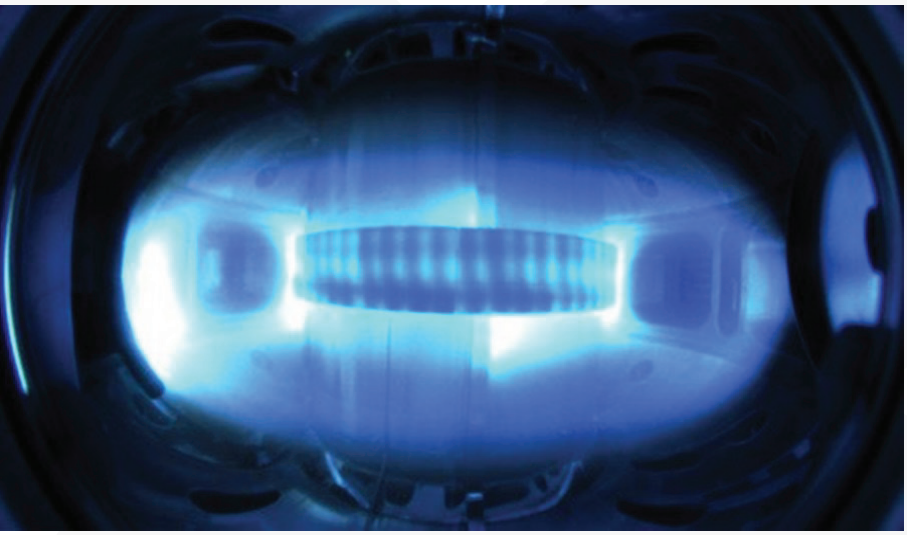
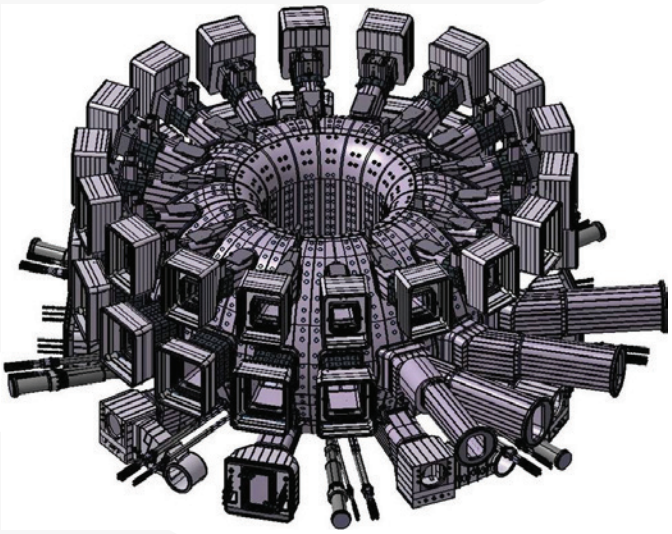
ESnet
ENERGY SCIENCES NETWORK

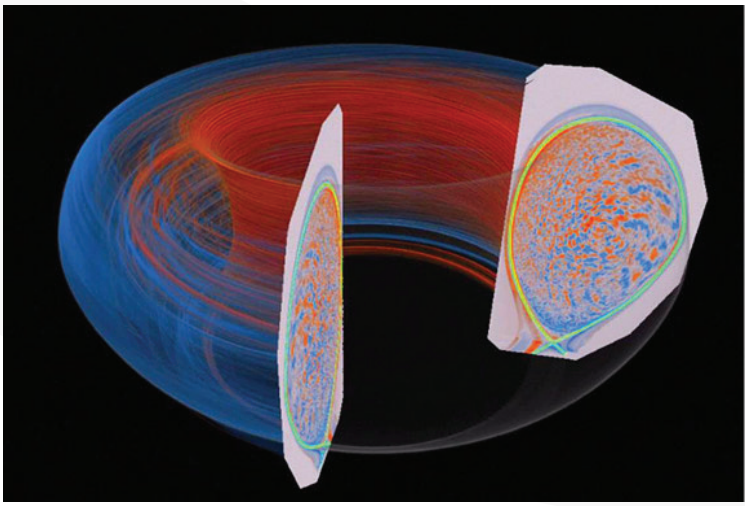
Fusion Energy Sciences Network Requirements Review

Final Report

August 12-13, 2014







BERKELEY LAB

U.S. DEPARTMENT OF ENERGY
Office of Science

Disclaimer

Fusion Energy Sciences Network Requirements Review Final Report

Office of Fusion Energy Sciences, DOE Office of Science
Energy Sciences Network (ESnet)
Germantown, Maryland
August 12–13, 2014

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

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

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

This is LBNL report LBNL-6975E.

Contents

Participants and Contributors

Choong-Seock Chang, PPPL (Fusion Simulations)

Dan Ciarlette, ORNL (ITER)

Eli Dart, ESnet (Science Engagement, Review Chair)

Vince Dattoria, DOE/SC/ASCR (ESnet Program Manager)

Michael Ernst, BNL (LHC experiments)

Paul Henderson, PPPL (PPPL Networking)

Mary Hester, ESnet (Science Engagement)

Steve Jardin, PPPL (TRANSP)

Stanley Kaye, PPPL (TRANSP)

Scott Klasky, ORNL (Simulations)

Randall Laviolette, DOE/SC/ASCR (SciDAC Partnerships)

John Mandrekas, DOE/SC/FES (FES Program)

David Schissel, General Atomics (DIII-D, International Collaborations)

Josh Stillerman, MIT/PSFC (Alcator C-Mod facility, International Collaborations)

Kevin Thompson, NSF (International Networking)

Jason Zurawski, ESnet (Science Engagement)

Report Editors

Eli Dart, ESnet: dart@es.net

Mary Hester, ESnet: mchester@es.net

Jason Zurawski, ESnet: zurawski@es.net

Executive Summary

The Energy Sciences Network (ESnet) is the primary provider of network connectivity for the US Department of Energy Office of Science (SC), the single largest supporter of basic research in the physical sciences in the United States. In support of the Office of Science programs, ESnet regularly updates and refreshes its understanding of the networking requirements of the instruments, facilities, scientists, and science programs that it serves. This focus has helped ESnet to be a highly successful enabler of scientific discovery for over 25 years.

In August 2014, ESnet and the Office of Fusion Energy Sciences (FES), of the DOE Office of Science, organized a review to characterize the networking requirements of the programs funded by the FES program office.

Several key findings resulted from the review. Among them:

1. During an experiment run, it is important to get actionable results from the approximately 15-minute inter-shot analysis quickly so that the analysis results can be used in the setup for the next shot. This places significant demands on the data transfer, data analysis, and collaboration systems used during the experiment. During a run, getting actionable results quickly is far more important than getting detailed results after a period of hours.
2. While the primary focus during experiments is on inter-shot analysis, it would be very useful to the experiment team to be able to run simulations at a remote HPC facility that could inform the later portion of an experimental run. Even though the turnaround time would be too great for inter-shot analysis, this capability would be very useful.
3. The International Research Network Connections (IRNC) program, funded by the National Science Foundation, is key to the collaborations between US and Asian sites. The IRNC program has funded and continues to fund trans-Pacific network connectivity devoted to science, which is a significant contributor to the success of these collaborations, including collaborations between US institutions and the EAST facility in China and the KSTAR facility in South Korea.
4. Collaboration tools (audio conferencing, video conferencing, screen sharing, instant messaging, etc.) are heavily used by the experimental fusion community, and are critical to the operation of the experiments. Remote participation in experiments occurs routinely, and collaboration technology is critical for effective remote participation.
5. The design of the data management systems for ITER is not yet complete, and is not expected to be complete for some time (possibly 5 years or more).

This report expands on these points, and addresses others as well. The report contains a findings section in addition to the text of the case studies discussed during the review.

Findings

Below are the findings for the FES and ESnet Requirements Review held August 12–13, 2014. These points summarize important information gathered during the 2014 FES–ESnet review.

- An experimental run is composed of multiple pulses lasting anywhere from a few seconds to many tens of seconds (some facilities are moving towards pulses lasting hundreds of seconds). Each pulse is called a "shot." Data are collected during the shot, and analyzed in the approximately 15-minute interval between shots. During an experiment run, it is important to get actionable results from the inter-shot analysis quickly so that the analysis results can be used in the setup for the next shot. This places significant demands on the data transfer, data analysis, and collaboration systems used during the experiment. During a run, getting actionable results quickly is far more important than getting detailed results after a period of hours. Data reprocessing is done overnight using more detailed analysis, but these results are used in post-experiment analysis rather than during the running of the experiment.
- While the primary focus during experiments is on inter-shot analysis, it would be very useful for the experiment team to be able to run simulations at a remote HPC facility. For example, running large-scale simulations concurrently on HPC resources during an experiment can help answer additional questions or verify experimental results. Even though actionable results come back in an hour or two, it would be very helpful to the experiment team (even though the turnaround time would not make the simulation run suitable for inter-shot analysis). In order to accomplish this, appropriate job queues would need to be available at HPC facilities, reliable high-performance data transfer would be needed, etc.
- The MDSplus software package is widely used in the experimental fusion community. It is used by Alcator C-Mod, DIII-D, and many other experimental fusion facilities (about 40 in total). MDSplus is largely transaction-based, which results in significant performance limitations when used over the wide area network (long distances increase the delay between transactions significantly, and MDSplus does not pipeline transactions). Some discussions took place at the review regarding enhancements that might be made to MDSplus to increase performance over long-distance networks.
- There is an ongoing discussion between DIII-D at General Atomics (GA) and EAST at the Chinese Academy of Sciences Institute for Plasma Physics (ASIPP) regarding the remote operation of a third shift of experiments at EAST by scientists at DIII-D. This would be beneficial because the third shift at EAST could occur during normal working hours on the US West Coast (where DIII-D is located). The viability of running remote experiments from GA has already been successfully demonstrated. This would increase the network activity between the two sites, and would also increase the reliance on the network for experiments.
- The National Science Foundation's IRNC program is key to enabling collaborations between US and Asian sites. The IRNC program has funded and continues to fund trans-Pacific network connectivity devoted to science, which is a significant contributor to the success of these collaborations, including collaborations between US institutions and the EAST facility in China and the KSTAR facility in South Korea.
- The experimental fusion sites are investigating the use of Amazon S3 to store large data ob-

jects. The data egress charges currently make the use of Amazon unattractive. If the data egress charges were to change significantly, that might affect the utility of Amazon's S3 service for the fusion community.

- There was a presentation on and subsequent discussion of the data movement and data distribution tools, architecture, planning, and deployment for the LHC experiments. Several aspects of the experience gained by the LHC experiments are likely to be useful in the design, deployment, commissioning, and operation of the data movement and analysis tools used to support ITER.
- Cloud services are being investigated for mid-range cluster computing (clusters with several hundred cores). Currently, however, it is less expensive for those sites to provide mid-range computing services in-house.
- The use of high-speed cameras is increasing at the experimental facilities. These generate significantly more data than the diagnostics that have been used in the past. At some sites high-speed camera data is the dominant consumer of storage, and dedicated network links are installed to carry this data within the site.
- The data rate at Alcator C-Mod doubles every 2.2 years. This has been consistent for the past 20 years.
- PPPL has deployed Globus, and has seen significant benefits from doing so. The performance of Globus is significantly higher than the previous tool (SCP), with individual users typically seeing performance improve by a factor of 100 times to 200 times. In one example case, Globus was 184 times faster than SCP.
- In partnership with ESnet, PPPL has made progress on designing and deploying a Science DMZ. Once the Science DMZ is in production, Globus data transfer nodes (DTNs) will be moved into the Science DMZ, which is expected to further enhance the data transfer capabilities at PPPL.
- There is a need to transfer simulation files from the ALCF to PPPL because of analysis capabilities present at PPPL that are not present at the ALCF. By deploying Globus and a Science DMZ at PPPL, data transfers between PPPL and ALCF are expected to improve.
- The perfSONAR suite of network test and measurement tools could provide significant benefit to some experimental fusion sites. In particular, the OWAMP test tool would be of use for early identification of LAN problems that might affect system performance during experiments.
- Collaboration tools (audio conferencing, video conferencing, screen sharing, instant messaging, etc.) are heavily used by the experimental fusion community, and are critical to the operation of the experiments. Remote participation in experiments occurs routinely, and collaboration technology is critical for effective remote participation.
- The discontinuation of the ESnet Collaboration Services (ECS) audio and video conferencing services is necessitating some painful adjustments in the fusion community. In the near term, the different fusion sites must all evaluate, procure, and deploy replacements for the services previously offered by ESnet as part of ECS. In the medium term, it is likely that the clients for multiple services will have to be supported by multiple sites because different sites are likely to deploy different systems. Some sort of GDS functionality will need to be supported by the sites, because the European collaborations continue to use GDS and interoperability with the European systems will be increasingly more important with the ITER project.
- Several international sites or facilities that might benefit from engagement by ESnet were identified at the meeting. These include W7-X in Germany, LHD in Japan, KSTAR in South Korea, and EAST in China.
- Both EAST and KSTAR are working toward much longer shot times (on the order of 300 seconds). In addition to increasing the volume of experimental data, this will affect the ways in which data must be transferred to other sites for productive intra-shot and inter-shot analysis.
- The International Collaboration Framework for Extreme-scale Experiments (ICEE) project is working on tools to facilitate the analysis of experiment data using HPC resources without requiring

file I/O (rather, RDMA is used). This will require support from the end sites (both the experimental facility and the HPC facility) as well as the networks that connect the experiment and HPC facilities for virtual circuit services to provide traffic isolation, bandwidth guarantees, and quality of service. ESnet's On-Demand Secure Circuits and Advance Reservation System (OSCARS) service is an ideal platform for providing these services in support of ICEE.

- An opportunity exists for collaboration between ESnet and ORNL on software-defined networking (SDN) and related technologies.
- The fusion community is a heavy user of NoMachine (NX) for remote access and remote visualization. In many cases, the use of NX eliminates the need to transfer data to a home institution for analysis.
- One of the scientists at MIT is a heavy user of NERSC and has a large volume of data stored at NERSC. As part of the MPO project, the user may require assistance in moving the data back to MIT for annotation. This activity is expected in the coming months.
- The design of the data management systems for ITER is not yet complete, and is not expected to be complete for some time (possibly five years or more).

Action Items

Several action items for ESnet came out of this review. These include:

- ESnet will put the MDSplus development team in touch with the developers of the SPADE data movement tool which might be useful for improving MDSplus performance over long-distance networks.
- ESnet will work with MIT/PSFC on a Science DMZ design that addresses tradeoffs between packet filtering policy and firewall appliances, and also incorporates IPv6 capabilities.
- ESnet will work with MIT/PSFC to assist with the design of a perfSONAR deployment.
- ESnet will work with members of the international collaboration involving the KSTAR facility in South Korea to help improve the networking and data transfer infrastructure there.
- ESnet will continue to work with the ICEE project to support the project's needs for bandwidth guarantees, quality of service, and virtual circuits.
- ESnet will continue to work with GA/DIII-D and PPPL on Science DMZ, perfSONAR, and related technologies in support of increased performance.
- ESnet will continue to develop and update the http://fasterdata.es.net site as a resource for the community.
- ESnet will continue to assist sites with perfSONAR deployments and will continue to assist sites with network and system performance tuning.
- ESnet will continue to support the development and deployment of perfSONAR.
- ESnet will continue to support the development and deployment of OSCARS to support virtual circuit services on the ESnet network.

