e robot arm with finger and vacuum grippers.

- **Automated Characterization tools:**
  - **Imaging system:** optical image characterization.
  - **UV-vis/fluorescence spectroscopy :** (polarized) absorption spectrum, fluorescent spectrum.

- **Electrical characterization system**: automated probe station and Keithley 4200 electrical characterization system.

- **Mechanical characterization**: automated sample loading stage and iNano indentation tester.

- **Electrochemical characterization system**: automated sampling and Gamry potentiostat.

- **HPC**: The Carbon computing cluster provides job scheduler, software for materials modeling across scales, and proprietary analytical software

ORNL

The CNMS currently has capabilities in automated synthesis and characterization:

- **AutoFlow(s) system:** Autonomous Continuous Flow reactor synthesis system with scalability, that consists of automated pumps, reactors and valves connected to characterization tools such as high-pressure liquid chromatography to enable optimization of reactions to produce polymers for sustainability applications. This system is fully configurable and addressable, with each

individual component networked to a main server in the lab through Wi-Fi on a local network, and an API for running autonomous synthesis routines with custom optimization.

- **Microscopy automation:** We currently have developed in-house tools for low-level control of scanning probe microscopy instruments (scanning tunneling microscopy and atomic force microscopy), utilizing FPGA cards and Python to develop abstractions and enable efficient programming for autonomous imaging and spectral capture. We have also developed the same types of platforms for control over the scanning beam in electron microscopy, enabling sparse sampling, atomic matter fabrication, feedback-governed crystallization, and for studying beam-matter interactions more generally.

HyperCT at the SNAP and NSLS-II Submicron Resolution X-ray Spectroscopy (SRX) Beamlines

For the purpose of this document, the focus remains imaging data sets, as they tend to generate very large data sets. The SNS SNAP high-pressure diffractometer is used to perform hyperspectral imaging experiments while the dedicated VENUS imaging beamline is being built. At SNAP, spatial resolution of $\sim$ 100 mm is achieved with wavelength resolution varying from 120 ns to 5 ms. The SNS beamline network is composed of a SNS imaging GPU computer (has access to archived data), SNS computer controls of beamline, edge node (has access to archived data). The SNS open research network is accessible by facility users from their home institution and is the central SNS data storage server, SNS translation node (for auto-reduction) and computational resources. One hyperspectral neutron computed tomogram, which is a fraction of an experiment, can contain $\sim$ 100,000 to 300,000 files. A tomogram data set is composed of tens of radiographic projections. Each hyperspectral radiographic projection is saved in a unique directory (associated to a unique run number which is associated with the measurement metadata). One hyperspectral neutron computed tomogram is approximately $\sim$ 100 GB (with the current detector technology).

The NSLS-II SRX is a micro/nano probe utilizing a high brilliance in-vacuum undulator at the NSLS-II. With two sets of focusing optics, X-rays are focused down to 200 nm to 2 μm spot for high-resolution imaging and spectroscopy experiments. The NSLS-II beamline network: is made of at acquisition and controls systems. The open research network is named the NSLS-II Science Network and Data Storage but also has computational resources, and connections to BNL HTSN. One hyperspectral X-ray fluorescence tomogram contain thousands of files. Depending on the imaging size, tens to hundreds of hyperspectral fluorescence images, or radiographs, are measured. For each radiograph, raw data is stored for each detector, line-by-line, as a separate file on the network storage. A scan database records the file names and locations, as well as the associated scan metadata. Secondary processing is done by collecting these raw files together to visualize the elemental composition for each radiograph into a single file. To calculate the 3D volume, these hyperspectral radiographs are reconstructed for each element of interest to visualize the volume. Files are typically stored in Hierarchical Data Format 5 (HDF5) format with compression enabled to save disk space. For a given hyperspectral tomography, we can expect to create 1,000 to 100,000 files and occupy 100 MB to 1 GB of disk space.

LBNL

The ALS is in the process of developing autonomous data collection in close collaboration with CAMERA. The first project is in collaboration with the Berkeley Synchrotron Infrared Structural Biology and focused on accelerating infrared spectroscopy. A second project applies the autonomous CAMERA algorithms to the optical beamline alignment of XPCS COSMIC.

Infrared Spectral Microscopy at the ALS

The Berkeley Synchrotron Infrared Structural Biology group at ALS, together with CAMERA has developed production-ready tools for users to perform their IR-Spectroscopy experiments autonomously. The intelligent decision-making has two main components:

- A Matrix factorization (most often PCA) for automated data analysis. Base-spectra are recomputed repeatedly on the current data set.

- UQ in combination with function optimization used for prediction and steering. The projections onto the first few base-spectra are approximated across the domain. That approximation and its UQ dictated the next measurement location.

XPCS Beamline Optics

CAMERA in collaboration with the beamline scientist are using Gaussian-process regression to efficiently align the optical elements of the XPCS beamline.

BNL

Autonomous X-ray Mapping at NSLS-II

In landmark experiments at BNL, the decision-making algorithm Gaussian-process (gpCAM) by CAMERA was deployed to control X-ray scattering imaging experiments. These studies were performed through a collaboration between BNL's National Synchrotron Light Source II (NSLS-II), the CFN and LBNL's CAMERA. Autonomous imaging was applied to a variety of heterogeneous materials problems, including polymer crystallization, nanoparticle film formation, electrospray deposition of polymer films [43], and additive manufacturing (3D printing). The same methodology can be applied to explore physical parameter spaces if one manufactures a sample library. Specialized synthesis methods enable the fabrication of one-dimensional or two-dimensional gradients of material properties, which thus enables the creation of combinatorial sample libraries. For instance, one can use thermal gradients to generate continuous libraries of annealing temperature, flow-coating for thickness or composition gradients, chemical gradients for libraries of surface chemistry, or electrospray deposition to generate arbitrary patterns of film composition and thickness. Gradient methods can frequently be combined to generate two-dimensional libraries where material parameters are varied continuously; the final sample thus represents an exhaustive slice through the high-dimensional space describing material synthesis/processing conditions. The outstanding challenge — to efficiently explore that slice — can be addressed using autonomous methods.

Robotic Synthesis at TMF

The Inorganic Nanostructures Facility offers facilities for automated synthesis and characterization of novel inorganic materials, including colloidal nanoparticles, two-dimensional materials, and hybrid inorganic-organic frameworks. Data generated in this facility includes: (1) synthetic recipes; (2) raw characterization data such as spectra, images generated by optical/Raman/and electron microscopes, X-ray diffraction patterns, and electrical measurement data, predominantly stored on local instrument computers; (3) raw simulation data; and (4) processed data stored on local computers and in more centralized electronic repositories, including: (1) electronic laboratory notebooks, e.g. Evernote and OneNote; (2) personal and shared cloud folders, e.g. Google Drive; (3) ScopeFoundry, as noted in the Imaging Facility plan, and (4) local and remote servers for Python Jupyter notebooks, and (5) centralized databases. For example, the Nimbus robots utilize the ESCALATE pipeline (MRS Communications 9, 846–859(2019), https://doi.org/10.1557/mrc.2019.72) for generating and capturing high-throughput perovskite synthesis data. Finally, TMF operates a commercial LIMS called Laboratory Execution and Analysis (Symyx/Unchained Labs) for generating, capturing, and storing data associated with two of its robotic synthesis platforms, HERMAN and WANDA. LEA has a modular architecture that stores data in a centralized LBNL Oracle database mediated by a TMF-based midlevel application server that mediates the transfer of data and metadata from client (robot) computers. Experimental designs are automatically stored in

the central database, along with process logs and, in some cases, characterization data. A strength of this system is its Library Studio software that allows for complex chemical recipes to be designed and encoded (in XML) by users unfamiliar with automation. Since this closed system does not have API's for open-source programming environments, it is difficult to integrate with ML algorithms and autonomous systems.

NSLS2

Today, the central Lustre file system is used for all experiments at the facility. The synchrotron light source has 28 beamlines in operation and 4 under construction. Staff and users have a home directory available everywhere—including workstations, data analysis servers, and Python Jupyter — so that configuration, conveniences, and access keys are naturally available wherever they work. The "original" copy of raw data is stored in a protected, internal system where it cannot be accidentally modified or deleted. This data is stored in a document model, with different experiments taking different approaches to document organization (i.e. number of files/documents per experiment). The documents will also reference external data that is too large to reasonably fit in the document, enabling rapid flow of information. The meaning of data, the size of data, the processing steps applied to data, and the analysis and interpretation of data vary depending on the beamline, the measurement technique, the detector(s) utilized, and the scientific goal. Raw data is generated primarily by two-dimensional (area), one-dimensional (strip), or point detectors as a part of scattering, imaging, or spectroscopy measurements. The data from all these detectors either takes the form of an image or histogram, or as a stream of time stamped photon detection events. Raw data may represent a scattering pattern, a transmission image, or spectra, for example. The size of the data may vary from a few megabytes to hundreds of terabytes per allocated beam time. Some form of data processing or reduction is usually performed after data is collected to transform the data from a technique or detector representation to an analyzable representation, such as a series of sinograms into a three-dimensional world-space volume, or spectra into elemental concentrations.

All the facility's raw data is available via HTTP(S) and Python Application Programming Interfaces (APIs) to support data portals and data science workflows. On a beamline by beamline basis, reduced data may be available through these same APIs. On a beamline by beamline basis, raw and/or reduced data may be made available in a traditional directory, with formats and naming conventions suitable for the beamline and user group. Data access, whether via APIs or filesystems, is granted on the basis of a proposal and a beamline. Beamline staff can access all the data on their beamline. Users can access data related to a proposal that they are listed on. Users can be added and removed from a proposal in the PASS, and their data access is updated accordingly. The facility has committed to storing data and to making it available to users for at least one year after the experiments and best effort beyond that. Our current plan is that 3.2 PBs will be commissioned in FY22.

To achieve the facility vision in the next several years, a complete suite of tools and capabilities in data infrastructure and computing need to be created and developed (see Figure 5.12.2). In addition to experimental control and data acquisition, these tools and capabilities include algorithms and AI/ML, scalable software libraries, workflow and orchestration tools, seamless real-time on-demand computing, network improvements, and discoverable data repositories.
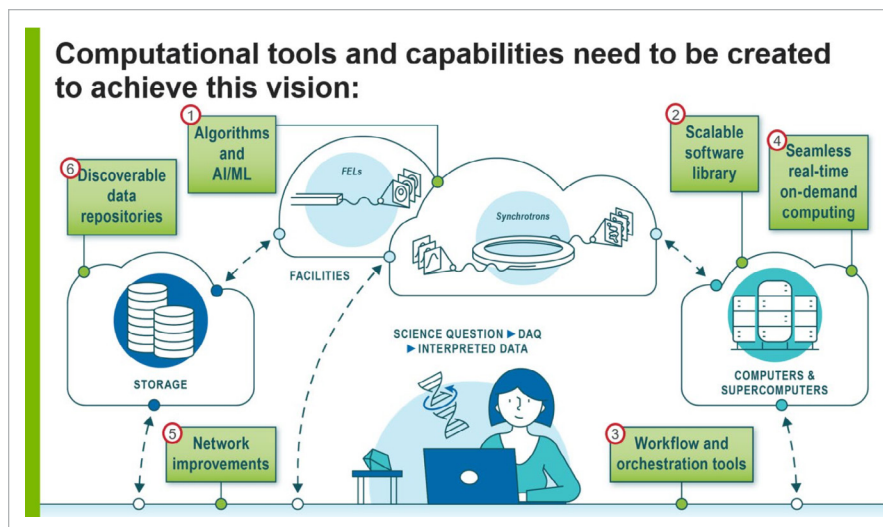
Figure 5.12.2: Autonomous Steering Vision

In the longer term, NSLS-II has identified 22 new beamlines to be developed to balance the beamlines portfolio and provide additional capabilities to meet the research needs of the broad scientific community. These include 12 enterprise beamlines based on mature techniques, 7 high-performance beamlines with additional cutting-edge capabilities, and 3 mission-specific beamlines to meet the needs of the targeted research communities. Together these new beamlines will support multimodal research and enhance the overall impact and productivity of the facility.

5.12.2.4 Process of Science

ANL
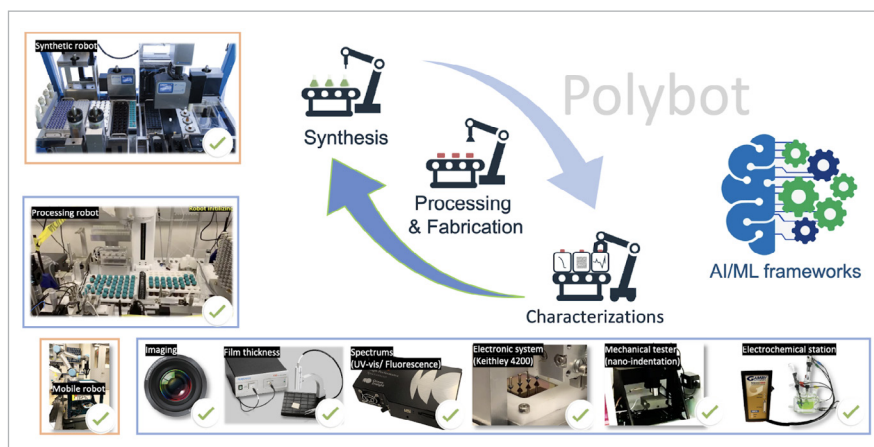
Figure 5.12.3: Polybot data pipeline at ANL/CNM

The Polybot platform consists of synthesis, processing, and characterization modules. The platform is designed to run closed-loop experiments and efficiently explore parameter space related to the synthesizability, processing conditions, and materials properties. Raw data is first stored locally on the computers that execute the experimental workflows. Filtering of faulty data is handled on the same computers. Preprocessed data is sent to computing clusters (e.g. the Carbon cluster), cloud storages, or other facilities via the network (communicated via RESTful APIs). Smaller data is typically sent as individual JavaScript Object Notation files whereas larger data is packaged and sent using data platform such as Globus.
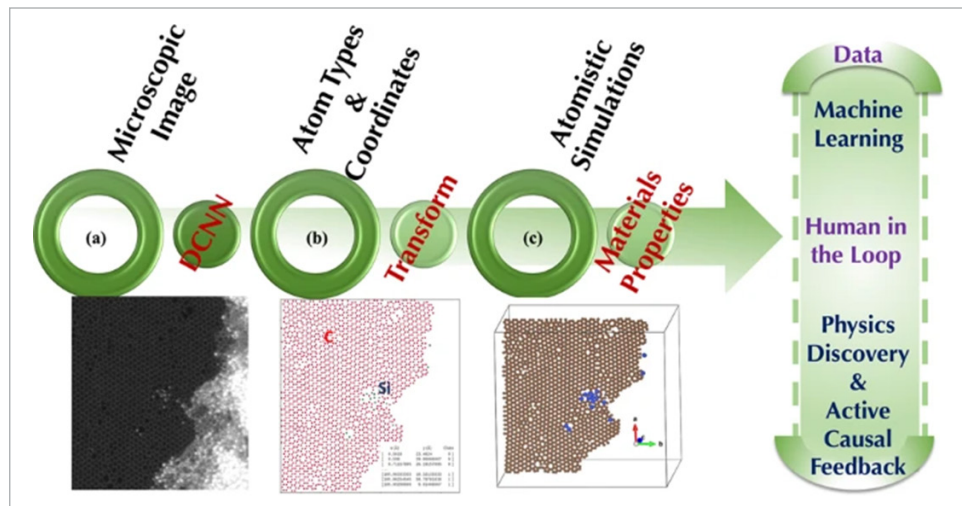
Figure 5.12.4: Example automated workflow from electron microscopy at the CNMS/ORNL

In this example, images are taken at the instrument, and then converted into atomic types and coordinates via an ensemble of models trained on simulated data. These are then fed into atomistic simulations for better understanding material properties and dynamics, and facilitating physics discovery. Here, data comes from both instrument and compute clusters, and training occurs on a GPU server. Inference occurs at the edge, on the instrument. As such, all levels of compute and data need to be accessible in real time.

Experiment and Instrument-Agnostic Autonomous-Experimentation Algorithms

It is CAMERA's goal to look at a host of scientific processes and abstract away details, to allow for agnostic tools to be developed. To make the tools fit into a variety of experimental frameworks, APIs have to be designed very flexibly to achieve generality but maintain user friendliness.

SNS at ORNL

Imaging capabilities can offer scale-bridging capabilities in energy materials that will advance the boundaries of knowledge in energy storage. SNS and NSLS-II will be used to study changes in chemistry, crystal structure and electrode morphology of a battery during operation. Performing characterization across multiple length and time scales is necessary to study the complex energy systems.

Networking between experts in energy storage materials, in X-ray, neutron and data processing/analysis experts is critical to understanding and mitigating degradation processes in electrochemical cells. However, networking is difficult due to the lack of central platform where data can be shared among collaborators on a same project, hindering scientific progress. An easily accessible platform that allows sharing and visualization, at a minimum, of processed data is needed. This is especially critical for imaging data as it tends to be very large.

ORNL are currently working on the development of intelligent acquisition and reconstruction of hyperspectral computed tomography at both SNS and NSLS-II called HyperCT. The principal goal of this research is to combine physics-based modeling with AI and ML algorithms to enable dramatic improvements in the throughput and performance of hyperspectral (i.e. multiple energies) computed tomography (HSCT) beamlines at DOE BES SUFs.

HyperCT's workflow involves a data measurement process that is initiated, the measured data is streamed and archived, and advanced reconstruction algorithms are implemented on a multi-GPU node that has access to the archive, giving a glimpse of the images from partial measurements. Such a streaming system is critical for

neutron tomography because it can take of the order of days for a complete measurement. Using this workflow, we recently completed an autonomous hyperspectral neutron computed tomography experiment at the SNS SNAP beamline and are using this data to develop and refine robust and efficient ML-informed reconstruction algorithms for sparse and low signal-to-noise data, which is a typical acquisition strategy at neutron sources (due to lower brightness than X-ray sources).

BNL

Autonomous Experimentation at NSLS2 Using Distributed Instruments

The scientific use cases present at NSLS2 are too diverse to specify here, albeit an effort has been made to specifics regarding software and networking advancements.

Present: AI/ML solutions are deployed using custom Pytorch, Tensorflow, JAX, or scikit-learn implementations for on the fly data analysis. These approaches include phase or property prediction, unsupervised data set decomposition and visualization, and anomaly detection for experimental aberrations. Adaptive experiments are performed at many beamlines across the facility using software tools in the Python ecosystem for reinforcement learning or Bayesian optimization. For example, spectroscopy experiments for searching samples of energy materials for maximal information gain or property optimization are performed combining measurement automation using Bluesky, Bayesian optimization using Gaussian processes at NIST (Gaithersburg, MD), and visualization remotely across the eastern seaboard. Data for such experiments is served using Tiled and experimental orchestration using a Kafka message bus.

Next: There is increasing demand for multimodal analysis that leverages multiple beamlines at NSLS2 and/ or measurements at other central facilities. These will require edge compute for analysis at the individual measurements, secure data transfer to distributed experimental orchestration, and remote access for users not at any facility. Simulation data to support experiments—by query of databases or on-demand computation—will be required at the time of measurements. Furthermore there are anticipated efforts in building more effective ways to engage the "human in the loop" through human–AI–facility interfaces. These will require case by case interfaces (graphical or otherwise), and secure access if used remotely.

Beyond: Scientific workflows are expected to be fully distributed across user facilities and laboratories. The measurement capabilities of beamlines and facilities are unique and scientific discovery often requires multiple types of analysis. It should be tractable for a user to synthesize a sample using standard techniques (e.g. Polybot) and trigger a sister synthesis engine to prepare an identical sample for measurement at multiple facilities, as well as robotic transport for the sample within a facility. This will fully leverage the diverse excellence at the many facilities in ESnet and maximize scientific output.

5.12.2.5 Remote Science Activities

Dashboard type web applications will be used for monitoring status of the instruments and for browsing.

BNL

NSLS2 at BNL

Present: Remote data processing using BNL networked compute, serving data analysis to remote users. This occurs using Guacamole (less frequently No Machine) for local monitoring or a JupyterHub for distributed monitoring and analysis.

Next: Remote data processing using deep learning with HPC or GPU clusters will need to send raw or preprocessed data to servers and which will reply with analysis results. These results can be communicated to on-site or remote users using similar technology as present. Autonomous platforms will need to communicate via messages both domestically and internationally, and potentially send raw data. Mail-in experiments programs will grow, where facility users are mailing samples to the facility for automatic or routine analysis.

Beyond: Enterprise beamlines with little to no direct interaction will be developed. These will expand the capacity for mail-in programs, leveraging autonomous analysis, reporting, and distributed compute.

ORNL

SNS

Today, the ORNL neutron facilities currently provide remote access to the beamlines as either viewer or as editor of an experiment. Users can set experiment parameters remotely, can start and visualize the progress of a measurement.

The next step involves the implementation of ML algorithms to users so they can make "intelligent" decisions based on data previously collected.

Beyond this goal, enabling live video feed and real-time data streaming and processing with high-rate detectors is envisioned.

5.12.2.6 Software Requirements

ANL

ANL/CNM primarily utilizes two storage technologies, block and object storage. The software solution for block storage access is Luster, and DataCoreSwarm for object storage. Both storage solutions utilize the Globus data platform services to share and transfer data with remote users or facilities. In addition, DataCore Swarm provides an S3 API to access, transfer, and share data. In the next two to five years, CNM plans to release custom RDMS to aid in the data curation, analysis, and publication of scientific data produce at CNM.

ORNL

At CNMS, we utilize local storage on a 1 PB system with file transfers through ftp/scp/Globus, as well as a smaller windows-based file server for backups of experimental data. We are trialing the use of DataFed as a federated data storage management solution. This is developed in-house and enables sharing of data with collaborators, as well as key features such as metadata queries and data provenance tracking.

At SNS, locally and remotely data resources management is already implemented. Data storage servers can be accessed data via SFTP. A suite of Python-based scripts based on Jupyter notebooks is also available on the data servers to help with data processing and analysis. Some of the experimental planning tools, such as optimization of sample composition and thickness, are also available as a web interface.

Within the envelope of our HyperCT project, we are presently developing tools to provide data processing and analysis for our imaging users at both SNS and NSLS-II. We utilize AVIZO/AMIRA (commercial software packages) for data 3-dimensional data visualization and some of the analysis for 3D data sets.

Over the next two to five years, most of the imaging software tools will need to be parallelized and it is anticipated having an edge node capability dedicated to imaging at both SNS and NSLS-II will be required since the data rate and size will only increase with time due to improvement in both facility beam flux and detector technology.

The ultimate goal is to provide users with publishable data, i.e., semi-automate data processing and analysis using ML methods that can be accessed from their institutions. Ideally, both codes and data would reside in the same location.

Center for Advanced Energy Research Applications

gpCAM (gpcam.lbl.gov) is used at ALS, NSLS2, SLAC, APS, TMF, and the Joint Genome Institute for intelligent decision-making.

BNL

NSLS2 at BNL

Central data storage of more than 3 PB annually is stored via Bluesky Databroker and accessed via the Tiled data access service. All the facility's raw data is available via HTTP(S) and Python APIs to support data portals and data science workflows. Data is stored immutably in a document model from the measurements with rich metadata. Tools for storing versioned analysis data, or processing pipelines are in development. On-site compute is available at the SDCC using a JupyterHub. This enables a scratch space for data analysis with access to central storage. The Bluesky experimental orchestration software enables AI/ML models to be linked into experiments. This can be done in-process using call-backs or out-of-process using message buses and a queue server. A specific tool of merit is Bluesky Adaptive. Integrally, this approach is agnostic to how the model is developed, enabling different users and groups to develop their tools specifically to their experimental needs with the software of their choice. Tiled takes this a step further by enabling on the fly data access with languages other than Python.

The Data Science and Systems Integration group at NSLS2 will continue to develop open-source software tools in the Bluesky project for data acquisition, management, and analysis. Enhanced data science capabilities will require a robust and rich sample database for historical data that maintains references to raw data, versioned analysis, and access rights for groups. A continued collaboration with SDCC will enable data science advancements using the internal network at BNL.

The software tools and programming languages that users will be using beyond the predictable horizon are not guaranteed. There is no expectation that data will commonly be processed using Python, Tensorflow, or Pytorch. Infrastructure that is developed for long-term support—especially at a user facility—should be capable of interfacing with presently nonexistent software.

5.12.2.7 Additional Network and Data Architecture

ORNL

At SNS and for the HyperCT project, sharing data between facilities is key in solving "big" science questions, and most importantly merge 3 or 4D imaging data between instruments located on the same site (i.e., imaging at the HFIR and at the SNS), or between facilities located on different sites (SNS and NSLS-II).

5.12.2.8 Use of Cloud Services

Cloud services (e.g., Box, Google Drive) will be mostly used for file sharing. For example, external collaborators who wish to access a subset of the data but do not have accounts at the institution where that data is stored. These services make it easy to set up sharing and automatically notifying the collaborators of any file updates.

However, cloud services tend to be slow at retrieving data and are rather expensive. Until prices are more competitive and data can be retrieved fast, there is no clear path forward to use these services for large data sets.

5.12.2.9 Data-Related Resource Constraints

None to report at this time.

5.12.2.10 Outstanding Issues

None to report at this time.

5.12.2.11 Case Study Contributors

Autonomous Experiment Steering for BES Facilities Representation

- **Alexander Hexemer**, *LBNL*, ahexemer@lbl.gov
- **Subramanian Sankaranarayanan**, *ANL*, skrssank@uic.edu
- **Phil Maffettone**, *BNL*, pmaffetto@bnl.gov
- **Hassina Bilheux**, *ORNL*, bilheuxhn@ornl.gov
- **Marcus Noack**, *LBNL*, MarcusNoack@lbl.gov
- **Kevin Yager**, *BNL*, kyager@bnl.gov
- **Yugang Zhang**, *BNL*, yuzhang@bnl.gov
- **Dale Huber**, *Sandia*, dlhuber@sandia.gov
- **Jennifer Hollingsworth**, *LANL*, jenn@lanl.gov
- **Jie Xu**, *ANL*, xuj@anl.gov
- **Henry Chan**, *ANL*, hchan@anl.gov
- **Kyle Kelley**, *ORNL*, kelleykp@ornl.gov
- **Rama Vasudevan**, *ORNL*, vasudevanrk@ornl.gov
- **Emory Chan**, *LBNL*, emchan@lbl.gov
- **Bogdan Vacaliuc**, *ORNL*, vacaliucb@ornl.gov
- **Singanallur Venkatakrishnan**, *ORNL*, venkatakrisv@ornl.gov
- **Stuart Campbell**, *BNL*, scampbell@bnl.gov
- **George Nelson**, *University of Alabama in Huntsville*, george.nelson@uah.edu
- **Charles Bouman**, *Purdue University*, bouman@purdue.edu
- **Gregery Buzzard**, *Purdue University*, buzzard@purdue.edu

5.13 BES Design and Development of Digital Twin Strategies

The digital twins case study brings together BES community representatives from X-ray light sources, neutron scattering facilities, NSRCs, and independent researchers, to discuss this emerging area of scientific investigation. The case study profiles the overall concept, some early results, and a discussion about what the future will hold.

5.13.1 Discussion Summary

- The rapid advancement in AI/ML algorithms, improved shared workflows, and the advent of exascale computational resources, make it possible to create a physically informed virtual platform to perform experimentation: digital twins.

- Digital twins represent physically accurate computational environments of experiments will help to guide in-silico experiments from conception to synthesis and measurements. Digital twins aim to enable offline design and optimization of all elements of the scientific process, from hypothesis to discovery, in situ or mainly prior to actual physical experiments.

- Digital twins:
  - Facilitate the creation of a virtual environment to exhaustively explore experimental controls
  - Import experimental read-outs at any time to provide instant synthetic read-outs
  - Create data that can iteratively be used as training for model improvements
  - Allow for a small subset of simulated experiments to be explored in actual experiments at the SUFs.

- The further development and application of digital twin simulations supporting experimental design and operation requires that a number of gaps in capabilities and infrastructure be filled, including sufficient and reliable networks, data storage resources, computing systems, transparent access to facilities via federated identity and shared data protocols, and software tools to operate across the DOE's distributed resource landscape.

- Data movement and data processing are critical for scientists and researchers, particularly once they have performed their experiments or simulations. Steps include:
  - Data acquisition via computational simulations or experiments
  - Data movement, cataloging, and archiving
  - Data analysis and visualization
  - Fitting to models and analysis of the processed data for structure, function or dynamics information

- Computational resources are ephemeral, and data is enduring, requiring long-lived allocations of storage resources. Public policy is further motivating this demand, requiring open access to data generated as part of federally funded research.

- The goal of a digital twin is to seamlessly extract information from the experimental data sets and use them as either input to dynamical simulations or as training data to improve models.

- The use of digital twins will increase the efficiency of the user experience at the SUFs by providing prior and concurrent guidance to users to run the most precise experiments.

- It is expected that as GPU usage continues to increase, either local cloud compute capabilities will grow to accommodate this demand, or our users will shift to external cloud services such as AWS or Azure. This may run into the dozens of instances/year level.

