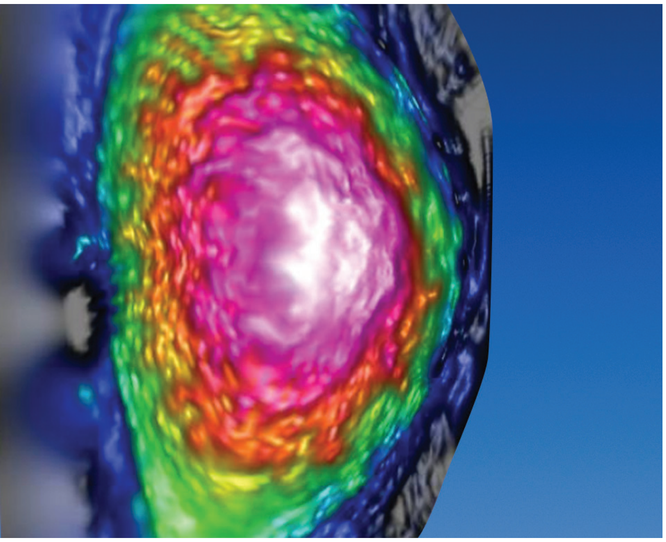
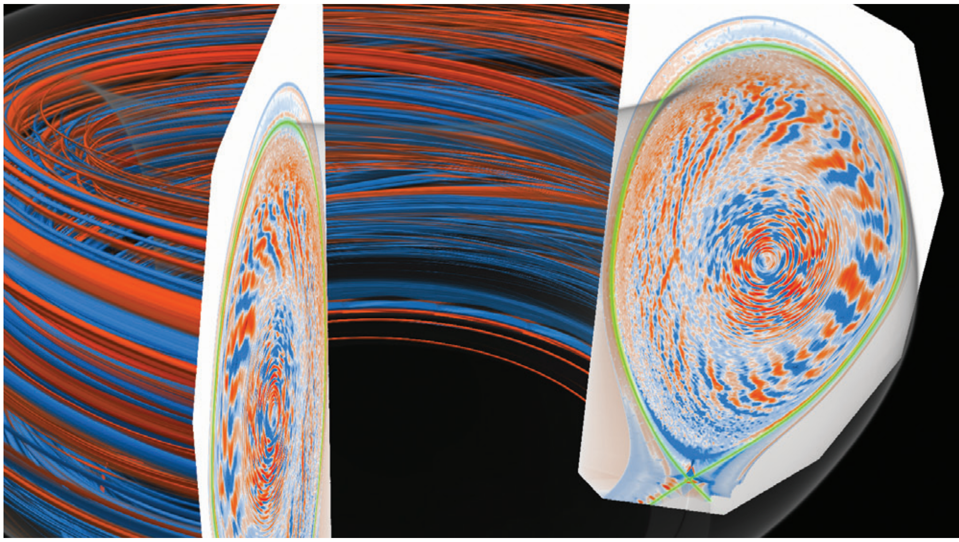
Fusion Energy Sciences Network Requirements

Office of Fusion Energy Sciences
Energy Sciences Network

Conducted December 8 and 9, 2011





Lawrence Berkeley National Laboratory

U.S. DEPARTMENT OF ENERGY
Office of Science

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.

Fusion Energy Sciences Network Requirements

Office of Fusion Energy Sciences, DOE Office of Science
Energy Sciences Network
Rockville, Maryland — December 8 and 9, 2011

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-XXXX.

Participants and Contributors

Richard Carlson, DOE/SC/ASCR (Network Research)

C.S. Chang, PPPL (Fusion Simulations)

Eli Dart, ESnet (Networking)

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

Steven Gitomer, NSF (Plasma Physics Program Director)

Martin Greenwald, MIT PSFC (Alcator C-Mod)

Paul Henderson, PPPL (PPPL Networking)

Steve Jardin, PPPL (Fusion Simulations)

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

Scott Klasky, ORNL (Simulations)

Alex Ryskin, University of Rochester (High Energy Density Laboratory Plasmas)

David Schissel, General Atomics (DIII-D)

Brian Tierney, ESnet (Networking)

Jason Zurawski, Internet2 (Networking)

Editors

Eli Dart, ESnet — dart@es.net

Brian Tierney, ESnet — bltierney@es.net

Table of Contents

1 Executive Summary

The Energy Sciences Network (ESnet) is the primary provider of network connectivity for the U.S. Department of Energy Office of Science, 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 December 2011, ESnet and the Office of Fusion Energy Sciences (FES), of the DOE Office of Science (SC), organized a workshop to characterize the networking requirements of the programs funded by FES.

The requirements identified at the workshop are summarized in the Findings section, and are described in more detail in the body of the report.

2 Findings

2.1 General Findings

Experiments at fusion facilities are collaborative enterprises. A tokamak control room is large, and an experiment involves large groups of people (the range of 20-45 is typical) working together in a fast-paced, highly interactive manner. Because significant members of the experiment (sometimes the experiment leader) are not on site, collaboration technologies are critical.