ESnet SC Requirements Review Background and Structure

Funded by the Office of Advanced Scientific Computing Research (ASCR) Facilities Division, ESnet's mission is to operate and maintain a network dedicated to accelerating science discovery. ESnet's mission covers three areas:

1. Working with the DOE SC-funded science community to identify the networking implications of instruments and supercomputers and the evolving process of how science is done.
2. Developing an approach to building a network environment to enable the distributed aspects of SC science and to continuously reassess and update the approach as new requirements become clear.
3. Continuing to anticipate future network capabilities to meet new science requirements with an active program of R&D and advanced development.

Addressing point (1), the requirements of the SC science programs are determined by:

(a) A review of major stakeholders' plans and processes, including the data characteristics of scientific instruments and facilities, in order to investigate what data will be generated by instruments and supercomputers coming online over the next 5–-10 years. In addition, the future process of science must be examined: How and where will the new data be analyzed and used? How will the process of doing science change over the next 5-–10 years?

(b) Observing current and historical network traffic patterns to determine how trends in network patterns predict future network needs.

The primary mechanism to accomplish (a) is through SC Network Requirements Reviews, which are organized by ASCR in collaboration with the SC Program Offices. SC conducts two requirements reviews per year, in a cycle that assesses requirements for each of the six program offices every three years.

The review reports are published at http://www.es.net/requirements/. The other role of requirements reviews is to help ensure that ESnet and ASCR have a common understanding of the issues that face ESnet and the solutions that it undertakes.

In August 2014, ESnet organized a review in collaboration with the FES Program Office to characterize the networking requirements of science programs funded by Fusion Energy Sciences.

Participants were asked to codify their requirements in a case-study format that included 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 is used. Participants considered three timescales in their case studies: 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).

More specifically, the structure of a case study is as follows:

- Background—an overview description of the site, facility, or collaboration described in the case study
- Collaborators—a list or description of key collaborators for the science described in the case study (the list need not be exhaustive)
  - Instruments and Facilities—this describes the ”hardware” of the science of the case study. Instruments and facilities might include detectors, microscopes, supercomputers, telescopes, fusion reactors, or particle accelerators. The instruments and facilities view of the case study provides the information about data rates, data volumes, location of data, origin of data, and so forth.
  - Process of Science—this section describes the ways in which scientists use the instruments and facilities for knowledge discovery. The process of science section captures aspects of data flow, instrument duty cycle, data analysis, workflow, and so forth.
  - Software Infrastructure - this section describes the software used to manage the daily activities of the scientific process in the local environment. It also includes tools that are used to locally manage data resources, facilitate the transfer of data sets from or to remote collaborators, or process the raw results into final and intermediate formats.
- Near-term Remote Science Drivers—a discussion of science drivers that are not local to the primary site or institution of the case study. These typically involve the use of the wide area network for some purpose (data transfer, remote control, remote access, etc). The time period for ”near-term” is 0-2 years, as above. This section has two subsections—instruments and facilities, and process of science.
  - Instruments and Facilities—this describes the ”hardware” of the science of the case study as above, except from a non-local perspective. Examples might include the use of remote HPC resources, or a particle accelerator at another facility.
  - Process of Science—this section describes the process of science as it pertains to the use of remote instruments and facilities.
  - Software Infrastructure—this section describes the software used to manage the daily activities of the scientific process in the wide area environment. It includes tools that are used to manage data resources for the collaboration as a whole, facilitate the transfer of data sets to or from remote collaborators, or process the raw results into final and intermediate formats. The objective is to capture the software tools that move data over the network.
- Medium-term Local Science Drivers—similar to near-term local science drivers, but with a time horizon of 2-5 years in the future. A good way to think of this is that the medium-term view incorporates the current technological paradigm or facility environment, rather than just the current budget horizon.
  - Instruments and Facilities—local instruments and facilities, with a 2–5 year time horizon. Specifically, what will change in the next 2-5 years?
  - Process of Science - local process of science, with a 2–5 year time horizon. Specifically, what will change in the next 2–5 years?
  - Software Infrastructure—local software infrastructure, with a 2–5 year time horizon. Specifically, what will change in the next 2–5 years?
- Medium-term Remote Science Drivers—similar to near-term remote science drivers, but with a time horizon of 2–5 years in the future.
  - Instruments and Facilities—remote instruments and facilities, with a 2–5 year time horizon. Specifically, what will change in the next 2–5 years?
  - Process of Science—remote process of science, with a 2–5 year time horizon. Specifically, what will change in the next 2–5 years?

  - Software Infrastructure—remote software infrastructure, with a 2–5 year time horizon. Specifically, what will change in the next 2–5 years?
- Beyond Five Years—this describes the strategic planning horizon, including new facilities, major technological changes, changes in collaboration structure or composition, etc.
- Network and Data Architecture—this section describes the use of specific networking resources (e.g. a Science DMZ) and how those resources are structured in the context of the science. In addition, if changes would significantly impact the science, they can be captured here.
- Data, Workflow, Middleware Tools and Services—anything not captured in the software infrastructure section can be captured here.
- Outstanding Issues—if there are current problems that should be brought to ESnet's attention, they are captured here.

The information in each narrative was distilled into a summary table, with rows for each timescale and columns for network bandwidth and services requirements. The case study documents are included in this report.

Office of Fusion Energy Sciences Overview

The Fusion Energy Sciences (FES) program mission is to expand the fundamental understanding of matter at very high temperatures and densities and to build the scientific foundation needed to develop a fusion energy source. This is accomplished through the study of plasma, the fourth state of matter, and how it interacts with its surroundings.

FES has four strategic goals:

1. Advance the fundamental science of magnetically confined plasmas to develop the predictive capability needed for a sustainable fusion energy source;
2. Support the development of the scientific understanding required to design and deploy the materials needed to support a burning plasma environment;
3. Pursue scientific opportunities and grand challenges in high energy density plasma science to better understand our universe and to enhance national security and economic competitiveness; and
4. Increase the fundamental understanding of basic plasma science, including both burning plasma and low temperature plasma science and engineering, to enhance economic competitiveness and to create opportunities for a broader range of science-based applications.

To achieve its mission and goals, FES is organized along four subprograms representing mutually supportive scientific areas:

- ***Burning Plasma Science—Foundations***: This subprogram supports fundamental experimental and theoretical research aimed at the resolution of magnetic fusion plasma science issues that will be encountered in the next generation of burning plasma experiments, including ITER. It includes research at the major fusion facilities, theory and simulation, and research on technologies needed to support the continued improvement of the experimental program and facilities. Program elements under this subprogram most responsible for driving FES networking requirements include the three major fusion facilities (the DIII-D tokamak at General Atomics in San Diego, CA, the National Spherical Torus Experiment – Upgrade (NSTX-U) at the Princeton Plasma Physics Laboratory, and the Alcator C-Mod tokamak at the Massachusetts Institute of Technology), and the advanced simulation projects in the FES Scientific Discovery through Advanced Computing (SciDAC) portfolio.
- ***Burning Plasma Science—Long Pulse***: The Long Pulse subprogram exploits capabilities available on new long-duration superconducting international facilities to advance our understanding on how to control and operate a burning plasma and also addresses the development of the materials required to withstand the extreme conditions in a burning plasma environment. This subprogram includes long-pulse international tokamak and stellarator research and fusion nuclear science and materials research. Program elements under this subprogram most responsible for driving FES networking requirements include the recently established multi-institutional collaborations at two superconducting tokamaks in Asia—the Experimental Advanced Superconducting Tokamak (EAST) in China and the Korean Superconducting Tokamak Advanced Research (KSTAR)

in Korea—and a new superconducting stellarator in Germany (Wendelstein 7-X) which will begin operation in 2015.

- ***Burning Plasma Science—High Power***: This subprogram supports the U.S. Contributions to the ITER project. ITER will be the world's first magnetic confinement experiment to achieve self-heated or burning plasmas and its mission is to demonstrate the scientific and technical feasibility of fusion as a future energy source. ITER is currently under construction in St. Paul-lez-Durance, France, by an international consortium consisting of the U.S., China, India, Japan, South Korea, the Russian Federation, and the European Union (the host). While the networking requirements of the ITER project are currently modest and driven by the interactions between the U.S. ITER Project Office (USIPO) at ORNL and the ITER Organization (IO) in France, ITER will be a major driver of FES networking requirements, especially across the Atlantic, when the machine achieves first plasma.
- ***Discovery Plasma Science***: This subprogram supports research that explores the fundamental properties and complex behavior of matter in the plasma state to improve the understanding required to control and manipulate plasmas for a broad range of applications. Research activities supported by this subprogram are carried out at academic institutions, private companies, and national laboratories across the country. The networking requirements of the science supported by this subprogram are rather modest and unlikely to impact the FES needs in the timeframe addressed by this program review.

Case Studies

Case Study 1

Fusion Facilities: International Perspective

1.1 Background

International collaboration has been a key feature of magnetic fusion energy research since its declassification in 1958. Over the last 30 years, formal multi-lateral and bilateral agreements have created, in effect, a single, loosely coordinated research enterprise. The fusion community; which traditionally included the United States, western Europe, Australia, Russia and Japan; has expanded in recent years to include eastern Europe, Korea, China, and India. Planning and program advisory committees typically have cross memberships, particularly among the most active nations—i.e., the United States, Europe, and Japan. Preparation for ITER has further strengthened cooperative research, especially through the ITPA (International Tokamak Physics Activity). Driven by improvements and broad deployment of network technology, the changes in modalities for collaborative research have been dramatic with remote access to data, and remote participation in planning and executing experiments, which are now routine.

However, despite technological advances, challenges remain. The increase of multi-national research teams has brought even more demanding challenges for network and network-based services. Moreover, 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.

1.2 Near-term Local Science Drivers

1.2.1 Instruments and Facilities

The U.S. runs three major experimental fusion facilities, the Alcator C-Mod device at MIT (see Section 3.10), DIII-D at General Atomics (see Section 2.10) , and NSTX-U at PPPL (see Section 4.8). All three have large extended research teams and run, essentially, as National User Facilities. In addition to their collaborators, the facilities carry out coordinated joint research under specific Department of Energy Office of Science targets, and as part of the ITPA.

1.2.2 Software Infrastructure

The MDSplus data system is widely used in the world-wide 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 log books are becoming ubiquitous. Rather than transferring data sets, remote data access and remote computer access are the preferred modes of operation.

1.2.3 Process of Science

See Section 1.1.

1.3 Near-term Remote Science Drivers

1.3.1 Instruments and Facilities

Below is a description of all the major international fusion facilities.

ASDEX-Upgrade (AUG) is a mid-sized divertor tokamak located at the Max-Planck-Institut f́ur 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 U.S. facilities, including on C-Mod (turbulence fluctuations; H-mode pedestal physics; ion-cyclotron range of frequencies, ICRF, heating, metallic first walls; and steady-state scenario development); DIII-D (divertor and pedestal physics; electron-cyclotron range of frequencies, ECRF, heating, 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 U.S. groups.

The ***Joint European Torus (JET)***, previously under the European Fusion Development Agreement (EFDA) and currently under the European Consortium for the Development of Fusion Energy (EUROfusion), 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 U.S. facilities include C-Mod (H-mode pedestal physics; scrape off layer, SOL, transport; self-generated core rotation; toroidal Alfven eigenmode, TAE, physics; and disruption mitigation); DIII-D (H-mode pedestal physics, especially edge localized modes, ELM, suppression; neoclassical tearing modes; resistive wall modes and rotation; and steady-state scenario development); NSTX-U (Alfven eigenmode physics, neoclassical tearing modes, and resistive wall mode research).

ITER is a partnership among seven parties (Europe, Japan, U.S., China, South Korea, Russia, and India) 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 more than 20 years. The current date for first operation is 2023. Collaboration during the construction phase discussed in another chapter of this report; the research phase discussed below under "beyond 5 years."

DOE has specifically funded two large international collaboration projects supporting collaborations between a team comprised of many U.S. 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 and development
- Technology for enhanced international remote operation and participation
- Simulations to support transferring scenarios from one device to another with very different operating characteristics

Korea Superconducting Tokamak Advanced Research (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 with PPPL (PCS, diagnostics, ICRF), ORNL (fueling), GA (PCS, data analysis, ECH), MIT (long-pulse data system), and Columbia University (data analysis). KSTAR had its first plasma in 2008, and U.S. scientists worked closely with KSTAR scientists in the last several experimental campaigns.

Experimental Advanced Superconducting Tokamak (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 U.S. 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 U.S. scientists to collaborate with EAST during the experimental campaign. Tools were deployed to rapidly transfer EAST's MDSplus data to an MDSplus server located in GA's Science DMZ and to then serve that data to approved U.S. scientists. In addition, near-real-time plasma control data from EAST was also transmitted to the GA Science DMZ and made available through a web interface.

Steady State Tokamak (SST-1) 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 with DIII-D focuses on areas of physics, plasma operation, theory, and electron cyclotron emission (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.

Large Hellical 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 U.S. collaborations on this device.

Wendelstein 7-X (W7-X) is a large ($B class) stellarator at Greifswald, Germany. Commissioning began in May 2014 and initial experiments will begin in 2015. The U.S. contributed to construction and MIT is

collaborating on data acquisition for diagnostic systems. A funding opportunity was just announced for participation in the research program.

A number of additional facilities are also targets of somewhat less intense collaboration including WEST (formerly Tore Supra) in Cadarache, France, TCV at the Centre de Recherches en Physique des Plasmas (CRRP) in Lausanne, Switzerland and Mega Amp Spherical Tokamak (MAST) at the Culham Science Center in the UK.

1.3.2 Software Infrastructure

There are three approaches to remote collaboration across the wide area network. 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 the data the user requests. However it suffers from wide-area transactional latency problems.

We are actively working to mitigate the performance problems for MDSplus over wide-area networks. This is being addressed on two levels. On the one hand we are providing tools, which fetch many related data items at once, reducing the number of high-level transactions across the network. At a lower level, MDSplus is being layered on UDP-based protocols, thereby reducing the latency and bandwidth penalty. These can both be further improved by providing local data caching.

1.3.3 Process of Science

The wide-area network plays a critical role in the ability of U.S. 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 U.S. 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, H.323 videoconferencing, and Skype have all been utilized. It should be noted that for some foreign collaborations (e.g., EAST), the ability to use traditional phone lines for conversations are not an option due to the prohibitive expense. Recently, there has been a trend to use H.323 for more formal larger meetings and these are facilitated by Multipoint Control Units (MCU) to bridge together numerous participants. The decision to no longer support ESnet's MCU has resulted in institutions seeking out their own solutions for video conferencing. Possible alternate solutions include Webex, zoom.us, OpenExchange, Blue Jeans, Fuze, and Vidyo. Many international collaborators use H.323-based MCUs and provide GDS dialing instructions for meetings. Each U.S. participant must now work out how to connect to these meetings.