- For most of the facility, the data storage and transfer speeds are currently sufficient, and are not expected to be significant concerns in the next two years. The main constraints are with the access to computation, which is particularly limited for the theory groups, and the changing nature of the HPC platforms which make older code quickly obsolete, requiring significant rewriting to accommodate new hardware upgrades.

- The lack of a uniform data pipeline standard that span facilities causes issues, such as differences in metadata associated with different data types and different levels of data manipulation. To reach a stable and universal environment standard data formats, software development framework, sample tracking, metadata capturing, and labeling are required.

- Coupling digital twin simulations with experimental steering will become an increasingly relied upon capability at the light sources as the decade progresses. Due to the high computational cost associated with data processing, reduction, and analysis, and of model training on such large data sets, the light sources will need to employ the use of geographically distributed computing facilities and data repositories. This necessitates high bandwidth, low latency network connections between the light sources, and remote computing and data storage centers.

5.13.2 BES Design and Development of Digital Twin Strategies Case Study

Viable frameworks to perform and test a complete experiments on a computer, versus using a physical scientific facility, are an emerging area of research. Termed "digital twins", these concepts represent physically accurate computational environments of experiments will help to guide in-silico experiments from conception to synthesis and measurements. This includes shared workflows and continuous updates from real experiments that facilitate the design of new experimental capabilities and execution of optimal experimental strategies to drive physics/chemistry knowledge acquisition.

5.13.2.1 Science Background

Understanding the structural and dynamical evolution of materials is central to a plethora of energy applications. There is a large fraction of the approximately 16,000 users at DOE's SUFs who are interested in accessing world-class spatiotemporally resolved characterization modalities and performing experiments to explore structural evolution as well as energy and mass transport (electrons, phonons, atoms, and ions) across material interfaces. This translates into an intense competition for instrument time amongst the users, meaning that each user can only perform a small number of actual spatiotemporally resolved experiments. The need for a physically informed strategy for offline design and optimization of these experiments is emerging as a grand challenge for SUFs. The rapid advancement in AI/ML algorithms, improved shared workflows and the advent of exascale computational resources now make it possible to create a physically informed virtual platform to augment and guide spatiotemporal experimental trials. Moreover, due to a confluence of compute capacity, data veracity/volume, and new algorithms, materials and chemical simulations are bridging the gap in matching experiments by providing viable frameworks to perform/test complete experiments on a computer, e.g., digital twins.

Digital twins aim to enable offline design and optimization of all elements of the scientific process, from hypothesis to discovery, in situ or mainly prior to actual physical experiments, thus greatly accelerating achieving scientific goals. This includes research and results on physically accurate surrogate models for traditional first principles modeling and simulation (e.g., fast approximants for first principles simulations, for materials interaction potentials, for efficient sampling, etc.), methods for efficient and accurate data inference and reduction (e.g., data fusion, compression, Bayesian inference, causal approaches, etc.), and approaches to match/register simulation data with experiments. For example, one envisioned utility of digital twins is to reduce the time to synthesis of materials/chemicals with desired properties. Assuming that a target structure is predicted from a first principles model, a series of syntheses could be carried out within a virtual environment, with a small subset of real experiments serving to ground and tune the synthesis models. Subsequently, optimal characterization schemes could also be found within a digital twin to isolate whether the structure of interest was generated; ideally, the analysis routines could be self-generated within the digital twin to further accelerate time-to-solution.

The latter is a particularly important aspect of facility and experimental design that enables both to be optimized in the same step.

Such digital twins are expected to have a profound impact on both the user experience and productivity as users will be better informed about the impact of their experimental controls on the expected read-outs. This will allow them to take maximum advantage of the allocated time for experimentation, thus accelerating scientific discovery.

5.13.2.2 Collaborators
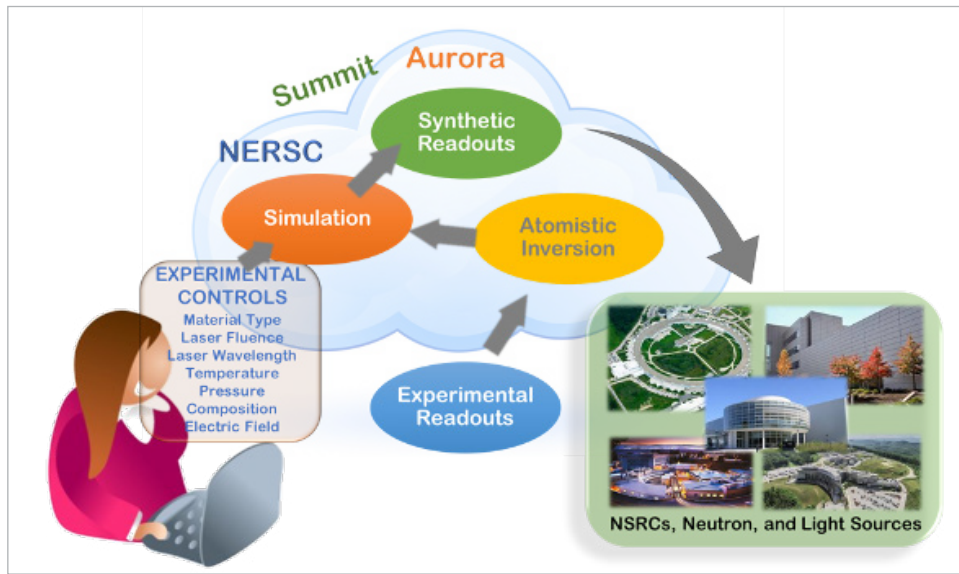
Figure 5.13.1: A typical example of a digital twin illustrating the various stages and facilities where experimental and synthetic data sets are expected to be created, shared, computed and stored

The digital twin shown in Figure 5.13.1 will facilitate the creation of a virtual environment by leveraging physical models with AI/ML algorithms and shared workflows to allow users to exhaustively explore experimental controls on high-performance computers (HPCs), import experimental read-outs at any time instant, perform inversion on them to provide starting configurations for simulations on supercomputing resources and obtain synthetic read-outs. A small subset that displays the most interesting phenomena can then be explored in actual experiments at the SUFs. These experiments can iteratively be used as training data for model improvement in the digital twin environment to close the loop. The success of such digital twin platforms hinge on seamless information extraction from experimental data sets that serve as inputs to physically accurate yet efficient dynamical models running on HPC and exascale machines. The digital twin allows users to exhaustively explore experimental controls and obtain read-outs — a small subset of the data that displays the most interesting physics and/or phenomena can be explored in actual experiments. The data from these experiments can iteratively be used as training data for model improvement to close the loop.

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
CNM	primary	data transfer via Globus, ANL Box, NST Server	GB-TB	daily	N	access to proprietary analytical software
CNMS	primary	data transfer via Globus	GB-TB	daily	N	Growing toward instrument control, will need access to proprietary software
CFN	primary	data transfer Globus, cloud storage, ad hoc methods	GB-TB	on demand	N	Data organization (format, metadata) needs to be defined/ agreed on a per-case basis
CINT	primary	cloud services, portable hard drive,	100GB-100TB	ad hoc	N	Data rates/volumes, network access
APS	primary	large disk system at the APS using the APS Data Management System, data transfer Globus, ANL Box	GB-TB	daily	N	access to proprietary analytical software
ALS	primary & secondary	cloud service, local database. Globus occasionally.	100 MB-100 GB per experiment or simulation. 1-10 TB/year	ad hoc, approx. daily	Y, through data mining, web	Access to proprietary software. Slow transfer of large data sets, esp. between data storage & computational resources. Lack of standardization. Scientific users not comfortable with databases. Compatibility with legacy equipment/software.
SNS/HFIR	primary	data transfer via Globus	GB-TB	daily	N	Ok for small size data transfers
LCLS (SLAC)	both primary and secondary	data portal, data transfer, and remote analysis on clusters at SLAC	~100 TB	Continuous	N	large data files and complex structured data make data sharing challenging

Table 5.13.1: Digital Twins Collaboration Space

The NSRCs benefit from an intrinsically strong interaction with their associated National Laboratory signature strengths in multiple areas and takes advantage of the distinctive capabilities of other DOE user facilities at, including the OLCF, the ALCF, NERSC, the SNS and the HFIR, the APS, the ALS, the National Synchrotron Light Source II (NSLS-II), LCLS-II, and SSRL. In particular, the CNMS emphasizes a strong link to neutron sciences, providing an environment for researchers to integrate neutron studies into nanoscience efforts. The CNMS uses its expertise in materials sciences (including polymer synthesis) and computational sciences towards the incorporation and development of materials-by-design approaches, and seeks new advances in imaging sciences that build on ORNL's demonstrated leadership in scanning probes, scanning transmission electron microscopy, He-ion microscopy, and atom-probe tomography.

Each year the NSRC supports approximately several hundred (~400-700 users annually per NSRC) and ~6,000 annually for APS from more than 100 different institutions spanning academia to industry, and from around the world. The user community is diverse, ranging from students who work closely with NSRC staff, learning unique skills from experts and gaining access to cutting-edge instrumentation as they advance their research,

to "PUs" who collaborate with staff to develop new capabilities and instruments that are then made available to the broad NSRC user community. About 10% of all users are theory users, who work with the staff of the CNMS Nanomaterials Theory Institute to gain access to expertise and computational resources. About 40% perform synthesis and/or nanofabrication and typically use a broad range of characterization tools to verify the quality of the synthesized materials or to investigate their novel properties. The remaining 50% of the users come to the NSRCs specifically for characterization, using a broad range of experimental tools.
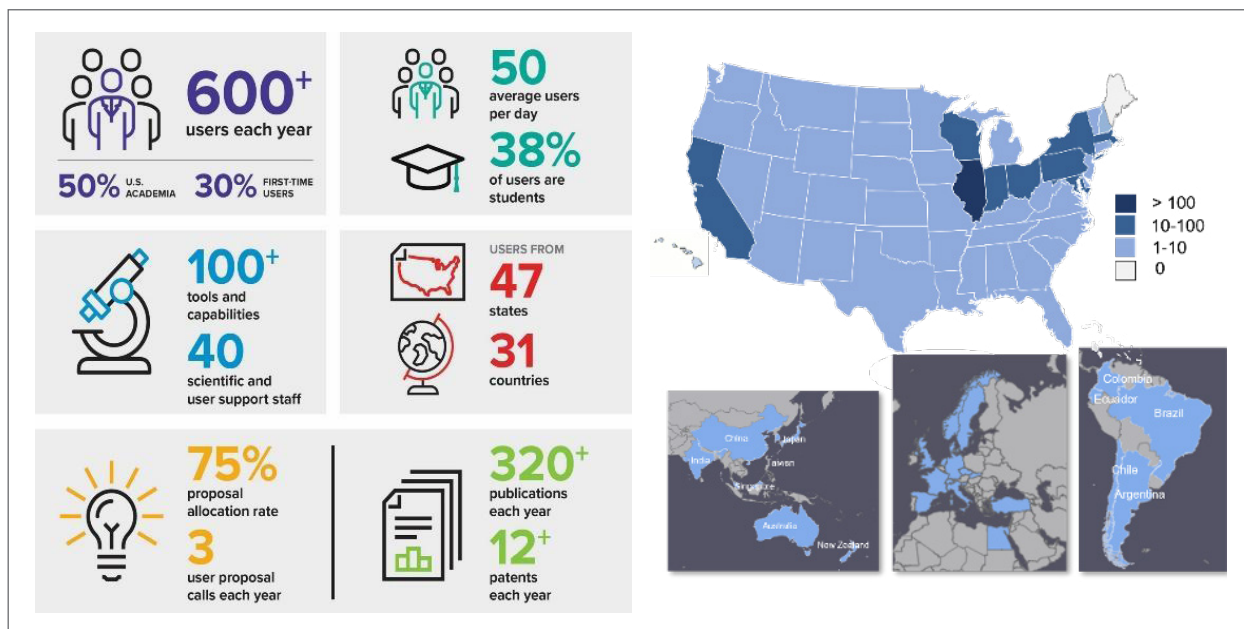
Figure 5.13.2: Geographical distribution and statistics of CNM user program. Other NSRC's have similar user statistics and geographical distribution.
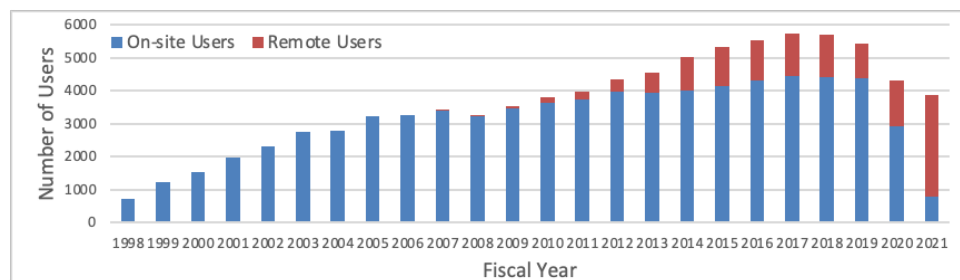
Figure 5.13.3: Total number of APS users

5.13.2.3 Use of Instruments and Facilities

The NSRC's offer a wide range of capabilities designed to enhance the study and creation of nanomaterials. From X-rays, neutron scattering, microscopy, spectroscopy, to clean room–based nanofabrication techniques, the NSRC's provide its staff and users with a broad combination of scientific resources involving data entries that vary in size from kilobytes to terabytes at collection speeds spanning from weeks to microseconds to nanoseconds. Data generated from these activities covers a broad range from surface optical and electron microscopy, optical microspectroscopy and laboratory source X-ray diffraction from the nanoscale to wafer-scale characterization to name a few. Our electron microscopy data flow includes structure and spectroscopy of materials using imaging, diffraction and energy loss spectroscopy (typically ~20 MB per entry, ~500k entries per yr). Recently installed high-speed electron microscopy leveraging a K2 detector (~103 frames per second) generates significantly higher

data volume and velocity, at 5 MB/entry and up to 720k entries/hour. Our X-ray microscopy capabilities include nano-focused coherent Bragg diffraction imaging and X-ray fluorescence spectroscopy (~0.1TB per entry, ~36 entries per yr) — these are expected to have 100x data rate increase due to the APS diffraction-limited source upgrade with first light in 2024 making full use of our EIGER2-X1M (1028x1062x32bit 2000fps) collection bandwidth — the planned approach for data collection depends on high-throughput ptychography recently demonstrated with locally deployed edge computing delivering real-time imaging that is continuously improved with HPC live-training algorithmic recognition.

The NSRC's also have significant focus on nanoscale synthesis. Our toolbox of synthesis techniques include chemical and nanoparticle synthesis methods, various gas- and liquid-phase deposition methods, general wet-lab and air-free chemistry techniques, and methods for the preparation of samples as films and surfaces. Nano-bio hybrid materials and biological assembly efforts center on nature-inspired materials that potentially can support energy conversion and energy transport, as well as understanding biosensing mechanisms in cell-like environments that are functionalized with engineered nanomaterials. We are able to hybridize metals, organics, semiconductors, and dielectrics with biomaterials to create nanobio metamaterials with unique properties. Data generated from these activities are primarily optical microscopy, electrical characterization and conventional and ultrafast optical spectroscopy (typically ~1GB/entry ~10entries/year). Our recent development of autonomous synthesis within the ANL Self Driving Lab Initiative creates a local control of an experimental process based on data recognition which requires transfer of data to HPC resources for external training of algorithmic approaches that are then locally deployed. These are covered in more detail as part of other facility reports and case studies. From a digital twins perspective, the goal is to develop complementary simulation methods and techniques that allow for synthetic extraction of characterization and synthesis information.

The NSRCs have numerous remote users in addition to on-site users who also utilize the above capabilities and the computational clusters. These users are primarily cyber-based and connect to the internal compute capabilities or leadership computing facilities using ssh combined with a SecurID and a PIN. Each user has an account on one of the compute systems and accesses the compiled codes, data archives, and necessary math libraries directly. Data from the calculations is typically analyzed and stored on the local compute systems, often for the duration of the project, and then transferred to the user's home. The size of the data/results collected varies from project to project and the compute system (100s GB to 100s TB). Online data analysis and reduction are important after which GridFTP can be effective for much of the needed data movement between computing resources.

The NSRC's also are home to a comprehensive suite of theory and modeling tools, a HPC cluster, and various software and modeling/simulation tools. The theory, modeling, and simulation methods include numerous density functional theory packages, molecular dynamics, Monte Carlo, genetic algorithm, finite difference time domain, and AI/ML. The High-Performance Computing Clusters (e.g. Carbon at CNM) provides GPUs and edge-computing capabilities. A recent upgrade of 22 nodes each with 2 NVIDIA TESLA V100 GPUs is geared towards AI workflows, including training of neural network models for image analysis and performing inference from high-speed electron microscopy data. Software and modeling/simulation tools include BLAST, FANTASTX (Fully Automated Nanoscale to Atomistic Structure from Theory and eXperiment), computational electrodynamics (e.g., Lumerical, MEEP), and other specialized analysis software and modeling expertise.

In addition to the NSRC's, there are experiments at DOE synchrotron sources which can also benefit from digital twins. For example, at the LCLS at SLAC, there are 8 different instruments that all collect large amounts of data for both online analysis, quick analysis but after data is stored locally, and long-term analysis. Data needs to be transferred to high performance computing facilities in some cases. In all cases, users logon to the machines at SLAC to perform analysis and/or transfer data. All experiments vary, but can be as much as 100Tb per experiment. Each file contains a number of device data in different forms per event (in this case, each X-ray pulse), such as million pixel images, intensity values, spectra, text files, voltages, pressures, etc. In the next few years, with the delivery of high-repetition-rate capability, the current repetition rate (120Hz) will advance to 1MHz for a 10,000x-fold increase in data, making use of high performance facilities critical.

The further development and application of using digital twins and simulation with experiments at the light sources requires that a number of gaps in capabilities and infrastructure across the light sources, Laboratories, and computing centers be filled, including sufficient, reliable bandwidth, sufficient and sustainable data storage resources, on-demand access to large-scale computing systems, transparent access to facilities and systems across the complex, including federated identity and shared data protocols, and software tools and infrastructure to facilitate the development of scientific computational and data workflows that operate across the DOE's distributed resource landscape.

The digital twin effort plans to leverage the collective expertise of dedicated theory groups at NSRCs and light sources and bring together the software codes and frameworks developed by them to realize the digital twins. In the next two to five years, we are working towards integrating physical models, AI-based inversion tools and shared workflows to develop digital twins for select experimental synthesis, microscopy, X-ray and neutron characterization studies. Our vision beyond five years is to realize fully automated workflows that will allow for seamless sharing of experimental and simulation data sets amongst the users and allow them to perform multiple types of analysis (including AI/ML) in a digital twin environment. In the 5-10 yr timeframe, we envision extending the digital twins to several other spectroscopic and X-ray modalities (e.g. XPCS) to facilitate virtual experiments that allows users to map experimental controls to the intended outputs and exhaustively perform synthetic experiments before the actual ones.

5.13.2.4 Process of Science

Scientific understanding enabled through the careful design of experiments or computational simulations for the collection and analysis of key data, is one of the primary products of user facilities. A digital twin framework takes advantage of the fact that many experimental capabilities at SUFs (NSRC's and light/neutron sources) characterize either time-dependent atomic positions and/or time-dependent structure factors (is the scattering vector, is atom i's form factor, and , its position at time t). The read-outs can be predicted from dynamical simulations if the underlying models are physically accurate on time and length scales comparable to the ones probed in experiment. Due to experimental resolution limits, those observables have to be reconstructed from coarsened and incomplete real and/or reciprocal space information, necessitating advanced techniques, such as phase reconstruction algorithms to be informed by atomistic methods. In this context, AI/ML tools and powerful workflows can provide a good solution to the inverse problem of obtaining atomic positions from experimental observables by augmenting the incomplete experimental data with atomistic simulations of the energy landscape. Furthermore, AI/ML assisted multifidelity scale bridging allows users to perform dynamical simulations on a "reconstructed" or "inverted" user-defined sample and exhaustively explore experimental controls and expected read-outs in a virtual environment.
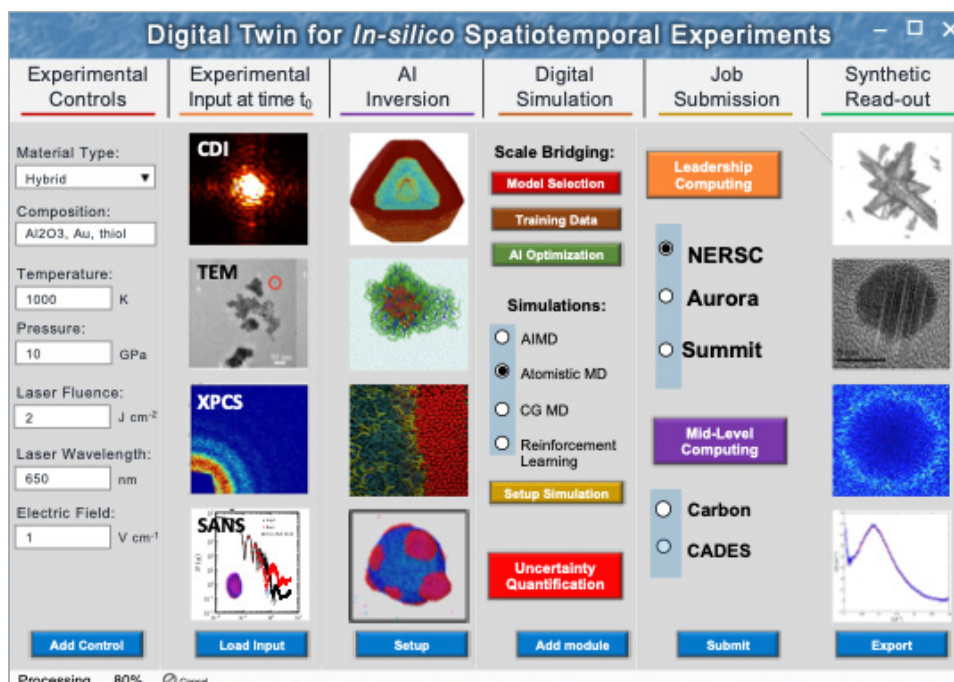
Figure 5.13.4: An example of the "digital twin" workflow being developed by NSRC staff

Data movement and data processing are critical for scientists and researchers, particularly once they have performed their experiments or simulations. Experimentally measuring and computationally simulating materials provide the major means for collecting scientific data and is the primary activity of users working at or with the NSRC's – the measurement techniques vary widely across the suite of instruments and theoretical/computational approaches, but they have some common steps involving data acquisition and analysis as listed below:

— Data acquisition via computational simulations or experiments

— Data movement, cataloging, and archiving

— Data analysis and visualization

— Fitting to models and analysis of the processed data for structure, function or dynamics information

The computational resources, on the other hand, are ephemeral, data is enduring, requiring long-lived allocations of storage resources. Public policy is further motivating this demand, requiring open access to data generated as part of federally funded research.

The goal of a digital twin is to seamlessly extract information from the experimental data sets and use them as either input to dynamical simulations or as training data to improve models. In the spatiotemporal domain, there are a variety of microscopy, spectroscopy and scattering techniques for gathering spatiotemporally resolved information at varying resolutions (Fig. 4). Recent SUF efforts have achieved ultrahigh temporal resolution down to attosecond scales, providing unprecedented precision in visualizing processes at the levels of electrons and atoms. In the spatial domain, likewise, a variety of tools allow for imaging at length scales ranging from atomic to mesoscale resolution. Real-space imaging modalities include transmission electron microscopy (TEM) and nonscanning probes that utilize the coherence of radiation to reconstruct the real-space density distribution, e.g., CDI. Combining time and space domain techniques, innovative spatiotemporal imaging tools such as UEM propel material characterization into a new era. On the other hand, physically accurate and efficient models are critical for digital twin's success. Each NSRC has a theory group with staff expertise in techniques spanning the

various different modeling and simulation scales. In principle, one can mimic the time-resolved experiments and extract both real and reciprocal space information by performing detailed materials simulations.

The use of digital twins will increase the efficiency of the user experience at the SUFs by providing prior and concurrent guidance to users to run the most precise experiments. For example, users may require extensive searching of the parameter space — of both computational models and experimental controls — requiring automated workflows in order to maximize the throughput of parameter combinations investigated (Fig. 5.13.4).

Without available workflows that couple an experiment's goals with known theory, digital twins have not been utilized in SUFs. In two to five years, applications of ASCR workflow research offers untapped opportunities to address the above challenges. Workflows can fill in the missing gaps in resolution in space and time by comparing multiple models and selecting the best model(s) to emulate an experiment. The utilization of facilities can be optimized by conducting parameter searches ahead of scheduled beam time. Workflows can help scientists derive the theory to explain empirical results by launching an ensemble of simulations with different initial conditions and searching for results that match the experiment. Beyond five years, we anticipate fully deployed digital twins at SUFs that have shared workflows operating in both forward and reverse modes. In the former, a user can switch between various multifidelity modeling scales and perform virtual simulations to extract synthetic real or reciprocal space data expected in experiments. In the latter, information extraction from experimental measurements is used to retrieve back structural information as well as facilitate continuous updates from real experiments (real or reciprocal space), which would be used as training data to improve the models for AI solutions to inverse characterization problems.

5.13.2.5 Remote Science Activities

In the case of a digital twin, most of the remote resources are computational in nature, e.g., utilizing compute clusters at one of the NSRCs or via leadership computing (NERSC, ALCF, OLCF). Other (noncomputational) remote resources being utilized typically are done via users communication to staff in real time during planning of experiment and its execution, alongside observation through Teams or Zoom. Direct steering of instruments is becoming possible.

5.13.2.6 Software Requirements

The CNM primarily utilizes two storage technologies, block and object storage. The software solution for block storage access is Luster, and DataCore Swarm for object storage. Both storage solutions utilize the Globus data platform services to share and transfer data with remote users or facilities. In addition, DataCore Swarm provides an S3 API to access, transfer, and share data. In the next two to five years, CNM plans to release custom RDMS to aid in the data curation, analysis, and publication of scientific data produced at CNM.

Online data analysis and reduction are important and much of this is accomplished within the particular application software. Other components may involve postprocessing via VisIt, VMD, VESTA, Matlab, P4Vasp, etc. Typical data formats are ASCII, XML, HDF5, binary, lossless compressed, etc. GridFTP can be effective for much of the needed data movement between computing resources, especially in the postprocessed form. Other modes such as NFS, SMB/CIFS, Globus Online, BBCP, FTP, SCP, can also be useful and are utilized.

CNMS data is being transitioned to a large PB storage server. CNMS also runs a local server with about 100 TB of storage which is a windows file system that our staff can access to back up their data. CNMS is currently trialing the use of DataFed, an ORNL-developed federated data management system for keeping track of both simulation and experimental data sets. DataFed is a software solution that links together geographically diverse file storage systems and makes all of them available through the same utility. It offers considerable features relevant for scientific information flow data provenance tracking, metadata searches and includes a Python API for automating routines. DataFed can work with Globus and other tools.

We also publish data sets through ORNL's constellation application, which provides a quick interface for hosting the file, and providing a doi. It is envisioned that this will be integrated into DataFed in the future.

For the experimental side, we are trialing the use of a gateway device to transfer files from the instruments to the data storage system, with the files automatically ingested by DataFed so that the relevant metadata is also captured, thus enabling keyword searches by metadata fields. The data storage is local to ORNL. Any remote data sets can be accessed through collaborators through either Globus endpoints or through scp/ftp, or through the DataFed interface (if the collaborator also utilizes DataFed).

For analysis of spectral and imaging data sets, the CNMS, with partners, developed the pycroscopy ecosystem of packages (github.com/pycroscopy). This software effort has amassed over 200k downloads in five years, and now consists of ~ a dozen important packages for different types of data analysis. All the packages are written in Python and are open-sourced. This ecosystem is divided primarily into three distinct areas, (1) Data input/output and utilities, (2) Generic imaging and spectroscopy analysis, and (3) More focused and domain-specific packages, as shown in Figure 5.13.1. The general feature of the ecosystem is a common 'currency' of data object (the sidpy data set object) that enables efficient processing storage and visualization.

5.13.2.7 Additional Network and Data Architecture

The Argonne campus network connects 100 buildings and 12 data centers in a multivendor switching and routing network. The campus has a new robust single-mode fiber cable plant that has plenty of capacity and ability to be expanded.

External network connectivity is provided by ESnet, MREN, Internet2, and Internet2 Peering Exchange (I2PX). ESnet recently migrated Argonne to the new ESnet6 network and has redundant optical transport and routers deployed on the Argonne campus in a redundant configuration supporting up to 400GE connections.

There are two primary networking node locations distributed between the north (Bldg 221) and south (Bldg 541B) sides of the campus to provide network redundancy. Generally, all buildings and data centers have connections to each of these networking nodes. The fiber used to connect to off-site providers is diverse and exits on the east and west sides of the campus.

The Argonne networks supports many connection speeds. The WAN supports 10GE, 40GE, 100GE, and 400GE. The campus core network support speeds of 1GE, 10GE, 40GE, and 100GE. Building connections are generally multiple 10GE with some 40GE. Data centers 10GE, 40GE, and 100GE. LAN connections within buildings are generally 1GE and 10GE.

