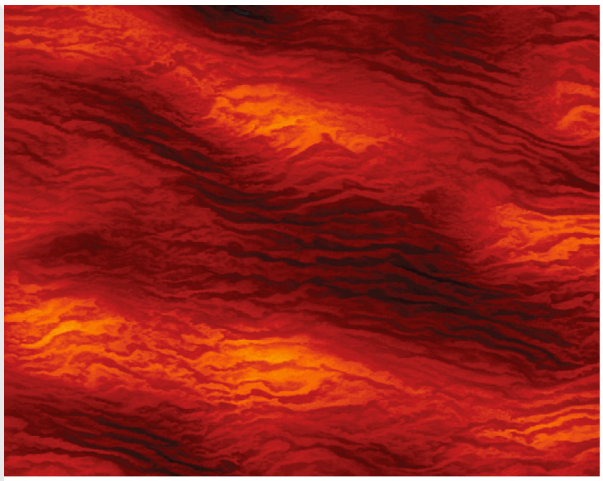
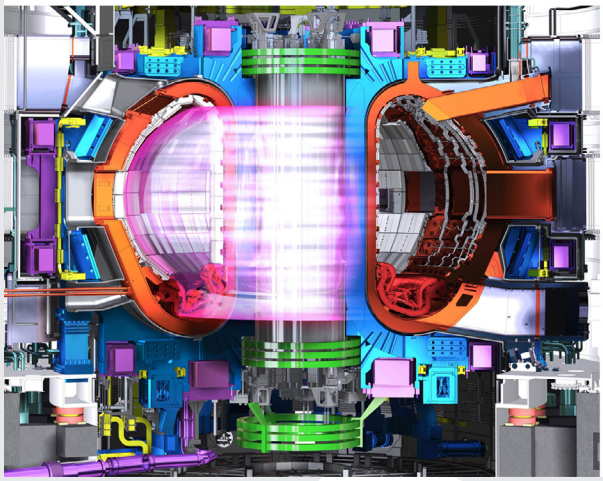
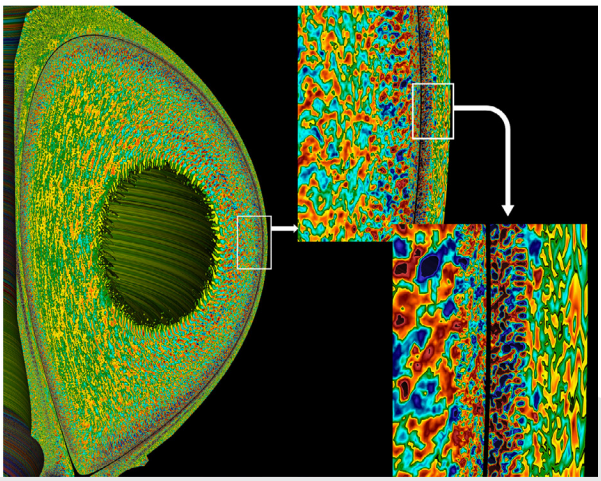
ESnet
ENERGY SCIENCES NETWORK

Fusion Energy Sciences Network Requirements Review

Final Report
April – October, 2021



BERKELEY LAB

U.S. DEPARTMENT OF ENERGY | Office of Science

ESnet
ENERGY SCIENCES NETWORK

Fusion Energy Sciences Network Requirements Review

Final Report
April – October, 2021

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

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

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

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

This is LBNL report number: LBNL-2001462

Cover Images:

Upper left: gyrokinetic simulation: MIT PSFC

Upper right: General Atomics

Lower left: Credit © ITER Organization, http://www.iter.org/

Lower right: Figure caption: Turbulence in edge plasma: PPPL, ALCF, and OLCF

Participants and Contributors

Fatema Bannat Wala, Lawrence Berkeley National Laboratory and Energy Sciences Network

Paul Bonoli, MIT Plasma Science and Fusion Center

Ben Brown, Department of Energy Office of Science

CS Chang, Princeton Plasma Physics Laboratory

Michael Churchill, Princeton Plasma Physics Laboratory

Brandon Cook , Lawrence Berkeley National Laboratory and the National Energy Research Scientific Computing Center

Doug Curry, Oak Ridge National Laboratory

Eli Dart, Lawrence Berkeley National Laboratory and Energy Sciences Network

Ahmed Diallo, Princeton Plasma Physics Laboratory

Bill Dorland, Princeton Plasma Physics Laboratory

Frederico Fiuza, SLAC National Accelerator Laboratory

Mark Foster, SLAC National Accelerator Laboratory

Richard Gerber, Lawrence Berkeley National Laboratory and the National Energy Research Scientific Computing Center

David Green, Oak Ridge National Laboratory

Chin Guok, Lawrence Berkeley National Laboratory and Energy Sciences Network

Walter Guttenfelder, Princeton Plasma Physics Laboratory

Susan Hicks, Oak Ridge National Laboratory

Saswata Hier-Majumder, Department of Energy Office of Science

Jerry Hughes, MIT Plasma Science and Fusion Center

James Kafader, Lawrence Berkeley National Laboratory and Energy Sciences Network

Scott Kampel, Princeton Plasma Physics Laboratory

Stan Kaye, Princeton Plasma Physics Laboratory

Josh King, Department of Energy Office of Science

Cornwall Lau, Oak Ridge National Laboratory

John Mandrekas, Department of Energy Office of Science

Orso Meneghini, General Atomics

Bill Miller, Department of Energy Office of Science

Ken Miller, Lawrence Berkeley National Laboratory and Energy Sciences Network

Inder Monga, Lawrence Berkeley National Laboratory and Energy Sciences Network

Raffi Nazikian, Princeton Plasma Physics Laboratory

Jeff Nguyen, General Atomics

Francesca Poli, Princeton Plasma Physics Laboratory

Juergen Rapp, Oak Ridge National Laboratory

Katherine Riley, Argonne National Laboratory and Argonne Leadership Computing Facility

Lauren Rotman, Lawrence Berkeley National Laboratory and Energy Sciences Network

Steve Sabbagh , Columbia University and Princeton Plasma Physics Laboratory

Brandon Savage, MIT Plasma Science and Fusion Center

David Schissel, General Atomics

Douglass Schumacher, The Ohio State University

Laurie Stephey, Lawrence Berkeley National Laboratory and the National Energy Research Scientific Computing Center

Josh Stillerman, MIT Plasma Science and Fusion Center

Andrew Wiedlea, Lawrence Berkeley National Laboratory and Energy Sciences Network

Dennis Youchison, Oak Ridge National Laboratory

Jason Zurawski, Lawrence Berkeley National Laboratory and Energy Sciences Network

Report Editors

Ben Brown, Department of Energy Office of Science: benjamin.brown@science.doe.gov

Eli Dart, ESnet: dart@es.net

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

Saswata Hier-Majumder, Department of Energy Office of Science: saswata.hier-majumder@science.doe.gov

Josh King, Department of Energy Office of Science: josh.king@science.doe.gov

John Mandrekas, Department of Energy Office of Science: john.mandrekas@science.doe.gov

Bill Miller, Department of Energy Office of Science: bill.miller@science.doe.gov

Ken Miller, ESnet: ken@es.net

Lauren Rotman, ESnet: lauren@es.net

Andrew Wiedlea, ESnet: awiedlea@es.net

Jason Zurawski, ESnet: zurawski@es.net

Table of Contents

1 Executive Summary

About ESnet

The Energy Sciences Network (ESnet) is the high-performance network user facility for the US Department of Energy (DOE) Office of Science (SC) and delivers highly reliable data transport capabilities optimized for the requirements of data-intensive science. In essence, ESnet is the circulatory system that enables the DOE science mission by connecting all of its laboratories and facilities in the 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 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.

This Review

Throughout 2021, ESnet and the Office of Fusion Energy Sciences (FES) of the DOE SC organized an ESnet requirements review of FES-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 their relationship to the FES 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 FES program has two goals: (1) expand the understanding of matter at very high temperatures and densities and (2) build the knowledge needed to develop a fusion energy source. Providing energy from fusion is one of the 14 Grand Challenges for Engineering in the 21st Century[1], and FES is the largest federal government supporter of research that is addressing the remaining obstacles to overcoming this challenge.

Together with its partner science agencies, FES supports a devoted workforce that has made impressive progress since the first fusion experiments over 60 years ago. Progress is made each day by scientists and engineers at DOE national laboratories, at universities, and in private industry. With public financial support for this fundamental research, fusion scientists are undertaking fundamental tests of fusion energy's viability using some of the most ambitious energy projects, the most powerful supercomputers, and the fastest networks in the world today.

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

- International fusion collaborations.
- Remote observation and participation of fusion facilities.
- General Atomics: DIII-D National Fusion Facility.
- MIT Plasma Science and Fusion Center (PSFC).
- Princeton Plasma Physics Laboratory (PPPL).
- Planning for ITER operation.
- Public-private partnerships in fusion research.
- Material Plasma Exposure eXperiment (MPEX) at Oak Ridge National Laboratory (ORNL).

1 https://www.engineeringchallenges.org/challenges/fusion.aspx

- Matter in Extreme Conditions (MEC) Experiment at the SLAC National Accelerator Laboratory (SLAC).
- LaserNetUS Program.
- Multi-facility FES workflows.
- Whole-device modeling (WDM) and FES high-performance computing (HPC) activities.

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 FES, ASCR, and ESnet. These terms are defined as follows:

- **Findings**: key facts or observations gleaned from the entire review process that highlight specific challenges, particularly those shared among multiple case studies.
- **Actions**: potential strategic or tactical activities, investments, or opportunities that 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 FES activities listed previously.
- ESnet Site Coordinators Committee (ESCC) members from FES activity host institutions, including the following DOE labs and facilities: Argonne National Laboratory (ANL), General Atomics (GA), LBNL, MIT PSFC, the National Energy Research Scientific Computing Center (NERSC), ORNL, PPPL, and SLAC.
- Networking and/or science engagement leads from the ASCR HPC facilities.
- DOE SC staff spanning both ASCR and FES.
- 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:

- Preparations for ITER
  - ITER contains over 50 major diagnostic packages, consisting of thousands of data channels, and will eventually produce 2 PB of raw data each day through a gradual increase in capability. ITER will require more than an exabyte of data storage by the mid-2030s, and this estimate does not include the volume of analyzed and simulated data that will also be

produced and archived. ITER will commence operations with much less data production per day (~ 20 TB) during the first phase of plasma operation (engineering commissioning, first plasma, and engineering operations) planned for 2026.

- ITER peak data production rates are not fully known as of 2021. However, aggregate estimates of a 20 TB/day data production rate have been made for the engineering operations phase. The ITER timeline, as of 2021, is as follows:
    - First plasma: Dec 2025.
    - Additional commissioning and construction: through Dec 2028.
    - Pre-fusion power operations (Phase 1): Dec 2028 through Jan 2030.
    - Pre-fusion power operations (Phase 2): June 2032 through Mar 2034.
    - Nuclear assembly: 2035.
    - Regular operations: Dec 2035.
- In present fusion facilities, a typical experiment is a collection of similar discharges executed over a single day or partial day, with each discharge typically lasting < 10 s. Initially, discharges in ITER will be of similar duration per pulse, but with the goal of reaching 300 s. by the mid-2030s. However, unlike existing experiments, ITER may run experiments over multiple days.
- Development and implementation of the policies and infrastructure that support data sharing is a crucial need for the FES community in preparation for ITER experimentation. Having access to that data in a timely manner is critical to advancing research and development activities, as well as remote participation in ITER operation.

- Scientific Data Management
    - As superconducting international experiments achieve truly long-pulse operation (> 100 s), it is important that ESnet provide the connectivity needed for the US fusion community to effectively access data from facilities around the world by contributing to secure trusted high-throughput data pipelines between these major international experiments and US hubs that can store and distribute the data and analysis capability to registered US collaborators.
    - The FES community has nearly adopted approaches where computation occurs as close to the experimental data storage as possible, typically the same location where the instrument is located. This approach, often called "edge computing," does not require experimental data to be moved from an instrument location to a remote HPC environment. It does create situations where a user, who may be representing a third location, bypasses their own home institution's computational capabilities when performing analysis. Edge computing may change the geometry of a workflow, depending on the location of resources and scientists in a network topology.

- Scientific Workflow
  - The time between experimental shots in magnetic fusion energy (MFE) tokamak experiments is critical to the overall workflow, placing extreme emphasis on network reliability and performance. Recent developments to the overall efficiency of the process mean less time to react and influence experimental direction. Networking is a critical component of a distributed workflow, and ESnet partners with the US fusion community to effectively access data from facilities around the world by developing secure trusted high-throughput data pipelines between these major international experiments and US hubs that can store and distribute the data and analysis capability to registered US collaborators.
  - The overall operation time of GA's DIII-D tokamak will remain similar for the next five years, and it is anticipated that the rate of acquiring new data will continue to increase. From 2010 to 2020, the total amount of DIII-D data increased by an order of magnitude.
  - MIT PSFC's Alcator C-Mod data archive is approximately 150 TB in size and remains heavily accessed by the FES community.
  - Gyrokinetic simulation will be a major research element during the exascale era of computation. Execution of this simulation at DOE HPC centers has the potential to produce data volumes beyond what the current generation of computing and storage is capable of handling. Effort to reduce data size is therefore required before results can be stored locally or transferred from ASCR HPC centers back to PPPL. Additionally, only some portions of the output can be viewed remotely due to the size of the data sets and the responsiveness of interactive tools that can be used to visualize.
- Remote Collaboration
  - Commercially available collaboration tools that support communication functions such as audio, video, and text chat (e.g., Discord, Zoom, etc.) are critical to the process of science for FES experiments and facilities. This trend started years prior to the COVID-19 pandemic, remained crucial for ongoing operation during the pandemic, and will remain a part of operation into the future provided the tools perform well and local staff support the use case. Enabling these tools through network peering relationships (directly and via cloud providers) is important for collaboration.
  - The FES community has adapted remote observation and participation use cases over time and found a number of software tools that work well, along with a number that are still challenging to use due to design or operational considerations. Some of these are commercial, others may be open source. X Window System, VNC, NoMachine, and others that allow for the ability to view, and occasionally control, remote resources often conflict with information security requirements. Performance of these tools, particularly over great distances, depends heavily on network latency and available bandwidth, both of which are hard to control on busy commodity or institutional networks. Future remote observation

and participation approaches will demand tools that offer similar feature sets, along with ways to validate and ensure network performance on an end-to-end basis.

- Multi-facility Computational Workflows and Use Cases
  - The ability to access live data streams from FES experiments will become necessary in the coming years, particularly as experimental facilities more routinely couple to collaborating computing facilities. This multi-facility model will require advanced software to link experimental resources to storage and computing via the network infrastructure. The increased collaboration will alleviate existing areas of friction, provided resources can be made available to operate at real time during experimentation.
  - The FES community has long relied on a computational paradigm that encourages the use of a single location: this may be a dedicated pool at an experimental facility, a local cluster, a DOE HPC center allocation, or a commercial facility. Unfortunately, workflows are typically designed to use only one locality, and may be affected if computational resources are not readily available at the specified location. A more efficient approach would be to pursue using computational resources in multiple locations simultaneously, even if it implies having to migrate data, or computational jobs, away from a preferred location. To distribute and manage computational demand and data mobility requirements, more standardization and resource sharing across the FES complex will be needed. Intelligent tools could be designed to be made aware of options, and better spread analysis to the available resources.
  - DOE HPC allocations for FES are subject to annual renewal, and this causes challenges for strategic planning or long-term investments in a particular computing capability or workflow architecture. If renewing at the same location is not possible, this often leads to complications in data and workflow migrating to alternate facilities: adapting software to run on different systems, granting accounts to existing users, and sending a majority of scientific data. Unified APIs and simplified methods to manage data between DOE HPC facilities could simplify the friction seen in these scenarios.
  - As the FES community prepares for ITER, the ability to leverage resources across the DOE SC landscape in a multi-facility paradigm (e.g., DOE HPC resources, analysis facilities, distributed users, all linked via ESnet) will become more important as data volumes far exceed the storage and processing capacity of any single location that participates in FES science. This integration of FES experimental facilities with that of DOE HPC resources via ESnet is critical to the success of the ITER collaboration. Exploring Science DMZ architectures at all FES facilities will be required to ensure that a baseline for data mobility can be achieved.
- International & Domestic Networking
  - FES research, development, and operational activities rely heavily on international connectivity provided by ESnet. The coming years will

see the commissioning of new experiments and the addition of new collaborators, and increases in data volume that will place particular emphasis on the reliability and capacity for ESnet's international connections to Europe, and peering relationships with providers that reach other parts of the world (e.g., the Asia-Pacific region, South America, and Africa).

  - The Experimental Advanced Superconducting Tokamak (EAST) in Hefei, China, is a significant international facility used by the US fusion research community. Operational considerations, such as data mobility to and from this facility, rely on the IPv6 communications protocol because it affords higher levels of performance. Ensuring IPv6 peering across ESnet infrastructure, and with international partners, is critical to the process of science for these interactions.
  - The FES community requires stable connectivity to a number of cloud-based communication services that facilitate the community's remote participation and collaboration use cases. These include, but are not limited to, audio and video conferencing (e.g., Discord, Zoom, etc.). ESnet provides critical paths to these commercial services.
  - ESnet connectivity is operationally critical for a number of FES facilities. Topological network backups, as well as capacity augmentations, will be required in future years to ensure continuous operation. Each FES facility relies on the ESnet connection to support research and education (R&E) connected activities domestically and internationally, as well as commercial peering to critical storage, audio, and video services that are used during the process of science.

- Software Infrastructure
  - Software licensure and import/export controls can complicate scientific workflows, particularly if approaches that are designed for single user/machine use cases are adapted to shared environments, such as an HPC facility. For example, a user of a shared resource often does not have the administrative rights to install and operate software that may require these permissions. This can prevent critical software from being run on resources that would accelerate the workflow, and prevent productivity for the process of science.
- Cybersecurity
  - FES workflows that span facilities (either experimental site to user, or experimental site to HPC facility) struggle with mechanisms to share and automate credential exchange required by cybersecurity policies. Such credential exchanges are common for data migration and analysis workflow tools. Improving the flexibility of FES workflows to use resources at other facilities will require modification of software mechanisms to cope with security requirements.
  - Remote collaboration within the FES community has unique cybersecurity requirements that affect current and future use cases. In particular, the requirements to support remote observation, remote participation, and remote control of any given experiment will dictate

the implemented security posture. Large international efforts such as ITER, which features 35 countries in collaboration, will challenge the implementation of baselines due to administrative and national boundaries that are involved. Particular focus will be given to account management, collaboration tools, and controls placed on data export.

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

- Preparations for ITER
  - The FES community will experience unprecedented data volumes in the coming years due to new experimental designs and changes to workflows that place heavy emphasis on networking to link distributed data, processing, and collaborators. It is recommended that the community consider starting a set of "data challenge" activities to support a number of use cases, which will prepare experiments and facilities for increasing data volumes and reveal gaps in the way that hardware and software are able to cope with the future readiness requirements.
  - It is estimated that the following wide-area networking requirements for different milestone years based on current projections to support the international community. ASCR, FES, and ESnet should evaluate these outbound requirements at the facility, and consider them when designing peering with GÉANT, or connectivity across the existing DOE transatlantic strategy:
    - 2023: 20 Gbps
    - 2027: 200 Gbps
    - 2031: 500 Gbps
    - 2035: 1.5 Tbps
- Scientific Data Management & Workflow
  - A number of current FES community approaches to the handling and management of scientific data could benefit from the experience gained by a cross-section of other DOE SC areas. It is recommended that collaborative groups begin the process of discussing scientific workflow and software support for FES data and networking preparedness at FES collaboration sites as ITER is commissioned.
  - ESnet will work with laboratories, sites, and FES collaborations to explore best common practices (BCPs) related to data architecture and mobility strategies through the Data Mobility Exhibition (DME) and other forms of coordinated "data challenges" within the FES community.
- Remote Collaboration
  - ESnet will work with the FES community to periodically review important remote collaboration tools and their network requirements to ensure that commercial peering and site capacities are matching expectations. Services such as collaboration, audio and video (e.g., Discord, Zoom, etc.), as well as computation and storage (e.g., Google Cloud Project, etc.) are critical to FES remote participation and

observation use cases, and are critical to a number of sites that are only connected to ESnet.

- Multi-facility Computational Workflows and Use Cases
  - FES collaborators are interested in pursuing more multi-facility workflows, provided there is time to share requirements and evaluate their effectiveness. A set of pilot demonstrations is recommended for the FES community, DOE HPC facilities, and ESnet , so that all parties can become more familiar with the process and adopt the procedure as routine.
- International and Domestic Networking
  - ESnet must work with the FES community to understand the international connectivity requirements of ITER, and will work with the French NREN RENATER or the pan-European REN GÉANT to deliver ITER data to US-based collaborators.
  - ESnet will continue to work with sites that host major FES use cases (e.g., PPPL, GA, and MIT PSFC) to investigate ways to augment primary and backup site connectivity options.
- Software Infrastructure
  - The FES community, as it prepares for activities such as ITER, should consider adopting hardware and software approaches that are used by other DOE communities (e.g., high-energy physics [HEP]) to implement a distributed data architecture consisting of a central data producer and numerous collaborators and analysis facilities.
- Cybersecurity
  - Implementation of broad cybersecurity policies can affect the performance of open scientific workflows that rely on data mobility between cooperating facilities. FES and ASCR must work with institutional CIOs and cybersecurity staff to understand the possible impacts, and recommend appropriate mitigations and strategies to afford compliance and protection without affecting performance. This work will influence future collaboration, including multi-facility workflows, and remote use cases that are regularly used in FES.

2 Review Findings

The requirements review process helps to identify important facts and opportunities from the programs and facilities that are profiled. The following sections outline a set of findings from the FES and ESnet requirements review starting in April 2021 and running through October 2021. These points summarize important information gathered during the review discussions surrounding case studies and the FES program in general. These findings are organized by topic area for simplicity and by common themes:

- Preparations for ITER
- Scientific Data Management
- Scientific Workflow
- Remote Collaboration
- Multi-facility Computational Workflows and Use Cases
- International and Transoceanic Networking
- Domestic Networking for Local and Wide-Area Use Cases
- Software Infrastructure
- Cybersecurity

2.1 Preparations for ITER

- The ITER tokamak, located in Cadarache, France, is the most ambitious fusion experiment ever undertaken. ITER is a magnetic confinement device where hydrogen isotopes are heated to temperatures up to 100 million degrees C, forming a plasma and forcing nuclei to fuse to create fusion energy. [Case Study 6]
- ITER brings together 35 nations and 7 major partners (China, the European Union, India, Japan, Korea, Russia, and the United States) to collaborate on this experiment, which will be designed to achieve sustained high-fusion power (500 MW, 500–550-second pulse) by the mid-2030s, and to achieve full steady-state operation thereafter. [Case Study 6]
- ITER contains over 50 major diagnostic packages, consisting of thousands of data channels, eventually producing in excess of 2 PB of raw data each day, with a gradual increase over time. ITER will require more than an exabyte of data storage by the mid-2030s, and this estimate does not include the volume of analyzed and simulated data that will be produced and archived. ITER will commence operations with much less data production per day (~ 20 TB) during the first phase of plasma operation (engineering commissioning, first plasma, and engineering operations) planned for 2026. ITER and the international fusion community will have time to learn and prepare for when peak data is expected in the mid-2030s. [Case Study 6]
- An important design philosophy for ITER analysis is embodied in the Integrated Modeling and Analysis Suite (IMAS) being developed at the

ITER organization (IO) under the guidance of the Integrated Modeling Expert Group (IMEG). The backbone of the IMAS infrastructure is a standardized, machine-generic data model that represents simulated and experimental data with identical structures. [Case Study 6]

- The US is supplying substantial hardware to the ITER facility including some of the superconducting magnets, power supplies, and various other components during the construction phase. In addition, there are seven key scientific instruments for plasma analysis that the US will supply and be responsible for during plasma operations: [Case Study 6]
    - Core Imaging X-ray Spectrometer
    - Electron Cyclotron Emission Radiometer
    - Low Field Side Reflectometer
    - Motional Stark Effect Polarimeter
    - Residual Gas Analyzer
    - Toroidal Interferometer/Polarimeter
    - Upper Infrared (IR)/Visible Cameras
- ITER will generate a range of "simulated" data covering every possible aspect of the ITER experiment beforehand, including first plasma experiments where extensive modeling has already taken place to understand the capabilities and limitations of all the first plasma diagnostics for interpretation and control. [Case Study 6]
- A major change with ITER is that experiments will need to be designed using a hierarchy of models of different physics fidelity in order to maximize the probability of success. A virtual experiment will essentially be created, consisting of models of the control system and vessel, plasma, heating, and diagnostic systems. Every conceivable contingency will need to be assessed and the control parameters adjusted to meet safety and performance requirements. [Case Study 6]
- In present US-based fusion facilities, a typical experiment is a collection of similar discharges executed over a single day or partial day, with each discharge typically lasting less than 10 s. The super-conduction tokamaks in China (EAST) and Korea (Korean Superconducting Tokamak Advanced Research [KSTAR]) have demonstrated operation with 100-second-long discharges. Sometimes an experiment can run over several days, but this is quite rare. Initial physics operation for ITER aims at 50- to 100-second-long discharges (phase I), up to 300-second-long discharges (phase II), and up to 500-second-long discharges (fusion plasmas) in the mid-2030s. Advanced operation targets for ITER could include high neutron fluence (1,000-second-long discharges) and steady-state (3,600-second-long discharges). However, unlike existing experiments, ITER may run experiments over multiple days. [Case Study 6]
- The US fusion community desires a combination of near real-time data during the actual plasma pulse, and then the rapid transfer of the bulk of the scientific data within ~5 minutes or less after the pulse is completed. This

provides opportunities for remote participants to complete essential analysis in time to inform the next pulse or several pulses thereafter in experiments that rely on that feedback mechanism. Today, 100-Gbps network connections at major scientific centers are not uncommon, and thus such a network throughput to the US starting in the first year is reasonable. This could grow in future years as fidelity increases for the produced data sets. [Case Study 6]

2.2 Scientific Data Management

- As superconducting international experiments achieve truly long-pulse operation (> 100 s), it is important that ESnet provide the connectivity needed for the US fusion community to effectively access data from facilities around the world by contributing to secure trusted high-throughput data pipelines between these major international experiments and US hubs that can store and distribute the data and analysis capability to registered US collaborators.. [Case Study 1]
- Development and implementation of the policies and infrastructure that support data sharing is a crucial need for the FES community in preparation for ITER experimentation. ITER will produce an unprecedented amount of data that will be of critical interest to the US FES community. Having access to that data in a timely manner is critical to advancing research and development activities as well as participating in remote operations of ITER. Development and implementation of policies and infrastructure supporting data sharing are a high-priority need. [Focus Groups]
- Heterogeneous data formats are problematic for the FES community and create a lot of work to support and adapt software that can be used at a variety of experimental facilities. Experimental instruments often have different data schema, which complicates creating software for data management and dissemination. This may create problems for future-proofing systems as well as for the creation of operating environment-agnostic software. [Focus Groups]
- Use of the Science DMZ architecture[1], Data Transfer Nodes (DTNs)[2], and the Modern Research Data Portal[3] deployed at participating FES sites is recommended to ensure systemic capabilities for scientific data mobility. These components allow for high-performance operation when supporting data transfer (bulk or streaming) as a part of the operational science workflow. [Focus Groups]
- The FES community has nearly entirely adopted approaches where computation occurs as close to the experimental data storage as possible, typically the same location where the instrument is located. This approach, often called edge computing, does not require experimental data to be moved from an instrument location to a remote HPC environment. The approach does create situations where a user, who may be representing a third location, will bypass their own home institution's computational

1 https://fasterdata.es.net/science-dmz/
2 https://fasterdata.es.net/science-dmz/DTN/
3 https://mrdp.globus.org

capabilities when performing analysis. Edge computing may change the geometry of a workflow, depending on the location of resources and scientists in a network topology Tools such as MDSplus and NoMachine NX facilitate this interaction, and this use case is expected to continue to grow in importance in the future. [Case Study 3, 4, 5]

- The Exascale Computing Project (ECP) WDM codes, once complete, will undergo a period of distributed community analysis. This simulation data will need to be available to the wider community for a minimum of five years to provide the source data that will be used to develop fusion surrogate models and digital twins. [Case Study 5]
- Public-private partnerships with non-DOE entities are funded to perform aspects of FES research; one example of this is the INFUSE program. Many of these entities are unfamiliar with mechanisms to interact with DOE SC facilities including ASCR HPC centers and ESnet. As a part of program onboarding, providing better information on DOE resources available through science engagement may encourage use of these facilities during the process of science. [Case Study 7]
- The MPEX experiment at ORNL is under design, and will be operational by 2027. [Case Study 8]
    - The standard short-pulse use case will produce:
        - An estimated 50 GB of scientific data per run day, with 100 run days per year. This is an estimated 5 TB of data per year.
        - Visible light cameras will be used for measuring the target surface and will produce raw video data streams at 1 GBps. Up to six cameras can be used at various angles during a run period and can generate just under 4 TB of raw data frames per hour, or up to 24 TB per hour if all cameras are operating.
        - A single IR camera can be used for measuring surface materials' interactions, and it is estimated to produce raw data rates at 9 GBps or 32 TB per hour.
        - Lastly, approximately 35,000 archived signals for operational data are stored in a relational database. The archived data consumes approximately 17 GB per day or 6.2 TB per year.
    - A second use case, consisting of a longer pulse (two weeks of continuous operation), has the potential to generate 1 PB of scientific experimental data. The camera rates listed above will apply as well, but will be limited to the two-week operational period.
    - MPEX will expose data via recommended mechanisms that ORNL and Oak Ridge Leadership Computing Facility (OLCF) support (e.g., HTTP portals, RSYNC, secure copy protocol [SCP]).
    - MPEX is designing an experimental workflow between the instrumentation and local computational and storage resources, and will approach data handling similarly to other large-scale experiments. The overall approach will be to save all "raw" data to archival storage, and then to create a triggering system to reduce information into formats that are easy to process and share.

- The Linac Coherent Light Source (LCLS) x-ray free electron laser (XFEL) is the SC user facility at SLAC that delivers state-of-the-art ultrashort X-ray pulses able to probe the characteristics of matter and the dynamics of physical processes at the atomic and molecular scale. The MEC instrument at LCLS combines the XFEL with high-power, short-pulse lasers to produce and study high energy density (HED) plasmas. MEC-U will have a dedicated infrastructure for reading out detectors, and a shared infrastructure for data reduction, online monitoring, and fast feedback. It will use resources supplied by either SLAC or remotely NERSC: [Case Study 9]
    - The underlying LCLS-II data management system, which MEC will take full advantage of, is designed to handle data rates of 100 GBps and produce 100 PB of data per year.
    - MEC data set sizes are highly dependent on the physics case being studied. Based on estimated laser pulses and beam allocations, it is expected that data sets for an experimental run could be a minimum of 10 GB to a maximum 100 TB with individual file sizes not exceeding 1 TB. The total number of files per experiment can range from a few hundred to 10,000 with a median of 3,000.
    - MEC data transfer will utilize LCLS systems, with the main data transfer tools being bbcp and XRootD[4] on-site data transfer hardware. Other tools are also supported on SLAC's DTNs: scp, sftp, rsync, and a Globus[5] endpoint for data transfers.
- The LaserNetUS virtual organization (VO) is loosely coupled, and sites vary in terms of data volume produced and mechanisms to collect, store, and disseminate data to users. [Case Study 10]
    - Typical shot output is several MB to as much as one GB. An entire experimental run consists of tens to hundreds of shots over the course of several days. The experiments produce scientific data files as well as camera output and may approach hundreds of GBs of data.
    - Managing the data is at the discretion of each site involved in the collaboration. Typical approaches could be requiring the use of portable media, integration to commercial cloud storage, or the ability to transfer data from network-enabled portal systems that are on premises.
    - Researchers are responsible for all aspects of data analysis and data reduction, which they do at their home institutions typically when an experiment has completed. These activities could include simulations, which are used to predict the outcome of experiments or the experimental data is used to guide and benchmark the simulations.
- DOE programs that span facilities and communities (e.g., INFUSE, LaserNetUS) do not typically require a data architecture review to facilitate sharing of experimental results. As a result, the solutions in this space can vary between facilities. Most of these facilities have developed approaches to

4 https://xrootd.slac.stanford.edu
5 https://www.globus.org

address data storage and sharing capabilities, and they have scaled with the current and near-term projections for data volumes. However, the lack of a cohesive and shared understanding of best practices will harm productivity as volumes of data increase. Having access to community-recommended approaches through science engagement, and potentially more efficient data transfer hardware and software, would benefit participants and lead to more efficient use of resources over time. [Case Study 7, 10]

- The ITER computing and data management model is still under development, but is expected to consist of a main data center located in France at the instrument for edge computing, and associated policies and infrastructure to manage distributed data dissemination to partners around the world. ITER data management will require coordination from the US FES community to ensure efficient and equitable access. [Case Study 6, 11]
- The FES community is exploring the use of commercial cloud services for a number of use cases. Some are easier to approach, and could be adapted to a cloud environment with minimal modifications; others require study to understand the set of costs (e.g., computation, storage, and integration resources) that would be associated.
    - MIT PSFC currently hosts historical data from the Alcator C-Mod project[6]. MIT PSFC has started to investigate if migrating this data to an off-premises cloud environment would be less expensive, and easier to manage, long term. Considerations for this potential migration are the overall costs associated with hosting remotely versus locally, and if the software tools that are used to access the data can function at the same level of performance within the cloud. The latter involves testing performance characteristics to ensure no adverse effects to scientific workflows that rely on the data archive. [Case Study 4]
    - GA has investigated some cloud providers as a way to manage backup data and some use cases. Cost and performance of cloud computing use is being explored to understand the tradeoffs between cloud use versus on-premises infrastructure operations and maintenance costs. [Case Study 3, 11]
    - PPPL has migrated some data analysis tasks into cloud storage, and is exploring others as they prepare for upgrades to NSTX-U and the affiliated computational and software requirements. [Case Study 5]

2.3 Scientific Workflow

- FES experimentation typically features three event horizons for data analysis: [Case Study 1]
    - **Automated real-time analysis:** performed as a part of the plasma control system and using local (e.g., to the instrument) computation resources due to latency and availability requirements. The results of this are available faster than the time required to start the next control cycle of an experiment (e.g., approximately 10 ms). Real-time visualization is also

6 https://www.psfc.mit.edu/research/topics/alcator-c-mod-tokamak

possible, although it is often "near" real time. In either analysis use case, the use of remote computational resources for this task is challenging, and avoided.

  - **Control room analysis:** a more sophisticated analysis procedure than the previous, which is designed to provide operators during experimental execution. Requires availability of local compute resources to ensure results are available in near-real time, since the overall goal is to guide adjustments to experiments between shots.
  - **Off-line analysis (e.g., "overnight"):** computationally more expensive routines are used in batch mode to extract more reliable properties of the plasma from sensor measurements. Computational resources at HPC facilities are routinely used for this purpose.

- The time between experimental shots in MFE tokamak experiments is critical to the overall workflow, placing extreme emphasis on network reliability and performance.
- Networking is a critical component of a distributed workflow. ESnet partners with the US fusion community to effectively access data from facilities around the world by developing secure trusted high-throughput data pipelines between these major international experiments and US hubs that can store and distribute the data and analysis capability to registered US collaborators. [Case Study 1, Focus Groups]
- The time between shots during a fusion experiment is limited to tens of minutes across the current generation of MFE tokamak experiments, implying that any analysis that can be done must be highly scheduled and responsive, or a risk exists that the output cannot be used to guide future shots. For this reason, many FES experiments rely on local, and instantly available, computational resources and tools versus leveraging other facilities in a coupled model. [Case Study 2]
- The overall operation time of GA's DIII-D tokamak will remain similar for the next five years, and it is anticipated that the rate of acquiring new data will continue to increase. From 2010 to 2020, the total amount of DIII-D data increased by an order of magnitude. [Case Study 3]
- MIT PSFC's Alcator C-Mod data archive is approximately 150 TB in size and remains heavily accessed by the FES community. There are ongoing efforts to understand how this can be kept active in the coming years, as the hardware that provides the archive will require maintenance or augmentation. Upgrading local hardware and software to modernize the portal or migration of the data to a dedicated facility remain possibilities. [Case Study 4]
- Gyrokinetic simulation will be a major research element during the exascale era of computation. Execution of this simulation at DOE HPC centers has the potential to produce data volumes beyond what the current generation of computing and storage at PPL can handle. As a result, effort to reduce data size is required before it can be stored locally, or transferred from ASCR HPC centers back to PPPL. Additionally, only some portions of

the output can be viewed remotely due to the size of the data sets and the responsiveness of interactive tools that can be used to visualize. In order to maximize productive use of XGC: [Case Study 5]

  - For example, the code XGC can curtail simulation output to adjust to the available memory regions of current DOE HPC resources, with the penalty of reducing fidelity. If required due to memory or networking constraints, XGC can make adjustments to conform to the capabilities of future resources.
  - PPPL and ASCR HPC facilities will require storage upgrades to offer temporary locations for XGC output. PPPL will double its capacity in the coming years to offer PBs of storage space.
  - PPPL is upgrading on-premises data architecture to install new data transfer hardware, is adopting Globus as a software package, has upgraded local storage, and will be working with ESnet to increase network capacity.

- XGC can produce a simulation of turbulence transport in an ITER-like plasma for a given equilibrium time slice using ORNL's Summit; this typically requires a day or more of run time and produces a data set that is approximately 50 PB in size. This volume must be reduced before storage or data transfer, and often only a small portion (typically 1–10 TB) can be sent back to PPPL. [Case Study 5]
  - Future machines are expected to produce data that can approach 300 PB in size.
  - Full data transfer for volumes this large would require multiple Tbps network connections on ESnet between the ASCR HPC facilities and PPPL.
  - Approaches to optimize bulk data transfer and streaming will be required even for reduced data sets.
- XGC is exploring ways to leverage cloud storage as a part of the experimental workflow. Due to the relative performance, as well as the volume and potential costs, it is not expected that cloud storage will replace local resources, but could be used to facilitate data backups, or use cases that require sharing. Additional work in this area could investigate cloud computing for multi-data set analysis. [Case Study 5]
- The MPEX project at ORNL has achieved CD-1 (approve alternative selection and cost range), and is in the design phase. It is expected that the project will be completed by 2027. [Case Study 8]
  - Data will be produced mainly on MPEX with its installed diagnostics. Some post-mortem analysis of material samples will take place in other locations by collaborators.
  - Collaborators will have access to raw and processed data on MPEX and might transfer parts of data for further analysis or processing.
  - It is expected that data long-term storage and archiving is managed at ORNL.

- The Matter in Extreme Conditions Upgrade (MEC-U) proposes a major upgrade to MEC that would significantly increase the power and repetition rate of the high-intensity laser system to the petawatt level. [Case Study 9]
    - The MEC-U project reached CD-1 in Q4 FY 2021 and will have an estimated CD-4 date of FY 2026. It is expected that the MEC-U data system will be complete and ready for beam time by June 2026.
    - MEC-U plans to utilize the existing LCLS-II cyberinfrastructure for operations and will be able to run concurrently without additional upgrades.
- LaserNetUS provides time to users to run laser-based experiments utilizing a collection of high-power, short-pulse lasers that are operated by 10 participating institutions and facilities. These laser systems are often combined with long-pulse "driver" lasers to achieve high density and pressure or with other beams. [Case Study 10]
    - The actual amount of data involved during a run is small (a few GB is common).
    - Each facility has its own research program that is, to varying degrees, separate from LaserNetUS and data associated with the facilities' local programs.
    - There is no standard approach to handle data mobility, and often facilities rely on nontechnical approaches (e.g., portable media) to transfer research data.
- ITER peak data production rates are not fully known as of 2021. However, aggregate estimates of a 20 TB/day data production rate have been made for the engineering operations phase. The ITER timeline, as of 2021, is as follows: [Case Study 6, 11]
    - First plasma: Dec 2025.
    - Additional commissioning and construction: through Dec 2028.
    - Pre-fusion power operations (Phase 1): Dec 2028 through Jan 2030.
    - Pre-fusion power operations (Phase 2): June 2032 through Mar 2034.
    - Nuclear assembly: 2035.
    - Regular operations: Dec 2035.
- The process used for FES simulation workflows is adapting as new codes are developed and more computational resources are made available to the FES community. The classic style of developing a single code base for a small set of machines is being replaced by models that create ensembles of many codes running on multiple machines. This has also been coupled to research to incorporate a greater number of variables and metrics, adjusting to new time and spatial scales, and overall attempts to create "reduced" data models. These adaptations are being driven by HPC allocations occurring at more locations but also by an increased focus preparing for new experimental facilities such as ITER. [Case Study 12]
- The FES community is interested in pursuing simulation workflows that will incorporate the use of artificial intelligence (AI) and machine learning (ML)

in the future, as the codes are adapted to run on next-generation machines and at a larger number of facilities. [Case Study 12]

2.4 Remote Collaboration

- The FES community has a long history of remote collaboration, which will continue as large international efforts (such as ITER, which features 35 countries in collaboration) come into operation. The community draws a distinction between three major types of remote use cases for its scientific workflows: [Case Study 2, & 6, Focus Groups]
    - **Remote observation:** being able to observe aspects of a running FES experiment/instrument, typically through camera views or observable electronic diagnostics. Remote observation is common at many FES facilities. Several considerations must be given to security policies and technologies used, but overall, this is a mature and supportable use case by many major FES experimental facilities. During the pandemic, this method was used around the world.
    - **Remote participation:** encapsulates the requirements of the previous category, but also adds the ability to communicate with local collaborators to influence the direction of experimentation (e.g., modifications that will be made prior to the next shot). Remote participation requires a closer relationship between participants. Examples include EAST and GA, and KSTAR and PPPL. This extra level of cooperation allows for a shared understanding of security considerations, along with goals for experimentation. Typically, the same tools can be used for communication and coordination.
    - **Remote control:** also encapsulates the previous two categories, but affords some level of control over the instrumentation during the experimental process. Remote control is uncommon due to the level of safety and security that is required to operate a FES facility/experiment. It may become more common, provided that the technologies (e.g., network performance, security, measurement/observation integrity, control infrastructure) can be validated and trusted.
- Remote use cases require various levels of technology and policy support to be successful. This comes in the form of either a dedicated environment or known toolsets along with specific information security policies that apply to both the source and users of the end-to-end workflow. [Case Study 2, Focus Groups]
    - It is desirable to make the experience "seamless" so that the process of science is not impeded by technical or policy difficulties; without these considerations in place, the use case will not be optimally productive and may not occur at all.
    - Much of the prior work is being done to support the upcoming ITER use case, which will rely on strong international partnerships.
    - Remote use environments are present at the three major facilities to support collaboration: GA, MIT PSFC, and PPPL.
    - PPPL is currently planning for the Princeton Plasma Innovation Center

(PPIC)[7], expected in 2027, which will feature dedicated spaces to support remote collaboration.

- Commercially available collaboration tools that support communication functions such as audio, video, and text chat (e.g., Discord, Zoom, etc.) are critical to the process of science for FES experiments and facilities. This trend started years prior to the COVID-19 pandemic, remained crucial for ongoing operation during the pandemic, and will remain a part of operation into the future provided the tools perform well and local staff support the use case. Enabling these tools through network peering relationships (directly and via cloud providers) is important for collaboration. Many collaborations and sites adopted these tools during COVID to ensure scientific work could continue. The rapid adoption and use of these tools to support GA's DIII-D was noted in particular. [Focus Groups, Case Study 3]
- Major FES facilities have invested resources into enabling complete remote observation and participation environments. Typically, these considerations include ample ways to transmit and receive audio and video, and in some cases support augmented reality, from remote facilities around the world (e.g., EAST, KSTAR, and eventually ITER), and collaborators that may be located domestically but unable to be in the same physical location. Upgrading domestic connectivity in the coming years to adapt to this continued remote participation, and the associated network requirements, will be required to support features such as high-definition displays and maintain stable latencies and bandwidth needs. [Focus Groups]
- The FES community has adapted remote observation and participation use cases over time, and found a number of software tools that work well, along with a number that are still challenging due to design or operational considerations. Some of these are commercial, others may be open source. X Window System , VNC[8], NoMachine, and others that allow for the ability to view, and occasionally control, remote resources often conflict with information security requirements. Performance of these tools, particularly over great distances, depends heavily on network latency and available bandwidth, both of which are hard to control on busy commodity or institutional networks. Future remote observation and participation approaches will demand tools that offer similar feature sets, along with ways to validate and ensure network performance on an end-to-end basis. [Focus Groups]
- Improvements to existing experiments and development of new scientific infrastructure are allowing for longer shot durations in the FES community. Historically a shot may have lasted only seconds, and future patterns indicate it may be possible to extend this to minutes, hours, or even days. Relatedly, the time between these shots can grow smaller, meaning a greater number of experimental results can be gathered during an experimental run along with larger data volumes for individual observations. These changes to experimental behavior will place more emphasis on networking when remote use cases are present. Collaborators will participate for potentially

7 https://www.pppl.gov/about/learn-more/capital-projects
8 https://www.realvnc.com/en/

longer periods of time, and the time between experiments will be critical to guiding next steps. Networks must be stable, predictable, and have ample capacity for these needs. [Focus Groups, Case Study 2, 8, 11]

2.5 Multi-facility Computational Workflows and Use Cases

- In the FES context, a multi/coupled facility workflow is not considered to be a pairwise facility transaction (e.g., experimental facility coupled with a DOE HPC facility via ESnet). For the FES community, the multi-facility use case implies several facilities working together collaboratively: [Focus Groups, Case Study 11]
  - Instrument and local operations staff at one or more locations.
  - Collaborating/participating groups at a number of remote facilities that are linked via communications tools and remote diagnostics to understand and observe experimental progress.
  - One or more computational and storage facilities, where dedicated analysis resources are available for diagnostics between shots.
  - All of these linked by network infrastructure that carries both communications and data transmission.
- The ability to access live data streams from FES experiments will become necessary in the coming years, particularly as experimental facilities more routinely couple to collaborating computing facilities. This multi-facility model will require advanced software to link experimental resources to storage and computing via the network infrastructure. The increased collaboration will alleviate existing areas of friction in getting things working, provided things can be done in real time. The areas of friction currently include: [Focus Groups, Case Study 11]
  - The increasing volume of data, on ever-increasing timescales as shot lengths increase, and time between shots decreases.
  - Adoption of modern data distribution and caching mechanisms to better disseminate and manage data volumes.
  - The ability of software and hardware tools to quickly ingest and process data, using locally and nationally available computational resources of various varieties of HPC and high-throughput computing (HTC).
  - The ability to provide prompt analysis outputs, which can be used to guide choices during active experimentation during cycles between shots.
- The FES community has long relied on a computational paradigm that encourages the use of a single location. This may be a dedicated pool at an experimental facility, a local cluster, a DOE HPC center allocation, or a commercial facility. Unfortunately, workflows are typically designed to use only one locality, and may be affected if computational resources are not readily available at the specified location. A more efficient approach would be to pursue using computational resources in multiple locations simultaneously, even if it implies having to migrate data, or

computational jobs, away from a preferred location. To distribute and manage computational demand and data mobility requirements, more standardization and resource sharing across the FES complex will be needed. Intelligent tools could be designed to be made aware of options and better spread analysis to the available resources. [Focus Groups, Case Study 5, 11]

- DOE HPC allocations for FES are subject to annual renewal, and this causes challenges for strategic planning or long-term investments in a particular computing capability or workflow architecture. If renewing at the same location is not possible, this often leads to complications in data and workflow migrating to alternate facilities: adapting software to run on different systems, granting accounts to existing users, and sending a majority of scientific data to another facility. Unified APIs and simplified methods to manage data between DOE HPC facilities could simplify the friction seen in these scenarios. Longer-duration (strategic) allocations of computing at ASCR facilities would also allow more effective software investments to be made by the FES community. [Case Study 5]
- DOE programs that span facilities and communities (e.g., INFUSE, LaserNetUS) do not include access to generalized pools of computational resources that can be utilized by participants. While it is possible for participants to pursue these resources independently from DOE HPC facilities, it is a secondary step that must be managed independently. Having access to computational resources, and potentially more efficient data transfer and analysis tools, would benefit participants and lead to more efficient use of resources over time. [Case Study 7, 10]
- Emerging and upgraded FES experiments, such as MPEX and MEC, will adopt the use of DOE HPC resources for some aspects of the experimental workflow. MPEX will leverage NERSC and OLCF, and MEC (via use of LCLS-II infrastructure) will continue to use NERSC. MPEX use of NERSC is not expected for several years, but will consist of TB to PB data transfers to analyze diagnostic data, output from experimental cameras, and simulation workflows. [Case Study 8, 9]
- As the FES community prepares for ITER, the ability to leverage resources across the DOE SC landscape in a multi-facility paradigm (e.g., DOE HPC resources, analysis facilities, distributed users, all linked via ESnet) will become more important as data volumes will far exceed the storage and processing capacity of any single location that participates in FES science. This integration of FES experimental facilities with that of DOE HPC resources via ESnet is critical to the success of the ITER collaboration. Exploring Science DMZ architectures at all FES facilities will be required to ensure that a baseline for data mobility can be achieved. [Case Study 11]
- The ability for DOE HPC facilities to address the requirements of an FES-initiated multi-facility workflow requires addressing several key areas: [Focus Groups, Case Study 3, 11]
    - Creating a dedicated pool of compute resources that can be accessed without having to wait in a queue, either local to experiment, at a coupled

facility, or dynamically allocated through other computational paradigms (e.g., grid or cloud-based). Analysis between shots has a very limited time window (10–15 minutes using current shot-length and expectations on the available time between shots), during experimental periods of several hours. The results of a shot are often used to influence the next; thus analysis operations must be available rapidly to support this use case. This bursty resource use often does not fit an HPC center's operational pattern.

- System-wide scheduling, namely ensuring that all components (computation, storage, networking, and software at all portions of the end-to-end path) are ready when the analysis procedure starts.
- Worker nodes on an HPC system must either have the ability to reach wide-area networking resources (to fetch data that is needed), or must have another transparent mechanism to otherwise retrieve a remote data set (e.g., use of caches, data lakes, or other pre-staging of experimental results).
- To facilitate secure use of the infrastructure, there must be the ability to implement automated methods to facilitate authentication on multiple systems in multiple locations.
- If the analysis process encounters a problem, the experimental staff must have a way to seek immediate help from the HPC facility, instead of a slower turnaround usually seen in trouble-ticket models. Due to the timescales for analysis between shots, there cannot be a lag time to understand and deal with system problems.
- APIs for computational systems must be aware of the multi-facility nature, and accommodate by allowing multiple observers and by supporting remote view operations (e.g., X Window System) for visualization.
- FES workflows would like the flexibility to be able to run at multiple DOE HPC facilities, which implies that having a unified system architecture/API that spans administrative boundaries would be preferable.
- Intelligent software stacks, similar to those seen in other DOE SC programs to manage multi-facility use cases, should be developed or adopted.
- The network(s) that link facilities must have mechanisms to guarantee performance (latency, bandwidth, etc.) to eliminate delays between shots.
- FES codes (current or future) used for workflow and analysis will need modification to understand and adapt to the workflow of the ASCR HPC facilities and FES experimental facilities. Changes could be handling security barriers more gracefully, making concessions to deal with local and remote data, and adapting to the architectures of multiple HPC environments.