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. Enhancements to MDSplus that reduce the required number of network transactions, as well as automated local caching schemes are being investigated.

Remote participation in international and U.S.-based experiments has the same critical time component. The techniques mentioned above are all used simultaneously to support an operating tokamak, placing even higher demands on the wide-area network, 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.

Despite improvements in intercontinental links and development of national networks, collaborators still report problems with link speed to sites in China, Korea and Japan. This information is anecdotal rather than systematic and is usually brought to our attention when U.S. scientists travel abroad. This suggests that expectations by researchers at some foreign labs are still relatively low. It is not clear if the problem is with the connection from lab to national backbone or with the local area network at these labs. As these issues are identified, solutions can be pursued. Significant improvements have been realized by using UDP-based tools for data transport. Automated data caching schemes could also be applied. Compressed X-Windows (NX from NoMachine), and Windows Remote Desktop provide an alternative to moving large data sets across the WAN.

Further development of tools, services and middleware would be particularly useful for support of international collaboration. The issues are similar to those needed for domestic collaboration but with added difficulty due to differences in technology, standards and policies in the various political entities involved. Needed capabilities include:

1. *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 site, and computes that data at a third site. (The National Fusion Collaboratory deployed this capability for data analysis within the U.S. domain.)
2. *Caching*: Smart and transparent caching will become increasingly important as data sets grow. By the time ITER is in operation, this capability will be essential. Good performance for interactive computing and visualization will require optimization of caching and distributed computing. At the same time, complexity needs to be hidden from end users.
3. *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.
4. *Network monitoring*: We need to be monitoring the network backbone as well as end-to-end connections. Tools for testing and visualizing the state and performance of the network should be readily accessible.

1.4 Medium-term Local Science Drivers

1.4.1 Instruments and Facilities

The local requirements for compute, storage, and network capabilities, are largely unchanged in this time period.

1.4.2 Software Infrastructure

Software maintenance and development will continue on MDSplus with foci on maintainability, performance, and long-pulse operation. Web-based data displayers, are being developed, and there is continuing work on web-based electronic log books and site-specific run planning and management software.

1.4.3 Process of Science

See Section 1.1.

1.5 Medium-term Remote Science Drivers

1.5.1 Instruments and Facilities

EAST: Over the next several years the operation of EAST will continue to expand both in the amount of data taken (the number of diagnostics will increase) as well as the amount of time that the machine is operated. The superconducting nature of the EAST tokamak allows for 24-hour-per-day operations for weeks at a time. There have been discussions between the U.S. and China for the U.S scientists to become actively involved in EAST's third shift operation (daytime in the U.S.). If this is pursued, then there is the possibility of a greater increase in the breadth and scope of this collaboration as well as an increase in the amount of network traffic.

KSTAR: In a similar fashion to EAST, as KSTAR continues to operate over the next five years more data will be available to remote participants and there will be greater opportunity to participate in experiments. In contrast to working on EAST, there has been no discussion regarding third shift operation of KSTAR. Therefore, we conclude for the time being, that the network requirements from the U.S. to EAST will exceed those of the U.S. to KSTAR.

W7-X: The U.S. has an active stellarator program centered at PPPL and ORNL. With the initial experiments at W7X scheduled to begin in 2015, the U.S. collaborations with this device will most likely increase the importance.

International Fusion Energy Research Center (IFERC): As part of ITER's Broader Approach, the IFERC is being built in Rokkasho, Japan. The purpose of this center is to complement the ITER project through various R&D activities in the field of nuclear fusion and will be capable of performing complex plasma physics calculations. With computational power above 1 Petaflop, the supercomputer will be ranked among the most powerful systems in the world, and at least ten times more powerful than any existing system dedicated to simulations in the field of fusion in Europe and Japan. The supercomputer, with a memory exceeding 280 TB and a high-speed storage system exceeding 5 PB, will be complemented by a medium term storage system and a pre/post-processing and visualization system. The full exploitation of this computer from the U.S. will rely on fast and reliable network connectivity between Rokkasho and fusion facilities in the U.S.

1.5.2 Software Infrastructure

See Section 1.3.2.

1.5.3 Process of Science

See Section 1.1.

1.6 Beyond 5 years

ITER research phase: Though the ITER experiment is not scheduled to start up for approximately 10 years, detailed planning has begun for the research program and for the data and communications systems needed to support that program. Estimates on data volume are based on extrapolation from the

current generation of experiments. A (hopefully) more accurate bottoms-up estimate will be carried out as work progresses on all ITER subsystems. Using a variety of methods, the current best guess is that ITER will acquire 1 TB per shot; 1–10 PB per year and will aggregate in the neighborhood of 100 PB over its lifetime. The requirements for off-site access have not been established, but the project is committed to full remote exploitation of the facility. Based on extrapolation from current practice, the project might be required to export 10–100 TB per day, during operation, with data rates in the neighborhood of 0.3–50.0 GBps. At the same time, a steady level of traffic for monitoring and control will be expected. However, this should be less than 10% of the numbers quoted above. In all cases, some form of intelligent caching is assumed so that large data sets are sent only once over intercontinental links. With reasonable effort, the projected data volumes could be accommodated today, so they are not expected to present particular difficulties in ten years time, assuming adequate resources are applied.

On the other hand, coordinating research in such a vast collaboration will likely be a formidable challenge. Differences in research priorities, time zones, languages and cultures will all present obstacles. The sort of *ad hoc*, interpersonal communications that are essential for the smooth functioning of any research team will need to be expanded tremendously in scope. The hope is to develop and prototype tools using the current generation of experiments and to export the technology and expertise to ITER.

JT60-SA: The JT-60SA ("Super Advanced") is a large, breakeven-class, superconducting magnet tokamak proposed to replace the JT-60U device at Naka, Japan. This program represents a coordinated effort between the EUROfusion and Japan Atomic Energy Agency (JAEA). Although there is a rich history of collaboration between the U.S. and Japan the extent of the U.S. involvement in this experiment is not known at the present time, but is expected to go.

1.7 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 our remote collaborations.

Personal interaction is critical to remote collaboration, especially international. Language, time zone differences, and simply not knowing the collaborators as well, exaserbate this. ESnet's descision to drop ECS from its portfolio will complicate this significantly.

1.8 Data, Workflow, Middleware Tools and Services

- GridFTP and FDT have both been used to try and decrease the time to transfer data from remote sites. Any tool or service that can reduce the time to transfer data over the WAN would be beneficial.
- Audio/Video conferencing services are used to conduct remote meetings as well as participate in remote experiments. Today H.323 is the most commonly used service. ESnet dropping their support of this is forcing the community to seek their own heterogeneous solutions for this service (e.g., OpenExchange, zoom.us, Blue Jeans, Vidyo, WebEx, Fuze, Skype).
- QoS for scheduled time period is not used presently but is desired for getting data in real time from remote experiments.
- UDT, an open source protocol layered on UDP, is being used to mitigate latency/transaction induced performance issues.
- Remote desktop protocols, such as NX (NoMachine), are used to allow remote users to utilize local computing resources. In many cases this obviates the need to move large data sets across the wide area network.

1.9 Outstanding Issues

Increased data bandwidth is needed. In the next several years multiple international experiments 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 needed to transfer between international and domestic sites are not very big. However, the real-time or near-real-time aspect of data transfers and between collaborating sites is very important.

QoS will be helpful. Real-time events are needed to coordinate the data transfer, remote control, synchronous data analysis, and coordination of high performance computing resources. While this type of data is not large in size, some type of guaranteed fixed-time-delivery of network packets is very beneficial for effective exploitation of remote experiments and domestic high performance data computational resources. For example, effective coordination of the transferring of ITER experimental data, scheduling of domestic data analysis resources (including exascale leadership class computers for ITER simulations and experiment data computation), and managing of on demand-burst, will rely on guaranteed fixed-time-delivery of events. 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. (E.g., ESnet peering with GLORAID on November 2010 decreased the latency about 25% between DIII-D and EAST.)

Integrated collaboration services, as were provided by ECS, are critical to collaborative international research. Multi-point video conferencing, telephone bridging, and screen sharing are now standard practice. International collaborations have been somewhat problematic due to heterogeneous collaboration tool sets. The decision to drop ECS, without replacement, exacerbates this heterogeneity. The experiments, due to local constraints, will pursue their own solutions, exposing them to learning curve, cost, and interoperability issues.

Table 1.1: Fusion experimental collaborations rely on real-time data streaming and data replication. While the total size of data is not large, due to the team-based nature of fusion experiments, the real-time aspect of data transfer, audio/video streaming is critical. When ITER comes online in about 10 years, the data traffic, both domestically and between Europe and the United States, will immediately increase.

Science Drivers			Anticipated Network Needs	
Science Instruments, Software, and Facilities	Process of Science	Data Set Size	LAN Transfer Time Needed	WAN Transfer Time Needed
0-2 years				
• Multiple remote experiment facilites	• Real-time data access and analysis • Team-based collaboration with data sharing, screen sharing and multi user high-definition videoconferences	• Data volume (2TB/day) • TCP/IP-based client server data, audio/video	• Consistent streaming 24x7	• Consistent streaming 24x7 (1–2 min delay is tolerable)
2–5 years				
• Multiple remote experiment facilites (New facilities will be added)	• Real-time data access and analysis • Team-based collaboration with data sharing, screen sharing and multi user high-definition videoconferences	• Data volume (2TB/day) • TCP/IP-based client server data, audio/video	• Consistent streaming 24x7	• Consistent streaming 24x7 (1–2 min delay is tolerable)
5+ years				
• Multiple remote experiment facilites (New facilities including ITER will be added)	• Real-time data access and analysis • Team-based collaboration with data sharing, screen sharing and multi user high-definition videoconferences	• Data volume (20TB day) • TCP/IP-based client server data, audio/video • Possible file-based simulation	• Consistent streaming 24x7	• Consistent streaming 24x7 (1–2 min delay is tolerable)

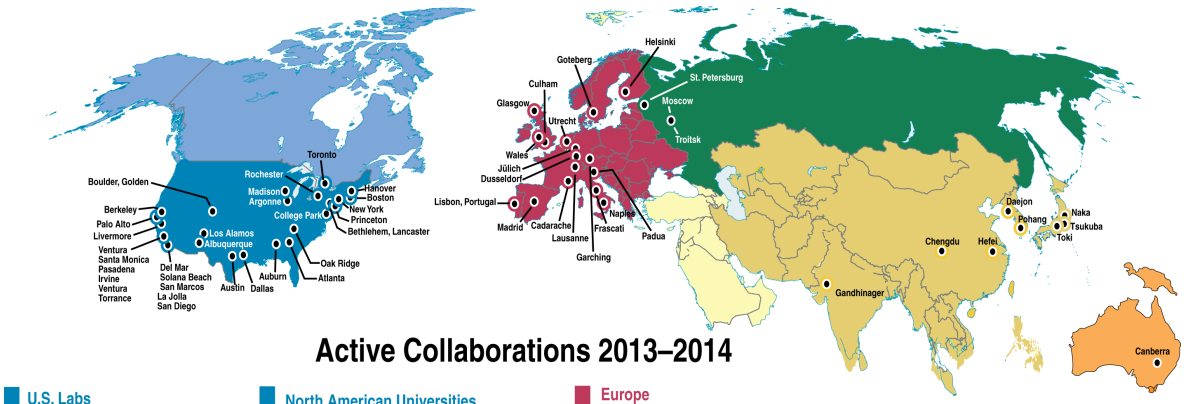
Case Study 2

General Atomics: DIII-D National Fusion Facility and Theory and Advanced Computing

2.1 Background

The DIII-D National Fusion Facility at General Atomics' 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 plasma control system has opened a new world of precise control of plasma properties and facilitates detailed scientific investigations. Its open data system architecture has facilitated national and international participation and remote operation. 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 impact ITER performance, and training new scientists for support of ITER experiments.

General Atomics also conducts research in theory and simulation of fusion plasmas in support of the Office of Fusion Energy Sciences 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 NERSC and OLCF.



Active Collaborations 2013–2014

U.S. Labs

LBNL (Berkeley, CA)
LLNL (Livermore, CA)
NASA Ames (Mtn View, CA)
NREL (Golden, CO)
ORNL (Oak Ridge, TN)
PPPL (Princeton, NJ)
SNL (Sandia, CA, NM)

U.S. Industries

ALITRON (Solana Beach, CA)
Calabazas Creek (San Mateo, CA)
CompX (Del Mar, CA)
CPI (Palo Alto, CA)
FAR-TECH, Inc. (San Diego)
General Atomics (San Diego)
Lightcraft Technologies (Bennington, VT)
Lodestar (Boulder, CO)
National Instruments (Austin, TX)
Tech-X (Boulder, CO)

North American Universities

Auburn U. (Auburn, AL)
Columbia U. (New York, NY)
College of William & Mary (Williamsburg, VA)
Courant Institute (NY)
Georgia Tech (Atlanta, GA)
Lehigh (Bethlehem, PA)
MIT (Cambridge, MA)
ORAU (Oak Ridge, TN)
ORISE (Oak Ridge, TN)
Palomar College (San Marcos,CA)
Purdue U. (W. Lafayette, IN)
U. Arizona (Tuscon, AZ)
UCB (Berkeley, CA)
UC Davis (Davis, CA)
UCI (Irvine, CA)
UCLA (Los Angeles, CA)
UCSD (San Diego, CA)
CIPS, U. Colorado (Boulder)
U. Maryland (College Park, MD)
U. Rochester (NY)
U. Texas (Austin, TX)
U. Toronto (Toronto, Canada)
U. Tulsa (Tulsa, OK)
U. Wisconsin (Madison, WI)
West Virginia U. (Morgantown, WV)
Yeshiva U. (New York, NY)

Europe

CCFE (Culham, United Kingdom)
CEA (Cadarache, France)
CFN-IST (Lisbon, Portugal)
CIEMAT (Madrid, Spain)
CRPP-EPFL (Lausanne, Switzerland)
D-TacQ Solutions (Scotland)
DIFFER (Nieuwegein, Netherlands)
ENEA (Frascati, Italy)
ENEA Consorzio RFX (Padua, Italy)
ENEA Consorzio CREATE (Naples, Italy)
EURATOM-FOM (Utrecht, Netherlands)
EURATOM-Tekes, Aalto U. (Helsinki, Finland)
Fusion for Energy (Barcelona, Spain)
FZ-Julich (Germany)
HH U. of Dusseldorf (Germany)
IFP-CNdR (Italy)
IPP,AS CR, EURATOM/IPP, Prague
ITER (Cadarache, France)
JET EFDA (Culham, United Kingdom)
Max Planck-IPP (Garching, Greifswald)
U. of Helsinki (Finland)
U. Padova (Padua, Italy)
U. Strathclyde (Glasgow, Scotland)
U. of York (York, United Kingdom)
VTT (Finland)

Russia

D.V. Efremov Institute (St. Petersburg, Russia)
RCC Kurchatov Institute (Moscow)

Asia

ASIPP (Hefei, China)
IPR (Gandhinagar, India)
JAEA (Naka, Japan)
KAIST (Daejeon, S. Korea)
Kyoto University (Japan)
KBSI-NFRI (Daejeon, S. Korea)
NIFS (Toki, Gifu-ken, Japan)
Pohang U. (S. Korea)
Seoul Nat. U. (S. Korea)
SWIP (Chengdu, China)
Tohoku University (Sendai, Japan)

Australia

Australia National U. (Canberra)

Figure 2.1: A wide base of collaborations establishes a foundation for a strong DIII-D program.