ANL WAN

The Argonne WAN network connectivity is provided by ESnet and MREN (Internet2 gateway, Peer Exchange (I2PX)) ESnet has optical and routing gear on the Argonne campus. Argonne has two Juniper MX960 border routers that are split between buildings 221 and 541B for redundancy and diversity.

Argonne uses ESnet OSCARS virtual circuits. (On-Demand Secure Circuits and Advance Reservation System)

The Argonne connections to ESnet are 2x100GE. The connections to MREN are 3x10GE. The border routers are connected to each other with 2x100GE connections. The connections to the campus Core network is 2x40GE and 40GE to the perimeter firewall.

Argonne has one perfSONAR connected to the border at 40GE.

Argonne deploys cyber security at the border with blackhole routing and network traffic capture taps for traffic collection and intrusion detection.

- **Present to two years**
  - — 400GE upgrade expected in FY22
- **Two to five years**
  - — Upgrade border routers
  - — Multiple 400GE
- **Five + years**
  - — Terabit WAN connections

ANL Science DMZ

The Argonne ScienceDMZ is composed of Juniper QFX series equipment that is connected directly to the Argonne border routers with redundant 2x100GE links. The ScienceDMZ provides high-speed connectivity between scientific organizations for the exchange of data without taking the traditional path through a firewall.

Connections in to the Science DMZ are a minimum of 100GE with most facilities connecting at 2x100G. This allows collaborators to exchange data in a high-speed environment that does not affect commodity network connectivity.

- **Present to two years**
  - — 400GE connectivity
- **Two to five years**
  - — Hardware upgrade
- **Five + years**
  - — Terabit connectivity

ANL Core

The Argonne core network is provided by 2 Cisco Nexus 7710s spread across buildings 221 and 541B for redundancy. These switches offer network speeds at 1GE, 10GE, 40GE, and 100GE. The Core provide connections to the buildings and many of the data centers on the Argonne campus.

- **Present to two years**
  - — No changes
- **Two to five years**
  - — Replace core devices to NextGen
- **Five+ years**
  - — Terabit connectivity

ANL LAN

The Argonne LAN consists of the building access and aggregation switches. The Laboratory has standardized on Cisco and Aruba switches in the LAN. There is a combination of 1GE, 10GE, and 40GE network speeds in the LAN. Desktops are connected at 1GE.

- **Present to two years**
  - — Continual refresh of switches older than five to seven years
  - — Wireless access point upgrades

- **Two to five years**

  - Consider an overlay network topology

  - Ongoing switch refreshes

  - Zero-trust networking

- **Five+ years**

  - Next-Gen

  - 10GE to the desktop

ANL Data Centers

The primary data center at Argonne consists of Cisco Nexus equipment in an ACI fabric. It connects to the Argonne core network at 4x40GE and offers 2x40GE connectivity to every ToR switch. There is a combination of 1GE, 10GE, 25GE, and 40GE to host in this space.

- **Present to two years**

  - Hardware refresh for ToR switches

- **Two to five years**

  - Spine switch replacement

- **5+ years**

  - Terabit connectivity

5.13.2.8 Use of Cloud Services

The NSRC's do not presently utilize cloud services for its scientific data infrastructure or for computations/data analysis. Cost-effectiveness was a major factor hindering the use of cloud computing services for CNM users. CNM does not have any immediate plans to utilize cloud services for our needs. Currently the CNMS makes limited use of cloud services, but this is expected to change. Our cloud service is a locally run OpenStack run by CADES, which provides small virtual machines to staff and users at the CNMS on an as-needed basis. These VMs can, and are, used for data processing and occasionally some more heavy parallel computation jobs (e.g., those requiring ~64 CPU cores), for special requests. Many of our users also utilize Colab, for access to a GPU for training. It is expected that as GPU usage continues to increase, either our local cloud compute capabilities will grow to accommodate this demand, or our users will shift to external cloud services such as AWS or Azure. This may run into the dozens of instances/year level. CNM along with the other NSRC's will continue to evaluate the capabilities and cost-effectiveness of using cloud services in the near term.

5.13.2.9 Data-Related Resource Constraints

For most of the facility, the data storage and transfer speeds are currently sufficient, and are not expected to be significant concerns in the next two years (perhaps with the exception of very particular experiments, such as 4D STEM). The main constraints are with the access to computation, which is particularly limited for the theory groups, and the changing nature of the HPC platforms which make older code quickly obsolete, requiring significant rewriting to accommodate new hardware upgrades. These are perhaps the most significant issues. Additionally, the different NSRCs and different facilities within each, have different schemes and different levels of implementation for acquiring, labeling, storing, and providing access to the heterogeneous data generated. The lack of data pipeline standards causes issues, such as differences in metadata associated with different data types and different levels of data manipulation. To capture and curate data from the NSRCs, many levels of technical details need to be worked out, such as data formats, software development framework, sample tracking, metadata capturing, and labeling. In addition, data sets from correlated measurements, and the corresponding simulations, need to be handled in a coordinated manner to extract synergistic information. These issues hamper the efficient

use of NSRC data for scientific discovery. There is a need for common standards and shared workflows for data across the NSRCs which will not only provide an effective data solution, but will also enable cross-center data sharing, augmentation, and manipulation.

Coupling 'digital twin' simulations with experimental steering will become an increasingly relied upon capability at the light sources as the decade progresses. Due to the high computational cost associated with data processing, reduction, and analysis, and of model training on such large data sets, the light sources will need to employ the use of geographically distributed computing facilities and data repositories. This necessitates high bandwidth, low latency network connections between the light sources, and remote computing and data storage centers.

5.13.2.10 Outstanding Issues

None to report at this time.

5.13.2.11 Case Study Contributors

BES Design and Development of Digital Twin Strategies Representation

- **Subramanian Sankaranarayanan**, *ANL*, skrssank@uic.edu
- **Christopher Mayes**, *SLAC National Accelerator Laboratory*, cmayes@slac.stanford.edu
- **Guimei Wang**, *LBNL*, gwang@bnl.gov
- **Sugata Chowdhury**, *Howard University*, sugata.chowdhury@howard.edu
- **Joshua Turner**, *SLAC National Accelerator Laboratory*, joshuat@slac.stanford.edu
- **Timmy Ramirez-Cuesta**, *ORNL*, ramirezcueaj@ornl.gov
- **Kevin Yager**, *BNL*, kyager@bnl.gov
- **Yugang Zhang**, *BNL*, yuzhang@bnl.gov
- **David Prendergast**, *LBNL*, dgprendergast@lbl.gov
- **Remi Dingreville**, *Sandia National Laboratories*, rdingre@sandia.gov
- **Bobby Sumpter**, *ORNL*, sumpterbg@ornl.gov

ESCC Representation

- **Linda Winkler**, *ANL*, winkler@mcs.anl.gov
- **Corey Hall**, *ANL*, chall@anl.gov

5.14 Multifacility Experimentation and Analysis Workflows: X-ray Light Source Perspective

The multifacility experimentation and analysis workflows, focused on X-ray light sources, brings together BES community representatives from X-ray light sources discuss this emerging area of scientific investigation. The case study profiles the overall concept, some early results, and a discussion about what the future will hold.

5.14.1 Discussion Summary

- BES-funded light sources often house multiple beamlines and instruments. Most operate 24/7 for 8-9 months a year with multiple user groups running concurrently. Experiments are diverse in their data size with a daily rate ranging from 10 MB (X-ray spectroscopy) to many TB (imaging and tomography). In many cases, data is transferred from the host facilities to the users for analysis, using portable drives or data transfer tools.

- There are a number of different sources/sinks for the data, depending on the workflow:

    — Synchrotron facility to DOE HPC facility

    — Synchrotron facility to light source or national lab

    — Synchrotron facility to university

    — Synchrotron facility to cloud computing/storage

- Synchrotron facility to an DOE HPC facility is the most heavily used for streaming analysis to process raw data or do post-experiment analysis, as well as archive historical data, train and retrain AI/ML, and use of simulations to inform experimental process.

- The Lab to Lab use case is exercised when multimodal analysis is desired, or if a lab/university has a specialized local computing resource that is used for analysis.

- The synchrotron facility to university use case is mostly used to do the slow transfer of data sets post experiment to a users' local storage/computing resources.

- The synchrotron facility to cloud computing/storage is not used by the facility but may be employed by users.

- The near real-time interpretation of structure revealed by X-ray diffraction requires significant computational resources. The analysis pattern is characterized by bursts of short jobs, requiring very short startup time. Current data collection rates are about 15 TB/day for high compute experiments. Within three to five years, it is expected that this rate will increase to 500 TB/day and in 5+ years to > 1 PB/day.

- During analysis execution at a remote HPC site the data stream can be about 1 Gb/s, but is likely to increase by at least a factor of 10x within three years as more sophisticated methods of visualizing intermediate results are implemented.

- Archiving data between experimental and computational facilities is common, and typical transfers are 8-24 Gbps but are expected to increase by a factor of at least 10 within a year and a factor of 100 within three years.

- Data collection rates are growing exponentially due to light source and detector upgrades, and computational requirements are also growing in proportion. Sending data from a light source facility to the HPC facility requires a data path of sufficient bandwidth.

- By the end of the decade, data aggregated at rates of multiple terabits per second (100-300 Gbps within three years and exceeding 1 Tbps in five years) may flow via ESnet from any of the light sources to any of the DOE HPC facilities. There is an urgent need to upgrade the light source

to ESnet connections at some of the laboratories to 400 Gbps and above to match the input bandwidth at ASCR computing facilities within the next two years.

- Key elements of a future data management strategy for the light sources include a common API for accessing network and computing resources, parallel data transfer tools, high-fidelity data transfer, network performance monitoring, reservations, and dynamic network provisioning.

- There is a growing need for remote acquisition and analysis at user facilities. The development and deployment of tools to facilitate remote data acquisition and analysis continues, despite the challenges of spanning instruments at one facility or instruments across multiple facilities.

- Remote (e.g., streaming) data analysis is often necessary due to the size and time constraints of bulk-data movement. This is compounded by users without access to dedicated computation or storage resources, or a lack of sophisticated software tools. It is often practical to use systems managed by experts at the facility, accessed via remote desktop tools or web applications like Jupyter.

- Data is managed using various facility-specific and instrument-specific solutions. Bluesky Databroker is becoming a common interface at some facilities, but it will never encompass all the data of interest.

- Data is transferred using, HTTP services (e.g. Bluesky Tiled), SFTP, and Globus.

- For some instruments and techniques, storage is cheap and the budget for storage sufficient. In other cases, aggressive downsampling and/or lossy compression are needed.

- It is routine to execute experiments where the physical instrumentation and the compute infrastructure are not colocated. When the compute infrastructure is used to inform experimental decisions (whether fully or semi-autonomously) the network transfer speed can be a significant limiting factor on throughput.

- An emerging pattern is the use of "online" data reduction, which can reduce data volume in order to be able to write to disk and stream to DOE HPC facilities over ESnet. Throughput from the detectors could exceed 1 TBps by 2028, and local data reduction will reduce by a factor of 10 to about 1 Tbps over ESnet.

5.14.2 Multifacility Experimentation and Analysis Workflows: X-ray Light Source Perspective Case Study

Four synchrotron facilities in the US are sponsored by the Scientific User Facilities Division (SUFD) at the Office of BES in the US DOE. Using electron storage rings and linear accelerators, these facilities produce intense photon beams with energies ranging from terahertz to hard X-rays using electron accelerators.

The raw data in an experiment at a light source is captured by a variety of detectors. The interpretation of the data is typically conducted by the user team, in the aftermath of the measurements, with some exceptions where in-line data analysis is critical to successfully conduct the experiment. Due to the growing complexities of the analysis workflow, coupling experimental facilities with dedicated analysis frameworks, often run at HPC facilities, is a new area of investigation for the future use of the BES community.

5.14.2.1 Science Background

Interaction of photons with matter produces signals that reveal subtle details of physical processes, mechanisms of complex chemical reactions, and intricate material properties. All DOE sponsored synchrotron sources operate as national user facilities, with capabilities — both cutting-edge instrumentation and the technical expertise to operate them — available to users from academia, industry, government agencies, and research institutions worldwide. Light sources user community conducts research in a broad range of scientific disciplines, ranging from biology and environmental science through basic and applied chemistry to condensed matter physics. The

majority of users gain access to BES user facilities via a merit-based peer-review process. Scientific proposals for access to the facility's capabilities are evaluated by the independent review committee. Calls for these proposals typically take place every 4 months. Typically, individual users' experiments last between several hours to 7 days When the users' research results are expected to be published in open literature, the access is free, with a small fraction of time allocated to proprietary research on a cost-recovery basis.

National Synchrotron Light Source II (NSLS-II) at BNL, which began operations in 2014 is the newest facility in the DOE complex. ALS and APS, at Berkeley and Argonne national laboratories respectively, plan considerable upgrades in 2023-2026, replacing their storage rings with state-of-the-art designs. These upgrades will bring our light sources into a new era of fully coherent and close-to-theory high brightness, which enable experiments for nanometer-scale spatial resolution, sensitivity to correlated processes with ultrashort time scales, and various techniques relying on beam coherency. LCLS, at SLAC National Accelerator Laboratory, the world's first hard X-ray free-electron laser, will be complemented with LCLS-II which will provide 1000x increase in photon pulse rate.

The raw data in an experiment at a light source is captured by a variety of detectors, such as ion chambers, scintillators, semiconductor (Si, Ge, CdTe) detectors, to name a few, which capture flux, energy, and location of the photons and the time of the events. Multiple, method-specific techniques are used to extract physically meaningful information about structure and dynamics in the specimen under study. Detector data is coupled with the data from the sample environment (e.g., temperature, pressure, etc.) which allows to infer complex structure-function relationships. The interpretation of the data is typically conducted by the user team, in the aftermath of the measurements, with some exceptions where in-line data analysis is critical to successfully conduct the experiment.

Day in the Life of a User – The Data Lifecycle at LCLS

Users submit proposals in response to calls twice a year. Each proposal is reviewed and ranked by the independent PRP. Top-ranking proposals are awarded beam time. Planning begins as soon as beam time is awarded. The LCLS facility begins discussions with users about their needs. Users may begin generating simulations, growing samples, or building instrumentation several months in advance of the beam time. Experiments typically last for 5 days of 12-hour shifts. Experiments are either on a day shift, 9 AM – 9 PM or a night shift, 9 PM – 9 AM and multiple experiments may run simultaneously. Experiment configurations at the beamlines are not static and change weekly. Users arrive on-site several days prior to their beamtime and spend time checking out the beamline, setting up for the experiment, or testing analysis workflows, which may require some access to computing resources. Any given experiment can have from a handful of collaborators to 70 – 80, some of which are on-site and driving the experiment during the beam time while the others are analyzing data from their home institutions and participating with on-site colleagues via telecommunication software.

In addition to data collection for scientific inquiry, data is also collected to calibrate the detectors, align the beamline, or tune sample delivery. By the last few days of beamtime, most of the beamline parameters have been optimized to extract the most science information, so average rates of useful data written to disk usually increase during the last days of an experiment. Some experiments require some coordinated excitation of the sample followed by data-taking. For example, a laser pulse may excite a material which is followed up by a precisely timed series of X-ray pulses to image the system at different time intervals following the stimulus. In this case, the data system coordinates this sequence of events and collects metadata that is stored alongside the instrument data to describe the conditions under which the experiment was done.

For 80% of experiments, local (SLAC) compute resources are sufficient to provide quasi-real-time analysis, but for 20% of experiments, we stream data to a DOE Leadership Class Facility (NERSC, OLCF, ALCF) for analysis. The intent is for beamline scientists to be able to keep up with the acquisition rate of their data and obtain analysis results within minutes of ending a period of data acquisition. After data is copied to the spindle-based offline storage, users may copy data to their home institution for analysis. Approximately a tenth of users copy

data to their home institutions, usually using Globus. Most users complete their offline analysis using SLAC-provided computing resources in the 4 months following their experiment. On average, a typical user will rerun over their entire data set up to 10 times. In some cases, users will also reanalyze data taken during a previous beamtime (up to 10 years ago) or combine results with data taken at other light sources to include (but not limited to) ALS, APS, NSLS-II, SSRL, SACLA, PAL, EuXFEL.

The timescales for generation of data vary from group to group but can be summarized as follows. Ahead of an experiment, users may generate simulations in collaboration with theorists. This data is privately managed by the users. During the experiment, beamline scientists work closely with experimenters to optimize the data acquisition (calibrate detectors, detector to IP distances, sample delivery, hit rate) and optimize optics/laser timing. Once those are setup, data acquisition begins, which is an interactive process. Data is acquired for about 10 minutes at a time, and analyses are run "live" during acquisition. The goal is to get an answer to a scientific question within a few minutes of stopping acquisition so the experimenters can decide how best to steer the experiment. Data quality is also assessed as part of this process: is the calibration/timing/alignment good? Is the data reduction performing optimally? Is the signal to noise as expected? If everything looks good, keep acquiring data. If not, modify the analysis code, take more data, and determine the source of the problem. Then, fix the problem and continue data taking. In some cases, there is a comparison between theory and experiment.

5.14.2.2 Collaborators

BES-funded light sources use a peer-reviewed proposal system to enable user access to their capabilities. Synchrotron facilities house multiple beamlines (approx. 30, 40 and 65 for NSLS-II, ALS, and APS, respectively) whereas LCLS operates 10 instruments. Most of the beamlines are run simultaneously 24/7 for 8-9 months a year. Multiple user groups run their experiments during operating periods, bringing the total number of users to thousands a year (1500, 2000, 5000 and 1000 annually for NSLS-II, ALS, APS, and LCLS respectively). Experiments are very diverse in their data size with a daily rate ranging from 10 MB (X-ray spectroscopy) to many TB (imaging and tomography). In many cases, data is transferred from the host facilities to the users for analysis, using portable drives or data transfer.

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
ALL USERS	varies, depending on whether they are a multimodal data producer or only a consumer	both portal and transfer	kB to TB depending on instrument	spanning slow post hoc analysis up to live-streaming real time feedback with low-latency requirements	Yes, transport modes include HTTP and Globus	Need partial access, streaming support, and speed

Table 5.14.1: Multifacility X-ray Collaborations

There are a number of different sources/sinks for the data, depending on the workflow and the needs of the experimenters. The data paths can be summarized as follows: 1) Synchrotron facility to ASCR compute facility, 2) Synchrotron facility to light source or national lab, 3) Synchrotron facility to university, and 4) Synchrotron facility to cloud computing/storage such as Google, Amazon AWS, or Microsoft Azure. The first, Synchrotron facility to ASCR facility is the most heavily used. It is used to do streaming analysis using ASCR computing to process raw data or do post-experiment analysis. A primary copy of the data set is always kept at synchrotron facility and a secondary copy is kept at NERSC. The synchrotron to ASCR facility workflow is used to archive the data sets at NERSC and to train/retrain AI/ML models that are being used by software at a beamline during experiments or to analyze data post experiment. An emerging use case is the use of simulations, performed at the ASCR facility, and then used during the experiment to inform data analysis. The second data path, lab to lab, is used when multimodal analysis is desired, or if a lab/university has a specialized local computing resource

that is used for analysis. The third data path, synchrotron facility to university, is mostly used to do the slow transfer of data sets post experiment to a user's local storage/computing resources. Finally, the synchrotron facility to cloud computing/storage is not used by the facility but may be employed by users. The synchrotron facility has investigated the possibility of using cloud resources but found that the costs associated with I/O were unsustainably high. However, synchrotron facility does not prevent users from making use of these resources if they wish, and if multimodal analysis is desired, cloud storage and computing becomes a more attractive option. In some cases, such as streaming real-time analysis during an experiment, there may be strong real-time constraints on the flow of data between the facility and a computing resource.

5.14.2.3 Use of Instruments and Facilities

Due to the variety of users, experiments at light sources will not cover all detection and data transfer and data interpretation schemes.

Use Case: XPCS

XPCS is a technique requiring highly coherent X-ray beam available at modern synchrotron sources. It enables researchers to investigate nanoscale and mesoscale dynamics of complex materials on timescales ranging from below milliseconds to minutes or hours. XPCS is based on measuring time correlation functions of the speckle fluctuations that occur when a coherent X-ray beam is scattered from a disordered sample. It can be used to measure equilibrium dynamics via the "usual" single-speckle intensity-intensity autocorrelation functions. When combined with 2D area detectors and a multispeckle detection technique, it can also be used to measure nonstationary, nonequilibrium dynamics via two-time correlation functions.

As an example of the current state of affairs, Coherent Hard X-ray Scattering (CHX) beamline provides state-of-the-art flux of coherent photons, leveraging exceptional brilliance of the NSLS-II storage ring. CHX employs a 2D detector with 4 x 106 pixels (Eiger 4K) operating at the frame rate of 1kHz. Speckle patterns are collected for 30 s. Analysis of the data is required to steer the experiment. The collected data is stored locally and analyzed using a 72-core, 1 TB RAM server, local to the instrument. With the available computer capabilities, the data analysis, which is a computationally intensive process takes ~15 min, reducing the duty cycle of the instrument to below 10 %. Enabling in-line analysis will require 100Gbps data transfer to a multinode cluster and optimization of the analysis code to enable parallelization or GPU utilization.

Use Case: LCLS X-ray Scattering Experiments

X-ray scattering experiments are a powerful tool to determine the molecular structure and function of unknown samples, such as COVID-19 viral proteins . In crystallography experiments, molecular structure is determined by merging the X-ray diffraction patterns from millions to billions of protein crystals exposed in random orientations. The near real-time interpretation of structure revealed by X-ray diffraction requires significant computational resources.
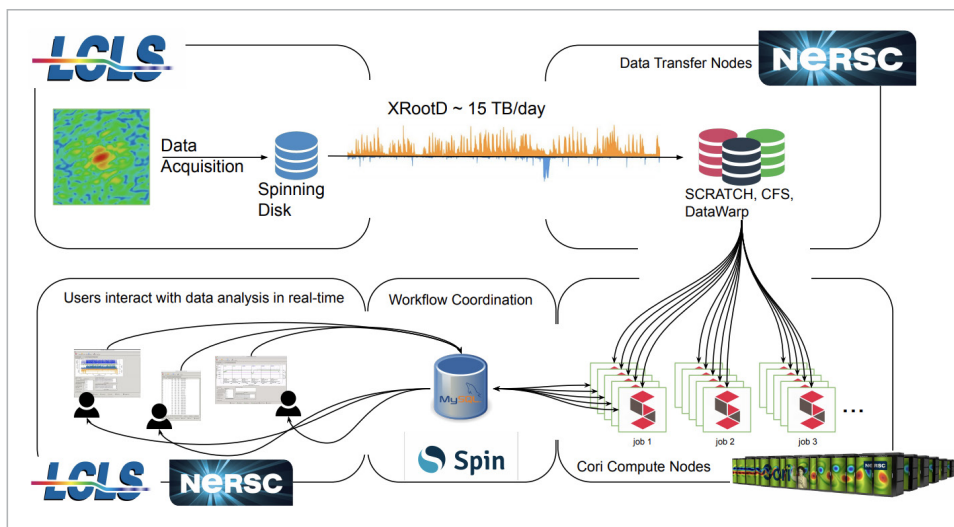
Figure 5.14.1: LCLS SFX/SPI example is a representative information extraction pipeline. In this example, raw data is acquired at the light source facility (**LCLS**), automatically transferred to an ASCR facility (**NERSC**) via ESnet, and analyzed using a scalable software library appropriate to the science domain. It is possible to achieve a turnaround time from data acquisition to full molecular reconstruction within 10 minutes. (ref arXiv.2106.11469v2)

The analysis pattern is characterized by bursts of short jobs, requiring very short startup time. Current data collection rates are about 15 TB/day for high compute experiments. Within three to five years, it is expected that this rate will increase to 500 TB/day and in five+ years to > 1 PB/day. During analysis execution at a remote HPC site, updates on job progress, intermediate results, and logfiles may be transferred back to the originating facility. After analysis jobs are complete, summary results are transferred and made available to users via a web browser at LCLS. In total, this data stream from NERSC back to LCLS is about 1 Gb/s, but is likely to increase by at least a factor of 10x within three years as more sophisticated methods of visualizing intermediate results are implemented. In addition to transferring data for running experiments to NERSC for fast-feedback processing, LCLS also uses ESnet to archive data at NERSC. Typical transfers are 8 to 24 Gbps but are expected to increase by a factor of at least 10 within a year and a factor of 100 within three years.

Such "big data" experiments are challenging to execute and often require fast turnaround between data acquisition and data analysis to enable experimenters to make informed decisions when driving experiments in order to make use of the limited resources of X-ray beam time and samples. Data collection rates are growing exponentially due to light source and detector upgrades, and computational requirements are also growing in proportion. In order to analyze data on experiment timescales, it is necessary to send data to remote high performance computing facilities such as NERSC, the ALCF, or the OLCF. Sending data from a light source facility to the HPC facility requires a data path of sufficient bandwidth.

In principle, by the end of the decade, data aggregated at rates of multiple terabits per second (100-300 Gbps within three years and exceeding 1 Tbps in five years) may flow via ESnet from any of the light sources (ALS, APS, LCLS, NSLS-II, SSRL) to any of the ASCR compute facilities (ALCF, NERSC, OLCF). There is an urgent need to upgrade the light source to ESnet connections at some of the laboratories to 400 Gbps and above to match the input bandwidth at ASCR computing facilities within the next two years.

Key elements of a future data management strategy for the light sources include a common API for accessing network and computing resources, parallel data transfer tools, high-fidelity data transfer, network performance monitoring, reservations, and dynamic network provisioning. Consideration should be given to defining a methodology for handling simulated and other data to enable multimodal analysis and for reporting meaningful intermediate analysis results back to the user without abusing the network.

Use Case: Autonomous Experiment Steering for Scientific Discovery at Synchrotron Light Sources

Autonomous experiment steering is an emerging mode of conducting science at the five US DOE-funded light sources. Within the light sources, this capability has the promise to provide automated setup of the source and sample alignment, intelligent data collection, quality verification, data reduction, and coupling of experimentally derived data with information derived from theory, models, and simulations. Autonomous experiment steering has the potential to unlock new materials science knowledge to, for example, better understand failure modes in materials, enable the synthesis of new materials, aid in the creation of purpose-built designer materials, and assist in additive manufacturing processes.

Feedback must often be obtained on timescales too short for humans to react in order to plan and steer experiments to, for example, catch rare events or see fast processes. Due to the intrinsic capabilities of the current and soon to be upgraded sources, coupled with high data rate detectors that generate large volumes of data at increasingly higher rates, advanced computational techniques, including AI/ML, must be employed to realize autonomous steering of light source experiments. These methods require the utilization of considerable supercomputing power to process data and train AI/ML models that may then be used to make real-time decisions using edge or local computing systems (see Figure 5.14.2).

Figure 5.14.2: Prototypical autonomous experiment ML data workflow. Data generated at a light source instrument (in this case at LCLS) is streamed directly to a supercomputing center (in this case ALCF). An ML model is trained on the fly. The trained model is deployed on an edge-computing device at the light source instrument to make experiment decisions in real time.

Networking must be able to connect light source instruments to edge, local, campus, and centralized computing facilities reliability, and with low latency. Terabit per second networking and beyond will be required to handle the large amounts of data expected in the coming years.

The further development and application of autonomous experiment steering at the light sources requires that a number of gaps in capabilities and infrastructure across the light sources, Laboratories, and computing centers be filled, including sufficient, reliable bandwidth, sufficient and sustainable data storage resources, on-demand access to large-scale computing systems, transparent access to facilities and systems across the complex, including federated identity and shared data protocols, and software tools and infrastructure to facilitate the development of scientific data workflows that operate across the DOE's distributed resource landscape.

Autonomous experiment steering will become an increasingly relied upon capability at the light sources as the decade progresses. Due to the high computational cost associated with data processing, reduction, and analysis, and of model training on such large data sets, the light sources will need to employ the use of geographically distributed computing facilities and data repositories. This necessitates high bandwidth, low latency network connections between the light sources and remote computing and data storage centers.

Unified solutions across the light sources are required in order to leverage efficiencies of scale, and to provide facility users with the ability to easily and transparently manipulate data across the light sources. A shared computational fabric for the complex should be developed that connects light source instruments (and other SUF) to a multitiered, distributed computing landscape, including edge, local, campus, and supercomputing centers, data repositories and archives, and facility user institutions in a seamless and transparent manner. This necessitates the development of advanced networking capabilities and increased networking bandwidth, sustainable and discoverable data repositories, on-demand real-time supercomputing access, and workflow and orchestration tools.

Use Case: Autonomous Experiment Steering for Scientific Discovery

Coupling of simulations and development of "digital twins" to light source experiments has the potential to unlock new materials science knowledge to, for example, better understand failure modes in materials, enable the synthesis of new materials, aid in the creation of purpose-built designer materials, and assist in additive manufacturing processes. It will also allow for a more efficient and optimum use of beamtime at the light sources.