2.6 International and Transoceanic Networking

- FES research, development, and operational activities rely heavily on international connectivity provided by ESnet. The coming years will see

the commissioning of new experiments, the addition of new collaborators, and increases in data volume that will place particular emphasis on the reliability and capacity for ESnet's international connections to Europe, and peering relationships with providers that reach other parts of the world (e.g., the Asia-Pacific region, South America, and Africa). [Case Study 1, Focus Groups]

- Networking to support ITER is still in a planning phase though the direction remains unclear; this includes aspects of domestic connectivity within France as well as international connectivity to support distributed collaborators. Options for connectivity could involve the French NREN RENATER[9] or directly connecting to the pan-European REN GÉANT[10]. ESnet can adapt to the connectivity options implemented by ITER once a plan is developed. It is anticipated that the ITPA (International Tokamak Physics Activity) will participate in this activity. [Case Study 1, 6]
- Preparing for ITER operation remains an important focus for the FES community. Current timelines, which may shift due to the ongoing COVID-19 pandemic, feature a phased construction model with key milestones occurring between 2025 and 2035. First plasma is currently expected in December of 2025, with the deuterium-tritium burn occurring in 2035. The next two years are critical for planning how the US FES community will prepare for ITER, with intense focus on the following aspects: [Case Study 1, 6, Focus Groups]
  - Identifying the expected volumes of data that are possible from the facility, and the expectations for being able to act on and handle activity bursts, during operational periods.
  - Adopting components of a scientific platform (e.g., software, computational hardware, storage) able to handle the data requirements locally and at distributed facilities. This includes work being performed via the ITER IMAS activities.
  - Putting in place a timeline for "data challenges" that can exercise the entire ecosystem of the ITER data architecture by simulating the volume and timing requirements using the operational tools.
- The EAST[11] in Hefei, China, is a significant international facility used by the US fusion research community. Operational considerations, such as data mobility to and from this facility, rely on the IPv6 communications protocol because it affords higher levels of performance. Ensuring IPv6 peering across ESnet infrastructure, and with international partners, is critical to the process of science for these interactions. [Case Study 3]

9 https://www.renater.fr/en/organization
10 https://www.geant.org/About
11 http://english.ipp.cas.cn/rh/east/

2.7 Domestic Networking for Local and Wide-Area Use Cases

- The FES community requires stable connectivity to a number of cloud-based communication services that facilitate members' remote participation and collaboration use cases. These include but are not limited to audio and video conferencing (e.g., Discord[12], Zoom[13], etc.). ESnet provides critical paths to these commercial services. [Focus Groups]
- ESnet connectivity is operationally critical for a number of FES facilities. Topological network backups, as well as capacity augmentations, will be required in future years to ensure continuous operation. Each FES facility relies on the ESnet connection to support R&E connected activities domestically and internationally, as well as commercial peering to critical storage, audio, and video services that are used during the process of science.
    - GA has a 10 G wide area network (WAN) connection to ESnet and a 1 G WAN backup connection through a commercial provider. Recent events, including a fiber cut in June 2021, have severely affected the ability of GA to perform daily operations. Upgrading the backup connection to support 10 G to ESnet is viewed as a critical requirement to science productivity. It is a high priority for the organization to ensure a diverse path exists, to support operations into the future. [Case Study 3]
    - MIT has a 1 G ESnet connection through the MIT campus, but is interested in upgrading due to increased use cases that rely on external connectivity to support remote computing and storage that exists off-site, increased levels of remote observation use cases, and serving more data from the Alcator C-Mod project. Upgrading the ESnet connection implies working with the MIT campus to upgrade local area network (LAN) and metro area network (MAN) connectivity. [Case Study 4]
    - PPPL networking requirements have steadily increased over the years as the facility has taken more active roles in existing global FES experiments, such as KSTAR, and prepares for the future requirements of ITER. PPPL currently connects through MAGPI[14], and has upgraded its local networking environment to accept a 100 G WAN connection from ESnet. PPPL is pursuing a primary ESnet 100 G connection, and would also like to pursue a backup connection through diverse paths and providers. [Case Study 5]
- The FES community is exploring the costs and usability of integrating cloud-provided storage and computation into scientific workflows, particularly at facilities that are not able to scale local resources due to cost, space, or lack of expertise. [Case Study 4, 5, 12]
- PPPL has a number of use cases that leverage the Google Cloud Platform (GCP)[15] for storage of data and the execution of software codes; the cloud-

12 https://discord.com
13 https://zoom.us
14 https://www.magpi.net
15 https://cloud.google.com

based storage can take up several TBs of space in the coming years. The usage patterns for this data are not intense: it is expected that some access will occur, but nothing that is part of an active scientific workflow. The usage can come from domestic and international partners. [Case Study 5]

- PPPL HPC workloads that utilize ASCR HPC facilities routinely are not able to perform at peak efficiency due to a number of limitations. Recent upgrades to the PPPL local network and data architecture are expected to alleviate the problems, but further testing will be needed. Some potential bottlenecks to peak efficiency with data mobility are: [Case Study 12]
  - Security infrastructure on PPPL campus was undersized for the expected data volumes and expected capacities. A recent upgrade should enable a higher level of performance.
  - Data transfer hardware was not regularly used. A recent upgrade to deploy purpose-built DTNs will become a part of several scientific workflows.
  - Data transfer software was not standardized, with projects using a mixture of tools that could not efficiently utilize the network and hardware. PPPL is moving toward more capable tools (e.g., Globus) for its DTN pool.
  - New use cases that mix bulk data movement, as well as real-time streaming, mean that the network and DTNs must be responsive to latency as well as bandwidth requirements.
  - Due to the volume of data produced, simulations that execute at DOE HPC facilities are now generating more output than can be easily stored on at the DOE HPC facility long term, or transferred back to PPPL in a timely manner, using the existing software and networking capabilities. PPPL is upgrading site capabilities (e.g., networking, storage, and tools that can be used for data mobility) to address the capability gap. But it will be necessary to scale DOE HPC centers in the future as exascale simulations begin and produce larger data volumes.

2.8 Software Infrastructure

- Software licensure, and import/export controls, can complicate scientific workflows, particularly if approaches that are designed for single user/machine use cases are adapted to shared environments such as an HPC facility. For example, a user of a shared resource often does not have the administrative rights to install and operate software that may require these permissions. This can prevent critical software from being run on resources that would accelerate the workflow, and prevent productivity for the process of science. [Focus Groups]
- MDSplus remains critical to the operation of the FES community, and is widely used and deployed at experimental and analysis facilities. Modifications to the core software have helped FES keep pace with increases in networking capabilities and computational availability. [Case Studies 3, 4, 5]
- FES simulation and theory workflows do not utilize MDSplus, and often rely on other tools that are native to the HPC facilities to accomplish data mobility tasks (e.g., Globus/GridFTP). Not all FES experimental facilities

have similar hardware or software capabilities available, which affects the efficiency of data transfer as a part of these workflows. [Case Study 4]

- The Transport Solver (TRANSP)[16] tool remains critical to FES analysis. TRANSP has the ability to use both MDSplus and Globus to accomplish computational and data mobility tasks, respectively. As a part of the process to define the ITER IMAS, TRANSP will undergo design and development to become compatible with the appropriate interface data structure (IDS) requirements. This marks an early step for the FES community to adopt universal standards for cataloging tokamak data standards. [Case Study 5, 6]
- OMFIT (One Modeling Framework for Integrated Tasks) is a modeling and experimental data analysis software used in the FES community. OMFIT will adapt existing workflows to advance modeling approaches that use HPC resources, and will be more widely deployed as the community prepares for ITER. It is expected that OMFIT will expand to allow for the use of more analysis codes, at more locations, with more participants. Improvements to the systems that handle data mobility, and ways to automate authentication and authorization, are expected. [Case Study 12]

2.9 Cybersecurity

- Use of the Science DMZ architecture, DTNs, and the Modern Research Data Portal is recommended to ensure network security, as well as high performance, when executing FES scientific workflows. [Focus Groups]
- Data and software licensing restrictions affect the ability to use data or software productively. Information security and privacy, as well as monetary/business concerns, are the primary sources of these impediments. [Focus Groups]
- Preparation for ITER is underway, with many US FES community members assisting in gathering requirements for eventual operation. Particular emphasis will be placed on information and cybersecurity policies, data dissemination policy and approach, remote observation and participation, and data volumes expected over time. [Case Study 1, 6]
- FES workflows that span facilities (either experimental site to user, or experimental site to HPC facility) struggle with mechanisms to share and automate credential exchange required by cybersecurity policies. Such credential exchanges are common for data migration and analysis workflow tools. Improving the flexibility of FES workflows to use resources at other facilities will require modification of software mechanisms to cope with security requirements. [Focus Groups, Case Study 1]
- Remote collaboration within the FES community has unique cybersecurity requirements that affect current and future use cases. In particular, the requirements to support remote observation, remote participation, and remote control of any given experiment will dictate the implemented security posture. Large international efforts such as ITER, which features 35 countries in collaboration, will challenge the implementation of baselines

16 https://transp.pppl.gov

due to administrative and national boundaries involved. Particular focus will be given to account management, collaboration tools, and controls placed on data export. [Case Study 2, Focus Groups]

- Public and private partnerships within FES, particularly those from the INFUSE program, often have differing levels of cybersecurity policy, which results in technology friction when scientific data access or exchange may occur. [Case Study 7]

3 Review Recommendations

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

- Preparations for ITER
- Scientific Data Management
- Scientific Workflow
- Remote Collaboration
- Multi-Facility Computational Workflows and Use Cases
- International and Transoceanic Networking
- Domestic Networking forLocal and Wide-Area Use Cases
- Software Infrastructure
- Cybersecurity

3.1 Preparations for ITER

- The FES community will experience unprecedented data volumes in the coming years due to new experimental designs and changes to workflows that place heavy emphasis on networking to link distributed data, processing, and collaborators. It is recommended that the community consider requesting regular participation in a set of "data challenge" activities to support a number of use cases to prepare experiments and facilities for increasing data volumes and reveal gaps in the way that hardware and software cope with the future readiness requirements: [Case Study 1 and 6, Focus Groups]
    - In the context of international collaboration with ITER, in the run-up to first plasma, test the tooling to move data from the instrument to the collaborators.
    - In the context of domestic collaboration, test the remote observation and participation use cases.
    - In the context of multi-facility workflows, test bulk data movement or streaming between experimental locations and distributed analysis facilities.
- It is recommended that ASCR, FES, and ESnet re-assess, via a formal assessment mechanism similar to the 2021 requirements review, the ITER data analysis and network requirements in advance of first plasma. It will also be important to engage the expected data I/O for computation in this assessment, given that decisions in the near future may have important implications for the US and other ITER members regarding the timeliness of data access and the quality of remote participation. [Case Study 6]

- It is estimated thatI the following wide-area networking requirements for different milestone years of ITER will be needed to support the international community. ASCR, FES, and ESnet should evaluate these outbound requirements at the facility, and consider them when designing peering with GÉANT, or connectivity across the existing DOE transatlantic strategy: [Case Study 6]
  - **2023**: 20 Gbps
  - **2027**: 200 Gbps
  - **2031**: 500 Gbps
  - **2035**: 1.5 Tbps
- It is expected that the US fusion community will access ITER data from domestic data mirrors located at DOE facilities, and that some or all the analysis and simulations of importance to ITER will need to be returned back to ITER institutional storage so the analysis and simulation products can be distributed to all the ITER parties. ASCR, FES, and ESnet will provide guidance and connectivity to the domestic facilities involved in this effort. [Case Study 6]
- Decisions have yet to be made by the DOE regarding ITER data centers in the US. However, it is expected that ITER's full data set will be mirrored at a site (or sites) in the US and used by US researchers. By the mid-2030s, storage is expected to reach the exabyte level. FES will consult with ASCR and ESnet when these choices are finalized to ensure connectivity can match the scientific output. [Case Study 6]
- It is anticipated that there will be some US researchers on-site at ITER, but the majority will be remotely located throughout the US. The ability for anyone to effectively participate in ITER experiments is predicated on timely access to the data. It is therefore critical that the requirements for ITER's data workflow be clearly stated as the ITER data workflow requirements pertain to remote participants so that FES, ASCR, and ESnet can prepare for that use case. [Case Study 6]
- Great potential exists for real-time data to enable effective remote participation in ITER experiments from the US. Coordination between FES, ESnet, ASCR, and the data I/O will be essential to design and deliver a data-streaming capability that best serves the needs of US participants during ITER operation. [Case Study 6]

3.2 Scientific Data Management

- A number of current FES community approaches to the handling and management of scientific data could benefit from the experience gained by a cross-section of other DOE SC areas. Lessons learned by members of the DOE HPC community, ESnet, and participants from other program areas could help to strategically influence the trajectory of FES community preparedness for ITER. It is recommended that two collaborative groups either be formed, or otherwise joined to ongoing discussions, to influence some of the discussions surrounding scientific workflow and software

support for FES data and networking preparedness at FES collaboration sites as ITER is commissioned. [Focus Groups]

- ITER will produce unprecedented amounts of data, beyond what many FES collaborators are accustomed to managing. Prior to ITER, a number of topics should be discussed by the FES community, the DOE HPC community, ESnet, and members of the ASCR and FES program offices. These topics could include:
  - Data formats and tooling.
  - Network and data architectures.
  - Data mobility approaches.
  - Data volume expectations.
  - Experimental timelines.
  - Data-sharing policies and procedures.
  - Cybersecurity.
  - FES software licensure and associated import/export controls.
  - "Data challenges" to exercise all of the above.
    - Independent of the ITER workflow, there is a general interest in pursuing FES modes of operation that leverage existing and future capabilities of ESnet and DOE HPC facilities. The FES community, the DOE HPC community, ESnet, and members of the ASCR and FES program offices are encouraged to address multi-facility workflows, with a focus on the following gaps:
    - Addressing the real-time needs through dedicated pools of compute resources that can be accessed without having to wait in a queue to support analysis between shots.
    - System-wide scheduling to ensure all components are allocated for analysis.
    - Worker nodes on HPC systems having the ability to directly or indirectly reach wide-area networking resources for data mobility reasons.
    - Uniform authentication procedures that span multiple systems.
    - Real-time availability of facility support staff during critical operations.
    - Computational APIs that can facilitate observers and remote view operations.
    - The ability to run FES workflows at multiple DOE HPC facilities in a more native fashion.
    - Intelligent software stacks to facilitate the multi-facility use case (e.g., software that is aware of data locality).
    - Networks that can offer guarantees on performance (latency, bandwidth, etc.).
- As ITER is implemented, new requirements to facilitate access and sharing of experimental data will develop. The FES community, and supporting

DOE facilities such as ASCR HPC centers and ESnet, must strive to work within these boundaries in the most seamless way possible or productivity will be harmed. This includes fast networks, the ability to use resources during time windows of experimentation, and adaptable software that can run in multiple locations and access data where it is located. Some of this discussion could involve the ITPA[1]. [Focus Groups, Case Study 1, 6]

- As FES approaches ITER, the adoption of IMAS/IDS for data representation is an important first step to unifying data formats. It is recommended that the FES community look to this as a future goal to standardize data from existing and future instruments, to unify the way that software and workflow can be implemented to address analysis. There is ongoing work with community members and industrial partners (e.g., Google) to create unified databases of FES data, which could benefit this goal. [Case Study 6, Focus Groups]
- The FES community could benefit from the creation of data hubs specialized to analysis and storage use cases needed in the FES community. The overall goal of these types of facilities would be to centralize and specialize on specific computational and storage tasks, adopting the known tools of the FES community, and servicing the desired data formats that are accepted by research groups. Pipelines in/out of these facilities can be well established to allow collaborators a level playing field to interact with the science. [Focus Groups]
- PPPL has implemented a new local network design to support scientific use cases and ways to manage experimental data. PPPL has requested assistance from ESnet to validate both network and data transfer performance. [Case Study 5]
- INFUSE has requested assistance from ESnet to provide a briefing for its community on scientific data management approaches for current awardees, and help in developing a BCP for future participants. Topics may include data transfer hardware and software, along with network design, security policy, and ways to interact with DOE SC resources such as HPC facilities. [Case Study 7]
- LaserNetUS does not have strict guidelines for the cyberinfrastructure and readiness for each facility that participates; therefore, no specific guidelines exist for the policies and procedures to address data management and mobility within, or between, facilities. It is recommended that ESnet provide a briefing for the community on scientific data management approaches, along with developing a BCP that can be used for emerging and upgrading facilities. [Case Study 10]

3.3 Scientific Workflow

- ESnet will work with the FES community and ASCR HPC facilities to explore multi-facility workflows in more detail and how they can be integrated into the scientific workflow specifically for use cases such

1 https://www.iter.org/org/team/fst/itpa

as control room analysis and other off-line computationally intensive operations. Software tools that manage the workflow are critical, and ESnet will help DOE FES understand and support modifications to existing tools, or the creation of new tools, that better adapt to the multi-facility environment. [Focus Groups, Case Study 11]

- ESnet will work with MIT PFSC and DOE FES to explore options for the Alcator C-Mod data archive. This may involve the use of advanced software and hardware tools for data mobility, or working with ASCR HPC and storage facilities as an alternative data location. [Case Study 4]
- ESnet will work with PPPL to validate its new data architecture, specifically the addition of new data transfer hardware and software in its science enclave. Participation in the DME will ensure PPPL is ready to handle the increasing data volumes from XGC. [Case Study 5]
- ESnet will work with ORNL as MPEX is designed to perform "data challenges" between the facility and external collaborators. Participation in the DME will ensure ORNL infrastructure is properly tuned between the experimental enclave and ESnet. In addition, the Modern Research Data Portal could be employed to make large-scale data sets from MPEX available to local computing resources at collaborating institutions. [Case Study 8]
- ESnet will continue to work with SLAC as LCLS-II is upgraded, so that experiments like MEC have fast and predictable paths to ASCR HPC facilities such as NERSC. [Case Study 9]
- ESnet will work with LaserNetUS to audit the data mobility of the 10 partner sites. ESnet can provide a BCP to these sites that describes a simple data architecture to link experimental results to high-performance hardware and software components designed to facilitate data sharing. Participation in the DME will ensure that the 10 LaserNetUS sites are capable of producing a baseline level of data transfer performance. [Case Study 10]
- ESnet will work with INFUSE to provide a BCP that describes ways that the participants can construct a simple data architecture to interface with ASCR HPC facilities. [Focus Groups Case Study 7]

3.4 Remote Collaboration

- ESnet will work with the FES community to periodically review important remote collaboration tools, and its network requirements, to ensure that commercial peering and site capacities are matching expectations. [Focus Groups, Case Study 2]
- ESnet can review FES collaboration tools and help evaluate what performance improvements may be possible, while still abiding by cybersecurity requirements for operation. [Focus Groups, Case Study 2]
- The FES community would benefit from a coordinated view of cybersecurity that can be used to manage communication and data exchange between facilities that are collaborating. Currently, the environment is defined on a per-facility basis, which can lead to mismatches in expectations, performance, and usability for the scientific workflow. FES facilities and

experiments, ASCR HPC facilities, and ESnet can collaborate to define a baseline that can be implemented to ensure security, while enabling higher levels of performance for data access and mobility. [Focus Groups]

3.5 Multi-Facility Computational Workflows and Use Cases

- FES collaborators are interested in pursuing more multi-facility workflows, provided there is time to share requirements and evaluate their effectiveness. A set of pilot demonstrations is recommended for the FES community, DOE HPC facilities, and ESnet, so that all parties can become more familiar with the process and adopt the procedure as routine. [Focus Groups, Case Study 11]
- The FES community should explore a more robust strategy for computation that builds upon existing and potential resources that exist within the DOE SC complex. By doing so, it will be possible to use readily available computational power more easily, versus suffering slowdowns due to resource shortages. This approach should leverage computational resources local to experimental facilities (e.g., GA, MIT PSFC, PPPL) used in conjunction with DOE HPC facilities. Future research will depend on the ability to effectively and efficiently utilize computational resources and increasing volumes of data. These investigations might include: [Focus Groups, Case Study 11]
    - **Generalized workflows:** mechanisms to port analysis workflows from locally available resources of a specific variety to others that could be distributed and of different architectures. This requires software intervention in most cases.
    - **Locational paradigms:** ability to use both local and remote computational resources simultaneously, instead of always running on only one variety.
    - **Computational paradigms:** adapting to use both HPC and HTC mechanisms to perform analysis.
    - **Data staging:** mechanisms to better migrate and stage experimental data where it can be used at centralized analysis facilities, or distributed to community members.

3.6 International and Transoceanic Networking

- Preparing for ITER operation remains an important focus for the FES community. Current timelines indicate that the facility's first plasma will occur in December of 2025, with full operation expected by 2036. The next two years are critical for US FES and ESnet planning and preparation for support to ITER operations. It is recommended that: [Case Study 1, 6, Focus Groups]
    - The FES community should inform ESnet of expected data volumes, operational patterns, collaborations and partnerships, changes to the implementation schedule, and components to be implemented for the scientific data architecture.

- ESnet will consider different network connectivity options between the US and European partners that will address the data mobility requirements that are produced via the aforementioned activities.
- ESnet will assist the FES community in implementing "data challenges" that can exercise the entire ecosystem of the ITER data architecture at the pre-operations times to prepare for operation.
- ESnet can help FES in identifying any bottlenecks that may exist facility to facility, as well as what the user population may experience through the use of advanced network monitoring tools.

- ESnet must work with the FES community to understand the international connectivity requirements of ITER, and will work with the French NREN RENATER or the pan-European REN GÉANT to deliver ITER data to US-based collaborators. [Case Study 1, 6]
- ESnet will work with the FES community, and the Globus project at the University of Chicago, to evaluate and understand the capabilities of Globus when deployed across international networks that utilize the IPv6 protocol. This activity will relate to the ongoing collaboration with the EAST in Hefei, China, and ways that data mobility performance can be improved. [Case Study 3]

3.7 Domestic Networking for Local and Wide-Area Use Cases

- ESnet will work with GA to investigate ways to augment site connectivity with a second 10 G connection to prevent operational disruption during outage of the primary. [Case Study 3]
- ESnet will work with MIT to investigate ways to augment site connectivity to upgrade the 1 G connection to something larger. This will require coordination with the MIT campus, as it is the local provider. [Case Study 4]
- ESnet will work with PPPL to upgrade the primary connection to 100 G, and investigate ways to augment site connectivity with a second 100 G through a diverse path to serve as a backup. [Case Study 5]
- ESnet will audit commodity connectivity that the FES community depends on to ensure performant and diverse paths are available for ongoing operational soundness. Services such as collaboration, audio and video (e.g., Discord, Zoom, etc.), as well as computation and storage (e.g., Google Cloud Project, etc.) are critical to FES remote participation and observation use cases, and are critical to a number of sites that are connected only to ESnet. [Focus Groups]
- ESnet will work with GA, MIT, and PPPL to perform tests associated with the DME[2], a framework that evaluates the ability of a facilities data architecture to be responsive to scientific data challenges. [Case Study 3, 4, 5, Focus Groups]

2 https://fasterdata.es.net/performance-testing/2019-2020-data-mobility-workshop-and-exhibition/2019-2020-data-mobility-exhibition/

3.8 Software Infrastructure

- The ITER software stack, which encompasses many aspects of the complete workflow including data mobility to support analysis activities, is still being designed and subject to participation by several FES community members. It is recommended that ESnet be involved early in some of these discussions to provide insight into data mobility concerns. This may be done in/around the ITPA. [Case Study 1]
- ESnet can assist FES facilities to adopt hardware and software approaches that are native to HPC facilities to accelerate simulation and theoretical FES workflows that require data mobility and which are known to be highly performant. These solutions can be to install and adopt known tools (e.g., Globus, MRDP) or potentially offer services operated by ESnet to foster data mobility improvements. [Case Study 4, Case Study 10, Focus Groups]

3.9 Cybersecurity

- Implementation of broad cybersecurity policies can on occasion affect the performance of open scientific workflows that rely on data mobility between cooperating facilities. FES and ASCR must work to understand the possible impacts, and recommend appropriate mitigations and strategies to afford compliance and protection without affecting performance. ESnet's Science DMZ approach to network perimeter implementation is a part of this approach, and is recommended for FES facilities and experiments. [Focus Groups]
- The FES community and DOE ASCR facilities must collaborate to understand and mitigate the impacts that data and software licensing restrictions have on scientific workflows. Sound policies must be implemented to ensure that software that is critical to FES operations can be used without restrictions that would harm productivity. [Focus Groups]
- The FES community and the DOE HPC centers should collaborate to understand the requirements, policies, and implementation of cybersecurity to facilitate workflows that span facilities (either experimental site to user, or experimental site to HPC facility). In particular, mechanisms to share and automate credential exchange are common for data migration and analysis workflow tools, and are critical to the overall process of science for FES. [Focus Groups, Case Study 1]
- The FES community, in collaboration with other international partners, must define a baseline for cybersecurity to support the remote observation, remote participation, and remote control use cases for facilities. Large international efforts such as ITER, which features 35 countries in collaboration, will require these policies to be conveyed to ensure smooth operation and broad participation to facilitate data sharing and overall participation. [Case Study 2, Focus Groups]

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 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 FES 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 FES of the DOE SC. These case studies were discussed virtually between April and October 2021. Although every effort was made to accurately capture the output of these discussions, each case study was not peer reviewed, and, as a result, the inclusion of some factual inaccuracies within these studies may be possible. Nevertheless, the inclusion of these raw case studies within the report is important, as they form the basis of many of the findings and recommendations above.

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:

- International fusion collaborations
- Remote observation and participation of fusion facilities
- GA: Dlll-D National Fusion Facility
- MIT PSFC
- PPPL
- Planning for ITER operation
- Public-private partnerships in fusion research
- MPEX at ORNL
- MEC Experiment at SLAC
- LaserNetUS Program
- Multi-facility FES workflows
- WDM and FES HPC Activities

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 International Fusion Collaborations

5.1.1 Discussion Summary

International fusion collaborations enable US researchers to explore critical science and technology issues at the frontiers of magnetic fusion research, using the unique capabilities of the most advanced overseas research facilities.

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

- As ITER is designed, there will be new requirements to facilitate access and sharing of experimental data. These includes fast networks, the ability to use resources during time windows of experimentation, and adaptable software that can run in multiple locations and access data where it is located. Some of this discussion may involve the ITPA.
- The ITER software stack, which encompasses many aspects of the complete workflow, including data mobility to support analysis activities, is still being designed and subject to participation by several FES community members
- FES workflows that span facilities (either experimental site to user, or experimental site to HPC facility) struggle with mechanisms to share and automate the credential exchange that is required by cybersecurity policies; this typically is required for workflow tools that attempt to migrate data and perform analysis.
- FES research, development, and operational activities rely heavily on international connectivity provided by ESnet. The coming years will see the commissioning of new experiments, the addition of new collaborators, and increases in data volume that will place particular emphasis on the reliability and capacity for ESnet's international connections to Europe, and peering relationships with providers that reach other parts of the world (e.g., the Asia-Pacific region, South America, and Africa).
- Networking to support ITER remains undecided and opaque; this includes aspects of domestic connectivity within France as well as international connectivity to support distributed collaborators. There are options for connectivity that could involve the French NREN RENATER (https://www.renater.fr/en/organization), or directly connecting to the pan-European REN GÉANT (https://www.geant.org/About).
- Preparing for ITER operation remains an important focus for the FES community. Current timelines indicate that the facility's first plasma will occur in December of 2025, with full operation expected by 2036. The facility will have periods of reduced operation through 2032, and full operation expected by 2036. The next years are critical for planning how the US FES community and how it will prepare for ITER.
- Improvements to existing experiments, and development of new scientific infrastructure, is allowing for longer shot durations in the FES community. Historically a shot may have lasted only seconds, and future patterns

indicate it may be possible to extend this to minutes, hours, or even days. Relatedly, the time between these shots can grow smaller, meaning a greater number of experimental results can be gathered during a run along, with increasing data volumes for each. This time between shots is critical to the experimental process, placing extreme emphasis on network reliability and performance.

5.1.2 International Fusion Collaborations Case Study

A major emphasis of US international collaborations is on superconducting facilities that are capable of true steady-state operation and large-scale fusion plasmas that are not currently accessible in the domestic program. Facilities such as EAST (China), KSTAR (Korea), WEST (W Environment in Steady-state Tokamak, France) and W7-X (Germany) offer access to devices that can in principle operate in steady-state. There are also strong collaborations with Joint European Torus (JET) in the UK, which is a more conventional pulsed plasma device that has plans to operate with deuterium-tritium fuel in late 2021.

5.1.2.1 Background

Since the FES requirements review in 2014[1], rapid progress has taken place in ITER construction with first plasma expected in 2026. These preparations have provided further impetus for enhancing coordination and collaboration between international and US fusion facilities. Such collaborations take place through multilateral bodies such as the ITPA and through bilateral international agreements between ITER member countries.

The pandemic has had a significant affect on travel in the last 15 months. While travel has decreased, improvements in network technology and remote collaboration tools such as Zoom have made it fairly easy to collaborate remotely and transfer modest data sets for further analysis. While for some things remote participation is still not effective, such as the operation of scientific instruments where on-site staff are usually required, many of the interactions such as participation in joint experiments and even leading experiments, can be carried out remotely. However, a key to effective remote participation in international experiments is the need for on-site researchers that invest their time and effort to integrate US researchers into the goings-on at the facility.

Given security concerns regarding the remote control of experiments and scientific instrumentation, it is difficult to imagine that this model of collaboration will change significantly in the future, for ITER or other facilities.

Looking to the next 5-10 years, the existing superconducting experiments (EAST, KSTAR, WEST, W7-X) will extend their pulse lengths to 10s of minutes. Also, new superconducting experiments coming on line (JT-60SA in Japan and ITER in France) will be able to operate for several hundred seconds, and are capable of producing orders of magnitude larger datasets than present experiments.

Gaining experience now with the routine transfer and analysis of large datasets (including real-time data) from existing international long-pulse experiments will be

1 Eli Dart, Mary Hester, Jason Zurawski, “Fusion Energy Sciences Network Requirements Review - Final Report 2014”, ESnet Network Requirements Review, August 2014, LBNL 6975E

important to prepare for ITER operation. While improvements in network technology and remote collaboration tools have enabled improved participation in international experiments since 2014, it is clear that significant challenges remain. An obstacle to progress is that collaborations that cross major administrative domains must cope with different choices for standards, as well as different policies for privacy, data access, remote participation and remote control. While rapid low latency data transfer was achieved as a proof of principle on KSTAR (see 2014 report) the day-to-day networking capability is far from achieving these performance levels.

5.1.2.2 Collaborators

This work will present three important representative experimental collaborations. The most significant of these is JT-60SA in Japan and ITER in France. However, given the history of data access issues in Japan and the very conservative approach in Japan towards data sharing, it is believed that the greatest progress in data streaming for off-site analysis can take place in devices such as EAST (China), KSTAR (S. Korea) and W7-X (Germany).

5.1.2.2.1 KSTAR in Dejeon, South Korea

KSTAR collaborators are widely distributed in the USA and these collaborators join in KSTAR experiments from multiple locations including PPPL (Princeton NJ), UCSD and GA (San Diego).

For KSTAR, it is estimated that about 20 US collaborators exist in total, though far less at any one time. It is routine to have situations where 2-3 US researchers participate in an experiment. Typically, the number of US participants is zero, i.e., the US is only engaged in KSTAR experiments when the KSTAR management reserves specific time for US researchers on the facility. This is because it still represents significant overhead on the part of the KSTAR researchers to facilitate US engagement, owing to language and administrative barriers. There is always a sense that external observers are "imposing" on the good will of the local staff and so engaging remotely is not performed more often than is necessary.

While the international connectivity to KSTAR consists of 100 Gbps connectivity, the use of a VPN is required to access key parts of the facility. As such, this can limit performance due to the large latency, and requirement to pass through security infrastructure.

A typical use case for KSTAR is to use software tools, such as NoMachine, or jump through a head node that allows for access to control and diagnostic systems. Along with this, there is the ability to use the MDSplus tool for transfer and analysis of data. KSTAR facilitates a limited remote control for experimentation: it is possible to modify settings for specific algorithms used for remote control of a specific experiment.

Data volumes are increasing, it is estimated that KSTAR produces approximately 30-40 TB per year, up from previous years data sets. For example, the 2020 run produced 24 TB over 3060 experimental shots (this gives an average of 6 GB per shot). Data mobility is possible through the Globus tool, and some US facilities will mirror KSTAR data sets nightly, with a possible increase in frequency in the future.

5.1.2.2.2 The Wendelstein 7-X (W7-X) Stellarator in Greifswald, Germany

The US collaborates with this facility, and local staff are responsible for operating the facility, sensors, acquisition systems, and for allowing data to be shared with off-site collaborators. W7-X collaborators are also widely distributed in the USA including PPPL, MIT (Cambridge MA), ORNL (Knoxville, TN) and the U. Wisc. (Madison WI).

For W7-X, it is believed that there is a similar number of US collaborators (up to 20) in multiple institutions, although more details should be collected. One collaboration of interest supports the gas puff imaging (GPI) diagnostic on W7-X. For this a group of MIT scientists and students work closely with approximately five close collaborators from Max-Planck-Institut für Plasmaphysik (IPP) -Greifswald, and have additional support from a few dozen other W7-X team members. The Institute for Plasma Physics (IPP) stores a primary and secondary copy of data generated with the GPI diagnostic, and provide access to the data via both MDSplus and an IPP CoDaC-specific method. Currently approximately 1 GB of data reside in a single data set, with of order 10 data sets per day created when running the experiment. Most analysis is performed close to where the data reside at IPP. Analysis by data 'owners' and collaborators informs the physics of boundary turbulence and transport in stellarators

5.1.2.3 Instruments and Facilities

The following sections will highlight some major facilities and interactions.

5.1.2.3.1 ASDEX Upgrade (AUG)

ASDEX Upgrade is a mid-sized divertor tokamak located at the Max-Planck-Institut für Plasmaphysik (IPP) in Garching, Germany. The primary mission for the machine has been support for ITER design and operation, with focus on integrated, high-performance scenarios, the plasma boundary and first wall issues. There are major collaborations in place with US facilities, including at the MIT/PSFC (Turbulence fluctuations, pedestal physics, ion cyclotron range of frequency (ICRF) heating, metallic first walls and steady-state scenario development); DIII-D (divertor and pedestal physics; cyclotron range of frequency heating and current drive and steady-state scenario development); NSTX-U (diagnostics development and turbulence studies). Important collaborations on theory and modeling are also in place with many US groups.

5.1.2.3.2 JET

JET, under the European Fusion Development Agreement (EFDA), is located at the Culham Science Centre, in Abingdon, United Kingdom. It is the largest tokamak currently in operation in the world. Major collaborations in place with US facilities include MIT/PSFC (pedestal physics, TAE physics and disruption mitigation); DIII-D (H-mode pedestal physics, especially edge localized mode (ELM) suppression, neoclassical tearing modes, resistive wall modes and rotation, steady-state scenario development); NSTX-U (Alfven Eigenmode physics, neoclassical tearing modes and resistive wall mode research).

5.1.2.3.3 ITER

ITER is a partnership among 35 nations to build the world's first reactor-scale fusion

device under construction in Cadarache, France. The ITER Project expects to finish major construction in 2018 and to operate for >20 years. The current date for first operation is 2023.

5.1.2.3.4 KSTAR

KSTAR is an all-superconducting tokamak experiment located at Daejeon, Korea. KSTAR's size, operation capabilities and mission objectives for the initial operating period will eventually be comparable to the present DIII-D tokamak. The main research objective of KSTAR is to demonstrate steady-state high-performance advanced tokamak scenarios. PPPL (plasma control system [PCS], diagnostics, ICRF), ORNL (fueling), GA (PCS, data analysis, electron cyclotron heating [ECH]), MIT (Long-pulse data system), Columbia U. (data analysis). KSTAR had its first plasma in 2008, and US scientists worked closely with KSTAR scientists in the last several experimental campaigns.

5.1.2.3.5 EAST

EAST, located at the Chinese Academy of Sciences, Institute of Plasma Physics (ASIPP), Hefei, China, is the world's first operating tokamak with all-superconducting coils. EAST is somewhat smaller than DIII-D but with a higher magnetic field so the plasma performance of both devices should be similar. Its mission is to investigate the physics and technology in support of ITER and steady-state advanced tokamak concepts. Major collaborations with US facilities include, GA (digital plasma control, diagnostics, advanced tokamak physics, operations support), PPPL (diagnostics, PCS), Columbia University (data analysis), MIT (long-pulse data system development) and the Fusion Research Center at the University of Texas (diagnostics, data analysis, theory). The collaboration with scientists from the United States was instrumental in their successful first plasma in September 2006. Since then, collaborations have continued in every EAST experimental campaign. During the 2014 campaign, GA deployed a Science DMZ based on the ESnet model to improve the ability of US scientists to collaborate with EAST during the experimental campaign. See the Use Case on Remote Operation and Control.

5.1.2.3.6 SST-1

SST-1 (Steady-State Tokamak) is located at the Institute for Plasma Research (IPR), in Bhat, India. It is the smallest of all the new superconducting tokamaks with a plasma major radius of 1.1 m, minor radius of 0.2 m, and plasma current of 220-330 kA. First plasma occurred on June 20, 2013. The main object of SST-1 is to study the steady-state operation of advanced physics plasmas. One collaboration is with DIII-D in the area of physics, plasma operation, theory, and ECE diagnostics. The recent re-working of the SST-1 superconducting toroidal field magnets increased the device's error fields. MIT personnel have carried out calculations of the 3D error field magnitude and the effect on plasma initiation. Collaborations on additional topics are under discussion. It is anticipated that this collaboration will grow to encompass other groups within the United States.

5.1.2.3.7 LHD

The Large Helical Device (LHD) is a large (R = 3.9 m, a = 0.6 m, B = 3 T)

superconducting stellarator device that began operating in 1998 at the National Institute of Fusion Science, Toki, Japan. There are active US collaborations on this device.

5.1.2.3.8 Wendelstein 7-X

W7-X is a large ($B class) stellarator at Greifswald, Germany. Commissioning began in 5/2014 and initial experiments will begin in 2015. The United States contributed to construction and MIT is collaborating on a GPI diagnostic for edge turbulence and on data acquisition for diagnostic systems. A funding opportunity was just announced for participation in the research program.

5.1.2.3.9 JT60-SA

The JT-60SA ("Super Advanced") is a large, breakeven-class, superconducting magnet tokamak nearing the end of construction in Naka, Japan. This program represents a coordinated effort between the EFDA and JAEA. Although there is a rich history of collaboration between the United States and Japan the extent of the US involvement in this experiment is not known at the present time, but is expected to grow.

5.1.2.3.10 Other Facilities and Interactions

A number of additional facilities are also targets of somewhat less intense collaboration including WEST (formerly Tore Supra) in Cadarache, France, Tokamak à configuration variable (TCV) at the Center for Research in Plasma Physics in Lausanne, Switzerland and the Mega Ampere Spherical Tokamak (MAST) at the Culham Science Center in the UK.

DOE has specifically funded two large international collaboration projects to support collaborations between a team of many US institutions, and (separately) KSTAR and EAST.

The first focuses on plasma-material interactions, and involves (at least) MIT (lead), GA, PPPL, LLNL, UCLA, UCSD, William & Mary. The topics cover:

- High-power, long-pulse radio-frequency (RF) actuators
- Tungsten divertor for long-pulse operations
- Disruption analysis and experiments
- Optimization of operational scenarios with a high-Z first wall
- Self-regulated plasma-surface interactions in long-pulse tokamaks
- Technology for enhanced remote participation

The second focuses on plasma scenarios and control, and involves GA (lead), MIT, PPPL, LLNL, UT Austin, ORNL, Lehigh, and UCLA. Principal topics in the collaboration include:

- Scenario development with superconducting coils and highly diverse heating and current drive
- Long-pulse sustained high-performance operation issues

- Consistency of long-pulse and high performance with metal walls and divertor materials
- Robust plasma control for long-pulse disruption-free scenarios
- Diagnostic development relevant to long pulse
- RF actuator modeling & development
- Technology for enhanced international remote operation and participation
- Simulations to support transferring scenarios from one device to another with very different operating characteristics

5.1.2.4 Process of Science

For the purpose of a case study, two collaborations will be identified, both on superconducting international experiments. The first experiment is KSTAR in Korea and the second is a newer device called W7-X in Germany. The US has strong collaborations in place with both facilities, including scientific instruments installed on these facilities and strong participation in joint experiments, including experiments which are led by remote participants in the USA.

In typical fusion experiments, a plasma pulse is created by energizing magnets, ionizing a gas, and applying a combination of heating and current drive methods that allows high pressure to be achieved for as long as the magnets are energized. The plasma duration is typically a <10 s in existing experiments but it can be much longer in international superconducting experiments.

Data is collected on up to several thousand sensors distributed around the machine and on many different timescales depending on the field being measured. At one end of the temporal domain, data can be collected at 5-10 MHz and on several hundred channels used to image plasma turbulence. At the other end, data can be collected on thermocouples embedded inside the walls to measure bulk thermal properties with a time constant of 10s of ms.

The data may be used by two workflows. First, the data can be used in real time to control the experiment. Second, the data can be stored and accessed after the pulse. It has been trending that more data is now being accessed in real time for advanced methods of plasma control, making use of improved architectures (*graphics processing unit*s [GPUs], field programmable gate arrays [FPGAs]) and improved algorithms (AI/ML methods) for determining the present and future plasma state.

While more data and more real-time analysis is occurring in plasma control, remote participants at present cannot easily access the control or pulse design computers on major facilities because of safety and security concerns. This will also be the case on ITER. This means that remote participants must continue to work with on-site operators to implement control system and pulse schedule changes (expected for KSTAR and W7-X).

In fusion experiments, there are three time scales for data analysis:

- The first timescale of analysis is automated real-time analysis of the data as part of the PCS. Real-time analysis is becoming more sophisticated as availability of fast, responsive, computing improves. For example, there is

the capability to compute the plasma stability from sensor input fast enough to control the experiment. This is a remarkable achievement and much more is coming with advances in hardware and algorithms. Likewise, at KSTAR the US has built an FPGA that takes 30-ish magnetic measurements from the plasma and characterizes the dominant magnetic instabilities and feeds this into the PCS. The goal is to develop control algorithms that use this information with other sensor data to steer the plasma discharge away from dangerous operating boundaries.

- Typically, the round-trip time for these signals to the US and back is too long to expect that HPC capacity in the US can be used to steer experiments in real time. The control of the experiment and all the computation required for that control must remain local. However, the signals feeding into the control system and the signals coming out of the control system can be monitored in real time for visual inspection. Also, in principle, these "steaming" data can be analyzed asynchronously from remote locations. However, there is a huge gap between principle and practice in this regard. While some proof of principle for fast transfer of large data volumes was demonstrated and documented in the 2014 report, the current standard of transfer is too inadequate and unreliable for such data streaming. This should be a capability that is developed in time for ITER operation and demonstrate practical solutions in existing long-pulse experiments.
- A second timescale of analysis is known as "control room" analysis. This is analysis that is perhaps too sophisticated for real-time control but very useful for humans to make decisions about what to do for the next plasma pulse based on an understanding of what happened in the last pulse. It is this level of "between-shot" analysis that is extremely useful for remote participants in experiments. Such analysis is typically carried out at the international facility on local computers. However, it cannot be expected that local computers will have all the compute capability that international collaborators will need and it will therefore be important to develop the network capability and sort out the security issues to enable rapid post pulse data transfer, or even better the streaming of data during a plasma pulse for off-site analysis (this latter capability will be particularly important for ITER).
- A third timescale of analysis typically takes place overnight or on longer periods of time such as months. At this level, computationally more expensive routines are used in batch mode to extract more reliable properties of the plasma from sensor measurements. While this is not relevant for plasma control, it is seen as necessary for publication of measurements in refereed journals or for further physics analysis. associated with advanced simulation codes and the detailed comparisons that have to be made between those codes and processed data.

While these different time scales are shrinking and more sophisticated analysis is taking place faster and with greater automation, it is expected that these categories of analysis will persist during ITER operation.

5.1.2.5 Remote Science Activities

The WAN obviously plays a critical role in the ability of US scientists to participate

remotely in experimental operations on any of the international machines discussed above. Network use includes data transfer as well as specialized services like a credential repository for secure authentication. Overall, the experimental operation of these international devices is very similar to those in the United States with scientists involved in planning, conducting and analyzing experiments as part of an international team.

Experimental planning typically involves data access, visualization, data analysis, and interactive discussions amongst the distributed scientific team. For such discussions, some form of video conferencing (e.g., Zoom, Skype) is utilized. Which technology that is used often depends on the technical capability of scientists at each end and on their experience.

Data analysis and visualization is typically done in one of three ways; either the scientist logs onto a remote machine and utilizes the foreign laboratories existing tools, the data is transferred in bulk for later local analysis and visualization, or they use their own machine and tools to remotely access the data. The widespread use of MDSplus makes the last of these techniques easier and more time efficient yet this is not possible at all locations.

Remote participation in international experiments has the same time critical component as does participating in experiments on US machines. The techniques mentioned above are all used simultaneously to support an operating tokamak placing even higher demands on the WAN, especially predictable latency. In addition to what was discussed above, information related to machine and experimental status needs to be available to the remote participant. The use of browser-based clients allows for easier monitoring of the entire experimental cycle. Sharing of standard control room visualizations is also being facilitated to assist the remote scientist to be better informed.

Representative Remote Science Activities:

- **MIT at TCV:** Data is being generated by diagnostics installed on TCV. Diagnostics will use ~30 cameras over the next 0-5 years. No major upgrades that will change the data load substantially are planned. Computer clusters are available at TCV. École Polytechnique Fédérale de Lausanne (EPFL) is set up to store all data from TCV. The data sets are similar to any pulse-based experiment. The maximum discharge length is 2 sec.
- **MIT at W7-X:** Data generated by GPI diagnostic installed on the W7-X stellarator - the plan is to extend the pulse length to ~100 s within the next 2 years and extend it to 30 minutes within the next 5 years. IPP Griefswald has the computational, storage, and network capabilities to handle the expected data load. Computer clusters and data storage are available to all users. However, firewall policies are quite restrictive, limiting off-site access.

5.1.2.6 Software Infrastructure

The MDSplus data system is widely used in the worldwide fusion community. It provides tools for local data management as well as remote data access. Web-based tools for run planning, run monitoring, and electronic logbooks are becoming ubiquitous. Rather than transferring data sets, remote data access and remote computer access are the preferred modes of operation.

There are three approaches to remote collaboration across the WAN. Traditional file transfer, or data extraction followed by transfer is used but does not fit into the interactive nature of operating a fusion experiment, where the results from one shot inform the decisions about the next shot. Nevertheless, this is used, and takes advantage of the traditional wide-area data transfer tools such as gridftp, remote screen access, NX, Windows Remote Desktop, and VNC work well for applications that do not require too much user interaction. In general, the pictures of the data are significantly smaller than the data itself. This approach avoids transferring large data sets across the network. Finally, remote data access using MDSplus allows for transfer of only the subset of data the user requests. However, it suffers from wide-area transactional latency problems.

5.1.2.7 Network and Data Architecture

Fusion experiments are highly interactive. Immediate results are fed back into the setup of subsequent shots. This makes performance, as opposed to throughput, the more important metric for wide-area collaboration. The main challenges are related to network latency. Any help ESnet can provide in this area would improve the success of remote collaborations.

Personal interaction is critical to remote collaboration, especially international. Language, time zone differences, and simply not knowing the collaborators as well, exacerbate this.

5.1.2.8 Cloud Services

Beyond what has been discussed, there are no major cloud services used during international collaborations.

5.1.2.9 Data-Related Resource Constraints

Increased data bandwidth is needed. In the next several years multiple international tokamaks will operate in long-pulse mode and will require continuous data replication and data access. Those experiments will have more diagnostics and increased time-fidelity.

Lowering the network latency is also important. Currently, the amount of data transferred between international and domestic sites is not very large. However, the real-time or near-real-time aspect of data transfers and between collaborating sites is very important. Increased peering with major Internet providers worldwide is helpful. In the past, shorter path and better peering helped with increased network throughput and decreased latency.

While for some activities such as participation in joint experiments and even leading experiments, remote meeting and participation capabilities have been adequate, remote participation is still not effective for the operation of scientific instruments where on-site staff are usually required.

5.1.2.10 Outstanding Issues

The following sections outline ongoing and expected areas of friction for international fusion collaborations now, and into the future.

5.1.2.10.1 Long-Pulse Support

As superconducting international experiments achieve truly long-pulse operation (> 100s), it is essential that ESnet provide the support needed for the US fusion community to effectively access data from facilities around the world by developing secure trusted high-throughput data pipelines between these major international experiments and US hubs that can store and distribute the data and analysis capability to registered US collaborators.

5.1.2.10.2 Increased Network Bandwidth

In the next several years multiple international tokamaks will operate in long-pulse mode and will require continuous data replication and data access. Those experiments will have more diagnostics and increased time-fidelity. Looking to the next 5-10 years, the existing superconducting experiments (EAST, KSTAR, WEST, W7-X) will extend their pulse lengths to 10s of minutes. Also, new superconducting experiments coming on line (JT-60SA in Japan and ITER in France) will be able to operate for several hundred seconds, and are capable of producing orders of magnitude larger datasets than present experiments.

5.1.2.10.3 Supporting the Virtual Control-Room and Interactive Experiments

Fusion experiments are highly interactive. Immediate results are fed back into the setup of subsequent shots. This makes performance, as opposed to throughput, the more important metric for wide-area collaboration. The main challenges are related to network latency. Any help ESnet can provide in this area would improve the success of remote collaborations.

5.1.2.10.4 Operationally Realistic Testing

Gaining experience now with the routine transfer and analysis of large datasets (including real-time data) from existing international long-pulse experiments will be important to prepare for ITER operation.

5.1.2.10.5 Federated Security

Technical and policy advancements to allow sharing of authentication credentials and authorization rights would ease the burdens on individual collaborating scientists. This sort of development is crucial for more complex interactions, for example where a researcher at one site accesses data from a second and computes on that data at a third site. (The Scientific Discovery Through Advanced Computing [SciDAC] funded National Fusion Collaboratory[2] project deployed this capability for data analysis within the US domain.).

5.1.2.10.6 Document and Application Sharing

Improved tools for sharing displays, documents and applications are already urgently needed. Cognizance of different technology standards and policies will be important. There have been difficulties navigating common document sharing rules within ITER, and these policies and practices should be reviewed.

2 D.P. Schissel, "Grid Computing and Collaboration Technology in Support of Fusion Energy Sciences," Physics of Plasmas 12, 058104 (2005).

Personal interaction is critical to remote collaboration, especially international. Language, time zone differences, and simply not knowing the collaborators as well, exacerbate this.

5.1.2.11 Case Study Contributors

International Fusion Collaborations Facilities Representation

- Jerry Hughes[3], MIT PSFC
- Raffi Nazikian[4], PPPL
- David Schissel[5], GA

ESnet Site Coordinator Committee Representation

- Scott Kampel[6], PPPL
- Jeff Nguyen[7], GA
- Brandon Savage[8], MIT PSFC

3 jwhughes@psfc.mit.edu
4 rnazikia@pppl.gov
5 schissel@fusion.gat.com
6 skampel@pppl.gov
7 nguyend@fusion.gat.com
8 bsavage@psfc.mit.edu

5.2 Remote Observation and Participation of Fusion Facilities

5.2.1 Discussion Summary

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

- The FES community has a long history of remote collaboration, which will continue as large international efforts (such as ITER, which features 35 countries in collaboration) come into operation. The community draws a distinction between three major types of remote use cases for their scientific workflows:
  - **Remote observation:** Being able to observe aspects of a running FES experiment/instrument, typically through camera views or observable electronic diagnostics. Remote observation is common at many FES facilities. There are several considerations that must be given to security policies and technologies used, but overall this is a mature and supportable use case. During the pandemic, this method was used around the world.
  - **Remote participation:** Encapsulates the requirements of the previous category, but also adds the ability to communicate with local collaborators to influence direction of experimentation (e.g. modifications that will be made prior to the next shot). Remote participation requires a closer relationship between participants. Examples include EAST and GA, and KSTAR and PPPL. This extra level of cooperation allows for a shared understanding of security considerations, along with goals for experimentation. Typically the same tools can be used for communication and coordination.
  - **Remote control:** Also encapsulates the previous two categories, but affords some level of control over the instrumentation during the experimental process. Remote control is uncommon due to the level of safety and security that is required to operate a FES facility/experiment. It may become more common, provided that the technologies (e.g. network performance, security, measurement/observation integrity, control infrastructure) can be validated and trusted.
- Remote use cases require various levels of technology and policy support to be successful. This comes in the form of either a dedicated environment or known toolsets along with specific information security policies that apply to both the source and users of the end-to-end workflow:
  - It is desirable to make the experience "seamless" so that the process of science is not impeded by technical or policy difficulties; without these considerations in place the use case will not be successful.
  - Much of the prior work is being done to support the upcoming ITER use case, which will rely on strong international partnerships.

    - Remote use environments are present at the three major facilities to support collaboration: GA, MIT PSFC, and PPPL.
    - PPPL is currently planning for the PPIC, expected in 2027, which will feature dedicated spaces to support remote collaboration.
- Improvements to existing experiments, and development of new scientific infrastructure, is allowing for longer shot durations in the FES community. Historically, a shot may have lasted only seconds, and future patterns indicate it may be possible to extend this to minutes, hours, or even days. Relatedly, the time between these shots can grow smaller, meaning a greater number of experimental results can be gathered during a run along, with increasing data volumes for each. These changes to experimental behavior will place more emphasis on networking when remote use cases are present. Collaborators will participate for potentially longer periods of time, and the time between experiments will be critical to guiding next steps. Networks must be stable, predictable, and have ample capacity for these needs.
- The time between shots during a fusion experiment is limited to 10s of minutes across the FES facilities, implying that any analysis that can be done must be highly scheduled and responsive, or there is a risk that the output cannot be used to guide future shots. For this reason, many FES experiments rely on local, and instantly available, computational resources and tools versus leveraging other facilities in a coupled model.

5.2.2 Remote Observation and Participation of Fusion Facilities Case Study

The international magnetic fusion research community has a long history of effective collaborative research going back to the 1958 meeting on Peaceful Uses of Atomic Energy in Geneva. The subsequent years have seen the collaborative environment consistently adopt new technology trends to facilitate information sharing and communication that spans international barriers. This case study will highlight some of the components of the approaches that are used during FES operations now and into the future.

5.2.2.1 Background

From a historical perspective, FES collaboration has been centered on exchanging scientists, ideas, and even data in a highly collaborative effort to advance the science of magnetically confined plasmas. The ITER project, currently under construction in France, is a great example of this collaborative spirit where 35 nations have joined together to build the world's largest tokamak.