Rapid analysis of data, interactivity of analysis, and deterministic data transfer behavior are very important to the experiments — one pulse or “shot” every 15 minutes means that the analysis of the data, the interpretation of the results by the experiment team, and the reconfiguration of the facility based on the results must all occur within 15 minutes. This places significant performance, reliability, and consistency demands on the networks and data transfer systems used.

Operations of the Experimental Advanced Superconducting Tokamak (EAST) (China) and Korea Superconducting Tokamak Advanced Research (KSTAR) (Korea) will become more demanding from a networking perspective in the coming years — data rates will increase, collaboration tools will become ever more important, and performance will be critical. Strategic engagement is needed with the Asian networks and experimental facilities to ensure successful collaborations. Virtual circuit technologies such as those currently deployed in the Energy Sciences Network (ESnet) and in some Korean science networks could be very helpful. It is recommended that a demonstration project be considered. In addition to the Asian facilities, many international fusion facilities have ongoing collaborations with U.S. institutions. International network connectivity is strategically important for these collaborations.

Middleware tools and services, such as federated security and network monitoring, are needed.

ITER will require significant networking and systems resources to reach its full potential. A set of data and service challenges, similar to those used in preparation for the Large Hadron Collider (LHC) experiments, should be considered.

Data transfer performance must be improved. This will require the cooperation and collaboration of networking, systems, and security personnel at multiple institutions.

The fusion facilities make use of the Argonne Leadership Computing Facility (ALCF), the National Energy Research Scientific Computing Center (NERSC), and the Oak Ridge Leadership Computing Facility (OLCF), and need good data transfer performance to/from these supercomputer centers. In addition, there is a need to transfer data between the supercomputer centers.

2.2 Collaboration Tools

The workshop featured an extensive discussion of collaboration tools. Better collaboration tools and better integration of those tools are needed. The case studies, in particular Alcator C-Mod but also DIII-D, provide significant detail about requirements both in the near term and for a research agenda.

The experiments asked for an expanded menu of technical support options for collaboration technologies.

3 Action Items

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

- ESnet will continue to work with the U.S. fusion facilities to increase data transfer performance from Asian facilities in support of international collaborations.
- ESnet will explore the feasibility of a virtual circuit demonstration involving KSTAR and a U.S. fusion facility.
- ESnet will explore additional enhancements and media integration to its videoconferencing service, ESnet Collaboration Service (ECS).
- ESnet will continue to develop and update the fasterdata.es.net site as a resource for the community.
- ESnet will continue to assist sites with Performance Service Oriented Network monitoring ARchitecture (perfSONAR) deployments and will continue to assist sites with network and system performance tuning.

In addition, ESnet will continue to develop and deploy the ESnet On-demand Secure Circuits and Advance Reservation System (OSCARS) to support virtual circuit services.

4 Workshop Background and Structure

The strategic approach of the Office of Advanced Scientific Computing Research (ASCR — ESnet is funded by the ASCR Facilities Division) and ESnet to define and accomplish ESnet's mission involves three areas:

1. Working with the Office of Science (SC) community to identify the networking implication of the instruments, 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. Anticipating future network capabilities to meet future 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) Exploring the plans and processes of major stakeholders, including data characteristics of scientific instruments and facilities; anticipating what data will be generated by instruments and supercomputers coming online over the next 5-10 years; and examining the future process of science: how and where will the new data be analyzed and used, and how the process of doing science will change over the next 5-10 years

b) Observing current and historical network traffic patterns and determining how trends in network patterns predict future network needs

The primary mechanism for accomplishing (a) is the SC Network Requirements Workshops, sponsored by ASCR and organized by the SC Program Offices. SC conducts two requirements workshops per year, in a cycle that repeats every three years:

- Basic Energy Sciences (2007, 2010)
- Biological and Environmental Research (2007, 2010)
- Nuclear Physics (2008, 2011)
- Fusion Energy Sciences (2008, 2011)
- Advanced Scientific Computing Research (2009)
- High Energy Physics (2009)

The workshop reports are published at http://www.es.net/requirements/.

The requirements workshops also ensure that ESnet and ASCR have a common understanding of the issues that face ESnet and the solutions that ESnet undertakes.

In December 2011, ESnet and the DOE SC Office of Fusion Energy Sciences (FES) held a workshop to characterize the networking requirements of FES-funded programs.

Workshop participants codified their requirements in a case-study format that included:

- A network-centric narrative describing the science

- The instruments and facilities currently used, or anticipated for future programs
- The network services needed
- The ways in which the network is used.

Participants considered three timescales in their case studies: the near term (immediately and up to 12 months in the future), the medium term (two to five years in the future), and the long term (more than five years in the future). 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.

5 Office of Fusion Energy Sciences

5.1 Introduction

The mission of the Fusion Energy Sciences (FES) program is to support fundamental research to develop the knowledge base for a new and attractive form of energy based on the nuclear fusion process, the same process that gives rise to the energy of the sun and stars. In addition, FES supports research focusing on the underlying sciences of plasma physics, the study of the fourth state of matter that is the central component of magnetically confined fusion systems; as well as the emerging field of high energy density physics (HEDP), the state of matter encountered in inertially confined fusion energy systems. Related work contributes to understanding astrophysics, geosciences, industrial low-temperature plasma processing, turbulence, and complex self-organizing systems.