The coupling of simulations with light source experiments can be split up into three main areas in the experimental life cycle. Firstly, before the experiment, simulations can be used to help prepare, plan, and determine if the experiment is even feasible. Secondly, during the allocated beam time, simulations can help guide and inform the strategy and guide the experiment. Finally, simulations can be used to aid in the data analysis in order to extract the maximum scientific information from the data.

Figure 5.14.3: Diagram that shows the flow chart of using a digital twin with light source experiments

Coupling simulations experiment steering will become an increasingly relied upon capability at the light sources as the decade progresses. Due to the high computational cost associated with data processing, reduction, and analysis, and of model training on such large data sets, the light sources will need to employ the use of geographically distributed computing facilities and data repositories. This necessitates high bandwidth, low latency network connections between the light sources and remote computing and data storage centers.

Unified solutions across the light sources are required in order to leverage efficiencies of scale, and to provide facility users with the ability to easily and transparently manipulate data across the light sources. A shared computational fabric for the complex should be developed that connects light source instruments (and other SUF) to a multitiered, distributed computing landscape, including edge, local, campus, and supercomputing centers, data repositories and archives, and facility user institutions in a seamless and transparent manner.

This necessitates the development of advanced networking capabilities and increased networking bandwidth, sustainable and discoverable data repositories, on-demand real-time supercomputing access, and workflow and orchestration tools. Also, as the majority of light source users are not experts in using computational facilities, there needs to be easy access and both in terms of obtaining the resources and performing the calculations.

As an example, here is the multimodal experiment we foresee becoming more common in the coming years. Recently, two instances of a single sample were placed simultaneous in two separate beamline instruments, which specialize in complementary techniques. The data streams from each instrument were fed to a single point of control, which orchestrated the sequence of measurements that each instrument should perform, synthesizing input from AI/ML agents and experimentalists. In this case, the two beamline instruments happened to be in the same facility, so all control and data communication took place over the local networks. In the future, we anticipate experiments that may combine two or more instruments from different facilities connected by wider networks.

5.14.2.4 Process of Science

Serial Femtosecond X-ray Crystallography (SFX) reveals the reaction mechanism by providing separate atomic structures for each metastable state, and several time points in between. SFX experiments offer huge benefits to the study of macromolecules, including the availability of femtosecond time resolution and the avoidance of radiation damage under operando conditions. SFX techniques will be instrumental in many experiments, ranging from the determination of macromolecules structure and dynamics, to the understanding and controlling of materials nucleation pathways, to the study of oxygenic photosynthesis.

In SFX, the structural information is derived from the diffraction data collected from a stream of individual crystals, with the primary feature extraction step consisting of measuring the Bragg spot intensities on each diffraction pattern. The main steps in the SFX algorithm are (1) identifying the Bragg diffraction spots, (2) deducing the geometry of the lattice repeat, (3) refining the model again and (4) summing the X-ray signal in each spot for further analysis.

Figure 5.14.4: Pictorial representation of the SFX pipeline

Based on a 10% hit rate and on the CCTBX workflow compute requirements, we project that the rate of useful SFX events will approach 5 kHz and the computing resources needed to reconstruct SFX data in LCLS-II will run routinely in the few tens of PFLOPS and some of the more demanding algorithms, like IOTA, will require several tens of PFLOPS.

Use Case: Heterogeneous Catalysis

A second illustrative example is in the area of heterogeneous catalysis, where functional systems are neither homogeneous, nor static. The evolution of atomic and electronic structure, the making and breaking of chemical bonds, and the exchange of vibrational energy through intermediate states ultimately determine functionality. These interactions further lead to dynamic restructuring of catalyst materials during reaction. Knowing the time evolution of the atomic and electronic structure of molecules and substrates, particularly near elusive transition states, is critical to developing a predictive understanding for catalysts design

Today we are unable to develop a complete picture of this structural evolution — neither on the ultrafast timescales relevant to atomic motion, nor the nsec to msec timescales characteristic of diffusion and materials evolution. LCLS has enabled the study of simple surface reactions, and reported the first observation of a surface transition state — based on studies of ideal crystals, with reactants prepared at high concentrations in vacuum. However, these studies do not address "working" catalysts which are typically of low dimensionality (e.g. nanoclusters of metals on an oxidic support), where coupled fluctuations of the electronic and atomic structure become increasingly important.

LCLS-II-HE, coupled with advanced computational approaches applied to massive data sets, will enable completely new approaches for simultaneously following both the atomic and electronic structure of heterogeneous catalysts in operation. In one such proposed approach, nanocatalysts are prepared (either with or without preadsorbed reactants) and interrogated pulse by pulse using a gas-phase jet, liquid particle injector, or translating tape. Multimodal characterization incorporates ultrafast X-ray spectroscopy (e.g. EXAFS, XES, or RIXS) or photoelectron spectroscopy to probes the electronic structure and chemical environment, while coherent X-ray scattering from the same hard X-ray pulse probes the atomic structure. Demonstration experiments in the soft X-ray range at FLASH and at LCLS highlight the promise of this approach for characterizing small heterogeneous ensembles of nanoparticles at the atomic scale using hard X-rays.

SPI

SPI is a promising emerging technique for such applications, Here, diffraction images are collected from individual particles, and are used to determine molecular (or atomic) structure, even from multiple conformational states (or nonidentical particles) in operating conditions that are inaccessible through other methods. These techniques will be instrumental in many areas of science, ranging from understanding and controlling nanomaterials self-assembly, to the chemistry and morphology of combustion aerosols to the coupled electronic and nuclear dynamics in heterogeneous (nano) catalysis.

However, determining structure from these experiments is challenging, since orientations and states of imaged particles are unknown and images are highly contaminated with noise. Furthermore, the number of useful images is often limited by achievable single-particle hit rates. Advanced algorithms currently under development (e.g. M-TIP by the CAMERA group) introduce an iterative projection framework to simultaneously determine orientations, states, and molecular structure from limited single-particle data by leveraging structural constraints throughout the reconstruction and offer a potential pathway to increasing the amount of information that can be extracted from single-particle diffraction.

A fundamental challenge in SPI is that the orientation of each imaged particle is unknown and must be recovered to determine structural information. Additionally, many samples display conformational flexibility and may exist in one of many possible structural states. To account for varying structural states and avoid a loss of resolution due to averaging of states, the diffraction patterns may need to be classified to the correct state. Furthermore, single particles scatter very few photons; hence the images are heavily contaminated by shot noise, often with less than a photon per Shannon pixel at high scattering angles.

5.14.2.5 Remote Science Activities

The COVID emergency accelerated growing interest in remote acquisition and analysis at user facilities. As a result, staff and users have increased comfort and fluency with remote work. This includes planning with distributed teams, using video conferencing and remote desktop tools, and collaborative authorship of both prose documents and software code. Even with the return to normal operations, facilities are offering more telework and remote arrangements to new hires and existing staff. The development and deployment of tools to facilitate remote data acquisition and analysis, prioritized during the initial months of the COVID emergency, continues today. Multimodal experiments, whether spanning instruments at one facility or instruments across multiple facilities, present many of the same challenges and benefit from the solutions.

Remote data analysis (where the bulk of the data stays near where it was acquired) is sometimes necessary even if the user visits the site. The data may be too large to practically copy or move. The user may not have storage resources to keep a copy on their own resources. Or, the user may not have computation resources to process data on their own resources. Finally, even with small data sets, users may not have software sophistication to set up the software environment they need to do their work. It may be more practical to use systems managed by experts at the facility, accessed via remote desktop tools or web applications like Jupyter.

At LCLS, operating experiments is the shared responsibility of beamline scientists, local staff, and on-site users. LCLS has web services, a data management system, and a data analysis infrastructure that allows for virtual access to all of the resources necessary to analyze the data taken at LCLS from a remote location. In most cases, all the same resources are available to users doing analysis at home as at SLAC. In most cases, the data remains in LCLS local storage and LCLS offline processing resource are used to analyze. Because of the uniqueness and changeability of LCLS instruments, LCLS does not envision remote operation of the instruments.

5.14.2.6 Software Requirements

The data is managed using various facility-specific and instrument-specific solutions. At NSLS-II and, increasingly, at other light sources, Bluesky Databroker is becoming a common interface, but it will never encompass all the data of interest.

Data is transferred using, HTTP services (e.g. Bluesky Tiled), SFTP, and Globus.

The accelerator systems and beamline instruments use the EPICS for low-level device control. The Bluesky Suite of tools has been broadly adopted for at NSLS-II high-level experiment control and increasingly, becoming a common interface at other light sources . Remote access tools, such as ssh, NX, and NoMachine, will continue to be utilized to enable remote access for LCLS experiments.

The basic requirements for a data management system typically include:

- Security: Authentication/Authorization (WebAuth using SLAC LDAP, Stanford SSO)
- Scan/run control (Bluesky, Python, epics, custom DAQ, Kafka message bus)
- File catalog (was iRods, migrating to Rucio by the end of 2022)
- Experiment metadata (mongodb, Kafka)
- Electronic Logbook (custom web application)
- Samples (mongodb)
- Workflow (Kafka, LSF/SLURM, AirFlow, APIS provides feedback)
- Feedback and reporting (custom web form)
- Instrument operator portals (MongoDB)

In addition, we also require support for:

- Composability/site specificity (custom web service, Kafka/Websocket)
- Scalability/Reliability (MongoDB, Kafka)
- Backup/restore (MongoDB — mongodump, mongorestore)
- Scriptability (Python, Kerberos endpoint, PAM endpoint)
- Ease of deployment/upgrades (Python+Javascript application)
- Ease of development (Conda, Kubernetes, document-based store with Python_Javascript)

Python is widely used for experimental software these days; it also provided many simple web frameworks. A service-oriented architecture based on a REST based web service API is commonly used; and Flask seems to be a popular choice for building REST services with Python. The principal requirements for the backend are scalability/reliability; MongoDB has built in sharding/replication and the ability to grow a cluster seamlessly. MongoDB also uses a document model; which significantly enhances the ease of development avoiding much of the boilerplate associated with other backends. In addition, a message oriented architecture where all changes of interest are published on a message bus, enables composability of applications. The ability to separate out site-specific features into serverless components that react to these messages and shape the business data to suit the needs of the site is also useful. Using Kafka as a message bus allows for clients to reliably continue from where they left off; this facilitates building reliable components. Kafka messages can also be routed to the frontend's using websockets to facilitate reactive UI's eliminating the need for polling.

5.14.2.7 Additional Network and Data Architecture

BNL HTSN

BNL has implemented a vendor agnostic, resilient, scalable and modular terabit per second (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 primary components:

- **Network perimeter**
  - Three diverse 100 Gbps circuits that peer with ESnet. 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 15-20 PBs of data monthly.
- **Science DMZ**
  - Supports open, high-speed WAN/Internet) access for all scientific collaborations throughout the BNL campus.
- **Science Core**
  - A Tbps DCI for data intensive collaborations at BNL. This network interconnect enables high-speed connectivity between collaborations such as ATLAS, STAR, PHENIX, CAD, CFN, NSLS-II, High-Performance Computing 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.

- **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.

  — 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 ASN but does not have Internal BGP (iBGP) peering's between them. Each Leaf or pairs of Leaves will require their own ASN.

- **Storage Core**

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

BNL Next-Generation Network Perimeter

The network perimeter at BNL is a high-speed and fault-tolerant network infrastructure that provides the BNL site connectivity to the Internet and various scientific wide area networks. It supports numerous data intensive collaborations such as BES, Biological and Environmental Research, HEP, and Nuclear Physics. It also supports critical campus services such as workstations, phone service, security, safety and monitoring and enterprise and cloud computing. Since being placed into production in 2013, the network perimeter has transmitted over 100+ PBs of data per year to numerous scientific collaborations worldwide.

In 2013 the BNL network perimeter was bleeding edge 100 GbE technology. Now, the hardware has reached 8+ years in age and it is no longer cost-effective to purchase additional hardware for these platforms. Newer platforms today support much greater 100 GbE interface densities along with supporting 400 GbE which will allow BNL to support all its data intensive collaborations well into the future. With these factors in mind, BNL will possibly procure a next-generation network perimeter within the next one to two years. This will prepare the laboratory to meet the missions needs of all the data intensive collaborations at BNL.

The SLAC Shared Science Data Facility (S3DF)

S3DF is a compute, storage and network architecture designed to support massive scale analytics required by the SLAC experimental facilities. S3DF will effectively replace several outdated SLAC systems. S3DF is considered an all-new environment that serves as a greenfield for modern technologies: storage, containerization, interactive workflows, data management, identity and access.

The S3DF infrastructure is optimized for data analytics and is characterized by large, massive throughput, high concurrency storage systems. Over the next decade, S3DF will deploy a few PFLOPS of CPU computing, tens of PFLOPS of GPU computing, hundreds of PBs of fast storage, and more than two exabytes of archiving capabilities.

A significant fraction of these capabilities will be dedicated to LCLS — more specifically 2 PFLOPS of CPU and 5 PFLOPS of GPU computational resources, and 50 PFLOPS of fast access storage, will be reserved for LCLS experiments.

The S3DF will be complemented by DOE HEC facilities (OLCF, ALCF and NERSC) for the most computationally demanding LCLS experiments. This approach of combining local and complex-wide facilities will allow sizing S3DF to satisfy the requirements of the majority of LCLS experiments, rather than being driven by the most demanding techniques, while at the same time mitigating the risks related to unplanned outages

at HEC facilities and reducing some of the complexities associated with running on a supercomputer (resource orchestration, users accounts, lack of privileges, priority management, container requirements, etc).

S3DF will be hosted in the SRCF. SLAC and Stanford have recently completed the plans for a new SRCF module (SRCF-II) that will double the current data center capabilities, for a total of 6 MW. The new data center module will be available to host the S3DF expansion by early spring 2023.
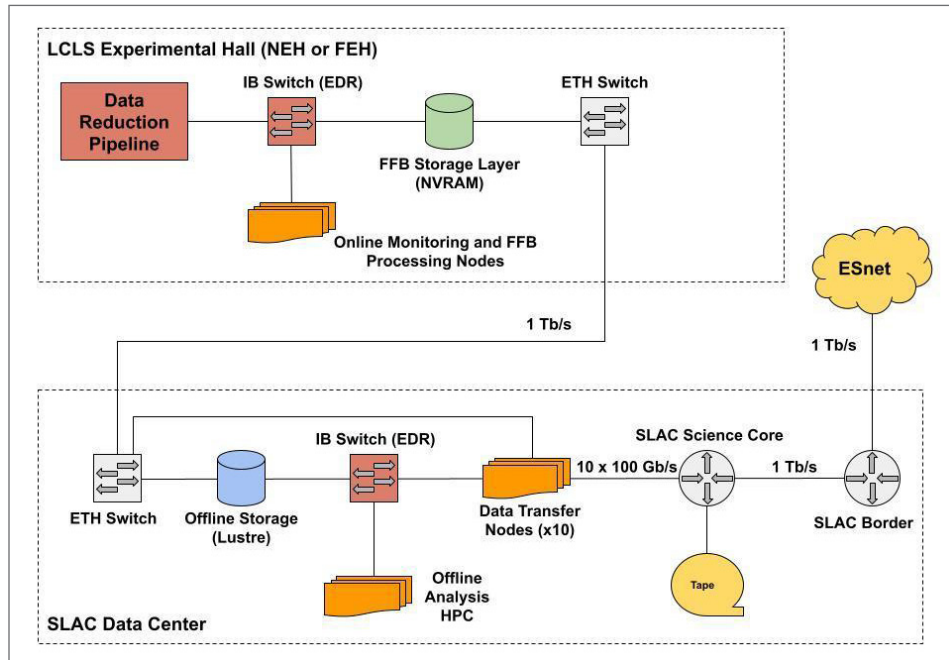
Figure 5.14.5: S3DF Diagram
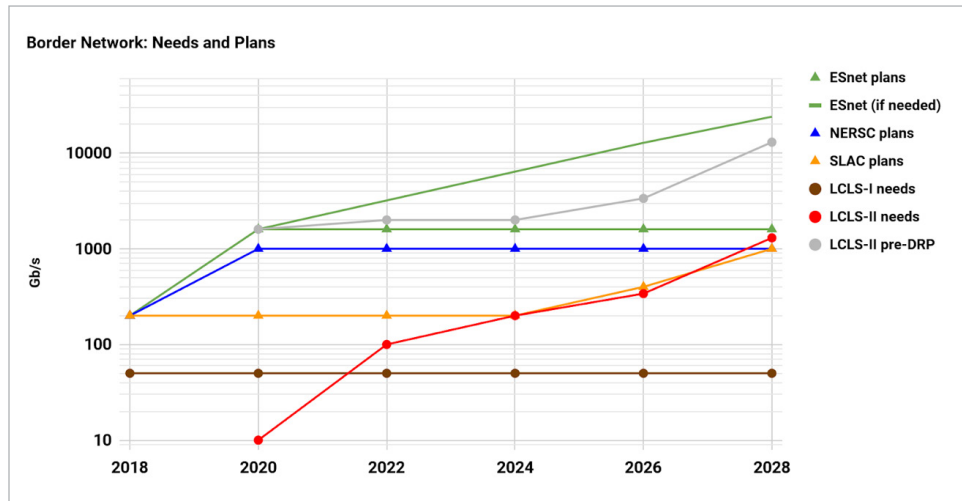
Figure 5.14.6: SLAC Projected Capacity Requirements

5.14.2.8 Use of Cloud Services

Commercial cloud services are used, primarily for blob storage and cloud VMs, with some experimental use of Kubernetes. At NSLS-II, both Microsoft Azure and Amazon AWS are used.

Presently, mostly used to support public demos and tutorials, such as the Users' Meetings, where data and compute resources need to be made available to users outside the lab network. In the future, applications may include a Content Distribution Network to enable faster delivery of data to nonlocal users; archival and, relatedly, a way to transfer stewardship of data from the facility to users (i.e., "If you want to keep it, it's moving to an account that you pay for."); and compute infrastructure for "burst-y" workloads that are more cost-effective in the Cloud.

5.14.2.9 Data-Related Resource Constraints

For many instruments and techniques, storage is cheap and the budget for storage sufficient. There are notable and interesting exceptions, where aggressive down sampling and/or lossy compression are needed. It is becoming more routine to execute experiments where the physical instrumentation and the compute infrastructure are not collocated. When the compute infrastructure is used to inform experimental decisions (whether fully or semi-autonomously) the network transfer speed can be a significant limiting factor on throughput. Locating and accessing specific data of interest, which may be a region of interest inside a large data set, is of interest. Streaming access to data, to kick off data processing while measurements are still in progress, is also of interest.

LCLS is doing online data reduction to reduce data volume in order to be able to write to disk and stream to NERSC over ESnet. Throughput from the detectors will exceed 1 TBps by 2028, and local data reduction will reduce by a factor of 10 to about 1 Tbps over ESnet to NERSC.

5.14.2.10 Outstanding Issues

None to report at this time.

5.14.2.11 Case Study Contributors

Multifacility Experimentation and Analysis Workflows: X-ray Light Source Perspective Representation

- **Eli Stavitski,** *BNL,* istavitski@bnl.gov
- **Daniel Allan,** *BNL,* dallan@bnl.gov
- **Maria Chan,** *ANL,* mchan@anl.gov
- **Jana Thayer,** *SLAC National Accelerator Laboratory,* jana@slac.stanford.edu
- **Wanli Yang,** *LBNL,* wlyang@lbl.gov

ESCC Representation

- **Vinny Bonafede,** *BNL,* bonafede@bnl.gov
- **Mark Lukasczyk,** *BNL,* mlukasczyk@bnl.gov
- **Rune Stromsness,** *LBNL,* rstrom@lbl.gov
- **Richard Simon,** *LBNL,* rsimon@lbl.gov
- **Mark Foster,** *SLAC National Accelerator Laboratory,* fosterm@slac.stanford.edu

5.15 Multifacility Experimentation and Analysis Workflows: Neutron Scattering Perspective

The multifacility experimentation and analysis workflows, focused on neutron scattering facilities, brings together BES community representatives from neutron scattering facilities discuss this emerging area of scientific investigation. The case study profiles the overall concept, some early results, and a discussion about what the future will hold.

5.15.1 Discussion Summary

- The future or neutron scattering science and the study of quantum materials will require large quantities of data, from several locations, comprised of simulations and data, needing to be co-analyzed and used in ML.

- Growing data volumes and rates provide a significant challenge, and current condensed matter physics tools are being replaced by more interdisciplinary team science approaches. Existing and future experiments demand powerful analysis capabilities and real-time feedback.

- Neutron facilities attract thousands of users, many of them will use multiple experimental and computational facilities. As advanced analysis includes AI/ML based methods sharing experimental and simulated data, networks will become increasingly important. This shift will also require 'programmatic' ways to search and retrieve large amounts of data.

- Large-scale analysis and data transfer capabilities are critical to the development of AI accelerated codes. Network capabilities to allow these collaborations, and data transfers, need to be put in place.

- Experimental and simulation data, computer resources, and collaborators, are located in different location requiring remote use.

- Future multifacility neutron experiments will require:

  — Simulations and training for neutron experiments (spectroscopy, reflectometry, and diffuse scattering) have data sets in the 10s to 100s of TB ranges.

  — Pre-trained AI will require to be extended in real-time during experiments. This involves new simulations and training which will be at different remote locations.

  — Experiment data sets from neutron instruments are in the 10s of GB to TB range and will need to be streamed for specialized AI to undertake analytics to recognize patterns and identify features.

  — The nature of specialized computational resources and codes mean that simulation and data tasks will be distributed over different locations.

  — ESnet provides the network infrastructure to bridge this gap. In addition, co-analysis with data sets from other facilities, especially microscopy, synchrotron, and XFEL data is needed.

- Workflows relying on ML will require to rethink searching, access and retrieval of data for training as well as data management related to the derived networks and products.

5.15.2 Multifacility Experimentation and Analysis Workflows: Neutron Scattering Perspective Case Study

Researchers are faced with increasingly complex and difficult scientific questions and challenges that require the use of data involving multiple experimental techniques as well as cutting edge modeling to solve. These involve large quantities of heterogeneous data from several institutions and facilities in the form of simulations

and measured data that need to be co-analyzed and involve the application of data science techniques such as ML. In addition, real-time simulations and analysis are demanded during experiments involving remote users and resources. A domain that typifies these challenges is quantum materials where the first such studies are being performed[1 2 3 4]. Coupling resources at dedicated analysis facilities, to those of experimental facilities, offers unparalleled opportunities to investigate new workflows.

5.15.2.1 Science Background

Quantum materials are being intensively studied due to their exceptional properties and new physical principles involved. To harness and control quantum coherence in these materials, and the interfaces between them requires rationally designing materials for novel technologies ranging from energy harvesting to devices for quantum information, and sensing beyond classical limits. Essential for this is undertaking codesign between high performance simulations based on state-of-the-art theory with materials and experiments.

A significant challenge is the growing data volumes and rates, which means that conventional condensed matter physics tools are being replaced by more interdisciplinary team science. Neutron and X-ray facilities epitomize the demands of cutting-edge experiments. These experiments demand powerful analysis capabilities and real-time feedback to aid researchers. Meanwhile high-throughput theoretical modeling of quantum materials is being performed by other team members at other locations using specialized computational codes and resources. Unsupervised learning techniques, including autoencoders and generative modeling, provide the means to identify patterns in data and simulations, relevant variables involved, and understandable low-dimensional descriptions of otherwise complex phenomena.

5.15.2.2 Collaborators

Neutron and X-ray user facilities attract thousands of users carrying our experiments. Many of them will use multiple experimental and computational facilities in their research. As advanced analysis includes AI/ML based methods sharing experimental and simulated data as well as potentially trained networks will become increasingly important.

Collaborators will come from national and international research facilities. While this case study focuses on quantum materials, the types of workflows discussed here will have broader applicability. Collaborators will have a varied computational skill set and resources at their home institution.

Key collaborators for this case study are the University of Tennessee, ORNL, ANL and NERSC.

---

[1] M. Doucet, A. M. Samarakoon, C. Do, W. T. Heller, R. Archibald, D. A. Tennant, T. Proffen, G. E. Granroth, Machine Learning for Neutron Scattering at ORNL, Machine Learning: Science and Technology, https://doi.org/10.1088/2632-2153/abcf88. (2020).

[2] Z. Chen, N. Andrejevic, N. Drucker, T. Nguyen, R. P. Zian, T. Smidt, Y. Wang, R. Ernstorfer, A. Tennant, M. Chen, M. Li, Machine Learning on Neutron and X-ray Scattering, Chem. Phys. Reviews 2, 031301 (2021); https://doi.org/10.48550/arXiv.2102.03024, ArXiv:2102.03024.

[3] A. Samarakoon, P. Laurell, A. Banerjee, S, Nagler, S. Okamoto, D. A. Tennant, Extraction of the interaction parameters for alpha-RuCl3 from neutron data using machine learning, Phys Rev Research 4 L022061 (2022).

[4] A.M. Samarakoon, D.A. Tennant, F. Ye, Q. Zhang, S.A. Grigera, Integration of Machine learning with Neutron Scattering: Hamiltonian Tuning in Spin Ice with Pressure, Communications Materials accepted (2022); https://arxiv.org/abs/2110.15817.

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
EXPERIMENTAL USER FACILITIES (E.G. SNS)	Primary	Data transfer, API access desired	10-200GB	Ad hoc in batches per experiment	No	API type search and transfer would streamline workflow
LEADERSHIP COMPUTATIONAL FACILITIES	Primary	Data transfer or none if facility is location of main analysis	10-500GB	Ad hoc, special needs for ML to be explored	No	API type search and transfer would streamline workflow
INSTITUTIONAL COMPUTING FACILITIES	Primary	Data transfer	10-200GB	Ad hoc, special needs for ML to be explored	No	API type search and transfer would stream-line workflow

Table 5.15.1: Multifacility Neutron Source Collaborations

5.15.2.3 Use of Instruments and Facilities

Science enabled by multifacility workflows will utilize experimental user facilities such as the SNS and the HFIR at ORNL as well as computational facilities. Computational facilities will include organizational resources such as the CADES at ORNL as well as leadership computing such as Summit (and soon Frontier) at ORNL. Scientists will choose experimental neutron and X-ray facilities based on the specific experimental problem to be addressed and similarly the computational facilities that match analysis and modeling needs.

As AI/ML based methods find broader acceptance and applications, the need for accessing large amounts of experimental and/or simulated data sets for training will become more important. This shift will also require 'programmatic' ways to search and retrieve large amounts of data and make them available for training which is very different from the more conventional data portal concept (web interface for user to search and download smaller amounts of data sets).

5.15.2.4 Process of Science

Magnetic quantum materials are examples where data science approaches and close integration of advanced modeling and experiment are allowing advances to be made on problems that previously were out of scope. Particularly, deeper understanding of the physical states of quantum materials and prediction and explanation of their behavior is being achieved, by identifying their Hamiltonian—the master equation linking theory to experiment—and making reliable predictions of experimentally accessible quantities.

$Dy_2Ti_2O_7$ — spin ice — is a material that has presented major challenges regarding understanding the underlying interactions in the material and its mysterious properties at low temperature for over two decades[5]. Demanding simulations are required to model the material and at the same time it is necessary to combine multiple data sets including diffuse neutron scattering, heat capacity, and susceptibility including at multiple magnetic fields for analysis. ML in the form of trained nonlinear autoencoders and generative models have been demonstrated to provide the ability to extract new physical understanding as well as quantifying accurate Hamiltonians. They have provided automated capabilities for complex data handling tasks such as removal of backgrounds from diffuse scattering data and extraction of the interactions[6] yielding new insight into the material

---

[5] S.T. Bramwell, M.J, Harris, The history of spin ice, J. Phys. Condens. Matter 32 (32) 37410 (2020)
[6] Samarakoon A.M., Barros K., Li Y.W., Eisenbach M., Zhang Q., Ye F., Dun Z.L., Zhou H., Grigera S.A., Batista C.D., Tennant D.A., Machine Learning Assisted Insight to Spin Ice Dy2Ti2O7, Nature Communications 11 892 (2020); https://doi.org/10.1038/s41467-020-14660-y.

and its out-of-equilibrium behavior at low temperatures[7]. Additionally, the use of generative models has facilitated the design of experiments, rapid exploration of complex phase diagrams and real-time analysis [4]. This work has culminated in the discovery of unforeseen new physical principles which otherwise were unlikely to be detected[8].

The material RuCl3 is a candidate for applications in topological quantum computing. It has presented extraordinary difficulties in determining its physical state and interactions. While inelastic scattering is the best experimental technique to determine this, the modeling involved is very demanding. HPC simulations are required and again ML approaches have been necessary to analyze the data [Samarakoon22_I]. Here co-analysis with light scattering, thermodynamic properties, and measurements at multiple temperatures and magnetic fields are needed which is not currently possible which means that while interactions are determined they are not fully validated.