Over the past decade, fusion research has seen an extension beyond remote scientists participating in experimental operation. Now, in some select cases, fusion researchers will operate and control a remote experiment. Thus, the abbreviation RCR has been extended in some cases from Remote Collaboration Room to Remote Control Room. What follows in this use case is an examination of **remote control and operation** as opposed to only **remote participation** with the realization that this unique capability is only possible in a few select instances. Cases to be studied include

- The operation of the EAST located in Hefei, China, from the RCR located at

GA in San Diego

- Operating an instrument attached to the DIII-D tokamak operated at GA by DIII-D team members off-site at MIT PFSC.

The scientific workflow for remote operation and control remains mostly the same as for remote participation:

- A fusion plasma is created within the experimental infrastructure
- Scientific data is acquired via sensors, cameras, diagnostics
- Scientific data goes through a first round of analysis to produce immediate results
- Scientific data is prepared for visualization, as needed
- The resulting data and visualizations are debated amongst the scientific team (local and remote) until a decision is reached on what changes to make before creating the next fusion plasma.

In this highly interactive environment, equal and timely access to all data is critical for all parties. Typically, the steps above must be completed in the limited time window between experimental shots.

5.2.2.2 Collaborators

The collaboration space for this use case is being limited to the primary actors described in Section 5.2.2.1: staff and users at GA, MIT PFSC, and PPPL domestically, and the EAST facility in China.

5.2.2.2.1 EAST to GA RCR

There are two groups of collaborations for the GA RCR and the collaboration zone. First, is the EAST facility itself which includes not only the on-site staff but also the data produced by the facility. The second group are those approved US scientists who are not in the RCR but are allowed to connect to the collaboration zone services. There are presently approximately 70 approved US scientists who can connect and use the services and they are located throughout the United States at national laboratories and major universities.

5.2.2.2.2 DIII-D at GA to MIT PFSC

As in Section 5.2.2.2.1, there are multiple collaborating groups in this relationship. The local staff at GA are responsible for operation of DIII-D, along with a number of scientists that may be actively participating in the experiment. The second group are the remotely located US scientists who are not in the RCR but are allowed to connect to the collaboration zone services at MIT PSFC.

5.2.2.3 Instruments and Facilities

5.2.2.3.1 EAST to GA RCR

Using available space (~300 ft^2), a dedicated Remote Control Room (RCR) has been constructed at GA in the same building that houses the majority of DIII-D's scientific staff. Originally designed for twelve participants, it presently has seats for eight as

usage patterns indicated more table space was required per person than was assumed. Each participant has a 24" LCD monitor with display cable, an Ethernet network cable, a mouse and keyboard, and a power outlet. Scientists are expected to bring their laptop computer and either run applications natively on the laptop or use X Windows to log onto one of the generally available Linux workstations.

Borrowing on ideas successfully deployed in the DIII-D control room, larger displays facing the participants are used for more experiment-centric quantities of interest that can change daily. The physical dimensions of the room allowed five 52" LCD TVs to be mounted on the wall facing the scientists. Four of these TVs are stacked in two rows of two and one TV is rotated 90° and placed adjacent to the stack of four. Given the elongated nature of modern tokamak design, the rotated TV with its 1.78 ratio of height to width is well suited for display of real-time plasma boundary calculation where typical machine elongation is ~1.8. Mounted to the right of the scientists are six 24" LCD monitors mounted in two rows of three each. These screens are used to display data associated with the plasma control system including raw diagnostic data and computed quantities. Each screen is divided into four quadrants with each one displaying a temporal history of two quantities resulting in a total of 48 traces.

A Collaboration Zone based on ESnet's Science DMZ model is deployed via a second network segment directly from the ESnet border router. This capability is used exclusively for GA collaboration with the EAST tokamak in China where large bulk data transfers of EAST data to this Science DMZ are accomplished when EAST is operating. Scientists in the GA RCR can then access EAST data in this Science DMZ. For security, ACLs are used on the collaboration zone router allowing only approved IP addresses (both IPv4 and IPv6) to connect to systems in the collaboration zone.

Dedicated computers in the GA RCR are used to display content on all screens. Contained within the Science DMZ is a DTN that is used to transport from EAST bulk data with the Aspera data transport software. The bulk files are actually MDSplus Tree files and they are placed directly in a dedicated MDSplus server reserved exclusively for EAST data. The data transfer sequence is automated so that data arrives in the collaboration zone in a timely manner. The messaging system RabbitMQ is used to coordinate the transfer process between multiple machines and to allow for the collection of statistics and status information. In addition to automated bulk transfer, end users are able to queue up data transfer requests using a Web portal with RabbitMQ background to request less common MDSplus tree data that are not automatically transferred. A Web Portal provides a method to manually request data transfers and automatic monitoring of transfers and experimental status. In addition to the automatic bulk transfer of MDSplus data, a portion of the plasma control data is moved directly to the collaboration zone in real time. This is accomplished via an Redis in-memory data structure server.

5.2.2.3.2 DIII-D at GA to MIT PFSC

The MIT PSFC has a dedicated RCR space, having an open plan with 14 workstations available for scientific computing on a networked linux cluster administered by the PSFC MFE division. Access is available off-site via Secure Shell (SSH) or NoMachine. Numerous computational tools such as IDL and MATLAB are maintained on the local cluster, and ready access is available to resources at off-site institutions. The facility

is also equipped with extensive video-conferencing capabilities for interacting with off-site colleagues, students, and collaborators.

The RCR provides the functionality for staff and students to watch and interact with real-time displays of data and video feeds, when they are provided by off-campus fusion-experiment control rooms. In particular, direct access to DIII-D is available on this network, allowing for full access to discharge data there. Full access to the full catalog of C-Mod data is granted from these workstations, via a connection to a PSFC server running MDSplus. The cluster has robust connectivity through ESnet to SC partner institutions. The RCR is routinely used by scientists, postdocs and graduate students for connectivity with operations at off-site facilities.

The key features of the RCR include:

- Four real-time displays and 14 workstations to be used for participation with remote experiments
- A sound-isolated video conference (VC) area with conference table, chairs, and large display Monitor
- Desk space and monitors for users who wish to connect laptops
- A common space with chairs that fosters discussion and the sharing of ideas and experience
- With these capabilities, the RCR mitigates risks associated with host facility schedule changes, thus helping to control travel expenditures. It also allows increased participation from PSFC staff and students who may be unable to travel for various reasons.

During COVID-19, the RCR platform has been used significantly for access to the MFE computing by PSFC and external researchers.

5.2.2.4 Process of Science

5.2.2.4.1 EAST to GA RCR

GA uses the RCR to operate and control EAST during their Third Shift. Remote operation means that the scientists conducting the experiment and the team programming the feedback control of the plasma pulse are not collocated in the EAST control room. US scientists not located in the GA RCR participate in the experiment but are not able to perform any hardware control functions. The strong collaboration between scientists at EAST in China and scientists in the United States, and GA specifically, has resulted in a number of common technologies including the MDSplus data management system and the PCS. The collaborative environment, shared technology, and 8-hour time difference between EAST and GA made third shift operation of EAST by GA an ideal opportunity for conducting novel fusion experiments.

The process of the science for running EAST's third shift is very similar to the process at other tokamaks. Prior to the operation date, an experiment coordinator prepared the experiment using a web-based wiki calendar to assign critical roles, such as session leader, physics operator, diagnostic coordinator and computer operator. Since the third shift started at 5 AM local time, advance coordination is critical. The coordinator also uploads a mini proposal and shot plan document to the calendar for everyone's benefit. The calendar and documents are displayed on one of the 24" monitor in the RCR.

The details of audio/video connections are also posted on the monitor. Throughout the experimental session, hardware/software plasma control adjustments are debated and discussed amongst the experimental team and made as required by the experimental science. Decisions for changes to the next plasma pulse are informed by data analysis conducted between plasma pulses. Thus, this mode of operation requires rapid data analysis that can be assimilated in near real time by a geographically dispersed research team.

Communication between EAST's team and remote participants are conducted using the Zoom audio/video conferencing service. The transition of Discord to utilizing Google Authenticator rendered Discord a non-viable solution for this collaboration. Two large TV displays are dedicated for Zoom video and screen sharing respectively. Advanced features of Zoom made the experience better for everyone, namely text chat for exchanging tactical information and break out rooms for ad-hoc conversations. Each participant in the RCR typically brings a laptop computer with a USB/analog headset in order to participate in the experiment. Prior to a run week, a training session is conducted with the US participants to ensure they are ready for the experiment (e.g., their computer, network connections, data access, audio/video). Similar to the physical control rooms, guests occasionally visited the RCR to casually follow the experiment or say hello to their counterpart at EAST.

The operation of EAST's Third Shift is expected to remain similar for the next five years. However, as with other operating facilities, the amount of data collected and analyzed will continue to increase, placing an additional strain on the networking and computational infrastructure.

5.2.2.4.2 DIII-D at GA to MIT PFSC

MIT co-operates the LLAMA diagnostic on DIII-D. LLAMA "keeper" for the run day, sited at MIT, checks in on diagnostic interlocks, temperatures using Ignition process control software. Device shutter can be actuated remotely. Other actuations, such as power supplies, must be done via contacting someone on-site at DIII-D. Using tools such as ReviewPlus (in IDL) and MATLAB tools running at DIII-D, but displaying locally, the diagnostic operator confirms correct signal acquisition shot to shot.

5.2.2.5 Remote Science Activities

5.2.2.5.1 EAST to GA RCR

As noted previously, the remote instrument is the EAST, located at the Chinese ASIPP, Hefei, China. EAST is somewhat smaller than DIII-D but with a higher magnetic field so the plasma performance of both devices should be similar but its superconducting coils allows it to run plasmas of significantly longer duration. Its mission is to investigate the physics and technology in support of ITER and steady-state advanced tokamak concepts. The collaboration with scientists from the United States was instrumental in their successful first plasma in September 2006. Since then, collaborations have continued in every EAST experimental campaign.

The Collaboration Zone acts as a data depot for EAST data in the United States and therefore any approved US scientist is able to access EAST data much faster than having to retrieve data from China. The US connection is provided by ESnet.

Given the highly successful collaboration with EAST, it is anticipated that this will continue into the foreseeable future. As with all operating tokamak, the amount of data will only increase over time placing more severe demands on the associated network and computer infrastructure.

5.2.2.5.2 DIII-D at GA to MIT PFSC

Information on DIII-D can be found in Section 5.3.

5.2.2.6 Software Infrastructure

A variety of software is used for remote collaboration:

- **MDSplus:** Written by a team led by MIT, the MDSplus Data Acquisition and Management software suite is used by a number of fusion devices around the world and remains under active development. It is a read/write repository for EAST and it is this bulk data in the form of Tree files that are moved to an MDSplus Server in the Collaboration Zone.
- The Aspera high-speed Fast and Secure Protocol (FASP) data transfer protocol is used to facilitate the bulk movement of MDSplus Tree files from EAST to the GA Collaboration Zone.
- The Aspera watchfolder daemon is used to mirror the source file system for new MDSplus Tree Files and transfer them to the target file system within the GA Collaboration Zone using the FASP protocol. The daemon monitors each new file 'cooldown' period, so it knows that the file I/O is complete. Additionally, an MD5 hash checksum is compared in order to transfer only modified files in the filesystem, saving bandwidth.
- A Redis in-memory data structure store at EAST is used to transmit in near-real-time data to a Redis server in the GA Collaboration Zone. All persistent data is stored in the MySQL relational database. The real-time information includes pulse information (latest number, time of, length), plasma current, and the data transfer status.
- The messaging system RabbitMQ is used to coordinate the data transfer process between multiple machines and to allow for the collection of statistics and status information. A portion of the data transfer sequence is automated so that data arrives in the collaboration zone in a timely manner. The remaining data set must be requested by end users via the web portal.

Once the data arrives in the collaboration zone, the software and methodology of analysis is similar for many other tokamaks. Users access the data via the MDSplus server in the collaboration zone through custom software typically written either in IDL or Python.

The software environment is not anticipated to change within the next five years with the possible exception of Aspera. If the network connection into EAST rises to 10 Gbps (see Section 5.2.2.9), then a re-examination of transfer software will be undertaken to determine if the commercial Aspera solution still makes the most sense.

5.2.2.7 Network and Data Architecture

5.2.2.7.1 EAST to GA RCR

The LAN for GA has been described extensively in Section 5.3.2.7. For the GA collaboration zone, a second 10 Gbps network segment that comes directly from the ESnet border router is used. There is no firewall in the path to or from the collaboration zone. Instead, based on ESnet's Science DMZ model, ACLs are used on the collaboration zone router allowing only approved IP addresses (both IPv4 and IPv6) to connect to systems in the collaboration zone.

5.2.2.7.2 DIII-D at GA to MIT PFSC

The MIT PFSC network is described in Section 5.4.2.7.

5.2.2.8 Cloud Services

Cloud resources are not actively used to support this use case, beyond the aforementioned use of potentially cloud-based communication tools.

5.2.2.9 Data-Related Resource Constraints

For the GA RCR and the Collaboration Zone as it relates to remote operation of EAST, the constraints associated with bulk data transfer speed are the most pressing. There are two main issues associated with data transfer. The first is that EAST connection to the Chinese WAN is limited to no more than 1 Gbps. The second is that the Aspera software used for bulk data transfer is a commercial software package whose license cost rises as the network throughput increases.

For the 1 Gbps EAST network limitation, discussions have been ongoing on how this network speed might be increased. This is of course a local matter for EAST staff to work through, but does involve a variety of issues including cost. When the present capability for the Collaboration Zone was deployed in 2015, Aspera was the clear winner as far as performance, ease of use, and robustness. In the intervening time, technology has evolved, and there may be clear open-source substitutes so that the increased cost of Aspera may not be an issue.

Assuming in the future that 10 Gbps is available from EAST to the GA Collaboration Zone, the desire will be to consume a significant fraction of that bandwidth for bulk transfers. DIII-D's ESnet connection is also 10 Gbps, so in the event that DIII-D and EAST are running at the same time, there will most likely not be enough available bandwidth. Thus, the request for a second 10 Gbps ESnet connection to DIII-D also has a significant role to play in the remote operation and control of EAST.

5.2.2.10 Outstanding Issues

There are no additional issues to report at this time.

5.2.2.11 Case Study Contributors

Remote Control and Operation of Fusion Facilities Representation

- Jerry Hughes[1], MIT PSFC

1 jwhughes@psfc.mit.edu

- Raffi Nazikian[2], PPPL
- David Schissel[3], GA

ESnet Site Coordinator Committee Representation

- Scott Kampel[4], PPPL
- Jeff Nguyen[5], GA
- Brandon Savage[6], MIT PSFC

2 rnazikia@pppl.gov
3 schissel@fusion.gat.com
4 skampel@pppl.gov
5 nguyend@fusion.gat.com
6 bsavage@psfc.mit.edu

5.3 GA: DIII-D National Fusion Facility

5.3.1 Discussion Summary

GA has been an international leader in magnetic fusion research since the 1950s. The DIII-D National Fusion Facility, operated by GA for the US DOE, is the largest magnetic fusion research facility in the US. DIII-D research has delivered multiple innovations and scientific discoveries that have transformed the prospects for fusion energy.

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

- As the FES community continues to adopt new approaches to computational analysis, there has been increased scrutiny on which mechanisms are scalable and work best for different types of workflows. Traditionally it has been the case that most analysis is done "closer" to where the experimental data resides rather than transferring data directly. In this paradigm, a user may be sitting at a site with ample local computational resources, but invokes software that runs "remotely" at a location that houses an instrument and dataset. Tools such as MDSPlus facilitate this interaction, and it is expected to remain an important use case to support in the future.
- There has been some overall FES community interest in cloud services. Some use cases are easier to approach, and could be adapted to a cloud with minimal modifications; others require study to understand the technical costs that would be associated. GA has investigated some cloud providers as a way to manage backup data, and some limited analysis use cases.
- The EAST in Hefei, China is a significant international facility that FES community members, such as GA utilize. Operational considerations such as data mobility to and from this facility, have been known to use IPv6 communications protocol because it affords higher levels of performance. Maintaining IPv6 peering across ESnet infrastructure, and with international partners, is critical to the process of science for these interactions.
- Recent advancements by the Globus project at the University of Chicago may allow operation utilizing the IPv6 protocol. If possible, this would open up an opportunity for GA to consider use of this tool for data mobility in their ongoing collaboration with the EAST in Hefei, China.
- ESnet connectivity is critical for FES facilities, and backups and capacity augmentations will be required in future years to ensure continuous operation. GA has a 10G WAN connection to ESnet, and a 1G WAN backup connection through a commercial provider. Recent events, including a fiber cut in June 2021, have severely affected the ability of GA to perform daily operations and upgrading the backup connection to support 10G to ESnet is viewed as a critical requirement to science productivity.
- The current 10G ESnet connection for GA is critical for the facility's

operation, and must be maintained at all costs. In addition to the ability to access R&E connected facilities domestically and internationally, commercial peering to enable cloud services that support storage, audio, and video are critical to the process of science.

- ESnet will work with GA to perform tests associated with the DME: a framework that evaluates the ability of a facilities data architecture to be responsive to scientific data challenges.
- The overall operation time of GA's DIII-D will remain similar for the next five years, and it is anticipated that the rate of acquiring new data will continue to increase. From 2010 to 2020 the total amount of DIII-D data increased by an order of magnitude.
- The ability for ASCR facilities to address an FES multi-facility workflows requires addressing several key areas. A working group consisting of ASCR HPC Facilities, ESnet, ESCC members, and FES community members is recommended to study some of the gaps. GA has experience in this area, and would participate.

5.3.2 GA: DIII-D National Fusion Facility Case Study

The DIII-D National Fusion Facility, operated by GA for the US DOE, is a world-leading research facility that is pioneering the science and innovative techniques that will enable the development of nuclear fusion as an energy source for the next generation.

DIII-D is the product of evolving fusion research at GA going back to the 1950s. Early tokamak designs, starting in the 1960s, were circular in cross-section, but GA scientists developed the "doublet," a configuration with an elongated hourglass-shaped plasma cross-section. The Doublet I, II, and III tokamaks in the 1970s and 1980s showed that this approach allowed for a hotter and denser stable plasma. Further research led to a modification of Doublet III in the mid-1980s to DIII-D's current D-shaped cross-section. Successes with this configuration inspired many other devices to adopt the D-shape, including JET (UK), TCV (Switzerland), ASDEX-U (Germany), JT-60U (Japan), KSTAR (Korea), and EAST (China).

5.3.2.1 Background

The DIII-D National Fusion Facility at GA's site in La Jolla, California is the largest magnetic fusion research device in the United States. The research program on DIII-D is planned and conducted by a national (and international) research team. The mission of DIII-D National Program is to establish the scientific basis for the optimization of the tokamak approach to fusion energy production. The device's ability to make varied plasma shapes and its plasma measurement system are unsurpassed in the world. It is equipped with powerful and precise plasma heating and current drive systems, particle control systems, and plasma stability control systems. Its digital PCS has opened a new world of precise control of plasma properties and facilitates detailed scientific investigations. A significant portion of the DIII-D program is devoted to ITER requirements including providing timely and critical information for decisions on ITER design, developing and evaluating operational scenarios for use in ITER, assessing physics issues that will affect ITER performance, and training new scientists for support of ITER experiments.

DIII-D's open data system architecture has facilitated national and international participation and remote operation. During experimental operation, experimental data and analyzed data is stored locally and assimilated by the team conducting the experiment throughout the day. Data is accessed via a client/server interface with data collected directly from the experiment stored as read-only and analyzed data as read-write. All DIII-D data is kept indefinitely and is backed up locally and to an off-site colocation facility.

GA also conducts research in theory and simulation of fusion plasmas in support of the Office of FES overarching goals of advancing fundamental understanding of plasmas, resolving outstanding scientific issues and establishing reduced-cost paths to more attractive fusion energy systems, and advancing understanding and innovation in high-performance plasmas including burning plasmas. The theory group works in close partnership with DIII-D researchers in identifying and addressing key physics issues. To achieve this objective, analytic theories and simulations are developed to model physical effects, implement theory-based models in numerical codes to treat realistic geometries, integrate interrelated complex phenomena, and validate theoretical models and simulations against experimental data. Theoretical work encompasses five research areas: (1) MHD and stability, (2) confinement and transport, (3) boundary physics, (4) plasma heating, non-inductive current drive, and (5) innovative/integrating concepts. Numerical simulations are conducted on multiple local Linux clusters (multiple configurations and sizes) as well as on computers at ALCF, NERSC, and OLCF.

5.3.2.2 Collaborators

The DIII-D Program is world renowned for its highly collaborative research program that engages collaborative staff at all levels of program management and execution.

User/Collaborator and Location	Do they store a primary or secondary copy of the data?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of dataset? (report in bytes, e.g. 125GB)	Frequency of data transfer or download? (e.g. ad-hoc, daily, weekly, monthly)	Is data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g. difficult tools, slow network)?
See Figure 5.3.1.	No	Client/Server API	45 GB/pulse, 2 TB/week	Demand is cyclical with DIII-D pulses but continual	Y, via Client/Server API	No

Table 5.3.1 – GA Data Relationships

The DIII-D Research Plan is founded on the extensive expertise of the research staff that comprises the DIII-D Research Team, which includes experimentalists and theoreticians from universities, national laboratories, and private industry around the world (see Figure 5.3.1). Approximately 830 researchers from around the world are active users of DIII-D data. These team members are from 137 institutions including 76 Universities (40 United States, 36 international), 33 National Laboratories (7 United States, 26 International), and 17 High Technology Companies.

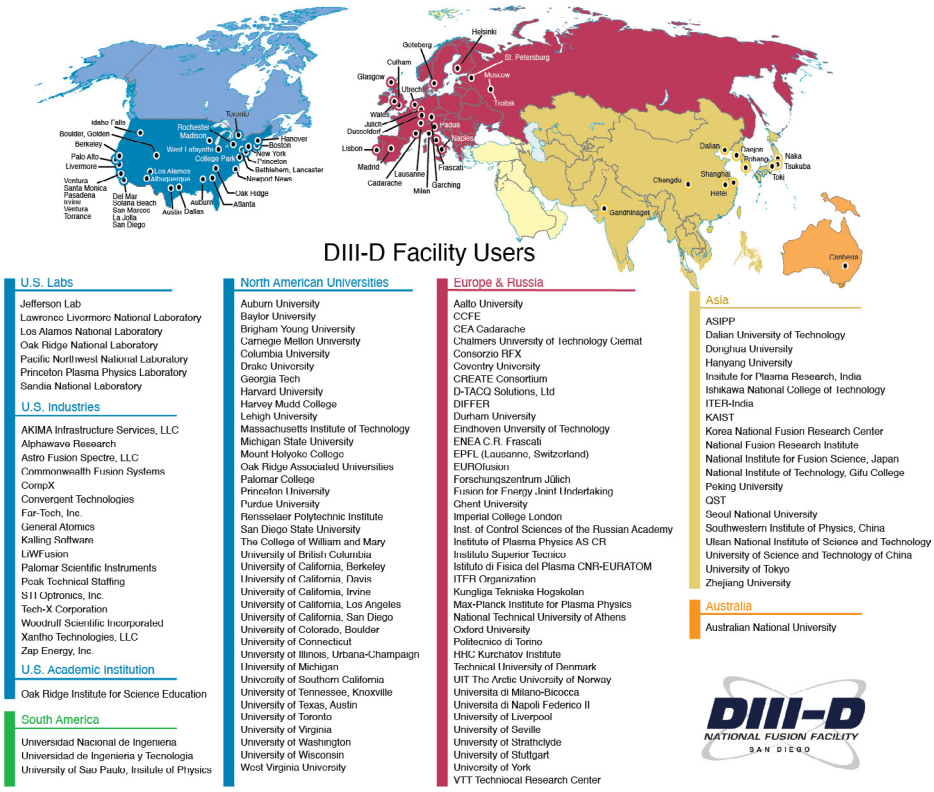


Figure 5.3.1 – DIII-D National Team

5.3.2.3 Instruments and Facilities

The DIII-D Tokamak, along with the approximately 100 diagnostics that are used to measure attributes of the plasma or the hardware infrastructure, comprise the DIII-D National Fusion Facility. The majority of the acquired data comes from these diagnostics. The Facility is typically being upgraded, diagnostics are added, and existing diagnostics are enhanced to acquire more data.

GA's connection to ESnet is currently 10 Gbps and the backup ESnet connection is at 1 Gbps. Discussions have been ongoing with ESnet for several years to upgrade the backup connection to 10 Gbps and to operate both connections simultaneously in a production environment.

The majority of the computing and storage devices are connected by a switched 10 and 1 Gbps Ethernet LAN. Network connectivity between the major computer building and the DIII-D facility is 20 Gbps. The major data repositories for DIII-D comprise approximately 1 PB of online storage with metadata catalogs stored in a relational database. Network connectivity to offices and conference rooms is at 1 Gbps on a switched Ethernet LAN. There are approximately 2000 devices attached to this LAN with the majority dedicated to the DIII-D experiment.

Like most operating tokamaks, DIII-D is a pulsed device with each pulse of high-temperature plasma lasting on the order of 10 seconds. There are typically 30 pulses per day and operates approximately 16-22 weeks per year. For each plasma pulse, up to 10,000 separate multi-dimensional measurements are acquired and analyzed

representing approximately 45 Gigabytes of data. Several proposed large-data diagnostics and the push to operate closer to the 20-22 weeks per year will only increase the amount of data generated.

The experimental data is accessed both locally and over the WAN through a client/server data management system. Rapid access to the experimental data, usage of data analysis tools, as well as audio/video-based collaboration tools creates significant network traffic during the experiment. DIII-D's data is made available to remote collaborators through several avenues. The first methodology is direct access to the data repositories through the secure client/server interface. The second technique is VPN that places the remote computer on the DIII-D network allowing full access to all services. The final technique is SSH access to a specific gateway computer and then from there SSH access to other nodes on the internal network. To facilitate the speed of interacting with graphical user interface (GUI) programs, the remote desktop software No-Machine is supported. DIII-D operates a Linux-based computational cluster (500 cores) with a Slurm scheduling system (interactive work is allowed as well). In addition, metadata is available via a variety of SQL and NoSQL databases through direct API calls or dynamic web pages.

Although the operation time of DIII-D will remain similar for the next five years, it is anticipated that the rate of acquiring new data will continue to increase. From 2010 to 2020 the total amount of DIII-D data increased by an order of magnitude. To keep up with this demand plus the increased usage of collaborative technologies, even within the local campus, the reach of 10 Gbps within the LAN is ever increasing. This is one of the main reasons that upgrading the backup circuit to 10 Gbps and operating it in parallel to the main connection is so important.

5.3.2.4 Process of Science

Throughout the experimental session, hardware/software plasma control adjustments are debated and discussed amongst the experimental team and made as required by the experimental science. The experimental team is typically 20–40 people with many participating from remote locations. Decisions for changes to the next plasma pulse are informed by data analysis conducted within the roughly 20 minute between-pulse interval. This mode of operation requires rapid data analysis that can be assimilated in near real time by a geographically dispersed research team.

The pulsed nature of the DIII-D experiment combined with its highly distributed scientific team results in WAN traffic that is cyclical in nature. Added onto this cyclical traffic is a constant demand of general scientific data analysis and the collaborative services mostly associated with Internet-based video/audio collaboration services. General meetings are conducted using Zoom and DIII-D operations includes Zoom but adds Discord as well. As the collaborative activities associated with DIII-D continue to increase, there is an increasing usage of collaborative visualization tools by off-site researchers that requires efficient automatic data transfer between remote institutions.

The scientific staff associated with DIII-D is very mobile in their working patterns. This mobility manifests itself by traveling to meetings and workshops, by working actively on other fusion experiments around the world, and by working from home. For those individuals that are off-site yet not at a known ESnet site the ability to efficiently

transition from a commercial network to ESnet becomes very important. Therefore, ESnet peering points are becoming a critical requirements area.

While the operation of DIII-D is expected to remain similar for the next five years, scientists will be increasingly focused on remote collaborations between DIII-D and other facilities. Presently, the DIII-D scientific team is actively involved in operations for the EAST tokamak in China and the KSTAR tokamak in the Republic of Korea. Over the next 5 years, the operation of these tokamaks will become even more routine and it is anticipated that the remote participation of DIII-D scientists will increase. These tokamaks will be operating at the same time as DIII-D, putting an increased strain on the WAN. Therefore, how ESnet peers, particularly with China and South Korea, will become increasingly important.

5.3.2.5 Remote Science Activities

For DIII-D, the need for real-time interactions among the experimental team and the requirement for interactive visualization and processing of very large simulation data sets are challenging. The remote aspect for DIII-D is collaborating scientists who are off-site. Otherwise, the entire DIII-D national Fusion Facility is located within one large LAN.

Related to scientific activities and access to resources, DIII-D makes no distinction between an on-site and remote team member. Thus, the process of the science described in Section 5.3.2.4 applies to both local and remote scientists.

5.3.2.6 Software Infrastructure

There are six main data repositories for the DIII-D National Fusion Facility:

- **PTDATA**: Written in the 1980s, it is unique to DIII-D and is used to manage data acquired directly from the DIII-D experiment (raw data). It is a write-once repository with a client/server API. The data resides on JBODs utilizing Zettabyte file system (ZFS) and connected to a Linux-based server.
- **MDSplus**: Written by a team led by MIT, the MDSplus Data Acquisition and Management software suite is used by a number of fusion devices around the world and remains under active development. It is a read/write repository and for DIII-D the majority of the data within is derived through analysis of raw data. The data resides on JBODs utilizing ZFS and connected to a Linux-based server.
- **Object Storage**: Presently using a DDN WOS appliance, the main data set contained within this object storage array are fast camera data taken during DIII-D experimental operations. Data access is via an S3 API layer and erasure encoding is for data protection.
- **User Files**: Multiple NFS mounts of ZFS-based files systems allow the DIII-D team to write code and analyze data on a large computational cluster.
- **Relational Databases**: Multiple relational databases (MS SQL, MySQL, etc.) are used for a variety of purposes. However, the main usage is for scientific metadata related to raw, analyzed data, and camera data. A large number of web applications running several Apache web servers use the relational database infrastructure in their backend.

- **NoSQL database**: A recent addition to the DIII-D data infrastructure, NoSQL databases (e.g., Redis) are used mainly to securely serve real-time data from the protected tokamak sub-networks to local and remote DIII-D team members. This data may not be contained within the PTDATA and MDSplus environments. For example, real-time plasma control signals or plant status information.

The scientific process in the wide-area environment is very similar to that in the local area environment. Tools that are used to access and manage data locally are the same that are used for remote scientists. Remote scientists can either make client/server calls to PTDATA or MDSplus to retrieve data locally or log into DIII-D resources and only transfer X Window System traffic over the WAN. Large data file transfers are not the workflow utilized by the DIII-D Team.

Utilizing either their own local or DIII-D compute resources, the DIII-D community writes their own analysis codes utilizing API calls to retrieve data. Historically, the commercial software IDL and Fortran have been used to create custom data analysis software. With the explosive growth of Python for scientific data analysis, a significant increase in the usage of Python has been seen within the DIII-D community.

The software infrastructure at DIII-D is not anticipated to change dramatically in the next five years.

5.3.2.7 Network and Data Architecture

The LAN is segmented into multiple Virtual-LANs (VLANS) where resources are logically grouped together (see Figure 5.3.2). A traditional tiered switching architecture is deployed with a six-member collapsed-core fabric of switches that provide high-speed switching between segments. Per-VLAN Spanning Tree (PVST+) is utilized to avoid switching loops. Some systems are placed into private LANs with layer 2 switching provided but not layer 3 routing. For cybersecurity, Network Access Control (NAC) is a requirement prior to a new user being admitted to the wired LAN and Wireless LAN. Users provide authentication from an approved device and are dynamically placed into their proper network segment. All inter-VLAN routing occurs at the core and at the enterprise firewall. Wireless LAN deploys lightweight access points with the controller being from one of the firewall instances. A high-availability pair of Fortigate firewalls running multiple virtual instances are deployed for North-South, East-West, and Tokamak traffic to further isolate and fine tune access policies.

DIII-D's link to the WAN is provided by ESnet with a 10 Gbps primary circuit and 1 Gbps backup circuit. Peak traffic as recently measured by ESnet approaches 2 Gbps to the WAN edge but given the pulse cycle of DIII-D experimental operations and if combined with EAST operation, this peak could be higher (daily average is 0.5 Gbps).

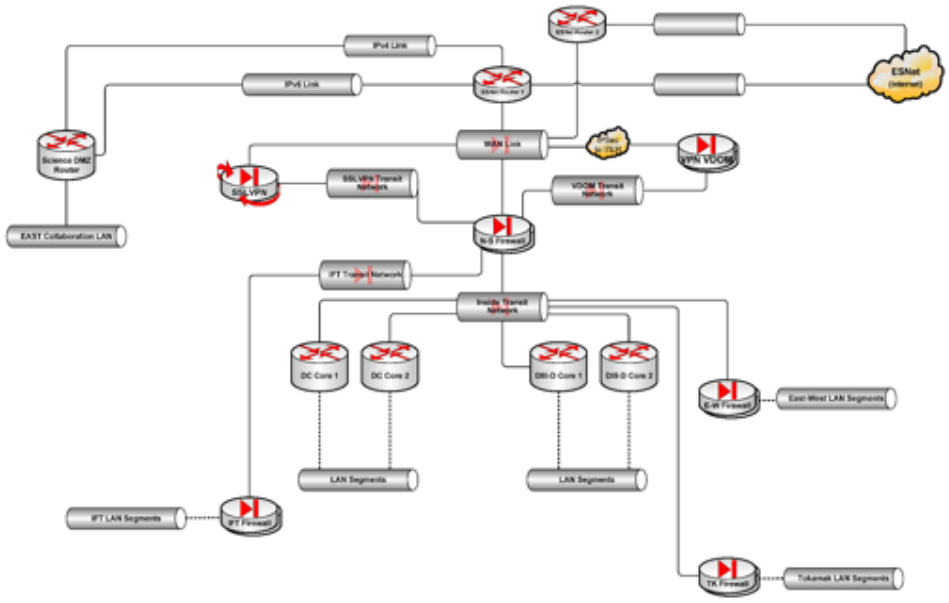


Figure 5.3.2 – GA/DIII-D Local Area Network

The edge firewalls form a Border Gateway Protocol (BGP) peer with a local ESnet router, which sends the egress traffic to ESnet's Sunnyvale location. Two different architecture models allow connectivity to DIII-D's remote collaborators, either via a direct login to a NoMachine server or via an IPsec VPN tunnel.

Local Network architecture follows a collapsed core, with core fabric members in the data center and in the DIII-D Facility building, roughly 1 mile apart geographically. The two physical locations are connected via underground fiber. NAC is implemented for end user access, and Science DMZs are applied to sequester sensitive systems where applicable. The policy-based firewall is responsible for permitting access between network segments. The most common edge links are 1 Gbps, though legacy equipment may be limited to 100 Mbps based on technological constraints. Performance monitoring is accomplished using such tools as SolarWinds, Icinga, and NetMRI.

A Science DMZ is deployed via a second network segment directly from the ESnet border router. This capability is used exclusively for DIII-D's collaboration with the EAST tokamak in China where large bulk data transfers of EAST data to this Science DMZ are accomplished when EAST is operating.

Based on the anticipated growth of DIII-D's usage of the WAN within the next 5 years, it is expected that there will be a need to utilize two 10 Gbps network connections to ESnet for production services.

5.3.2.8 Cloud Services

There are several commercial cloud services that are or will become critical to the operation of the DIII-D National Fusion Facility. First on the list is Zoom, which has replaced the functionality of the ESnet Collaboration Services. A Zoom Meeting Connector is deployed in the DIII-D Data Center so that all meeting traffic including

video, voice, and data sharing goes through an on-premise Zoom Meeting Connector. However, user and meeting metadata are managed in Zoom's public cloud. Therefore, this is a hybrid cloud environment and is critical to allow the distributed DIII-D community to meet and scientifically collaborate.

A recent addition to DIII-D's cloud service is the usage of Discord during DIII-D operations. Voice, video, text chat, and shared screens are all used to mimic the DIII-D control room environment. Without Discord, it would have been much harder to operate during the COVID-19 pandemic. More detail is provided on how DIII-D has operated throughout the stay-at-home-order in Section 5.3.2.10.

Although not exactly a cloud, DIII-D has transitioned its off-site disaster recovery strategy to use a collocation service provided by Hurricane Electric in their Fremont datacenter. DIII-D owned hardware (network switch, server, storage arrays) have been placed in a dedicated rack and 1 Gbps connectivity is provided to ESnet. Data is replicated to this off-site storage periodically, 24/7 and 365 days a year. Access to this Hurricane Electric colocation center is therefore critical for DIII-D to remain current in its disaster recovery plan.

For a number of years, the DIII-D organization has run GitLab services on the LAN providing a central code repository for software development and version control. This service is migrating to the cloud-based GitHub environment, and given how central to the scientific enterprise is code development, access to GitHub will be critical to the DIII-D facility.

Presently, the DIII-D Facility is in the process of deploying some Microsoft Office 365 Services via an Azure cloud tenant. SharePoint will be used for enhanced document sharing and collaboration capability and the transition of electronic mail to Exchange Online in Office 365 is beginning to be examined. It is likely that access to this Microsoft cloud will become critical to the facility's operation.

5.3.2.9 Data-Related Resource Constraints

There are no data-related resource constraints, now or anticipated in the future, that will affect DIII-D's scientific productivity. This assumes that the trajectory of technology refreshing, including expanding storage and computing capability, continues to follow the planned path.

5.3.2.10 Outstanding Issues

The most critical outstanding issue is that GA's backup connection to ESnet is only at 1 Gbps. When the primary 10 Gbps link to the WAN fails, the speed of the backup circuit is not able to handle all of the load as peaks above 1 Gbps are routinely observed. In addition, the recent transition to Hurricane Electric's Fremont datacenter for DIII-D's disaster recovery strategy places a nightly load on the WAN that is greater than 1 Gbps. Figure 5.3.3 shows WAN network utilization during the latest primary circuit outage on June 4, 2021. Discussions have been ongoing for a number of years on different possible solutions but the time has come to transition the backup circuit to 10 Gbps. The vision is, that in this environment, DIII-D would operate with both 10 Gbps connections simultaneously in a production environment.

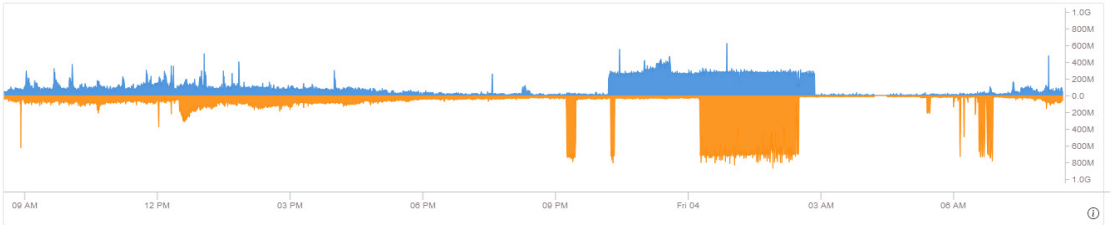


Figure 5.3.3 – GA's backup connection to ESnet (1 Gbps) saturates when the primary circuit is down (June 4, 2021). Blue is to site, orange is from site.

The other issue is how the COVID-19 pandemic and stay-at-home order affected how DIII-D operates. DIII-D experiments are routinely conducted throughout the year with the control room being the focus of activity. Although experiments on DIII-D have involved remote participation for decades, and even have been led by remote scientists, the physical control room always remained filled with ~40 scientists and engineers all working in close coordination. The severe limitations on control room occupancy required in response to the COVID-19 pandemic drastically reduced the number of physical occupants in the control room to the point where DIII-D operations would not have been possible without a significantly enhanced remote participation capability.

Leveraging experience gained from GA operating EAST remotely from San Diego, the DIII-D Team was able to deploy a variety of novel computer software solutions that allowed the information that is typically displayed on large control room displays to be available to remote participants. New audio/video solutions using Discord were implemented to mimic the dynamic and ad-hoc scientific conversation that are critical in successfully operating an experimental campaign on DIII-D. Secure methodologies were put into place that allowed control of hardware to be accomplished by remote participants including DIII-D's digital PCS. Enhanced software monitoring of critical infrastructure allowed the DIII-D Team to be rapidly alerted to issues that might affect operations. Existing tools were expanded and their functionality increased to satisfy new requirements imposed by the pandemic.

Finally, given the mechanical and electrical complexity involved in the operation of DIII-D, no amount of software could replace the need for "hands on hardware." A dedicated subset of the DIII-D team remained on-site and closely coordinated their work with remote team members which was enhanced through extensions to the wireless network and the use of tablet computers for audio/video/screen sharing. Taken all together, the DIII-D Team has been able to conduct very successful experimental campaigns in 2020 and 2021.

Because of DIII-D's history of supporting remote participants, there was not the need for a major network and data-sharing overhaul. The usage of Discord for real-time ad-hoc DIII-D facility communication meant another cloud service became critical (note, Discord is not available in some countries). New data, not traditionally shared with remote participants, was made available. The total number of these different data sets was large yet taken together the total data rate is not large and thus not a major impact on the network. Remote login and VPN capability capacity were increased, but followed the existing architectural deployment. The reach of the wireless LAN was significantly extended and the deployment seamlessly joined the existing networking capability.

Given the large international DIII-D Team, any increased peering with major Internet providers worldwide will be helpful to the DIII-D National Fusion Facility. In the past, shorter paths and better peering helped with increased network throughput and decreased latency. (E.g., ESnet peering with Gloriad in November 2010 decreased the latency about 25% between DIII-D and EAST).

5.3.2.11 Case Study Contributors

GA: DIII-D National Fusion Facility Representation

- Dr. Richard Buttery[1], GA
- David Schissel[2], GA

ESnet Site Coordinator Committee Representation

- Jeff Nguyen[3], GA

1 buttery@fusion.gat.com
2 schissel@fusion.gat.com
3 nguyend@fusion.gat.com

5.4 MIT PSFC

5.4.1 Discussion Summary

The MIT PSFC seeks to provide research and educational opportunities for expanding the scientific understanding of the physics of plasmas, the "fourth state of matter," and to use that knowledge to develop useful applications. The central focus of PSFC activities has been to create a scientific and engineering base for the development of fusion power. A diverse set of non-fusion plasma research areas and related technologies and applications are also actively pursued at the PSFC.

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

- FES use of cloud services is still being explored. Some use cases are easier to approach, and could be adapted to a cloud with minimal modifications; others require study to understand the technical costs that would be associated. Alcator C-Mod data, housed at MIT, is being explored as a possible cloud use case. There are concerns regarding if the cloud will be scalable enough to address some of the tools that currently operate on this data - many of which rely on smaller transactions to extract portions of a data set versus an entire bulk or streaming use case.
- MDSplus remains critical to the operation of the FES community, and is widely used and deployed at experimental and analysis facilities. Modifications to the core software have helped FES keep pace with increases in networking capabilities, and computational availability.
- The FES community has adopted approaches where computational analysis is often done "closer" to where the experimental data resides rather than transferring data directly. In this paradigm, a user may be sitting at a site with ample local computational resources, but invokes software that runs "remotely" at a location that houses an instrument and dataset. Tools such as MDSplus facilitate this interaction, and it is expected to remain an important use case to support in the future.
- FES simulation and theory workflows do not utilize MDSplus, and often rely on other tools that are native to the HPC facilities to accomplish data mobility tasks (e.g., Globus/GridFTP). Not all FES experimental facilities have similar hardware or software capabilities available, which can affect the efficiency of data transfer as a part of these workflows.
- ESnet can assist MIT PSFC adopt hardware and software approaches that are native to HPC facilities to accelerate simulation and theoretical FES workflows that require data mobility. These solutions can be to install and adopt known tools (e.g., Globus, MRDP), or potentially offer services operated by ESnet to foster data mobility improvements.
- ESnet connectivity is critical for FES facilities, and backups and capacity augmentations will be required in the future years to ensure continuous operation. MIT PSFC has a 1G ESnet connection, through the MIT

campus, but is interested in upgrading due to increased use cases that rely on external connectivity to support remote computing and storage, as well as increased levels of remote observation use cases. Upgrading the ESnet connection implies working with the MIT campus to upgrade LAN and MAN connectivity.

- The FES community is exploring ways that cloud-provided storage and computation could be integrated into scientific workflows, particularly at facilities that are not able to scale local resources either due to cost, space, or lack of expertise to operate long-term storage pools. Investigations are underway to understand the costs and usability for FES workflows.
- MIT PSFC's Alcator C-Mod data archive is approximately 150 TB in size, and remains heavily accessed by the FES community. There are ongoing efforts to understand how this can be kept active in the coming years, as the hardware that provides the archive will require maintenance or augmentation. Upgrading local hardware and software to modernize the portal, or migration of the data to a dedicated facility remain possibilities.
- The FES community should explore ways to better utilize computational resources that exist at collaborator sites, as well as DOE HPC facilities, as future research depends on the ability to effectively and efficiently utilize computational resources and increasing volumes of data.
- ESnet will work with MIT PSFC to perform tests associated with the DME: a framework that evaluates the ability of a facilities data architecture to be responsive to scientific data challenges.

5.4.2 MIT PSFC Case Study

PSFC researchers study the use of strong magnetic fields to confine plasma at the high temperatures and pressures required for practical fusion energy. This research is conducted using on-site experimental facilities, theory and simulation, and collaboration with researchers at other facilities. PSFC scientists, students, and engineers perform experiments and develop technologies to confine and heat the plasma and to manage the interactions between the plasma and the reactor materials.

5.4.2.1 Background

5.4.2.1.1 MFE

In the area of MFE the PSFC collaborates on both domestic and international experimental facilities, with research objectives aimed at preparing for burning plasma experiments on the horizon: readiness for SPARC, ITER and pilot plants are significant drivers for the research. Awards from FES support collaboration on DIII-D, ASDEX Upgrade, WEST, JET, TCV, W7-X etc. Significant science theme areas include (1) RF Actuators for Fusion, (2) Disruption Science, (3) Science of ELM-suppressed Regimes, (4) Integrated Studies of the Tokamak Core and Edge, (5) Transport Physics and Profile Prediction and (6) Material Assessment for Compact Fusion Power Plants and Plasma-Materials Interactions. Data streams originate from

- Alcator C-Mod data archive
- smaller devices still operating at PSFC

- off-site instruments/facilities

Raw data support analysis in support of the theme areas above, and others. They are accessed by internal researchers and external collaborators. Raw and analyzed data also support efforts to perform physics-based simulations. Data, especially from mid- to large fusion devices, may be utilized for analysis/simulation several years after the particular experiment is complete. Manuscripts continue to be written about Alcator C-Mod, five years after its last plasma discharge was run.

5.4.2.1.2 Alcator C-Mod Data Archive

Data from more than twenty years of Alcator C-Mod operation provide information in a unique region of parameter space (high magnetic field, high-power density, high current density, short current relaxation time, high electron density, high neutral opacity and high absolute pressure, with large edge temperature and density gradients, produced solely by RF actuators and free from core particle and momentum sources) no longer accessible to currently operating tokamaks. This information forms a critical contribution to a variety of studies, including multi-machine studies sanctioned by the ITPA in support of ITER. Access to the C-Mod archive is desired by international and domestic collaborators performing research on ITPA tasks, including the leaders of joint experiments. C-Mod data is also sought perennially for contributions to US DOE Joint Research Targets in MFE, which aspire to apply results from C-Mod, DIII-D and NSTX/NSTX-U to common tokamak physics problems. Finally, a number of theoretical and modeling collaborators depend on C-Mod data to execute their existing projects. In short, a critical community need is served through the preservation and distribution of C-Mod data.

5.4.2.1.3 MDSplus

MDSplus is a collaborative software development project that facilitates the data acquisition and data management tasks, and creates a platform that is used by most US-based FES experiments, and many non-domestic experiments. MIT leads the development effort with major collaborators from RFX/Padua and W7X/Greifswald.

The software provides a network API to access fusion data sets, as well as metadata storage so that data retains usefulness over long time periods. A primary use case for MDSplus is archival data sets, including Alcator C-Mod (1991--2016) which is still online and accessible.

There are thousands of downloads/installs per year of the MDSplus package.

5.4.2.1.4 Theory and Computation

The Theory and Computation at MIT PSFC has a number of investigations that span a number of physics topics (e.g., local gyrokinetic simulations of turbulence, integrated simulation of RF actuators). This group often produces large sets of simulation data, that are typically generated and stored at off-site facilities (e.g., NERSC, Massachusetts Green High-Performance Computing Center [MGHPCC]). Transfer of data and analysis occurs locally at PSFC, using tools such as SCP.

5.4.2.2 Collaborators

A number of collaborators maintain a relationship with the MIT PSFC. Some examples of collaborating institutions who have contributed to analysis/simulation of the C-Mod

data archive in the previous three years:

- GA
- PPPL
- ORNL
- Lawrence Livermore National Laboratory (LLNL)
- The College of William & Mary
- The University of Texas
- University of California, San Diego (UCSD)
- Max Plank IPP, Germany
- Culham Centre for Fusion Energy (CCFE)

Collaborators with the Theory division include:

- CompX
- Tech-X
- Lawrence Livermore National Laboratory (LLNL)
- Lodestar Research
- ORNL
- PPPL
- Rochester Polytechnique Institute (RPI)
- The University of Georgia
- The University of Illinois Urbana Champaign
- The University of Texas
- The University of Maryland
- LBNL

Data from most PSFC MFE projects, including C-Mod, are accessed using the MDSplus system (www.mdsplus.org), which provides data access through a simple application program interface (API) adapted for many common programming languages. Remote access is provided by MDSIP, a software-based network layer that allows the API to store or retrieve data using the internet IP protocol.

Data from the Center for Science and Technology with Accelerators and Radiation (CSTAR) facility with DIONISOS for material characterization is stored locally on the control computer hard drive. The data files can be backed up and shared using the MIT-sponsored Dropbox license and can be protected with further backup using MIT's Crashplan cloud service.

5.4.2.3 Instruments and Facilities

5.4.2.3.1 PSFC Computing Infrastructure

The MIT PSFC has a substantial complement of computer equipment, local network infrastructure, software and backup capability that includes:

- Office and laboratory space with a high-speed connection to Internet2 and ESnet
- Desktop Workstations (over 100) & Servers Local Area Network (~30)
- Switched 1 Gbps Ethernet to each desktop and 10 Gbps backbone
- Data Storage:
  - Local – ~500 TB + 60 TB in mirrored backup
  - Backup – Local tape archive + Enterprise wide – CrashPlan and TSM
- Extensive Software:
  - MDSplus (developed at the MIT PSFC and supported by DOE)
  - SVN server
  - Hg server
  - IDL
  - Python
  - PHP
  - MATLAB
  - OpticStudio (Zemax) optical design software
  - COMSOL finite element analysis software
  - SolidWorks computer-aided design (CAD) software
  - SolidEdge CAD software

5.4.2.3.2 Alcator C-Mod Data Archive

MIT PFSC continues to preserve the entirety of the Alcator C-Mod data archive, including raw and analyzed data. This requires ongoing maintenance of hardware for storing and serving these data, and software tools for PSFC researchers and external collaborators to access needed data for further analysis and modeling. It is anticipated that preserving and providing these data to the community will continue indefinitely.

The archived data from 25+ years of Alcator C-Mod operations are on line and available to users. The archive is served to users by two Linux servers and is approximately 135 TB in size. Two copies of the data are kept on line to protect against both hardware failure and accidental user data modification. In addition, it is both archived and backed up using a campus provided off-site service. A relational database provides summary, index and annotations information. The current hardware supporting this consists of:

- 2 data servers with 200 TB disk arrays
- 32 user workstations
- Virtual Server for remote access
- Virtual Server for relational databases
- (desktop computers / laptops for users)
- Networking for all of the above

PSFC IT staff maintains and supports this hardware. Plans are in place to replace aging hardware. Storage is currently end of life, and will require replacement in the next year. Servers are six or more years old, and are to be replaced within 2 years. A set of workstations used to interact with the C-Mod archive, and which provide general computing needs for collaborations, will be partially replaced in each of the following 5 years.

The software used to access and analyze these data includes:

- MDSplus - maintained by DOE grant DE-SC0008737
- Python - open source
- MATLAB - provided by MIT
- IDL - licensed
- Nomachine - licensed

5.4.2.3.3 CSTAR Laboratory

The CSTAR laboratory is jointly run by MIT Department of Nuclear Science and Engineering and the MIT Plasma Science and Fusion Center (PSFC). The laboratory contains substantial equipment that supports research in plasma-material interactions and advanced materials for fusion devices. Over the next 5 years, CSTAR is proposed for a support role in expanded materials R&D for the DIII-D program.

5.4.2.3.4 RCR

The MIT PSFC has a dedicated RCR space, which was prepared and equipped in 2018 using DOE funding. The RCR has an open plan with 14 workstations available for scientific computing on a networked Linux cluster administered by the PSFC MFE division. Access is available off-site via SSH or NoMachine. Numerous computational tools such as IDL and MATLAB are maintained on the local cluster, and ready access is available to resources at off-site institutions. The facility is also equipped with extensive video-conferencing capabilities for interacting with off-site colleagues, students, and collaborators.