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. 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 (Figure 2.1). Approximately 350 researchers from around the world are active users of DIII-D data, and there were 443 authors on DIII-D papers in 2011–2012. These team members are from 83 institutions including, 39 Universities (26 U.S., 13 international), 34 National Laboratories (7 U.S., 27 international), and 10 High Technology U.S. companies.

2.3 Near-term Local Science Drivers

2.3.1 Instruments and Facilities

The General Atomics' connection to ESnet is at 10Gbps with major computing and storage devices connected to a switched 10Gbps and switched 1Gbps Ethernet LAN. Network connectivity between the major computer building and the DIII-D facility is 20Gbps. The major data repositories for DIII-D comprise approximately 250TB of online storage with metadata catalogues stored in a relational database.

Network connectivity to offices and conference rooms is at 1Gbps and 100Mbps 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 funding limits operations to approximately 15 weeks per year. For each plasma pulse, up to 10,000 separate multi-dimensional measurements are acquired and analyzed representing 30 Gigabytes of data. The experimental data is accessed both locally and over the WAN. 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.

2.3.2 Software Infrastructure

The vast majority of DIII-D's raw data is stored in PTDATA, a locally written client/server data management system. Large frame rate diagnostic camera data is presently stored outside of PTDATA in the camera's native format and is served to users by a simple NFS directory mount. A large amount of DIII-D's analyzed data is stored in MDSplus, a client/server data acquisition and management system in use at a number of fusion facilities worldwide. Metadata catalogues are stored in a relational database for rapid searching. Programs written in Fortran, IDL, Matlab, Python, and C do the majority of data analysis and visualization. A variety of Linux clusters and workstations are available for computation to support automatic processing immediately after data is acquired, to interactive analysis in the DIII-D control room, to large-scale analysis/simulation done on multi-node systems. A large central file system mounted via NFS gives the user community a unified home area from all compute systems as well as a repository for code input and output files.

2.3.3 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 15 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.

2.4 Near-term Remote Science Drivers

2.4.1 Instruments and Facilities

DIII-D's data is made available to remote collaborators through two 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 GUI programs, the remote desktop software No-Machine is supported.

2.4.2 Software Infrastructure

The scientific process in the wide area environment is very similar to that in the local area environment. Tools that are used to manage data locally are the same that are used to manage data 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-Windows traffic over the wide area network.

2.4.3 Process of Science

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 the collaborative services mostly associated with several different types of videoconferencing with the majority H.323 based. As the collaborative activities associated with DIII-D continue to increase there becomes an increasing usage of collaborative visualization tools by off site researchers that requires efficient automatic data transfer between remote institutions.

The highly distributed nature of the DIII-D National Team requires the usage of substantial remote communication and collaboration technology. There are a total of eight Polycom H.323 end-points within the DIII-D complex: five conference rooms, one office, the DIII-D control room, and the DIII-D remote control room. In addition, the majority of these locations have the ability to use software based collaboration tools (e.g., Skype). On their own, scientists also utilize a variety of technology to communicate with audio/video to the desktop.

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.

2.5 Medium-term Local Science Drivers

2.5.1 Instruments and Facilities

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 2008 to 2012 the total amount of yearly data taken at DIII-D increased by four fold. To keep up with this demand plus the increased usage of collaborative technologies, even within the local campus, discussions have begun to increase the reach of 10Gbps within the LAN.

To reduce the cost of data storage, an examination of object storage technology is presently underway. It is not clear at this time if a private cloud model will be adopted or a more of a “cold” repository to alleviate the burden on the main central file system. The driving force behind these considerations is to reduce the average dollar per terabyte cost that the project is spending on storage.

2.5.2 Software Infrastructure

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

2.5.3 Process of Science

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.

2.6 Medium-term Remote Science Drivers

2.6.1 Instruments and Facilities

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 will be particularly challenging. Some important components that will help to make this possible include easy to use and manage user authentication and authorization framework (e.g., certificate authority), global directory and naming services, distributed computing services for queuing and monitoring, parallel data transfer between remote institutions, and network quality of service (QoS) in order to provide guaranteed bandwidth at particular times or with particular characteristics.

The unexpected termination of ESnet's Collaboration Services (ECS) will most likely result in the adoption of a commercial cloud based company to DIII-D's collaboration services rather than buying large hardware to be located in DIII-D's data center. In addition, such a cloud service will allow collaboration from a wide range of devices (e.g. laptops, tablets, phones) with the potential to increase the strain on DIII-D's ESnet network connection.

To reduce data storage costs (as described in Section 2.5.1), an initial examination is being undertaken of using commercial cloud storage such as that offered by Amazon Web Services (AWS). If such a model is adopted for some of DIII-D's data, the movement of data to and from the cloud service provider will add an additional burden to the ESnet network.

2.6.2 Software Infrastructure

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

2.6.3 Process of Science

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 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 with particularly China and South Korea will become increasingly important.

2.7 Beyond 5 years

In the outlying years, it is anticipated that the DIII-D will add more diagnostics, which will create more network load. Multiple international tokamaks will be fully operative with a rich diagnostic set and ITER, located in France, will be close to coming on line, and will operate in long pulse mode. With DIII-D operating to assist ITER it is possible to imagine the DIII-D scientific team working on numerous tokamak simultaneously placing a further strain on the WAN and creating a need for efficient peering to our Asian and European partners.

2.8 Network and Data Architecture

There has been some initial work on high-performance data transfers from the EAST tokamak in China to General Atomics to support U.S. scientific collaboration on EAST. Following the guidelines of ESnet,

a Science DMZ has been configured and appropriate software deployed to automate the transfer of data to allow near-real-time participation of U.S.-based scientists in EAST operation. Data transfers from EAST to General Atomics using both IPv4 and IPv6 have been utilized and specialized bulk data transfer software has been deployed to decrease the transfer time by a factor of 300.

Based on the anticipated growth of DIII-D's usage of the wide area network within the next 5 years, it is expected that there will be a need to increase DIII-D's ESnet connectivity beyond the present 10Gbps to either 40 or 100Gbps

2.9 Data, Workflow, Middleware Tools and Services

The increased collaboration of scientists on remote operating tokamak will continue to place a burden on the network and the tools used to move large amounts of data over the wide area network. Since participating in the science of an operating tokamak is a very time constrained environment, the amount of data to be examined combined with the short amount of time creates a large data rate. Thus, even though the total amount of data in fusion energy sciences is not large compared to other disciplines (e.g. high energy physics), the time requirement does create a higher data rate. Networks, services, and tools that support automated large data flows to support remote experimental operation will be critical to the success of fusion energy sciences in the outlying years.

2.10 Outstanding Issues

In general, increased data bandwidth is needed as DIII-D generates more data in the future and the number of remote participants continues to increase.

Lowering the network latency is also important. Currently, the amount of data needed to transfer between DIII-D and its partner sites are not very big. However, the real-time or near-real-time aspect of data transfers and between collaborating sites is very important.

Quality of Service (QoS) will be helpful. Real-time events are needed to coordinate the data transfer, remote control, synchronous data analysis, and coordination of high performance computing resources. While this type of data is not large in size, some type of guaranteed fixed-time-delivery of network packets is very beneficial for effective exploitation of remote experiments and domestic high performance data computational resources.

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. (E.g., ESnet peering with GLORIAD on November 2010 decreased the latency about 25% between DIII-D and EAST.)

Integrated collaboration functionality, as were provided by the ESnet Collaboration Services (ECS), are critical to collaborative research. Multi-point video conferencing, telephone bridging, and screen sharing are now standard practice for the DIII-D research community. The termination of ECS has resulted in DIII-D allocating resources to evaluate alternate solutions to replace the functionality of ECS. It appears that the U.S. fusion community will end up adopting a variety of solutions resulting in a heterogeneous collection of collaboration tool sets. This heterogeneity will place an additional burden on the DIII-D researcher who collaborates with a variety of U.S. institutions. Any activity to help push the community back to a uniform tool set would be highly beneficial to the U.S. fusion energy sciences research community.

Table 2.1: Table summarizing this case study's network expectations and foreseeable science drivers for the next five years and onward.

Science Drivers			Anticipated Network Needs	
Science Instruments, Software, and Facilities	Process of Science	Data Set Size	LAN Transfer Time Needed	WAN Transfer Time Needed
0-2 years				
• DIII-D Tokamak • Collaboration on other experiments • SciDAC/FSP simulation and modeling • Assistance in ITER construction	• Real time data access and analyses for experimental steering • Shared visualization • Remote collaborative technologies • Parallel data transfer	• Data volume (2TB/day) • Data set composition: TCP/IP-based client server data, audio/video	• Consistent streaming 24x7	• Consistent data streaming 24x7 (1–2 min delay is tolerable)
2–5 years				
• DIII-D Tokamak • Collaboration on other experiments • SciDAC/FSP modeling • ITER construction support and preparation for experiments	• Real time data analysis for experimental steering combined with simulation interaction • Real time visualization interaction among collaborators across U.S.	• Data volume (5TB/day) • TCP/IP-based client server data, audio/video	• Consistent streaming 24x7	• Consistent data streaming 24x7 (1–2 min delay is tolerable)
5+ years				
• DIII-D Tokamak • Collaboration on other experiments • ITER experiments	• Real time remote operation of the experiment • Comprehensive simulations	• Data set (20TB/day) • TCP/IP-based client server data, audio/video • Possible file-based simulation	• Consistent streaming 24x7	• Consistent data streaming 24x7 (1–2 min delay is tolerable)

Case Study 3

Plasma Science and Fusion Center: Alcator C-Mod Tokamak

3.1 Background

The Plasma Science and Fusion Center (PSFC) is a large interdisciplinary research center located on the MIT campus in Cambridge, MA. Its major facility is the Alcator C-Mod tokamak—one of the three major devices in the U.S. magnetic fusion energy program. The PSFC has a number of smaller research facilities as well, including LDX (Levitated Dipole Experiment). Research on these devices has relevance outside the fusion program, particularly to space and astrophysical plasma physics. The Plasma Science division at PSFC carries out a broad program of theory and computational plasma physics. The computational work emphasizes wave-plasma interactions and turbulent transport.

3.2 Collaborators

Collaborators on the C-Mod facility represent a level of effort of approximately 20 FTE. These scientists (and a few students from collaborating universities) provide diagnostic equipment, propose and carry out experiments and perform simulations in support of C-Mod research. With more than 35 collaborating institutions in total, the major collaborating institutions include:

- Princeton Plasma Physics Laboratory
- University of Texas, Austin
- Los Alamos National Laboratory
- Lawrence Livermore National Laboratory
- General Atomics
- University of California, San Diego
- University of California, Los Angeles
- ITER organization
- Max-Planck-Institut für Plasmaphysik (IPP), Garching, Germany
- French Alternative Energies and Atomic Energy Commission (CEA), Cadarache, France
- University of Tromso, Norway
- Lodestar Research Corporation, Boulder, CO

- University of Tokyo, Japan
- York University, UK

3.3 Near-term Local Science Drivers

3.3.1 Instruments and Facilities

At the Alcator C-Mod facility, research is carried out in the areas of turbulent transport, plasma-wall interactions, MHD and RF heating and current drive. A significant portion of machine time is devoted to answering questions connected to design and operation of the ITER device, now under construction in Cadarache, France. The C-Mod team is international, with collaborators at more than 35 institutions in the United States and abroad. C-Mod is also an important facility for graduate training. Due to funding uncertainty there are only 12 PhD students currently carrying out thesis research, but new students are now being accepted.

PSFC has about 1,500 network-attached devices, more than half associated with the C-Mod team and experiment. The infrastructure is switched 1 and 10 Gbps Ethernet, with 1 gigabit connectivity to all workstations and desktops. The C-Mod experiment is directly supported by about 10 multi-core linux servers for data acquisition, storage and analysis and 80 linux workstations for users. A great deal of additional equipment is used for real-time monitoring and control. The experiment conducts 30–40 "shots" per day, each storing about 12 GB of data. All experimental data is maintained on disk, with approximately 70 TB currently archived. This data is duplicated on a second RAID array and backed up on local tape archives as well as on MIT's TSM tape backup system. (Since the PSFC computers are in ESnet address space, this latter backup process generates traffic to and from the ESnet-MITnet interconnect located off the MIT campus in Boston.) Higher level data is maintained in SQL databases, which hold several million records.

The experimental team makes extensive use of the DOE computational facilities at NERSC and OLCF, as well as a local computer cluster with 600 cores (with various upgrades planned). PSFC researchers are also actively involved in several SciDAC collaborations.

3.3.2 Software Infrastructure

Data collection, management, and remote dissemination for the Alcator C-Mod experiment is done with the MDSplus data system. This system, developed at MIT, has become a *de facto* standard in the fusion community. It is used by about 40 experiments around the world. Rather than transfer entire data sets, the system provides remote users with the same data reading and data writing APIs as are used locally. Since the protocols are transactional, latency rather than bandwidth dictates performance.

3.3.3 Process of Science

The experimental team works in a highly interactive mode, with significant data analysis and display carried out between shots. Typically 25–45 researchers are involved in experimental operations and contribute to near real-time decisions between shots. Thus a high degree of interactivity with the data archives and among members of the research team is required.

3.4 Near-term Remote Science Drivers

3.4.1 Instruments and Facilities

Activity centers around a set of national and international experiments (the latter described in more detail in Section 1.9). The C-Mod team works closely with the experimental groups on the DIII-D and NSTX tokamaks (at GA and PPPL respectively), as well as a large set of widely dispersed collaborators. PSFC researchers also make intensive use of DOE computing facilities, principally at NERSC and OLCF.

The PSFC WAN connection is through a local gigabit link to an on-campus interconnect with MITnet, of which ESnet and Internet2 peer. The local fiber infrastructure would allow this link speed to be increased further, with only moderate effort and expense if traffic warrants. The ESnet connection is shared by SC researchers at MIT, particularly the Lab for Nuclear Sciences (LNS) funded by the Offices of High Energy Physics and Nuclear Physics.

3.4.2 Software Infrastructure

The MDSplus data system is used for both local and remote data management. We are working on addressing wide area network performance issues with two thrusts:

- Acknowledged network transactions are subject to network latencies. By reducing the number of exchanges, by transferring more complex data structures in each transaction, better wide area performance can be achieved.
- TCP/IP performance is limited by network latency and packet loss. Layering our protocols on UDT, a UDP-based transport protocol, improves wide-area performance.

Note: We would like to optimize performance rather than throughput.

3.4.3 Process of Science

As noted above, fusion experiments are highly interactive—regardless of whether researchers are on or off site. Remote researchers can lead experiments, control diagnostics (measurement systems) and trigger data analysis tasks.[1] Fast and efficient data access is clearly a requirement for this mode of operation.