Figure 5.15.1 shows the ML-based workflow suitable for neutron scattering that was used in both studies above. Multiple steps are performed manually as not all parts of the data pipeline are currently automated. A crucial bottleneck is that the codes for training the neural networks for Dy2Ti2O7 and RuCl3 produced around 40Tb of data, which requires HPC resources to compute with experimental data sets of order 200Gb. It is not possible currently to move the simulations quickly to different locations for retraining of the ML nor can additional runs be scheduled to calculate new simulations based on findings from experiments during those experiments. Further, ML can be directly run on the data, but this again means data streams transporting large amounts of live data to remote locations. Compounding this our experience shows that using all available data on materials including from light sources, other neutron sources, and lab-based measurements is important for analysis. This means accessing and moving around this other data as well as collaborative use of orthogonal simulation codes for these techniques again on remote computational facilities would also require integration.
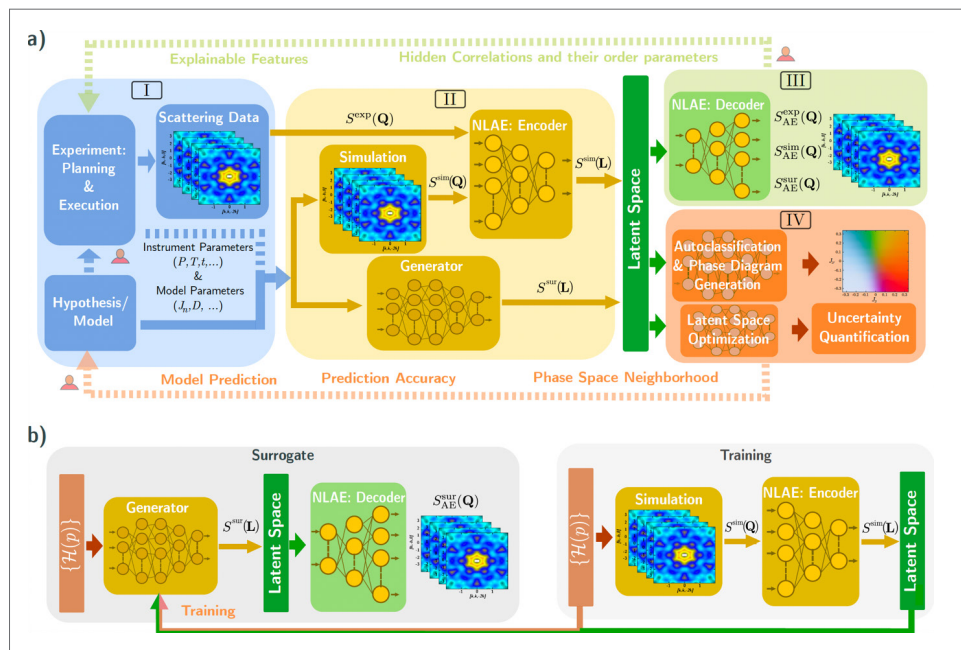
Figure 5.15.1: Schematic overview of machine-learning integration into the direct and inverse scattering problem (a). The ML workflow used here to drive the scattering experiment with automated data analysis and feeding back vital information.

[7] A. M. Samarakoon, B. Klemke, R. A. Borzi, R. Feyerherm, F. Ye, Q. Zhang, Z. Dun, H. Zhou, T. Egami, C. Castelnovo, L. Jaubert, R. Moessner, S.A. Grigera, D. A. Tennant, Structural spin glassiness in spin ice Dy2Ti2O7, Phys Rev Research, under review (2021); ArXiv: 2107.12305; and Anomalous magnetic noise in imperfectly flat landscape in the topological magnet Dy2Ti2O7, PNAS 119 (5) e2117453119 (2022); https://doi.org/10.1073/pnas.2117453119

[8] J. Hallen, C. Castelnovo, S.A. Grigera, D.A. Tennant, R. Moessner, Universal noise from monopole motion on dynamical fractals. Science accepted (2022)

The workflow is split into four main sections: (I) scattering experiment design and optimization; (II) parameter space exploration and information compression; (III) structure or property predictions; and (IV) parameter space predictions. Section II links to both III and IV via latent space {L}, a compressed version of the large pixel space. Dashed lines with a silhouette indicate parts of the flow that currently still require some human intervention. The latent space representations, S(L), are used in surrogates that bypass expensive calculations. (b) Schematic design of the surrogate model used to predict S(L) and S(Q) for a model with a given set of parameters, H(p). It comprises a radial basis network, mapping parameter space to latent space and a decoder to reconstruct S(Q) from latent space representations. The training of the surrogate is done based on a set of S(L) obtained from a set of models at different parameters, H(p), using Monte Carlo simulations and Nonlinear Auto-Encoder encoding. These surrogates are used for exhaustive searches of parameter space, identifying phases and phase transitions, and predicting optimal regions for experimental study. More simulations are done iteratively in the areas of interest, and the surrogates are trained accordingly to improve their prediction accuracy.

New Developments

Other developments that are pushing forward the need to achieve large-scale analysis and data transfer capabilities are the development of AI accelerated codes. By using ML massive atomistic simulations and complex quantum simulations that previously were too expensive or slow to impact experiments are now becoming available. That means that the ESnet capabilities to allow these collaborations and data transfers need to be put in place. In addition to this, there is increasing evidence that the integration of ML into measurements results in significant improvements in speed and outcomes.

Science Workflow
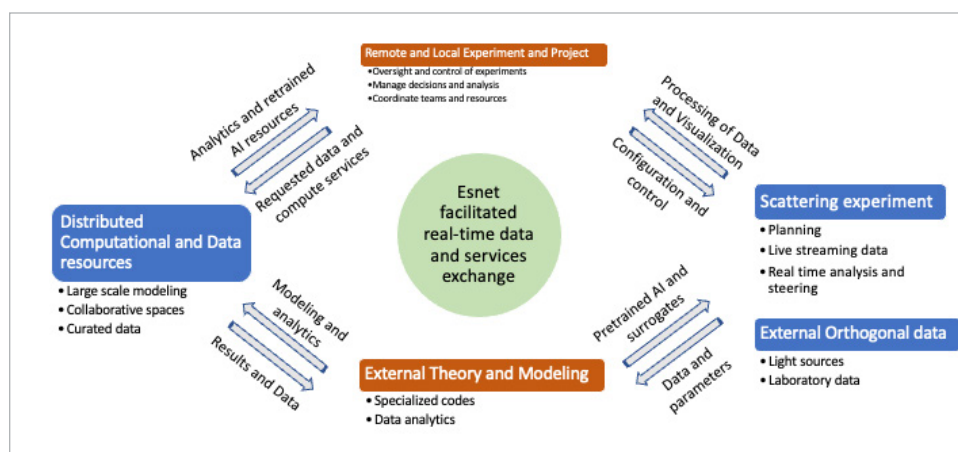
Figure 5.15.2: Scientific workflow involved: The wider networking and tasks need to realize the kinds of AI integrated experiment and analysis workflows illustrated in Figure 5.15.1.

Requirements are teams of theorists and experimentalists to have the ability to share data, simulations, and data analytics capabilities in real time to effectively conduct experiments and maximize the scientific value of multifacility co-analysis. Realizing this capability requires ESnet fast data transfers and archiving capabilities between supercomputing and edge-computing resources. Organizational tools and data centers that provide specialized theory and simulations to train neural networks as surrogates and the ability to mine and extend simulations and training data sets are also required. Finally, intensive data analytics on experiment data during a beamtime to identify features and aid experimental decision-making can involve sending data to distant institutions for rapid analysis.

5.15.2.5 Remote Science Activities

Experimental and simulation data, computer resources and collaborators are located in different locations requiring remote use. Partially accelerated by COVID, more facilities now offer remote experiments. In the context of cross-facility workflows, the question is 'what is remote'?

5.15.2.6 Software Requirements

Recent developments in using AI and ML based techniques require vast amounts of training data. These can be simulation data or experimental data often in some reduced form. Depending on the computational cost of simulations, simulation data will be shared as well. While the typical use in the past centered around finding and downloading a few data sets for analysis, use of ML will require 'API' type capabilities to find, aggregate and use/move large amounts of data for training of neural networks.

In addition, trained networks will be published and should be made available for reuse in the future. While not solely a software issue, the need for metadata related to trained networks and tools to use them including e.g., for transfer learning and such are new challenges that need to be addressed.

5.15.2.7 Additional Network and Data Architecture

Key requirements for simulations and training for neutron experiments (spectroscopy, reflectometry, and diffuse scattering), are in the 10s to 100s of TB ranges. Pre-trained AI such as generative models will require to be extended in real time during experiments. This involves new simulations and training which will be at different remote locations. Meanwhile the experiment data sets from neutron instruments are in the tens of GB to TB range and will need to be streamed for specialized AI to undertake analytics to recognize patterns and identify features. The nature of specialized computational resources and codes mean that simulation and data tasks will be distributed over different locations. ESnet provides the network infrastructure to bridge this gap. In addition, co-analysis with data sets from other facilities, especially microscopy, synchrotron, and XFEL data is needed.

5.15.2.8 Use of Cloud Services

It is expected that cloud services related to the workflow, data archiving and ML might meet the needs outlined here. However, this will need to be determined in detail as work related to this case study progresses.

5.15.2.9 Data-Related Resource Constraints

Workflows relying on ML will require rethinking searching, access and retrieval of data for training as well as data management related to the derived networks and products.

5.15.2.10 Outstanding Issues

Currently realizing any workflow discussed here requires a one-off staging of data, modifying of scripts and represents more a demo or prototype than a scientific workflow solution across multiple experimental and computational facilities as well as local universities and institutions. Traditionally we think of scientists downloading one or few data sets through a portal. However, as discussed here a cross-facility workflow employing ML methods requires a different approach to data search and retrieval e.g., though a generic API as well as expanded sets of metadata facilitating easy assembly of training data.

As applications of ML grow, there will be an increasing number of trained networks that require storage, access and the needed metadata to allow reuse.

5.15.2.11 Case Study Contributors

Multifacility Experimentation and Analysis Workflows: Neutron Scattering Perspective Representation

- **Alan Tennant**, *University of Tennessee*, dtennant@utk.edu
- **Steve Johnston**, *University of Tennessee*, sjohn145@utk.edu
- **Cristian Batista**, *University of Tennessee*, cbatist2@utk.edu
- **Thomas Proffen**, *ORNL*, tproffen@ornl.gov
- **Shelly Ren**, *ORNL*, rens@ornl.gov
- **Steve Hartman**, *ORNL*, hartmansm@ornl.gov
- **Garrett Granroth**, *ORNL*, granrothge@ornl.gov
- **Willem Blokland**, *ORNL*, blokland@ornl.gov
- **Kipton Barros**, *LANL*, kbarros@lanl.gov
- **Yingwai Li**, *LANL*, yingwaili@lanl.gov
- **Anjana Samarakoon**, *ANL*, asamarakoon@anl.gov

5.16 Multifacility Experimentation and Analysis Workflows: NSRC Perspective

The multifacility experimentation and analysis workflows, focused on NSRCs, brings together BES community representatives from the NSRCs to discuss this emerging area of scientific investigation. The case study profiles the overall concept, some early results, and a discussion about what the future will hold.

5.16.1 Discussion Summary

- 4D cameras can operate at 350 Gigabits per second, and can produce 5 PBs of data per year. These types of data rates will only increase in electron microscopy due to the increased proliferation and continued development of high-speed electron detectors.

- New detectors based on CMOS designs are currently being installed at multiple national laboratories.

- The 4DCamera Distillery is a DOE-funded project across all NSRCs to develop and deploy methods and tools based on AI/ML to analyze electron scattering information. The expected increases to data velocities and volumes present significant challenges for moving, storing, and processing data.

- Most of the data from microscopy collaboration is first reduced before it is shared over cloud storage. Globus is often used to transfer files to local storage to run analysis on HPC systems.

- The size of data sets for frontier electron microscopy experiments (tomography, 4D, high frame rate) makes it infeasible for users to download the full data set, and impractical to browse the data set looking for information of interest. This points to the need for remote interfaces for data browsing and analytics.

- Scientific characterization tools have been pivotal to increasing our understanding of nanoscience. Substantial bottlenecks exist between the data streams emanating from these tools, and the feedback from theoretical/simulation insights. Directly coupling the microscope data to simulations in real time to assist the experimenter is of crucial importance in the quest towards automation and autonomous materials/physics discovery platforms.

- Significant computational needs beyond edge compute will be required in the future, including GPUs.

- Taking advantage of midcapacity computing, and DOE HPC capabilities, can generate 0.5-5 GBs of data per calculation that can quickly add up to few TBs of data. A dedicated pipeline with fast data-transfer rates connecting edge computers that are close to experiments, with computational infrastructure will drive simulations "on the fly" to both guide experimental investigation, as well as accelerate materials as well as fundamental physics discoveries.

- Networking is critical to achieve autonomous and "smart" characterization tools. Data captured during synthesis using characterization tools can be processed to update experimental conditions. As these approaches are computational expensive, data needs to be transferred from the instrument to computation, and back, to provide instructions during autonomous data capture.

- Simulations that can be used to guide and explain the observations encountered in the experiment are critical, and require fast processing and networks access to be fully realized.

- In many instances, data reduction is achieved at the source during the experiment. The reduced data is usually stored on local drives and given to the user through USB keys or other media. Experimental and simulation data is not usually shared immediately, but this paradigm is rapidly changing.

- Current remote science activities are primitive, but adapting. Remote engagement among scientific collaborators is necessary, and future facilities will need to support remote progress. An outstanding challenge is to provide a remote instrument control experience that is commensurate with that one can obtain while local to the instrument.

- Remote data access has seen a much broader uptake of cloud storage solutions at the facilities, both institute-provided local solutions, as well as the use of commercial cloud vendors. Future network solutions would ideally integrate seamlessly and provide high-speed connections to typical storage systems.

- Future lab instruments should expose a standard API, so that users could devise integrated experimental workflows.

- The software being used across facilities is currently highly heterogeneous. There is not currently a single, integrated plan for the software systems that will be deployed.

- NSRCs host commercial tools from a multitude of vendors. This makes it inevitable to interact with a large number of different software systems, many of which use proprietary data formats. It remains a significant challenge to integrate these tools with automated workflows.

- Data rates for next-generation electron microscopes exceed typical local infrastructure capabilities. As microscope data rates will increase yet further in future iterations, it is necessary for future network and storage infrastructure to be up to the task of handling these data rates.

- Use of cloud storage solutions has increased, but is still done in an ad hoc manner. In the long term, DOE as a complex will need to answer questions about the long-term archival storage of facility data. If facilities are to be responsible for long-term storage of user data, then there will need to be consideration and commitments regarding facility mission, funding for storage, and network infrastructure for access.

- The size of electron microscopy data sets challenges current computational methods. Most analysis pipelines are insufficient for real-time analysis and impractical for post facto analysis. Significant improvements in computational workflows must be developed and integrated.

- Cross-facility integrated data sets face a substantial challenge with respect to acquiring and preserving suitable metadata. Improved software tools for acquiring, transferring, and browsing/searching through metadata are necessary.

5.16.2 Multifacility Experimentation and Analysis Workflows: NSRC Perspective Case Study

Advancements in NSRC experimental technology has led to an increase in the volume and quantity of scientific data produced, and raised serious issues regarding the approach to computing and storage in the long term. This case study will review examples from the NSRCs that involve multifacility workflows, the relative success in their execution, and plans for the future

5.16.2.1 Science Background

Direct Electron Detectors based on CMOS Image Sensor processes have been transformational in electron scattering and microscopy – culminating in the 2017 Nobel Prize for cryo-electron microscopy. Concurrent with the transformational science that direct electron detectors enable is the rich and large amount of data that they produce. At the leading edge of this trend, the newly installed 4D Camera at LBNL operates at 350 Gigabits per second and is estimated to produce about 5 PBs of data per year. The origin of this extreme data rate is that the 4D Camera was designed to collect as many scattered electrons as possible from a typical electron microscope experiment, a feature critical for understanding the structure of beam sensitive materials, nanomaterials, and a wide range of energy-relevant technologies. As the 4D Camera demonstrates, these types of data rates will only

increase in electron microscopy due to the increased proliferation and continued development of high-speed electron detectors aimed at high-speed imaging and high-resolution analytical measurements.
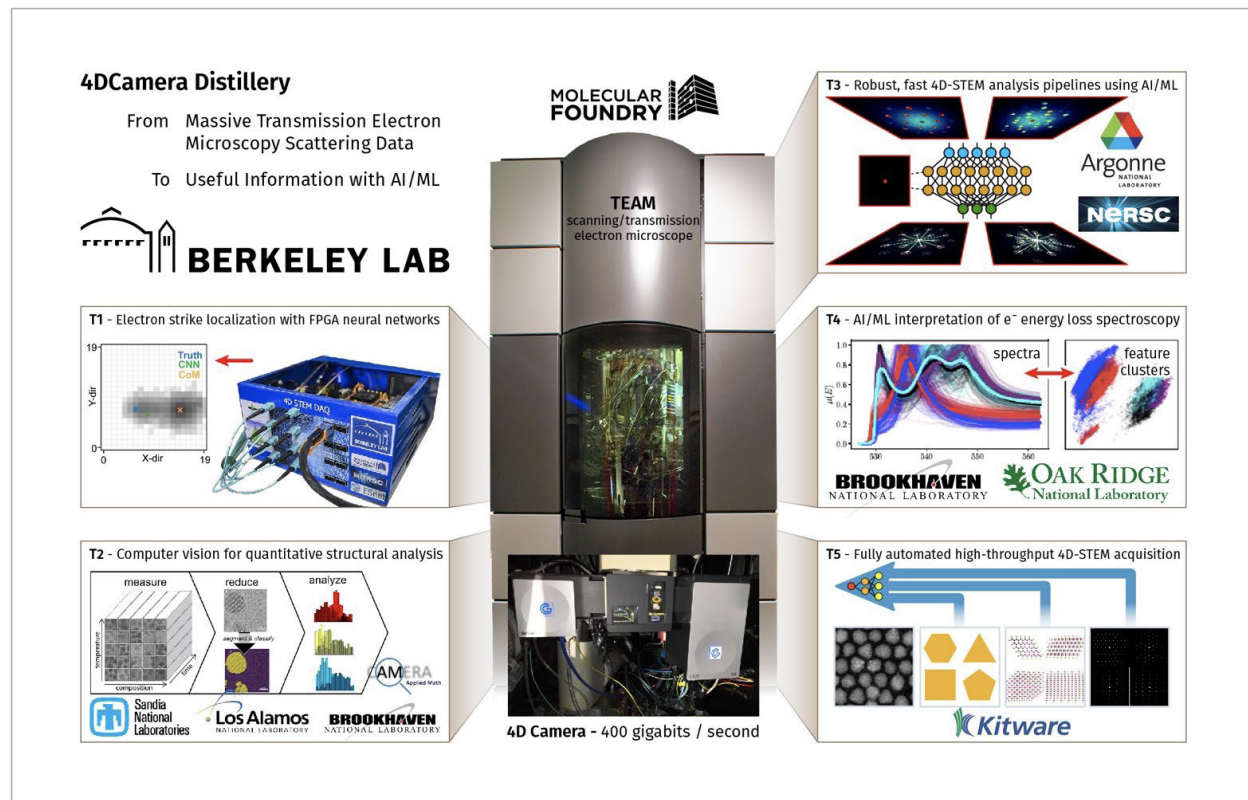
Figure 5.16.1: LBNL 4DCamera Design

For example, new detectors based on CMOS designs are currently being installed at multiple national laboratories. For instance, BNL is installing a 4D Camera and at LBNL there is a new detector project to design, fabricate and implement a novel, very high-speed ($\sim$MHz) electron spectroscopy detector for electron energy loss spectroscopy. The data rate will be similar to the 4D Camera, (350Gbitps). As direct electron detectors are upgradable options for existing electron microscopes, their proliferation is possible at all electron microscopy research centers.

The 4DCamera Distillery is a recently funded project across all DOE NSRCs to develop and deploy methods and tools based on AI/ML to analyze electron scattering information from the data streams of fast direct electron detectors. The Distillery program consists of five thrusts that address both the critical need for data

reduction tools for these detectors, as well as scientific opportunities to create new modes of measurement and experimentation in the electron microscope enabled by fast electron detection. Direct electron detectors such as the 4D Camera recently deployed at TMF at LBNL produce massive volumes of data (up to 3 TB/min) at incredibly high sampling rates (87 kHz). These data velocities and volumes present significant challenges for moving, storing, and processing data.

5.16.2.2 Collaborators

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
BNL (EXAMPLE OF PRIMARY SITE)	primary	Data transfer (Globus, cloud services, ad hoc)	10-100 TB	daily	N	Data rates challenge current systems.
USERS	secondary	various (Globus, cloud services, ad hoc)	10 TB	Ad hoc	N	Size of data sets makes it hard to organize and browse. Full download of data sets impractical and inefficient.

Table 5.16.1: Multifacility NSRC Collaborations

Collaborations for microscopy projects are typically those where the user/collaborator provides the sample to be investigated. Alternatively, the samples could be synthesized in house, and some characterization that is not available locally may be performed externally at other facilities (for example, at a synchrotron for quantitative measurements of composition and structure of oxide films). Most of the data from such collaborations is first reduced before it is shared over cloud storage such as Dropbox, although in the case of large simulation files, Globus is often used to transfer files to local storage, if for example this is required to run analysis on HPC systems.

The size of data sets for frontier electron microscopy experiments (tomography, 4D, high frame rate) makes it infeasible for users to download the full/raw data set, and impractical to browse the data set looking for information of interest. This points to the need for remote interfaces for data browsing and analytics.

5.16.2.3 Use of Instruments and Facilities

Consider that scientific characterization tools, and in particular forms of scanning probe and electron microscopy, have been pivotal to increasing our understanding of nanoscience by enabling functional properties to be correlated with microstructural and atomic features of samples. Despite the proliferation of such tools, substantial bottlenecks exist in terms of the analysis of the data streams emanating from these tools, and the (much longer time) feedback from theoretical/simulation insights that can provide knowledge about the physical mechanisms underpinning the observed relationships and phenomena. Directly coupling the microscope data to simulations in real time to assist the experimenter is of crucial importance in the quest towards automation and autonomous materials/physics discovery platforms. A simple example of such a workflow is shown in Figure 5.16.2.
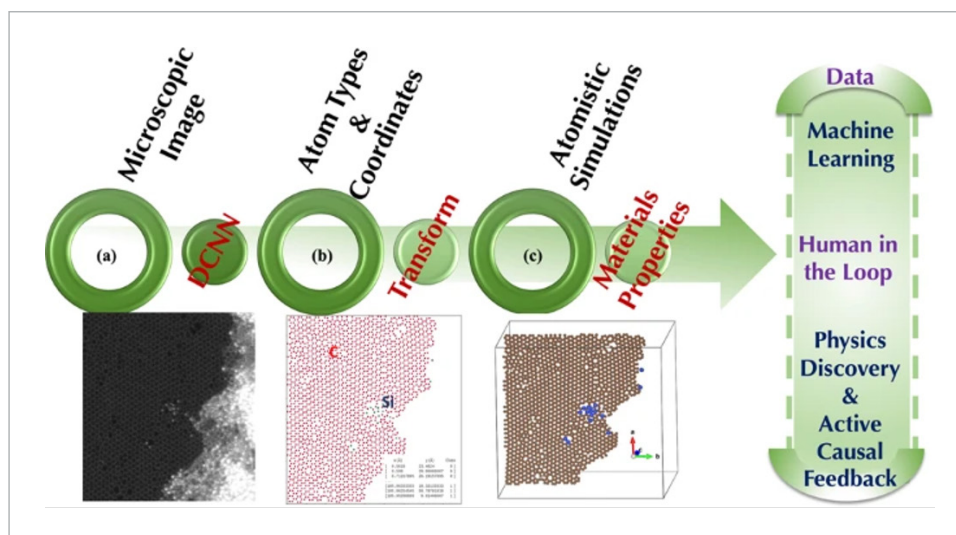
Figure 5.16.2: From Ghosh et al., npj Comp. Mater. 8, 74 (2022). (a) Deep learning models take input images and convert them into atomic coordinates. These are then used as input for simulations, which can then be run to provide insight into the physics and relevant chemistry, energetics of the different structures observed and also provide guidance towards feedback to manipulate the structures observed towards predefined targets.

The data rates on traditional STEM are lower (~20 MBps), but the same can be applicable to 4D-STEM which can provide data rates of >100 Gbitps and which will need reduction at the edge before further processing can be accomplished. At the same time, significant computational needs beyond edge compute will be required, including GPUs for tasks such as model training (step a in Fig. 2), and to run molecular dynamics codes, or cluster compute for first principles approaches such as density functional theory. These calculations can range from 1 to 10s of nodes, taking advantage of both midcapacity computing as well as DOEs HPC capabilities, including the hybrid CPU/GPU architecture available in the latter. Such calculations generate 0.5-5 GBs of data per calculation that can quickly add up to few TBs of data with concurrent embarrassingly parallel high-throughput screening approaches. A dedicated pipeline with fast data-transfer rates connecting edge computers that are close to experiments, with such midcapacity and HPC infrastructure in national labs. has the much needed, yet untapped potential, to drive specific simulations "on the fly" to both guide experimental investigation, as well as accelerate materials as well as fundamental physics discoveries.

The data from microscopy experiments is typically stored immediately on a NAS device, and then transferred to dedicated storage for longer terms (typically about five years).

5.16.2.4 Process of Science

Networking is critical to achieve autonomous and "smart" characterization tools, and for physics discovery, as shown in Figure 5.16.2. Essentially, data captured at synthesis and characterization tools can be processed and the experimental conditions updated, via suitable algorithms such as Bayesian optimization. However, such algorithms are compute hungry and therefore, data needs to be transferred from the instrument to the GPU cluster for computation, and then fed back to the instrument as directions for adjusting the instrument parameters to achieve optimal sampling and autonomous data capture.

Simulations that can guide and explain the observations encountered in the experiment are also critical. At this point, most simulations are performed either before or after the experiment is completed, and little to no feedback is available, either to experiment or theory.

With respect to data analysis and reduction, in many instances, data reduction is achieved at the source during the experiment, i.e., with simple techniques such as signal averaging. This (reduced) data is usually stored on

local drives and then given to the user, through USB keys or other media. Experimental and simulation data is not shared immediately and usually there is a considerable time lag; we do envision (as in Figure 5.16.2) that this paradigm is rapidly changing. Although standardized file structures are available for much of the simulation data, this is not universal, and the problem is exacerbated on the experimental side where every vendor chooses their own (often proprietary) standard. This makes data sharing more difficult, leading to the need to write 'converters' to a common file format, such as hdf5. Much of this work has been completed in the past few years. The future will revolve around (a) automating the pipelines for translation of the file to the standard selected, automating data processing based on the files generated, automating the theory-experiment matching, and then presenting visualization to the user so they can better understand the differences between the experiment and the simulated results, for immediate feedback to both domains. At this point, most of the data resides in separate storage boxes that are unlinked. In five years, it is expected this will no longer be the case.

5.16.2.5 Remote Science Activities

Currently, remote science activities are relatively primitive; however, it is already clear that this is an area where facilities will see significant growth over the coming years. A transition has been spurred in large part by the COVID-19 pandemic, where remote engagement among scientific collaborators became necessary. In the wake of this event, it is clear that future facilities will need to support remote progress of various kinds.

Computing work has a long history of leveraging remote user access, and this mode of operation will undoubtedly continue. Here, one can expect an increasing desire for remote interfaces to simplify launching and managing of jobs, as well as remote data analysis platforms suitable for nonexperts. As always, a reliable and robust network infrastructure will be necessary for HPC users to make best use of facilities.

On the experimental side, we are beginning to see uptake of remote science activities, including remote engagement (video conferencing between scientists), remote instrument control, and remote data access. All of these require increasingly robust and low-latency network connections. For NSRC facilities, which host a heterogeneity of instruments from a variety of vendors, the remote control method may often be relatively inefficient (e.g., remote desktop), which correspondingly requires an overprovisioned network infrastructure, with respect to data rates and latency. An outstanding challenge is to provide a remote instrument control experience that is commensurate with that one can obtain while local to the instrument. E.g., video-rate data does not transfer smoothly over typical remote desktop interfaces. Video-streaming solutions may be required. Remote data access has seen a much broader uptake of cloud storage solutions at the facilities, both institute-provided local solutions, as well as the use of commercial cloud vendors. Future network solutions would ideally integrate seamlessly and provide high-speed connections to typical storage systems.

Looking beyond the short term, it is possible to imagine multifacility experiments that leverage instruments across multiple national laboratories. Providing a smooth user experience in such cases would require fast and efficient remote systems, as previously mentioned, but also integrated with each other. Ideally, the various lab instruments would each expose a standard programmatic control interface (API), so that users could devise integrated experimental workflows. The behind-the-scenes network infrastructure would need to correspondingly efficiently transfer and aggregate data sets, presenting them to the user in a coherent way.

5.16.2.6 Software Requirements

The software being used across facilities is currently highly heterogeneous. There is not currently a single, integrated plan for the software systems that will be deployed.

Of note is that NSRCs host commercial tools from a multitude of vendors. This makes it inevitable to interact with a large number of different software systems, many of which use proprietary data formats. It remains a significant challenge to integrate these tools with automated workflows, requiring cumbersome workarounds both for control and for data access.