The RCR provides the functionality for staff and students to watch and interact with real-time displays of data and video feeds, when they are provided by off-campus fusion-experiment control rooms. In particular, direct access to DIII-D is available on this network, allowing for full access to discharge data there. Full access to the full catalog of C-Mod data is granted from these workstations, via a connection to a PSFC server running MDSplus. The cluster has robust connectivity through ESnet to SC partner institutions. The RCR is routinely used by scientists, postdocs and graduate students for connectivity with operations at off-site facilities.

The key features of the RCR include:

- Four real-time displays and 14 workstations to be used for participation with remote experiments
- A sound-isolated VC area with conference table, chairs, and large display Monitor
- Desk space and monitors for users who wish to connect laptops

- A common space with chairs that fosters discussion and the sharing of ideas and experience

With these capabilities, the RCR mitigates risks associated with host facility schedule changes, thus helping to control travel expenditures. It also allows increased participation from PSFC staff and students who may be unable to travel for various reasons.

5.4.2.3.5 HPC Facilities

The MIT PSFC has substantial HPC resources in the form of the PSFC@Engaging cluster. A GPU cluster has also been added, for both ML and gpu-accelerated physics codes. In addition, MIT has a sizable computing time allocation (45,000,000 compute hours) at the NERSC in Berkeley, CA. Example usage of NERSC computing time includes high-fidelity nonlinear gyrokinetic studies of tokamak plasmas. Computing resources for less computationally intensive activities are available on local clusters, as described below.

PSFC@Engaging

The PSFC@Engaging computational cluster consists of a 100 compute node subsystem integrated into the "Engaging Cluster," which is located at the MGHPCC in Holyoke, MA. The PSFC subsystem is operated as part of the "Engaging Cluster" with access to a 2.5 Petabyte parallel file system. The total PSFC subsystem is 3200 cores with 12.8 Terabytes of memory.

This 100 node subsystem is connected together by a high-speed, non-blocking Fourteen Data Rate (FDR) Infiniband system. This Infiniband system is capable of 14 Gbps with a latency of 0.7 microseconds. This network is non-blocking; thus, each node has immediate access to each other node as well as to the parallel file system.

Each compute node in the subsystem is configured with 2 Intel E5, Haswell-EP processors at 2.1GHz, 16 cores per processor, for a total of 32 cores per node. Each node has 128 GB DDR4 of memory with 1.0 TB on the local disk. The individual compute nodes are very similar to the compute nodes in the Cori–Haswell partition at NERSC. However, high-fidelity nonlinear gyrokinetic simulations will require the use of larger facilities such as NERSC.

PSFC GPU Cluster

In the fall of 2020, the PSFC acquired a 6 node cluster of gpu enabled computers, each of which has 4 gpu accelerator cards. Half of the gpu cards are RTX cards that are appropriate for ML tasks and half are V100 cards with double precision floating point appropriate for the development of gpu-accelerated physics codes. The nodes come equipped with a software stack of gpu enabled ML libraries and compilers.

5.4.2.4 Process of Science

The following use cases describe typical data workflows for MIT PSFC:

- Raw device data (e.g., Alcator C-Mod profiles) can be accessed via MDSplus and transferred to a computing environment for analysis. The computing could be supplied by PSFC or by off-site resources.
- Data streams from MIT PSFC can be transferred as inputs to higher order analysis framework (e.g., a transport simulation)

- Simulations can be performed with increasing levels of physics fidelity (e.g., a gyrokinetic simulation of fluxes)
- The results of analysis and simulation workflows can be interrogated (e.g., transport calculations from experiment vs simulation compared)

These workflows occur on a spectrum of different platforms, and involve data transfer in a number of different ways and utilizing different actors:

- PSFC MDSplus Servers
- PSFC MFE Workstations
- TRANSP Server at PPPL
- PSFC MFE Workstations
- NERSC CORI

5.4.2.5 Remote Science Activities

MIT PSFC researchers collaborate in a significant way on both domestic and international facilities, including DIII-D, NSTX-U, AUG, JET, EAST, WEST, TCV and W7-X.

Typical usage of off-site facilities:

- SSH connections to workstations off-site for visualizing data, performing data analysis using the off-site computing capabilities
- transfer of limited data sets for local analysis
- monitoring and control of off-site diagnostics

The latter instance can be described by example: MIT co-operates the LLAMA diagnostic on DIII-D. LLAMA "keeper" for the run day, sited at MIT, checks in on diagnostic interlocks, temperatures using Ignition process control software. Device shutter can be actuated remotely. Other actuations, such as power supplies, must be done via contacting someone on-site at DIII-D. Using tools such as ReviewPlus (in IDL) and MATLAB tools running at DIII-D, but displaying locally, diagnostic operator confirms correct signal acquisition shot to shot.

Additional instruments require some remote interaction on DIII-D, ASDEX Upgrade, TCV, and W7-X. It is expected that there will be additional instruments added on a 2-5 year horizon, as much as doubling current usage.

5.4.2.6 Software Infrastructure

5.4.2.6.1 Local and Remote Data Management

- MDSplus - local experiment data, APIs for remote experiment data.
- HDF5 - Modeling codes, Publication data attachments
- Network Common Data Form (NETCDF) - Modeling codes
- Relational Databases (SQLSERVER, MySQL)
- Web-based experiment logbooks (C-Mod, DIII-D, NSTX)

5.4.2.6.2 Data Transfer

- MDSplus
- Globus
- sftp
- scp

5.4.2.6.3 Data Processing

- MDSplus
- PiScope[1]
- python
- Paraview[2]
- MATLAB
- idl

5.4.2.6.4 Future Tool Use

- **Present to 2 years:**
  - The technologies will largely remain the same. New capabilities of the tools will be exploited as they are available. There is a definite shift, from proprietary software (MATLAB, idl) to python-based tools
- **Next 2-5 years:**
  - This is an evolution of the previous
  - Larger data sets
  - Better remote data access
  - Better remote / cloud computing / VDI graphics
  - Lifting of covid restrictions will allow us to get back to using shared collaboration spaces at the PSFC (RCR), Conference rooms (with video conferencing …)
- **Beyond 5 years:**
  - Possible shift to off premise computing (cloud)
  - Possible shift to off premise data storage (cloud)
  - Highly interactive remote collaboration

5.4.2.7 Network and Data Architecture

The MIT PSFC Local Area Networking environment is a mixture of 10 Gbps LAN connections, and 1 Gbps connectivity to workstations. Typical usage within the environment is a mixture of enterprise use cases (e.g., web traffic, email, etc.) and scientific use cases (e.g., remote access, data transfer, etc.). The MIT PSFC Wide-Area Networking environment is limited to 1 Gbps that is delivered via a connection on the

1 https://piscope.psfc.mit.edu/index.php/Main_Page
2 https://www.paraview.org/

MIT campus. MIT is connected to the Northern Crossroads (NoX), where the National and International connections can be reached (e.g., ESnet, Internet2, etc.). See Figure 5.4.1 for more information.

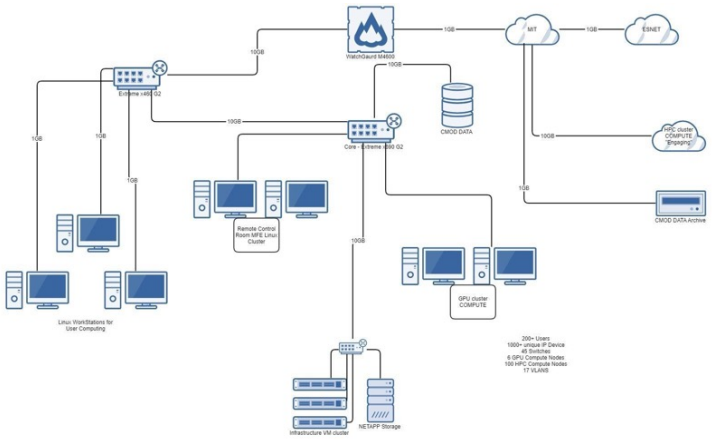


Figure 5.4.1 – MIT PSFC Network Diagram

The coming years will see changes to the computing infrastructure that may affect networking capabilities. Enterprise storage solutions are being adopted, along with off-site disaster recovery

replication, which will affect the network VM infrastructure and data-sharing approaches. Along with this, investigations into the use of cloud services (See Section 5.4.2.8) are possible to enable remote data center functions. The biggest obstacle in this is the unknown cost and the migrations of capital expenses to operational expenses. With much of the network and compute industry is moving to the cloud, PSFC IT will need to evaluate migrating to the Hybrid Cloud.

Covid-19 occupancy restrictions have increased reliance on remote communication tools (zoom.us), regular short meetings (remote) to mitigate downsides of not seeing each other, and reduced business travel.

5.4.2.8 Cloud Services

Covid-19 has caused us to support users who are not physically present at the lab. It points out the need to virtualize many local resources, and points out the possibility of not doing so on premise. MIT PSFC will be exploring use of cloud for other network services like storage, compute, LDAP and more.

5.4.2.9 Data-Related Resource Constraints

The current data storage and servers are end of life (EOL) and will be replaced. This, coupled with increased demand for storage as use cases increase in data volume over time has MIT PSFC

Next 2-5 year anticipating insufficient storage space, and will need to consider expansion.

In terms of network demands, it is anticipated that both bandwidth and QoS will become an issue as cloud-based computing is adopted.

5.4.2.10 Outstanding Issues

MIT PSFC would like to explore the use of ESnet domain name service (DNS) services.

5.4.2.11 Case Study Contributors

MIT PSFC Case Study Representation

- Josh Stillerman[3], MIT PSFC
- Jerry Hughes[4], MIT PSFC

ESnet Site Coordinator Committee Representation

- Brandon Savage[5], MIT PSFC

3 jas@mit.edu
4 jwhughes@psfc.mit.edu
5 bsavage@psfc.mit.edu

5.5 PPPL

5.5.1 Discussion Summary

The U.S. DOE PPPL participates in a number of experiments and programs within FES. This case study will be used to highlight two in particular:

- XGC
- The National Spherical Torus Experiment Upgrade (NSTX-U)

Additional contributions from PPPL authors appear in other sections as joint efforts that span facility boundaries. PPPL faces a number of challenges in the coming years, and is preparing strategic solutions to prepare. The growing volume of data produced by simulation, the expanding needs of remote collaboration, and the impacts of large international experiments that will produce and share unprecedented data output are all heavily influencing the cyberinfrastructure design for PPPL. In all cases, networking provided by ESnet will remain a critical component, and PPPL is preparing to upgrade primary paths, and seek adequate backups, to ensure continuous operation.

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

- Gyrokinetic simulation will be a major research element during the exascale era of computation. The data produced during runs of this simulation can grow to volumes beyond what is capable for current computing storage to handle. As a result of this, effort to reduce data size is required before it can be stored locally, or transferred from ASCR HPC centers back to PPPL. Additionally, only some portions of the output can be viewed remotely due to the size of the data sets and the responsiveness of interactive tools that can be used to visualize. To properly support XGC:
  - XGC has the ability to limit output to fit within memory regions of current (and future) ASCR HPC resources
  - PPPL and ASCR HPC facilities will require storage upgrades to offer temporary locations for XGC output. PPPL will double their capacity in the coming years to offer PBs of storage space.
  - PPPL is upgrading their data architecture to install new data transfer hardware, is adopting Globus as a software package, has upgraded local storage, and will be working with ESnet to increase network capacity.
- XGC can produce a simulation of turbulence transport in an ITER-like plasma for a given equilibrium time slice using ORNL's Summit in 2 days of run time, but the resulting data set is approximately 50 PB in size. This volume must be reduced before storage or data transfer, and often only a small portion (typically 1-10TB) can be sent back to PPPL.
  - Future machines are expected to produce data that can approach 300 PB in size.
  - Full data transfer for volumes this large would require multiple Tbps network connections on ESnet between the ASCR HPC facilities and

PPPL.

  - Approaches to optimize bulk data transfer, and streaming, will be required even for reduced data sets.

- XGC is exploring ways to leverage cloud storage as a part of the experimental workflow. Due to the relative performance, as well as the volume and potential costs, it is not expected that cloud storage will replace local resources, but could be used to facilitate data backups, or use cases that require sharing. Additional work in this area could investigate cloud computing for multi-data set analysis.
- The TRANSP tool remains critical to FES analysis, and can provide interpretive and predictive simulations of a full tokamak discharge. TRANSP has the ability to use both MDSplus and Globus to accomplish computational and data mobility tasks, respectively. As a part of the process to define the ITER IMAS, TRANSP will undergo design and development to become compatible with the appropriate IDS requirements. This marks an early step for the FES community to adopt universal standards for cataloging tokamak data standards.
- The ECP-WDM code, once complete, will undergo a period of distributed community analysis. This simulation data will need to be available for a minimum of 5 years to provide data that will be used to develop fusion surrogate models and digital twins.
- DOE HPC allocations for FES are subject to annual renewal, and this causes challenges for strategic planning or long-term investments in a particular computing capability or workflow architecture. If renewing at the same location is not possible, this often leads to complications in data and workflow migrating to alternate facilities: adapting software to run on different systems, granting accounts to existing users, and sending a majority of scientific data to another facility. Unified APIs and simplified methods to manage data between DOE HPC facilities could simplify the friction seen in these scenarios. Longer-duration (strategic) allocations of computing at ASCR facilities would also allow more effective software investments to be made by the FES community.
- PPPL networking requirements have steadily increased over the years as the facility has taken more active roles in existing global FES experiments, such as KSTAR, and prepares for the future requirements of ITER. PPPL currently connects through MAGPI, and has upgraded their local networking environment to accept a 100G WAN connection from ESnet. They are pursuing a primary ESnet 100G connection, and would also like to pursue a backup connection through diverse paths and providers.
- ESnet will work with PPPL to upgrade the primary connection to 100G, and investigate ways to augment site connectivity with a second 100G through a diverse path to serve as a backup.
- ESnet will work with PPPL to validate their new data architecture, specifically the addition of new data transfer hardware and software in their science enclave. Participation in the DME will ensure PPPL is ready to

handle the increasing data volumes from XGC.

- PPPL has a number of use cases that leverage the GCP for storage of data and the execution of software codes; the cloud-based storage can take up several TB of space in the coming years. The current usage patterns for the data are not intense, but this may grow in the coming years as AI/ML informed simulations may be added to workflows. The usage can come from domestic and international partners.
- PPPL has migrated some data analysis tasks into cloud storage, and is exploring others as they prepare for upgrades to NSTX-U and the affiliated computational and software requirements. There is an effort to provide container-based versions of tools (e.g., TRANSP) as an alternative to running on PPPL computing resources.

5.5.2 PPPL Case Study

The U.S. DOE PPPL is a collaborative national center for fusion energy science, basic sciences, and advanced technology. PPPL is dedicated to developing the scientific and technological knowledge base for fusion energy as a safe, economical and environmentally attractive energy source for the world's long-term energy requirements. The Laboratory has three major missions:

1. to develop the scientific knowledge and advanced engineering to enable fusion to power the United States and the world;
2. to advance the science of nanoscale fabrication for technologies of tomorrow; and
3. to further the development of the scientific understanding of the plasma universe from laboratory to astrophysical scales. PPPL's expertise in fusion and plasma science.

PPPL's five core capabilities reflect its expertise and the role it plays in DOE missions:

- Plasma and FES
- Systems Engineering and Integration
- Large-Scale User Facilities/Advanced Instrumentation
- Mechanical Design and Engineering
- Power Systems and Electrical Engineering

For 70 years, PPPL has been a world leader in magnetic confinement experiments, plasma science, fusion science, and engineering. As the only DOE national laboratory with a FES mission, PPPL aspires to be the nation's premier design center for the realization and construction of future fusion concepts. PPPL also aims to drive the next wave of scientific innovation in plasma nanofabrication technologies to maintain US leadership in this critical industry of the future. Further, Princeton University and PPPL develop the workforce of the future by educating and inspiring world class scientists and engineers to serve the laboratory and national interest.

5.5.2.1 Background

This case study will profile two major projects that PPPL operates for the FES program:

- XGC
- The National Spherical Torus Experiment Upgrade (NSTX-U)

5.5.2.1.1 XGC

XGC is a gyrokinetic particle-in-cell code, which specializes in the simulation of the edge region of magnetically confined thermonuclear fusion plasma. The simulation domain can include the magnetic separatrix, magnetic axis and the biased material wall. The goal of the XGC program is not only to provide the edge component of the high-fidelity, kinetic WDM that relies on coupling of multiple codes including a core-edge coupling (see Section 5.12), but also to function as a single whole-volume kinetic code on exascale and post-exascale computers, with plasma heating/current-drive and material-wall interaction modules coupled in, to predict fusion energy production from first-principles-based models. Such a code could provide high-fidelity predictive understanding of future fusion reactor performance and assist building of more reliable surrogate models/digital twins that can be utilized in timely analysis and planning of next experiments. The headquarters of the XGC program is the Theory Department and the Computational Science Department at PPPL, funded by DOE FES.

When the ECP-WDM code is complete, the data needs to be community analyzed that are distributed over US. Important simulation data needs to be stored for over 5 years to provide data base for the development of fusion surrogate models and digital twins.

Since XGC is an extreme scale code that scales well to the maximal capability of the 200PF Summit and is expected to scale similarly well on the upcoming exascale computers, the data it produces is and will be big. For example, a simulation of turbulence transport in an ITER-like plasma for a given equilibrium time slice on Summit produces about 50 PB of particle data for two-days of wall-clock time. Since such a large amount of data cannot be saved in the OLCF scratch filesystem, normally limited to 2 TB of mesh data. It is desirable to move this data to PPPL for an in-depth interactive physics analysis after each one-day simulation. From the upcoming exascale computers, it is anticipated that moving about 10 TB physics data to PPPL after each simulation. If a mechanism to stream data analysis from an exascale HPC memory to PPPL cluster memory is utilized, it will be possible to deal with up to 250 PB/20 hours, which is about 3.5 Tbps. If this data can be moved at full capacity of 100 Gbps ESnet, this corresponds to about 0.3% of the particle data. The streaming-analyzed data will have to be lossy-compressed for storage at PPPL or on cloud. At a compression ratio 10, it would have saved 75 TB of streamed exascale-HPC data per simulation.

5.5.2.1.2 NSTX-U

The National Spherical Torus Experiment Upgrade (NSTX-U) is an innovative magnetic fusion device that was constructed by the PPPL in collaboration with the ORNL, Columbia University, and the University of Washington at Seattle.

NSTX-U is exploring the potential for energy production through thermonuclear fusion in spherical tokamak plasmas. Experiments are performed on the device, and measurements are made of both the experimental engineering systems and the plasma by a large number of sensors and diagnostics. Furthermore, the measured plasma data is most often used as a basis for analysis by large simulation codes, and the results of these computer simulations are also part of the data/analysis ecosystem. The measured

information is transferred to storage devices on a centralized computer cluster within a data storage framework called MDSplus. This framework is used for a number of domestic and international fusion experiments, facilitating ease of collaboration among the various experiments themselves.

The research community for NSTX-U consists not only of PPPL researchers but also of scientists from 18 collaborating institutions domestically. Most collaborators provide diagnostics; the others provide analysis codes. The data is long-lived; that is, it is stored, sometimes for decades, on the easy-to-access MDSplus platform. This is necessary as there is always a backlog of data that needs to be analyzed. The data from all the researcher diagnostics/analysis codes are stored in this centralized repository where it can be accessed at will by any member of the research team. The data is regularly backed up for preservation.

As part of the experiment, the data is further analyzed and refined - in some cases - multiple times for input to simulation codes. An example is the TRANSP code, a PPPL developed and maintained equilibrium and transport solver for tokamak discharge analysis. TRANSP is currently used on tokamak experiments worldwide. Each TRANSP run generates a NETCDF file that is archived either at PPPL or at hosting facilities. Multiple TRANSP runs can exist for each experimental tokamak discharge.

5.5.2.2 Collaborators

The collaboration space for both XGC and NSTX-U is broad, and features domestic and international participants.

5.5.2.2.1 XGC

User/Collaborator and Location	Do they store a primary or secondary copy of the data?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of ofofdataset? datasetdataset? (report in bytes, e.g. 125GB)	Frequency of data transfer or download? (e.g. ad-hoc, daily, weekly, monthly)	Is data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g. difficult tools, slow network)?
France, ITER headquarters	within 2 years: Secondary copy 2-5 years: Primary copy 5 yrs and beyond	Data transfer Portal	100 GB 75 TB 100TB	Monthly Quarterly	N Y, via portal	Network speed can be a limiter of timely collaboration
Univ, Colorado Boulder	Present: Secondary copy 2-5 years: Primary copy	Data transfer Portal	100GB 75TB 100TB	Monthly Monthly Monthly	N Y, via portal	Network speed can be a limiter
Lodestar, Colorado Boulder	Secondary copy	Data transfer	100GB	monthly	N	

5.5.1 – XGC Data Relationships

XGC features major collaboration with several domestic and international sites, along with other users (not reflected in the table) with smaller data mobility needs.

5.5.2.2.2 NSTX-U

The NSTX-U research community consists of over 130 physicists from both the United States and internationally. The funded collaborators, those that provide either diagnostics or analysis codes specifically for NSTX-U are from the United States from the following institutions:

- PPPL, Princeton University, Princeton NJ
- Princeton University, Princeton NJ
- Nova Photonics, Princeton NJ
- MIT, Cambridge MA
- Penn State University, State College PA
- Lehigh University, Bethlehem PA
- Johns Hopkins University, Baltimore MD
- The College of William and Mary, Williamsburg VA
- ORNL, Oak Ridge TN
- University of Tennessee, Knoxville TN
- University of Wisconsin, Madison WI
- University of Texas, Austin TX
- UCLA, Los Angeles CA
- UC Irvine, Irvine CA
- Lawrence Livermore National Laboratory, Livermore CA
- GA, San Diego CA
- University of Washington, Seattle WA

The data/analysis workflow for the collaborators is generally the same as for PPPL researchers (diagnostic data reduced/analyzed on-site and transferred to MDSplus). Some of the theory collaborators are using their own codes but are running them on the PPPL system and the results are transferred to the NSTX-U MDSplus repository.

There is one exception for collaborators who are doing theory/simulation calculations at their home institutions, using their own codes and storing the results on their own computers. This data is not needed for NSTX-U operations.

One exception is for TRANSP runs that are run on the PPPL cluster by PPPL users, which are stored locally.

Given that the bulk of the data and analysis are stored and run locally at PPPL, filling out Table 5.5.2 is N/A for this situation.

	Do they store a primary or secondary copy of the data?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of dataset? (report in bytes, e.g. 125GB)	Frequency of data transfer or download? (e.g. ad-hoc, daily, weekly, monthly)	Is data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g. difficult tools, slow network)?
(list from east to west)						
(from US, to outside) TRANSP	primary	File transfer protocol (FTP) fetch via globus, fully automated as the run is completed	1 Gbps average for each run	daily	N	
in the US, both East and West coast TRANSP	primary	FTP fetch via globus, fully automated as the run is completed	1 Gbps average for each run	daily	N	

Table 5.5.2 – NSTX-U Data Relationships

5.5.2.3 Instruments and Facilities

To accomplish the goals of XGC, it is necessary to use instruments that are primarily located off-site, at DOE HPC facilities. Section 5.3.2.1 will outline some of this, with more information regarding the remote nature of the operations provided in 5.5.2.1. NSTX-U is located at PPPL, and has minimal remote requirements.

5.5.2.3.1 XGC

XGC is designed to run on large computational resources, namely the DOE HPC facilities. It can be run on smaller institutional resources, but due to the computational and storage requirements, the majority of time is spend using resources at ALCF, OLCF, and NERSC.

This section will mention many of these remote resources, since they are the primary instruments for XCG operation. Section 5.5.2.5 will add additional detail regarding the impact to networks like ESnet.

Present-2 years:

XGC runs on the 200PF Summit at OLCF, 11PF Theta at ALCF and 30PF Cori at NERSC. XGC will also have access to the exascale Frontier at OLCF, 100PF Perlmutter at NERSC and ~40PF Polaris at ALCF in 1 year.

Next 2-5 years:

XGC plans to use the 1.5EF Frontier at OLCF, 1 EF Aurora at ALCF, and 100PF Perlmutter. The data and streaming rate will be similar to those from Frontier, as described in the current 1-2 requirement.

Beyond 5 years:

XGC's usage of the exascale HPCs will become more intense, with a few wall-clock days of simulation per study. XGC will contain at least 10X more number of species for longer wall-clock days of simulation. Plasma heating/current drive codes will be

coupled in. If there are post-exascale machines available beyond 5 years from now, XGC will try to utilize them for bigger science studies.

5.5.2.3.2 NSTX-U

Table 5.5.3 describes key diagnostics for NSTX-U, and the groups working to provide.

NERSC is sometimes used for large computational simulations.

5.5.2.4 Process of Science

XGC produces simulation output, which can be significant in terms of number of files, and volume of data that must be transmitted between the DOE HPC centers and PPPL. NSTX-U by comparison will produce smaller data volumes that are mostly contained within PPPL, with the exception of collaborators that may access data through MDSplus.

Key diagnostics operational in FY16 that will be recommissioned for initial plasma operation	Provider
Magnetics for equilibrium reconstruction*	PPPL
PFC thermocouples*	PPPL
PFC Langmuir probe*s	PPPL
Multi-Pulse Thomson Scattering (MPTS)*	
Toroidal CHERS (T-CHERS)*	PPPL
Fission Chamber Neutron Detectors*	PPPL
Plasma TV Cameras*	PPPL
Halo Current Detectors	PPPL
High-Frequency Mirnov Arrays	PPPL
RWM/Locked Mode Sensors	PPPL
Edge Rotation Diagnostic (ERD)	PPPL
HAL Spectrometer	PPPL
Real-Time Velocity (RTV) diagnostic	PPPL
Tangential Bolometer	PPPL
Motional Stark Effect-Collisionally Induced Fluorescence (MSE-CIF)	Nova Photonics
Beam Emission Spectroscopy	Univ. of Wisconsin-Madison
Fast Ion D-alpha Arrays (FIDAs)	UC Irvine
Solid State Neutral Particle Analyzers (SSNPAs)	UC Irvine
Scintillator neutron detectors	UC Irvine
Extreme ultraviolet (EUV) spectrometers*	LLNL
EIES (Filterscopes)*	LLNL
Visible Survey Spectrometer	LLNL
ENDD	LLNL
LADA	LLNL
DIMS & DIBS	LLNL
2-D Divertor Fast Cameras	LLNL
Two-color intensified 2-D cameras (TWICE 1 & 2)	LLNL
Fluctuation Reflectometry	UCLA
Divertor Spectroscopy (UV-VIS-NIR)	Univ. of Tennessee-Knoxville
Ultra Soft X-Ray array (USXR)	The Johns Hopkins University
Multi-Energy Soft X-Ray array (ME-SXR)	The Johns Hopkins University
Key diagnostics planned for completion in FY2021 and FY2022 to be ready for first experimental campaign	**Provider**
MPTS Calibration Probe	PPPL
Real-Time MPTS	PPPL
HHFW Reflectometer	PPPL
Far Infrared Toroidal Interferometer Polarimeter (FIReTIP)	UC-Davis
High-k Scattering	UC-Davis

IR Thermography of Lower Outer Diverter	Univ. of Tennessee-Knoxville
Doppler Backscattering/Cross-Polarization Scattering	UCLA
Diagnostics planned for completionin FY2023 and FY2022 to be ready for second experimental campaign	**Provider**
Main Plasma Resistive Bolometer	PPPL
Beam Emission Spectroscopy Deep Core	Univ. of Wisconsin-Madison
Pulse Counting SSNPA	UC Irvine
Transmission Grating Imaging Spectrometer	The Johns Hopkins University
Supersonic Gas Injector (SGI)	PPPL
IR Thermography of Upper Outer Diverter	Univ. of Tennessee-Knoxville
Wide Angle IR Thermography	Univ. of Tennessee-Knoxville
Diagnostic planned for completion in FY2024 to be ready for third experimental campaign	**Provider**
Faraday Effect Polarimeter-Interferometer	UCLA

Table 5.5.3

5.5.2.4.1 XGC

Present-2 years:
On Summit, the number of files is about 250,000 and the number directories is about 2,000. The maximum size of a file is about 10GB. On Frontier, the number of files will be about 2,500,000 and the number of directories will be about 20,000, with the maximum size of a file being about 100GB.

Reduced size physics-data output is up to 2 TB from Summit and 10 TB from Frontier per simulation, which are desired to be transferred to PPPL for timely physics analysis.

Next 2-5 years:
On Frontier, the number of files will be about 2,500,000 and the number of directories will be about 20,000 with the total number of data size 500 PB per simulation.

The data workflow requirement will be the same as for the Frontier requirement in the 1-2 years time frame. XGC can be well established for exascale computing on Frontier and Aurora, streaming data analysis from an exascale HPC memory to a PPPL cluster memory is planned to be used.

Beyond 5 years:
On Frontier, the number of files will be about 25,000,000 and the number of directories will be about 200,000, with the total number of data size 2EB per five-days of simulation.

It is anticipated that the data network and science analysis workflow will handle at least 5 times more data amount than what is needed in 2-5 years period.

5.5.2.4.2 NSTX-U

Data from the diagnostics listed above will be synthesized to provide a comprehensive view of the plasma characteristics. Other plasma properties will be inferred/computed from various data as inputs to large simulation codes. Networking is critical in being able to have collaborators access data and results, perform their own calculations and then transfer the results to the local MDSplus. In addition, networking is critical for enabling remote participation of collaborators in experiments.

5.5.2.5 Remote Science Activities

The primary remote use case for the PPPL case study is XGC, which relies on external computation with DOE HPC facilities. NTSX-U and a number of other FES projects, rely on remote communication with experiments with cloud-based communication platforms.

5.5.2.5.1 XGC

XGC utilizes resources at DOE HPC centers, thus the majority of work occurs during the remote execution of code, and the need to retrieve data results across the ESnet network.

Present-2 years:

When the XGC exascale computing is established, it is planned to use streaming data analysis from an exascale HPC memory to a PPPL cluster memory. The available source of streaming data per simulation can be up to 500 PB/20 hours, which must be significantly reduced for a timely transfer via ESnet. If it is possible to move the data at full capacity of 100 Gbps ESnet, PPPL would analyze about 0.1% of the particle data. The streaming-analyzed data will have to be lossy-compressed for storage at PPPL or on cloud. At a compression ratio 20, it is possible to save 75 TB of streamed exascale-HPC data per one-day simulation. An Adaptable I/O Systems (ADIOS) -based web portal will be used for streaming data analysis, compression, provenance, and storage workflow.

Next 2-5 years:

The available source of streaming data per simulation can be up to 500 PB/20 hours, which must be significantly reduced for a timely transfer via ESnet. If it is possible to move the data at full capacity of 100 Gbps ESnet, PPPL would analyze about 0.1% of the particle data. The streaming-analyzed data will have to be lossy-compressed for storage at PPPL or on cloud. At a compression ratio 20, it is possible to save 75 TB of streamed exascale-HPC data per one-day simulation. An ADIOS based web portal will be used for streaming data analysis, compression, provenance, and storage workflow.

Beyond 5 years:

It is anticipated that XGC may be used remotely by ITER scientists at EU, Japan, Korea, etc. using their own exascale supercomputers. It is desirable to have the data flow to PPPL at the above rate.

5.5.2.5.2 NSTX-U

Remote resources at this point are limited to conferencing. Once the experiment begins to run, it is expected that there will be both video and audio connections to the experiment control room, with an expanded communication network.

5.5.2.6 Software Infrastructure

While PPPL uses a variety of data transfer protocols including SCP, BBCP, FTP, etc, the primary and recommended tool over the last several years has been Globus/Grid FTP. PPPL has a 10g Globus server in the PPPL DMZ, with direct connectivity to HPC storage servers. A new, 100g capable server is currently being tested and tuned. It is expected to be deployed in the next several months.

Performance monitoring in the WAN is handled by PerfSonar, currently running on the

PPPL internal network. A new, 100g capable PerfSonar node is also being deployed and will be capable of testing in all areas of the PPPL core network (outside, DMZ, inside).

Internally, bandwidth monitoring is handled by the PRTG software package, which provides current and historical throughput data for key points within the PPPL network.

5.5.2.6.1 XGC

Currently, XGC is using Globus to move the physics data to PPPL from the computing facilities. Files are in ADIOS-BP format. A prototype DELTA framework has been developed for remote data flow with streaming analysis capability. A web-based data management protocol, based on an eSimMon dashboard, is under development combining ADIOS and DELTA capabilities into it, which will be operational within 2 years.

Future use cases for XGC will center on the web-based data management protocol, which combines in the ADIOS2, DELTA, eSimMon technologies. XGC simulation on exascale computers will be operated like a large experimental facility, in which the simulation scenario will be jointly developed by distributed collaborators across US and different continents. The simulation data will be streamed in real time to the distributed collaborators for aggregated simulation steering information and timely scientific discovery.

5.5.2.6.2 NSTX-U

NSTX-U software is highly dependent on the diagnostics being examined. These codes are written by the responsible groups and physicists that operate on each diagnostic. Additionally, there are some larger codes that are used to reduce data as input for more comprehensive calculations, and a commercial database application that is used to store important time-slice data.

5.5.2.7 Network and Data Architecture

PPPL has historically not been able to support data throughput requirements for some experiments and collaborations and has seen data sent elsewhere for storage and analysis. PPPL is hoping to work with ESnet to upgrade the primary internet connection to 100g in the next 12 months.

On the WAN side, PPPL currently has 10g connections to ESnet via NYC and Washington DC via Magpi dark fiber. A 10g connection to Princeton University is scheduled to be upgraded to 100g later this month. All routing is currently static.

A Science Laboratories Infrastructure project for infrastructure is moving through the approval process which will provide a variety of improvements to network infrastructure, including hardware redundancy in the data center, replacement of aging copper and fiber infrastructure, and additional WAN connectivity. A fiber ring topology is also proposed to improve overall resiliency. PPPL also plans to move from static routing at the edge to BGP with ESnet and Princeton University.

PPPL currently has a PerfSonar node on its internal network running at 10g. A new PerfSonar server capable of 1/10/25/40/100g testing has been installed and is currently being tuned and tested.

In addition to a 10g Globus node in the DMZ, an second Globus server built for internal data movement within PPPL is online with limited use. PPPL expects this to change with the start of NSTX-U operations, and that a new server will be needed to facilitate internal large-scale data transfers associated for the experiment.

PPPL does not currently have a science DMZ, but as previously mentioned recently upgraded to a high-end firewall. PPPL plans to complete performance testing to determine real throughput capabilities in the LAN, MAN, and WAN. If performance improvements are necessary, the plan is to test the Palo Alto Application Override feature which promises line speed capabilities. Other options include Arista Direct Flow and Direct Flow Assist functionality. A Science DMZ is also being considered if other options do not provide the necessary results.

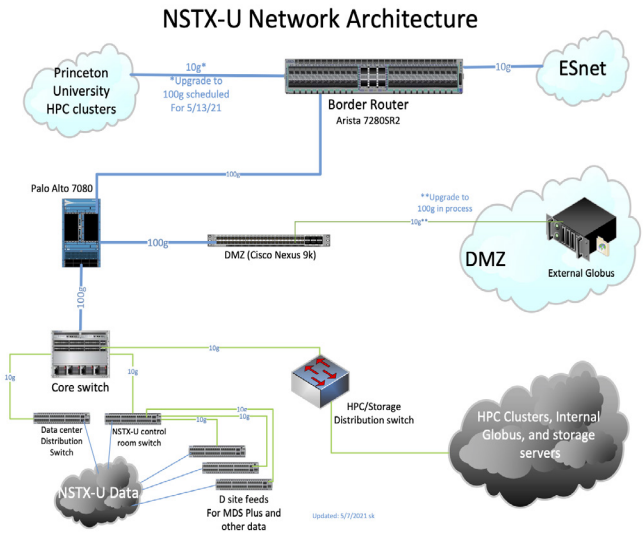


Figure 5.5.1 - PPPL Network Architecture

In the next 2-5 years, PPPL also plans to upgrade the internal network design from a flat layer 2 topology to Layer 3 routing with an interior gateway routing protocol. This represents a significant change from the current design and will require close consultation with the PPPL cyber security team to meet existing and future requirements.

Five years from now, it is likely that data throughput requirements will steadily increase for NSTX-U, with more dramatic increases in other areas. The ITER international collaboration is expected to be ready to begin operations in this general timeframe. And while these data requirements will not be fully known for some time, it is entirely possible that 100g may not be sufficient to support data movement and analysis requirements. PPPL is planning to be a US data hub for ITER and will closely monitor developments as they occur. PPPL wants to be prepared to increase throughput capabilities above 100g if necessary.

PPPL is actively pursuing new initiatives, including public/private partnerships in areas such as microelectronics and liquid metals. The PPIC, the first new building on the PPPL campus in almost 40 years, is expected to be completed in several years to

support new experiments and initiatives. The PPPL network team must be prepared to meet these data throughput requirements that stem from these areas of expected growth.

5.5.2.7.1 XGC

The PPPL HPC environment connects to the core network, which has seen many upgrades, with the PPPL border router, core switch and core firewall all replaced in the last 18 months. The most recent of these upgrades was the move to a Palo Alto 7080 firewall which occurred on 4/17/2021. All devices in the PPPL network core are now running with 100g uplinks.

HPC connectivity to the core is provided by an Extreme Networks distribution layer switch. The uplink is currently 10g, but will be upgraded to 100g later this year. All devices in HPC are on the same VLAN. A Globus server in the DMZ is dual homed with a direct connection to storage servers in the HPC VLAN. Globus will also be upgraded to 100g later this year.

Most recent HPC growth has actually occurred at Princeton University. A new GPU based cluster, Traverse, is 80% dedicated to PPPL. A new cluster, Stellar, is currently being installed and will also be dedicated primarily to PPPL. Globus at 100g and the new 100g connection to Princeton University will provide for dramatically improved data-movement capabilities to support data modeling.

The network supporting the HPC environment is currently being reviewed. The PPPL network team is working with the manager of HPC to clarify future growth on the PPPL and Princeton University campuses, and the resulting network requirements. PPPL expects new network hardware and expansion of VLANs in this environment as this process moves forward.

While precise numbers are not yet known, the PPPL network team is preparing for a significant increase in data throughput requirements. The PPPL flagship experiment, NSTX-U, is expected to come online next year. PPPL expects data flows to the storage network within the HPC environment will increase dramatically when operations resume. PPPL is also moving to a full user facility model supporting a variety of fusion energy research and expects to add many more experiments, including public/private partnerships in areas such as microelectronics and liquid metals. All of these will likely increase networking requirements in HPC.

5.5.2.7.2 NSTX-U

NSTX-U is PPPL's flagship experiment, and runs across a layer 2 LAN environment with a core firewall serving as the primary router for the internal network. The network is segmented with VLANs. While the experiment has devices across the network, two primary VLANs support NSTX-U: diagnostics and controls. Devices on these networks run primarily at 100m and gigabit speeds. Uplink speeds are primarily a mix of gigabit and 10 gigabit. Data flows are primarily internal, sending diagnostic and camera data to HPC storage servers and MDS Plus data trees. The experiment is expected to begin operations sometime next year.

The core network has seen many upgrades, with the PPPL border router, core switch and core firewall all replaced in the last 18 months. The most recent of these upgrades

was the move to a Palo Alto 7080 firewall which occurred last month. All devices in the PPPL network core are now running with 100g uplinks.

Over the next two years, PPPL plans to begin upgrading network uplinks to the data center at 25 or 100g based on expected throughput. Initial upgrades will focus on the primary star point for NSTX-U data in a control room. PPPL plans to upgrade this connection from 10 to 100g later this year. Other connections will also be upgraded as the network design is completed and more details on requirements are available.

5.5.2.8 Cloud Services

PPPL is exploring a number of options in the cloud computing space, from a number of different providers. The facility completed the Authorization to Operate (ATO) process for GCP, and has integrated this into a number of use cases across the lab. Additional ATOs are being performed for Microsoft Azure and Amazon Web Services, with the goal of trying to adapt other scientific workflows in the future. The three cloud computing services will allow for a number of potential use cases in the future, as the facility and research teams test different data handling capabilities.

In the general case, any cloud services that PPPL considers must undergo a security impact assessment in order to receive authorization. This begins with PPPL's IT cyber team, who evaluates the product, the security capabilities of the environment, and culminates in writing an impact assessment. The risk based assessment of the product is passed to the Princeton Site Office for evaluation from the cyber team and the Chief Information Officer for discussion and approval.

5.5.2.8.1 XGC

XGC's usage of cloud services will mainly be in the form of data storage capability, as the data volumes are fast outpacing the ability to store locally or at DOE HPC facilities. A critical evaluation point to the adoption of cloud approaches is the speed at which data can be uploaded and downloaded from the resources: a requirement will be the ability to access cloud data at the rate of a few minutes for 1 TB of data. Data analysis capability in the cloud, could also prove to be very useful, so that only the analyzed data can be transferred to PPPL, but this is still being investigated.

5.5.2.8.2 NSTX-U

NSTX-U will explore cloud computing and storage that is made available at PPPL, but does not have any specific plans at this time that can be enumerated.

5.5.2.9 Data-Related Resource Constraints

A shared resource constraint that spans the two use cases is a lack of long-term storage space to support growing data needs. PPPL is working to improve this for the entire facility. The current technology planning is designed to increase PPPL's storage capacity by more than double to handle current and short term storage needs, but also redesigning to support a highly scalable and fast environment that can be built out over time. PPPL will be standing up a new 3 PB storage system this fiscal year to support all science activities.

5.5.2.9.1 XGC

The data generated from XGC when run at DOE HPC facilities is significant, as described in earlier sections. Timely physics productivity is currently constrained by the slow data transfer speed from the computing facilities to PPPL, and data storage capacity at PPPL. As the exascale computers are used for science runs in a year from now, this issue will become more severe.

The upgrades to the PPPL storage infrastructure will have mechanisms to support localized and remote data transfer use cases. This will allow transfer to and from PPPL at high speeds, and the ability to handle large datasets. This new storage infrastructure will support 100 Gbps+ locally, with a new 100 Gbps capable DTN included to support higher transfer speeds.

XCG is looking to expand upon the general PPPL data architecture by investigating the deployment of multiple DTNs, as needs arise. This would facilitate an ingress data storage level, to act as a fast caching tier, and then flow off to general storage tier. This will help with leveraging faster external data pathways to the DOE HPC facilities, and make remote data available as fast as required per use case.

5.5.2.9.2 NSTX-U

NSTX-U does not have additional data-related constraints beyond what has been described, and will benefit from the plans mentioned above that describe storage and data mobility upgrades at PPPL.

5.5.2.10 Outstanding Issues

There are no additional issues to report at this time.

5.5.2.11 Case Study Contributors

PPPL Representation

- CS Chang[1], PPPL
- Michael Churchill[2], PPPL
- Bill Dorland[3], PPPL
- Walter Guttenfelder[4], PPPL
- Stan Kaye[5], PPPL
- Francesca Poli[6], PPPL

ESnet Site Coordinator Committee Representation

- Scott Kampel[7], PPPL

1 cschang@pppl.gov
2 rchurchi@pppl.gov
3 bdorland@pppl.gov
4 wgutten@pppl.gov
5 kaye@pppl.gov
6 fpoli@pppl.gov
7 skampel@pppl.gov

5.6 Planning for ITER Operation

5.6.1 Discussion Summary

Given the extensive experience developed by ESnet in meeting US networking needs for large collaborative and international scientific projects, it would be particularly important for ASCR, ESnet and FES to perform a formal assessment of the ITER data analysis and network requirements well in advance of ITER first plasma. It will also be important to engage the IO in this assessment, given that IO decisions in the near future may have important implications for the US and other ITER members regarding the timeliness of data access and the quality of remote participation.

5.6.2 Planning for ITER Operation Case Study

The ITER tokamak is the most ambitious fusion experiment ever undertaken. ITER is a magnetic confinement device where hydrogen isotopes are heated to temperatures up to 100 million degrees, forming a plasma and forcing nuclei to fuse to create fusion energy. ITER brings together 35 nations and 7 major partners (China, the European Union, India, Japan, Korea, Russia and the United States) to collaborate on building the world's largest tokamak, designed to achieve sustained high fusion power (500 MW, 500-550 s) by the mid-2030s, and to potentially achieve full steady-state operation thereafter. ITER is located in Cadarache, France, only a 350 km drive from CERN, the location of another major global scientific collaboration with significant US participation on the Large Hadron Collider (LHC)

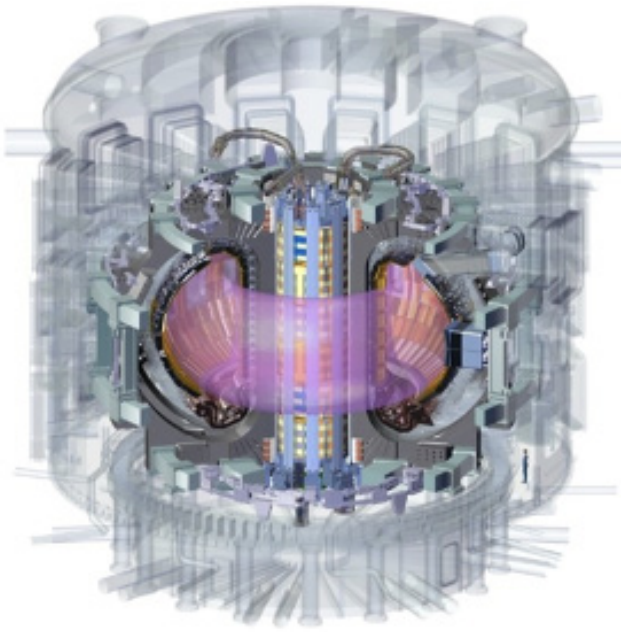


Figure 5.6.1 - Schematic diagram of the ITER tokamak being constructed in Cadarache, France. (https://www.iter.org/)

5.6.2.1 Background

ITER is first and foremost a scientific experiment, with over 50 major diagnostic packages consisting of thousands of data channels, producing in excess of 2 Petabytes of raw data each day and requiring more than an exabyte of data storage by the mid-2030s.

This estimate does not include the volume of analyzed and simulated data that will be produced and archived. Given the volume of data anticipated from ITER diagnostics and analysis, and the widely distributed scientific teams engaged in every aspect of machine operations and scientific research, ITER must employ state-of-the-art data management solutions to meet its mission goals.

Fortunately, ITER will commence operations with much less data production per day (~ 20 TB) during the first phase of plasma operation (engineering commissioning, first plasma, and engineering operations) planned for 2026. Therefore, ITER and the international fusion community will have time to learn and prepare for when peak data is expected in the mid-2030s. Nonetheless, plans are currently being made for ITER's IT infrastructure to support data storage and analysis needs during this early period and these decisions will likely affect future capabilities and expansion plans. Consequently, the IO, and specifically the IT and IMAS group, are actively engaged in discussions with ITER Members to refine data management and analysis plans for first-plasma. An initial and very productive meeting with FES, ITER IT and ESnet took place in July of 2021, with the expectation that these communications would continue.

All ITER (like present fusion) experiments will follow a general pattern: capture of raw data and the processing of such data to produce calibrated physical quantities. Such analysis is typically performed on-site with computing resources close to where the data is located. From there, more complex analysis and comparisons of data to simulations will take place mostly off-site due to the limited resources the IO will have to support the ITER research needs.

An important design philosophy for ITER analysis is embodied in the IMAS being developed at the IO under the guidance of the IMEG. The backbone of the IMAS infrastructure is a standardized, machine-generic data model that represents simulated and experimental data with identical structures. IMAS will serve to bring together calibrated data and simulated data in one framework in order to develop a rigorous statistical approach to the inferences drawn from measurements. Some simulated data will be generated locally, but many and also the most sophisticated simulations will be performed off-site using leading HPC capacity in the ITER member countries, and all such data, simulated and raw, and all codes and metadata associated with such analysis, will be stored in the IMAS framework to guarantee traceability and reproducibility.

Perhaps for the first time in the history of fusion research, ITER will generate a range of "simulated" data covering every possible aspect of the ITER experiment beforehand, including first plasma experiments where extensive modeling has already taken place to understand the capabilities and limitations of all the first plasma diagnostics for interpretation and control[1]. Extensive simulations will be used to assist in the planning of experiments, the design of the control system and control diagnostics, and in the analysis of data after the experiments. While some modeling does take place before experiments in present devices, the transition to ITER will represent a stark contrast with current standards.

While ITER is a major international collaboration, within ITER there will be many smaller collaborations focused on each scientific instrument including their operation, calibration and physical interpretation. The US will provide seven diagnostic packages

1 J Sinha et al., Development of synthetic diagnostics for ITER First Plasma operation (2021) Plasma Phys. Control. Fusion Vol. 63 084002: https://doi.org/10.1088/1361-6587/abffb7

to ITER and each of these will be operated by an international team of collaborators, some on-site in Cadarache and many off-site. At the peak of its operation in the mid-2030s, ITER will likely have thousands of collaborators distributed around the world who will require access to the data generated by the ITER instruments and to simulated data generated by multiple HPC systems worldwide.

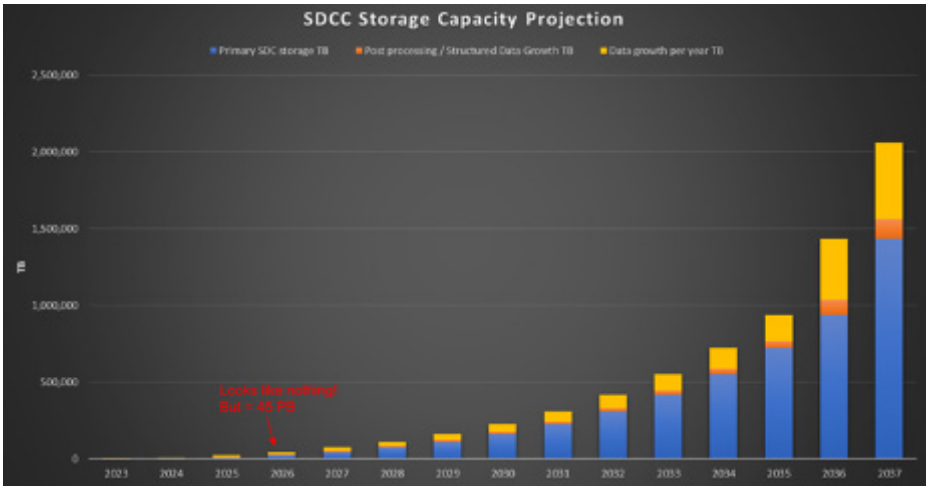


Figure 5.6.2 - ITER will generate >2 PB of data a day and require exabyte storage by the mid-2030s (courtesy Peter Kroul, Computing Center Officer, ITER).

Storage (PB)	2023	2027	2031	2035
primary SDC storage/yr	1.0	45.	225.	725.
Raw data growth/yr	2.0	25.	65.	170.
structured data growth/yr	0.5	5.	15.	45.
total capacity /yr	3.5	75.	305.	1,040.

Table 5.6.1 - Current data growth projections per year in PB (courtesy ITER). By 2035 the storage requirement will exceed an exabyte per year.

5.6.2.2 Collaborators

The main participants in ITER are the seven major parties (China, the European Union, India, Japan, Korea, Russia and the United States) comprising 35 nations. Today, the details of the operational phase of ITER are under discussion between the IO that is building ITER in Cadarache, France and the seven ITER members. These discussions will make recommendations to the ITER Council on the structure, roles and responsibilities of the ITER partners and IO.

As for the scientific instruments that the US will provide to ITER, it is not yet known which countries will collaborate with the US in their operation. Such information may be forthcoming by the next ESnet review of fusion network needs. However, it is useful to list the major diagnostic instruments that the US will supply to ITER as currently planned.

5.6.2.3 Instruments and Facilities

5.6.2.3.1 Core Capabilities

ITER will be perhaps the most complex machine ever built. The ITER facility represents an integration of many separated systems, each of which will be highly instrumented and monitored in real time to ensure that the overall integrated operation of the facility is consistent with the availability, and within the operational limits, of each component. The main chamber, where the fusion reactions will take place, comprises several major systems including the vacuum vessel, cryostat, plasma facing components, including the divertor that handles the bulk of the power exhaust from the plasma, and superconducting magnets. Other systems include the tritium breeding blanket(s), cryogenic systems for the magnets, power supplies, heating and current drive systems and vacuum and fueling systems, to name a few. Real-time data on all of these systems will be generated and fed into the control system to ensure that the safety of each system and overall facility is not compromised. For example, if the heat load on the tungsten divertor is predicted to reach operational limits (say nearing 10 MW/m2), then the control system will be programmed to recognize the issue and respond by a combination of actions including but not limited to the injection of impurities for further power dissipation, the movement of the plasma strike point away from the high heat flux region, and/or the safe termination of the discharge if necessary. Here the emphasis is placed on prediction, which must be built into the control system in order to anticipate events of high probability and to avoid or mitigate their effects. This is just one example of the complex decisions that the ITER control system must make during operations, and it (the control system) will depend on accurate real-time data from each system as well as models of each system to anticipate and mitigate adverse events.