Multi-institutional collaborations are a critical part of the research carried out at the PSFC. In addition to remote researchers who use facilities at MIT, scientists and students at the PSFC are actively involved in experiments at laboratories around the world. As noted above, the MIT theory groups are involved in several nation-wide computational projects and rely on use of remote supercomputers. MIT also supports the MDSplus data system that is installed at about 40 fusion facilities world-wide.

All groups at the PSFC make active use of collaboration technologies. Five conference rooms are set up for videoconferencing and are used for all regular science and planning meetings. In addition, videoconferencing is available from the C-Mod control room and used to support remote participation. In recent years, off-site session leaders led 5–10% of runs. Videoconferencing software is also installed on many office computers.

The PSFC has made significant use of the ESnet-provided collaboration services. The H.323 video-conference facilities were used for both scheduled and *ad-hoc* collaboration. Data/screen sharing were regularly used to broadcast visuals from presentations. We had hoped that over the next five to ten years there would be an expansion of these services in both technical and support areas (see below

[1] Fusion has a long history of this work. As a historical note, remote operation of tokamak diagnostics was first employed in 1992 and full remote operation of a tokamak was demonstrated in 1996 when a group of scientists from MIT and LLNL ran the C-Mod experiment from a control room in California.

for details). ESnet's decision to drop support for these services will force us to replace them as best we can. This will lead to heterogeneous solutions in the community, which is not optimal.

MIT is in the process of migrating its phone system to SIP-based VoIP systems. These systems offer the possibility of supporting the next generation of collaboration tools. Taking advantage of an MIT pilot program, we have been able to integrate new tools into the normal workflow. One aim is to improve *ad-hoc* interpersonal communications, which, we believe, limits the effectiveness and engagement of remote participants. We can anticipate similar technology migration at all collaborating sites in coming years. Collaboration tools based on the SIP protocol would seem to offer a method for seamless integration of needed services.

3.5 Medium-term Local Science Drivers

3.5.1 Instruments and Facilities

Overall, operations on our experiments will be similar over the next 5 years. Historically, data rates on the C-Mod experiment have increased by a factor of 10 roughly every 6 years. This is expected to continue.

3.5.2 Software Infrastructure

To provide for the ongoing expansion in data rates, we have begun planning for upgrades to the local area network. We are building out our 10G links to servers and 10 G network backbone. Workstation and desktop connectivity will remain at 1Gbps for the near future.

Working with the MIT Information Services group, we expect to continue and expand SIP-based tools to fully integrate our data and telecommunications networking. Over this period a complete migration from traditional telephony to VoIP is anticipated.

3.5.3 Process of Science

We would not anticipate major changes in the science processes in this time period. Experimental operation will continue to be highly interactive and involve the simultaneous interactions of a large fraction of the research team.

3.6 Medium-term Remote Science Drivers

3.6.1 Instruments and Facilities

A set of new international facilities (EAST, KSTAR, and SST-1) are now in operation and are entering a physics exploitation phase. Others will begin operation during this period (W7-AX in 2014, JT60-SA in 2016).

3.6.2 Software Infrastructure

See Section 3.5.2.

3.6.3 Process of Science

Collaboration with off-site researchers will continue grow over the next 5 years. Funded collaborations are in place with other domestic facilities at PPPL and GA as well as several international facilities—most notably EAST (China), KSTAR (Korea) and AUG (Germany). The W7-X stellarator in Germany will commence operation this year and several proposals for collaboration with that facility are planned. Activities in support of ITER construction will be centered around the U.S. ITER Program Office and will probably not drive much additional traffic to MIT in this time period.

3.7 Beyond 5 years

As we approach ITER operations (about 8 years from now), there will be increased network traffic associated with preparation for the research program, data challenges and diagnostic development.

As stated above, the removal of ESnet managed collaboration tools will make addressing future collaboration needs (which are evolving as anticipated) problematic. Future collaboration tool needs include:

- Global directory services
- Centrally administered conferences (call out)
- VoIP/SIP collaboration tools
- Screen sharing presentation tools
- Recording and Playback
- Instant messaging (collaboration and meeting setup)
- Higher quality multipoint video
- Presence/availability information
- Better integration of above elements
- Integration with authorization tools

Support Needs

- Quickly determine the state collaboration services
- Communicate this state to meeting participants (Instant Messaging)

Operation of these tools should not be too complicated or expensive. Typically remote collaborators do not have full-time staff support to initiate and monitor every remote session. The current trouble ticket response system is not timely enough when there is a meeting taking place or about to start with remote participants. In the case of technical difficulties, decisions about canceling a meeting or remote session needs to be made promptly.

3.8 Network and Data Architecture

MIT, in partnership with General Atomics and Lawrence Berkeley National Laboratory, is working on a DOE-sponsored big data initiative entitled: “Automated metadata, provenance catalog, and navigable interfaces: ensuring the usefulness of extreme-scale data.” The project is building a set of tools that support data tracking, cataloging, and integration across a broad scientific domain. The system documents workflow and data provenance in the widest sense, providing information about the connections and dependences between the data elements. Data, from large-scale experiments and extreme-scale computing are expensive to produce and may be used for high-consequence applications. The main

network requirements for the project are video conferencing and screen sharing, which the team uses extensively.

3.9 Data, Workflow, Middleware Tools and Services

Our collaborations use a wide range of middleware tools and services. The de facto standard for remote data access is MDSplus, which was developed within the U.S. fusion community. A subset of Globus tools are used to implement a secure layer for data access. Authentication is based on X.509 certificates and a "FusionGrid" certificates from the Open Science Grid certificate authority. Also created by the fusion collaborator is a distributed authorization system (ROAM) – which allows the creation of managed resources with a flexible set of privileges defined and assigned to end users by individual resource managers. As noted, we have made extensive use of the ESnet collaboration services.

There is a real need for better integration of remote participation tools. The integration mentioned above needs to encompass all modes of machine mitigated interpersonal communication. The goal is to enable users to initiate conversations on the most appropriate media then move from tool to tool (voice, video, IM, email, screen sharing, data sharing, etc.) seamlessly and as needed. Tools need to be standards-based, modular, role-aware, presence-aware and web friendly in a multi-platform environment. This becomes more difficult in the new environment.

Improvements in cybersecurity are also needed. This includes federated authentication, single sign-on and better credential life-cycle management (creation, renewal, revocation).

Our limited experience with commercial cloud services for data sharing with China has not been positive. The transfer times to and from the "cloud" for large data sets are prohibitive. We would welcome any help that ESnet could provide either using the commercial vendors, or with similar services they could provide.

3.10 Outstanding Issues

Issues related to existing collaboration services H.323

- We will need to provide and pay for multi-point video conferencing service.
- The choice of vendors for this service will be done on a site-by-site basis making integration, and advanced functionality difficult.
- Endpoints will change from hardware video conferencing codecs to computer-based systems.
- GDS (global dialing) is still the preferred mechanism for many of our collaborators to specify connection directions for meetings. We will have to provide our own registered gate keepers or use alternate dialing schemes to join these meetings.

Screen Sharing

- ESnet-provided ReadyTalk is heavily used in the community.
- As with H.323 sites will have to provide (and pay for) replacement service.
- These services will likely be heterogeneous in the community.
- There are learning curve, and startup inefficiencies with any new solution.

Chat, Presence and Directory Service

- These services would greatly benefit our collaborative research environment.

- It will be very difficult to integrate and deploy unified chat, presence, and other services in an environment where each institution or facility deploys their own collaboration services - interoperability is a much greater barrier in this case.

Table 3.1: Table summarizing this case study's network expectations and foreseeable science drivers for the next five years and onward.

Science Drivers			Anticipated Network Needs	
Science Instruments, Software, and Facilities	Process of Science	Data Set Size	LAN Transfer Time Needed	WAN Transfer Time Needed
0-2 years				
• C-Mod Tokamak • Collaboration on other national and international facilities • Simulation and modeling	• Incoming and outgoing remote participation on experiments • Use of remote supercomputers, remote databases • Active use of collaboration technologies	• Data volume (300 GB/day) • Data set composition 1,500 files; largest about 1 GB	• Less than 3 minutes	• Bursty; small portion of data set transferred with no noticeable delay • e.g., 20 MB in 1 second • Endpoint; see note
2–5 years				
• Prep work for ITER • Additional international collaborations	• Increased use of collaboration technologies including SIP/VoIP • Involvement in development of next major US experiments	• Data volume (1TB/day) • Data set composition 3,000 files; largest about 2 GB	• Less than 3 minutes	• Bursty; small portion of data set transferred with no noticeable delay • e.g., 40 MB in 1 second • Endpoint; see note
5+ years				
• Research on ITER • Additional international collaborations • Possible new facilities for materials PWI, or FNS	• preparation for ITER, research on ITER • Increased emphasis on cyber security due to regulatory issues on ITER	• Data volume (3–5 TB/day) • Data set composition 3,000 files; largest about 4 GB	• Less than 3 minutes	• Bursty; small portion of data set transferred with no noticeable delay • e.g., 80 MB in 1 second • Endpoint; see note

Case Study 4

Princeton Plasma Physics Laboratory

4.1 Background

PPPL physicists develop and run 8–10 different major massively parallel physics codes, mostly at the National Energy Research Scientific Computing Center (NERSC/LBNL), the Oak Ridge Leadership Computing Facility (OLCF/ORNL), or at the Argonne Blue Gene/P supercomputer (ANL). PPPL scientists are collaborators on several SciDAC projects including the Center for Simulation of Plasma Microturbulence (CPSM), the Center for Simulation of Wave-Plasma Interactions (CSWPI), the Center for Extended Magnetohydrodynamic Modeling (CEMM), the Center for Edge Physics Simulation (EPSI), and the Center for Nonlinear Simulation of Energetic Particles in Burning Plasmas (CSEP). There are also typically several INCITE and ALCC awards to PPPL scientists each year.

Massively parallel codes study different aspects of physical phenomena that occur in fusion confinement configurations and are the state of the art for both scientific content and computational capabilities. The codes mostly divide into three types. The microturbulence codes study the development and effects of fine-scale turbulent fluctuations in the core of the confinement region that lead to increased particle, momentum, and energy loss in tokamaks and stellarators. Among these are GTC, GTS, and GYRO. The edge physics codes study the physics at the boundary between the core plasma and the surrounding vacuum region. The leading codes in this area are XGC0 and XGC1. The macrostability codes solve the extended magnetohydrodynamic equations to study the onset and evolution of device-scale global instabilities over long timescales. Among these are M3D, M3D-K, and M3D-C1. (The M3D-K hybrid code is used to simulate energetic particle-driven Alfven instabilities and energetic particle transport in tokamak plasmas and its requirements tend to be intermediate between GTS and M3D-C1). We focus here on one code of each category: GTS, XGC1, and M3D-C1.

In addition to the massively parallel physics codes that are run remotely, PPPL maintains an extensive local computing capability for running serial jobs and those requiring modest numbers of processors. The local facility is also used extensively to debug and postprocess the massively parallel jobs. In addition, the PPPL local facility is the home of the TRANSP interpretive and predictive transport code package. TRANSP is used by tokamak physicists worldwide to interpret experimental data from experiments and to predict the operation of future experiments.

In FY 2013 there were 5,675 TRANSP runs made that accounted for about one million hours of CPU time. About half of the TRANSP jobs were run by scientists external to PPPL. Access is provided to PPPL for running TRANSP via the Fusion Grid. Although TRANSP was originally a serial code, it is being parallelized incrementally, so that now about one-third of the submitted jobs use MPI-based parallelism, usually with 8–128 processors. While TRANSP has been installed at several remote sites, most of the runs are made on the PPPL cluster (OpenScienceGrid), even if submitted remotely. This allows for rapid

debugging and turnaround by the TRANSP support staff should the run fail.

4.2 Collaborators

Most of the massively parallel jobs are run at LBNL, ORNL, or ANL and the data is stored and analyzed where the job is run via remote connections (using NX) to scientists at their desks. Smaller jobs are typically run and analyzed at PPPL, as are all the TRANSP jobs. TRANSP is routinely run by a number of physicists for various purposes, which include interpretive analysis of existing datasets, predictive runs for testing theoretically-based transport models, for developing operational scenarios in present and future devices, and with the recent implementation of the free boundary equilibrium solver, for developing algorithms for control of parameters such as rotation and current profiles. These control algorithms will eventually be implemented into the NSTX-U real-time control system. TRANSP is being used as a predictive tool for developing operational scenarios for about 500-second-long ITER discharges.

4.3 Near-term Local Science Drivers

4.3.1 Instruments and Facilities

Approximately 5500 processing cores and 850TB of storage are available locally at PPPL. This provides local computing resources and storage for small simulations. While processing around 160,000 jobs per year, 40% are single CPU jobs, 50% utilize 2-64 CPUs, with the remaining 10% utilizing between 64 and 512 CPUs. About 25 million CPU hours were consumed in FY 2013.

Efficient internal networking is important for file access and interprocess communication, but wide area access is also important, as 50% of registered users are located offsite at collaborative institutions, both within the United States and overseas. Offsite users access data and facilities unique to PPPL, including NSTX data, the TRANSP (tokamak transport code) processing environment, and other collaborative capabilities.

To support this collaborative research, PPPL enjoys a 10Gbps connection with ESnet. PPPL is not taxing this connection presently, with uptime and availability a more important concern than raw bandwidth. This is especially true since PPPL has moved core services, like email, to the internet “cloud,” and will be migrating more services in the near future. PPPL also has a backup connection to ESnet’s New York router, at 10Gbps, which is automatically utilized if the main 10Gb link to Washington DC is interrupted.

Thus with its current ESnet bandwidth of 10Gbps, PPPL is comfortably meeting the current needs of its research mission. PPPL needs to extend that high-speed capability further into the PPPL campus so that the transfer of large data sets to local storage can be accomplished in a timely manner as well. A dedicated Globus server has been deployed, as has the *bbcp* file transfer tool. Extensive testing with perfSONAR has been completed, and an internal network rework, with ESnet’s excellent support, is in the design phase to reduce bottlenecks to efficient data transfer.

4.3.2 Process of Science

A typical simulation code GTS, a particle-in-cell gyrokinetic code, today employs about 40 billion particles and a mesh of approximately 400 million node points and is run for about 10,000 time steps on 100,000 processors. Storage requirements for each time step is dominated by the particle data: 4 x 1010 particles x 8 bytes x 12 variables = 4 TB. If particle data from every time step is saved, it would require 40 PB of storage. However, normally only the mesh data is saved for post-processing. If mesh data is saved every 10 time steps, this would require 4 x 108 (mesh size) x 8 (bytes) x 4 (variables) x 103 (time steps) = 12.8 TB of data to be saved for one run.