5.16.2.7 Additional Network and Data Architecture

Note aforementioned data rates for next-generation electron microscopes (350 GBps) exceed typical local infrastructure capabilities. Using such systems requires bespoke solutions, including local SSD arrays as cache to capture data, and high-speed (fiber-optic) links to institutional data storage systems. As microscope data rates will increase yet further in future iterations, it is necessary for future network and storage infrastructure to be up to the task of handling these data rates.

5.16.2.8 Use of Cloud Services

Use of cloud storage solutions has increased dramatically at the NSRCs. However, it is currently mostly done in an ad hoc manner, where individual scientists manage data for the tools under their purview.

In the near term, we anticipate increasing use of institutional and commercial cloud storage solutions. This will see increased data traffic to cloud storage, and increasing need for automated computational workflows that can ingest data from cloud storage solutions.

In the long term, DoE as a complex will need to answer questions about the long-term archival storage of facility data. If facilities are to be responsible for long-term storage of user data, then there will need to be consideration and commitments regarding facility mission, funding for storage (whether institutional or commercial), and network infrastructure for access.

5.16.2.9 Data-Related Resource Constraints

The size of electron microscopy data sets challenges current computational methods. For collection of megapixel images at kilohertz frame rates, most analysis pipelines are insufficient for real-time analysis and impractical even for post facto analysis. Significant improvements in computational workflows must be developed and integrated. These may arise through improved algorithms, machine-learning methods, hardware acceleration (e.g., GPUs, FPGAs, etc.), or improved use of HPC. It is likely that all of these approaches will see broader use (in different contexts).

Cross-facility integrated data sets face a substantial challenge with respect to acquiring and preserving suitable metadata. Data aggregation and automated analysis can only occur when metadata and provenance are saved correctly. Currently this is not done on the majority of tools. Improved software tools for acquiring, transferring, and browsing/searching through metadata are necessary.

5.16.2.10 Outstanding Issues

None to report at this time.

5.16.2.11 Case Study Contributors

Multifacility Experimentation and Analysis Workflows: NSRC Perspective Representation

- **Kevin Yager,** *BNL,* kyager@bnl.gov
- **Yugang Zhang,** *BNL,* yuzhang@bnl.gov
- **Ryan Wixom,** *Sandia,* rwixom@sandia.gov
- **Jianxin Zhu,** *LANL,* jxzhu@lanl.gov
- **Subramanian Sankaranarayanan,** *ANL,* ssankaranarayanan@anl.gov
- **Maria Chan,** *ANL,* mchan@anl.gov
- **Bobby Sumpter,** *ORNL,* sumpterbg@ornl.gov
- **Panchapakesan Ganesh,** *ORNL,* ganeshp@ornl.gov
- **Edward Barnard,** *LBNL,* ESBarnard@lbl.gov

5.17 Use of the ESnet for Quantum Simulations of Materials and Molecules

Quantum simulations of materials and molecules is an emerging area of research with unique data needs. This case study, featuring a cross-section of BES community members, will outline some of the current state, and future goals of this work.

5.17.1 Discussion Summary

- Quantum simulations of materials and molecules use a variety of platforms, encompassing midsize compute cluster and storage systems owned by universities, and high-performance architectures at national facilities.

- FPMD simulations generate data sets of tens to hundreds of terabytes.

- Generated data may not reside on the HPC facilities where they have been created for future access, curation and reuse, which necessitates data transfer between collaborators.

- Globus is used through ESnet for the transfer of data sets at different stages of investigation.

- We expect this hybrid model to persist in the era of quantum computers where part of the calculations will be carried out on classical architectures and part on quantum hardware, with the need to transfer between the two platforms.

- A future hybrid computing strategy will require tighter integration of classical and quantum resources. In this case, a quantum processor becomes a co-processor on which certain tasks are offloaded, but will require real-time data movement. The data transfer would be small, but real-time transfer will be critical between resources.

- ML surrogates will require highly expensive quantum simulations. Training these models will require training data collected through massively distributed simulations on leadership-class supercomputing facilities. The data set sizes, and bandwidth requirements, will require efficient movement of GB to TB of data.

- Simulations on small clusters generate initial structure information for larger simulations. Data is order 10 MB per file, with tens of thousands of files.

- Simulations at HPC facilities can range on the order from 1 GB to 10 TBs, with the number of files and directories can range from 1 to tens of thousands, with total data set size around tens of TBs.

- Once exascale computers become seriously available to more users, the size of all of simulation data will likely be multiplied by 100- to 1,000-fold.

- A number of quantum simulation users and collaborations currently use multiple facilities and locations to accomplish science goals that cross the DOE, commercial, and university complex. Data sharing is primarily challenged by data formats and searchability, and for certain classes of simulations the size of the data to be transferred remains a significant issue to solve.

- Future usage of computing will require interaction between multimodal experimental and multifidelity theoretical approaches. The urgent need is for on-demand computing, interoperable smart workflows across various computing platforms, along with long-term cloud storage solutions.

- Most of the data storage and transfer speeds are appropriate for quantum simulation, and this is unlikely to change within the next several years.

- Data accessibility, searchability, and usability is a key issue for quantum simulation, and improvements to enable high throughput scientific workflows will be required.

- There are significant needs for software that performs well on heterogeneous computers. In addition, cloud computing could be an area of need in the future in terms of resources and interoperability.

- There is a need for experienced data/computing personnel for streamlining of workflows, data management and custodians that are flexible to respond to specific needs of different projects/applications.

5.17.2 Use of the ESnet for Quantum Simulations of Materials and Molecules Case Study

Quantum simulations of materials and molecules use a variety of platforms, encompassing midsize compute cluster and storage systems owned by universities, and high-performance architectures at national facilities. Most problems require the transfer of data between university clusters or national laboratory resources and HPC facilities, both at the time of investigation and at the completion of the project, when storage of data is planned (often data may not reside on the HPC facilities where they have been created for future access, curation and reuse).

5.17.2.1 Science Background

The study of physical and chemical problems using atomistic and quantum simulations typically includes the determination of structural properties of complex systems, for example by carrying out FPMD, possibly coupled with advanced sampling techniques, and the study of electronic and transport properties, for example by performing advanced calculations with methods beyond density functional theory. FPMD simulations and the complex workflows surrounding molecular and materials simulations generate data sets of tens to hundreds of terabytes, depending on the system and length of the simulations[1].

In most studies, simulations are performed on different platforms: for example, on midsize clusters to generate starting configurations and HPC facilities for production runs, with the need to move large data sets back and forth between facilities. Another example pertains to running ensemble-based simulations on HPC facilities and sampling that data to perform property-related computations on midsize clusters, as part of a complex workflow. Most often Globus is used through ESnet[2] for the transfer of data sets at different stages of the investigation, not only for MD trajectories but also for electronic structure data.

In essence any large-scale simulation of materials and molecules is always hybrid, with components on midsize compute clusters and components on HPC facilities, and hence with the need of transferring data. Some analysis can be carried out on the fly, but not all analyses are amenable to immediate processing. In fact, data should be made available for possible reprocessing or new processing when an investigation is completed. Data curation is also expected to be hybrid. Even in the case when data is made completely available and accessible on a paper-paper basis (e.g., through Qresp[3]), it may not be further analyzed on the platform where it is made available, but rather necessitates transfer for additional processing.

We expect this hybrid model to persist in the era of quantum computers where part of the calculations will be carried out on classical architectures and part on quantum hardware, with the need to transfer between the two platforms, at least in the era of near-term noisy quantum computers. Below are specific use cases that emphasize the multifaceted science challenges and requirements that are faced in quantum simulations of materials and chemicals.

---

[1] http://www.quantum-simulation.org/reference/index.htm
[2] https://fasterdata.es.net/data-transfer-tools/globus/
[3] https://paperstack.uchicago.edu/explorer

Simulation of Quantum Algorithms for Materials Properties

An important component of the case study, particularly as quantum computers become better, is the use of them to provide key information – electronic energies, gradients, etc. for, e.g., FPMD simulations. Efficient quantum algorithms need to be developed and tested, and the benchmarking of the performance of these algorithms – including the Variational Quantum Algorithm (VQE) and various forms of Quantum Phase Estimation (QPE) – is crucial. The simulations can require large amounts of memory (e.g., on the order of 2N Bytes, where N is the number of qubits (spin-orbitals in electronic structure simulations, ideally > 30 for moderate-sized molecules, making the problem memory and computationally challenging). Such calculations will nonetheless suggest the path forward for actual hybrid classical/quantum computer simulations. Approaches to accelerate large-scale simulations, including databases for initial VQE starting points will also be pursued. A different strategy that needs to be considered is a much tighter integration of classical and quantum resources. A good example is the recent Google work on quantum Monte Carlo sampling. Here a quantum processor becomes a co-processor on which certain tasks are offloaded, but will require real-time data movement. Similar approaches are being explored in quantum chemistry. While the data transfer in such a case is small, real-time transfer is critical.

ML Surrogates to Accelerate First Principles Simulation

A next-generation predictive capability for computational materials design will require highly expensive quantum simulations that include the established suite of simulation methods such as density functional theory, coupled cluster, and multireference frame techniques. The computational cost of these simulation techniques increases very rapidly with the accuracy required, and applications to materials design are frequently limited by computational resources. A promising approach is the development of ML models that act as low-cost, high-fidelity surrogates. Training these models will require a collection of diverse and high-quality training data. Such training data can be collected through massively distributed simulations on leadership-class supercomputing facilities, tied together by an active learning procedure that guides exploration in chemical space. As the resulting simulation data may be broadly useful to a variety of ML research efforts, it will be useful to be able to quickly distribute the simulation data to researchers throughout the DOE complex. We anticipate that the training of future ML surrogate models will incorporate more information about the electronic wave function, and excited-state properties, which increases the data set sizes, and bandwidth requirements. High-quality data sets of 102-104 gigabytes will become the norm, and we anticipate a need to rapidly transfer these data sets throughout the DOE complex.

Computational Simulations of Emergent Phenomena in Strongly Correlated Quantum Materials

Emergent phenomena including unconventional superconductivity and quantum spin liquids, in strongly correlated quantum materials have a profound impact in the next-generation devices for energy and quantum applications. On the one hand, these phenomena originate from the competing interactions and couplings among multiple degrees of freedom. Therefore, these phenomena defy the descriptions by the current density functional theory, which has been quite successful for many systems without strong correlations (e.g., normal metals, semiconductors). On the other hand, these phenomena are cooperative in the thermodynamic limit and are absent in the atomic scale. There are three types of computational finite-size approaches on strongly correlated systems described for example Hubbard model: exact diagonalization (ExactDia), quantum Monte Carlo (QMC), density matrix renormalization group (DMRG). ExactDia calculates exactly the ground and low-lying excited states of the model, while QMC and DMRG approach larger systems by discarding nearly irrelevant high-energy quantum states to save the memory space. For example, ExactDia scales exponentially with the number of sites whereas DMRG scales exponentially with system width. However, in contrast to the ExactDia, QMC and DMRG results are not always guaranteed to be reliable. For example, there are conflicting conclusions from QMC and DMRG as to the existence of unconventional superconductivity in the simplest two-dimensional Hubbard model. Therefore, a revival of exact calculations on a larger system is needed. This implies a physical memory of hundreds of terabytes, and very fast data exchange. Specifically, the ability to reach 40-60 sites for a two-dimensional repulsive Hubbard model (t-J model) will significantly improve the scaling analysis, which enables

an independent determination of the superconductivity instability. Additionally, larger simulations can be needed to adequately treat the complexity of real physical systems based on experimental observations, for example to study the atomic-scale inhomogeneity observed in high-temperature cuprates. We note that such high-quality calculations of these strongly correlated systems can serve as a reference for quantum computing algorithms on noisy intermediate-scale quantum hardware.

Structurally Complex Strongly Correlated Materials with Ab Initio Accuracy

Multifunctional correlated materials with competing degrees of freedom (spin, charge, topology, etc.) have applications ranging from energy-efficient computing to novel qubit materials. While key milestones in theory have identifying structure-property relationships in pristine systems (e.g., strain tuning of multiferroicity, pressure-driven superconductivity), understanding their emergent properties in the presence of structural distortions, atomic disorder and reconstruction, and in amorphous materials remains a challenge. Computational complexity appears at several stages in the process ranging from the large unit cells required to the inclusion of correlations at an appropriate level and requires multiscale physics approaches. For example, atomic reconstruction with twist in moiré heterostructures observed in electron microscopy experiments needs to be included for accurate modeling of the emergent physics of these systems requiring large unit cells. Amorphous materials can be simulated using ab initio molecular dynamics of multiple large quench-melt ensembles to represent bonding networks/coordination. These large-scale atomic configurations can generate 100 GB-TBs of data in themselves during calculation. Resulting atomic data is commonly shared with collaborators and/or refined (e.g. electron microscopy data, X-ray/neutron diffraction data) Further calculations including the construction of Wannier-based tight-binding models (e.g. to include correlations at level of dynamical mean-field theory (DMFT), calculation of topological surface states), and the use of methods for treating the resulting large matrices (e.g. Kernel Polynomial methods) to calculate electronic structure additionally generate large-scale data sets (GBs-TB) that need to be simultaneously manipulated and stored. A combination of HPC resources is needed for such workflows with efficient data transfer between these resources needed — e.g., large-scale structural optimizations parallelize on CPU/GPU nodes with large storage requirements and are ideal for systems like Perlmutter, while Wannier-based tight-binding calculations and DMFT still require large storage but with much reduced CPU requirements (e.g., local institutional clusters with ~10s of nodes).

High-Throughput QMC Based Many-Body Simulations of Heterogeneous Solids and Interfaces

Functional properties of materials in energy and quantum technologies (e.g., batteries, photovoltaics, ferroelectric-capacitors, semiconductors, quantum devices etc.) are largely determined by the intrinsic defects and solid-solid (or solid-liquid) interfaces present in them, that can in principle range from 10sÅ to 10-100nm. Many of the systems of technological interest, such as complex oxides or magnetic 2d interfaces, require accurate many-body approaches to describe their properties. While accurate ab initio electronic structure approaches, such as QMC, using QMCPACK (www.qmcpack.org), can in principle describe the ground-state properties exactly, they remain limited today to systems with linear dimensions of ~10-20Å. QMC scales as $\sim N^3$, where N is the total number of electrons in the problem, limiting the size of the system on current HPC architectures. A single typical QMC calculation even for such a system size will use O(100 GB) of RAM (per node), and run on 2k-4k nodes on DOE HPC machines (even on hybrid machines such as Summit/Theta/Perlmutter), for ~10 hrs and use ~100 MB – ~100 GB storage (depending on observables stored – e.g., wave function files and related quantities such as 1RDM can take more space than charge density). We envision the need for generating more regular QMC benchmark data sets to both predict ground-state (and some excited-state) properties, as well as to inform other approximate electronic structure methods (e.g., DFT functionals, GW approximation, model parameters in lattice models, force-field parameters etc.). While the above estimates are for a single determinant-based Slater-Jastrow wave function, with one fixed nodal structure, improvements to the wave function are possible by performing nodal optimization and/or adding additional determinants, which increases the cost (and storage) by a factor of ~100x or more. This provides a compelling reason to increase availability of compute infrastructure as well as data-storage infrastructure on HPC machines, with pipelines that offer fast data-transfer

from node-to-node as well as from HPC compute nodes (or high performance storage system (HPSS) archival systems) to dedicated HPC data-reduction/analysis machines, before it is moved to midcapacity computers in national labs/universities where visualization is performed to obtain scientific inference from the simulations. Efficient use of this infrastructure also requires workflows to prepare/submit/collate data, marching through the different simulation steps, that are interoperable across midcapacity, HPC compute/storage facilities, using data-transfer nodes when applicable. We envision these workflows to also be eventually integrated to AI/ML surrogate models or at least a GP process for smart choices of materials/parameters to explore for a targeted scientific outcome.

Development and Understanding of Complex Catalytic Environments

Catalysts are essential to the industrial complex by efficiently taking raw materials and feedstocks to commodity and specialty chemicals required for our continuing technological advancement. Complex catalytic environments must be simulated accurately to understand existing chemical phenomena and to develop new catalysts that are more efficient or that produce new chemicals and materials. These complex environments must include highly accurate and realistic pictures of the reactions that are occurring in solvents and on surfaces, especially in confined surfaces such as zeolites, mesoporous silicon nanoparticles, and metal organic frameworks. These simulations require multiscale and multiphysics approaches that incorporate many configurations of the system (ensembles) over time (such as in molecular dynamics). Additional insight and properties from these configurations are often obtained by additional simulations on subsets of the configurations. Choosing these configurations wisely and launching them to different computational resources is required — resulting in complex workflows where data must be transported between different computational facilities and researchers. These needs will only be extended as quantum simulations become the everyday workhorse for AI investigations of molecular simulations and chemical predictions.

Designing Novel Molecular Crystals for Direct Air Capture (CO2 Capture from Air)

The direct air capture of carbon dioxide is a potentially transformative technology for addressing climate change. A few select materials solutions are being investigated for industrial application, but the field is still very much in its infancy, and there is a substantial need for the discovery of new adsorbents exhibiting a greater separation capacity, long-term adsorption/desorption cycling stability, and fast uptake kinetics. A major hurdle here is the high cost of regenerating the material to release the captured CO2, which currently limits applicability on a large scale. Computational simulations serve as a virtual experimental platform to understand the thermodynamics and kinetics of the CO2 adsorption and release driven by cooperative chemical reactions in a dynamic and complex chemical environment. To extract rate constants, simulations will need to be highly accurate, and high levels of theory with a large computational complexity will need to be used to obtain the desired results. Dynamics simulations will be critical to understand the transformation of molecular crystals in the presence of carbon dioxide and water, and to understand the transformation when carbon dioxide gets released through heating or other thermodynamic means. Since the processes involve chemical transformations not suited for classical MD approaches, ab initio dynamics combined with classical molecular dynamics simulations will be required to gain the needed scientific insight. So far, simulations will provide fundamental insights. To design novel molecular crystals will require the generation of large data sets, that can be used with ML and inverse design strategies.

First Principles Simulation of Heterogenous and Hybrid Materials

These simulations are hybrid from start to end and they will continue to be hybrid also when using quantum computers, hence the necessity of transferring terabytes of data at different stages of the work. In addition, these simulations require multiple plans for data storage because some studies are long enough that storage availability changes during the course of the investigation.

5.17.2.2 Collaborators

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
HPC FACILITIES (NERSC, OLCF, ALCF)	Primary and secondary	Usually data transfer (Globus)	Dependent on use case below, GB-TB	Daily	Y, data transfer	Data formats, searchability of data
AMES/BNL/ LBNL/PNNL	Secondary	Data transfer, although data portals will be in the future	10s of GB to TBs	Weekly to monthly per collaborator	Y, data transfer	Data formats, searchability of data
ANL/LBNL/PNNL/ LANL	Primary and secondary	Globus	10s of GB to TBs	Weekly to monthly	Y, data transfer	Compatibility of data formats for quantum simulators
BNL	Secondary	Data transfer (Globus)	10s of GB to TBs	Ad hoc	Y, data transfer	N/A
LANL	Primary and Secondary	Data transfer and portable hard drive	102–104 GB	Ad hoc, roughly monthly	N	Data formats, searchability of data, fast node response time for I/O
LBNL (NCEM, MF, MSD), NERSC, UC BERKELEY	Primary	Data transfer (Globus)	10 GB to TBs	Weekly-monthly	Y, data transfer	Format, readability, transferability for different applications
QMC-COLLABORATORS (NATL. LABS, LED BY ORNL, AND US UNIVERSITIES)	Secondary	Data transfer portal	10s GB to 10s TB	Weekly	Y, data transfer	Data query, fast nodes to download (upload) few TB files

Table 5.17.1: Quantum Collaboration Scenarios

It is important to note that the types of science and users/collaborations described above involve multiple facilities/locations to accomplish the science. So, these should be considered examples of typical use cases. Many university partners are involved in these collaborations (too many to number) and data sharing is primarily challenged by data formats and searchability – although in some types of simulations, the size of the data to be transferred could be an issue. Appropriate levels of security should also be considered during data transfer while not significantly impinging on the timely completion of work.

5.17.2.3 Use of Instruments and Facilities

Quantum simulations of materials and molecules use a variety of platforms, encompassing midsize compute cluster and storage systems owned by universities, and high-performance architectures at national facilities (NERSC, ALCF, and OLCF). Most problems require the transfer of data between university clusters or national laboratory resources and HPC facilities, both at the time of investigation and at the completion of the project, when storage of data is planned (often data may not reside on the HPC facilities where they have been created for future access, curation and reuse). Data produced at the facilities can range widely based on the type of simulations and computational workflows and campaigns needed to accomplish the science. Often times, the simulations on the smaller clusters (at the national laboratories or universities) are used to generate initial structure information for larger simulations. This data tends to be on the order of 10 MB per file and for large campaigns there can be tens of thousands of these files. For the HPC facilities, current file sizes can range on the order from 1 GB to 10 TBs, the number of files and directories can range from 1 to tens of thousands with total

data set size maxing out in the 10s of TBs. This will likely not change significantly until the exascale computers become more available and scientific challenges are tackled at this scale. Once these systems become seriously available to more users, the size of all of the data will likely be multiplied by 100 to 1000-fold.

The future of computing in the two to five year timeframe and beyond will require both active and passive interaction between multimodal experimental and multifidelity/physics theoretical approaches on nationally federated experimental, computing and storage facilities, handled by smart workflows. The urgent need of the hour is on-demand computing, with interoperable smart workflows across various HPC as well as midcapacity computing platforms, along with long-term cloud storage solutions. As an example of active interactions, workflows that spawn expensive yet accurate quantum simulations of materials on high-end HPC clusters, such as using the QMC method, could be requested in a targeted manner based on errors assessed using lower-fidelity modeling approaches performed on midcapacity clusters that are both available on demand and linked via fast data-transfer nodes to synchrotron and microscopy (user) facilities, nationwide, which are themselves guided by AI/ML algorithms running on edge clusters at these facilities. As an example of passive interaction, the stored experimental data sets corresponding to different characterization measurements from small/big experimental facilities are constantly used in the background to assess low-fidelity simulation stored results, to inform simulation protocols that bridge different length/timescales, as well as provide training data to improve the AI/ML models that drive experiments and perform data reduction at these facilities on their edge and/or local-hardware, enabling predictive simulations of functional properties of materials. We foresee that such interoperable smart workflows will be needed to guide future synthesis and characterization experiments in the right direction for autonomous discovery of new materials with desired properties, leading to rapid identification of the optimal technological solutions to our pressing global challenges.

5.17.2.4 Process of Science

The unifying theme of the simulations and instrumentation described above is that of scientific discovery involving analysis and sharing of the many simulations and analysis among teams of scientists. Already, the common mode of operation is for multiple team members to work together on data produced from multiple sources (both hardware and software) to understand the different physical underpinnings of a science challenge. While workflows are starting to be used in these teams, it is still not uncommon for each scientist (or local group of scientists) to use their own methods and worry about the integration with other data later in the process. It is clear that this is changing and will be a challenge for the next several years and into the timeframe beyond the next five years. Unfortunately, there is no one scientific workflow that is universal for all simulations. However, there are some commonalities.

- Often it is the case that a scientist will want to search existing data to understand the simulations that have been performed before. This is currently a difficult task due to varying issues such as data formats, provenance and trustworthiness of the data, and public accessibility and searchability of data across federated data sources. Beyond the next five years, we anticipate that this will become more possible as common data formats (or at least converters between formats) become more available. From a network point of view, this will mean much more data being exchanged between scientists in an asynchronous manner than the current case of explicit transfer of data between scientists.

- This data is then manipulated into a form where local resources are used to develop initial targets (parameters) for simulations on HPC or (depending on the nature of the simulation) cloud computing. This may involve data transformation or a set of simplified simulations. In the next several years and into the future, this will likely continue since many local sites are still committed to providing local computing resources.

- This initial data is then transferred to an HPC resource for highly accurate and/or large capacity simulation. Currently much of this is accomplished through direct data transfer and is largely human directed (often using scripts). This is changing as more users are starting

to adapt workflow systems that allows multiple types of simulations and/or the same type of simulation on multiple molecular configurations. Ultimately, beyond the next five years one could anticipate that many users will use workflow systems that will allow them to run many simultaneous simulations on varying resources to accomplish the computation for the scientific challenge. This will mean that the data exchange will become more rapid fire and that it will likely involve many network paths since more computational resources than just one primary one will be used to accomplish the science.

- Data reduction currently happens mostly on local resources — meaning that the simulation data is often transferred from an HPC resource to a local computer for analysis. With more federated data and the need to combine multiple data streams, the need for these data transfers will either increase significantly or better ways of visualizing and analyzing data will need to be developed. In the former approach, the current model will continue but with multiple copies of the same data on multiple computational platforms.

- We expect a need for on-demand midcapacity computing and long-term storage solutions that are woven with fast data-communication networks and smart interoperable workflows, with connections to HPC facilities as/when needed for capacity computing, AI/ML model training and big-data reduction and visualization.

- HPC computing and experimental facilities are becoming more and more integrated, as simulations become a critical input to data analysis at facilities such as the neutron and X-ray sources. Real-time analysis, computing and data movement will be critical. Approaches to this integration are being developed under the super facility model, where ESnet is a critical component.

Overall, it is anticipated that there will be more frequent communication required and many more disparate sources of data to accomplish complex scientific workflows.

5.17.2.5 Remote Science Activities

The current and near future remote resources will include the DOE HPC facilities as well as remote access to data sources/repositories (such as the MP, NIST Materials Data Repository, etc.). Beyond five years, cloud computing could become more used as computational chemistry and materials science software becomes more available on those resources. It is also possible that cloud computing/data sharing will become more used for data analysis and visualization.

5.17.2.6 Software Requirements

One very important current and future need related to software is the ability to share and transform data into formats that can be used in shared visualization and data analysis without extensive user input. Various efforts within DOE, NSF and NIST are underway to standardize data formats for data storage and sharing, there is still a lack of uptake in the community to use the standards and in the development of software that can facilitate the use by the end user. Progress is being made, with many of the computational chemistry and materials science centers who are developing software tools that are interoperable, or providing clear data APIs. With a growing adaptations of data standards by the scientific community one can anticipate that software and standardized formats will need to be supported at the various computing sites.

5.17.2.7 Additional Network and Data Architecture

Quantum simulation can be made into an analysis or characterization tool to work together with other experimental facilities. In an integrated scientific discovery process, large-scale real-time computing is in demand, so easy access to a HPC center from an experimental facility is needed. A network with unified cybersecurity policies might be necessary for such seamless connection to happen.

5.17.2.8 Use of Cloud Services

Cloud computing is not currently used extensively in this community. However, it is possible that more cloud computing will be used to analyze complex amounts of data and for large scientific campaigns of smaller sized individual simulations. The reader is also referred to section 8 of the case studies of the NSRCs for additional information.

5.17.2.9 Data-Related Resource Constraints

Currently, most of the data storage and transfer speeds are appropriate and this is unlikely to change within the next several years. As discussed previously, data accessibility, searchability, and usability is a key issue that is our community is facing and will need improvements to enable high throughput scientific workflows.

An additional concern will be to ensure that the large software ecosystem in the chemical and materials science are able to continue running well on the ever-changing computing resources available now and into the future. In particular, there are still significant needs for software that performs well on heterogeneous computers (although significant strides have been made with the ASCR funded Exascale Computing Project, the BES-funded Computation Materials and Chemical Sciences programs). In addition, cloud computing could be an area of need in the future.

There is a need for experienced data/computing personnel for streamlining of workflows, data management and custodians that are flexible to respond to specific needs of different projects/applications.

5.17.2.10 Outstanding Issues

None to report at this time.

5.17.2.11 Case Study Contributors

Use of the ESnet for Quantum Simulations of Materials and Molecules Representation

- **Giulia Galli**, *ANL*, gagalli@uchicago.edu
- **Kipton Barros**, *LANL*, kbarros@lanl.gov
- **Panchapakesan Ganesh**, *ORNL*, ganeshp@ornl.gov
- **Stephen Gray**, *ANL*, gray@anl.gov
- **Sinéad Griffin**, *LBNL*, sgriffin@lbl.gov
- **Wibe de Jong**, *LBNL*, wadejong@lbl.gov
- **Theresa Windus**, *AMES*, twindus@ameslab.gov
- **Qin Wu**, *BNL*, qinwu@bnl.gov
- **Jianxin Zhu**, *LANL*, jxzhu@lanl.gov

5.18 The Materials Project (MP): Status and Future Directions

The MP is a multi-institution, multinational effort to compute the properties of all inorganic materials and provide the data and associated analysis algorithms for every materials researcher free of charge. The ultimate goal of the initiative is to drastically reduce the time needed to invent new materials by focusing costly and time-consuming experiments on compounds that show the most promise computationally.

5.18.1 Discussion Summary

- The MP was founded in 2011, and now serves a community of over 200,000 registered users.

- Data needs from MP users range from highly frequent requests for common objects (100s of GB to TB sized downloads) to less frequent, but larger, data objects (10s to 100s of TB).

- The need to transfer data to and from MP is expected to increase as third-party contributions of data grow.

- The MP infrastructure was initially supported by LBNL and NERSC compute and data infrastructure, and is now transitioning to AWS cloud resources to serve the growing user base and usage patterns.