The US is supplying substantial hardware to the ITER facility including some of the superconducting magnets, power supplies and various other components during the construction phase. In addition, there are seven key scientific instruments for plasma analysis that the US will supply and be responsible for during plasma operations. Some of these systems are of great importance to understanding the state of the plasma and ensuring the safety of the facility, such as the IR cameras that will be used to inform the shape and gas control system of impending overheating of the walls. A listing of the key scientific instruments to be supplied by the US and their use in ITER experiments is indicated below:

- Core Imaging X-ray Spectrometer - Measures the core spatial profiles of ion temperature and toroidal rotation with three x-ray crystal spectrometers.
- Electron Cyclotron Emission Radiometer - Measures the spatial profile of the electron temperature; provides time resolution adequate for use as a sensor in feedback control of predicted instabilities.
- Low Field Side Reflectometer - Provides primary measurement of electron density spatial profile in the pedestal (edge) region of the plasma; contributes to determining the plasma pressure gradient that influences stability and transport.
- Motional Stark Effect Polarimeter - Determines the internal magnetic field of the plasma from measurement of the polarization direction of light emitted by the heating neutral beams; spatial variation of this magnetic field directly affects the stability of the plasma.

- Residual Gas Analyzer - Measures the concentrations of neutral gases during a plasma discharge in the divertor exhaust duct and in the main chamber; capable of distinguishing the fuel gases (D, T) from the He exhaust.
- Toroidal Interferometer/Polarimeter - Determines the electron density by probing the plasma with a CO^2 laser; to be used as a sensor in a feedback system to control various fueling sources.
- Upper IR/Visible Cameras - Includes five endoscopic mirror systems for viewing divertor targets from the upper port region and to detect hot spots on these targets using visible and IR imaging sensors; Operates with the plasma shape/position control system to prevent melting of tungsten divertor targets.

In addition, there are over 50 separate diagnostic packages being delivered to ITER by the seven ITER partners, and each package can have hundreds or thousands of sensor signals. This heterogeneous set of signals must be processed at various levels and the raw and processed data will be made available to the international research community for detailed analysis. As mentioned earlier, ITER is likely to produce in excess of 2 PB/day aggregated from over 50 separate instruments at peak operation.

Given the volume of real-time data that will be generated on the status of each system and on the measured properties of the plasma, careful planning will be needed to ensure that the highest priority data is available in near real time to off-site collaborators in the US who will be participating in ITER's operation. There is more on this topic in section 5.6.2.10 on outstanding issues.

5.6.2.3.1 First-plasma and Engineering Operation Diagnostics and Sensors

During first plasma and engineering operations, it should be possible, in principle, to transmit a significant quantity of facility and plasma data in near real time. This is because first-plasma will have a modest set of diagnostics compared to full ITER operation and because many auxiliary systems and their extensive sensor arrays (like negative ion neutral beams and ICRH) will not be installed. The first plasma diagnostics are needed primarily to establish the main parameters of the plasma, up to 1 MA at 2.65 T with up to 8 MW of ECH. The emphasis for the diagnostics will be on characterizing the plasma breakdown and equilibrium during ramp-up of the current and to ensure that runaway electron interactions with materials are detected (hard X-rays). The magnetics diagnostics will be fully available for equilibrium reconstructions and high-bandwidth magnetic measurements will be able to detect MHD modes. Given the short duration of the plasma pulses (< 10s), the vacuum vessel will need to be included in the equilibrium reconstructions, and the combination of all diagnostics, magnetic and non-magnetic, will be needed to validate models of plasma breakdown and current ramp-up. In general, the imaging diagnostics (IR camera for surface temperature measurements) may take up the most bandwidth. On the other hand, real-time compression methods are readily available (think Netflix). Table 5.6.3 shows plasma diagnostics taken during engineering commissioning.

Measurement Requirement	Installed Diagnostic Systems
Magnetics for position, velocity, shape and MHD mode structure	Magnetics System Electronics and Software Continuous External Rogowski Outer Vessel Coils Steady State Sensors Flux Loops Inner Vessel Coils Diagnostic Sensors
Line averaged electron density (toroidal polarimeter/ interferometer)	Desntiy Interferometer (single channel)
Runaway electron detection device (hard X-rays)	Hard X-ray Monitor
Impurity identification and influxes (visible and near UV spectroscopy including H and visible bremsstrahlung), partial systems	H_α/Visible in EPP12 Vaccuum Ultra-Violet Survey Visible Spectroscopy Reference System (partial) – temporary X-ray Crystal Spectrometer
Visible/IR TV viewing (spectroscopically filtered), partial coverage	Visible/IR Equatorial in EP16 (temporary) Visible/IR Equatorial in EPP12 (partial)
Torus pressure and gas composition (torus pressure gauges, RGA)	Pressure Gauges (temporary)
Toroidal Field (TF) Mapping	Temporary set of magnetic pick-ups for TF mapping
Machine protection	Tokamak Structural Monitoring System Stray ECRH detector

Table 5.6.3 - First plasma diagnostics, ITER Research Plan, p.57.

Exact estimates of the peak data production rate are difficult to come by, however aggregate estimates of 20 TB/day data production rate have been made for the engineering operations phase. On the other hand, except for data requiring very large bandwidth (for MHD activity or camera data), a good fraction of low-bandwidth data relevant to systems monitoring and plasma equilibrium analysis can be transmitted at a reduced rate, say 10 kHz. Assuming only 1% of peak wire speed for a 100 Gbit/s data pipeline from ITER to a US data center, it will be possible to accommodate several thousand channels of data transfer near real time at 10 kHz with 16 bit signal resolution. Thus, the possibility for streaming a good fraction of ITER data during engineering operations is a distinct possibility, however this will require careful planning and preparation.

5.6.2.4 Process of Science

At its peak operation in the mid-2030s, ITER will produce > 2 petabytes of data each operating day. While the actual data production during first-plasma operation will be significantly less (anticipating 20 TB per day during engineering operations) the process of science will not change significantly, at least in its main outlines.

There are three important and distinguishable processes involved in a fusion experiment:

1. the planning and execution of the experiment
2. the first-cut analysis and assessment of the data in the control room
3. the further processing of the data and its use in constraining or validating models and simulations after the experiment

5.6.2.4.1 Experimental Planning

Experimental planning often involves the review of prior experiments or databases of

prior experiments. Typically, in the United States, the data is archived in an MDSplus[2] tree and saved to a central storage system at the site of the facility. Generally, the physicists use their collective judgement for assessing what needs to be done in the next experiment. Intuition is sometimes augmented by more sophisticated modeling. Once the proposal is approved, the scientists are assigned a "physics operator" whose task it is to determine how best to achieve the required conditions, using a combination of tools to design the discharge trajectory and control parameters.

A major change with ITER is that experiments will need to be designed using a hierarchy of models of different physics fidelity in order to maximize the probability of success. A virtual experiment will essentially be created, consisting of models of the control system, vessel, plasma, heating and diagnostic systems. Every conceivable contingency will need to be assessed and the control parameters adjusted to meet safety and performance requirements.

ITER operation, therefore, represents a major shift from the current standard of facility operations. The intense analysis before ITER experiments will mean few, if any, outside experimental proposals from collaborations; ITER operation will focus on mission needs which means devoting more time to planning mission experiments than is currently the norm, and individual missionexperiments may extend for weeks or even months, tied to major project milestones and system commissioning activities.

From the network perspective, it can be expected that a significant amount of pulse design activity will take place at ITER, and in the several ITER analysis centers worldwide before experiments are performed. The experimental planning interval may be months, in which case the ITER members will need to carefully plan the level of physics fidelity they will use in their modeling according to their HPC capacity. Also, depending on the physics fidelity, the experimental planning can produce enormous amounts of data (particularly if gyrokinetic models are used for turbulence and transport simulation and full-physics models for key diagnostics). In the US it is likely that network and computer availability will need to be planned with ESnet and leadership-class computing facilities well ahead of the planned experiment. The analysis, including the high-fidelity simulation data will need to be integrated with IMAS and pushed back to ITER for final assessment.

5.6.2.4.1 Experimental Performance

In present facilities a typical experiment is a collection of similar discharges executed over a single day or partial day, with each discharge typically lasting between 10s and 100s. Sometimes an experiment can run over several days but this is quite rare. Initially, discharges in ITER will be of similar duration per pulse, but with the goal of reaching 500s by the mid-2030s. However, unlike existing experiments, ITER will likely have experimental sessions that last for weeks or even months at a time on a given experiment. This is because ITER will have only a few very high-priority milestones and limited machine time for their accomplishment. Also, for the safe operation of ITER, including power handling of the exhaust to the walls and the diverter, there will be limited flexibility to alter the plasma shape and perhaps other parameters.

For current fusion experiments, the plasma pulse sequence and control settings are

2 T. Freudian, et al., MDSplus yesterday, today and tomorrow (2018) Fusion Engineering and Design Vol. 127 106–110, https://doi.org/10.1016/j.fusengdes.2017.12.010

often adjusted from shot to shot with minimal advanced planning. The experimental teams on current experiments can range from 50 to 200 people, with many participating from remote locations. Plans for the next plasma pulse are often informed by rudimentary control room analysis performed on a local unix cluster in the short interval (~20 minute) between plasma pulses. However, the sophistication of the analysis that can be performed between plasma pulses is changing because of the availability of HPC facilities and fast networks. Several proof-of-principle tests[3] [4] [5] have been conducted for fusion experiments which demonstrates the potential of modern networks and HPC infrastructure to provide sophisticated analysis to the control room to help guide experimental decisions. A related development is the push towards more automation of data preparation. Many diagnostic systems still require manual intervention to produce useful calibrated data, thus delaying the accessibility of the data for informing decisions in the control room.

ITER expects to provide all its data both in real time for the control system and for between-shot analysis. Careful planning is required to ensure that this data can be accessed by US remote users rapidly enough to potentially inform control room decisions, knowing that this mode of operation will be more carefully scripted than other FES collaborations.

For ITER, the relative amount of data analysis on-site versus off-site is still unknown but it is anticipated that the bulk of the analysis will be performed off-site in various data and analysis centers around the world. By using HPC infrastructure in the member countries and advanced networks, more data can be analyzed and more sophisticated analysis can be performed during experiments.

An important question is how fast data can be pulled from ITER and how much computing capacity can be reserved in the US when experiments are running at ITER. Ideally, the US would like a combination of near-real-time data during the actual plasma pulse and then the rapid transfer of the bulk of the scientific data within ~5 minutes or less after the pulse is completed. This would provide opportunities for US remote participants to perform analysis in time to inform the next pulse or several pulses thereafter, noting that the interval between pulses will be approximately 20 min - 1 hour on ITER. It is noted that ITER operation may be scripted far in advance, so control room deviations using instant analysis may not influence experimental direction. Table 5.6.2 lists required network throughput from the IO to the United States for the acquired raw data in Table 5.6.1 to be transferred within 5 minutes after making some assumptions on the number of shots obtained in a year. Today, 100 Gbps connections for major scientific centers are not uncommon and thus such a network throughput to the US starting in the first year is reasonable, provided the system is carefully designed to deliver routine data transfer at close to wire speed. But a significantly greater and increasing capacity will be required in outlying years, requiring detailed and advanced planning.

---

3 R.M. Churchill et al., A Framework for International Collaboration on ITER Using Large-Scale Data Transfer to Enable Near-Real-Time Analysis (2021) Fusion Science & Technology, Vol. 77, pp. 98-108: https://doi.org/10.1080/15361055.2020.1851073

4 R.M. Churchill et al., A Framework for International Collaboration on ITER Using Large-Scale Data Transfer to Enable Near-Real-Time Analysis (2021) Fusion Science & Technology, Vol. 77, pp. 98-108: https://doi.org/10.1080/15361055.2020.1851073

5 M. Kostuk, et al., Automatic Between-Pulse Analysis of DIII-D Experimental Data Performed Remotely on a Supercomputer at Argonne Leadership Computing Facility (2017) Fusion Science and Technology Vol. 74, 135-143: https://doi.org/10.1080/15361055.2017.1390388

In fact, during first plasma and engineering operations, it may be feasible to transmit a significant fraction of all ITER data in near real time at reduced bandwidth employing methods under development by ASCR to support large-scale collaborative and international scientific experiments[6]. (See also Appendix A for ITER first-plasma diagnostics). The IO in Cadarache is beginning to make plans now for data and network system requirements for first plasma. They have begun discussions with FES and ESnet technical experts to understand US data and analysis needs, and these discussions will continue until final requirements are completed. The discussion above emphasizes the importance of an early dialogue with the IO, FES and ESnet to address data needs in time for the first plasma.

	2023	2027	2031	2035
Estimated Required Network Throughput (Gbps)	20	200	500	1500

Table 5.6.2 - An estimate of the required network throughput from the IO to the US based on the data quantities in Table 5.6.1 to be transferred in 5 minutes. Assumptions were made on how ITER would operate yielding a total number of shots per year.

5.6.2.4.1 Experimental Analysis

ITER will have long contiguous periods of non-operation for maintenance and major upgrades (see Figure 5.6.3). There is roughly one year of operation for two years of maintenance and upgrades. This means that an enormous amount of analysis, publication, preparation, and experimental planning will take place during these non-operating periods. It also means that the capabilities of ITER will evolve enormously from one major campaign to the next as major new systems come online during the non-operating periods. It is tempting to think that activity at ITER may quiet down during these times, but the opposite may be more correct due to the very large amount of analysis, modeling and planning involved in preparing for the next campaign.

While it is expected that the US will access ITER data from their own data mirrors, some or all the analysis and simulations of importance to ITER will need to be pushed back to ITER storage and mirrored back out to all the ITER parties. Part of the design principle for the IMAS is to guarantee traceability and reproducibility of all significant analysis. How this will play out for complex data-intensive simulations is to be seen, but there may be the need to invest in high-performance Science DMZ DTN clusters[7] to make the transfers manageable. The knowledge gained from remote operation of EAST third shift from the US[8] and from collaborative experiments on KSTAR, JET and other international facilities, will be valuable in designing this deployment. Therefore, depending on the fidelity of the modeling and simulations employed, the amount of data that needs to be transferred from and to ITER could be substantial during operations and shutdown periods.

6 STREAM2016: Streaming Requirements Workshop Final Report, Office of Advanced Scientific Computing Research, DOE SC, Tysons, Virginia, March 22-23, 2016. https://www.osti.gov/servlets/purl/1344785/

7 E. Dart, et al., The Petascale DTN Project: High Performance Data Transfer for HPC Facilities, arXiv:2105.12880

8 D.P. Schissel, et al., Remote third shift EAST operation: a new paradigm (2017) Nucl. Fusion Vol. 57, 056032, https://doi.org/10.1088/1741-4326/aa65a8

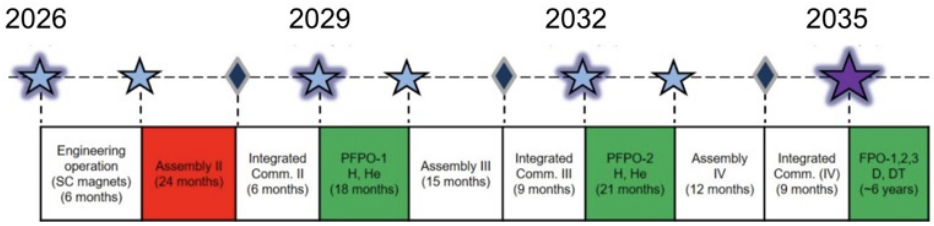


Figure 5.6.3 - Timeline from the start of engineering commissioning including first-plasma (2026) to fusion power operation (2035).

5.6.2.5 Remote Science Activities

While ITER experiments have not yet begun, much of the US scientific community intends to participate remotely in ITER research. Therefore the US team will need to be able to have real-time interactions among the experimental team as well as conduct interactive visualization and processing of large data sets. Access to real-time data has proven highly valuable with the remote EAST work and it is anticipated the same will be true for ITER. Thus, there is a three-phase vision for data transfer from ITER to the US The first is the smaller (though by no means small) real-time data, the second is the key physics quantities transferred on a timescale that allows examination between pulses, and the third is everything else.

As with fusion research today, large leadership-class computing facilities will be used to support the US research effort but it is anticipated that smaller clusters at the collaborating institutions will contribute significantly as well. Decisions have yet to be made by DOE regarding ITER data centers in the US. However, it is expected that ITER's full data set will be mirrored at a site (or sites) in the US and used by US researchers. By the mid-2030s, storage is expected to reach the exabyte level.

In addition to data generated at ITER and transmitted to the US, it is expected that a significant volume of simulated data generated and stored in the US and mirrored back to ITER, to be stored in the IMAS framework. The data centers should have substantial local analysis computing available, since computing must follow the data. This would imply midrange computing facilities at the data centers in the few to 10s of petaflop.

It is also expected that advanced visualization methods will be developed that allow for remote visualization of large and complex data sets, combining modeling and experimental data. Visualization becomes an important issue when rapid assessment of multi-model data and modeling analysis must take place in a limited time for the support of ITER operations. Major computing capability is also needed to assist with data visualization, as well as dedicated visualization infrastructure.

5.6.2.6 Software Infrastructure

The fusion community is currently developing various workflow tools to manage data analysis, modeling and simulations, making use of software libraries of analysis tools and providing standardized data I/O interfaces. These tools (like OMFIT[9]) are proving very popular with the research community. Other workflows are being developed to enable tightly coupled simulation packages to work together in integrated modeling,

9 O. Meneghini et al., Integrated modeling applications for tokamak experiments with OMFIT (2015) Nucl. Fusion Vol. 55, 083008, https://doi.org/10.1088/0029-5515/55/8/083008

from platforms like TRANSP up to midrange simulation codes developed by the DOE SciDAC projects that can typically run on medium scale computers currently in the 1-2 petaflop range. Further into the future, the ECP is producing even more tightly coupled multi-physics models that will require leading edge HPC infrastructure.

Looking forward to ITER operation, it is expected that a range of computation activities from the local linux cluster or several petaflop-level computing resources and then to the leadership-class national computing facilities for advanced simulation and analysis.

5.6.2.7 Network and Data Architecture

It is too early to comment on the data architecture; however, it is expected that the primary network for transmitting data from ITER to the US and for then subsequently serving data to the US user community will be through ESnet connecting to the European ring network that currently supports the LHC experiment.

5.6.2.8 Cloud Services

The fusion community has not really explored cloud services for data analysis, however there are developments in this area. Cloud-based services will likely evolve considerably by the time ITER comes online.

5.6.2.9 Data-Related Resource Constraints

It is too early to comment, given that the first plasma is in late 2026. However, the experience of the HEP community will be useful when the time comes.

5.6.2.10 Outstanding Issues

ITER has not yet finalized requirements for scientific data access and network needs for first-plasma in 2026. Current estimates suggest up to 20 TB of data per day will be generated during Engineering Operations. Engineering operation follows first-plasma and could achieve up to 1 MA of plasma current with up to 8 MW heating power (ECH), and with an array of important diagnostics including full magnetics for equilibrium analysis. The IO is seeking input from the ITER members (including input rom ESnet through FES) on data needs in order to finalize data storage and network requirements.

The ITER Project Requirements state very generally that computing resources for data processing must be provided by ITER and that a separate archive must be provided >50 km from the primary storage site. ITER is installing a 100 Gbps network connection from Cadarache to Marseille which is over 50 km away, where the separate data archive will be located. Marseille is also the location where the ITER network will connect to the European ring network that serves the LHC at CERN. At this point of connection, ESnet is very familiar with the network and can use their accumulated experience with CERN to help prepare for US ITER networking needs.

From the US perspective, it is anticipated that there will be some US researchers on-site at ITER but the majority will be remotely located throughout the US The ability for anyone to effectively participate in ITER experiments is predicated on timely access to the data. It is therefore critical that the requirements for ITER's data workflow be clearly stated as it pertains to remote participants. As has been stated previously, ITER's data can be grouped into different categories delineated by the time criticality of the access. The first data to reach US researchers and engineers should be

the real-time data that is accessible to the control room and used in the control system for real-time analysis. It was this type of real-time remote data that was effective in allowing the DIII-D National Fusion Facility to operate remotely under COVID-19 restrictions[10]. In addition, if real-time data streaming can be expanded to many more signals, including those available to the PCS, then more sophisticated analysis can be performed by remote participants in the US to support ITER operations, including but not limited to near-real-time equilibrium reconstruction and post-equilibrium analysis.

There is great potential for real-time data to enable effective remote participation in ITER experiments from the US. Coordination between FES, ESnet, ASCR and the IO will be essential to design and deliver a data-streaming capability that best serves the needs of US participants during ITER operation.

The next group of data would be a bulk transfer of select (if not all) datasets. This one-time transfer would be to a centralized US storage facility that could then serve data to the rest of the US community. This model allows rapid transfer (the Science DMZ model) and eliminates the competition on limited transatlantic network throughput (the data is transferred once rapidly for all to then share). This data needs to arrive in a very timely manner so that it can be digested by the remote scientists on the same timescale as their counterparts within the ITER control room. Ideally, this data will arrive between plasma pulses for use in analysis that can support ongoing experiments. If remote scientists do not have the data, then it is impossible for them to practically participate in the scientific conversation with their on-site colleagues. Given the quantity of ITER data and the short period of time in which to make the transfer, the networking requirements between the IO and the US are substantial. Early on, during first plasma experiments, present 100 Gbps networking capability will satisfy those needs. However, this assumes that a large fraction of the theoretical bandwidth can be utilized on a routine basis, which will require careful planning to ensure that all components are designed and tested to meet this level of performance. But as the ITER project progresses, the network requirements will rise substantially, reaching the TB/s level. Clearly, this type of networking infrastructure needs to be planned well in advance.

It should be noted that for some researchers, it will be more efficient to remotely log into ITER computer resources and utilize a remote desktop session to do data visualization and analysis. On the other hand, it is envisioned that the IO computing resources will be limited compared to ITER member resources, so careful planning is needed to provide the necessary bulk data transfer needs. The bulk transfer of data does not preclude remote login to the IO, but instead adds an important degree of flexibility to support operations that would not be possible without the timely transfer of data.

Finally, the remainder of the ITER data requiring extensive on-site processing (e.g., using local HPC) would be transferred, perhaps within hours or overnight, so that the US Feeder center(s) would have a complete mirror of the ITER data to service US users. This final transfer of highly processed data could be facilitated using either local computing resources or distributed resources (e.g., NERSC) accessed via ESnet. The remote operation of EAST, KSTAR and other fusion experiments, and the marrying

10 U.S. DOE. 2021. Office of Science User Facilities: Lessons from the COVID Era and Visions for the Future. Report from the December 2020 Roundtable. https://doi.org/10.2172/1785683

of leadership-class computing facilities to tokamak operations, demonstrated by US researchers, shows that the above vision can be successfully implemented.

5.6.2.11 Case Study Contributors

Planning for ITER Operation Representation

- Raffi Nazikian[11], PPPL
- David Schissel[12], GA

ESnet Site Coordinator Committee Representation

- Scott Kampel[13], PPPL
- Jeff Nguyen[14], GA

11 rnazikia@pppl.gov
12 schissel@fusion.gat.com
13 skampel@pppl.gov
14 nguyend@fusion.gat.com

5.7 Public-Private Partnerships in Fusion Research

5.7.1 Discussion Summary

DOE FES provides funds for business awards to assist applicants seeking access to the world class expertise and capabilities available across the US DOE complex. This is one component of the Innovation Network for Fusion Energy (INFUSE), a DOE initiative to provide the fusion industrial community with access to the technical and financial support necessary to move new or advanced fusion technologies toward realization with the assistance of the national laboratories. The objective of INFUSE is to accelerate basic research to develop cost-effective, innovative fusion energy technologies in the private sector.

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

- The INFUSE program features public-private partnerships with non-DOE entities that are funded to perform aspects of FES research. Many of these entities are unfamiliar with mechanisms to interact with DOE SC facilities including ASCR HPC centers and ESnet.
- DOE programs that span facilities and communities (e.g., INFUSE) do not typically require a data architecture review to facilitate sharing of experimental results; solutions in this space can vary between facilities. While organic approaches have scaled to date, the lack of a cohesive and shared understanding of best practices as data volumes increase will begin to harm productivity. Having access to community recommended approaches, and potentially more efficient data transfer hardware and software, would benefit participants and lead to more efficient use of resources over time.
- INFUSE has requested assistance from ESnet to provide a briefing for their community on scientific data management approaches for current awardees, and help in developing a BCP for future participants. Topics may include data transfer hardware and software, along with network design, security policy, and ways to interact with DOE SC resources such as HPC facilities.
- DOE programs that span facilities and communities (e.g., INFUSE) do not include access to generalized pools of computational resources that can be utilized by participants. While it is possible for participants to pursue these resources independently from DOE HPC facilities, it is a secondary step that must be managed independently. Having access to computational resources, and potentially more efficient data transfer and analysis tools, would benefit participants and lead to more efficient use of resources over time.

5.7.2 Public-Private Partnerships in Fusion Research Case Study

The INFUSE program will accelerate fusion energy development in the private sector by reducing impediments to collaboration involving the expertise and unique resources available at DOE laboratories. This will ensure the nation's energy, environmental and security needs by resolving technical, cost, and safety issues for industry.

5.7.2.1 Background

The INFUSE (Innovation Network for Fusion Energy) program is a DOE SC public-private partnership managed through the SC FES Program to facilitate collaboration between DOE national laboratories fusion researchers, and private companies pursuing commercial fusion energy development. The director and deputy director are located at ORNL and the PPPL, respectively. INFUSE is governed by a 10-member committee known as the "Point-of-Contacts Committee," and POC members are appointed at each of the 10 participating DOE labs. The INFUSE program's goal is to lower the hurdles and challenges faced by private companies when interacting with the government labs, so that decades of fusion plasma research and technology development can be transferred from the labs to private industry interested in commercializing the technology for energy production. This knowledge focuses on electricity, hydrogen and process heat.

INFUSE was started in 2019, loosely modeled after the GAIN (Gateway for Accelerated Innovation in Nuclear) program developed by DOE NE a few years earlier. INFUSE provides the cyber infrastructure for evaluation of Request for Assistance applications in which a company requests that DOE fund a national laboratory to collaborate with them to solve a scientific issue that is hindering their progress to develop fusion energy. The RFAs are reviewed following DOE SC protocols from two annual RFA submission periods or cycles. One is in the fall of the year, and the second is in early summer. There are five topical technology areas in which RFAs can be considered consistent with the mission space of DOE SC. Funding awards are made directly and only to the national laboratory to assist a company through a Cooperative Research and Development Agreement (CRADA) in which the company must provide a 20% cost share for the project either with in-kind work or direct cash payments to the laboratory.

INFUSE itself is not a direct user of ESnet services; however, collaboration between the companies and the INFUSE laboratories could use ESnet services regarding HPC simulations and real-time analysis of large confinement experiments scattered throughout the US and several foreign countries. For this exercise, INFUSE requested four of the larger private companies to participate. The foreign-controlled companies, General Fusion of Canada and Tokamak Energy (ST40) of the U.K. chose not to respond. However, case study input from the two US companies, TAE Technologies located in California and Commonwealth Fusion Systems (CFS) located in Massachusetts, was received. Below are specifics from each of these companies (text provided by each company):

5.7.2.1.1 TAE Technologies

TAE Technologies, Inc develops breakthrough solutions to complex problems. The company was founded in 1998 to develop and distribute safe, cost-effective commercial fusion energy with the cleanest environmental profile. With over 1,400 issued and pending patents and over $750 million USD in private capital, TAE is making major contributions to the development of a transformational energy source capable of sustaining the planet for thousands of years. The primary research performed is well within the interest of the DOE FES, and it is making practical strides to address key research needs outlined in the US DOE sponsored Toroidal Alternates Panel report

(2008)[1] and Research Needs Workshop (2009)[2].

TAE Technologies combines accelerator physics and plasma physics to solve the challenge of fusion. To this end, TAE has developed the advanced beam-driven field-reversed configuration (FRC) concept over the last 20 years in the C-1[3], C-2[4 5], C-2U[6 7], and C-2W (aka Norman)[8] series of experimental devices, with concurrent development of simulation capability[9]. The company's next-step Advanced FRC experiment, dubbed Copernicus, is currently in development. Copernicus will be a reactor-scale prototype which will use Hydrogen plasmas to demonstrate the viability of net-energy production with D-T fuel. TAE Technologies will then construct a final prototype to demonstrate net-energy gain of p-B11 fuel.

The C-2W experiment has a comprehensive suite of diagnostics that includes over 700 magnetic sensors, four interferometer systems, two Thomson scattering systems, ten types of spectroscopic measurements, multiple fast imaging cameras, bolometry, reflectometry, neutral particle analyzers, and fusion product detectors; over 4000 raw signals are collected into an MDSplus database with each experimental shot[10 11]. The signals are analyzed using a variety of computational methods including reduced model physics analysis and ML. All raw experimental data is stored permanently with redundancy.

A hierarchy of simulation models including 0D, 1D, 2D, and 3D representations using fluid based and particle-in-cell algorithms are used to perform predictive and interpretive simulations of the experiments. These simulations are performed in an in-house HPC cluster and at DOE Leadership Computing Facilities (LCF). Raw simulation data is stored temporarily for analysis; reduced simulation data is stored permanently with redundancy.

5.7.2.1.2 Commonwealth Fusion Systems

CFS enables worldwide clean energy for everyone, creating a sustainable environment

1 D. N. Hill and R. Hazeltine, "Report of the FESAC Toroidal Alternates Panel," US DOE, Washington DC, 2008.

2 R. Hazeltine, "Research needs for magnetic fusion energy sciences, Report of the Research needs Workshop (ReNeW)," US DOE, Washington DC, pp. 171–227, 2009.

3 N. Rostoker, M. Binderbauer, E. Garate, and V. Bystritskii, "Formation of a field reversed configuration for magnetic and electrostatic confinement of plasma," US6891911B2, May 10, 2005.

4 M. W. Binderbauer et al., "Dynamic Formation of a Hot Field Reversed Configuration with Improved Confinement by Supersonic Merging of Two Colliding High-$\textbackslashbeta$ Compact Toroids," Phys. Rev. Lett., vol. 105, no. 4, p. 045003, Jul. 2010, doi: 10.1103/PhysRevLett.105.045003.

5 M. Tuszewski et al., "A new high performance field reversed configuration operating regime in the C-2 device," Physics of Plasmas (1994-present), vol. 19, no. 5, p. 056108, May 2012, doi: 10.1063/1.3694677.

6 M. W. Binderbauer et al., "A high performance field-reversed configuration," Physics of Plasmas (1994-present), vol. 22, no. 5, p. 056110, May 2015, doi: 10.1063/1.4920950.

7 H. Y. Guo et al., "Achieving a long-lived high-beta plasma state by energetic beam injection," Nature Communications, vol. 6, p. 6897, Apr. 2015, doi: 10.1038/ncomms7897.

8 H. Gota et al., "Formation of hot, stable, long-lived field-reversed configuration plasmas on the C-2W device," Nucl. Fusion, vol. 59, no. 11, p. 112009, Jun. 2019, doi: 10.1088/1741-4326/ab0be9.

9 S. A. Dettrick, D. C. Barnes, and Belova, E. V., "Simulation of Equilibrium, Stability, and Transport in Advanced FRCs," presented at the accepted for IAEA 2020, Proceedings of the 28th IAEA int. Conf, Nice 2021, paper TH/P2-19 (International Atomic Energy Agency, Vienna, 2020).

10 T. Roche et al., "The integrated diagnostic suite of the C-2W experimental field-reversed configuration device and its applications," Review of Scientific Instruments, vol. 92, no. 3, p. 033548, Mar. 2021, doi: 10.1063/5.0043807.

11 M. C. Thompson, T. M. Schindler, R. Mendoza, H. Gota, S. Putvinski, and M. W. Binderbauer, "Integrated diagnostic and data analysis system of the C-2W advanced beam-driven field-reversed configuration plasma experiment," Review of Scientific Instruments, vol. 89, no. 10, p. 10K114, Oct. 2018, doi: 10.1063/1.5037693.

for current and future generations. CFS takes the proven scientific foundation of tokamaks and are developing new, high-field superconducting magnets to significantly reduce the size, cost, and iteration cycle of tokamaks to net fusion energy. The CFS vision is to systematically risk-retire the elements of a simplified, compact, high-field tokamak-based power reactor into a marketable and manufacturable product in a timeline that is consistent with reducing greenhouse gases and at a cost that can provide a wide adoption and return to investors. Toward that end, the CFS product roadmap has three phases: (1) to develop high-field magnets based on high-temperature superconductors culminating in large-bore magnets, this phase is currently underway and will complete in mid-2021. (2) Incorporate those magnets into a DT-burning DIII-D-sized tokamak called SPARC to demonstrate Q>2, >50MW of fusion energy production, and high-field plasma operating scenarios in a pulsed operation. CFS anticipates this phase to be completed by 2025. (3) Incorporate the developed technologies and others in heat extraction, sustainment, blankets and materials and understanding into a net-electricity pilot plant called ARC (affordable, robust, compact) fusion reactor that is a demonstration of a commercial product at approximately 200MW electric power.

Collaborators include US National Laboratories, universities, private companies and foreign organizations, and stakeholders include private investors, DOE, Nuclear Regulatory Commission and Massachusetts state leadership.

The SPARC facility (projected to come online in 2025) will be a unique platform on which to test reactor-relevant plasma physics. The SPARC device will not only demonstrate the reactor-relevant physics but will do so in a manner that makes the plasmas and machine conditions conducive to scientific study and quantification at a level that is sufficient to build confidence on projections to ARC. Various plasma data will be collected from SPARC experiments and analyzed depending on the diagnostics. Validation of codes with plasma data will also be performed.

5.7.2.2 Collaborators

5.7.2.2.1 TAE Technologies

TAE Technologies, Inc (TAE) is a privately held company in the fusion space with significant collaborations in the private and public realms. On the private side, TAE has a long-running ML collaboration with the Google AI team. Novel ML techniques are used both to optimize the experimental performance[12] and to reconstruct the plasma state from the experimental data using Bayesian Inference (BI)[13]. To this end, the entire experimental database is mirrored to cloud daily using a custom built rsync-like interface to Google cloud, with up to ~9 TB uploaded daily. BI is performed by the Google AI team on a 25 Petaflop arsenal of 200 NVIDIA Tesla V100 GPUs. Reduced data is downloaded back to the TAE computing center.

On the public side, TAE has several active collaborations with PPPL and ORNL funded by DOE through the INFUSE program. TAE has had at various times HPC allocations at DOE LCF using awards such as INCITE (at ALCF), ALCC (at ALCF and NERSC), ERCAP (at NERSC), and Director's discretion (at OLCF, ALCF, and NERSC). TAE

12 E. A. Baltz et al., "Achievement of Sustained Net Plasma Heating in a Fusion Experiment with the Optometrist Algorithm," Scientific Reports, vol. 7, no. 1, Art. no. 1, Jul. 2017, doi: 10.1038 s41598-017-06645-7.

13 M. Dikovsky et al., "Reconstruction of fusion plasma state with a Plasma Debugger," p. BM10.002, 2018.

mainly tries to do data analysis and visualization on the LCF visualization servers, through Visit and Jupyter clients, thus avoiding data transport. However TAE does sometimes feel the need to download the whole simulation output which can be in the 0.1 – 1 TB range per simulation. At present, rsync is the main transfer tool, however this is not as efficient as desired, and TAE may explore improvements in both tools and data transfer infrastructure.

TAE also has private-public collaborations funded from the private side, with domestic public partners including UC Irvine, PPPL, LLNL, TUNL, UCLA, U. Wisconsin, U. Washington, and U. Florida, and international partners including Nihon University, Japan, and Budker Institute of Nuclear Physics, Russia. Data transfer is typically not an issue with those collaborations as they are usually hardware-oriented.

User/Collaborator and Location	Do they store a primary or sesecondary copy of the data?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of dataset? (report in bytes, e.g. 125GB)	Frequency of data transfer or download? (e.g. ad-hoc, daily, weekly, monthly)	Is data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g. difficult tools, slow network)?
Brookhaven National Laboratory	Y	Email and USB	200 MB	Ad-hoc	N	N
Princeton Plasma Physics Laboratory	Y	Email / Google Drive	0.1-1 GB	Ad-hoc	Y Email/ Google Drive	N
Lawrence Livermore National Laboratory	Y	Email / Google Drive	10 MB	Bi-weekly	Y Email / Google Drive	N
Oak Ridge National Laboratory	Y	Email / Google Drive	1-10 GB	per-test, camera data	N	N
Idaho National Laboratory	Y	Email / Google Drive	1-10 MB Can go up to 10 GB	Bi-weekly	Y Email / Google Drive	N

Table 5.7.1 – CFS Data Relationships

5.7.2.2.2 CFS

Timescale	Facilities/ Instruments	Data sets (file size, # files, total data set size)
0-2 yrs	Modeling (Gyrokinetic and alpha particle simulations)	400 GB per year
2-5 yrs	Modeling (Gyrokinetic and alpha particle simulations)	1 TB per year
>5 yrs	Modeling (Gyrokinetic and alpha particle simulations)	1 TB per year

Table 5.7.2 – CFS Data Sizes

5.7.2.3 Instruments and Facilities

5.7.2.3.1 TAE Technologies

The current C-2W experiment has a comprehensive suite of diagnostics that includes over 700 magnetic sensors, four interferometer systems, two Thomson scattering systems, ten types of spectroscopic measurements, multiple fast imaging cameras, bolometry, reflectometry, neutral particle analyzers, and fusion product detectors; over

4000 raw signals are collected into an MDSplus database with each experimental shot. The signals are analyzed using a variety of computational methods including reduced model physics analysis and ML. All raw experimental data is stored permanently with redundancy.

5.7.2.3.2 CFS

Timescale	Facilities/ Instruments	Data sets (file size, # files, total data set size)
0-2 yrs	Modeling (Gyrokinetic and alpha particle simulations)	400 GB per year
2-5 yrs	Modeling (Gyrokinetic and alpha particle simulations)	1 TB per year
>5 yrs	Modeling (Gyrokinetic and alpha particle simulations)	1 TB per year

Table 5.7.2 – CFS Data Sizes

5.7.2.4 Process of Science

5.7.2.4.1 TAE Technologies

TAE has developed the advanced beam-driven FRC concept over the last 20 years in the C-1, C-2, C-2U, and C-2W (aka Norman) series of experimental devices, with concurrent development of simulation capability. The company's next-step Advanced FRC experiment, dubbed Copernicus, is currently in development. Copernicus will be a reactor-scale prototype which will use Hydrogen plasmas to demonstrate the viability of net-energy production with D-T fuel. TAE Technologies will then construct a final prototype to demonstrate net-energy gain of p-B11 fuel.

5.7.2.4.2 Commonwealth Fusion Systems (CFS)

Timescale	Facilities/ Instruments	Science workflow	Data analysis, reduction methods
>5 yrs	SPARC Tokamak	Generating burning plasma data	Various analysis methods depending on the diagnostic. Validation of codes with plasma data.

Table 5.7.3 – CFS Science Process

5.7.2.5 Remote Science Activities

5.7.2.5.1 TAE Technologies

TAE has no remote science drivers at this time.

5.7.2.5.2 CommonwealthFusion Systems (CFS)

Timescale	Data management software tools	Purpose
>5 yrs	MDSplus	SPARC experimental data handling

Table 5.7.4 – CFS Remote Science

5.7.2.6 Software Infrastructure

5.7.2.6.1 TAE Technologies

TAE uses MDSplus and MySQL databases to store experimental data, and HDF5 and ADIOS2 libraries to store simulation data. Experimental data is uploaded to Google cloud by custom rsync-like Google transport software. Simulation data is analyzed

on-site using DOE LCF visualization servers, or downloaded from DOE LCF machines by rsync or scp. TAE is exploring the use of other databases for storage and retrieval of experimental analysis and simulation data, and for simulation and analysis workflow reproducibility.

5.7.2.6.2 CFS

Timescale	Network capabilities
0-2 yrs	CFS is building a new campus and SPARC tokamak in Devens MA. Employees will be moving to a new campus in 2022. Network capabilities will be determined over the next several months.
2-5 yrs	TBD
>5 yrs	TBD

Table 5.7.5 – CFS Software

5.7.2.7 Network and Data Architecture

5.7.2.7.1 TAE Technologies

TAE has a dedicated optical fiber to a collocated HPC center. TAE is beginning to investigate the use of Science DMZ tools and capabilities.

5.7.2.7.2 CFS

Timescale	Network capabilities
0-2 yrs	CFS is building a new campus and SPARC tokamak in Devens MA. Employees will be moving to a new campus in 2022. Network capabilities will be determined over the next several months.
2-5 yrs	TBD
>5 yrs	TBD

Table 5.7.6 – CFS Networking

5.7.2.8 Cloud Services

5.7.2.8.1 TAE Technologies

TAE currently makes heavy use of Google cloud through collaborators at Google AI.Novel ML techniques are used both to optimize the experimental performance and to reconstruct the plasma state from the experimental data using BI. To this end, the entire experimental database is mirrored to cloud daily using a custom built rsync-like interface to Google cloud, with up to ~9 TB uploaded daily. BI is performed by the Google AI team on a 25 Petaflop arsenal of 200 NVIDIA Tesla V100 GPUs. Reduced data is downloaded back to the TAE computing center.

5.7.2.8.2 CFS

Timescale	Cloud service plans
0-2 yrs	TBD
2-5 yrs	TBD
>5 yrs	TBD

Table 5.7.7 – CFS Cloud Usage

5.7.2.9 Data-Related Resource Constraints

5.7.2.9.1 TAE Technologies

There are no data-related resource constraints at this time.

5.7.2.9.2 CFS

Timescale	Data-related constraints
0-2 yrs	None as of May 2021
2-5 yrs	TBD
>5 yrs	TBD

Table 5.7.8 – CFS Resource Constraints

5.7.2.10 Outstanding Issues

TAE would like to use ESnet and GridFTP or Globus depending on availability.

Communication and coordination between ESnet and FES Public-Private Partners is limited, and in some cases, there may be services and capabilities that ESnet or DOE UF may be able to provide in support to these partnerships which may boost efficiency (particularly in the area of data management) now, or in future. Recommend that a periodic touch-in/ESnet SET outreach occur between ESnet and the INFUSE POC Committee be established to identify whether such opportunities and needs exist.

5.7.2.11 Case Study Contributors

Public-Private Partnerships in Fusion Research Representation

- Youchison[14], ORNL
- Ahmed Diallo[15], PPPL
- Bob Mumgaard[16], CFS
- Dan Brunner[17], CFS
- Brandon Sorbom[18], CFS
- Alex Creely[19], CFS
- Matthew Reinke[20], CFS
- Sean A. Dettrick, TAE Technologies, Inc.
- Jack M. Margo, TAE Technologies, Inc.

ESnet Site Coordinator Committee Representation

- Susan Hicks[21], ORNL
- Scott Kampel[22], PPPL

14 youchisondl@ornl.gov
15 adiallo@pppl.gov
16 bob@cfs.energy
17 dan@cfs.energy
18 brandon@cfs.energy
19 alex@cfs.energy
20 mreinke@cfs.energy
21 hicksse@ornl.gov
22 skampel@pppl.gov

5.8 MPEX at ORNL

5.8.1 Discussion Summary

The MPEX is a next-generation linear plasma device that will support study of the way plasma will interact long term with the components of future fusion reactors. MPEX represents a shift from the historical direction of the plasma-material interaction field, which for many years focused on the effect that materials had on plasma, but not on the effect that plasma had on materials

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

- The MPEX experiment at ORNL is under design, and will be operational by 2027. The MPEX project at ORNL is currently in the DOE 413.3b project phase and has passed CD-1.
- The standard short-pulse use case will produce:
  - An estimated 50 GB of scientific data per run day, with 100 run days per year. This is an estimated 5 TB of data per year.
  - Visible light cameras will be used for measuring the target surface, and will produce raw video data streams at 1 Gbps. Up to six cameras can be used at various angles during a run period and can generate just under 4 TB of raw data frames per hour, or up to 24 TB per h if all cameras are operating.
  - A single IR camera can be used for measuring surface materials interactions, and it is estimated to produce raw data rates at 9 Gbps or 32 TB per h.
  - Lastly, there are approximately 35,000 archived signals for operational data stored in a relational database. The archived data consumes approximately 17 GB/day or 6.2 TB/year.
- A second use case, consisting of a longer pulse (2 weeks of continuous operation), has the potential to generate 1 PB of scientific experimental data. The camera rates listed above will apply as well, but will be limited to the 2 week operational period
- MPEX will expose data via recommended mechanisms that ORNL and OLCF support (e.g., HTTP portals, RSYNC, SCP). It is expected that data long-term storage and archiving is managed at ORNL.
- MPEX is designing experimental workflow, and will approach data handling similar to other large-scale experiments: saving "RAW" data to archival storage, and generating a system to reduce information to formats that are easy to process and share.
- Data will be produced mainly on MPEX with its installed diagnostics. Some post-mortem analysis of material samples will take place in other locations by collaborators. Collaborators will have access to raw and processed data on

MPEX and might transfer parts of data for further analysis or processing.

- ESnet and ORNL will collaborate on participating in the DME as MPEX is designed to perform data trials between the facility and external collaborators. This will ensure ORNL infrastructure is properly tuned between the experimental enclave, and ESnet.
- As an emerging experiments, MPEX will adopt the use of DOE HPC resources for some aspects of the experimental workflow. This is expected to be in the form of NERSC and OLCF, although there are ongoing discussions as MPEX is implemented. MPEX could potentially transfer TB to PB volumes of diagnostic data, output from experimental cameras, and simulation workflows to an external DOE HPC facility.

5.8.2 MPEX at ORNL Case Study

The MPEX is a linear plasma device to address the challenges of plasma-material interactions for future fusion reactors. MPEX will be a world-leading facility, able to expose materials and components to fusion reactor diverter relevant plasma conditions. This includes the capability to test material samples, which have been pre-irradiated with neutrons in fission reactors like the High Flux Isotope Reactor (HFIR) for example. MPEX will be a steady-state device able to expose components to long pulses up to deuterium ion fluences of 1e+31 per square meter. MPEX is unique with its capabilities to reach the plasma conditions expected in diverter plasma utilizing a novel high-power helicon plasma source, as well as an electron heating and ion heating system. The plasma exposed materials will be monitored with in-situ diagnostics as well as dedicated surface analysis tools, e.g., FIB/SEM, IBA-NRA in-vacuo to provide information on the surface evolution and hydrogen transport in the material for example.

5.8.2.1 Background

The MPEX project ($120M) is currently in the DOE 413.3b project phase and has passed CD-1. It is expected that the project is completed by 2027, and scientific exploitation will start subsequently. All information given via this case study is related to the strategic planning event horizon of five years into the future.

The user community will be domestic fusion and material scientists from ORNL, UT-K, UCSD, INL, GA, PPPL, MIT, UW-Madison, Penn State, UIUC as well as international partners with the IEA Plasma Surface Interaction Facilities collaborators network, and ITER. Collaborators will have access to raw and processed data on MPEX and might transfer parts of data for further analysis or processing. It is expected that data long-term storage and archiving is managed at ORNL.

Data will be produced mainly on MPEX with its installed diagnostics. Some post-mortem analysis of material samples will take place in other locations. This will include specific material analysis for thermo-mechanical tests, TEM lift-outs, thermal desorption spectroscopy and high resolution microscopy. The amount of data from these post-mortem analysis techniques will be small in comparison to the data accumulated during the plasma exposure.

5.8.2.2 Collaborators

Collaborating institutions listed in Table 5.8.1 are assumed to participate in a 4-day experiment, 8 hours a day. This will happen sporadically, maybe once every 2 years per institution. Each day is estimated to be 50 GB of data. Collaborating institutions will want a data transfer of that data from the host ORNL institution database to their remote location.

User/Collaborator and Location	Do they store a primary or secondary copy of the data?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of dataset? (report in bytes, e.g. 125GB)	Frequency of data transfer or download? (e.g. ad-hoc, daily, weekly, monthly)	Is data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g. difficult tools, slow network)?
UCSD	No	data transfer	200 GB	ad-hoc	no	no
INL	No	data transfer	200 GB	ad-hoc	no	no
Penn State	No	data transfer	200 GB	ad-hoc	no	no
UW Madison	No	data transfer	200 GB	ad-hoc	no	no
UIUC	No	data transfer	200 GB	ad-hoc	no	no
UT-Knoxville	No	data transfer	200 GB	ad-hoc	no	no
Japan collaboration	No	data transfer	200 GB	ad-hoc	no	no
IEA collaboration	No	data transfer	200 GB	ad-hoc	no	no
PPPL	No	data transfer	200 GB	ad-hoc	no	no
General Atomics	No	data transfer	200 GB	ad-hoc	no	no
MIT	No	data transfer	200 GB	ad-hoc	no	no

Table 5.8.1 – MPEX Data Relationships

5.8.2.3 Instruments and Facilities

MPEX has a wide range of instrumentation ranging from visible cameras, fast IR cameras, thermocouples, laser diagnostics, visible spectrometers, scanning electron microscopes, and high-energy light ion beams. Each component is capable of producing data that is relevant to the outcome of an experimental run, and will have an affiliated workflow.

MPEX is estimating that the total scientific data output can be broken down in the following manner:

- 50 GB of scientific data per run day with 100 run days per year. This is an estimated 5 TB of data per year.
- Visible light cameras for measuring the target surface produce raw video data streams at 1 Gbps. Up to six cameras can be used at various angles during a run period. Typical runs periods range from several hours, to up to two weeks in duration. These cameras generate just under 4 TB of raw data frames per hour, or up to 24 TB per h if all cameras are operating.
- A single IR camera for measuring surface materials interactions produces raw data rates at 9 Gbps or 32 TB per hour.
- There are approximately 35,000 archived signals for operational data stored in a relational database. The archived data consumes approximately 17 GB/

day or 6.2 TB/year.

All of the aforementioned operational data is stored and maintained on RAID clusters to permit access to archived data over the life of the facility. MPEX plans to store all data locally at ORNL. Remote users will need an ORNL UCAMS account to remotely access MPEX user network, which will have access to all the data. The remote user will also be provided with a SCP service to transfer data from ORNL to the remote location if desired.

5.8.2.4 Process of Science

The MPEX science program, the mechanism that will be used to grant time for experimentation, will be based on proposals that can be written by in-house scientists and domestic collaborators. It is also expected that within the frame of international collaboration networks, like the IEA PSI IA, ITPA (for ITER), US-JP bilateral agreements, that joint experiments with international partners will be carried out. Experiments will be executed by ORNL operations staff, and collaborators will be able to participate in experiments locally or by remote participation options. The remote participation use case will require the ability to maintain large data flows that show results and communication channels during the run, or just after the plasma pulse.

Data analysis methods will convert raw signals from diagnostics (cameras, lasers, microscopes) into useful scientific data. This includes standard algorithms for image processing and manipulating time-series datasets. Correlation, clustering, and regression analysis will often be used to often understand the results from the generated scientific datasets.

Simulations will often be used to understand experimental data either by local MPEX scientific staff or remote collaborators. This can range from finite element analysis to particle-in-cell to molecular dynamic simulations. These simulations may use shared local ORNL clusters or remote supercomputing facilities provided by other funding sources (e.g., OFES/ASCR SciDACs for supercomputing time on NERSC, ALCF, ORNL overhead for clusters).

5.8.2.5 Remote Science Activities

The primary remote use case for MPEX will be the use of DOE HPC resources, such as NERSC, that can be used for the analysis of experimental data sets of the creation of simulation data. Along with the use of centralized DOE HPC resources, smaller cluster resources at ORNL CADES will be leveraged .

Any user with an MPEX account is able to observe all operations data in real time via a remote login. A single server is dedicated to remote operations and resides external to the facility firewall but within the ORNL network. This server is restricted to receiving copies of operational data only. Data writes from this server are not permitted. Access to this server is achievable from any workstation within the ORNL network as long as the user has an MPEX account. Access to this server from outside of the ORNL network requires a user to first establish a remote connection to the ORNL network.

5.8.2.6 Software Infrastructure

The Hierarchical Data Format (HDF5)[1] is used to manage operations and scientific data. HDF5 provides a series of APIs that can be accessed via C, C++, Fortran,

and Java. It is designed to handle large time-series data sets, metadata, and is able to simplify integrating a large number of analysis tools including the operations archive, MATLAB, Mathematica, IDL, and MDS+.

All of the MPEX facility data is stored locally on RAID arrays. Standard data transfers to external facilities will be accomplished using the HTTPS 2.0 protocol, along with

tools like rsync, scp, and sftp.

5.8.2.7 Network and Data Architecture

The networking top support MPEX is separated into four sections. Dedicated networks are allocated to support:

- Operations
- Process control applications
- Diagnostics
- Users

Separating the networks ensures operational data transfers do not impede real-time functional requirements for process control applications. A dedicated diagnostic network provides dedicated services to high-throughput devices without being interrupted by operational data transfers. The user network provides timing synchronization through NTP for user equipment and data sets. User computing devices are not permitted on the facility network for security measures, so this network provides a timing reference only. A connection to the ORNL network is provided through the operations network. All network connections use 1-10 Gbps links depending on the resources that are connected to a particular switch. A 10 Gbps link is used to bridge the facility network to the ORNL network.

Currently, ORNL connects to ESnet via redundant 100 Gbps connections, one to the ESnet hub on-site at ORNL and another to the ESnet router in Nashville. The Nashville connection utilizes the ORNL optical line system from the ORNL secondary border router to the ESnet router in Nashville. ORNL also maintains a connection to the Southern Crossroads in Atlanta that is currently 10 Gbps but is in the process of being upgraded to 100 GBps. Current connectivity is depicted in Figure 5.8.1.