An XGC1 simulation for the C-Mod tokamak (MIT) uses about 100 billion particles, and a mesh of about 5 million node points, and it runs for 10,000 time steps on 170,000 processors on Jaguar for a one-day job or 120,000 cores on Hopper for a two-day job. The restart file writes out all the particle data every 1,000 time step and its file size is (1011 particles) x (9 variables) x (8 byte) + field data = about 10 TB at every 1,000 time step, or 100 TB total. If particle data from every time step is saved, it would require 100 Petabytes of storage. The physics-study files for spatial field variables from grid nodes is written every 10 time steps, and the file size of each time step is (5 x 107 data points) x (5 variables) x (8 byte) = 2 GB. The total file size of the grid field data (coming from a 10,000 time step of simulation) is 2 TB. One simulation wall time is about one day on Jaguar and two days on Hopper. Particle data is also needed for a more complete understanding of underlying physics, such as wave-particle interaction in phase space. However, it is prohibitively expensive at the present time to write out the 10 TB particle data at every 10 time steps, as this would total 10 PB. The I/O of XGC1 utilizes the ADIOS library, which enables parallel I/O of more than 200 Gbps. Hence, writing out the restart file takes about 5 min, and the local OLCF file-system network speed is fast enough. For transfer of the last restart file and the physics-study files from the scratch file system to a local server in one hour, the local area network speed requirement is about 16 Gbps.

The M3D-C1 code utilizes a fully implicit algorithm that allows it to take large time steps as required to simulate slowly growing instabilities. A typical run today will use 3 x 105 high-order finite element nodes to represent a tokamak. A large job will run for 750 hours on 1536 processors for a total of 1.1 million processor-hours. Each node requires 12 numbers to represent a single scalar variable, and there are typically 8 scalar variables resulting in 30 x 106 words or 2GB of data generated each time step. This is also the size of a restart file. Typically, not all of this data is stored; however making a movie requires data from at least 100 time steps for a file size of 200GB. The restart files are written with ADIOS and the graphics files are written with parallel HDF5.

TRANSP facility typically produces a 20GB file from its larger runs.

4.4 Near-term Remote Science Drivers

PPPL utilizes ESnet to access data at the supercomputer centers and other domestic and international locations and to provide access to the PPPL computing facilities for users worldwide.

4.4.1 Instruments and Facilities

PPPL researchers are heavily involved with offsite fusion projects within the United States and overseas, particularly the C-Mod and DIII-D experiments in the United States, the JET experiment in England, KSTAR in Korea, EAST in China, and ITER in France. A current project allows PPPL-based researchers to quickly analyze results from KSTAR and EAST and return valuable analysis to the operation staff local to the experiments. This capability is vital to fusion research since the newest reactors are those built overseas.

Access to the TRANSP analysis tools are provided to physicists worldwide via the Open Science Grid (OSG), and it is being actively used, with TRANSP runs from outside PPPL accounting for approximately half the total number of runs. The file size from larger TRANSP runs is 20GB, and is usually transferred back to the user's local storage.

The current 10Gbps connection enjoyed by PPPL is sufficient to support this research, as we have not seen any peaks approaching even a quarter of the peak theoretical throughput.

PPPL has deployed the popular NX product which optimizes XWindows traffic over long hauls, thus making it often unnecessary to transfer large data files back to the a local site.

4.4.2 Process of Science

Mostly, data analysis for the massively parallel codes is performed at a remote supercomputing site where the data is generated. However, for advanced analysis and visualization, the whole physics-study files need to be transferred to a local server. The data size of the whole physics-study file can range greatly depending on the code and the number of time points. However, in the examples cited above, it is about 4TB for GTS, 4TB for XGC1, and 200GB for M3D-C1.

In some situations, transferring the restart file (4 TB for GTS, 10 TB for XGC1, and 2 GB for M3D-C1) between OLCF and NERSC is required. For the largest of these, XGC1, and assuming 4 hours of transfer time, it requires 5 Gbps of network speed.

MDSplus is the main data archival and storage utility; it's near-universality (i.e., it is being used internationally) make the transfer of data to or from remote collaborators seamless. It also contains implicit functionality that keeps track of all levels of data, from raw to fully processed, including updates.

One set of data not amenable to MDSplus is time-varying 2D fast camera image data, videos of which are stored in their own format. The data is typically analyzed using procedures developed within the IDL, MATLAB and PYTHON frameworks. These procedures are highly transferable to remote collaborators as well as within the local group members. Data displays and discharge information are web-accessible through browsers for use either during the run or after. Finally, the real-time Run Display Wall plots can be accessed by remote users via the PPPL cluster. Large theory codes are often used, and key to timely completion of these runs is speed and number of processors.

Data transfer rates to remote collaboration sites in Korea and China are much less than to domestic sites like NERSC and ORNL. While 2.5Gbps to NERSC is achievable in a memory-to-memory copy, the best result to Korea is about 400Mbps (using perfSONAR testing).

4.5 Medium-term Local Science Drivers

4.5.1 Instruments and Facilities

PPPL's local computing and storage resources will continue to grow to meet the need for small to mid scale jobs. Computing resources will increase in capability as the density of new high-core count processors increases. Storage will grow at its historic rate of 30% annually, so storage resources will reach approximately 1.2PB in two years, and close to 2PB in 5 years.

4.5.2 Process of Science

In five years, the number of particles and mesh points used by each of the codes GTS, XGC1, and M3D-C1 will increase by an order of magnitude. Also, new physics and new variables will be added. The data size is anticipated to be about 10 times the present levels. If we require the same transfer time, this will require 10 times the transfer rates.

With NSTX-U operation, full TRANSP runs for every good plasma discharge will be required. This can be initiated during each operational day, and continue even after operations have concluded for the day. Consequently, this will require partially automated run submissions and dedicated processors. It is also expected that the demand for TRANSP runs from EAST, KSTAR, JET D-T operation and ITER preparation will increase during this time period. For the ITER work, this means good connection to the ITER Integrated Data Model.

Tools to display and analyze data from remote facilities are pretty much the same as for local use. A key difference may be the requirement to write site-specific conversion routines when data storage at the remote facility is not easily adaptable to the PPPL environment and tools. Downloading, compiling and running large theory codes from off-site has been challenging on occasion. GITHUB is used in some

cases for code downloads, but more standardization and ease-of-use is required. Remote participation in experiments outside PPPL will require reliable audio/video capabilities and fast access to remote data. The latter could be accomplished by real-time mirroring of data, which has already been done successfully on other U.S. facilities, but is not optimal for very remote sites, like those in China and Korea. Good and reliable remote access requires solving any bottlenecks due to cybersecurity issues, with the objective of maintaining security without degrading performance.

4.6 Medium-term Remote Science Drivers

4.6.1 Instruments and Facilities

PPPL's wide area network traffic is typically well under 5% of capacity. We do not foresee any changes in the next few years that will require an increase in bandwidth from the current 10Gbps capability.

4.6.2 Process of Science

Local science drivers are expected to remain the same over the next 2 to 5 years in terms of data acquisition, analysis and management, although with at least one order of magnitude increase in data once NSTX-U starts operating. Depending on licensing agreements, emphasis in the use of IDL vs. MATLAB vs. PYTHON may shift. Real-time processing of a number of data sets will be required for implementing and optimizing various control schemes, including disruption and beta control and current and rotation profile control. The use of large simulation codes is expected to increase, in conjunction with further development and comprehensiveness of these codes (XGC1, GTS, M3D-C1, to name a few). The size of the restart file for XGC1 is forecast to grow to 24TB in the next 2-5 years.

In addition to the ability to download, compile and run simulation codes from off-site effectively, a key challenge for remote science drivers is the ability to develop effective communications between PPPL and remote experiments. This is driven by strong and active collaborations expected between PPPL and experiments in Asia, Europe and the U.S. The requirements for this include reliable video and audio connections for a "Remote Control Room," and reliable and fast access to remote data. Development of computer capability for remote control of some aspects of the experiment or diagnostic configurations would be beneficial. More open access to information from experiments in sensitive countries (through trusted sites) is required.

In that context, we are having discussions about the best method of addressing large data sets generated at remote experimental sites. Method A is to transfer these data sets to local storage, then act upon them using local compute resources. Method B is to act upon the data sets using compute resources at the experimental site, and only connect to the resulting data graphs and other visualization tools via low-bandwidth video tools like NoMachine/NX.

Method A requires optimized network transfer mechanisms, and considerable local storage. The best we've been able to achieve for memory-to-memory copy of data from KSTAR, for example, is 562Mbps. Disk-to-disk copy is drastically slower at 3.2Mbps.

However, Method B has shown itself to be very usable with minimal delays or cutouts. This method has been used with KSTAR and EAST with acceptable results. Perhaps we should strive to optimize video traffic to/from these very remote sites (possibly using QoS and/or dedicated circuits).

4.7 Beyond 5 years

It is expected that TRANSP use, especially for KSTAR (Korea), EAST (China), and ITER preparation, will increase substantially.

However, the main fusion science driver beyond five years is the ITER project in Cadarache, France. Optimistic sources cite a startup date of 2023, with full operation in 2026. The estimated data production is estimated to be 100–2000TB/day. Transfer of this data to the United States will be required, but it may be that the entire data set need not be transferred and the percentage to be transferred remains to be determined.

Remote experiment participation will be very important for U.S. researchers collaborating with ITER, KSTAR, EAST, and other remote experiments. This participation will depend upon high quality video and audio connections.

4.8 Data, Workflow, Middleware Tools and Services

For transferring data over the network, scientists use *bbcp*, *gridftp*, *scp*, and *ftp*. A Globus server has also been setup at PPPL for transfer to/from the DOE Leadership Computing sites. Scientists at PPPL use VisIt, IDL, and MATLAB for visualizing data remotely (for example at NERSC). NX is well supported for connections to remote sites, which greatly facilitates the response time when viewing data. OSG is used to provide access to TRANSP for remote users.

Table 4.1: Table summarizing this case study's network expectations and foreseeable science drivers for the next five years and onward.

Science Drivers			Anticipated Network Needs	
Science Instruments, Software, and Facilities	Process of Science	Data Set Size	LAN Transfer Time Needed	WAN Transfer Time Needed
0-2 years				
• GTS • XGC 1 • M3D-C • TRANSP	• Core microstability • Edge microstability • Global stability	• 4 TB (graphics) • 4 TB (graphics) • 200 GB (graphics) • 10–50GB	• Transfer restart files from scratch to archive disk in 1 hour	• Transfer graphics files to local host in 1 hour
2–5 years				
• GTS • XGC 1 • M3D-C • TRANSP	• Core microstability • Edge microstability • Global stability	• 12 TB (graphics) • 24 TB (graphics) • 2.4 TB (graphics)	• Transfer restart files from scratch to archive disk in 1 hour	• Transfer graphics files to local host in 10 hour
5+ years				
• Existing codes and new codes	• Higher resolution and new physics couplings	• 24 TB (graphics)	• Transfer restart files from scratch to archive disk in 1 hour	• Transfer graphics files to local host in 1 hour

Case Study 5

U.S. ITER Project

5.1 Background

The United States is a partner nation in ITER, an unprecedented international collaboration of scientists and engineers working to design, construct, and assemble a burning plasma experiment that can demonstrate the scientific and technological feasibility of fusion energy. The U.S. ITER project is a DOE Office of Science project hosted by Oak Ridge National Laboratory in Tennessee. Partner laboratories are Princeton Plasma Physics Laboratory and Savannah River National Laboratory. ITER's other partners are the People's Republic of China, the European Union, India, Japan, the Republic of Korea, and the Russian Federation. Involvement in ITER provides significant benefits for the United States for a limited investment (less than 10% of construction costs): The United States has access to all ITER technology and scientific data, the right to propose/conduct experiments, and the opportunity for U.S. universities, laboratories and industries to design and construct parts.

The U.S. ITER project engages more than 400 companies and universities in 40 states plus the District of Columbia. As of March 2014, over $616M has been awarded to U.S. industry and universities, and obligated to DOE national laboratories. Over 80% of U.S. ITER project funding is spent in the United States. During the construction of ITER components throughout the United States and work with other ITER partners, collaborative tools are heavily used. As ITER moves from basic infrastructure to commissioning and operations, additional network needs will be required for data transfer and monitoring of ITER components, analysis of commissioning data, and eventually, the sharing and analysis of scientific experimental data.

5.2 Collaborators

The U.S. ITER project is an intersection of two collaborations. Worldwide, it is the U.S. domestic agency for the international ITER project. With the international collaboration, the U.S. ITER project collaborates with the ITER International Organization (ITER-IO) and ITER domestic agencies. Nationally, U.S. ITER coordinates all of the engineering, R&D, fabrication, delivery, and project management efforts between the partner laboratories, vendors, and the ITER international efforts.

5.3 Near-term Local Science Drivers

5.3.1 Instruments and Facilities

Development and operation of test stands supporting pellet injection, RF and vacuum systems, and Control, Data Access and Communications (CODAC) system will continue over the next few years. In addition, general engineering and project management efforts will continue for the life of the ITER project.

The local area network infrastructure consists of in-building 1 Gbps Ethernet connected to a lab-wide 1-10 Gbps fiber optic network infrastructure. Test stands and the CODAC development network are private networks protected by firewalls that limit ingress and egress of data.

In addition to our development needs, our video and web collaborative infrastructure consists of:

- ORNL ITSD supplied Bluejeans Cloud-based videoconference as a service (www.bluejeans.com)
- ORNL ITSD supplied Cisco 4215 MCU and VCS/VCSe pair with redundancy.
- ORNL ITSD supplied Lync 2013 servers
- Polycom MGC+50 Multipoint Control Unit for general U.S. ITER needs

The U.S. ITER primary Internet access is via the ORNL ESnet primary 100 Gbps link and a backup 10 Gbps link to an additional ESnet hub. In addition to our ESnet service links, we are also connected to SoX, the Southern Crossroads Network at 10 Gbps, which also peers to ESnet and Internet2.

U.S. ITER servers consist of two sets of servers located in the U.S. ITER data center and test stands throughout ORNL. The first set of servers are general purpose IT systems for project support and collaboration. Project support systems include:

- U.S. ITER project web, application, database, and file servers
- U.S. ITER Document Management System (iDocs)
- Videoconferencing infrastructure

The general purpose IT servers contain 25 TB of data storage and is growing at a rate of about 5 TB/year. All project data is backed up to the central ORNL/ITSD backup system in a different location.

The U.S. ITER technical servers are used for CODAC development and ENOVIA/CATIA computer-aided design (CAD) model repository replication. They support:

- ITER collaborative network servers for ENOVIA replication
  - Up to 13TB of data replication to ITER-IO has been needed for daily synchronization
  - A Riverbed Steelhead WAN optimization tool is used for increase data throughput for general IT systems that are not designed for more efficiency through high latency links
  - 20TB of storage

The U.S. ITER test stands that are in operation or being created over the next 0-2 years are:

- CODAC development cubicle
- Pellet Injector Lab
- Prototype vacuum pump testing
- Virtualized CODAC servers for training purposes
- ICH test loop
- ECH test loop
- ECH waveguide component test

The test stands are segmented from the standard ORNL network and connected via firewalls. Normally, most test stands will not generate a great deal of traffic outside of their local network. The CODAC development cubicle will require periodic synchronization of CODAC code with ITER-IO.

5.3.2 Software Infrastructure

U.S. ITER uses a number of engineering and project management tools. Most of these tools are commercial off the shelf (COTS) packages. A few are custom applications created and written by ORNL for custom project needs.