- Community contributions through a dedicated platform remains a critical research driver that now results in millions of submissions that can be efficiently processed.

- Integrating cloud and DOE infrastructure in a seamless manner is possible, and can serve as a model for other projects that may need to scale infrastructure beyond the needs of DOE compute centers, while still making optimal use of DOE resources.

- Despite the flexibility, the cloud resources can be cost prohibitive in storing and indexing very large data sets, or performing long-running computations in batch. The MP model leverages DOE infrastructure for high-throughput computations and large-scale data management services: the core subset of the data is then made available via cloud resources, while less commonly used related data or source raw data can be handled directly through DOE infrastructure. A networking pipeline between the two is still facilitated by ESnet.

- MP currently relies on NOMAD for public dissemination of raw calculation output.

- Services like Spin at NERSC have also been valuable, but suffer from geographic exposure and no uptime guarantees. Seeing services like this extended to multiple sites to allow greater redundancy would also be a benefit.

- MP's transition to the cloud and its described architecture have proven to provide low latency with high reliability and security. An important continuation will be to connect resources in the private subnets of MP's AWS VPC with resources on-premise at LBNL/NERSC, or other ESNET-connected HPC facilities.

- High bandwidth/low latency connections between AWS and LBNL/NERSC resources will be important to ensure that data can be ingested into the cloud resources in a timely manner at all DOE facilitates, delivered via ESnet connectivity.

- Having no dedicated connectivity between NERSC resources and AWS private network presents a significant roadblock for progress in transitioning to a seamless cloud experience. Enabling direct access and data transfer between MP's cloud resources at AWS and on-premise resources at other DOE facilities would significantly reduce the burden of contributing and linking data.

- Transferring data between DOE facilities remains cumbersome; MP is trying to move toward accessing other facilities from the cloud, and would like to see templates developed in this use case to allow other facilities to set up cloud resources and then connect on the VPC level within AWS regions.

5.18.2 The MP: Status and Future Directions Facility Profile and Case Study

As a public initiative supported by funding from the US DOE, the mission of the MP is to compute properties of all known materials and remove guesswork from materials design. Experimental research and the discovery of novel materials for applications such as energy storage, advanced electronics, and structural components can be targeted to the most promising compounds by screening and applying ML algorithms on MP's computational data sets. MP now serves a community of over 200,000 registered users.

5.18.2.1 Science Background

MP started in 2011, supported by LBNL and NERSC compute and data infrastructure. Data production and ingestion have been managed through multiterabyte MongoDB databases at NERSC, along with development clusters with web services hosted on the NERSC Science Gateways/Spin Platform. This approach served MP well for over a decade but with a community now having grown substantially (>200,000 registered users) and heavy use of both the online gateway and API, transitioning the MP infrastructure to the cloud became a cornerstone to the longevity of MP. It was evident that the MP collaboration and broader community needed to be empowered to contribute data and apps, and that responsibilities for different parts of the MP stack needed to be separated in order to scale with demand.

As a result, MP underwent a major effort to modernize its web infrastructure and transition it to the Cloud. The project designed, developed and deployed a microservices-based network architecture on the AWS cloud. The main web portal is in the process of transitioning to the new AWS based service. The new cloud architecture allows MP to meet modern requirements for high availability and rapid solutions. The MP use case is centered around enabling a new, hybrid networking infrastructure that enables seamless integration of cloud resources with computing and data infrastructure hosted by LBNL, NERSC and other DOE facilities through ESnet.

Overall, MP presents a unique opportunity to help develop and establish protocols for data transfer and management to scientific communities. Its model of a hybrid DOE-based and cloud architecture represents an opportunity to design scientific data management and dissemination systems that make the best use of all available tools. Data needs from MP users range from highly frequent and globally distributed requests for a core data set (i.e., ~35 million monthly requests) to less frequent requests for larger raw data objects. The need to transfer data to and from MP is expected to only increase as the MPContribs platform, enabling third-party contributions of data, continues to serve as a materials data provider and partnerships with ESnet may form a key enabler of this transformation.

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
CORE TEAM	Both	website, MongoDB, HPSS, CFS	hundreds of GBs to hundreds of TBs	All frequencies	No	No link between NERSC resources and AWS private network presents roadblock (site-to-site VPN required)
WEBSITE USERS	Secondary	Portal	100GB Atlas, 5TB S3	Daily	No	
API USERS	Secondary	REST API / Python client	see website	Daily	users can contribute small structure files for calculation	
MPCONTRIBS USERS	Both	portal, rest api, Python client	currently in the tens of GB but potential to grow quickly	daily, ad hoc	Yes, in the case of contributed data. Upload happens via REST API to Atlas (cloud hosted MongoDB) database.	
LIGHTSOURCES	Both	portal, rest api, Python client	currently in the tens of GB but potential to grow quickly	daily, ad hoc	Yes, in the case of contributed data. Upload happens via REST API to Atlas (cloud hosted MongoDB) database.	

Table 5.18.1: MP Collaborators

5.18.2.2 Collaborators

MP collaborators are listed in Table 5.18.1.

Community contributions through our dedicated platform MPContribs[4] are another powerful example for the use of such a connection. MPContribs has been reliably available to the public since late 2018, and has now matured to process half a million submissions within a day. In particular, it has started serving the light sources (e.g., ALS, NSLS-II) as well as ML communities. MPContribs has become an integral part of MP's next-generation website, underpinning the effort to disseminate data and apps contributed by the materials sciences community (see for example MOF and Catalysis Explorers) while leaving full ownership and control over the data with contributors. Enabling direct access and data transfer between MP's cloud resources at AWS and on-premise resources at other DOE facilities would significantly reduce the burden of contributing and linking experimental as well as computational data to materials and apps on MP. For example, it opens up the potential for deeply integrated, automated and bidirectional data pipelines between DOE light sources and MP.

5.18.2.3 Instruments and Facilities

NERSC infrastructure continues to support MP operations in compute, data backup and archive, local databases, and web services via the NERSC Spin platform. Specifically:

- Raw calculation outputs from the VASP code are stored (or transferred) by group members and collaborators to NERSC CFS. A significant bottleneck is easily transferring calculations run externally to NERSC, since this typically requires new account setup, Globus training, and the like.

---

[4] https://mpcontribs.org

- Command line tools developed by MP for data handling prepare raw data for organized archival via HPSS. The HPSS archive becomes the "source of truth" for all subsequent data building and dissemination to enable reproducibility.

- Periodically, subsets of files are transferred from HPSS to CFS or elsewhere to enable data processing via the XFER queue at NERSC.

- Parsing of raw data into a more versatile format goes into a central MongoDB collection managed by MP.

- MP "builders" use dedicated LBNL IT machine for high-throughput generation of derived MongoDB databases and collections corresponding to variety of materials properties.

- For active research work, MP runs and maintains dashboards to allow scientists to track the progress of their calculations. Individual scientists also have their own dedicated MongoDB databases for research purposes.

- MP processing pipeline also syncs to Google Drive and NOMAD, an external repository hosted in Europe and focusing on raw materials output files. NOMAD provides internal and external access to raw VASP output files as an extra layer of backup.

5.18.2.4 Process of Science

Data generated at NERSC and other locations is stored centrally and disseminated to the community through our online portal[5] and through our API. A schematic of the sequence of operations is provided in Figure 5.18.1.
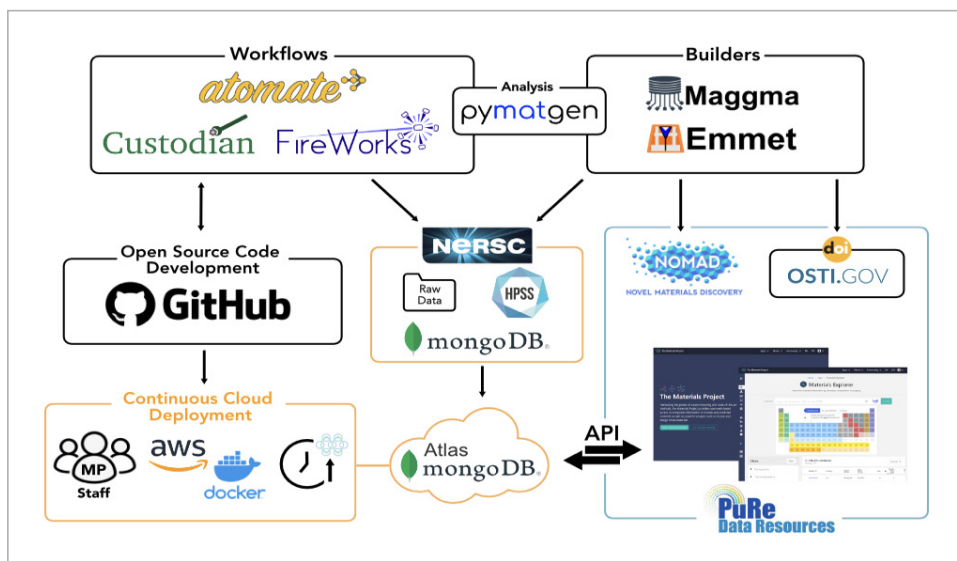
Figure 5.18.1: MP Process of Science

In pioneering the use of this hybrid model, MP can integrate cloud and DOE infrastructure in a seamless manner. Indeed, we expect that MP's design may serve as a model for other projects that may need to scale infrastructure beyond the needs of DOE compute centers such as NERSC while still making optimal use of DOE resources. While the AWS cloud will increase availability and allow us to scale to meet the growth in users, these cloud resources can be cost prohibitive in storing and indexing very large data sets, or performing long-running computations in batch. In our model we leverage DOE infrastructure for high-throughput computations and large-scale data management services. The core subset of the data is then distilled and made available via cloud resources, while less commonly used related data or source raw data can be handled directly through LBNL/

---

[5] http://www.materialsproject.org

NERSC infrastructure. MP can use the high bandwidth connectivity provided by ESnet to connect access DOE resources with AWS.

Opportunities include an ESnet connection to user resources at other DOE facilities to allow direct access to raw data on user infrastructure from MP services running at LBNL. This will allow, for example, development of GUIs that could significantly reduce the burden of extracting high-level data and integrating it within the MP website and API for broader sharing with the community. Similarly, data coming from beamlines at lightsources could be directly accessible to MP tools in the cloud, allowing more seamless transfer of data if select DOE resources were accessible via MP's AWS VPC.

5.18.2.5 Remote Science Activities

Remote lightsources (and other DOE resources like HPC, file storage, DBs) can use MP interfaces to "pull" data into MP through VPN site-to-site to ESnet.

5.18.2.6 Software Infrastructure
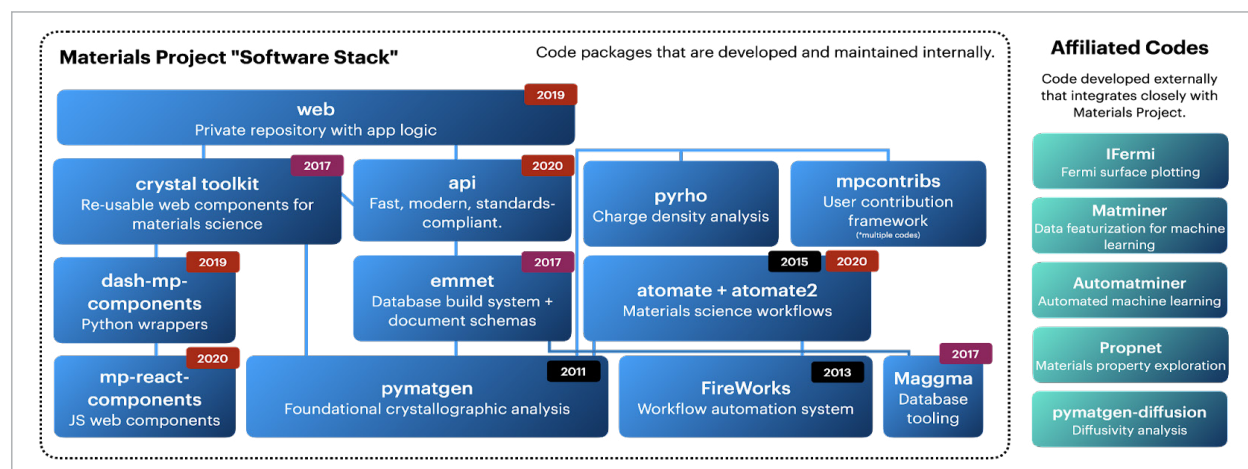
Figure 5.18.2: MP Software Infrastructure

All software tools are developed as open-source codes and allow community contributions. The full MP software stack is now quite extensive, and includes:

- Pymatgen for calculation definitions, structure objects, and materials analyses
- FireWorks, a general purpose workflow software for defining, managing, and executing millions of complex workflows (dispatch and manage jobs on HPC infrastructure)
- Atomate defines materials science workflows using the FireWorks infrastructure
- Emmet builds databases from original "tasks" collections (see 3) and implements API server with definitions for API documents/schema (with maggma for database tooling)
- MPContribs repo contains implementation for web portal, API server and client
- MP web and API services use cloud resources; atomate and FireWorks are run on HPC

5.18.2.7 Network and Data Architecture
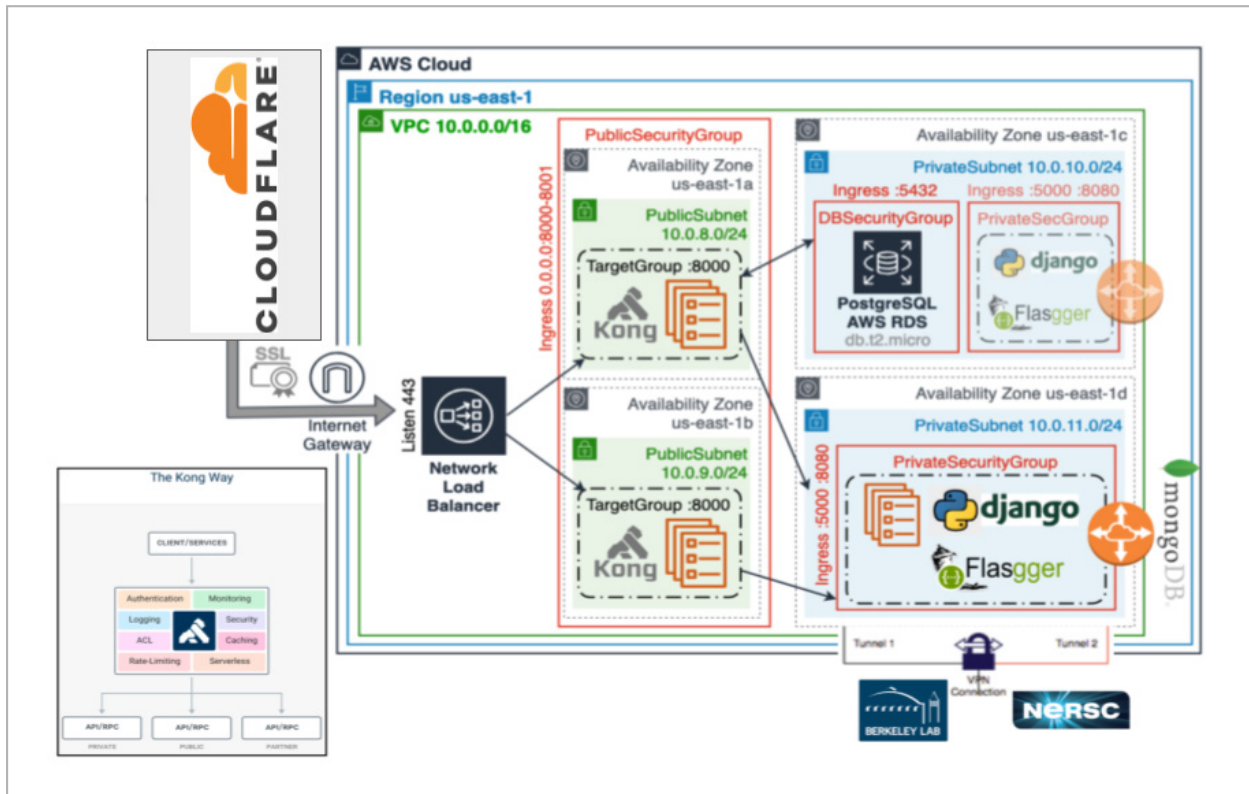
Figure 5.18.3: MP Data Architecture

MP's transition to the cloud and its described architecture have proven to provide low latency with high reliability and security. An important continuation in that vein is to connect resources in the private subnets of MP's AWS VPC with resources on-premise at LBNL/NERSC (or other ESNET-connected HPC facilities). Possibilities for implementing this include a private virtual gateway with a stable/reliable VPN or AWS DirectConnect connection.
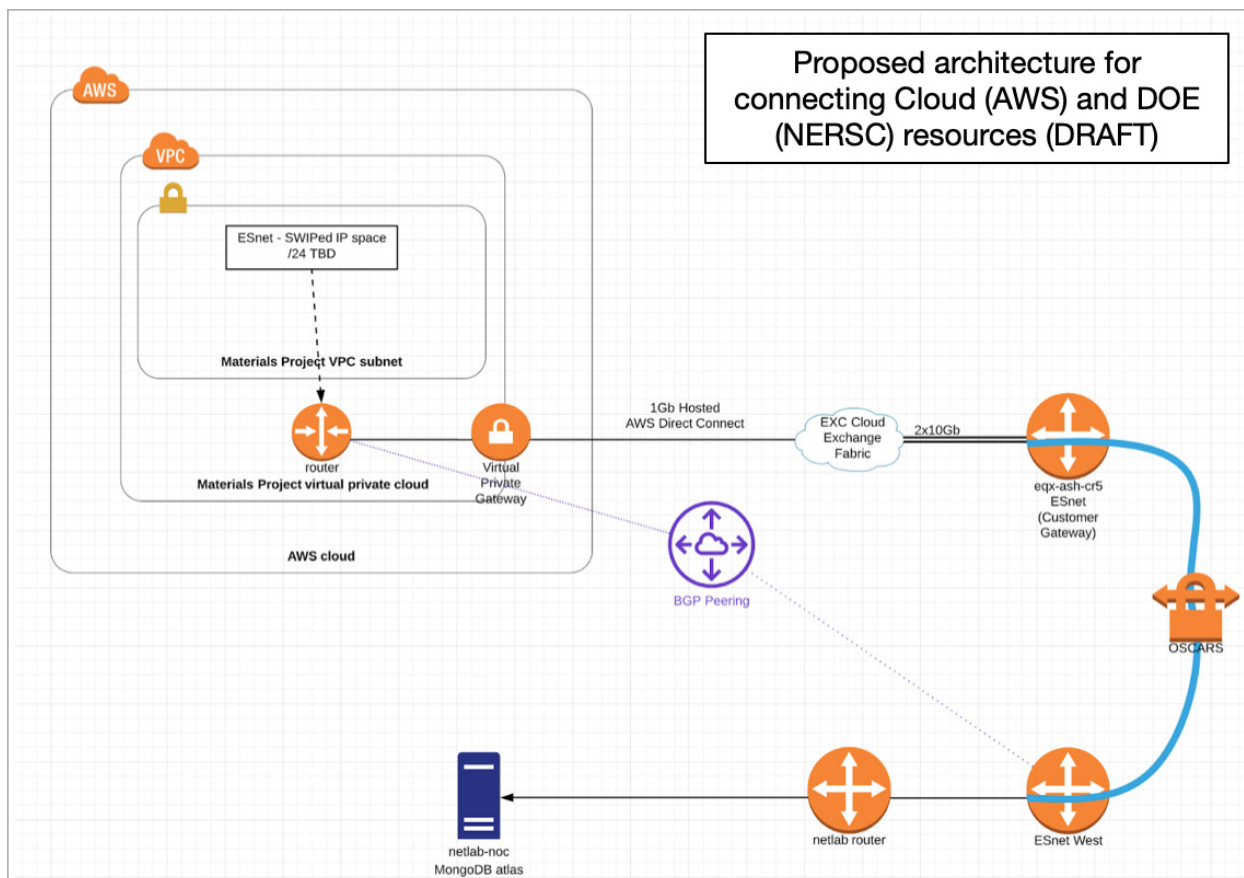
Figure 5.18.4: MP Cloud Architecture

MP can use this connection to access MongoDB and disk resources in the LBNL/NERSC network for data that is less frequently needed, and does not require high availability on AWS, but would be prohibitively expensive to host on Atlas MongoDB. MP would be accessing data from its website and APIs in real time through a direct connection to LBNL/NERSC, and requires a low-latency connection between AWS and LBNL/NERSC to provide end-to-end access to this data. Additionally, a high bandwidth connection between AWS and LBNL/NERSC resources will also be important to ensure that data can be ingested into the cloud resources in a timely manner.

Early discussions with ESnet yielded plans for a proof of concept demo, which could be a set up to show that we can establish and use the connection. We include the demo plans as a reference point: ESnet could set up the hosted connection in Equinix Cloud Exchange Fabric. A low bandwidth connection would likely suffice for this pilot project. For demo/testing purposes, a MongoDB resource would be needed in the ESnet network. The goal is to design a network path between the MP VPC instance in the AWS EAST-1 and EAST-2 regions and by way of ESnet transport, connect a stub network instantiated in the AWS cloud to LBNL and NERSC. This would deliver a single prefix (public ESnetClassless Inter-Domain Routing) through a dedicated EBGP peering at the LBNL and NERSC network edge.

5.18.2.8 Cloud Services

The MP cloud stack consists of a collection of microservices based on modern container technology, which exchange traffic within the network of a VPC. Each set of dependent microservices is deployed on AWS serverless compute engine Fargate with IP addresses in dedicated VPC subnets. Using Fargate removes the need to own, run and manage the compute infrastructure and instead allows MP to focus development on its

applications with security groups controlling traffic flow from the Internet and between subnets. MP's cloud architecture employs private VPC endpoints to make use of free intraregion traffic and improved latency/reliability between its Fargate microservices and other AWS services as well as the cloud-monitoring service Datadog. This traffic includes about ~5.6 TBs of data and 3.8M objects stored in S3 Object Storage. All of MP's API microservices use a private peering connection to a managed MongoDB Atlas instance in the same AWS region reducing traffic costs by a factor 10. The MongoDB instance currently holds about 140 GB of queryable and indexed data. Cloudflare's global network secures and accelerates MP's services before traffic arrives at the AWS network load balancer. Current (pre-public release) daily cloud costs add up to about $60/day. As MP scales to full traffic, cloud costs are expected to increase primarily due to AWS egress costs. The majority of these are likely to be covered by LBNL's egress waiver from AWS.

While the current deployment is on the AWS cloud, care has been taken to ensure services are portable, and could be deployed on other clouds as necessary, depending on future requirements.

5.18.2.9 Data-Related Resource Constraints

We are currently relying on NOMAD (an EU-funded scientific project) for public dissemination of raw calculation output, since they provide a platform and means to publicly share and transfer large amounts of data that are currently outside our means.

Transferring data between DOE facilities remains cumbersome; we are trying to move toward accessing other facilities from the cloud, and would like to see templates developed in this use case to allow other facilities to set up cloud resources and then connect on the VPC level within AWS regions (or equivalent for other cloud providers).

Services like Spin at NERSC have also been very valuable but suffer from geographic exposure and no uptime guarantees, seeing services like this extended to multiple sites to allow greater redundancy would also be a benefit.

5.18.2.10 Outstanding Issues

None to report at this time.

5.18.2.11 Facility Profile and Case Study Contributors

The MP Representation

- **Kristin Persson,** *LBNL,* kapersson@lbl.gov
- **Anubhav Jain,** *LBNL,* ajain@lbl.gov
- **Matt Horton,** *LBNL,* mkhorton@lbl.gov
- **Patrick Huck,** *LBNL,* phuck@lbl.gov
- **Bobby Sumpter,** *ORNL,* sumpterbg@ornl.gov

ESCC Representation

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

6 Focus Groups

A core component of the ESnet requirements review process displaced by the COVID-19 pandemic was the opportunity to hold impromptu conversations with colleagues. These could occur during the oral case study review period (and involve topics being presented or stumbled upon), but were also equally likely to occur before, during, or after the physical meeting. The importance of these interactions cannot be overstated, as they often resulted 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 May 2022, the BES requirements review team convened two virtual focus groups. The general plan for these meetings was to:

- Gather together small groups of case study authors during predefined time periods, using virtual tools.

- Prepare the groups by having them review outlines of their case studies and research focus areas (if they were unfamiliar).

- Structure a conversation to review areas of research, and then seed conversation with a set of topics that were found to be common across all case studies in the 2022 BES requirements review.

During these focus sessions, the BES requirements review team acted as a moderator for the conversation, but let discussion flow organically toward topics of mutual interest. The goals were to:

- Allow emerging projects and facilities to ask questions of the established BES 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 BES program area.

6.2 Organization

The BES requirements review featured 18 case study groups. The optimal way to organize focus groups was to offer two events, and invite all parties that were available to attend; this organizational assumption acknowledged the fact that not all participants could attend both. Similar discussion topics were available at each event, but the chosen topics could differ drastically depending on participation. The events were as follows:

- Focus group 1 was held on May 5, 2022.
- Focus group 2 was held on May 31, 2022.

The agenda for each event was designed to be simple and dedicated to keeping a majority of the event available for attendee discussion. A brief introduction from the BES requirements review team and an overview of each meeting's purpose . The remainder of the 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 were as follows:

1. Supporting "remote" control and operation: use cases that started pre-pandemic, and have continued beyond. How are users managing new workflows?

2. Challenges in supporting multi- or coupled facility workflows: are there complications in pairing BES experimentation with ASCR computing and storage via ESnet.

3. Data storage locality, quantity, and mobility: storage needs are outpacing capabilities to generate and analyze BES data. What is being done, and are there suggestions to improve the national data architecture.

4. Computation that has real-time, near real-time, or offline requirements: use cases that require computation during the process of experimentation and simulation. Limitations to current practices, and thoughts on how to improve. How this affects and encourages network use.

5. Software to perform analysis, sharing, simulation, or other use cases: current, near-term, and future needs and development opportunities that may leverage network resources.

6. Future networking and service requirements: capacity, traffic expectations, future services.

7. Facility upgrades and potential changes to data volumes and rates: How are users modifying workflows to prepare, and how can networks be adjusted to fit the new requirements?

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 who attended.

6.3 Outcomes

The following sections highlight the areas of discussion and relevant findings and recommendations that emerged during the talks. Some are directly related to the structured conversation, but others came out of natural discussions that may have strayed from the topic areas.

6.3.1 Focus Group 1

- The polling during the meeting produced the following discussion topics that were of interest to the assembled group:

- Data storage locality, quantity, and mobility.

- Software to perform analysis, sharing, simulation, or other use cases.

- Computation that has real-time, near real-time, or offline requirements.

- Facility upgrades and potential changes to data volumes and rates.

During this period of discussion, several notable items were brought up, and these are detailed in the following subsections.

6.3.1.1 Data Storage Locality, Quantity, and Mobility

From the case study review, two primary data flow paths are present in the BES workflows:

- Acquisition from instrument/detector, and flowing to some form of local/buffer storage.

- Migration from local/buffer to long-term storage, analysis, or sharing with collaborators.

The first workflow pattern, between the instrumentation and local capabilities, is universal throughout the BES community with every facility having a set of common approaches and mechanisms. The second workflow pattern, between local capabilities and more centralized resources, varies considerably depending on size and relationships. For example, at X-ray facilities such as ALS and APS, each may have workflows that push

large amounts of data to remote HPC facilities. The ESnet staff leading the discussion was most interested in understanding the specific approaches that facilities were using to support the use cases, along with understanding what works well, and what could use improvement.

All of the participants agreed that the two use cases are critical to the workflow. In fact, during the polling exercise that ranked the topics, many noted that it was hard to pick one to start with because they are all of high importance, and even the lowest ranked item ("Future Networking Needs") was still critical, since most of the other topics must build upon that topic.

Large X-ray facilities (e.g. APS, ALS, NSLS-II, etc.) have adapted to the data mobility challenges over time, and tried many different approaches to get to where they are today. Many anticipate that the answers they use today will be a stepping stone to changes in the future. Active research and development is being done to experiment with new tools and approaches as data volumes increase and the user community adapts to new ways of working with large facilities (e.g., remotely).

Interest in continuing to pursue workflows that involve major HPC centers is universal: most experimental facilities anticipate needing to leverage external resources for analysis workflows, along with emerging use cases that use AI and ML to steer experimentation in the future. HPC centers will continue to have a larger amount of computing and storage resources than an experimental facility, and support high-throughput data mobility tools that a number of scientific communities (including BES) use as a part of their workflows.

Negative aspects to relying on external HPC facilities that were mentioned during discussion are:

- Tying a workflow to resources not directly controlled by an experiment means that if the resources are not available (e.g., due to scheduled or unscheduled downtime), there is a risk of experimental delay, or failure. This could be mitigated with the ability to migrate to other resources at the same facility, or others that may support a similar API.

- HPC allocations are not automatic, and must be applied for on cycles. When they run out, it can sometimes be hard to request more. Having a more stable way to rely on HPC resources at a given facility (or several interchangeable facilities) is desirable.