1 https://support.hdfgroup.org/HDF5/doc/H5.intro.html

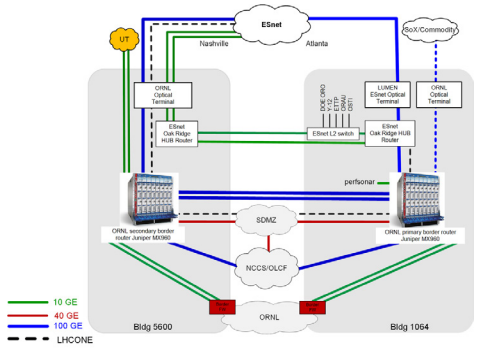


Figure 5.8.1 – Current ORNL Network and Data Architecture

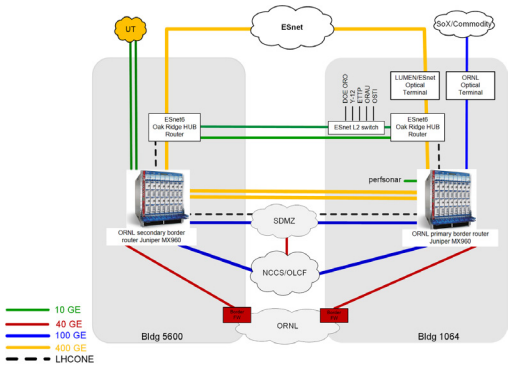


Figure 5.8.2 – Future ORNL Network and Data Architecture

With the installation of ESnet6 routers at ORNL this connectivity is expected to change in the summer of 2021. ORNL will then connect to the two ESnet6 routers located on-site at ORNL. These connections are expected to initially be 100Gbps but transition to 400Gbps with the upgrade of the ORNL border routers in FY21/22. The anticipated ESnet6 connectivity in the next 2-5 years is depicted in Figure 5.8.2.

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.

5.8.2.8 Cloud Services

MPEX does not anticipate the use of cloud resources at this time. Future considerations for the posting of public-facing data sets into cloud resources may be considered beyond the strategic planning timeframe.

5.8.2.9 Data-Related Resource Constraints

The "long-pulse" use case for MPEX, consisting of a 2 weeks continuous operation, has the potential to generate 1 PB of scientific experimental data in addition to data collected from sensors and cameras. The ability to store and transfer this volume of

data will require significant local support, along with any transient or remote resources that are used to support the use case.

5.8.2.10 Outstanding Issues

MPEX does not have any other outstanding issues to report at this time.

5.8.2.11 Case Study Contributors

MPEX at ORNL Representation

- Phil Ferguson[2], ORNL
- Juergen Rapp[3], ORNL
- Doug Curry[4], ORNL
- Cornwall Lau[5], ORNL
- David Green[6], ORNL

ESnet Site Coordinator Committee Representation

- Susan Hicks[7], ORNL

2 fergusonpd@ornl.gov
3 rappj@ornl.gov
4 curryde@ornl.gov
5 lauch@ornl.gov
6 greendl1@ornl.gov
7 hicksse@ornl.gov

5.9 MEC Experiment at SLAC

5.9.1 Discussion Summary

The MEC experiment, collocated with the LCLS XFEL is located at SLAC. The overall scientific goal of the instrument is to deliver ultrashort X-ray pulses in order to probe the characteristics of matter.

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

- The LCLS XFEL is an open-access user facility at SLAC that delivers ultrashort X-ray pulses able to probe the characteristics of matter. The MEC instrument at LCLS combines the XFEL with high-power, short-pulse lasers to produce and study HED plasmas.
    - The MEC-U proposes a major upgrade to MEC that would significantly increase the power and repetition rate of the high-intensity laser system to the petawatt level
    - The CD-1 of the MEC-U was completed Q4 FY2021 and the upgrade has an estimated duration of five years from CD-1 to CD-4. It is expected that the MEC-U data system will be complete and ready for beam time by June 2026.
    - MEC-U plans to use all existing LCLS-II cyberinfrastructure
- The MEC-U facility at SLAC LCLS-II will have a dedicated infrastructure for reading out detectors, and a shared infrastructure for data reduction, online monitoring, and fast feedback. It will use resources supplied by either SLAC, or remotely NERSC
    - The underlying LCLS-II system, which MEC will take full advantage of, is designed to handle data rates of 100 Gbps and produce 100 PB of data per year.
    - MEC dataset sizes are highly dependent on the physics case being studied. Based on estimated laser pulses and beam allocations, it is expected that datasets could be a minimum of 10 GB, to a maximum 100 TB with individual file sizes not exceeding 1 TB. The total number of files per experiment can range from a few hundred to 10,000 with a median of 3000.
    - MEC data transfer will utilize LCLS systems, with the main data transfer tools being bbcp and XRootD on-site data transfer hardware. Other tools are also supported on SLAC's DTNs: scp, sftp, rsync, and a Globus endpoint for data transfers.
- ESnet will continue to work with SLAC as LCLS-II is upgraded, so that experiments such as MEC have fast and predictable paths to ASCR HPC facilities such as NERSC.

5.9.2 MEC Experiment at SLAC Case Study

The LCLS XFEL is an open-access user facility at SLAC that delivers ultrashort X-ray pulses that are nine orders of magnitude brighter than any prior source, able to probe the characteristics of matter with unprecedented spatial and temporal precision. The MEC instrument at LCLS, funded by the DOE SC, Office of FES, combines the XFEL with high-power, short-pulse lasers to produce and study HED plasmas, and to develop the fundamental understanding of plasmas and matter in extreme environments. This has driven a remarkably rich array of high-profile scientific results with applications in fusion energy, isotope production, advanced materials, and medical and nuclear technology.

5.9.2.1 Background

The MEC-U Project at the focus of this case study proposes a major upgrade to MEC that would significantly increase the power and repetition rate of the high-intensity laser system to the petawatt level (PW, 1015 Watts) at 10 Hz, increase the energy of the shock-driver laser to the kilojoule level (kJ), and expand the capabilities of the MEC instrument to support groundbreaking experiments enabled by the combination of high-power lasers with the world's brightest X-ray source.

The particular strength of the MEC instrument is to combine the unique LCLS X-ray beam with high-power optical laser beams, and a suite of dedicated diagnostics tailored for this field of science (including an X-ray Thomson scattering spectrometer, an extreme ultraviolet (XUV) spectrometer, a Fourier domain interferometer, and a VISAR system). While the large vacuum target chamber makes the end station very versatile, it has been designed to service key scientific areas including HED physics, shock physics, and Warm Dense Matter physics.

As an open-access scientific instrument, the MEC experimental program consists of individual experimenter-led measurements. For experiments involving the LCLS X-ray laser access is governed by the MEC Program Advisory Committee (PAC), whereas for optical laser only experiments access is obtained through the LaserNetUS consortium. Beam time requests are typically five 12h shifts in duration.

The experimental detectors and diagnostics at MEC consist primarily of OPAL optical cameras, ePix10k cameras, an Andor Neo 5.5MP camera, a Princeton camera, and particle spectrometers, and produce known data types such as waveforms or megapixel 2D images. The data can be processed and analyzed in real time to enable important feedback and beam time decisions, e.g., laser and X-ray beam tuning, moving detectors/samples, and evaluating whether or not sufficient statistics have been accumulated. The data is then transferred to local storage where it can be accessed and transferred by the experimenters/collaborations to their institutions and computing resources for analysis.

5.9.2.2 Collaborators

MEC is a scientific instrument at SLAC. Access and use of the MEC instrument is proposal driven and granted on the basis of scientific merit. For experiments involving the LCLS X-ray laser access is governed by the MEC PAC, whereas for optical laser only experiments, access is obtained through the LaserNetUS (See Section 5.10) consortium. Beam time requests are typically five 12h shifts in duration.

Proposal PIs include researchers at US universities, US national laboratories, and international institutions. Data produced during the approved beam time for a given experiment is transferred from the facility to the analysis point, which is generally the home institution (be it a university or laboratory) of the experimenter team or collaboration.

User/Collaborator and Location	Do they store a primary or secondary copy of the data?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of dataset? (report in bytes, e.g. 125GB)	Frequency of data transfer or download? (e.g. ad-hoc, daily, weekly, monthly)	Is data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g. difficult tools, slow network)?
US University-based PIs	secondary	data transfer	10 TB (10 GB – 100 TB)	ad-hoc	No	N/A
US National lab based PIs	secondary	data transfer	10 TB (10 GB – 100 TB)	ad-hoc	No	N/A
International PIs	secondary	data transfer	10 TB (10 GB – 100 TB)	ad-hoc	No	N/A

Table 5.9.1 – MEC Data Relationships

5.9.2.3 Instruments and Facilities

This case study describes the MEC-U facility at LCLS. The MEC-U project involves a major upgrade to MEC that will significantly increase the power and repetition rate of the high-intensity laser system to the petawatt level at 10 Hz, increase the energy of the shock-drive laser to the kilojoule level (kJ), and expand the capabilities of the MEC instrument to support groundbreaking HED experiments enabled by the combination of high-power lasers with the world's brightest X-ray source. The MEC-U Project will provide more than an order of magnitude increase in power for both the short- and long-pulse lasers, and will involve the installation of a highly versatile pair of target chambers and associated state-of-the-art diagnostics in a dedicated new experimental cavern. It will leverage the increase in the maximum X-ray laser photon energy provided by LCLS-II, which will enable atomic-scale structure measurements, as well as shielding more of the plasma self-emission directed at detectors. A petawatt high-power laser and an upgraded long-pulse laser will be used to transform the understanding of plasma physics and extreme material science in the mission space of FES, providing precision tests of underlying theory and numerical models by exploiting the unique ability of coherent X-rays to probe the dynamic response of transient systems down to the atomic level.

The CD-1 of the MEC-U was completed in Q4 FY2021 and the upgrade has an estimated duration of five years from CD-1 to CD-4. It is expected that the MEC-U data system will be complete and ready for beam time by June 2026. Once MEC-U has completed construction, it will be operated by a local operations team with technical support provided by the SLAC staff. Given the anticipated schedule, this report relates to the 5+ year timescale (strategic planning).

The MEC-U facility will have a dedicated infrastructure for reading out detectors and a shared infrastructure for data reduction, online monitoring, and fast feedback (FFB). The data center is also a shared resource and may be supplied by either local resources at SLAC or by remote HPC, as illustrated in Figure 5.9.1. MEC-U will take full advantage of all the components used by LCLS-II instruments. The LCLS-II system

is designed to handle data rates of 100 Gbps and produce 100 PB of data per year. This will comfortably accommodate the expected needs for MEC-U. Below more detail is provided for each of the components of the infrastructure.

Data Acquisition (DAQ) and Data Reduction Pipeline (DRP). The experimental detectors and diagnostics at MEC-U, such as ePix cameras, Princeton camera, and OPAL optical cameras produce known data types such as waveforms or megapixel 2D images. For MEC-U, this data must be read out at 1 to 10 Hz. The estimated aggregated data rate is 2 Gbps. The 10 Hz repetition rate of the laser does not present a unique challenge in this respect, as these requirements are well within the capabilities of the LCLS-II data system: readout rates from 1 Hz to 1 MHz and recording of aggregated data rates > 200 Gbps. For the relatively modest data rates expected at MEC-U the DRP is not anticipated to be needed, but can be enabled if necessary.

Prompt analysis of the data is critical for MEC-U experiments, because such information is required for important decisions, e.g., laser and X-ray beam tuning, moving detectors/samples, and evaluating whether or not sufficient statistics have been accumulated. The MEC-U facility will provide users with a rapid, flexible, and easy-to-use real-time signal processing and data analysis tools. The Analysis Monitoring Interface (AMI) provides real-time (< 1 s) monitoring of the acquired data and data quality. It can also be used for performing analysis, such as averaging, filtering, and other generic manipulations of data including region of interest selection, masking, projections, integration, contrast calculation, and hit finding. The real-time data analysis rate is comparable to the maximum expected data production rate of 10 Hz. In experiments running at maximum data rate, it is possible that only a fraction of the events can be analyzed in real time.

Data management. Dataset sizes are highly dependent on the physics case being studied. Based on the statistics of current MEC experiments, taking into account the increase in laser repetition rate to 10 Hz, and assuming an average 5 day beam time allocation, it is expected that an experiment's aggregate data size to range from a minimum of 10 GB to a maximum 100 TB. Individual file sizes are < 1 TB. The total number of files per experiment can range from a few hundred to 10,000 with a median of ~3000. A metadata manager will store information about experimental runs, logbook, run parameters, etc. The data management system will provide automatic data transfers between different storage resources and remote sites (e.g. NERSC and SDF@SLAC), archiving of raw data and storage space management, and processing of the experimental run data at LCLS or remote (HPC) sites.

Storage and Compute Resources:

- Fast Feedback (FFB). An FFB layer, which is a standard HPC system, offers dedicated processing resources to the running experiment in order to provide quasi real time (< 1 min) feedback about the quality of the acquired data. FFB provides the first persistent storage layer in the data flow chain and also offers the first processing layer where the users can access the full data set. The FFB layer processing nodes are physically located close to the FFB layer storage nodes to allow high-throughput Infiniband connections. NVME-SSD technology is used for implementation of the FFB storage layer, allowing for very high I/O rates. As opposed to spindle based storage, these technologies excel at handling the high concurrency generated by

the DAQ writers, the data movers, and the FFB analysis. The current performance of the FFB layer is 50 Gbps write and 120 Gbps read. The system is expected to handle I/) rates of 100 Gbps writing and 200-300 Gbps reading needed for LCLS-II-HE. A compute cluster is connected to the storage, which currently comprises about 5000 cores. The current size of the FFB storage layer is 500 TB, but will increase to at least 1 PB. The data is kept on the FFB only for the duration of the experiment. For experiments with very high data rates the lifetime of the data in the FFB storage can be as short as a few hours;

- Off-line Analysis Cluster. A compute cluster of about 3200 cores is shared by all experiments is used for off-line analysis. It also includes a limited number of GPU processing. Currently the I/O rates are ~10 Gbps but will increase to above 100 Gbps. The cluster provides storage for raw data and a user writable space that is used for data processing. The storage is based on hard disk drives and its current size is 6 PB. Storage capacity will increase to ~100 PB for LCLS-II-HE and the computational power will increase accordingly. The data stays on disk for at least four months. Data that has been purged from disk can be restored by a user from the tape archive. The data is kept in the archive for 5-10 years.
- Networking. The FFB storage and compute systems are connected with 100 Gbps Inifiniband networks and the analysis cluster has 40 Gbps Infiniband. The networking between the experimental halls currently include multiple 10 Gbps Ethernet connections, but will be upgraded to 100 Gbps. The LCLS systems are connected to the SLAC Network using 2x100 Gbps connections and the Stanford Linear Collider Network is also connected with 2x100 Gbps to ESnet. All networks will be upgraded and it is expected that by 2027 when LCLS-II-HE is operational, the network will require 1 Tbps.
- Data transfer. The data management system handles the data transfers with the LCLS system and certain remote HPC system (e.g. NERSC). The main data transfer tools are bbcp and XRootD. The experimental users are responsible for transferring the data to their home institutions. They are provided DTNs, where they can login and use the tools of their choice: scp, sftp, rsync, bbcp. A Globus endpoint is also provided for data transfers.

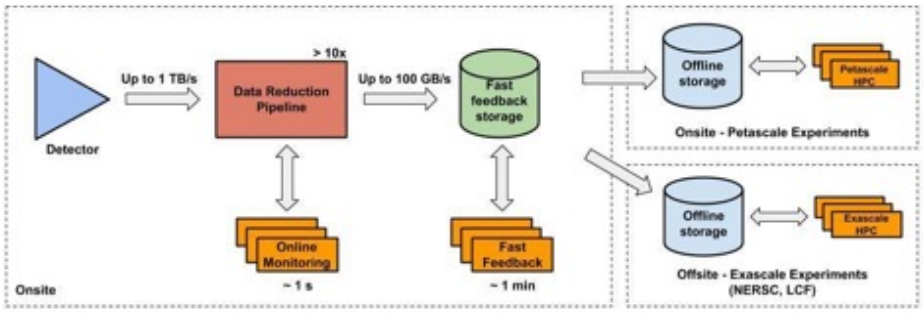


Figure 5.9.1: MEC-U Data System Main Components and Data Flow.

5.9.2.4 Process of Science

There are two primary workflows for MEC experiments. The first is the real-time signal processing and data analysis workflow and is common to all experiments. The second workflow is the data analysis by the experimenters and their group/collaboration after the beam time. A brief description of both of these workflows is given below.

During the experimental beam team, data must be accessed and analyzed quickly to allow users to iterate their experiments and extract the most value from scarce beam time. MEC experimenters will take advantage of the AMI provided by the LCLS Data Acquisition (DAQ) system to perform basic real-time data analysis, such as displaying detector images at 1 Hz, making projections of the detector image on-the-fly onto specified axis, making the trending plots of some diode readout, background subtracting, and simple mathematical operations on the data. Users may also integrate their own code to perform other sophisticated or device-specific processing by writing Python code for AMI or to run in the Photon Science ANAlysis 2 (psana2) framework at LCLS.

The second workflow involves the analysis of data by the experimenter(s) after the beam time. The processing of MEC data is not considered computationally (*central processing unit* [CPU] or network) intensive and is usually done at the institutions of the experimenters. This typically involves transferring the data to the experimenter institution and the analysis can include different experiment dependent steps. The analysis of the experimental data and the scientific discovery process will often involve integration with numerical simulations of the experiment or some of its relevant processes. Depending on the type of experiment, these simulations can be very computationally demanding (as for example particle-in-cell simulations of the interaction of intense lasers with solid targets or molecular dynamics 5simulations of shock compressed materials) and will require access to large-scale HPC resources, such as NERSC or ALCF. Access to these resources is independent of the experimental beam time.

5.9.2.5 Remote Science Activities

Remote HPC resources are available for all MEC experiments as a shared resource. If needed, these HPC resources can support near real-time analysis (< 10 min) of data bursts and fast turnaround on large data sets exceeding 10 Gbps, such as those anticipated in LCLS-II-HE. It is not anticipated that data rates at MEC-U will necessitate use of these large-scale resources for data analysis. Given the modest size of the data produced at MEC, off-line analysis is typically done at the institutions of the experimenters. However, if use of HPC resources becomes a need for MEC-U, these will be available as MEC-U is part of the LCLS data management system.

5.9.2.6 Software Infrastructure

MEC users require an integrated combination of data processing and scientific interpretation. During the experimental beam team, this must be carried out quickly to allow users to iterate their experiments and extract the most value from scarce beam time (workflow 1). After the experiment is conducted, the second workflow involves off-line analysis of the data by the experimenters.

The AMI provides real time (~ 1 s) analysis of the acquired data. Users primarily

interact with AMI through a GUI. The GUI allows the user to display and analyze information instantaneously through a set of simple operations that can be cascaded to achieve a variety of monitoring measures. The GUI can be used to perform such standard tasks as displaying detector images and waveforms and displaying data as histograms, strip charts, scatter plots. It can also be used for performing averaging, filtering, and other generic manipulations of data including region of interest selection, masking, projections, integration, contrast calculation, and hit finding. AMI can be used to view raw or corrected detector images and perform such tasks as background subtraction, detector corrections, and event filtering. AMI supports single event waveform plots and image projections that can be averaged, subtracted, and filtered, and it includes an algorithm for simple edge finding using a constant fraction discriminator. Displays of waveforms and images can be manipulated by adding cursors and doing cursor math or waveform shape matching. Users may also integrate their own code to perform other sophisticated or device-specific processing, either by building a C++ module plug-in for AMI, or writing Python code to run in the psana2 framework.

Psana2 is the main programmatic analysis code supported by LCLS-II. It is derived from the psana framework developed for LCLS-I and facilitates these essential features:

- Moving data from persistent storage to memory
- Handling – transparently to users – the perfectly parallel nature of LCLS data (for most) LCLS experiments, each event can be processed independently)
- Handling detector calibrations
- Invoking science-specific algorithms

Psana2 can analyze data both off-line and online: in addition to the ability to analyze persistent data files produced by the DAQ system with latency of a few minutes, it can analyze in-memory real-time DAQ monitoring data with latency of ~ 1 second. Since experiments and experimenters generally change multiple times per week, necessitating that algorithms and configuration parameters also change at that rate, software development must be as easy as possible. The LCLS-II analysis framework is capable of scaling to HPC should MEC-U develop a need for this feature in the future.

In order to ensure scalability to high rates and interoperability of code between online and off-line in LCLS-II, the psana analysis framework has been modified for handling I/O, parallelism, and calibration. This improves the robustness and efficiency of the online monitoring system as well because of the shared code base. To improve the flexibility of the online monitoring, and to minimize the side effects of online monitoring on data taking, AMI processes are able to connect and disconnect to/from the DAQ on-the-fly, that is without the need to stop and restart data collection.

In experiments using low repetition-rate laser systems (e.g., the kJ optical laser with a 20 minutes cooling time between shots), users often use their own codes or third party applications (like radial integration for Debye-Sherrer rings) to extract additional information critical to inform the next shots.

For off-line data analysis (workflow2), application specific software developed by the user community can be used to analyze processed data. This is followed by a number of experiment dependent steps, which may include comparison with simulations. The

processing of MEC/MEC-U data is not considered computationally (CPU or network) intensive as is usually done at the institutions of the experimenters.

5.9.2.7 Network and Data Architecture

The Campus and WAN networking that supports MEC is built on top of the LCLS infrastructure. Figure 5.9.2 describes this in detail, along with providing some of the upgrade plans for future years.

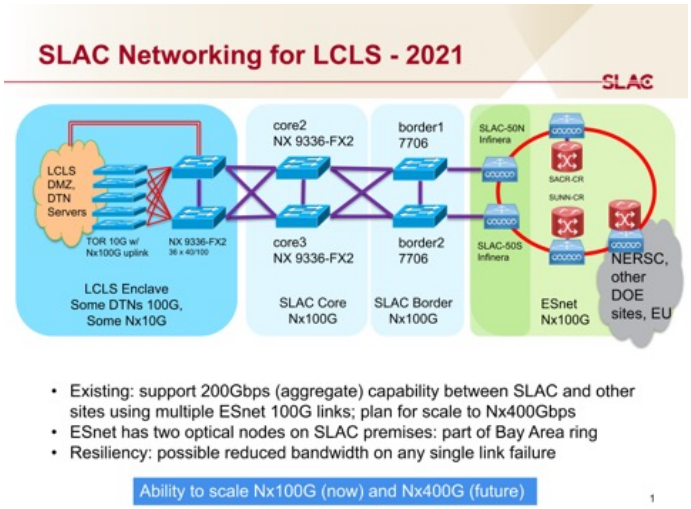


Figure 5.9.2 - SLAC Networking to support LCLS

The local network architecture for MEC is centered around the data acquisition, processing, and storage as described in section 5.9.2.3. The MEC-U will take advantage of LCLS-II's powerful data management system, which has the ability to handle both the automatic workflows of data through various storage layers — such as long-term data archiving — and user requests through a web portal, such as restoring data from tape. The data management architecture, which allows transparent integration of both local and, if needed, external HPC facilities (such as NERSC), is illustrated in Figure 5.9.2.

The MEC-U instrument will have a dedicated database for the experiments to store configuration parameters for the detectors, DAQ, and detector calibrations needed for data processing. There is a subsystem responsible for capturing all experiment metadata, including but not limited to experiment configuration, runs, run parameters, experimenter comments, experiment questionnaires and user-defined data. These metadata are programmatically accessible through Web Applications and Services.

The MEC-U data system will require readout of detectors from < 1 up to 10 Hz, with an estimated aggregated data rate of 2 Gbps. As a trigger, most devices will accept LCLS timing fibers or transistor-transistor logic pulses and will be read out via standard protocols. These requirements are well within the capabilities of the LCLS-II data system: readout rates from 1 Hz to 1 MHz and recording of aggregated data rates > 200 Gbps.

Data acquired at MEC is sent to an NVRAM-based data cache where the data

can automatically be viewed by online monitoring nodes and are made available on-demand to users for FFB analysis. Calibration data may be accessed as needed to support data analysis on the FFB storage and computing layer. The file movement and catalog subsystem is responsible for the data files produced by the experiments. A file manager keeps a catalog of all data files generated by an experiment, manages the space usage on local storage resources, and tracks the location of the files on different storage resources. The file manager automatically transfers files to and from different storage resources via DTNs. The file manager may utilize local network resources to automatically transfer data to SLAC off-line computer resources or it may utilize ESnet to transfer data to HPC computing, storage, and tape resources.

Because of the relatively low throughput of MEC-U compared to other LCLS-II experiments, the MEC-U is expected to require only the dedicated local storage and analysis resources to operate. However, since the local and remote infrastructure are resources shared by the entire facility, MEC-U will be able to make use of them if needed in the future.

The data from the experiments will be available on disk for the first 4 months after data collection. After that, it can be restored from the tape archive. The data is kept in the archive for 5-10 years.

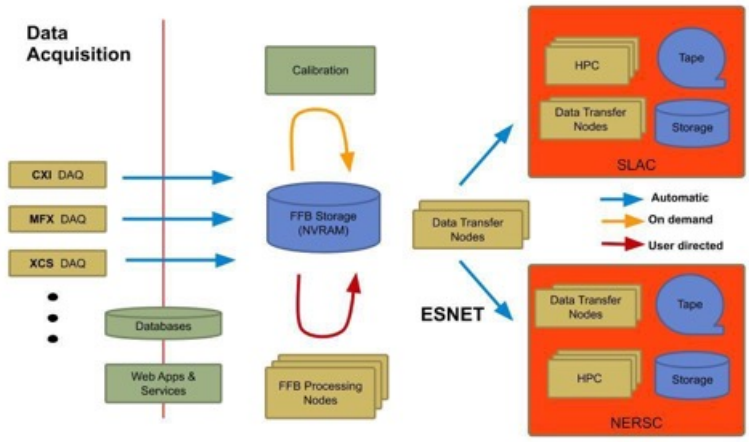


Figure 5.9.3: MEC and LCLS-II Data Management Architecture

5.9.2.8 Cloud Services

MEC currently does not use cloud services for real-time analysis of the data produced, and do not plan to use cloud services for MEC-U. Data analysis carried out by experimenters at home institutions may or may not use cloud resources. Given that the computational needs of this analysis are currently modest, it is expected that the demand for use of cloud services in the data analysis to be limited.

5.9.2.9 Data-Related Resource Constraints

The MEC-U project does not anticipate or foresee future network or data-related constraints to meet the project scientific goals.

5.9.2.10 Outstanding Issues

The MEC-U project does not have any other outstanding issues to report at this time.

5.9.2.11 Case Study Contributors

MEC Experiment at SLAC Representation

- Frederico Fiuza[1], SLAC
- Wilko Kroeger[2], SLAC
- Jana Thayer[3], SLAC
- Eric Galtier[4], SLAC

ESnet Site Coordinator Committee Representation

- Mark Foster[5], SLAC

1 fiuza@slac.stanford.edu
2 wilko@slac.stanford.edu
3 jana@slac.stanford.edu
4 egaltier@slac.stanford.edu
5 mark.foster@stanford.edu

5.10 LaserNetUS Program

5.10.1 Discussion Summary

LaserNetUS is a program established by the department's Office of FES to help restore the US's once-dominant position in high-intensity laser research.

LaserNetUS will provide US scientists increased access to the unique high-intensity laser facilities at ten institutions: University of Texas at Austin, The Ohio State University, Colorado State University, The University of Michigan, University of Nebraska-Lincoln, University of Rochester, SLAC, LBNL, Lawrence Livermore National Laboratory, and Université du Québec.

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

- The LaserNetUS VO is loosely coupled, and sites vary in terms of data volume produced, and mechanisms to collect, store, and disseminate data to users:
  - Laser capability dictates factors such as power, pulse length, and number of shots that can be run during an experimental period.
  - Typical shot output is several MB to as much as a GB. An entire experimental run, consisting of 10s to 100s of shots over the course of several days (which produce both scientific data files, as well as camera output), may approach 100s of GBs.
  - Managing the data is at the discretion of each site. Typical approaches could be requiring the use of portable media, integration to enterprise cloud storage, or the ability to transfer data from network-enabled portal systems that are on premises.
  - Site users are responsible for data analysis and data reduction, which they do at their home institutions. This includes simulations, which are used to predict the outcome of experiments or the experimental data is used to guide and benchmark the simulations.
- The LaserNetUS does not maintain a suggested set of policies and procedures to address data management and mobility within, or between, facilities.
- ESnet can assist FES facilities adopt hardware and software approaches that are native to HPC facilities to accelerate simulation and theoretical FES workflows that require data mobility. These solutions can be to install and adopt known tools (e.g. Globus, MRDP), or potentially offer services operated by ESnet to foster data mobility improvements.
- LaserNetUS provides time to users to run laser based experiments utilizing a collection of high-power, short-pulse lasers that are operated by 10 participating institutions and facilities. These laser systems are often combined with long-pulse "driver" lasers to achieve high density and pressure or with other beams:

- The actual amount of data involved during a run is small (a few GB is common).
- Each facility has its own research program that is, to varying degrees, separate from LaserNetUS and data associated with the facilities' local programs.
- There is not a standard approach to handle data mobility, and often facilities rely on non-technical approaches (e.g., portable media) to transfer research data.

- DOE programs that span facilities and communities (e.g., LaserNetUS) do not include access to generalized pools of computational resources that can be utilized by participants. While it is possible for participants to pursue these resources independently from DOE HPC facilities, it is a secondary step that must be managed independently. Having access to computational resources, and potentially more efficient data transfer and analysis tools, would benefit participants and lead to more efficient use of resources over time.
- DOE programs that span facilities and communities (e.g., LaserNetUS) do not typically require a data architecture review to facilitate sharing of experimental results; solutions in this space can vary between facilities. While organic approaches have scaled to date, the lack of a cohesive and shared understanding of best practices as data volumes increase will begin to harm productivity. Having access to community recommended approaches, and potentially more efficient data transfer hardware and software, would benefit participants and lead to more efficient use of resources over time.

5.10.2 LaserNetUS Program Case Study

LaserNetUS provides time to users to run laser based experiments. The actual amount of data involved during a run is currently relatively small (a few GB is common), and there is little issue with storage or transmission of this data. Each facility has its own research program that is, to varying degrees, separate from LaserNetUS and data associated with the facilities' local programs are not considered here. The experimental data volume is likely to grow in the future, however, and this is discussed.

Although users might work collaboratively with LaserNetUS personnel, publication of results is generally their responsibility so far as evaluation of research progress is concerned. Publication usually requires computer simulations using HPC and each user group does this at their own institutions, which are broadly dispersed. The data associated with this effort dwarfs that of the experiments and, although not directly part of the data handling needs of LaserNetUS, has been considered here in case this is of interest.

5.10.2.1 Background

LaserNetUS is a new network of the nation's (and Canada's) highest power lasers, established in autumn, 2018. A formal Vision and Mission Statement reads, in part, "The Mission of LaserNetUS is to re-establish US scientific leadership in laser-driven HED and High Field optical science by advancing the frontiers of laser-science research, providing students and scientists with broad access to unique facilities and

enabling technologies, and by fostering collaboration among researchers and networks from around the world." The statement goes on to list components of this mission: Supporting end users, Expanding the user base, Fostering closer ties between the research community and industry, and training and education. Science goals are ultimately defined by the end users, but this case study will address these components by allocating scarce resources to best enable the network to perform state-of-the-art research using high-power lasers. LaserNetUS is entirely supported by the DOE SC, FES.

The primary lasers of LaserNetUS are high-power, short-pulse lasers whose highest intensities reach 1022 W/cm2. The brightness of these lasers permits them to drive matter to extreme temperatures and pressures and the short-pulse duration (as short as 20 x 10-15 s) permits excitation with greatly reduced expansion of the target while resolving extraordinarily fast processes. These laser systems are often combined with long-pulse "driver" lasers to achieve high density and pressure or with other beams (optical, laser derived particle, or laser derived x-ray and gamma-ray) to manipulate and probe a physical system. The work done by LaserNetUS is part of the subfield of plasma physics and often referred to as high energy density science (HEDS) or physics (HEDP). These facilities permit study of relativistic matter where the electronic system of a target is relativistic, but the target still acts as a medium with collective modes. This is a fundamentally interesting state of matter that is also found in systems of astrophysical interest, such as near a black hole. The lasers of LaserNetUS can also create warm dense matter, matter that is intermediate between the condensed matter and plasma phases, and that can be found in the cores of planets or during inertial confinement fusion. Finally, the end users study the generation of secondary radiation which includes beams of energetic electrons and ions, sprays of neutrons, anti-matter, x-rays and gamma rays. This radiation can rival traditional means of generating such radiation and can be used as powerful sources for experiments not possible with lasers alone. Many potential applications are being developed, both for the scientific community (e.g., electron accelerators, ultrafast x-ray tomography) and for society (e.g., neutron sources).

An experiment at LaserNetUS involves some choice of optical system that explores one of these modalities. There may be only one facility that can perform a given experiment, or several, but usually one or two are most appropriate. The lasers of LaserNetUS vary in pulse energy, duration and shape, spatial mode, wavelength, pulse contrast (sharpness of pulse turn-on), and repetition rate. Experiments measure some combination of the reflected and transmitted light and emission of energetic particles and electromagnetic radiation. Up to ~6 primary diagnostics may be fielded at the same time (usually less) as well as a large number of diagnostics measuring the performance of the laser itself. Experiments can be performed "shot-on-demand" where each laser shot occurs after significant setup or the experiment can be performed "rep-rated" with data collection rates being clock driven. Most diagnostics yield their results in the form of an image (e.g., particle spectrometers, x-ray intensity and spatial profile), but will often include oscilloscope traces and simple numerical readouts (e.g., energy). Text information describing the shot is common. This data is initially held by the facility in diverse ways (e.g., on the recording instrument, local laptop or desktop, or uploaded to a server) and is generally in the range of Gbpshot to several GB for an entire experimental run consisting of multiple shots. Shot counts range from a

10's of shots to thousands, although the latter are still less common. The results from each shot may be saved or just the combined result of many, say to produce an x-ray image with good signal to noise. The data belongs to the users. They often leave the facility with their data, perhaps in an external drive, or they access it using the Cloud after the facility has uploaded it. These are the two primary mechanisms, but there is variation. For example, the Laboratory for Laser Energetics (LLE) (Omega EP) provides a sophisticated web-based interface. Although the facilities maintain the data for an extended time (currently not publicly specified), the users are responsible for the subsequent life history of their data. They are generally expected to publish and to conform to publication standards.

Publication requires extensive analysis. Direct analysis of the experimental data (e.g., background subtraction, calibration, extraction of reduced measures) is usually performed using desktop/laptop computers at the facility and afterward. It is very common for advanced analysis to involve extensive computer modeling using HPC, most commonly using particle-in-cell (PIC) codes or hydrodynamic (hydro) codes. These involve dozens or, more commonly, hundreds to thousands of cores running for hours to days. These simulations generate large amounts of data up to TBs per simulation. This is the province of the users and is not a responsibility of the facilities nor does this generally use facility resources. However, since it is a crucial step in the eventual conclusion of a research effort, it is described in Section 5.10.2.10, as well.

5.10.2.2 Collaborators

5.10.2.2.1 Facilities List

PI	Institution	System*	Email
Bob Cauble, Félicie Albert	LLNL Jupiter Laser Facility (JLF)	Titan/Comet	cauble1@llnl.gov
Todd Ditmire	University of Texas Austin	Texas Petawatt (TPW)	tditmire@physics.utexas.edu
Gilliss Dyer	SLAC	MEC	gilliss@slac.stanford.edu
Karl Krushelnick	University of Michigan	Hercules/Zeus	kmkr@umich.edu
François Legaré, Jean-Claude Kieffer	INRS Advanced Laser Light Source (ALLS)		francois.legare@inrs.ca, legare@emt.inrs.ca
Jorge Rocca	Colorado State University (CSU)	ALEPH	jorgerocca9@gmail.com
Thomas Schenkel	LBNL Berkeley Lab Laser Accelerator Center (BELLA)	BELLA	t_schenkel@lbl.gov
Douglass Schumacher	Ohio State University (OSU)	Scarlet	onald her.60@osu.edu
Donald Umstadter	University of Nebraska Lincoln	Diocles	onald.umstadter@unl.edu
Mingsheng Wei	University of Rochester Laboratory of Laser Energetics (LLE)	Omega EP	mingsheng@lle.rochester.edu

Table 5.10.1 – LaserNetUS Facilities

The 10 facilities of LaserNetUS are listed in Table 5.10.1. The facilities should be considered as collaborators since, together, they form the network. However, although they work together in many ways, they generally do not collaborate on a given user run. Thus, the highly collaborative aspect of LaserNetUS as a network does not have a critical effect on workflow as defined in the provided documents. Users are not required to be collaborators of the facilities. Although many choose to include selected facility personnel as full collaborators on their team, some do not. Either way, this consideration also does not have a critical effect on workflow. Accordingly, the

remaining discussions in this document are about Users and not collaborators.

As a network, LaserNetUS does not generally keep distinct records on the amount of data generated for each run, when it is transferred, or how the data was transferred to the users. In principle this information can be recovered, but it would be a significant effort. The tables provided below are used to provide an assessment of the geographical range.

To date, LaserNetUS has had three calls for proposals. Users get laser time by responding to the approximately-annual call for proposals. Their proposals are evaluated for scientific merit by a proposal review panel (PRP) that is convened by DOE and is independent of the facilities. The facilities are consulted by the PRP to assess technical feasibility only. The first two calls for proposals were held in 2019, Cycle 1 and Cycle 2. The 3rd call, Cycle 3, was delayed due the pandemic but has now been mostly concluded and the results announced.

Future collaborations (users): It is expected to have a call for proposals every 12 months, not necessarily aligned with the calendar or fiscal year for the time being due to the disruption caused by the pandemic. LaserNetUS is currently funded through 2023 so two more calls for proposals are assured. Given the results of Cycle 3, >60 additional runs are likely to be supported by these additional runs in total.

5.10.2.2.2 Experimental Run Summary

Cycles 1 and 2:

- 49 user experiments total awarded
- Most experiments have been completed (a few remain that were delayed due to the pandemic).
- >200 users participated in the experiments, including people who did not participate in on-site running of the experiment but participated in the technical development, theoretical work, or numerical computational studies.

Cycle 3

- >33 user experiments awarded (some final decisions still pending).
- The cycle is just getting underway with almost all experiments yet to be performed.
- The total number of participants across all three cycles is now 399, and this figure will likely grow.
- The tables below have been populated with the Cycle 3 PIs awarded run time.

5.10.2.2.3 Data Specifics

The following is a discussion of size of dataset, data access method, and frequency of data transfer by facility. The following are typical answers, sometimes given as a range, to these questions as determined by discussions with the facilities. The primary data access methods are:

- Storage on user provided media (e.g., external disk drive)
- The Cloud, with Google Drive and Box being common.

Some data is typically transferred every day that data is collected. A significant fraction of a run may involve setup and optimization with no or little data being collected.

Institution	Run duration (weeks) (this includes setup time and exploratory or optimizaton work	Typical dataset per run	Number of runs for Cycle 3 (current cycle)
LLNL Jupiter Laser Facility (JLF)	4	0.1 GB	2
University of Texas Austin	4	0.1 – 1 GB	3
University of Michigan	3-4	~20 GB	0
INRS Advanced Laser Light Source (ALLS)	2-3	~20 GB	5
Colorado State University	4	1 – 120 GB	7
LBNL Berkeley Lab Laser Accelerator Center	2-4	1-50 GB	4
Ohio State University (OSU)	3-4	~20 GB	4
University of Nebraska Lincoln	4	~20 GB	3
University of Rochester Laboratory of Laser Energetics (LLE)	1-2 days	1.5 – 3 GB	6

User/Collaborator and Location for Cycle 3 Run PI/Spokesperson and home institution listed. Each PI is the head of a participating group.	Do they store a primary or secondary copy of the data?	Facility Insitution for Cycle 3 (U. Michigan is currently down for upgrades.)
Alexey Arefiev, UCSD	Primary	OSU
Wendell Hill, U. Maryland	Primary	OSU
Louise Willingale, U. Michigan	Primary	OSU
Mihail Cernaianu, ELI-NP, Romania	Primary	OSU
Kazuki Matsuo, UCSD	Primary	LLE
Gaia Righi, UCSD	Primary	LLE
Matthew Edwards, LLNL	Primary	LLE
Mario Manuel, General Atomics	Primary	LLE
Shuzhong Zhang, Princeton	Primary	LLE
Andreas Schmitt-Sody, AFRL	Primary	JLF
Christopher McGuffey, General Atomics	Primary	JLF
Dean Rusby, LLNL	Primary	U. Texas
P.M. King, LLNL	Primary	U. Texas
Edison Liang, Rice University	Primary	U. Texas
Antoine Snijders, LBNL	Primary	LBNL
Razzy Simpson, MIT	Primary	LBNL
Christopher Thornton, STFC, UK	Primary	LBNL
E. Grace, Georgia Tech	Primary	LBNL
Byung-Kook (Brian) Ham, GIFS	Primary	INRS
Marianna Barberio, ALLS, Canada	Primary	INRS
Amina Hussein, U. Alberta, Canada	Primary	INRS
Yong Ma, U. Michigan	Primary	INRS
	Primary	INRS
Sophia Malko, PPPL	Primary	Col. State
Iain Wilkinson, HZB, Germany	Primary	Col. State
Bedros Afeyan, Polymath Research	Primary	Col. State
Alexander Thomas, U. Michigan	Primary	Col. State

User/Collaborator and Location for Cycle 3 Run PI/Spokesperson and home institution listed. Each PI is the head of a participating group.	Do they store a primary or secondary copy of the data?	Facility Insitution for Cycle 3 (U. Michigan is currently down for upgrades.)
M. Evans, U. Rochester	Primary	Col. State
Derek Alexander Mariscal, LLNL	Primary	Col. State
H. M. Milchberg, U. Maryland	Primary	Col. State
Nilson Vieira	Primary	U. Nebraska
Matthias Fuchs, U. Nebraska-Lincoln	Primary	U. Nebraska
Meriame Berboucha, Stanford	Primary	U. Nebraska

Table 5.10.3 – LaserNetUS Facility Contacts

5.10.2.3 Instruments and Facilities

The tables 5.10.4 and 5.10.5 cover the present circumstances. The laser repetition rate determines the maximum rate of data collection. Multiple lasers or configurations are separated by a semicolon. There are many lasers and laser modes of operation in the network; not all are listed, but the list is representative.

Associated with the lasers are a large number of laser diagnostics, as opposed to experimental diagnostics, some of which produce data that becomes part of the User's dataset. However, the contribution to the overall size of the data set is typically small, MBs.

Institution	Lasers (abbreviated description, see https://www.lasernetus.org/ for full description)	Laser Repetition Rate
LLNL Jupiter Laser Facility (JLF)	0.5 ps, 1053 nm, 10 J	15 per hour
LLNL Jupiter Laser Facility (JLF)	0.7 ps, 1053 nm, 130 J	2 per hour
University of Texas Austin	140 fs, 1057 nm, 120 J	1 per hour
University of Michigan	815 nm, 30 fs, 15 J	1/min
University of Michigan	30 fs, 810 nm, 3 J	5 Hz burst
INRS Advanced Laser Light Source (ALLS)	22 fs, 800 nm, 4 J; 40 fs, 800 nm, 45 mJ; 50 fs, 1200-2100 nm, 5 mJ	2.5 Hz
Colorado State University	45 fs, 400 nm, 10 J; 800 nm, 26 J	3.3 Hz
LBNL Berkeley Lab Laser Accelerator Center	30 fs, 815, nm 40 J	1 Hz
LBNL Berkeley Lab Laser Accelerator Center	50 TW	5 Hz
Ohio State University (OSU)	30 fs, 800 nm, 10 J	1/min
Ohio State University (OSU)	30 fs, 800 nm, 0.3 J	10 Hz.
University of Nebraska Lincoln	30 fs, 805 nm, 20 J	0.1 Hz
University of Nebraska Lincoln	30 fs, 810 nm, 0.3 J	10 Hz
University of Rochester Laboratory of Laser Energetics (LLE)	0.1 ns, 351 nm, 100 J (4 beams); 0.7 ps, 1054 nm, (300 J, 500 J) (beam 1, beam 2)	1/90 min

Table 5.10.4 – LaserNetUS Laser Types & Capabilities

Table 5.10.5 is a highly abbreviated description of the experimental diagnostics. Each facility has many diagnostics whose form and configuration might change significantly (or simply disappear or reappear) over the course of a Cycle. Primary focus is on diagnostics that produce images and thus results in the most data. Multiple diagnostics may be fielded.

Institution	Experimental Diagnostics (highly abbreviated description, see https://www.lasernetus.org/ for fuller description)
LLNL Jupiter Laser Facility (JLF)	Particle spectrometers, x-ray spectrometer, optical mode.
University of Texas Austin	Particle spectrometers, optical mode.
University of Michigan	Magnetic spectrometer, x-ray imaging
INRS Advanced Laser Light Source (ALLS)	X-ray diagnostics, pump-probe shadowgraphy, electron spectrum
Colorado State University	X-ray spectrometers, Thomson parabola spectrometer, Si photodiode array.
LBNL Berkeley Lab Laser Accelerator Center	Magnetic electron spectrometer, Thomson parabola spectrometer.
Ohio State University (OSU)	Particle spectrometers, optical imaging, pump-probe shadowgraphy, x-ray spectrometer.
	Electron spectrometers and imaging, x-ray spectrometers
University of Rochester Laboratory of Laser Energetics (LLE)	Vast collection of diagnostics, most with image based outputs.

Table 5.10.5 – LaserNetUS Diagnostics

The laser system controls are generally not available to the users but are run by trained staff. The users usually have access to the experimental chamber(s) and diagnostics, perhaps with staff guidance. All diagnostic handling at LLE is done by staff, however.

Most facilities have particle spectrometers, radiochromic film (RCF) stacks, x-ray imagers, and ways of profiling the laser beam before and after its interaction with the targets; all of these diagnostics are image based. (Diagnostics for the lasers themselves are not included here.)

All of the laser facilities regularly upgrade their laser systems and diagnostics in ways that do not significantly affect data handling as addressed in this document. All of the laser facilities are considering long range upgrades that could significantly affect operation or data handling, but these upgrades are generally not yet green-lighted or are being implemented in stages.

Some future upgrades in progress:

- JLF is finishing an upgrade which will be completed this year. Laser operation and control will be significantly improved but data production may not be significantly changed.
- The University of Michigan is undergoing an upgrade that will be complete in ~3-4 years which will substantially change the laser capability and data production rate, however, the role of this system in LaserNetUS is not yet established.
- INRS is switching to a deeper charge-coupled device (CCD) camera and a 4x increase in data is expected in the 1 year time frame. An additional increase of a factor of 2-3x is hoped for on the 2-5 year time frame.
- Colorado State University
    - Currently each shot produces 5 MB of data primarily in 1 image file.
    - 0-2 years: Planning is underway to be able to regularly collect data at 1 Hz or higher for some experiments. This could yield 0.6 GB/minute. Total resulting dataset would depend on run duration, which might be less than

the current 4 weeks since runs could be completed more quickly. (Note that much of a run is spent building the experiment and optimizing it.)

- 2-5 years: Achieve 5000-10,000 shots per day for solid density targets.

Most facilities use diagnostics that store the data that is collected (e.g., a camera or CCD array used as a particle detector or reflected laser diagnostic with storage on a controller or controlling laptop). This data is usually backed up in various ways: department server or User provided drive. The data is provided to the User as discussed previously: Users usually either bring their own external storage and/or a Cloud-based mechanism is used to transfer the data to the users. LLE has its own sophisticated web-based data transfer interface. Data collected at LLE is archived permanently using local servers.

Due to the large number of facilities, the large number of diagnostics at each facility, and the widely varying configuration for every run, a complete specification is difficult. However, there are commonalities. A varying number of small files are produced, for example, text files and Excel spreadsheets listing various run parameters and calibration parameters. These are typically small, 1 MB or less. The largest files are the image files. Examples of typical runs are given below. The following sections break out some of this detail.

5.10.2.3.1 INRS

A 10 day experiment studying laser wakefield (electron acceleration) will result in 22 GB of data. The data is collected using CCD detectors, perhaps integrating over 50 laser shots for each image. The data is contained in typically 200 - 360 image files. The data resides on the data collection computer with backup to a local desktop computer. The data is provided to the user using a local ftp type service. No commercial Cloud use.

5.10.2.3.2 Colorado State University

The number of diagnostics varies significantly for User runs with the entire dataset varying from a few GB to > 100 GB. Most of the data is in the form of image files from RCF stacks, image plates, CCD detectors and cameras. As an example of the high side of the range: one experiment ran at ¼ Hz using a gas jet target. 4-5 days of running resulted in a 125 GB dataset, mostly consisting of 4-5 MB image files. Over 5000 laser shots were collected per day. Data was transferred to the user by using an external drive provided by the user and Cloud services. A transfer was done each day.

5.10.2.3.3 University of Nebraska

Cameras typically produce most of the data and data files. Each camera has 1024x1640 pixels with 12 bits of data per pixel = 2.4 MB per camera per shot contained in a single image file. With 5 cameras and ~200 shots this yields 2.3 GB per day in 1000 files. This in turn corresponds to ~20 GB per typical run. Sometimes 16 bit pixel cameras are used, resulting in bigger files.

5.10.2.3.4 LLE

Each shot produces about 150 MB with an average of 9-10 shots/day and 1-2 days per LaserNetUS run. Each run produces 1.5 – 3 GB. Most of the data is in the form of

images. The largest can be up to 100 MB (for example, if a 100 layer RCF stack detector was used). Data is archived locally and downloaded by the User using a sophisticated web interface.

5.10.2.4 Process of Science

Networking currently does not play much of a role in enabling the science except for providing the resulting datasets to the user via the Cloud or ftp-like mechanism. The Users are responsible for data analysis and data reduction, which they do primarily at their home institutions. This includes, in particular, simulations. Simulations are used to predict the outcome of experiments or the experimental data may be used to guide and benchmark the simulations, after which the simulations can elucidate the mechanisms at play in the experiment. Shared resources are not a major factor for LaserNetUS operation.

Given the great importance of simulations for analyzing LaserNetUS data, some aspects of this are discussed in Section 5.10.2.10, even though this activity is not formally a part of LaserNetUS itself.

5.10.2.5 Remote Science Activities

Remote resources are not generally used in LaserNetUS operation. However, during the current pandemic, some facilities allowed the Users to participate remotely via Zoom. Normally, Users are expected to run their experiment on-site with support from facility personnel. (There is significant variation in this across the various facilities.) LLE was able to run all experiments remotely. Colorado State University and the U. of Nebraska supported remote experiments as well. OSU supported one remote experiment. It is not yet clear to what degree remote participation by Users will continue in the post-pandemic period, but some remote operation is expected to become a part of standard operations.

5.10.2.6 Software Infrastructure

Generally, data management software is not used. Commercial Cloud services are used as discussed in multiple Sections in this document. It is common for the software that came with the data collection apparatus (eg. scanners, CCD detectors, cameras) to be used for initial inspection and handling of the data. ImageJ is frequently used (recall that the largest component of a data set is usually image files). A wide variety of other software is used for initial inspection: MATLAB, Python-based tools, locally written scripts using various modalities and so forth.

The facility may process the raw data (background subtraction, calibration), but the main data processing is done by the Users at their home institutions or on their own laptops and is not a function of LaserNetUS, although LaserNetUS personnel may be heavily involved in the analyses or in discussion of the results over an extended period.

A major component of data analysis involves numerical simulation involving HPC by the Users using resources available to them through their home institutions or though funding agencies and government supported services. This is discussed in Section 5.10.2.10.

5.10.2.7 Network and Data Architecture

The facilities typically use LANs and networking is used to transfer data to the users.

Networking is also used for local control of the laser and experimental systems, although the control component does not generally involve large data throughputs. The tools listed in the instructions for this section are not generally used.

5.10.2.8 Cloud Services

Cloud services are a primary means for transferring data from the LaserNetUS facilities to the Users (and also between members of a given User group), as discussed in multiple Sections. The most common Cloud services are Google Drive and Box, although Microsoft One Drive is sometimes used now. Data analysis, computing and education components are not a significant component of LaserNetUS operation currently. No significant changes are actively planned for the specified time frames. However, it is likely that LaserNetUS will eventually include an education component to its operations.

5.10.2.9 Data-Related Resource Constraints

The LaserNetUS facilities do not currently report significant resource constraints. The current datasets produced during a run are readily handled using the currently available computing, storage, and networking. This is certainly less true of LaserNetUS Users when they are analyzing their datasets at their home institutions, as discussed in Section 5.10.2.10.

The various laser systems operate at a range of repetition rates. Several facilities have PW-class lasers that currently operate at 1 Hz or higher: Colorado State University, INRS, LBNL, University of Nebraska. Additional facilities have <100 TW lasers operating at 1-10 Hz. Most experiments are not able to employ the full repetition rate of these lasers and those that do typically do not do so for extended periods of an hour or more. However, some experiments are now being run at high repetition rate for extended periods. Possibly in the near term and certainly in the 2-5 year time frame, such operation is expected to be common. An example is discussed for Colorado State University in Section 5.10.2.3.