U.S. ITER uses a 3D CAD system from Dassault Systems called Computer Aided Three-dimensional Interactive Application, or CATIA. CATIA utilizes a database and file storage system to store the models called ENOVIA. Due to the latency between Oak Ridge, TN and Cadarache France of approximately 160 ms, an ENOVIA replication system was created to allow local access to the ENOVIA system. Therefore, the network traffic is reduced to a single daily bidirectional update of the changes. All ITER domestic agencies except Europe utilize this replication capability.

For document and records management, U.S. ITER uses a copy of the ITER Document Management system (IDM) located at U.S. ITER and is called iDocs. iDocs is accessible by the U.S. ITER project office and ORNL personnel working on U.S. ITER by directly accessing the web interface. All other partners access it via a Citrix interface.

3D visualization is needed to help visualize the complex structures of ITER to the team. A product from IC.IDO (www.icido.com) is used for this. In addition, it allows the remote co-visualization over high-latency links with little degradation of quality. This has been demonstrated with Wendelstein 7-X group in Germany. In addition to the IC.IDO technology, visualizations of ITER capabilities have been demonstrated using the Everest visualization facility at ORNL.

U.S. ITER's project management tools consists of COTS tools such as Oracle Primavera P6, Cobra, Microsoft SharePoint and other general IT products. In addition, web-based tools developed at ORNL are used to support the project.

5.3.3 Process of Science

U.S. ITER's process of science is based on collaboration and management of the construction project. As mentioned before, U.S. ITER is a confluence of two collaborations: the national collaboration and the international collaboration.

U.S. ITER works with ITER-IO and other domestic agencies on the integration and engineering of the overall design and scheduling of the component and activities. U.S. ITER works with partner laboratories and vendors to accomplish the goals by collaborating with them to coordinate the design and fabricate the necessary deliverables.

5.4 Near-term Remote Science Drivers

5.4.1 Instruments and Facilities

The U.S. ITER primary Internet access is via the ORNL ESnet primary 100 Gbps link and a backup 10 Gbps link to an additional ESnet hub. In addition to our ESnet service we are also connected to SoX, the Southern Crossroads Network at 10 Gbps which also peers to ESnet and Internet2.

U.S. ITER's WAN needs normally require quality connections to U.S. ITER partner laboratories, vendors such as General Atomics, ITER-IO in Cadarache France, and ITER domestic agencies in China, Europe,

India, Japan, Republic of Korea, and the Russian Federation. Our current needs for connectivity for collaboration and design are being met.

5.4.2 Software Infrastructure

Since U.S. ITER is a collaboration of partner laboratories, and with the ITER-IO in general, all of the U.S. ITER tools are essentially "wide area network" based. That has been the design philosophy of U.S. ITER. (See Section 5.3.2 for a summary.)

The most significant tools we use that are significant from a wide area network perspective is the ENOVIA replication due to latency issues, and the collaborative tools to allow remote conferencing to make meetings possible over such long distances.

At this time, ITER is not replicating experimental data since it is still under construction.

5.4.3 Process of Science

See Section 5.3.3.

5.5 Medium-term Local Science Drivers

5.5.1 Instruments and Facilities

In 2–5 years U.S. ITER and partner laboratories will start to see more participation in analysis from existing facilities in order to begin developing even better collaboration capabilities for remote experiment participation in ITER and for developing U.S. ITER components. As components get delivered, remote monitoring of the CODAC output will start to occur at lower data rates. This is important since travel is expensive and time consuming. In the future, replication of the IDM system and the U.S. version called iDocs will occur. This will ensure that all records for the project are captured as well as the 3D CAD models in ENOVIA that are already being replicated.

5.5.2 Software Infrastructure

See Section 5.3.2 for summary.

5.5.3 Process of Science

Many of the tools from years 0–2 will be the same as in 2–5 years. The additions will be newer collaborative tools and the beginning of monitoring and data transfer tools. The process of selecting tools has not been completed and will be based on COTS tools and other scientific facilities' tools that have dealt with these issues, such as those used at the CERN Large Hadron Collider.

5.6 Medium-term Remote Science Drivers

5.6.1 Instruments and Facilities

In 2–5 years U.S. ITER will evolve from a purely engineering and fabrication project to a commissioning project as components are delivered and installed. Therefore the collaboration aspects will start to outpace the engineering and 3D CAD needs in terms of the volume of data and participation. Collaborative tools will evolve and data transfer needs will start to occur.

5.6.2 Software Infrastructure

See Section 5.3.2 for a summary. All of the tools will continue to be used. Collaborative tools will continue to develop beyond standard videoconference and desktop sharing.

5.6.3 Process of Science

All of the current processes will remain but an emphasis on the installation and commissioning will replace design and fabrication for some components. Remote access, data transfer, and remote facility monitoring will become more dominant in the next 2–5 years.

5.7 Beyond 5 years

Beyond the five-year timeframe, ITER will move from concept to a physical facility to perform fusion research. Along the way, it will go through commissioning and subsystem operations phases. U.S. ITER support personnel will be monitoring this remotely to reduce travel and increase partition in the efforts. Demonstration and development of the diagnostics systems will start to produce large data sets, and data reduction and analysis will begin. Eventually, full operations and experiments will occur. This will greatly increase ITER network requirements to the terabytes per day range with fast data transfers to U.S. systems for data reduction and analysis for next-shot planning and longer term research.

This will require robust collaboration tools, data replication tools, and analysis tools and facility infrastructure throughout the United States for the U.S. participants to remotely collaborate. High-speed analysis of ITER experiments will help reduce the cycle time from shot to shot. Using U.S. and international computational capabilities, the next shot will be based on data analyzed from the previous shot. The estimated data production is estimated to be 100–2000TB/day. Transfer of this data to the United States will be required, but it may be that the entire data set need not be transferred and the percentage to be transferred remains to be determined.

In addition to experimental data, operational monitoring data will be transferred to the United States for analysis of the ongoing operation of U.S.-supplied systems and for remote read-only “control room environment” for experiments to aid in the process of experimentation.

5.8 Network and Data Architecture

Data transfer nodes and large storage pools for data storage will be needed. In addition, ITER will need computational capabilities to analyze the data and merge it with simulated data. Visualization capabilities to render the output in 3D to understand the time evolution of ITER shots will be used by researchers from desktops, tablets, and visualization facilities, such as Everest at ORNL.

5.9 Data, Workflow, Middleware Tools and Services

Many of the data workflow and middleware tools will be leveraged from existing big data projects around the world including those from high energy physics and astronomy where large replicated data sets are placed on the various continents for lower latency access and disaster recovery needs.

A new driver in terms of analysis and possibly storage needs are cloud-based capabilities. Capabilities such as Amazon Web Services and Microsoft Azure will become a commodity computation capability that can be used for fusion. While this will not replace supercomputing, it can be an efficient mechanism to bridge the boundary between smaller computation on desktops or small clusters to the large supercomputers. As costs go down, this needs to be viewed as a very flexible option. ESnet is encouraged to create high-speed capabilities to bridge these cloud services to DOE and university facilities to increase throughput. Negotiating a reduction in data transfer costs with the cloud services when using ESnet network capabilities would help drive down costs and be able to increase the value proposition of using cloud services.

5.10 Outstanding Issues

U.S. ITER has always had latency issues due to the world-wide nature of the project. We have overcome some of this by using wide area network optimization tools such as the Riverbed Steelhead device to increase communication throughput. This device mitigates problems with devices that only use standard network methods; such as standard IT servers. It would be helpful if a standard mechanism was used by (and through) ESnet and the world's research facilities to help standardize IT needs. While custom designed data transfer nodes can be used for scientific data, other more mundane IT systems need WAN optimization as well.

QoS and real-time data transfers will become an issue throughout the U.S. ITER network needs. Using QoS can help mitigate issues related to VoIP, VideoIP, and real-time data needs. Therefore, ESnet should ensure QoS to each facility and collaborative network is maintained and encouraged to each device.

Table 5.1: Table summarizing this case study's network expectations and foreseeable science drivers for the next five years and onward.

Science Drivers			Anticipated Network Needs	
Science Instruments, Software, and Facilities	Process of Science	Data Set Size	LAN Transfer Time Needed	WAN Transfer Time Needed
0-2 years				
• Collaboration systems (VC, web conferencing, etc) • ENOVIA/CATIA 3D CAD models • Project files and records • CODAC development cubical • Test stands	• Collaborative tools used to extend project influence throughout U.S. and ITER domestic agencies • Replication of 3D CAD system and CODAC development cubical from ITER-IO to U.S. and other domestic agencies	• CAD database is 200TB • Project file storage is 25TB	• LAN access is typically for general project access, data replication, and test stand/-CODAC operation transfer for a single replication	• 13 TB ENOVIA transfer for a single replication from ITER-IO to Oak Ridge • WAN transfer used for video-conferencing
2–5 years				
• Continue previous requirements • Increased participation in remote experimentation and data transfer develop ITER collaborative capabilities	• Continue previous requirements • Moving towards transferring more data for collaborating on experiments to CAD system and CODAC development cubical from ITER-IO to U.S. and other domestic agencies	• Continue previous requirements • Additional collaboration needs and data transfer for beginning of commissioning requirement for U.S. analysis	• Continue previous requirements • Increases will keep pace with additional WAN requirements transfer for a single replication	• Continue previous requirements • Increase in WAN data transfer requirement by 5-10TB
5+ years				
• Continue previous requirements • Operational collaboration software for remote experimental participation and possible diagnostics control • Experimental data replication between ITER-IO and U.S.	• Continue previous requirements • Move from project construction orientation data for collaborating and operating facility; CAD system and data sets and operational data sets will become dominant	• Continue previous requirements • 10–20TB per operating shot with multiple shots per day commissioning	• Continue previous requirements	• Continue previous requirements • The estimated data production is about 100–2000TB/day • Data transfers to the U.S. will be required, but may only need partial data relocation; this is still to be determined

Case Study 6

International Collaboration Framework for Extreme-scale Experiments

6.1 Background

Large-scale data exploration at the Korea Superconducting Tokamak Advanced Research (KSTAR) facility in Daejon, South Korea is based on international collaborations. As this collaboration produces more and more data, the existing workflow management systems are hard pressed to keep pace. A necessary solution is to process, analyze, summarize and reduce the data before it reaches the relatively slow disk storage system, a process known as in transit processing (or in-flight analysis). We have been researching and developing new data handling techniques for collaborative workflow systems and have been integrating filtering/indexing techniques in our WAN data staging, which we have built into ADIOS.

The keys to a successful collaboration system are in time-to-knowledge, which involves WAN data movement from remote facilities to ORNL and NERSC, so that data can be analyzed in near-real-time (NRT) and feedback from the workflows are sent back over to the researchers close to the experimental facility. There are several key factors which must be understood in order to expedite this knowledge processing of data. They typically are

1. What is the lifetime of the instability, and how quickly can we understand the instability? Typical instabilities which we are looking to investigate in our framework range from disruptions, to Edge Localized Modes, etc. The data from KSTAR is typically on high-velocity, low-volume except for fast-camera data. Typical fast camera data come from KSTAR electron cyclotron emission imaging (ECE-Imaging) and soft X-ray imaging. The setup of this diagnostic is shown in Figure 6.1. The data from this diagnostic comes in quickly and in order. The goal is to ultimately take this data and compare it to simulations such as XGC1 or Bout++. Since this data comes in quickly (microseconds), we believe that the data needs to be quickly moved (every second of the shot) to the collaborators at ORNL and NERSC. The data needs to be analyzed to detect blobs and track the blobs, and then find (or launch) simulations to compare this data. A comparison of the data is shown in Figure 6.2. This workflow involves the indexing/filtering of the image data over at the side from Korea, reduction of the data chunks which satisfy the query (for the location of the blobs), the data movement of the queries, and then finally the removal of noise and the generation of the features. Once this is done, the data is then compared to previous simulations, and if one does not exist then a new simulation could be performed. In this scenario the data is connected to a buffer server which has a 10 Gbps link. Since the lifetime of the shot ranges from tens to hundreds, we envision that this data will move point to point from KSTAR to Korea Institute of Science and

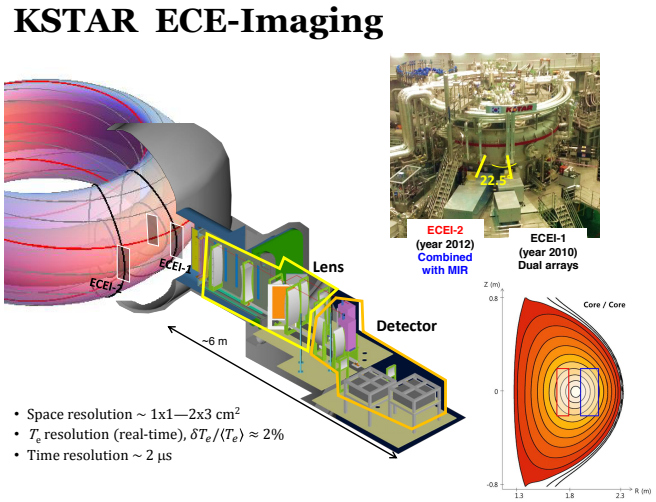


Figure 6.1: The ECE-Imaging diagnostic on KSTAR.

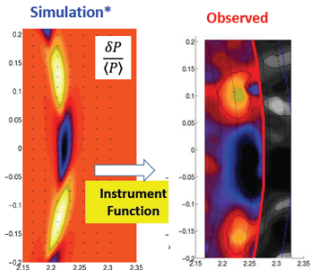


Figure 6.2: ECE data compared to Simulation data.

Technology Information (KISTI) and then over to ORNL and PPPL. The total data from just this one diagnostic is 4 GB.

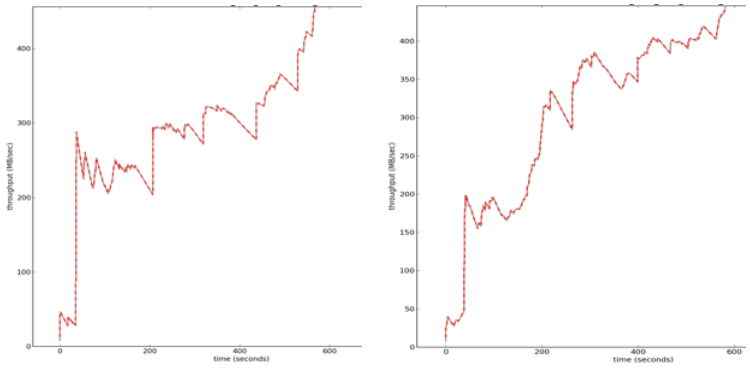
2. The goal from example 1 only involves one diagnostic, but there are many diagnostics which are becoming available. Furthermore, the data processing locally will need to be in the order of 1 ms. Figure 6.3 shows the expected data generation rates, and it is clear that the image data will be the largest data generators for KSTAR.
3. A critical aspect of International Collaboration Framework for Extreme-scale Experiments (ICEE) is working with both experimental and simulation data. Ultimately the data from simulations need to move to the experimental facility if they are to be mined for NRT processing of the data during long-pulse experiments. The questions which still remain are
   (a) Will fusion facilities be able to hold large simulation data, and if so, how much? Data from an XGC1 simulation can approach a petabyte, and this large scale data movement will mostly occur from file-based data movement. We envision that KISTI will be a temporary place to hold simulation data for NRT decisions.