Both of these items were featured in discussions as a part of the DOE ASCR Integrated Research Infrastructure (IRI) activity, taking place throughout 2022. In particular, there is a strong push to unify the computational, storage, and networking portions of common DOE workflows to use a similar set of API hooks and other technology components. This would facilitate workflow migration to available resources, and unify the way that resources can be requested and reserved.

Multifacility workflows also contain a deep tie to networking resources, given that many facilities are separated by geographical location. As BES workflows become more time dependent, particularly those that may try to utilize computational output in a real-time fashion, networking is a critical component that must perform well consistently. If a primary link becomes saturated or fails, the workflow could also fail.

ESnet can work with the major facilities to set up backup networking options (e.g., L2VPN service similar to other science communities such as HEP) if this functionality would help smooth out uncertainty about future workflows. Overlay networking options, similar to the Large Hadron Collider Open Network Environment (LHCONE), could allow multiple BES collaborating facilities a stable and performant way to communicate built on top of the traditional ESnet network infrastructure. Active research and development is being done to expose networking APIs to intelligent data movement tools (e.g., Rucio) such that provisioning paths (and backups) can be done as a part of the overall workflow. In addition to smarter tools aware of network capabilities, there has also been interest in exploring on-demand (e.g. streaming) data movement options.

An additional complication to data sharing comes when administrative boundaries are crossed. Identity management continues to be a friction point for BES users. This has been exacerbated in a multifacility workflow model where a user may need to authenticate to multiple locations, using multiple identities, on a regular basis.

Some tools are smarter about their ability to cache credentials, but this still does not solve the root problem of forcing users to perform the steps multiple times on a regular basis. A unified way to handle identity is still highly desirable.

The NSRCs and neutron sources note that while their data volumes are not growing to the current levels experienced (or anticipated) at X-ray facilities, the volumes are still projected to grow beyond their current technology demands. For example, SNS is constructing a new instrument (VENUS) that is projected to create 20–30 TB of raw image data in a single day. There will be mechanisms to reduce this locally, creating usable data sets that are a fraction of the size. Preparations are being made at the facility to adapt to the new volumes by upgrading networking capacities and making additional storage available.

AI and ML are hungry consumers of research data, at least in the early phase of their use when they are trained in anticipation of a new use case. These needs are currently viewed as self-contained to a given facility or use case, but there are efforts to broaden the effort by sharing more training resources across individual facility use cases. This will also increase the data mobility requirements: both at BES facilities and HPC centers that run the algorithms. BES will employ more use cases for AI and in the future and has experimented with both edge computing in or close to the border of the instrumentation, as well as centrally located at large computing facilities. ESnet maintains a testbed that is exploring edge-computing options, as well as the aforementioned research into smarter APIs to interact with network functions.

A last area of conversion in this topic centered on the growing demands of storage, and the temporal nature of allocations. In the future, having access to research data that spans facility boundaries, without having to worry about who may own or maintain the data, is desirable. The DOE ASCR IRI activity is also exploring these topics, and may recommend a more central way to store and manage research data that spans the administrative boundaries of a specific experimental facility. If this facility becomes a part of the ASCR ecosystem, ESnet would support connectivity to it.

6.3.1.2 Software to Perform Analysis, Sharing, Simulation, or Other Use Cases

Several X-ray facility case studies had common visions for the future of architecture, software, and workflow components in the current and future use cases of the BES community. DISCUS is the architectural vision that will integrate computation, software, instruments, and networking in a standard way. Bluesky will be an attempt to capture information and intent during the experimentation process to allow recording and reporting in a standard way.

The NSRCs and neutron sources currently employ a heterogenous approach, which must change to be more uniform. This comes from the environment in which many operate: a number of different types of instruments that each have a different software stack (some of which may be vendor specific). This environment does not lend itself to creation of a standard workflow without creating a lot of custom software and APIs. Many of the NSRCs are at a tipping point where the amount of data produced by some of the instruments is growing beyond current capabilities. Thus a standard approach to deal with current (and future) technology with unified software, computation, and storage is desirable.

With the growing data volumes, local networking must also be upgraded: typically the 1 Gbps control networks must be increased to 10 Gbps to allow for more seamless operation. As this occurs, the possibilities of integrating to national-level resources (such as HPC facilities) is also within reach. To use large HPC facilities means having to adhere to their standards for software and data management, which is also a forcing function to change the way that NSRCs operate. ESnet will continue to collaborate as necessary to link facilities that can benefit from multifacility workflows.

A general comment surrounding BES facilities (e.g., X-ray light source, neutron scattering, or NSRC) is that many of their designs and operational patterns are closely tied to the era in which they were designed. Each has different funding models, and different data production expectations. As some scale upwards, they are learning

valuable lessons that can be shared with others that have yet to reach those levels. In some cases a facility may be on the bleeding edge of a technology requirement, such as adopting software that requires intense engineering knowledge. Getting to that level of operation could take subject-matter experts away from their core competency of their science just to learn how to adapt the technology. The BES community does not want to be experts in software or networking engineering, and relishes adopting BCPs that the ASCR community can recommend. This will scale collaboration in the future. Examples include the adoption of the Science DMZ approach to networking, or the Globus platform for data mobility.

NSRCs, even with smaller data volumes, still gravitate toward some of the software and workflow solutions that other larger facilities have pioneered. In particular, the organizational structure of some of the software workflow tools works well locally, as well as nationally. These small productivity wins can outweigh the requirements of having near unlimited levels of bandwidth. For example, an older electron microscope may still produce data volumes in the MB to GB range and process the results using local storage and computation resources, but the ability to use intelligent software to automate the workflow is valuable.

Some NSRCs are located at higher-security facilities, and because of this must use in-house resources. Security compliance can be tricky to manage at a number of BES facilities, and some struggle with the need to accommodate the requirements but also make data accessible. It can be a struggle to migrate even small data sets to users in ways that do not violate our security requirements. This event has resonated by showing lots of national-scale resources are available, and it is a worthwhile exercise to explore available options to leverage HPC facilities and ESnet for future collaborative efforts.

6.3.1.3 Computation That Has Real-time, Near Real-time, or Offline Requirements

BES community members are considering many different approaches to their computational needs. An area of emerging interest for a number of years has been how computational power can be placed at or near the detector, particularly to refine the part of the workflow that performs data reduction. Using FPGAs (such as Xilinx or Altera) to do some of this work is of interest, but requires expertise that many in the community do not have readily available. In addition to data reduction, the aforementioned use cases that are employing AI and ML to steer experimental direction will leverage some mixture of local or edge-computing resources. Given the special nature of this work, namely using very customized training data that could have a large volume, physically close resources may be preferred. ESnet maintains an FPGA testbed, is also experimenting with edge-computing approaches, and will be working to figure out ways to deploy this resource close to sites so that it may be possible to perform similar functions for facilities and science experiments that could benefit.

Other ways to solve the problem of computational availability will be to allocate other forms of computational resources (e.g., CPU or GPU) at the facilities, or to leverage the multifacility workflow model by leveraging ESnet and DOE HPC facilities. The latter use case raises a problem that is not quite solved: how can real-time needs for an instrument preempt the operation of a shared resource such as those found at the HPC facilities. Some BES community members are working with the HPC facilities to implement ways to have real-time critical jobs be advanced through the scheduling system such that they can meet deadlines to ensure experimental success. This is critical since delivering the information fast is a requirement for the use case. The aforementioned DOE ASCR IRI activity has identified this use case in their investigation as well; a number of communities are interested in being able to address real-time computational needs.

As mentioned in previous topics, the BES community lacks the subject-matter expertise to explore new ways to address computational challenges. The use of FPGAs, or the manipulation of job queues, requires expertise that takes time to build. The BES community leverages HPC expertise from each of the facilities when possible, but finds it hard to grow this within the community of users who also know the underlying science use case.

Also revisited was the area of friction surrounding HPC allocation. As stated before, HPC allocations are not automatic, and must be applied for on cycles. When they run out, it can sometimes be hard to request more. Having a more stable way to rely on HPC resources at a given facility (or several interchangeable facilities) is

desirable. BES would like to explore just having a permanent allocation for use, instead of having each facility (or PI) have to request each time. In the future, BES may have fewer dollars to spend on storage and computation. Thus having a permanent way to address computing challenges will alleviate some of this pressure on the workflow. It will also lend itself to simplified use cases, using standard APIs, and less bespoke ways to address the overall challenge of data reduction, analysis, or storage approaches.

A last area of conversation involves the emerging streaming data use cases for analysis. One concept that has stopped BES, as well as a number of other communities, has been the challenges in having HPC worker nodes stream data directly from a remote repository. The typical use case has always been to perform a bulk-data movement to scratch storage, since worker nodes do not often have the ability to perform a wide area transfer directly. Research by DOE HPC facilities to enabling different approaches that may help this problem is ongoing and includes ways to indirectly fetch data (via the Cray Slingshot Interconnect), deploying in-network caches to bring data closer to HPC centers, or creating overlay networks that link resources on private VPN networks. These approaches are being explored by other DOE communities, including facilities and experiments from HEP.

6.3.1.4 Facility Upgrades and Potential Changes to Data Volumes and Rates

The majority of facility case studies indicate growing data volumes in the coming years as facility upgrades and new instruments are installed. The new generation of detectors is both faster and more accurate, which leads to the data they produce growing at exponential scales. With larger and more accurate data comes a need to increase network speeds and increase the available computing power at HPC facilities that serve the experiments.

LCLS at SLAC will experience significant data volume growth of the raw data capabilities (prior to any attempts to reduce the data sets). Needing Tbps capabilities in the coming years due to these upgrades is expected. SLAC is in the process of identifying the needs and timelines to increase its ESnet connectivity to reach NERSC for the analysis workflow. SLAC is in discussions with ESnet to upgrade the connectivity from two 100 Gbps connections in the short term. In the medium term, SLAC and ESnet will work to implement 400 Gbps connections. SLAC will continue to implement local upgrades, including connecting the new Shared Data Facility when the building is complete.

To better align upgrades of technology components in the future, understanding how the relationship between upgrade capabilities (e.g., detector area captured, readout rate, etc.) will influence the requirements for networking, CPU memory and bandwidth requirements, and overall storage volumes is useful. Currently there is not a one to one relationship between any of these items; the BES community notes that any technology upgrades that can be provided by ESnet or the HPC facilities will ultimately improve the scientific process until the next bottleneck is reached.

BES facilities and experiments commonly learn how new technology will benefit them, and then increase their productivity accordingly. For example, if computation doubles, they may try to double how much experimentation they accomplish in the same window. The same can be said for networking: with the availability of faster and more predictable networks, addressing data mobility between facilities will become easier, and people may send more data. This scalability will have limitations however, based on the previous discussions on workforce availability. A safe assumption will be that scientific output will track closely with technology upgrades, and the overall quality of experimentation will increase as well. For example, being able to validate a hypothesis with double the amount of data will strengthen the findings.

Another area of discussion brings up the nature of understanding both the successes of research (e.g., the published results), as well as the failures that lead to the successes. For example, NSRCs engaged in the robotic synthesis of materials may learn valuable lessons about approaches that do not work. Keeping accurate records of failure, and being able to consult them before starting new experiments, is valuable to the overall process of science.

6.3.2 Focus Group 2

The polling during the meeting produced the following discussion topics of interest to the assembled group:

- Computation that has real-time, near real-time, or offline requirements.

- Challenges in supporting multi- or coupled facility workflows.

- Data storage locality, quantity, and mobility.

- Supporting "remote" control and operation.

During this period of discussion, several notable items were brought up, as described in the following sections.

6.3.2.1 Computation That Has Real-time, Near Real-time, or Offline Requirements

The ESnet team started this topic area seeking some clarification regarding the major types of computation workloads, and suggested hardware types that will support autonomous experiment steering. From the submissions, and discussions with authors, there are two "real time" requirements to consider going forward: one for the AI and ML components that will send feedback to the physical (or virtual) scientific instrumentation, and another that can be used to provide real-time analysis and visualization for operators and users of the instruments. Discussion of these two use cases revealed that edge computing (e.g., anything with less than milliseconds of latency and response time) will be required to support AI and ML needs during operations, while the analysis to support human users can be slower (e.g., seconds of latency and response time) and will continue to be provided by centrally managed computation within a facility, or in some cases at a coupled HPC center. Participants noted that in practice the most successful implementations have been with dedicated CPU and GPU resources to support a tight "feedback loop" for the AI and ML use cases, and a slightly slower (and lower fidelity) workflow that used institutional computation to support visualization or longer-term analysis.

Operating these forms of computation comes with certain costs and risks. The real-time requirements placed on the AI and ML components imply that CPU or GPU resources must always be available, with almost instantaneous access required during the time of experimentation. Sharing this resource with several simultaneous uses could be challenging if the wait time grows beyond what is acceptable to keep the autonomous steering pipeline productive. Of equal consideration is the role of cybersecurity: edge-computing resources used by multiple parties must be trustworthy and verifiable during times of operation. Given the expected larger volumes of data from instruments and simulations in future years, designing computational infrastructure directly engaged with networking resources is the suggested way forward to ensure an efficient path to process data in a secure and predictable fashion.

The institutional computing requirements to support visualization and analysis workflows (either locally or remotely) are more well-known and accepted by the BES community, but the use of edge computing is still emerging as a new paradigm to support. ESnet inquired more about how the use of edge computing will become more integrated into the standard workflows in the coming years; in particular if it is anticipated that edge resources will become highly customized to specific use cases (such as steering a specific instrument or analysis), or they will be general purpose to facilitate a wider variety of potential users.

Most BES research on computation centers on the development of a generalized platform that can run an environment (e.g., a container with a set of microservices), which is then available for rapid deployment on demand. The generalized nature ensures that the environment has many options for deployment, but may not be as fast or specialized to a specific resource, which results in some loss of efficiency. Future development will explore the other possible direction: investing time into building custom hardware and software; particularly if it results in significant gains in efficiency for a given facility or instrument (e.g., compression, feature extraction, decision support analysis).

A future vision centers on the notion of a seamless experience where data flows freely from an instrument or simulation directly to the AI and ML components, to minimize the latency and increase the overall response

time. Beyond autonomous experimental steering, the work to support the creation of facility and instrument digital twins, and the research into quantum computing, will also leverage edge-computing resources to support AI and ML workloads. The BES community is moving toward a model where simulation and experimental data will both become input to advanced algorithms that will help to guide the direction of physical or virtual experimentation outcomes. It was noted that within the DOE IRI program, a number of use cases and subgroups asked a similar question to determine if a unified API for generalized edge computing would be the most efficient way to utilize the resources, of if these concepts are best seen as bespoke to individual users. Participants noted that even within the concept of the edge, distinctions may be made between "near" and "far" edge, where one is extremely specific, and the other is more generalizable.

Lastly, discussion returned to cybersecurity topics, specifically flows that related to the implementation of policies on access, identity, and permissions for the components of a modern BES workflow. Participants noted that for multifacility workflows to become more standardized, thought should be given to how independently managed resources and users can be integrated in a fashion that does not reveal their disjointed nature. For example, if a user at a BES beamline is to embark on a multifacility workflow, the following barriers often exist:

- The user may be from the parent laboratory, but often is not, and must go through steps to establish a local operating identify to access facility instrumentation, computation, and storage permissions prior to experimentation start.

- During experimentation, the user can utilize the aforementioned resources with a single identity and permissions, but may encounter friction if computation is required beyond the facility: use of a DOE HPC resource has another process to apply for an allocation, and receive a second identify.

- Creating workflows that link the two facilities will now involve the stitching together of multiple accounts and permissions, and may not be fully automated if restrictions are placed on data mobility between the facilities.

- Lastly, operating computational and storage workloads at one DOE HPC resource does not guarantee the same components will function the same at others. No common operating environment or workflow is available. This is another conversation ongoing in the ASCR IRI activity.

These considerations are large barriers to the adoption of autonomous steering approaches because they impart friction into a process that requires agility. It is often the case that DOE facilities (e.g., BES or ASCR) do not share common definitions of security and policy, which complicates the design and prototype development activities discussed during the requirements review. While the participants do not have all of the answers, the common ground of requiring a baseline form of AAA implementation that spans facilities is desired, and it should apply at a facility, instrument, computation, storage, and data mobility level to ensure friction-free operation. This will prevent circumvention of security, and allow for a more automated environment so the community can move away from offline forms of operation (e.g., the use of removable media and shared login credentials).

6.3.2.2 Challenges in Supporting Multi- or Coupled Facility Workflows

Pivoting away from computational topics, the conversation gradually focused more on the challenges that are facing multifacility operation. In addition to the inherent friction that comes from applying disjointed security policies and practices, the adoption of tools that are agnostic to location of individual components has been crucial to furthering these advanced workflows. Participants cited Jupyter notebooks, which can be easily deployed as a container into many forms of centralized, institutional, and local compute environments, as an important step to empowering users to adopt automated workflows during experimentation. Users that are not savvy with computing or coding can now use standard templates (https://neutronimaging.pages.ornl.gov/tutorial/notebooks/, https://analysis.sns.gov) to accomplish a number of standard experimental workflow portions: real-time analysis during an experiment, data mobility to off-site locations, and post-experimental analysis. A number of the BES

facilities have utilized this approach, particularly during the pandemic years, since it is easy to support remotely and is portable to a number of other experimental and HPC facilities. Some users noted the wide variety of other tools that support Jupyter, including Papermill[1], which has allowed users to link together multiple notebook instances and build more complex workflows. Some participants in the focus group noted that the next logical step for the concept of an interactive notebook initiated by the user is one tied specifically to an instrument. This would further eliminate the need for individual users to be subject-matter experts in programmatic concepts, and would allow them to focus more on the science outcomes, versus having to constantly construct the workflow at the start of each experimental use case.

The Jupyter notebooks do have some friction that was noted by participants. Watching tutorials, particularly for some users, is a necessary step that may not be undertaken prior to the first visit of a facility. Additionally, the tools could be simplified even more with enhancements to the GUI to clarify certain aspects. Lastly, the underlying software toolchain that supports the tools (e.g., Python) can be a delicate environment that requires special care when upgrading to not break functionality. A future consideration for the effectiveness of this tool will be scalability as the data volumes from instruments increase. For example, handling data sets of MB to GB sizes is typically within the boundaries of most Jupyter deployments, but approaching the TB and PB range will be a challenge for the underlying hardware environments. Future iterations of Jupiter may require adaptation to fetch data on demand versus complete importation, along with more intelligent APIs that can account for caching and prefetching of important data regions to not focus as much on the issue of data locality. The work on Bluesky and Tile is addressing some of these concerns.

This discussion topic finished in reviewing some of the challenges to regular use of DOE HPC resources in the future of BES experimentation. Previous exploration has seen this work go from concept, to pilot, to near regular use cases that have shown experimental and cost efficiency in the coupling procedure. BES participants do have future concerns, though, particularly with regards to budgets: as their data and computation needs grow, it is harder to purchase, deploy, and maintain local computing; thus the appeal of using DOE HPC resources is growing. The allocation process for users remains competitive, so considering other approaches where a facility applies for an overarching allocation that can be distributed to users in a more uniform fashion may make sense. This is a different model than some DOE HPC facilities are used to, and may challenge some notions of use and cost. The alternative would be for BES (or other programmatic areas) to continue to provide their own computing, which over time could be more expensive in terms of capital and operating expenditures. As the requirements for computation grow there will be some friction to work out, particularly for the use cases that require specialized resources (e.g., GPUs to support AI and ML). The participants are in agreement that all forms of computing available will be leveraged now, and in the future. Standardizing the offerings across the DOE ecosystem is also recommended to ensure that workflows are portable, and resources can be used uniformly.

6.3.2.3 Data Storage Locality, Quantity, and Mobility

A common theme across all of the BES case studies was the growth of experimental and simulated data sets, and the strain this will put on data storage requirements in the coming years. ESnet started this topic area by inquiring if general consensus existed on how BES facilities are handling data storage: attempting to store all data in perpetuity, putting in place sensible policies to cycle out the oldest data as space becomes limited, or time-based allocations that delete based on age alone. Participants noted that all are possible, and no standard exists based on facility. Smaller facilities with limited resources often follow the latter two approaches. Larger facilities, especially those with long-standing relationships with DOE HPC facilities, typically try to hang on to data forever, even if it means migrating to slower offline storage. In the general case the cost of storage has dropped, but the curation steps remain expensive (e.g., staff time to monitor and migrate data, along with updating catalogs). In the general case, most instruments and facilities have the ability to store multiple TB of data on live storage, and can then filter older data to institutional or offline storage as needed. Others noted extensive R&D efforts for new machines to develop more sophisticated data filters to reduce the footprint of raw data.

---

[1] https://papermill.readthedocs.io/en/latest/

Trying to establish a well-known centralized location, possibly a DOE HPC facility, that can serve as permanent storage was raised. In practice this idea has been discussed before, but faces some barriers. Export controls and other security policy differences between facilities remain large hurdles to a universal solution. A secondary problem is the lack of a unified data storage and metadata cataloging framework; these issues remain challenging since automating this is often hard, and experimental users often skip this step in the rush to acquire and publish research results. Some advancements, such as the standardized use of storage and computation for all instruments in a facility, have furthered the challenge of proper categorization, but there is a long way to go to span all of the BES ecosystem.

6.3.2.4 Supporting "Remote" Control and Operation

The last topic was remote operation for BES facilities. The pandemic accelerated this use case across the entire ecosystem, and in locations where it was previously unavailable, it is now a first-order operational mode that will remain for years to come. Several participants offered their individual experiences, which can be summarized as:

- The initial lift of enabling remote viewing was challenging from a security standpoint for some facilities. Operating remote viewing tools (e.g., NoMachine, VNC) had to be done in a careful fashion to abide by institutional security policies.

- The aforementioned issue with managing multiple identities also remains, and was made harder when users could not visit a facility to receive all credentials.

- Some instruments will never be fully remote operated, which implies that users must still coordinate with local staff during experimentation. This places a lot of emphasis on the power of remote collaboration tools (e.g., Zoom, WebEx, Bluejeans, Teams).

This section completed with additional discussion on authentication, in particular the adoption of multifactor procedures, and how the user population has responded to their requirement and implementation. MFA has become more widespread, and is used in a number of commodity applications (e.g., banking, use of commercial email services); thus its use for instruments at DOE facilities is not unheard of. The complications still come from having to use multiple MFA systems, for multiple identities, in the course of doing a single experiment. There is hope that the ASCR IRI discussions will help to describe and codify requirements and implementation strategies in the future.

List of Abbreviations

ACI	application centric infrastructure
AI	artificial intelligence
ALCF	Argonne Leadership Computing Facility
ALS	Advanced Light Source
ANL	Argonne National Laboratory
APS	Advanced Photon Source
APS-U	Advanced Photon Source Upgrade
ARI	ARPES and RIXS Imaging
ASCR	Advanced Scientific Computing Research
ASN	autonomous system number
AWS	Amazon Web Service
BAG	Block Allocation Groups
BCP	best common practice
BES	Basic Energy Sciences
BLAST	Bridging Length/Timescales via Atomistic Simulation Toolkit
BNL	Brookhaven National Laboratory
BTR	beam time request
CADES	Compute and Data Environment for Science
CAMERA	Center for Advanced Mathematics for Energy Research Application
CAT	collaborative access teams
CCD	charge coupled device
CCTBX	Computational Crystallography Toolbox
CDI	coherent diffraction imaging
CDN	Content Distribution Network
CFN	Center for Functional Nanomaterials
CFS	Community File System
CHX	Coherent Hard X-ray
CINT	Center for Integrated Nanotechnologies
CMOS	complementary metal–oxide–semiconductor
CMS	complex materials scattering
CNM	Center for Nanoscale Materials
CNMS	Center for Nanophase Materials Sciences
CPU	central processing unit
CryoEM	cryogenic electron microscopy
CSI	Computer Science Initiative
DAC	direct air capture
DAQ	data acquisition
DCI	Data Center Interconnect
DFT	density-functional theory

DISCUS	Distributed Infrastructure for Scientific Computing for User Science
DMFT	dynamical mean-field theory
DMRG	density matrix renormalization group
DOE	Department of Energy
DOE SC	DOE Office of Science
DOI	digital object identifier
DTN	Data Transfer Node
EB	exabyte
eBGP	External Border Gateway Protocol
ECXF	Equinix Cloud Exchange Fabric
ENOS	ESnet R&D into network operating systems
EPICS	Experimental Physics and Industrial Control System
ESCC	ESnet Site Coordinators Committee
FAIR	findability, accessibility, interoperability, and reusability
FCFS	first come first served
FLOPS	floating point operations per second
FPGA	field-programmable gate array
FPMD	first principles molecular dynamics
FTP	file transfer protocol
FTS	first target station
FY	fiscal year
GISAXS	grazing incidence small angle scattering
GPFS	General Parallel File System
GPU	graphical processing unit
GU	General User
GUI	graphical user interface
HDD	hard disk drive
HEC	High End Computing
HEDM	high-energy diffraction microscopy
HEP	High Energy Physics
HEX	high energy engineering X-ray scattering
HFIR	High Flux Isotope Reactor
HPC	high-performance computing
HPCS	High Performance Computing Services
HPSS	high performance storage system
HSCT	high-speed computed tomography
HTC	high throughput computing
HTSN	High Throughput Science Network
ICNM	In-situ Characterization and Nanomechanics
IDF	instrument definition files
IOC	input/output controllers

IOTA	Integration Optimization Triage and Analysis
IP	Internet Protocol
IR	Infrared
IRI	Integrated Research Infrastructure
IT	Information technology
LAN	local area network
LANL	Los Alamos National Laboratory
LBNL	Lawrence Berkeley National Laboratory
LCF	Leadership Computing Facility
LCLS	Linac Coherent Light Source
LDAP	lightweight directory access protocol
LEA	lab execution and analysis
LHCONE	Large Hadron Collider Open Network Environment
LIMS	Laboratory Information Management System
LN	liquid nitrogen
LSDCSC	Light Source Data and Computing Steering Committee
LSTM	long short-term memory
MB	megabyte
MBA	multibend achromat
MCP	multi-channel plate
MD	molecular dynamics
MDF	Materials Data Facility
MFA	Multi-factor authentication
ML	machine learning
MOF	metal–organic frameworks
MP	Materials Project
MREN	Metropolitan Research and Education Network
NAS	network attached storage
NCEM	National Center for Electron Microscopy
NERSC	National Energy Research Scientific Computing Center
NFS	network file system
NIH	National Institutes of Health
NISQ	noisy intermediate-scale quantum
NIST	National Institute of Standards and Technology
NPON	Nanophotonics and Optical Nanomaterials
NScD	ORNL's Neutron Sciences User Office
NSF	National Science Foundation
NSRC	Nanoscale Science Research Center
OLAP	online analytical processing
OLCF	Oak Ridge Leadership Computing Facility
OPI	operator interface

ORNL	Oak Ridge National Laboratory
OSCARS	On-Demand Secure Circuits and Advance Reservation System
PASS	Proposal Allocation, Safety, and Scheduling System
PB	petabyte
PCA	principal component analysis
PFLOPS	petaflop, which is 10^{15} floating point operations per second
PI	principal investigator
PIN	personal identification number
PRP	Proposal Review Panel
PU	Partner Users
QMC	Quantum Monte Carlo
QMS	Quantum Materials Systems
R&E	research and education
RA	Rapid Access
RDMS	research data management systems
RDP	remote desktop protocol
RIXS	resonant inelastic X-ray scattering
RT	room temperature
SAC	Scientific Advisory Committee
SAN	storage area network
SCS	SLAC Computing Services
SDCC	Scientific Data and Computing Center
SDK	software development kit
SFTP	secure file transfer protocol
SFX	serial femtosecond crystallography
SLAC	SLAC National Accelerator Laboratory
SNAP	SNS Neutrons And Pressure
SNL	Sandia National Laboratories
SNS	Spallation Neutron Source
SOA	service-oriented architecture
SPEAR	Stanford Positron Electron Asymmetric Ring
SPI	single-particle imaging
SPICE	Spectrometer Instrument Control Environment
SRCF	Stanford Research Computing Facility
SRX	submicron resolution X-ray
SSD	solid state drives
SSH	secure shell protocol
SSRL	Stanford Synchrotron Radiation Light
STEM	scanning/transmission electron microscopes
STS	second target station
SUF	Scientific User Facility

SUFD	Scientific User Facilities Division
SXN	soft X-ray nanoprobe
TB	terabyte
TEAM	transmission electron aberration-corrected microscope
TEM	transmission electron microscope
TES	thermal emission spectrometer
TMF	The Molecular Foundry
ToR	top of rack
UEC	Users' Executive Committee
UEM	ultrafast electron microscope
UI	user interface
UQ	uncertainty quantification
VENUS	versatile neutron imaging instrument
VM	Virtual machine
VPC	virtual private cloud
VPN	virtual private network
VQE	Variational Quantum Eigensolver
VUV	vacuum ultraviolet
WAN	wide-area network
XAS	X-ray absorption spectroscopy
XES	X-ray emission spectroscopy
XFEL	X-ray free electron lasers
XPCS	X-ray photon correlation spectroscopy
XRF	X-ray fluorescence