The amount and nature of the data produced varies widely from run to run currently and this is expected to be the case for the foreseeable future. Indeed, the versatility of LaserNetUS that results in this variation is one of its strengths. However, a data output rate of 100 GB/day might be a good estimate for the time when continuous, high repetition-rate operation becomes commonplace. Processing of this data is the User's responsibility and will likely stress the resources of some Users. This is nominally outside the scope of LaserNetUS which ends once the data is transferred to the User.

However, real-time analysis of the data will be critical to guide experiment. Computing will then become a strained resource for the relevant LaserNetUS facilities. It is not yet determined what kind of real-time analysis will be desired or how it will be implemented, but this is expected to be a significant problem in the near future which must be solved if full utilization of LaserNetUS capability is to be achieved in this time frame.

5.10.2.10 Outstanding Issues

The Users of LaserNetUS make heavy use of computer simulations to design their experiments and to analyze the results of an experiment. A common design task is to determine target and laser parameters, for example, a target might consist of multiple

layers, each of which plays a different role (shield, ablator, detector, etc.), and simulations are required to determine the number of layers and their thicknesses as well as the needed laser conditions (number of beams, energy, angle of incidence). A large variety of processes take place at the same time due to the high power of the lasers used, making it very difficult to employ analytic theory. Simulations can suggest which of these processes was (or will be) most important in a given experiment or provide information about the target that cannot be directly measured (space and time varying temperature distribution, collisionality, etc.) The validity of a simulation requires careful choice of numerical parameters (e.g., space and time resolution, choice of difference equations, etc.) and compromises of physicality in the representation (e.g., dimensionality of the simulation, amount of deviation from the parameters actually used in the experiment). An additional choice that must be made is the selection of results from the simulation to be saved. Storage limitations always prevent the saving of everything and most of the simulation results are not saved (e.g., perhaps only selected time steps are preserved). A mistake here may require rerunning the simulation if more information is subsequently found to be necessary. Unless a previous simulation can be used, a simulation campaign usually requires a substantial design phase to select and validate the simulation design, followed by the simulations that provide the actual data that will go into a target design or publication. The two most common simulation types for the experiments of LaserNetUS are PIC (a kinetic simulation method) and hydrodynamic ('hydro', a fluid method), with PIC arguably the more common. Both methods integrate the equations of motion by discretizing time and space and solving difference equations that approximate the true differential equations specified by physical law. Both generally require a large number of processors (dozens to hundreds to thousands to 10's of thousands) running for an extended time (hours to days to weeks).

Due to the wide variety of experiments performed at LaserNetUS and the equally wide variety of simulation design choices that can be made, there is enormous variation in the type and scope of the simulations performed to design an experiment and to analyze one. If a simulation must be designed from scratch, the design and execution of a simulation campaign usually takes longer than the experiment itself, perhaps by 1 or 2 orders of magnitude. The scope of the simulations varies by many orders of magnitude depending on the choice of dimensionality (1D, 2D, and 3D PIC and hydro simulations are all used, sometimes within the same campaign) and the decisions made on how well to attempt to maintain fidelity to the experiment (e.g., full or reduced target size) or physicality (e.g., time and space resolution may be chosen so that some processes cannot be resolved). These choices have a profound effect on the computational and storage requirements. There is no typical set of conditions, but some ranges used can be described. A key issue is whether a similar simulation or set of simulations to those needed has been designed before, reducing the design phase. It appears to be common in current LaserNetUS experiments that a substantial design phase is often required. Finally, this work is performed by LaserNetUS Users and their collaborators, usually working from their home institutions and using local and remote computational resources.

The discussion below is based on interviews with LaserNetUS users and facility personnel who are themselves users of other facilities, as well as a knowledge of the literature.

5.10.2.10.1 Data Production

It is typical to produce between 100 GB to 500 TB of data during the simulation phase. This entire range is fully used with the middle of the range (10's of TB) perhaps most highly populated. PB's are less common for LaserNetUS, but simulations of this scale are performed in the HEDS community, for example, for experiments associated with NIF (LLNL).

During the design phase most of the data may be discarded. The above range represents the amount of data that might be preserved for an extended time (1-2 years) until publication. After publication, some or most of this volume might be discarded, depending on the working style of the investigator. All researchers preserve the information needed to rerun the simulation and most preserve the smaller subset of simulation results that actually goes into the specific figures of a publication.

5.10.2.10.2 Frequency

Typically most LaserNetUS runs have an associated simulation campaign of this scope. Exceptions can include LaserNetUS runs designed to develop a new diagnostic or technique for future campaigns or other facilities. A list of the number of runs for Cycle 3 is provided in Section 5.10.2.2.

5.10.2.10.3 Data Distribution

A single simulation will commonly produce hundreds of files, each often in the range of 10's of GB. A small number of very small files, such as text files, are used to control the simulation or are generated by it. Dozens of simulations may be performed throughout a simulation campaign.

5.10.2.10.4 Data Storage

One of the two most common approaches is to leave the data on the machine that produced it, doing all analysis remotely. In this approach, large transfers of data do not occur. Since supercomputer facilities often purge their storage systems on some basis, the data is either mostly deleted or moved to tape backup with the latter often provided by the supercomputer facility. The other approach is to transfer the data to the home institution for processing. Here, if not eventually mostly deleted, the data is backed up using local servers or external hard drives.

5.10.2.10.5 Computational Requirements

Up to several million CPU-hours is generally needed to complete a campaign. This processing is used to run multi-processor jobs to perform the simulations. The number of cores ranges from 100's to 100's of thousands, with this entire range commonly used.

Additional processing is often required to post-process the results, although this tends to be a small load compared to the simulation itself. Running a large number of single processor jobs, as is done in other fields such as some areas in chemistry, is not a common modality. GPUs are increasingly being used, but do not yet constitute the majority modality. The machines used range from local clusters to facility supercomputers to remote supercomputers. NERSC is one example but no single system appears to dominate.

5.10.2.10.6 Software Infrastructure

A large number of simulation codes are used. Some are commercial, but more common are proprietary codes developed by the user (and often shared) or true open-source codes. Examples of this are EPOCH (PIC) and FLASH (hydro). The simulation codes may come with visualization software, but a wide range of analysis codes are used and specialized analysis codes written specifically for a simulation campaign are common. The most common analysis approach (if there is a most common one) is the use of Python-based platforms, but codes such as MATLAB are commonly used.

5.10.2.10.7 Resource Constraints

Currently the most important resource constraint described by the interviewees is computer time. Resource storage and data transfer can be an issue, but is not currently considered the primary limiting factor. One interviewee described this as a substantial obstacle. This may change as ML becomes more common.

5.10.2.10.8 Cloud Usage

This is not a large part of total resource usage, but services such as Google Drive and Box are frequently used to share highly processed results with collaborators.

Computation is not currently a target.

5.10.2.10.9 Future Needs

0-2 years:

Factor of 2 increase in computing resources will be needed to keep up with the state-of-the-art.

2-5 years:

An order of magnitude increase in computing will be needed.

A very common theme that came up is the increasing use of ML. Although currently not common, this is expected to be a primary modality on this time scale. The full consequences of this are not known. The number of simulations required is expected to increase by a factor of 100 or perhaps much more compared to current campaigns. (10,000 simulations is currently common for ML.) Post-processing, currently an important but minority fraction of total computing, will become a primary use of computing resources in its own right. This ties into the expected increase in available experimental data referred to previously when continuous, high repetition-rate experiments become more common. As stated earlier, this change in experimental work is underway now, and is expected to result in an order of magnitude increase in available data on this time scale. Currently, supercomputing is not used to directly process the experimental data (e.g., background subtraction, calibration), but this may change with ML.

5.10.2.11 Case Study Contributors

LaserNetUS Case Study Representation

- Félicie Albert[1], LLNL
- Alex Arefiev[2], UCSD
- David Blackman[3], UCSD
- Stepan Bulanov[4], LBNL
- Bob Cauble[5], LLNL
- Nick Czapla[6], OSU
- Todd Ditmire[7], University of Texas Austin
- Gilliss Dyer[8], SLAC
- Sylvain Fourmax[9], INRS
- Reed Hollinger[10], Colorado State University
- Andreas Kemp[11], LLNL
- Jean-Claude Kieffer[12], INRS Advanced Laser Light Source (ALLS)
- Karl Krushelnick[13], University of Michigan
- François Legaré[14], INRS Advanced Laser Light Source (ALLS)
- Remi Lehe[15], LBNL
- Jorge Rocca[16], Colorado State University
- Thomas Schenkel[17], LBNL
- Douglass Schumacher[18], The Ohio State University
- Alec Thomas[19], University of Michigan
- Petros Tzeferacos[20], University of Rochester LLE
- Donald Umstadter[21], University of Nebraska Lincoln

---

1 albert6@llnl.gov
2 aarefiev@eng.ucsd.edu
3 drblackman@eng.ucsd.edu
4 sbulanov@lbl.gov
5 cauble1@llnl.gov
6 czapla.4@buckeyemail.osu.edu
7 tditmire@physics.utexas.edu
8 gilliss@slac.stanford.edu
9 sylvain.fourmaux@inrs.ca
10 reed.hollinger@colostate.edu
11 kemp7@llnl.gov
12 legare@emt.inrs.ca
13 kmkr@umich.edu
14 francois.legare@inrs.ca
15 rlehe@lbl.gov
16 jorgerocca9@gmail.com
17 t_schenkel@lbl.gov
18 schumacher.60@osu.edu
19 agrt@umich.edu
20 p.tzeferacos@rochester.edu
21 donald.umstadter@unl.edu

- Jean-Luc Vay[22], LBNL
- Mingsheng Wei[23], University of Rochester LLE
- Anthony Zingale[24], OSU

22 jlvay@lbl.gov
23 mingsheng@lle.rochester.edu
24 zingale.10@buckeyemail.osu.edu

5.11 Multi-Facility FES Workflows

5.11.1 Discussion Summary

A number of pilot use cases and demonstrations have been conducted over the years to couple FES workflows to existing DOE HPC facilities. This experimentation had the modest goals of trying to reduce the number of deployed HPC resources within the FES ecosystem, and utilize higher performing and more well supported resources. Early efforts identified several areas of improvement, and future goals indicate a desire to continue, provided that some areas of friction can be reduced.

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

- In the FES context, a "Multi/Coupled Facility Workflow" is not considered to be a pairwise operation between two specific entities across a network substrate, as in other use cases (e.g., a Light Source using ESnet to reach an ASCR HPC facility). FES views the multi-facility use case as having numerous points:
  - Instrument and local operations staff at once location
  - Collaborating / Participating groups at a number of remote facilities which are linked via communications tools and remote diagnostics to understand and observe experimental progress
  - One or more computational and storage facilities, where dedicated analysis resources are available for inter-shot diagnostics
  - All of these linked by network infrastructure that carries both communications and data transmission
- FES use of cloud services is still being explored. Some use cases are easier to approach, and could be adapted to a cloud with minimal modifications; others require study to understand the technical costs that would be associated.
- The ability to access live data streams from FES experiments will become necessary in the coming years, particularly as experimental facilities more routinely couple to collaborating computing facilities. This multi-facility model will require advanced software to link experimental resources to storage and computing via the network infrastructure.
- The FES community would rather not see all analysis default to using local computational resources. However, to distribute and manage computational demand, there will need to be more unification and resource pooling across the FES complex to allow for fungible operation.
- The ability for ASCR facilities to address an FES multi-facility workflows requires addressing several key areas:
  - Creating a 'dedicated' pool of resources that can be accessed without having to wait in a queue
  - System-wide scheduling; namely ensuring that all components

(computation, storage, networking, & software - at all portions of the end-to-end path) are ready when the analysis procedure starts.
    - Worker nodes on an HPC system having ways to retrieve a remote data set
    - Security of the infrastructure must have automated hooks to facilitate the need to authenticate on multiple systems in multiple locations
    - APIs for computational systems must be aware of the multi-facility nature - and accommodate by allowing multiple observers, and by supporting remote view operations
    - Flexibility to be able to run at multiple DOE HPC facilities
    - Intelligent software stack to manage multi-facility use cases
    - The network(s) that link facilities must have mechanisms to guarantee performance (latency, bandwidth, etc) to eliminate delays during the workflow between shots
- The FES community should explore ways to better utilize computational resources that exist at collaborator sites, as well as DOE HPC facilities, as future research depends on the ability to effectively and efficiently utilize computational resources and increasing volumes of data.
- As the FES community prepares for ITER, the multi-facility use case will become more important as the ITER data volumes will far exceed the storage and processing capacity of any of the major FES facilities. Integration with DOE HPC facilities is critical. Exploring Science DMZ architectures at all FES facilities will be required to ensure that a baseline for data mobility can be achieved.
- The ITER computing and data management model is still under development, but is expected to consist of a main data center located at the instrument, and some set of policies and technology that will be adopted to manage distributed data dissemination to partners around the world. ITER data management will require coordination from the US FES community to ensure efficient and equitable access.
- ITER data rates are still projected to be 50 Gbps (400 Gbps) at peak operation. The ITER timeline, as of 2021, is as follows:
    - First plasma: Dec 2025
    - Additional commissioning and construction: Through Dec 2028
    - Pre-fusion power operations (Phase 1): Dec 2028 through Jan 2030
    - Pre-fusion power operations (Phase 2): June 2032 through Mar 2034
    - Nuclear assembly: 2035
    - Regular operations: Dec 2035

5.11.2 Multi-Facility FES Workflows Case Study

In order to expand the quality, variety, and quantity of analysis performed for fusion experiments, the use cases here describe workflows to send data generated at experimental machines to remote computing centers in near real time for further analysis/modeling. This use case will focus specifically on the use of remote computing

centers for analysis and support during experimental operation, in near real time to aid researchers in providing rapid analysis and shot assessment required to make control-room decisions on the direction of the experiment. The analysis proposed here is in support of the experiments and is distinct from the analysis/simulation of the data which can come many days or weeks after an experiment is run.

5.11.2.1 Background

An example of such a workflow performed in 2020 is shown in Figure 5.11.1, in collaboration with many researchers including at KSTAR (Minjun Choi), ESnet (Eli Dart), and NERSC (Laurie Stephey). Data from the electron cyclotron emission imaging (ECEI) produced at a rate of 8 Gbps was streamed from the KSTAR tokamak in Korea to the Cori supercomputer at NERSC in California. A spectral analysis (cross-coherence, cross-phase, autocorrelation for all channel pairs) of the ECEI data was run in parallel on time chunks of the data, with the data transfer taking less than 3 minutes, and the total time to completion 10 minutes. If run in serial on a single CPU the same analysis would normally take 12 hours. This demonstration of HPC accelerated analysis of KSTAR data provides a valuable proof of principle that such remote analysis of streaming data can provide a valuable resource for the execution of experiments where rapid decisions must be made in the control room based on incomplete information. As more and more data is being integrated into the real-time control systems on KSTAR and in superconducting experiments worldwide, the streaming of such data for accelerated analysis using national HPC infrastructure or dedicated FPGA/GPU architectures can add great value to the shot assessment due to the limited resources available on-site at the facility.

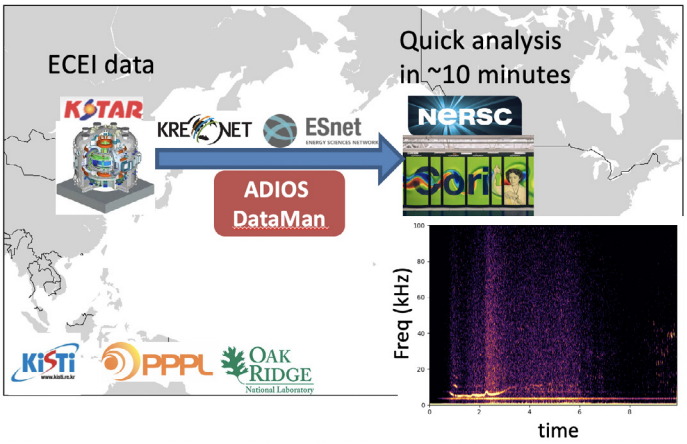


Figure 5.11.1 – KSTAR-NERSC workflow, transferring in near real-time ECEI data from the KSTAR tokamak in Korea, and performing a number of spectral analyses in time for between-shot inspection.

US magnetic confinement fusion researchers have various collaborations with experimental fusion devices across the world. Examples of existing experiments that will be discussed here are:

- KSTAR tokamak at the Korea Institute of Fusion Energy in Daejeon, South Korea
- DIII-D tokamak at GA in San Diego, California
- ITER tokamak (in construction) in Cadarache, France

These experiments typically run plasma shots of duration ~10-100s, with ITER aiming for 500s. Each shot is followed by a cooldown of ~10 minutes on current devices, leading to about 30-40 shots on a typical 8-hour run day.

Tokamaks have many disparate diagnostic instruments for measuring various properties of the fusion plasma (e.g., magnetic probes for MHD activity, scattering and beam diagnostics for fluctuations, etc.). These diagnostics have unique layouts, number of channels, sampling rates, etc. Measurements are typically processed in two distinct ways, first they are digitized and then stored locally on-site and/or used in the PCS in real time . Generally, various predetermined real-time or post pulse analyses are performed on-site, and scientists view, interpret, and use diagnostic measurements and analysis to understand the experiment behavior, and adjust experiment inputs for the next shots.

It should be noted up front that while there are some current remote example workflows, these tools and workflows are still being developed and are not at a daily production-level yet.

5.11.2.2 Collaborators

Table 5.11.1 maps some of the potential data volumes from experimental sites to affiliated computational facilities.

User/Collaborator and Location	Do they store a primary or secondary copy of the data?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of dataset? (report in bytes, e.g. 125GB)	Frequency of data transfer or download? (e.g. ad-hoc, daily, weekly, monthly)	Is data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g. difficult tools, slow network)?
KSTAR	Primary	Data portal and data transfer	~40 Gbps/shot	ad-hoc	-	Global network, can see variability
NERSC	Secondary	Data transfer	~8 Gbps/shot	ad-hoc	y	
DIII-D	Primary	Data transfer	~40 Gbps/shot	ad-hoc	-	
Argonne NL	Secondary	Data transfer		ad-hoc	y	
ITER[1]	Primary	Data transfer	~50 TB/shot	ad-hoc	-	Dependent on agreements in place, TBD

Table 5.11.1

5.11.2.3 Instruments and Facilities

This section will focus on the three aforementioned use cases: **Present-2 years:**

KSTAR is a mid-size, long-pulse tokamak, with maximum pulse lengths on the order of 100's of seconds, but typical on order of 10s.

- KSTAR Experimentation and NERSC Computation
- DIII-D Experimentation and ALCF Computation
- ITER Experimentation and TBD Off-Site Computation

Each of this will be considered for the three time quanta used for the review:

1 ITER will have 1000 second long pulses, generating 50 TB/shot, but not expected to generate at this level till after 2030. Plasma operation begins 2026, and diagnostics will be brought up slowly increasing the data output.

- Present - 2 years
- Next 2-5 years
- Beyond 5 years

5.11.2.3.1 KSTAR Experimentation and NERSC Computation

Present-2 years:

KSTAR is a mid-size, long-pulse tokamak, with maximum pulse lengths on the order of 100's of seconds, but typical on order of 10s.

NERSC Cori supercomputer is a Cray XC40 with a peak performance of about 30 petaflops. It has two partitions, a Haswell partition of 2,388 nodes with 2.3 GHz Intel Xeon processors (32 cores and 128 GB memory per node), and a KNL partition with 9,688 nodes with 1.4 GHz Intel Xeon Phi proessors (68 cores and 96 GB DDR4 and 16 GB MCDRAM per node). A shared Lustre parallel file system of 30 PB is available.

Next 2-5 years:

Various upgrades are planned, but key fundamentals will stay the same in this time frame.

Beyond 5 years:

There is little known about the experimental expectations, or the abilities of the computational resources for this window of time.

5.11.2.3.1 DIII-D Experimentation and ALCF Computation

Present-2 years:

The DIII-D National Fusion Facility is a DOE User Facility and the largest operating magnetic fusion device in the US. A more detailed description of DIII-D is in the accompanying use case. For DIII-D, the fusion plasma typically lasts from 5 to 10 seconds and is diagnosed by ~100 different systems. Data is acquired both during and after the pulse, is stored in local repositories and is available via a client/server architecture. Analysis of this acquired data is performed both automatically and by scientists who are participating in the experimental operation. Analyzed data is written into an MDSplus data management system and is also available via a client/server API.

Cooley has a total of 126 compute nodes; each node has 12 CPU cores and one NVIDIA Tesla K80 dual-GPU card. Aggregate GPU peak performance is over 293 teraflops double precision (using base GPU clocks), and the entire system has a total of 47 terabytes of system RAM and 3 terabytes of GPU RAM.

Next 2-5 years:

Although DIII-D undergoes yearly modifications and upgrades, the key fundamentals discussed in this use case will not change in the next 2-5 years.

Beyond 5 years:

There is little known about the experimental expectations, or the abilities of the computational resources for this window of time.

5.11.2.3.1 ITER Experimentation and TBD Off-Site Computation

Present-5 years:

ITER will not operate till 2026, and little is known at this time about operational patterns and expected computational load.

Beyond 5 years:

ITER will be the largest tokamak ever built and is scheduled to achieve first plasma by 2025 and the full fusion-power campaign scheduled to begin in 2035. Maximum pulse lengths will be on the order 500s. Data rates currently are estimated to be 50 Gbps, generating 10's of terabytes of data per shot and petabytes per day.

5.11.2.4 Process of Science

This section will focus on the three aforementioned use cases:

- KSTAR Experimentation and NERSC Computation
- DIII-D Experimentation and ALCF Computation diagnostics will be brought up slowly increasing the data output.

5.11.2.4.1 KSTAR Experimentation and NERSC Computation

Tests of streaming the data from a particular diagnostic (the Electron Cyclotron Emission imaging diagnostic, ECEI) from KSTAR to NERSC and PPPL for in-between shot analysis have been carried out. This diagnostic data measures temperature fluctuations on fast timescales (500 kHz - 1 MHz). The analysis carried out includes generating processed movies of the fluctuation measurements, and more detailed spectral analysis (e.g., cross-coherence). Visualization of results has been performed using various tools, including matplotlib and a newer web-based dashboard.

This data is streamed using a framework developed at PPPL/ORNL called DELTA, which utilizes the ADIOS2 framework for WAN transfers, and asynchronous message-passing interface (MPI).

5.11.2.4.1 DIII-D Experimentation and ALCF Computation

Combining the operation of DIII-D with automatic data analysis at ALCF has been demonstrated to be possible on a fast enough time scale for the analysis to be used by scientists in the DIII-D control room. The analysis code called SURFMN calculates the magnetic structure of the plasma using Fourier transform. Increasing the number of Fourier components provides a more accurate determination of the stochastic boundary layer near the plasma edge by better resolving magnetic islands, but requires 26 minutes to complete using local DIII-D resources, putting it well outside the useful time range for between-pulse analysis. These islands relate to confinement and ELM suppression, and may be controlled by adjusting coil currents for the next pulse. ALCF ensured on-demand execution of SURFMN by providing a reserved queue, a specialized service that launches the code after receiving an automatic trigger, and with network access from the worker nodes for data transfer. Runs are executed on 252 cores of ALCF's Cooley cluster and the data is available locally at DIII-D within three minutes of triggering.

5.11.2.5 Remote Science Activities

This section will focus on the three aforementioned use cases:

- KSTAR Experimentation and NERSC Computation
- DIII-D Experimentation and ALCF Computation

5.11.2.5.1 KSTAR Experimentation and NERSC Computation

The remote instrument currently in use is the Cori supercomputer at NERSC. Cori is a Cray XC40 with a peak performance of about 30 petaflops. Cori is comprised of 2,388 Intel Xeon "Haswell" processor nodes, 9,688 Intel Xeon Phi "Knight's Landing" (KNL) nodes. A real-time queue is in place that gives access to 6 compute nodes immediately, useful for these asynchronous streaming applications.

5.11.2.5.1 DIII-D Experimentation and ALCF Computation

The remote instrument supporting DIII-D operations is ALCF's Cooley cluster that has 126 compute nodes. Each of these compute nodes has 12 CPU cores and one NVIDIA Tesla K80 dual-GPU card. The entire system has a total of 47 TB of system RAM and 3 TB of GPU RAM. Interconnects between compute nodes is an FDR InfiniBand network.

5.11.2.6 Software Infrastructure

This section will focus on the three aforementioned use cases:

- KSTAR Experimentation and NERSC Computation
- DIII-D Experimentation and ALCF Computation

5.11.2.6.1 KSTAR Experimentation and NERSC Computation

A general framework called DELTA (github.com/rkube/delta) was created to bring together different data transfer pieces, along with parallelization of analysis codes for easy adaptation to new use cases. The data transfer itself utilizes the ADIOS2 framework (github.com/ornladios/adios2), with the DataMan transfer protocol, which enables streaming global WAN transfers, without file disk writes, along with the ability to programmatically switch to file based for testing purposes. The DataMan protocol also allows adjusting the tradeoff between speed and data delivery reliability. The DELTA framework utilizes the mpi4py package for asynchronous processing and parallelization of the data streams. The code is written in Python for ease of use, and wider applicability.

5.11.2.6.1 DIII-D Experimentation and ALCF Computation

For data access external to the DIII-D LAN, the MDSplus data acquisition and management software is the methodology to retrieve both acquired and analyzed data. Details of MDSplus usage are summarized in the DIII-D use case. Data retrieval is facilitated by the MDSplus API. For this use case, a root process of SURFMN running on Cooley connects directly to the MDSplus server at DIII-D and requests the necessary input data. A single process of SURFMN then performs all network communication, and then sends (receives) the relevant data to (from) all other processes via MPI. Once the analysis is complete, SURFMN once again connects to MDSplus,

this time using the unique runID as a key indicating where to store the results. Additionally, it makes a final connection to DIII-D's metadata catalog, updating the runID with metadata about the run, e.g., success or failure, user input and runtime parameters. It also posts a message that the calculation has been completed, and that data is available. At this point, DIII-D scientists in the control room can see the new analysis results and begin to investigate as needed.

5.11.2.7 Network and Data Architecture

This section will focus on the three aforementioned use cases:

- KSTAR Experimentation and NERSC Computation
- DIII-D Experimentation and ALCF Computation

5.11.2.7.1 KSTAR Experimentation and NERSC Computation

A lot of KSTAR data is stored in the MDSplus database format common in fusion tokamaks. The networking architecture within KSTAR network and connected to KREOnet has been upgraded to 100 Gbps, for a clear 100 Gbps line from KSTAR to NERSC utilizing KREOnet and ESnet.

5.11.2.7.1 DIII-D Experimentation and ALCF Computation

The LAN for DIII-D has been described extensively in the DIII-D Use Case. ANL, which houses the ALCF, is connected to ESnet at 2x100GE.

5.11.2.8 Cloud Services

There are no cloud resources used for analysis a NERSC to support KSTAR operation, or ALCF to support DIII-D operation.

5.11.2.9 Data-Related Resource Constraints

In the context of the DIII-D and ALCF use case, there were no immediate data-related resource constraints that have been identified. However, if analysis use cases are identified where very large computation would better support DIII-D operation, it is easy to imagine where insufficient data transfer performance could be a substantial detriment. It is not so much the large amount of data that DIII-D requires but rather the short amount of time (~20 minutes) available to analyze data, for the experimental team to interpret the results, and to reach a decision on what DIII-D parameters to change for the next plasma pulse. In such a time constrained environment, rapid data transfer becomes very important.

5.11.2.10 Outstanding Issues

There have been a number of discussions with leaders at major computational facilities on the challenges of using a large supercomputer to support an operating fusion experiment. Most of these issues center around how leadership-class computing facilities are designed to support codes that take a very long time to run and are best suited to a queuing model and job-scheduler. In contrast, an operating magnetic fusion experiment needs on-demand computing where rapid turnaround time of results is critical. Given the timing requirements, execution must be fully automated. Furthermore, input to an analysis code is going to come from some type of data retrieval API requiring a fast network connection from the data repository to the compute

worker node. How to provide such services is an active area of ongoing research and dialogue.

5.11.2.11 Case Study Contributors

Multi-Facility Workflows Representation

- CS Chang[2], PPPL
- Michael Churchill[3], PPPL
- Steve Sabbagh[4], PPPL
- David Schissel[5], GA

ESnet Site Coordinator Committee Representation

- Scott Kampel[6], PPPL
- Jeff Nguyen[7], GA

2 cschang@pppl.gov
3 rchurchi@pppl.gov
4 sabbagh@pppl.gov
5 schissel@fusion.gat.com
6 skampel@pppl.gov
7 nguyend@fusion.gat.com

5.12 WDM and FES HPC Activities

5.12.1 Discussion Summary

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

- OMFIT is a modeling and experimental data analysis software used in the FES community. OMFIT will adapt existing workflows to advance modeling approaches that use HPC resources, and will be more widely deployed as the community prepares for ITER. It is expected that OMFIT will expand to allow for the use of more analysis codes, at more locations, with more participants. Improvements to the systems that handle data mobility, and ways to automate authentication and authorization, are expected.
- The process of FES simulation workflows has, and will continue to change in the coming years as new codes are developed and more resources become available. The classic style of developing a single code base for a small set of machines is being replaced by models that create ensembles of many codes running on multiple machines. This has also been coupled to research to incorporate a greater number of variables and metrics, adjusting to new time and spatial scales, and overall attempts to create "reduced" data models. These adaptations are being driven by HPC allocations occurring at more locations, but also an increased focus preparing for new experimental facilities such as ITER.
- FES simulation will incorporate the use of AI and ML in the future, as the codes are adapted to run on next-generation machines and at a larger number of facilities.
- PPPL HPC workload that utilize ASCR facilities routinely are not able to perform at peak efficiency due to a number of limitations. Recent upgrades to the PPPL local network and data architecture are expected to alleviate the problems, but further testing will be needed. Some potential bottlenecks to peak efficiency with data mobility are:
    - Security infrastructure on PPPL campus was undersized for the expected data volumes and expected capacities. A recent upgrade should enable a higher level of performance.
    - Data transfer hardware was not regularly used. A recent upgrade to deploy purpose-built DTNs will become a part of several scientific workflows.
    - Data transfer software was not standardized, with projects using a mixture of tools that could not efficiently utilize the network and hardware. PPPL is moving toward more capable tools (e.g., Globus) for their DTN pool.
    - New use cases that mix bulk data movement, as well as real-time streaming, mean that the network must be responsive to latency as well as bandwidth requirements.
    - Simulations run at ASCR HPC facilities are now generating more data than can be easily stored and transferred using existing capabilities. The

upgrades at PPPL, and ongoing upgrades to ASCR HPC facilities, will ensure there are some mechanisms to scale the requirements into the future as exascale simulations become more common.

- The FES community is exploring ways that cloud-provided storage and computation could be integrated into scientific workflows, particularly at facilities that are not able to scale local resources either due to cost, space, or lack of expertise to operate long-term storage pools. Investigations are underway to understand the costs and usability for FES workflows.

5.12.2 WDM and FES HPC Activities Case Study

5.12.2.1 Background

This case study will combine the collective works of several PIs from the FES community. Their work spans the overall field of HPC use, and the relationship to FES research as a whole. Due to the overlapping nature of some of the facilities featured in this document, references to prior sections that discuss the overall technical capabilities of a site or project will be used.

5.12.2.1.1 SciDAC Program

The US DOE SciDAC program[1] was created to bring together many of the nation's top researchers to develop new computational methods for tackling some of the most challenging scientific problems.

Scientific computing, including modeling and simulation, is crucial for research problems that are not solvable by traditional theoretical and experimental approaches, are hazardous to study in the laboratory, or are time-consuming or expensive to solve by traditional means. Beyond the scientific computing and computational science research embedded in the SC Core Programs, SC invests in a portfolio of coordinated research efforts directed at exploiting the emerging capabilities of HPC. The research projects in this portfolio respond to the extraordinary difficulties of realizing sustained peak performance for those scientific applications that require HPC capabilities to accomplish their research goals.

The most recent focus for SciDAC , as of 2017, is to enable scientific breakthroughs on pre-exascale computing architectures. The partnerships now include DOE's Office of Nuclear Energy in addition to all 6 SC programs. SciDAC projects pursue computational solutions to challenging problems in climate science, fusion research, HEP, nuclear physics, astrophysics, material science, chemistry, particle accelerators, and nuclear fuels, and ensure that progress at the frontiers of science is enhanced by advances in computational technology, most pressingly, the emergence of the hybrid and many-core architectures and ML techniques. The SciDAC program is recognized, both nationally and internationally, as the leader in accelerating the use of HPC to advance the state of knowledge in science applications.

SciDAC projects are collaborative basic research efforts involving teams of physical scientists, mathematicians, computer scientists, and computational scientists working on major software and algorithm development to conduct complex scientific and engineering computations on leadership-class and high-end computing systems at

1 https://www.scidac.gov/about.html

a level of fidelity needed to simulate real-world conditions. Research funded under the SciDAC program addresses the interdisciplinary problems inherent in HPC and problems that cannot be addressed by a single investigator or small group of investigators.

5.12.2.1.2 MIT PSFC

The MIT PSFC is fully described in Section 5.4. What follows is a brief overview of some HPC activities undertaken by the facility.

5.12.2.1.2.1 HPC and SciDAC

These SciDAC activities use HPC to develop advanced reduced models that can be used in WDM calculations.

- RF SciDAC: Development of model hierarchies for RF wave-particle interactions in tokamaks through HPC. Model hierarchies consist of both first principle and reduced models.
- Multiscale Gyrokinetic SciDAC: investigation of multiscale (ion and electron) gyrokinetic transport in tokamak core and edge using continuum models (GS2, Gyrokinetic Electromagnetic Numerical Experiment [GENE], and GKEYLL).
- AToM: investigation of mutli-scale (ion and electron) gyrokinetic transport in tokamak core using CGYRO.
- Building ML models to accelerate RF simulations: Building a large database of simulations using GENRAY/CQL3D which are used as training/testing for ML surrogate models. Tens of thousands of simulations run on Engaging and at NERSC.

5.12.2.1.2.2 International Collaborations

There are three major international collaborations: WEST, EAST, and W7-X

- EAST: Through modeling and experiment, RF actuator control and its extension to high-performance plasmas in the EAST tokamak is investigated, in connection with lower hybrid current drive experiments conducted on Alcator C-Mod. HPC simulations (TorLH) of lower hybrid (LH) wave propagation in the EAST Device and ray tracing / Fokker Planck simulations (GENRAY-CQL3D) of LH current drive aimed at constructing databases on which control level (reduced) algorithms can be trained.
- WEST: Applying RF tools to experimental program and modeling at WEST. High-fidelity simulations of ICRF wave coupling and propagation in WEST using the Physics Equation Translator for MFEM finite element based framework. Simulations of LH current drive using the ray tracing / Fokker Planck model GENRAY-CQL3D.

5.12.2.1.3 M3DC1

M3D-C1, developed by PPPL, is a code that solves the extended-magnetohydrodynamic (MHD) equations, which is a model that describes plasma as

electrically conducting fluids of ions and electrons. This code is primarily used for calculating the equilibrium, stability, and dynamics of fusion plasmas, but has also been used for a variety of other applications, including astrophysical applications. In particular, M3D-C1 is designed to address some of the most critical challenges confronting tokamak plasmas: large-scale instabilities, which significantly degrade thermal confinement; and disruptions, which cause complete loss of energy confinement and which may cause damage to reactor-scale tokamaks such as ITER.

M3D-C1 builds upon some of the design principles pioneered by M3D, but the two codes are developed independently and do not share source code. The "C1" in M3D-C1 refers to the C1 property of its finite elements, which ensures that both the value and the derivatives of fields are continuous across mesh element boundaries.

Advanced numerical methods are employed in M3D-C1 to permit the efficient solution of its numerical model over a broad range of temporal and spatial scales. These methods include the use of high-order finite elements on an unstructured mesh; fully implicit and semi-implicit time integration options; physics-based preconditioning; and parallelization via domain decomposition and the use of scalable parallel sparse linear algebra solvers.

5.12.2.1.4 ECP-WD

The goal of the ECP-WDM program is the high-fidelity, first-principles-based WDM that relies on coupling of multiple codes including on exascale computers. Until 2023, the ECP project will perform Tcore-edge coupling utilizing a core delta-f gyrokinetic code GENE or Electromagnetic Gyrokinetic Turbulence Simulation (GEM) and an edge total-f gyrokinetic code XGC. The eventual goal, after 2023, is to include plasma heating/current-drive and material-wall interaction modules, to predict fusion energy production from first-principles-based models. The headquarters of the ECP-WDM program is the Theory Department and the Computational Science Department at PPPL.

When the ECP-WDM code is complete, the data will be analyzed by a community distributed across the US.. Important simulation data needs to be retained for 5 years supporting the development of fusion surrogate models and digital twins.

The edge component code XGC will dominate the computing and data needs of ECP-WDM due to the intrinsic multiscale full-f nature of the edge physics and the required unstructured triangular grid to describe the complicated edge geometry. Since XGC will occupy about half of the whole plasma volume, the network and data requirement of the ECP-WDM code will be about half of those by XGC, as described in the Case #5. A simulation of turbulence transport in an ITER-like plasma for a given equilibrium time slice by ECP-WDM on Summit will produce about 20 PB of particle data for two-days of wall-clock time. Since such a large amount of data cannot be saved in the OLCF scratch filesystem, only about 2 TB of mesh data can be retained. It is desirable to move this data to PPPL and collaborative users for an in-depth interactive physics analysis after each one-day simulation. From the upcoming exascale computers, there will be a requirement to move about 10 TB physics data to PPPL and distributed users after each simulation. If streaming mechanisms for data analysis from an exascale HPC memory to remote cluster memory are used, it will require dealing with streaming particle data whose rate can be up to 120 PB/20 hours, which

corresponds to 1.6 Tbps. If the data can be moved at full capacity of 100 Gbps ESnet, it would be about 0.6% of the particle data. The streaming-analyzed data will have to be lossy-compressed for storage at PPPL or on cloud. At a compression ratio 10, it would save 75 TB of streamed exascale-HPC data per simulation.

5.12.2.1.5 OMFIT

OMFIT[2] in the context of integrated simulations for interpretative and predictive modeling. To zeroth order, OMFIT can be considered as a framework for loosely coupled workflows. A typical, example of such a workflow follows:

- Data (simulated or experimental) is loaded into the framework. OMFIT does not differentiate between local or remote access of data. Remote data is transferred where OMFIT resides (such as either a personal laptop or an HPC system).
- Data is manipulated within the framework.
- Data is moved to some other (local or remote) system for (local or remote) execution of software. OMFIT does not differentiate between local or remote execution.
- This process may be repeated.

5.12.2.2 Collaborators

5.12.2.2.1 SciDAC Program

There are currently 2 SciDAC institutes with 24 participating institutions and a total annual funding of $12 million. The mission of the SciDAC-4 institutes is to provide intellectual resources in applied mathematics and computer science, expertise in algorithms and methods, and scientific software tools to advance scientific discovery through modeling and simulation in areas of strategic importance to the US DOE and the DOE SC.

Specific goals and objectives for the SciDAC institutes are to support, complement, or develop the following:

- Tools and resources for lowering the barriers to effectively use state-of-the-art computational systems.
- Mechanisms for taking on computational grand challenges across different science application areas.
- Mechanisms for incorporating and demonstrating the value of basic research results from applied mathematics and computer science.
- Plans for building up and engaging the nation's computational science research communities.

5.12.2.2.2 MIT PSFC

The MIT PSFC is fully described in Section 5.4. What follows is a brief overview of some HPC activities undertaken by the facility.

2 https://omfit.io

User/Collaborator and Location	Do they store a primary or secondary copy of the data?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")		Frequency of data transfer or download? (e.g. ad-hoc, daily, weekly, monthly)	Is data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g. difficult tools, slow network)?
RF SciDAC partners located at LLNL, MIT, PPPL, ORNL, CompX (Del Mar, CA), Tech-X & Lodestar Research (Boulder, CO), RPI (Troy, NY), UIUC (Champaign, IL),	Primary	Data portal and Data transfer. The project has immediate access to a projects directory at NERSC with 20 TB maximum storage aside from the High-Performance Storage System (HPSS), which is limitless.	25-30 GB at NERSC.	Data stays mostly at NERSC	N	No known issues
Multiscale gyrokinetic SciDAC Center partners: located at PPPL, UT Austin, U of Maryland, MIT, and LLNL	Primary	Data portal and data transfer	4 terabytes on NERSC; limited amount data are transferred	Data stays mostly at NERSC	N	No known issues
AToM integrated modeling SciDAC partners are located at GA, LLNL, MIT, ORNL, and PPPL.	Primary	Data portal and data transfer		Data stays mostly at NERSC	N	No known issues
AI / ML SciDAC on Building machine learning models to accelerate RF simulations partners located at LBNL, MIT, and PPPL	Primary	Data portal and data transfer	1 GB	Data stays mostly at NERSC and on the PSFC Engaging cluster	N	No known issues
EAST International Collaboration partners located at CAS-IPP (Hefei, China), GA (San Diego), Lehigh U., LLNL, MIT, and PPPL	Primary	Data portal and data transfers on computers located at the CAS-IPP, MIT (Engaging cluster), and NERSC	1 - 10 GB	Experimental data stays mostly on data servers at CAS-IPP and MIT. Simulation data remains mostly at NERSC and on the MIT (Engaging) and CAS-IPP (Shenma) computing clusters	N	Data transfers from CAS-IPP (Hefei, China) back to the US can be slow. Interactive workflows involving X-window displays can be very slow.

User/Collaborator and Location	Do they store a primary or secondary copy of the data?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")		Frequency of data transfer or download? (e.g. ad-hoc, daily, weekly, monthly)	Is data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g. difficult tools, slow network)?
WEST International Collaboration partners located at MIT, ORNL, and PPPL	Primary		Hundreds of datasets have been produced in first 6 months of project. Each dataset is a few MB. Unlikely to be steady rate.	Experimental data stays mostly on data servers at WEST (Cadarache, Fr) and MIT. Simulation data remains mostly at NERSC and on the MIT (Engaging) computing clusters	N	No known issues

Table 5.12.1 – MIT PSFC Data Relationships

5.12.2.2.3 M3DC1

M3D-C1 data is generated by a number of different users on several HPC platforms, including NERSC and local clusters at Princeton University/PPPL (Princeton, NJ) and GA (San Diego, CA). This data is typically analyzed in place, but is occasionally transferred for the purposes of long-term storage or for restarting simulations on a different platform. NERSC HPSS in particular is often used for long-term storage. Datasets vary in size depending on the use case, and can range from ~10 GB to ~1 TB. A typical representative case would be a few hundred GB.

User/Collaborator and Location	Do they store a primary or secondary copy of the data?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of dataset? (report in bytes, e.g. 125GB)	Frequency of data transfer or download? (e.g. ad-hoc, daily, weekly, monthly)	Is data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g. difficult tools, slow network)?
M3D-C1 Users at PPPL	Primary	Data Portal	100 GB	Monthly	N	
M3D-C1 Users at GA	Primary	Data Portal	10 GB	Monthly	N	
M3D-C1 Users at NERSC	Primary	Data Portal	1 TB	Monthly	N	

Table 5.12.2 – M3DC1 Data Relationships

5.12.2.2.4 ECP-WD

User/Collaborator and Location	Do they store a primary or secondary copy of the data?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of dataset? (report in bytes, e.g. 125GB)	Frequency of data transfer or download? (e.g. ad-hoc, daily, weekly, monthly)	Is data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g. difficult tools, slow network)?
WDM Users at PPPL	Primary	Data transfer Data portal	Current: 1TB 2-5 years: 10TB 5 and beyond: 100TB	Monthly	Y	

User/Collaborator and Location	Do they store a primary or secondary copy of the data?	Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe "other")	Avg. size of dataset? (report in bytes, e.g. 125GB)	Frequency of data transfer or download? (e.g. ad-hoc, daily, weekly, monthly)	Is data sent back to the source? (y/n) If so, how?	Any known issues with data sharing (e.g. difficult tools, slow network)?
WDM Users at The University of Texas at Austin	Primary	Data transferData portal	Current: 1TB 2-5 years: 10TB 5 and beyond: 100TB	Monthly	N	
WDM Users at University of Colorado Boulder	Primary	Data transferData portal	Current: 1TB 2-5 years: 10TB 5 and beyond: 100TB	Monthly	N	
Other major lab Users		Data transfer Data portal	2-5 years: 10TB 5 and beyond: 100TB	Monthly	N	

Table 5.12.3 – ECP-WD Data Relationships

5.12.2.2.5 OMFIT

Figure 5.12.1 illustrates the nationality of the over 1000 users and 70+ developers from over 35+ institutions[3].

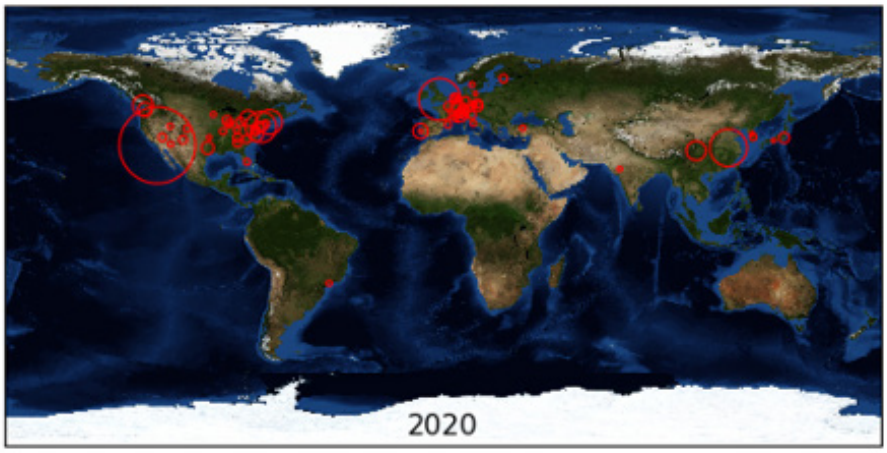


Table 5.12.1 – OMFIT Data Relationships

5.12.2.3 Instruments and Facilities

5.12.2.3.1 SciDAC Program

There are two major facilities to highlight within SciDAC:

- **FASTMath** — Frameworks, Algorithms, and Scalable Technologies for Mathematics
- **RAPIDS—SciDAC** Institute for Computer Science and Data

The FASTMath Institute develops and deploys scalable mathematical algorithms and

3 https://omfit.io/contributors.html

software tools for reliable simulation of complex physical phenomena and collaborates with domain scientists to ensure the usefulness and applicability of FASTMath technologies.

The RAPIDS Institute solves computer science and data technical challenges for SciDAC and SC science teams, works directly with SC scientists and DOE facilities to adopt and support RAPIDS technologies, and coordinates with other DOE computer science and applied mathematics activities to maximize impact on SC science.

5.12.2.3.2 MIT PSFC

For most MIT PSFC HPC needs, there is reliance on NERSC Facilities (Cori[4]), specifically the new Perlmutter [5]system that will be coming online at NERSC in 2021. At MIT there is also a 3200 core computing cluster (the PSFC Engaging CLuster) co-located in Holyoke, MA as well as a new GPU platform located on the MIT campus at the PSFC.

PSFC@Engaging

The PSFC@Engaging computational cluster consists of a 100 compute node subsystem integrated into the "Engaging Cluster," which is located at the MGHPCC in Holyoke, MA. The PSFC subsystem is operated as part of the "Engaging Cluster" with access to a 2.5 Petabyte parallel file system. The total subsystem is 3200 cores with 12.8 Terabytes of memory. This cluster is used to run advanced RF simulation tools such as GENRAY/CQL3D, TORIC, and TorLH, advanced gyrokinetic codes such as GYRO, GS2, GENE, and Gkeyll, and ML algorithms. In addition, Python-based frameworks such as πScope and Petra-M (piscope.psfc.mit.edu) have been implemented on the cluster for managing the workflows of the RF simulation codes.

This 100 node subsystem is connected together by a high-speed, non-blocking FDR Infiniband system. This Infiniband system is capable of 14 Gbps with a latency of 0.7 microseconds. This network is non-blocking, thus each node has immediate access to each other node as well as to the parallel file system.

Each compute node in the subsystem is configured with two Intel E5, Haswell-EP processors at 2.1GHz, 16 cores per processor, for a total of 32 cores per node. Each node has128 GB DDR4 of memory with 1.0 TB on the local disk. The individual compute nodes are very similar to the compute nodes in the "Cori – Phase 1" system at NERSC.

PSFC GPU extension

A test cluster for gpu acceleration of physics modeling consisting of six compute nodes with a total of 24 gpu cards is available. The acquisition of these gpu nodes follows the trend of the PSFC having local compute resources for training and preparation of use of larger facilities. This set of gpu resources will serve several purposes for the PSFC. The 24 gpus cards are a mixture of two types. One type is optimized for ML algorithms and the second for computations though it also can be applied to ML. These will be immediately useful to ML researchers at the PSFC and are of sufficient size to significantly accelerate the workflows in Deep Learning for disruption prediction and regression models for LH actuators.

4 https://www.nersc.gov/systems/cori/
5 https://www.nersc.gov/systems/perlmutter/

- The first type of gpu is an NVIDIA Quadro RTX 6000 with 24 GB of memory and PCIe 3.0 x16 (8.0GT/s). There will be 4 of these gpus per node with three nodes.
- The second type of gpu is an NVIDIA Tesla V100S with 32 GB of memory and a PCIe 3.0 module, has full precision and is of the type being considered for the next NERSC system, Perlmutter. There will be 4 of these gpus per mode with three nodes. This gpu system will serve as a development platform for members of the PSFC to prepare their codes and workflows for the Perlmutter system expected to come online in two phases in 2021 and 2022.
- Finally, this small gpu system will help inform possible expansions of it into a larger gpu cluster to continue to provide the dual platforms the PSFC has used of local and NERSC resources.

Datasets produced at NERSC vary in size from ~ 1 GB to several TB. Some simulations archive data in multi-nested directories consisting of 100's of files.

5.12.2.3.3 M3DC1

M3D-C1 is a code used for modeling fusion plasmas. This code is primarily run on HPC resources at Princeton University/PPPL (Princeton NJ), GA, and NERSC. Output file number can vary considerably depending on use case, with some cases generating ~100 in one directory and other cases generating ~10k files across ~100 directories. Total data set size per simulation is typically in the range of 100 GB–1 TB. Across all users, M3D-C1 typically generates a few 10s of TB of data per year per facility, although much of this is not saved due to limits on storage capacity. It is expected that the number of M3D-C1 users to increase steadily over the next five years. If enough storage were available, M3D-C1 users could easily be generating 100 TB per year now, and 1 PB per year in a few years. Accessing and transferring data to long-term storage facilities is a primary obstacle here.

5.12.2.3.4 ECP-WD

Present-2 years

ECP-WDM runs on the 200PF Summit at OLCF, 11PF Theta at ALCF and 30PF Cori at NERSC. ECP-WDM App will also have access to the exascale Frontier at OLCF, 100PF Perlmutter at NERSC and ~40PF Polaris at ALCF in 1 year.

On Summit, the number of files is about 100,000 and the number of directories is about 1,000. The maximum size of a file is about 5GB. On Frontier, the number of files will be about 1,000,000 and the number of directories will be about 10,000, with the maximum size of a file being about 50GB.

Next 2-5 years

ECP-WDM plans to use the 1.5EF Frontier at OLCF, 1 EF Aurora at ALCF, and 100PF Perlmutter. The data and streaming rate will be similar to those from Frontier, as described in the current 1-2 requirement. On Frontier, the number of files will be about 1,000,000 and the number of directories will be about 10,000 with the total number of data size 300 PB per simulation.

Beyond 5 years

ECP-WDM's usage of the exascale HPCs will become more intense, with a few wall-clock days of simulation per study. ECP-WDM will contain at least 10X greater number of species for longer wall-clock days of simulation. Plasma heating/current drive codes will be coupled in. If there are post-exascale machines available beyond 5 years from now, ECP-WDM will try to utilize them for bigger science studies. On Frontier, the number of files will be about 10,000,000 and the number of directories will be about 100,000, with the total amount of data generated per five-day simulation being approximately 1EB.

5.12.2.3.5 OMFIT

Different users use different instruments and/or facilities. OMFIT has public installations that are kept up-to-date at the largest fusion facilities worldwide[6]. Most users run these and use the computing clusters they are installed on. However, a non-negligible fraction also uses their own personal installations (e.g., on laptop).

Most OMFIT sessions are interactive (e.g., data is not collected between-shot DIII-D analyses or OMFIT batch runs). To date 470 users have run OMFIT at GA, and the aggregated data of their zipped save project files is illustrated in the figure. Interactive OMFIT sessions are mostly run on the interactive nodes of the IRIS cluster, a shared resource where the four interactive nodes have 256 GB of RAM and 32 cores each. Typical number of simultaneous OMFIT sessions on the cluster average on the order of 50 per node (200 total).