6.1.1 Current Network Connectivity Testing from NISN to NERSC

In our recent data throughput testing from one National Institute of Supercomputing and Networking (NISN)/KISTI machine to one NERSC/PDSF machine over a 10 Gbps link, transferring 260GB with 488 files in total. The cumulative average throughput over time was achieved about 450MB/sec (3.6 Gbps), over the shared network. GridFTP with adaptive module is used for transfers.

Data Type	Storage	Campaign	P. Length	Data/shot	Time/shot
Engineering data (EPICS data)	Main	1.1TB (2008)			
		420GB (2009)			
		637GB (2010)			
		675GB (2011)			
Experimental data (MDSplus data)	Main	204GB (2008)			
		503GB (2009)	3.5s		
		1.2TB (2010)	5s		
		1.7TB (2011)	10s	~1GB (max 1.5GB)	~2min
		> 2TB (2012)	>10s		
			300s	~ 100GB	
Diagnostic TVs	IMAGE	19GB (2008)			
		42GB (2009)	3.5s		
		230GB (2010)	5s		
		3.9 TB (2011)	10s	~ 2GB	
		>4TB (2012)			
ECEI	Distribute	3.7TB (2011)	10s	3.6GB	~2min
		> 7TB (2012)	>10s	7GB	
		?	300s	?	

Figure 6.3: KSTAR data generation.



(a) (b)

Figure 6.4: (a) Measured at July 4, 2014, 05:15AM KST, starting with 20 concurrent transfers, each having 4 parallel streams, up to 64 concurrent transfers increased by 4 concurrency, (b) measured at July 6, 2014, 12:50PM KST, starting with 20 concurrent transfers, each having 8 parallel streams, up to 128 concurrent transfers increased by 4 concurrency.

6.2 ICEE Requirements

Since our project is concerned with NRT decisions, low-latency data movement is critical, and file-based data movement is not optimal for our investigations. We see that we need to move the 10 TB of data from KSTAR to ORNL/PPPL/NERSC during the 100 s shot. This equates to 100 GB/s, but does not take into account the high level of data reduction which can be done during the experiment. We see that the common workflows that are performed with the fast imaging data utilize less than 10% of the total data size. Thus, we will need ultimately to move approximately 10 GB/s, and process the data directly from memory-to-memory, without having the file system step in the way.

6.3 Collaborators

ICEE is supporting collaborators from PPPL, LBNL, and ORNL. Data will move from KSTAR to KISTI to ORNL.

Case Study 7

Fusion Simulations: XGC Program

7.1 Background

Experimentalists found over thirty years ago that the fusion performance in a tokamak core is determined by the edge plasma condition. Understanding the edge plasma condition and its nonlocal effect on the core plasma condition has been a difficult issue because of the non-thermal equilibrium, multi-scale nature of the edge plasma and the global nature of the edge-core interaction that must include turbulence physics in the whole plasma volume from magnetic axis to the material wall. In order to address this issue, the XGC program was launched in 2005 as part of the SciDAC program, Proto-FSP Center for Plasma Edge Simulation (CPES). The XGC program has continued as the main simulation tool in the SciDAC-3 Center for Edge Physics Simulation (EPSI).

The XGC program has three kinetic particle codes XGC0, XGCa and XGC1; simulating tokamak plasmas in realistic diverted magnetic field geometry with the neutral particle recycling at the material wall. XGC codes simulate the background plasma physics together with the perturbed physics in multi-scale by using the full-function (full-f) particle technique, instead of the popular perturbed function (delta-f) technique in which the background plasma is not simulated. By containing more physics than the delta-f codes, the XGC simulations require more powerful computing, produce bigger data, and thus requiring much more network bandwidth.

XGC0 usually runs on Hopper at NERSC, using up to 80,000 cores for 20 hour or less wall-clock job. XGC1 mainly runs on Jaguar at OLCF, using up to 170,000 cores for one-day wall-clock job for DIII-D plasma (longer for C-Mod). XGC1 also runs on Hopper, using up to 120,000 cores for two-day wall-clock job for DIII-D plasma (longer for C-Mod) with restart submissions. XGC1 produces large size data, which are handled by the Adaptive IO System (ADIOS) framework. The data sizes for these simulations are increasing rapidly as more physics is included in the simulation. The network requirements, as described in this case study, are based on XGC1 since it produces much more data than XGC0. For the time being, the present HPC systems (Jaguar and Hopper) allow XGC1 to simulate ion-scale turbulence for DIII-D, C-Mod and NSTX plasmas. Within a couple years, it is expected that HPC systems will allow XGC1 to simulate electron-scale turbulence on the present experimental devices.

XGC0 is a drift-kinetic particle code with a 1D electrostatic potential solver. XGCa is a gyrokinetic particle code with a 2D electrostatic potential solver. XGC1 is a gyrokinetic particle code with a 3D electromagnetic field solver with microturbulence solutions. All three of them are highly scalable HPC codes. Since XGC1 solves for plasma macro-tubulence, it has the greatest number of grid cells and marker particles, hence producing the biggest data to communicate over the network. For this reason, this case study will be based upon the requirements set by the XGC1 code.

7.2 Collaborators

Within SciDAC-3 EPSI the XGC program performs collaborative research across 11 national laboratories and universities: PPPL, ORNL, LBNL, MIT, Rensselaer Polytechnic Institute (RPI), Rutgers University, Lehigh University, The University of Texas at Austin, University of Colorado, University of California San Diego, and University of California Davis. Among these, only PPPL, ORNL, Rutgers University, and The University of Texas are actually executing the code runs on major U.S. HPC systems. The rest of the participants get the simulation data transferred for physics or algorithm analyses.

7.3 Near-term Local Science Drivers

7.3.1 Instruments and Facilities

For capability computing, XGC1 uses Titan and Mira, with the production runs utilizing up to 88% of the maximum heterogeneous Titan capability and $\frac{1}{3}$–$\frac{2}{3}$ of the maximum Mira capability. A capability computing normally lasts for a few wall-clock days, using the checkpoint restarts. For smaller size capacity computing, XGC1 uses Edison and Hopper at NERSC, utilizing 10,000 to 100,000 cores. The capacity computing at NERSC can take a few days of wall-clock time to finish one physics study. We normally store the large checkpoint and analysis files in the storage space provided at the HPC sites. We would like to move some important checkpoint data (a couple terabytes per check pointing) and analysis data (up to 1TB) to remote sites. Numerous small size runs are executed on PPPL's computing cluster for developing new physics routines for XGC1, for optimizing input parameters before submitting large runs on HPC systems, and for small-size XGC0 and XGCa science runs.

7.3.2 Software Infrastructure

We normally use MATLAB to analyze data and to produce preliminary visualizations. For a more detailed visualization VisIt is used by ORNL collaborators. SCP has been used to facilitate the transfer of small size data, which is found to be inadequate for the transfer of large size analysis and checkpoint data. We plan to use Globus in the near future.

7.3.3 Process of Science

A typical one-day XGC1 simulation with kinetic ions, electrons and neutral particles for the NSTX tokamak uses about 60 billion particles, and a mesh of about 8 million node points, and it runs for 8,000 time steps on 16,384 heterogeneous Titan nodes (about 20 petaFLOPS) for a one-day job; or using 30 billion particles on 262,144 Mira cores (3.3 petaFLOPS) for 2,500 time steps. The restart file writes out all the particle and field data at every 500 time steps. Its file size is (0.6x1011 particles) x (9 variables) x (8 byte) + field data = about 5 TB at every 500 time steps for total of 80 TB from Titan; or about 2.5 TB at every 500 time steps for total of 12.5TB from Mira. If the particle data from every time step are saved for the maximal physics study, it would require 40 PB of storage per one-day simulation on Titan. The physics-study files for spatial field variables from grid nodes only are (5 x 0.8x107 data points) x (5 variables) x (8 byte) = 1.6 GB per time step. The total file size of the grid field data (coming from an 8,000 time step of one-day simulation) is 13 TB from Titan, or 4 TB from Mira. Since it is prohibitively expensive at the present time to write out all the physics data, we limit the physics data output to about 1TB at the present time. If we want to transfer one important restart file per one-day run, which is about 5TB for Titan and 2.5 TB for Mira, the average data that need to be transferred to PPPL after a one-day simulation is 4TB.

7.4 Near-term Remote Science Drivers

7.4.1 Instruments and Facilities

XGC1 uses ESnet. As described above, one restart file is about 5 TB from Titan, totaling 80 TB for a one-day simulation. At the present time, moving all the restart files to local storage is impractical because of the WAN and the PPPL storage limit. Thus, we tend to erase the past restart files. If we need to revisit the simulation at a later time, we have to start the simulation from the beginning.

Data analysis at OLCF and NERSC is well-supported. Thus, we do not need to move all the analysis data (1TB per one-day run) to PPPL from OLCF and NERSC. Only part of the data, such as the graphics data, is moved to PPPL and to the collaborating institutions.

However, the data analysis at Mira is not well-supported. About 1 TB of physics data need to be transferred to PPPL. Using SCP, this has been a problem. We will be using Globus in the near future to resolve this issue. As the I/O and the network speed increases in the future, we can increase the physics data size.

7.4.2 Software Infrastructure

We use ADIOS-BP file format. We transfer these files using SCP, which is the most time-consuming bottleneck in the worksflow. In switching to Globus, we are expected to achieve much faster data transfer performance. We will also use the technology developed in the ICEE project (International Collaborative for Extreme-scale Experiment).

7.4.3 Process of Science

Since the data size is large in the XGC1 simulation, we tend to analyze the simulation data at OLCF and NERSC using MATLAB. VisIt is used for visualization by the collaborators. However, the data analysis at Mira is not well-supported. We need to move about 1TB of data from ALCF to PPPL for an adequate physics study per one-day run. This has been a problem using SCP. We hope that Globus can solve this issue. We also use MATLAB for the local data analysis.

7.5 Medium-term Local Science Drivers

7.5.1 Instruments and Facilities

XGC codes will be running on the new HPC systems at OLCF, ALCF and NERSC, which are expected to be 10X more powerful than the present HPC systems. The data size is expected to grow 10X, accordingly. A new on-the-fly, on-memory type of data analysis method is needed, that will limit the output data size to be much below the 10X rate. Most of the transferred data to local sites will be in the graphics or in a reduced form. Local data analysis will be in a different format.

7.5.2 Software Infrastructure

We expect to use the networking tools developed from the ICEE project. We also expect to use a portable dashboard tool for job monitoring, data transfer and analysis.

7.5.3 Process of Science

As the HPC power is enhanced by 10X and the data size is enhanced in proportion, it will not be practical to rely on the usual post processing tools for data analysis. We will be executing the data analysis and visualization on the simulation memory. Thus, the total amount of data on disk to be moved to the local site is not expected to grow by 10X.

7.6 Medium-term Remote Science Drivers

7.6.1 Instruments and Facilities

XGC codes will be running on the new HPCs at OLCF, ALCF and NERSC, which are expected to be 10X more powerful than the present HPCs. The data size is expected to grow 10X, accordingly. A new on-the-fly, on-memory type of data analysis will perform while XGC codes are running on HPCS at OLCF, ALCF, and NERSC.

7.6.2 Software Infrastructure

We expect to use the networking tools developed from the ICEE project. We also expect to use a portable dashboard tool that exists near the data source for job monitoring, data transfer and analysis.

7.6.3 Process of Science

As the HPC power is enhanced by 10X and the data size is enhanced in proportion, it will not be practical to rely on the usual post processing tools for data analysis. We will be executing the data analysis and visualization on the simulation memory. Thus, the total amount of disk data to be moved to the local site is not expected to grow by 10X.

7.7 Beyond 5 years

Beyond 5 years, most of our simulation will be focused on the ITER plasma, which will produce10X-100X more data. The simulation platform will be the new leadership class HPCs at OLCF, ALCF, and NERSC, which are 10X more powerful than the existing HPCs. The exascale computer will be 100X more powerful than the present HPCs. Since the produced data would be so big, a new paradigm network and data architecture will be needed. In order to reduce the output data size, in the first place, XGC codes will be aggressively moving into the on-memory, on-the-fly data management, analysis, and visualization using the ADIOS and DataSpaces framework.

7.8 Network and Data Architecture

We need to solve the firewall issue. A new concept needs to be developed for the network security.

7.9 Collaboration Tools

Videoconferencing services and similar services are vital for the collaborative advancement of fusion science. A technology that allows screen sharing, and visual and voice contact for almost unlimited number of participants at reasonable or free cost needs to be developed and maintained.

7.10 Data, Workflow, Middleware Tools and Services

The XGC program work requires tools like Globus or other data transfer tools, automated data transfer toolkits, distributed data management tools, etc. We will be testing out the Globus toolkit and other data transfer tools in the near future. We also plan to rely upon the tools developed by the ICEE project.

7.11 Outstanding Issues

A collaboration dashboard, which is directly connected to the node-memory data, will be in need in the future. The dashboard shall have the data movement, analysis and storage function.

Table 7.1: Table summarizing this case study's network expectations and foreseeable science drivers for the next five years and onward.

Science Drivers			Anticipated Network Needs	
Science Instruments, Software, and Facilities	Process of Science	Data Set Size	LAN Transfer Time Needed	WAN Transfer Time Needed
0-2 years				
• Titan at OLCF and Mira at ALCF are to be used for capability computing • Edison and Hopper at NERSC are to be used for capacity	• MATLAB will still be the main analysis tool • VisIt will be the main visualization tool	• 4TB/data set • 1–50TB/data set depending on experimental device to be simulated. Data on ITER plasma will be 10X DIII-D • 1 physics data file (0.5–10 TB) and 1 restart data file 2–40 TB)	• 10 TB in 1 hour, once a day	• 10 TB in 10 hours • PPPL, ORNL, MIT, Univ. Texas at Austin, Lehigh U., Rutgers U. UCSD, UC Davis, U. Colorado
2–5 years				
• Titan, Mira Edison until the Coral procurement • New 10X HPC systems at OLCF, ALCF and NERSC	• We plan to use the new tools developed at ICEE project • We will be moving into the on-memory, on-the-fly data management, analysis, and visualization	• 10TB/data set • 5–100 TB/data set depending upon experiment • 1 physics data file (2–5 TB) and 1 restart data file 10–100 TB	• 20 TB in 1 hour	• 20 TB in 10 hours • Collaborating, sites (e.g., data exchanged with sites A, B, and C)
5+ years				
• New 10X HPC systems at OLCF, ALCF, and NERSC from Coral procurement • Exascale computer(s)	• It will be necessary to have the on memory, on-the-fly data management, analysis, and visualization capability • We are aggressively moving into direction using the ADIOS and DataSpaces technology	• 10-1000 TB per data set • One physics file • Restart data will need to be handled differently. Its size may become too big to be moved through I/O	• 100 TB in 1 hour	• 100 TB in 10 hours • PPPL, ORNL, MIT, Univ. Texas at Austin, Lehigh U., Rutgers U. UCSD, UC Davis, U. Colorado