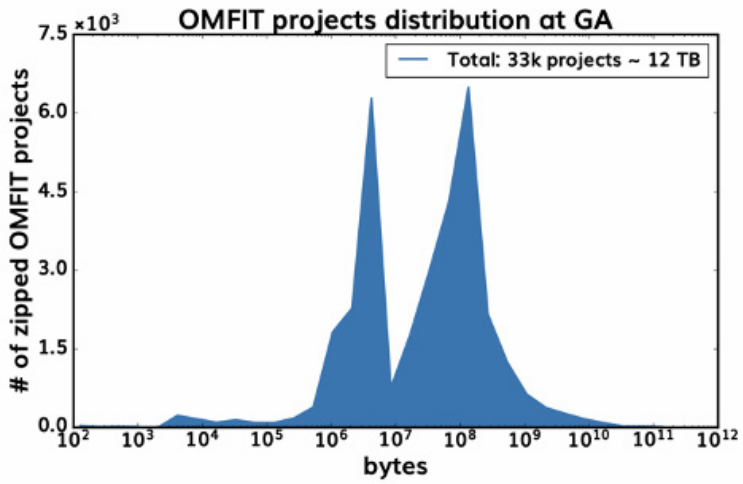


Figure 5.12.2 – OMFIT Project Distribution

5.12.2.4 Process of Science

5.12.2.4.1 SciDAC Program

As a national computational resource for scientific discovery, SciDEC workflows vary depending on the needs of each individual project participant.

6 https://omfit.io/run.html#publicly-available-installations]

5.12.2.4.2 MIT PSFC

A common, exemplar "network heavy" workflow is as follows: A python-based framework called "PiScope" that runs locally on workstations at MIT to access experimental data from the Alcator C-Mod data system is used to create input files from the experimental data for high-performance RF wave - particle simulation codes (mostly ray tracing and Fokker Planck solvers). These simulations are then run on remote platforms including the PSFC Engaging Cluster in Holyoke, MA and NERSC. The input files for the simulations are typically ASCII and the output files are both ASCII and NETCDF formats. Data is transferred to and from the remote platforms via SCP.

The PiScope workflow has been found to be so efficient that is has enabled science advances in a number of areas including the study of LH range of frequency waves are scattered by coherent blob-like turbulence in the edge or scrape-off layer of a tokamak (Alcator C-Mod and the EAST Tokamak). These studies rely on being able to perform thousands to tens of thousands ray tracing / Fokker Planck simulations in order to obtain meaningful statistics and explore high dimensional input parameter spaces. Similarly, the PiScope workflow has been used to construct simulation databases consisting of the order of 100,000 ray tracing / Fokker Planck simulations in order to train reduced models for RF current drive, usable for tokamak control level applications.

5.12.2.4.3 M3DC1

M3D-C1 simulations model both actual and hypothetical fusion experiments. Workflow typically involves taking data (< 1 GB) from an experiment to initialize the M3D-C1 simulation, and then analyzing the M3D-C1 output which represents the predicted evolution of the plasma. Raw output data represents spatially resolved magnetic and thermodynamic quantities. This is analyzed locally (generally without data transfer) using scripts written in Python, IDL, C++, and MATLAB. Due to storage restrictions, data is only stored at a fraction of the actual time resolution of the code (typically full data only from every ~100th time step or so is stored).

5.12.2.4.4 ECP-WD

Present to 2 years

ECP-WDM currently runs on the 200PF Summit at OLCF. ECP-WDM will also have access to the exascale Frontier at OLCF, 100PF Perlmutter at NERSC and ~40PF Polaris at ALCF in 1 year. Reduced size physics-data output is up to 2 TB from Summit and 10 TB from Frontier per simulation, which are desired to be transferred to PPPL and user facilities for timely physics analysis. When the exascale computing is established, streaming data analysis from an exascale HPC memory to a PPPL and user cluster memory will be employed. The available source of streaming data per simulation can be up to 250 PB/20 hours, which must be significantly reduced for a timely transfer via ESnet. If it is possible to move the data at full capacity of 100 Gbps ESnet, 0.2% of the particle data could be analyzed. The streaming-analyzed data will have to be lossy-compressed for storage at PPPL or on cloud. At a compression ratio 10, it is possible to save 75 TB of streamed exascale-HPC data per one-day simulation. An ADIOS based web portal will be used for streaming data analysis, compression, provenance, and storage workflow.

2-5 years

The data workflow requirement will be the same as for the Frontier requirement in the 1-2 years time frame. ECP-WDM can be well established for exascale computing on Frontier and Aurora, the use of streaming data analysis from an exascale HPC memory to a PPPL cluster memory is planned. The available source of streaming data per simulation can be up to 250 PB/20 hours, which must be significantly reduced for a timely transfer via ESnet. If the data can be moved at full capacity of 100 Gbps ESnet, it would be possible to analyze about 0.2% of the particle data. The streaming-analyzed data will have to be lossy-compressed for storage at PPPL or on cloud. At a compression ratio 10, it is possible to save 75 TB of streamed exascale-HPC data per one-day simulation. An ADIOS based web portal will be used for streaming data analysis, compression, provenance, and storage workflow.

Beyond 5 years

ECP-WDM's usage of the exascale HPCs will become more intense, with a few wall-clock days of simulation per study. Runs will contain at least 10X greater number of species for longer wall-clock days of simulation. Plasma heating/current drive codes will be coupled in. If there are post-exascale machines available beyond 5 years from now, ECP-WDM will try to utilize them for bigger science studies. It is anticipated that the data network and science analysis workflow will handle at least five times more data than needed in the 2-5 years period.

5.12.2.4.5 OMFIT

OMFIT is a framework that enables all kinds of integrated analyses and predictive simulations. It is explicitly cited in 150+ journal publications for a wide variety of applications[7 8].

5.12.2.5 Remote Science Activities

5.12.2.5.1 SciDAC Program

Virtually all participation in SciDAC are via remote access to HPC resources.

5.12.2.5.2 MIT PSFC

The primary remote resources that are utilized for HPC needs are NERSC, located at LBL (Cori and in the future Perlmutter) and the 3200 core PSFC Engaging Cluster co-located in Holyoke, MA.

In the present 2-years, next 2-5 years, and beyond 5 years it is anticipated that no major changes in how Cori, and subsequently Perlmutter, will be used. Similarly in the present 2 years it is anticipated that no major change in how the PSFC Engaging cluster will be used. However in the 2-5 year period and beyond the 5 year period it is expected the PSFC Engaging cluster will transition from a CPU-based to a GPU based system in order to aid MIT researchers in preparing to use Perlmutter. It is not clear how this latter transition will affect data transfers to and from the PSFC Engaging cluster however.

7 https://scholar.google.com/scholar?oe=utf-8&cites=18003378855044027491
8 https://scholar.google.com/scholar?oe=utf-8&cites=1329848731563781128

5.12.2.5.3 M3DC1

M3D-C1 data is typically analyzed in place on the HPC platform that was used to generate the data. Data transfer to remote resources is usually done for the purpose of long-term storage or for continuing a simulation using a different HPC platform (for example, if a user's allocation is spent on one platform). Significant changes to these patterns in the near term are not foreseen.

5.12.2.5.4 ECP-WD

Present – 2 years

ECP-WDM currently runs on the 200PF Summit at OLCF. ECP-WDM will also have access to the exascale Frontier at OLCF. Reduced size physics-data output is up to 2 TB from Summit and 10 TB from Frontier per simulation, which are desired to be transferred to PPPL and user facilities for timely remote physics analysis. When the ECP-WDM exascale computing is established, there is a plan to use streaming data analysis from an exascale HPC memory to a PPPL and User cluster memory. The available source of streaming data per simulation can be up to 250 PB/20 hours, which must be significantly reduced for a timely transfer via ESnet. If the data can be moved at full capacity of 100 Gbps ESnet, about 0.2% of the particle data could be analyzed. The streaming-analyzed data will have to be lossy-compressed for storage at PPPL or on cloud. At a compression ratio 10, it is possible to save 75 TB of streamed exascale-HPC data per one-day simulation. An ADIOS based web portal will be used for streaming data analysis, compression, provenance, and storage workflow.

2-5 years

The data workflow requirement will be the same as for the Frontier requirement in the 1-2 year timeframe. ECP-WDM can be well established for exascale computing on Frontier and Aurora, there is a plan to use streaming data analysis from an exascale HPC memory to a PPPL cluster memory. The available source of streaming data per simulation can be up to 250 PB/20 hours, which must be significantly reduced for a timely transfer via ESnet. If the data can be moved at full capacity of 100 Gbps ESnet, it is possible to analyze about 0.2% of the particle data. The streaming-analyzed data will have to be lossy-compressed for storage at PPPL or on cloud. At a compression ratio 10, it is possible to save 75 TB of streamed exascale-HPC data per one-day simulation. An ADIOS based web portal will be used for streaming data analysis, compression, provenance, and storage workflow.

Beyond 5 years

ECP-WDM's usage of the exascale HPCs will become more intense, with a few wall-clock days of simulation per study. ECP-WDM will contain at least 10X more species and data will be collected over longer simulation durations than at present. f In addition, codes will include Plasma heating additions to current drive codes. If there are post-exascale machines available beyond 5 years from now, ECP-WDM will try to utilize them supporting more detailed simulation campaigns. It is anticipated that the data network and science analysis workflow to remote PPPL and user facilities will handle at least five times more data than what was needed in the 2-5 years period.

5.12.2.5.5 OMFIT

OMFIT remote capabilities rely on SSH connections (with the possibility of multi-hop tunneling) and support for two-factor authentication.

Data transfer between where OMFIT is run and remote experimental facilities is often where things can be slow. However, for data analyses and simulations that are especially heavy (e.g. for diagnostics or simulations generating large amounts of data) users tend to run OMFIT where the data is located. Also, local caches (both in memory and disk) alleviate some of these issues.

5.12.2.6 Software Infrastructure

5.12.2.6.1 MIT PSFC

An overview of the MIT PSFC software landscape can be found in Section 5.4.2.6.

5.12.2.6.2 M3DC1

M3D-C1 users utilize scp and globus for file transfers among different HPC facilities. M3D-C1 data is analyzed using various scripts written in Python, IDL, C++, and MATLAB. Significant changes in the use of these data transfer methods is not foreseen.

5.12.2.6.3 ECP-WD

While PPPL uses a variety of data transfer protocols including SCP, BBCP, FTP, etc, the primary and recommended tools over the last several years has been Globus -Grid FTP. PPPL has a 10g Globus server in the PPPL DMZ, with direct connectivity to HPC storage servers. A new, 100g capable server is currently being tested and tuned. It is expected to be deployed in the next several months.

Performance monitoring in the WAN is handled by PerfSonar, currently running on the PPPL internal network. A new, 100g capable PerfSonar node is also being deployed and will be capable of testing in all areas of the PPPL core network (outside, DMZ, inside).

Internally, bandwidth monitoring is handled by the PRTG software package, which provides current and historical throughput data for key points within the PPPL network.

Present-2 years

Currently, ECP-WDM is using Globus to move the physics data to PPPL from computing facilities. Files are in ADIOS-BP format. A prototype DELTA framework has been developed for remote data flow with streaming analysis capability. A web-based data management protocol, based on an eSimMon dashboard, is under development combining ADIOS and DELTA capabilities into it, which will be operational within 2 years.

2-5 years and beyond

The web-based data management protocol will be fully operational, which combines in the ADIOS2, DELTA, eSimMon technologies. ECP-WDM simulation on exascale computers will be operated like a large experimental facility, in which the simulation scenario will be jointly developed by distributed collaborators across US and different continents. The simulation data will be streamed in real time to the distributed

collaborators for aggregated simulation steering information and timely scientific discovery.

5.12.2.6.4 OMFIT

An overview of the GA software landscape can be found in Section 5.3.2.6.

5.12.2.7 Network and Data Architecture

5.12.2.7.1 MIT PSFC

An overview of the MIT PSFC networking landscape can be found in Section 5.4.2.7.

5.12.2.7.2 M3DC1 & ECP-WD

An overview of the PPPL networking landscape can be found in Section 5.5.2.7.

5.12.2.7.3 OMFIT

An overview of the GA networking landscape can be found in Section 5.3.2.7.

5.12.2.7.4 ORNL

An overview of the ORNL networking landscape can be found in Section 5.8.2.7.

5.12.2.8 Cloud Services

5.12.2.8.1 MIT PSFC

An overview of the MIT PSFC cloud usage landscape can be found in Section 5.4.2.7.

5.12.2.8.2 M3DC1 & ECP-WD

An overview of the PPPL cloud usage landscape can be found in Section 5.5.2.7.

ECP-WDM's usage of cloud services will be in the data storage, since a fast data access from the big physics data is becoming a constraint on scientific discovery. One condition for use of cloud data storage is the need for fast sustained access speeds. Additionally, being able to perform data analysis via cloud memory could be very useful, so that only the analyzed data could be transferred to PPPL and other user facilities.

5.12.2.8.3 OMFIT

An overview of the GA cloud usage landscape can be found in Section 5.3.2.7.

Network team does not deploy cloud services currently. It is expected to keep devices on premise where possible.

5.12.2.8.4 ORNL

An overview of the ORNL cloud usage landscape can be found in Section 5.8.2.7.

5.12.2.9 Data-Related Resource Constraints

5.12.2.9.1 MIT PSFC

An overview of the MIT PFSC resource constraints can be found in Section 5.4.2.9.

5.12.2.9.2 M3DC1

An overview of the PPPL resource constraints can be found in Section 5.5.2.9.

5.12.2.9.3 ECP-WD

ECP-WDM's physics datasets from the leadership-class computers are large, as described in earlier sections. Timely physics productivity is currently constrained by the slow data transfer speed from the computing facilities to PPPL/Users and data storage capacity at user facilities. As the exascale computers are used for science runs in a year from now, this issue will become more severe.

Current activities will increase PPPL's storage capacity by more than double to handle current and short term storage needs, but also redesign to support a highly scalable and fast environment that can be built out over time. PPPL will be standing up a new 3 PB storage system this fiscal year, which will be ready for localized and remote data transfer to and from PPPL at high speeds and for large datasets. This new storage infrastructure will support 100 Gbps+ locally, with a new 100 Gbps capable DTN included to support higher transfer speeds.

Long term would be to potentially stand up multiple DTNs, as needs arise. But also to stand up an ingress data storage level, to act as a fast caching tier, and then flow off to the general storage tier. This will help with leveraging faster external data pipes and make remote data available as fast as required per use case.

5.12.2.9.4 OMFIT

Network team is not currently seeing any data transfer performance issues. Link utilization is often well below thresholds for best practice. One exception to this is a set of tap links for network monitoring and security.

5.12.2.10 Outstanding Issues

There are no additional issues to report at this time.

5.12.2.11 Case Study Contributors

WDM & Other HPC Activities Representation

- CS Chang[9], PPPL
- David Green[10], ORNL
- Orso Meneghini[11], GA
- Paul Bonoli[12], MIT PSFC

ESnet Site Coordinator Committee Representation

- Susan Hicks[13], ORNL

9 cschang@pppl.gov
10 greendl1@ornl.gov
11 meneghini@fusion.gat.com
12 bonoli@psfc.mit.edu
13 hicksse@ornl.gov

- Scott Kampel[14], PPPL
- Jeff Nguyen[15], GA
- Brandon Savage[16], MIT PSFC

14 skampel@pppl.gov
15 nguyend@fusion.gat.com
16 bsavage@psfc.mit.edu

6 Focus Groups

A core component of the ESnet requirements review process that was displaced by the COVID-19 pandemic was the opportunity to hold impromptu conversations with colleagues. These could occur during the 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 June 2021, the FES 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 pre-defined 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 2021 FES requirements review.

During these focus sessions, the FES 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 FES 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 FES program area.

6.2 Organization

The FES requirements review featured 12 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 Tuesday, June 8.
- Focus group 2 was held on Wednesday, June 23.

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 FES requirements review team and meeting purpose started each, and then 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. Multi/Coupled Facilities Workflows: pairing FES experimentation with ASCR computing and storage via ESnet.
2. Supporting "Remote" Participation in FES: reviewing the current technology requirements, and ways these can be improved for domestic and international FES experimentation.
3. Future Networking — International Focus: transatlantic capacity to support ITER (expected rates and timelines), along with any changes to support transpacific needs (EAST, KSTAR).
4. Cybersecurity for Science Facilities: approaches to securing the infrastructure, yet still allowing for high-performance data sharing.
5. Data Access/Sharing Policy and Implementation: tradeoffs between security, 6. performance, and usability for managing FES data. Ways a holistic "data architecture" can be implemented (software, hardware, network, etc.).
6. Future Networking — Domestic Focus: capacity expectations and timelines to support networking at GA, MIT, PPPL, or other sites that are doing FES work.
7. Cloud Potpourri: computing and storage use cases, experimentation, interest, barriers to adoption.
8. Software and Computing Stack: current, near-term, and future needs and development opportunities; ways that other DOE SC locations leverage resources.
9. Storage Crisis: needs outpacing capabilities to generate and analyze.

A piece of polling software was utilized to gauge the relative interest in each topic area during the meeting. This was done to gain an understanding of what mattered to those who were represented in the room. The interest could be based on things they wanted to hear more about (potentially from other attendees), things they were concerned with implementing, or things they felt they could share experience with. Each focus group came to different conclusions about what topics mattered most, and as a result, each focus group's conversation flowed more naturally toward the strengths and weaknesses of those that attended.

6.3 Outcomes

The following sections highlight the areas of discussion and relevant findings and 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 Access/Sharing Policy and Implementation.
- Supporting "Remote" Participation in FES.
- Future Networking — International and Domestic Focus.
- Multi/Coupled Facilities Workflows.
- Data Access/Sharing Policy and Implementation.

During this period of discussion, several notable items were brought up.

6.3.1.1 Multi/Coupled Facilities Workflows

- The FES community is interested in exploring ways to make multi-facility workflows operate in a more routine and seamless manner in the future. Given the availability of computational and storage resources that far exceed local institutional capability, it makes operational sense to adapt to these resources provided that some expectations can be met. In particular, the ability to access dedicated resources, with strict time limits that are related to experimental operation, will be required to ensure success. Other factors, such as adapting current software tools to the environments, along with understanding the limitations and requirements related to cybersecurity, will be important.
- In addition to multi-facility operation, FES acknowledges that local computation is still an important part of the ecosystem. However, this should not always be viewed as the default mode of operation. Efforts to unify the software stack and computational resource allocation should result in allowing jobs to execute in locations with available resources, and ways to handle the resulting data flow across high-speed networks.

6.3.1.2 Supporting "Remote" Participation in FES

- **The FES community has a long history of remote collaboration and expects to continue this into the future. There are now several distinct patterns to remote use cases to consider:**
  - Remote observation: being able to observe aspects of a running FES experiment/instrument, typically through camera views or observable electronic diagnostics. Remote observation is, and will remain, common at many FES facilities. During the pandemic, this method was used around the world.
  - Remote participation: one can observe aspects of a running FES experiment/instrument similarly to remote observation, but remote participation adds the ability to communicate with local collaborators to influence direction of experimentation. Remote participation requires a closer relationship between participants, and this extra level of cooperation allows for a shared understanding of security considerations, along with goals for experimentation.

    - Remote control: uses the same considerations of both prior categories, but affords some level of control over the instrumentation during the experimental process. Remote control is currently uncommon due to the level of safety and security that is required to operate a FES facility/experiment.
- The FES community will continue to use remote use cases, provided that the technology support is in place to facilitate this. Considerations include fast networks with stable latency to support remote audio, video, chat, and diagnostics, along with accessible platforms that can be used by collaborators around the world.
- Changes to experimental behavior (e.g., lengthening shot time, or shortening time between shots) may complicate remote viewing in some cases, making the tools and networks more critical to the overall process.

6.3.1.3 Future Networking — International Focus

- ITER requirements are still being calculated in the run-up to first plasma at the site and continuous operations years after that. It is expected that connectivity to support ITER will far exceed the current capacity across the Atlantic that ESnet current supports. ESnet, DOE ASCR, and DOE FES will evaluate and provide solutions to the capacity and locality concerns in the coming years.
- Beyond capacity to support ITER, international connectivity is a strong driver for several FES experiments. A number of facilities in the Asia-Pacific region (EAST, KSTAR) as well as in Europe (W7-X) routinely use ESnet-maintained network peering to ensure operational success.

6.3.1.4 Data Access/Sharing Policy and Implementation

- FES is interested in exploring the concepts of data challenges in the run-up to ITER. This set of exercises will incrementally prepare participating facilities for the increase in data volumes by testing the hardware and software that support FES operation.
- The FES community, in collaboration with ESnet and ASCR HPC facilities, is recommended to consider working together to understand and support issues surrounding data sharing and dissemination from ITER. The effort is still early, and there is time to define the policy and technical implementation of a number of scientific workflows that will rely on ITER data sharing. Without a clear strategy, complications to seamless and efficient data access to this critical project could occur.

6.3.2 Focus Group 2

The polling during the meeting produced the following discussion topics that were of interest to the assembled group:

- Multi/Coupled Facilities Workflows.
- Supporting "Remote" Participation in FES.
- Future Networking — International Focus.

- Cybersecurity for Science Facilities.
- Data Access/Sharing Policy and Implementation .
- Future Networking — Domestic Focus.

During this period of discussion, several notable items were brought up.

6.3.2.1 Multi/Coupled Facilities Workflows

- The FES community remains interested in exploring approaches that will make multi-facility workflows more routine. Early investigations by GA and ALCF, as well as PPPL and NERSC, found a number of friction points that prevented regular and efficient execution of FES workflows. Some areas that can be improved in the future include:
  - Exploring ways to create a dedicated pool of compute resources that can be accessed without having to wait in a queue. FES analysis (which usually occurs between experimental shots) has a very limited time window of around 10–15 minutes. The results of analysis performed on a shot are often used to influence the next run of a given experiment. Thus the steps that contribute to analysis must be available rapidly.
  - The ability for worker nodes to directly access data streams that may be non-local, either through streaming, caching, or other emerging technologies (e.g., Slingshot interconnect[1]) is required.
  - Mechanisms to enable more direct technical support during operational periods. Due to the high-profile nature of the computation during experimentation, problems cannot be filed into a best effort trouble-ticket queue.
  - The ability to support system-wide scheduling; namely, ensuring that all components (computation, storage, networking, and software, at all portions of the end-to-end path) are ready when the analysis procedure starts.
- FES collaborators are interested in pursuing more multi-facility workflows, provided there is time to share requirements and evaluate their effectiveness. A set of pilot demonstrations is recommended, so the FES community becomes more familiar with the process and adopts the procedure as routine.

6.3.2.2 Supporting "Remote" Participation in FES

- A number of commercially available collaboration tools (audio, video, chat) are critical to the process of science for FES experiments and facilities. This trend started years prior to the pandemic, was crucial for ongoing operation during, and remains a part of operation into the future. Enabling these tools through ESnet's network peering relationships (directly, and via cloud providers) is important for ongoing collaboration and operation.
- Major FES facilities (e.g., GA, PPPL, MIT PFSC) have invested considerable resources into enabling remote participation environments. Typically, these considerations include ample ways to transmit and receive

1 https://www.hpe.com/us/en/compute/hpc/slingshot-interconnect.html

audio and video from remote facilities around the world (e.g., EAST, KSTAR, and eventually ITER), and collaborators that may be located domestically but unable to be in the same physical location. Upgrading domestic connectivity in the coming years to adapt to this continued remote participation will be required.

- Some remote collaboration tools work better than others in practice; this can be due to flaws in design or the age of the tool and how often it may receive updates. X Window System, VNC, NoMachine, and others allow for the ability to view, and occasionally control, remote resources, but only work well when information security, as well as network bandwidth and latency, can be controlled and guaranteed. Future remote observation and participation approaches will demand tools that offer similar feature sets.

6.3.2.3 Future Networking — International Focus

- ESnet connectivity remains critical for FES facilities, and backups and capacity augmentations will be required in the future years to ensure continuous operation.
  - GA has a 10 G WAN connection to ESnet, and a 1 G WAN backup connection through a commercial provider. Recent events, including a fiber cut, have severely affected the ability of GA to perform daily operations and upgrading the backup connection to support 10 G to ESnet is viewed as a critical requirement to science productivity.
  - MIT has a 1 G ESnet connection through the MIT campus, but is interested in upgrading due to increased use cases that rely on external connectivity to support remote computing and storage, as well as increased levels of remote observation use cases. Upgrading the ESnet connection implies working with the MIT campus to upgrade LAN and MAN connectivity.
  - PPPL has upgraded its local networking environment to accept a 100 G WAN connection from ESnet, and is interested in learning if a backup connection can also be acquired through diverse paths and providers.

6.3.2.4 Future Networking — Domestic Focus

- Preparing for ITER remains an important focus for the FES community. Current timelines indicate that the facility's first plasma will occur in December of 2025, with full operation expected by 2035. There will be periods of reduced operation that will occur between 2029 and 2032, and full operation is expected by 2035. The next two years are critical for planning how the US FES community will prepare for ITER and should focus on:
  - Identifying both the expected volumes of data that are possible and the expectations for being able to act on and handle activity bursts, during operational periods.
  - The FES community adopting a platform (e.g., software, computational hardware, storage) able to handle the data requirements locally and at distributed facilities.
  - ESnet implementing network connectivity between the US and European partners to address the data mobility requirements.

- Putting in place a timeline for data challenges that can exercise the entire ecosystem of the data architecture by simulating the volume and timing requirements using the operational tools.
- Identifying any bottlenecks that may exist facility to facility, as well as what the user population may experience.

6.3.2.5 Data Access/Sharing Policy and Implementation

- The policies and approaches surrounding data formats and sharing, particularly as FES prepares for ITER, remain a high priority to normalize. ITER will produce an unprecedented amount of data that will be of interest to the US FES community, and having access to that data in a timely manner, and a uniform set of tools to operate and analyze the results, is critical to advancing research and development activities.
- Heterogeneous data formats are problematic for the FES community and create a lot of work to support and adapt software that can be used at a variety of experimental facilities. The adoption of IMAS/IDS for data representation is an important first step to unifying data formats. It is recommended that the FES community look to this as a future goal to standardize data from existing and future instruments, to unify the way that software and workflow can be implemented to address analysis.
- FES could benefit from the creation of "data hubs" specialized to services that are needed in the FES community. These data hubs would feature dedicated computing and storage resources and common software tools, and would be designed to handle FES formatted data sets. The end goal would be to establish pipelines in and out of these facilities, which would allow collaborators a level playing field to interact with the science.

6.3.2.6 Software and Computing Stack

- Software licensure, as well as import/export controls, can complicate scientific workflows, particularly if approaches that are designed for single user/machine use cases are adapted to shared environments such as an HPC facility. For example, a user of a shared resource often does not have the administrative rights to install and operate software that may require these permissions. This can prevent critical software from being run on resources that would accelerate the workflow, and prevent productivity for the process of science.

6.3.2.7 Cybersecurity for Science Facilities

- Implementation of cybersecurity requirements can occasionally affect the performance of open scientific workflows that rely on data mobility via networks. The FES and ASCR communities must work to understand these impacts, and recommend appropriate mitigations and strategies to afford compliance and protection, without affecting performance. ESnet's Science DMZ approach to network perimeter implementation is a part of this approach, and is recommended for FES facilities and experiments.

Appendix A: International Connectivity

Throughout the 2021 FES requirements review process, the case for international networking needs has come to the forefront to support nearly every case study for aspects of the workflow. These needs can be categorized as follows:

- **Experimental source located apart from analysis facilities.** Scientific instruments, such as tokamaks, computing resources, etc., have a single source, and often rely on an analysis activities that are physically separated. Global collaboration often means that international networks are a critical part of the process of science.
- **Inter-collaboration information sharing.** Other portions of the scientific workflow (distributed analysis on intermediate formats, production of simulation data, backups, etc.) may involve international collaborators.
- **User-level data sharing.** Users of scientific data are worldwide and are not always known a priori.

The following sections will highlight specific findings from the review, along with supplemental information on international connectivity from the R&E community. Some of the links are funded via the DOE (e.g., ESnet); others come from the NSF and foreign collaborators (e.g., GÉANT, Rede Nacional de Ensino e Pesquisa [RNP], NORDUnet, etc.).

A.1 Current State and Near-Term Plans for the International R&E Circuits

International connectivity for the R&E community is provided by a number of different providers and funding sources, and is delivered through several exchange points located around the country. These facilities feature connectivity to domestic R&E and commercial carriers, which link many of the FES facilities.

A.1.1 Domestic Exchange Points

There are a number of domestically located exchange points where network providers establish peering with each other. This fabric of connectivity allows for a seamless transfer of scientific network traffic between cooperating providers:

- **MANLAN:** New York, New York[1].
- **WIX:** Washington, DC.
- **Starlight:** Chicago, Illinois[2].
- **Pacific Wave:** Los Angeles, California and Seattle, Washington[3]
- **AMPATH**: Miami, Florida[4].

ESnet maintains connectivity to these locations, as well as peering with providers that are present, to ensure that traffic can reach critical international locations

1 https://internet2.edu/network/global-networks-and-partnerships/man-lan-new-york-and-wix-virginia-exchange-points
2 http://www.startap.net/starlight
3 http://pacificwave.net
4 https://ampath.net

A.1.2 Transatlantic Networking

As of October 2021, there were 10 100 G circuits, providing an aggregate of 1 T of R&E capacity, between the United States and Europe as shown in Figure A.1. These links are supported by the DOE, NSF, Internet2[5], CANARIE (Canadian National Research and Education Network [NREN])[6], GÉANT (European NREN)[7], SURF (Dutch NREN)[8], and NORDUnet (Nordic NREN)[9]. During the third quarter of 2021, these links averaged 17.67 Gbps and transferred over 174 PB of data. Many of these networks collaborate regularly through established consortia[10 11].

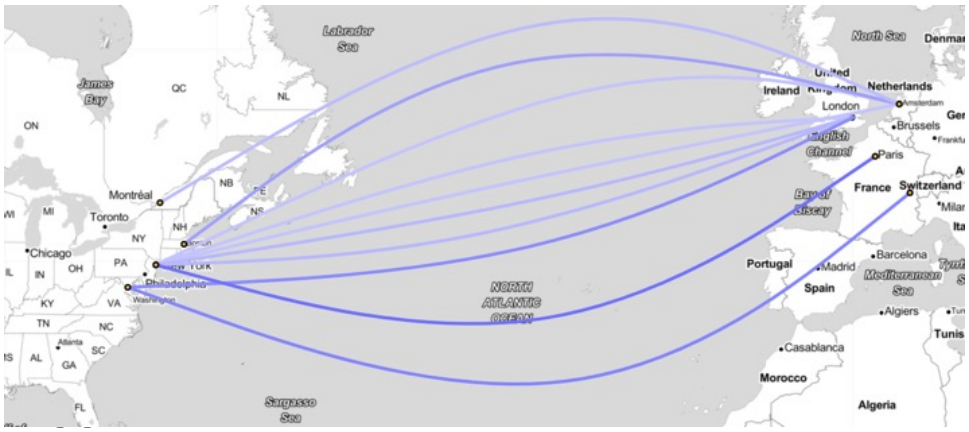


Figure A.1 – Current R&E networks between the US and Europe. Data available live at http://ana.netsage.global.

A.1.3 Transpacific Networking

In Asia, the Asia Pacific Ring Consortium jointly supports connectivity (shown in Figure A.2) for roughly 400 G of capacity between the US and Asia as well as 10–20 G between Guam and Singapore, and Guam and Hong Kong. In late 2020, the SingAREN/Internet2 link between Singapore and Los Angeles was replaced by a SingAREN-managed circuit that runs between Singapore, to Tokyo, and then to Los Angeles (on a different cable than the Science Information Network [SINET] Tokyo-LA capacity)[12 13]. In 2021, the path between Guam and Singapore was upgraded to 100 G. Depending on Federal Communications Commission regulators, the Guam–Hong Kong and Sydney–Hong Kong paths may be upgraded to 100 G in 2022 as well. Currently, these links are underutilized, but the diversity of paths is needed for redundancy and resilience in the earthquake and tsunami-prone Ring of Fire region.

5 https://internet2.edu/network/global-networks-and-partnerships
6 https://www.canarie.ca/about-us
7 https://www.geant.org/Networks
8 https://www.surf.nl/en
9 https://www.nordu.net
10 https://internet2.edu/network/global-networks-and-partnerships/advanced-north-atlantic-ana
11 https://gna-re.net
12 https://www.singaren.net.sg
13 https://www.sinet.ad.jp/en/aboutsinet-en

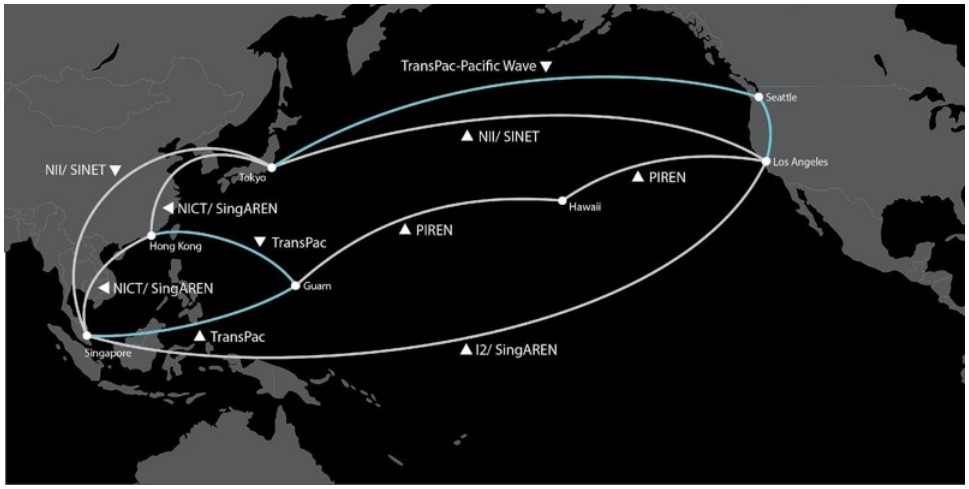


Figure A.2 – Current R&E networks between the US and Asia. Data available live at http://aponet.netsage.global.

A.1.4 South American Networking

Between the US and South America, R&E networking is primarily supported via an NSF International Research and education Network Connections (IRNC) award to Julio Ibarra entitled "Americas-Africa Lightpaths Express and Protect (AmLight-ExP)."[14] Figure A.3 shows the current (2022) network map, consisting of 600 Gbps of upstream capacity between the US and Latin America, and 100 Gbps to Africa. Overall, it is possible to leverage a total of more than 2 Tbps of international connectivity using the *AmLight Express* (green line), *AmLight Protected* (white line), plus waves provided by *RedClara*[15] and *SANReN*[16] and *TENET*[17] (pink line). These connections are delivered via points of presence (PoPs) in Florida, Brazil, Chile, Puerto Rico, Panama, and South Africa. Future plans consist of adding 200 Gbps from Sao Paulo to Boca Raton in 2023, and adding an additional PoP in Atlanta Georgia.

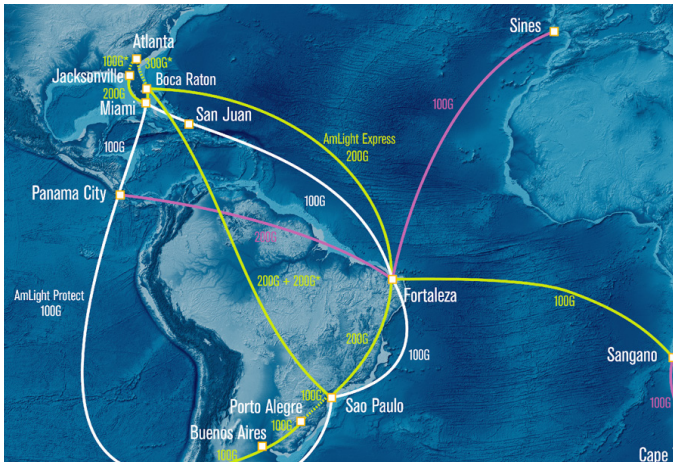


Figure A.3 – R&E networks between the United States and South America

14 https://ampath.net
15 https://www.redclara.net/index.php/en/
16 https://www.sanren.ac.za
17 https://www.tenet.ac.za

A.2 Case Study Findings

A.2.1 International Fusion Collaborations

A number of fusion collaborators that leverage instruments are located internationally. Some of these are located in Europe, and leverage transatlantic connectivity (described in Section A.1.2), and other are located in Asia and leverage transpacific connectivity (described in Section A.1.3).

The European collaborations are as follows:

- The Wendelstein 7-X (W7-X) stellarator in Greifswald, Germany.
- AUG at the Max-Planck-Institut für Plasmaphysik (IPP) in Garching, Germany.
- JET and MAST at the Culham Science Centre, in Abingdon, United Kingdom.
- ITER (formerly the International Thermonuclear Experimental
- Reactor) in Cadarache, France.
- WEST, whose name is derived from Tungsten (e.g., the chemical symbol "W") Environment in Steady-state Tokamak, and formerly referred to as Tore Supra, in Cadarache, France.
- TCV at EPFL in Lausanne, Switzerland.

All of these facilities utilize the GÉANT[18], the pan-European data network for the research and education community, to access ESnet and Internet2 in the United States. Each will connect to NRENs in their respective countries to reach GÉANT:

- Germany uses DFN[19].
- The UK uses JANET[20].
- France uses RENATER[21].
- Switzerland uses SWITCH[22].

The Asia/Pacific collaborations are as follows:

- KSTAR in Dejeon, South Korea.
- EAST at ASIPP, Hefei, China.
- Steady State Tokamak (SST-1) at the IPR in Bhat, India.
- The LHD at the National Institute of Fusion Science, Toki, Japan.
- The Japan Torus-60 (JT60-SA) in Naka, Japan.

All of these facilities utilize APAN[23] (Asia Pacific Advanced Network) to leverage international links and peering to reach ESnet and Internet2 in the US. Each will connect to NRENs in their respective countries to reach APAN:

18 https://geant.org
19 https://dfn.de/en/
20 https://www.jisc.ac.uk/janet
21 https://www.renater.fr
22 https://www.switch.ch
23 https://apan.net

- Japan uses SINET[24]
- South Korea uses KREONET[25].
- China uses both the China Education and Research Network (CERNET[26]) and the China Science and Technology Network (CSTNET[27]).
- India uses the National Research and Education Network of India (ERNET India[28]).

At this time, no significant upgrades in connectivity or peering are anticipated to support the needs of the case studies presented in this review.

A.2.2 Remote Observation and Participation of Fusion Facilities

The majority of remote participation use cases are expected to utilize domestic connectivity to support viewing domestic instruments, via domestic users. The anticipated international use instrumentation and operator use cases are:

- KSTAR in Dejeon, South Korea, and NSTX-U at PPPL in the US.
- EAST in Hefei, China, and DIII-D at GA in the US.
- ITER in Cadarache, France, and the potential for several sites in the US.

As described in Section A.2.1, ample international connectivity is available to support these use cases, and at this time no significant upgrades in connectivity or peering are anticipated to support the needs of the case studies presented in this review.

A.2.3 GA: DIII-D National Fusion Facility

As described in Sections A.2.1 and A.2.2, GA serves as both an instrument source (e.g., DIII-D) in an active collaboration with the staff at EAST in Hefei, China, as well as operating EAST remotely. Ample international connectivity is available to support these use cases, and at this time no significant upgrades in connectivity or peering are anticipated to support the needs of the case studies presented in this review.

It is also likely that other users based internationally could download scientific data, but a fine-grained analysis has not been performed to give specific examples.

A.2.4 MIT PSFC

MIT PSFC is not actively engaged in any international collaborations to remotely operate instrumentation. It is likely that users based internationally could download scientific data (e.g., Alcator C-Mod), but a fine-grained analysis has not been performed to give specific examples.

A.2.5 PPPL

As described in Sections A.2.1 and A.2.2, PPPL serves as both an instrument source (e.g., NSTX-U) in an active collaboration with the staff at KSTAR in Dejeon, South Korea, as well as operating KSTAR remotely. Ample international connectivity is available to support these use cases, and at this time no significant upgrades in

24 https://www.sinet.ad.jp/en/top-en
25 https://www.kreonet.net
26 https://www.edu.cn/english/
27 http://www.cstnet.net.cn
28 https://ernet.in

connectivity or peering are anticipated to support the needs of the case studies presented in this review.

It is also likely that other users based internationally could download scientific data, but a fine-grained analysis has not been performed to give specific examples.

A.2.6 ITER (Initially the International Thermonuclear Experimental Reactor)

As described in Sections A.2.1, ITER (when constructed) will be served by the GÉANT backbone network, and the RENATER NREN, within Europe. During the initial design and testing of ITER tools and capabilities, existing transatlantic connectivity options will be used to evaluate workflows to US-based collaborators. The data distribution patterns of ITER are currently still being decided, which will influence the frequency and volume of data destined for US-based locations, making overall volume expectations hard to predict.

It is anticipated that the DOE will work closely with ITER operations to augment connectivity in the coming years to anticipate ITER data needs.

A.2.7 Public-Private Partnerships in Fusion Research

A number of entities funded by INFUSE could be located internationally, but at the time of this writing the case study subjects do not leverage any international locations for the production of data, use of simulations, or pressing of results.

It is likely that use cases for this program can be located, and leverage international resources, but a fine-grained analysis has not been performed to give specific examples.

A.2.8 MPEX at ORNL

The case study does not leverage any international locations for the production of data, use of simulations, or pressing of results; all data products are produced domestically at DOE facilities.

It is likely that users based internationally could download scientific data, but a fine-grained analysis has not been performed to give specific examples.

A.2.9 MEC Experiment at SLAC

The case study does not leverage any international locations for the production of data, use of simulations, or pressing of results; all data products are produced domestically at DOE facilities.

It is likely that users based internationally could download scientific data, but a fine-grained analysis has not been performed to give specific examples.

A.2.10 LaserNetUS Program

The majority of LaserNetUS resources are located domestically, but a single facility in the partnership is located in Québec, Canada: the INRS ALLS. Users accessing this facility from US-based locations will use both the CANARIE network, the national R&E provider for Canada, and ORION, the regional R&E provider for Québec. ESnet and Internet2 in the US both connect to CANARIE in multiple locations, facilitating high-bandwidth capabilities.

It is likely that users based internationally could download scientific data from the US-based instruments, but a fine-grained analysis has not been performed to give specific examples.

A.2.11 Multi-Facility FES Workflows

As described in Sections A.2.1, A.2.2, A.2.3, and A.2.5, a number of use cases will leverage an international instrument and domestic collaborators and computing resources. Ample international connectivity is available to support these use cases, and at this time no significant upgrades in connectivity or peering are anticipated to support the needs of the case studies presented in this review.

A.2.12 WDM and FES HPC Activities

As described in Sections A.2.1, A.2.2, A.2.3, and A.2.5, ample international connectivity is available to support the WDM use case. At this time, no significant upgrades in connectivity or peering are anticipated to support the needs of the case studies presented in this review.

Appendix B: DOE HPC Facilities and Networking

The DOE SC operates a number of HPC and high-performance networking (HPN) facilities to meet critical mission needs.

B.1 HPC Facilities

ASCR operates three HPC user facilities:

- NERSC at LBNL, which provides HPC resources and large-scale storage to a broad range of SC researchers.
- Two LCFs, which provide leading-edge HPC capability to the US research and industrial communities:
  - OLCF at ORNL.
  - ALCF at ANL.

ASCR facilities couple computing resources with large-scale, state-of-the-art storage, networking and software tools essential for computational scientific research. Additionally, system software, communications, math libraries, and applications must scale to meet the extreme size of the facilities. ASCR computational facilities are accessible to researchers through an open peer-review process. Once permitted access, users schedule time, run experiments, analyze results, and interface with facility support staff that aid researchers in the scientific discovery process. HPC is critical to advancing scientific discovery.

B.2 HPN Facilities

ASCR's high-performance scientific network facility, ESnet, is among the fastest in the world and is dedicated to making scientific progress completely unconstrained by the physical location of instruments, people, computational resources, or data. The network delivers unparalleled infrastructure, capability, and tools uniquely designed to address the special needs of scientific data movement. As science grows, it continues to push the limits of information communication and drives innovations and development for future high-performance scientific networks. ESnet is connected to the three major computational facilities with multiple 100 Gbps connections, with plans to upgrade to 400 Gbps after ESnet6 is fully commissioned, and the sites can accept connection upgrades, in 2022.

B.3 LAN and WAN Block Diagrams

The network architecture of the major HPC facilities changes frequently, but follows general design patterns:

- Data ingest mechanism, consisting of dedicated resources (e.g., DTNs) that can be used to send or receive scientific data between the facility and collaborators.
- A secured perimeter to protect internal resources.

- Fast file systems that link the data transfer resources and the computational resources.

Figure B.1 depicts the WAN architecture, which typically consists of multiple connections on the border (to both ESnet and other network providers) that are then distributed to data transfer resources and the rest of the internal network. Other factors (e.g., firewalls, gateways, etc.) exist for other affiliated functionality.

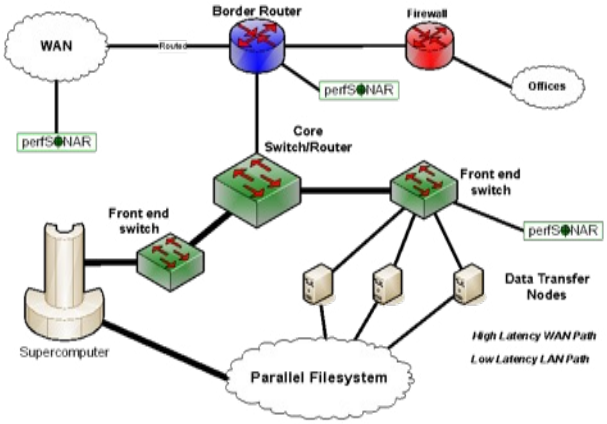


Figure B.1 – Block Diagram of Typical HPC Facility WAN Infrastructure

Figure B.2 shows a LAN-level view of a typical HPC facility, focusing on the interconnection fabric between major HPC components, and how they interact with the WAN-facing services.

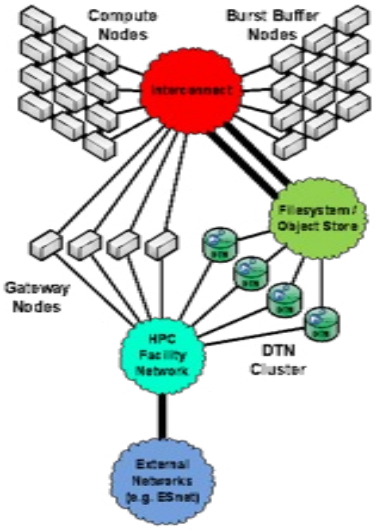


Figure B.2 – Block Diagram of Typical HPC Facility LAN Infrastructure

List of Abbreviations

ADIOS . . . . Adaptable I/O Systems
AI . . . . . artificial intelligence
ALCF . . . . . Argonne Leadership Computing Facility
AMI. . . . . analysis monitoring interface
ANL . . . . . Argonne National Laboratory
ARC . . . . . ARC (affordable, robust, compact) fusion reactor
ASCR . . . . . Advanced Scientific Computing Research
ASIPP . . . . . Academy of Sciences, Institute of Plasma Physics
ATO. . . . . authorization to operate
AUG . . . . . ASDEX Upgrade
AWS . . . . . Amazon Web Services
BCP. . . . . best common practices
BGP. . . . . Border Gateway Protocol
BI . . . . . Bayesian inference
CAD . . . . . computer-aided design
CCD . . . . . charge-coupled device
CFS. . . . . Commonwealth Fusion Systems
CPU . . . . . *central processing unit*
CSTAR . . . . Center for Science and Technology with Accelerators and Radiation
DAQ . . . . . data acquisition
DELTA . . . . aDaptive rEaL Time Analysis of big fusion data
DME. . . . . Data Mobility Exhibition
DOE . . . . . Department of Energy
DRP . . . . . Data Reduction Pipeline
DTN . . . . . Data Transfer Node
EAST . . . . . Experimental Advanced Superconducting Tokamak
ECEI . . . . . electron cyclotron emission imaging
ECH . . . . . electron cyclotron heating
ECP . . . . . Exascale Computing Project
ECS. . . . . ESnet Collaboration Services
EFDA . . . . . European Fusion Development Agreement
ELLA . . . . . Europe and Latin America
ELM . . . . . edge localized mode
ESCC . . . . . ESnet Site Coordinators Committee
FASP. . . . . Fast and Secure Protocol
FDR . . . . . Fourteen Data Rate
FES. . . . . Fusion Energy Sciences
FFB. . . . . fast feedback
FPGA . . . . . field programmable gate arrays
FRC . . . . . field-reversed configuration
FTP. . . . . *File Transfer Protocol*
GA. . . . . General Atomics
GAIN . . . . . Gateway for Accelerated Innovation in Nuclear
GCP . . . . . Google Cloud Platform
GÉANT. . . . GEometry ANd Tracking
GENE. . . . . Gyrokinetic Electromagnetic Numerical Experiment
GPI . . . . . gas puff imaging
GPU . . . . . *graphics processing unit*
GUI. . . . . graphical user interface
HED. . . . . high energy density
HEDP. . . . . high energy density physics
HEDS. . . . . high energy density science
HEP . . . . . high-energy physics
HPC . . . . . high-performance computing
HPN . . . . . high-performance networking
HPSS . . . . . High-Performance Storage System
HTC . . . . . high-throughput computing
ICRF. . . . . Ion cyclotron range of frequency
ICRH . . . . . Ion Cyclotron Resonance Heating
IDL. . . . . Interactive Data Language
IDS . . . . . interface data structure
IEA . . . . . International Energy Agency
IMAS . . . . . Integrated Modeling and Analysis Suite
IMEG . . . . . Integrated Modeling Expert Group
INCITE . . . Innovative and Novel

	Computational Impact on Theory and Experiment
INFUSE	Innovation Network for Fusion Energy program
IO	ITER organization
IP	Internet protocol
IPP	Institute for Plasma Physics
IPR	Institute for Plasma Research
IR	infrared
IT	information technology
ITER	International Thermonuclear Experimental Reactor
ITPA	International Tokamak Physics Activity
JAEA	Japan Atomic Energy Agency
JBOD	just a bunch of disks
JET	Joint European Torus
KNL	Knight's Landing
KSTAR	Korean Superconducting Tokamak Advanced Reactor
LAN	*local area network*
LBNL	Lawrence Berkeley National Laboratory
LCF	Leadership Computing Facilities
LCLS	Linac Coherent Light Source
LH	lower hybrid
LHC	Large Hadron Collider
LHD	Large Helical Device
LLAMA	Lyman-alpha Measurement Apparatus
LLE	Laboratory of Laser Energetics
MAGPI	Mid-Atlantic Gigapop in Philadelphia for Internet2
MAN	metro area network
MAST	Mega Ampere Spherical Tokamak
MB	megabyte
MDS	MIT/Model Data System
MEC	Matter in Extreme Conditions
MFE	Magnetic Fusion Energy
MFEM	Modular Finite Element Methods
MGHPCC	Massachusetts Green High-Performance Computing Center
MHD	extended-magnetohydrodynamic
ML	machine learning
MPEX	Material Plasma Exposure eXperiment
MPI	message-passing interface
NAC	network access control
NERSC	National Energy Research Scientific Computing Center
NETCDF	Network Common Data Form
NFS	Network File System
NoSQL	not only SQL
NoX	Northern Crossroads
NREN	National Research and Education Network
NSF	National Science Foundation
NTP	Network Time Protocol
NYC	New York City
OLCF	Oak Ridge Leadership Computing Facility
OMFIT	One Modeling Framework for Integrated Tasks
ORNL	Oak Ridge National Laboratory
OSU	Ohio State University
PAC	Program Advisory Committee
PB	petabyte
PCS	plasma control system
PI	principal investigator
PIC	particle-in-cell
POC	point of contacts
PPIC	Princeton Plasma Innovation Center
PPPL	Princeton Plasma Physics Laboratory
PRP	proposal review panel
PRTG	Paessler Router Traffic Grapher
PSFC	Plasma Science and Fusion Center

QoS . . . . . *quality of service*
R&E . . . . . research and education
RAID . . . . . redundant array of independent disks
RAPIDS . . . SciDAC Institute for Computer Science and Data
RCF . . . . . radiochromic film
RCR . . . . . remote control room
RF . . . . . radio frequency
SC . . . . . DOE Office of Science
SciDAC. . . . Scientific Discovery Through Advanced Computing
SCP. . . . . secure copy protocol
SDC . . . . . ScaleIO Data Client
SINET . . . . Science Information Network
SSH. . . . . Secure Shell
TB . . . . . terabyte
TCV. . . . . Tokamak à Configuration Variable
TDRS . . . . . Tracking and Data Relay Satellite
TDRSS . . . . Tracking and Data Relay Satellite System
TRANSP. . . Transport Solver
UC. . . . . University of California
VC. . . . . video conference
VISAR. . . . . Velocity Interferometer System for Any Reflector
VLAN . . . . . Virtual LAN
VM . . . . . virtual machine
VO. . . . . virtual organization
VPN. . . . . *virtual private network*
WAN. . . . . wide-area network
WDM . . . . . Whole-Device Modeling
WEST. . . . . W Environment in Steady-state Tokamak
XCG . . . . . X-point Gyrokinetic Code
XFEL . . . . . X-ray free electron laser
XGC . . . . . X-Point Included Gyrokinetic Code
ZFS . . . . . Zettabyte file system