To carry out its mission, FES supports research activities involving more than 1,100 researchers and students at approximately 67 universities, 10 industrial firms, 11 national laboratories, and 2 federal laboratories, distributed over 31 states. These activities include efforts in experiment, theory, and advanced computation, ranging from single-investigator research programs to large-scale national and international collaborative efforts.

5.2 Major Facilities

At the largest scale, the FES program supports world-class magnetic confinement facilities shared by national teams of researchers to advance fusion energy sciences at the frontiers of near-energy producing plasma conditions. Each of the major facilities offers world-leading capabilities for the study of fusion-grade plasmas and their interactions with the surrounding systems. The Department’s three major fusion physics facilities are: the DIII-D tokamak at General Atomics (GA) in San Diego, California; the Alcator C-Mod tokamak at the Massachusetts Institute of Technology (MIT) in Cambridge, Massachusetts; and the National Spherical Torus Experiment (NSTX) at the Princeton Plasma Physics Laboratory (PPPL) in Princeton, New Jersey. These three major facilities are operated by the hosting institutions but are configured with national research teams of local scientists and engineers, researchers from other institutions and universities, and foreign collaborators. In addition to the FES major facilities, a range of small innovative experiments at universities and national laboratories explore the potential of alternative confinement concepts.

5.3 ITER

U.S. participation in ITER is a Presidential Initiative to build and operate the first fusion science facility capable of producing a sustained burning plasma. ITER’s mission is to demonstrate the scientific and technological feasibility of fusion energy for peaceful purposes. ITER is designed to produce 500 MW of fusion power at a power gain Q >10 for at least 400 seconds, and is expected to optimize physics and integrate many key technologies needed for future fusion power plants. The seven ITER parties (China, European Union, India, Japan, Russia, South Korea, and the United States) represent over half of the world's population. The European Union is hosting the site for the international ITER Project at Cadarache, France. Through ITER, the FES program is pushing the boundaries in large-scale international scientific collaboration.

5.4 International Collaborations

In addition to their work on domestic experiments, scientists from the United States participate in leading-edge experiments at international fusion facilities in Europe, Japan, China, South Korea, the Russian Federation, and India — the ITER members — and conduct comparative studies to enhance our understanding of the underlying physics. These facilities include the world's largest tokamak, the Joint European Torus (JET) in the United Kingdom; a stellarator (the Large Helical Device [LHD] in Japan); a superconducting tokamak (Tore Supra in France); and several smaller devices. The United States is also collaborating with South Korea on KSTAR and with China on research using the new long-pulse, superconducting, advanced tokamak EAST. These collaborations provide both a valuable link with the 80% of fusion research that is conducted outside the United States and a firm foundation to support ITER activities.

5.5 Advanced Computations

High-performance computing has played an important role in fusion research since the early days of the fusion program. NERSC — SC's production scientific computing facility — started as the Magnetic Fusion Energy Computer Center (MFECC). Currently, most FES advanced computational projects are supported under the auspices of SC's Scientific Discovery through Advanced Computing (SciDAC) program. The goal of FES SciDAC projects is to advance scientific discovery in fusion plasma physics by exploiting the emerging capabilities of terascale and petascale computing and associated progress in software and algorithm development, and to contribute to the FES long-term goal of developing a predictive capability for burning plasmas. The current FES SciDAC portfolio includes eight projects, set up as strong collaborations among 29 institutions, including national laboratories, universities, and private industry. Of these, five focus on topical science areas while the remaining three — which are jointly funded by FES and ASCR and are known as Fusion Simulation Prototype Centers — focus on code integration and computational framework development in the areas of edge plasma transport, interaction of RF waves with MHD, and the coupling of the edge and core regions of tokamak plasmas. The FES SciDAC projects' success, combined with the emerging availability of even more powerful computers and a need to develop an integrated predictive simulation capability for the needs of ITER and burning plasmas, have led FES to propose a new computational initiative, the Fusion Simulation Project (FSP). The FSP, to be initiated in the future, will develop experimentally validated computational models that can predict the behavior of magnetically confined plasmas in the regimes and geometries relevant for practical fusion energy, by integrating experimental, theoretical, and computational research across the FES program and taking advantage of emerging petascale computing resources.

6 Alcator C-Mod Tokamak at the MIT Plasma Science & Fusion Center

6.1 Background

The Plasma Science & Fusion Center (PSFC) is a large, interdisciplinary research center on the MIT campus in Cambridge. 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 includes a number of smaller research facilities, such as the Versatile Toroidal Facility (VTF), an experiment studying collisionless magnetic reconnection. Research on these devices has relevance outside the fusion program, particularly to space and astrophysical plasma physics. The Plasma Science Division at the PSFC carries out a broad program of theory and computational plasma physics. The computational work emphasizes wave-plasma interactions and turbulent transport.

6.2 Key Local Science Drivers

6.2.1 Instruments and Facilities

The largest activity at the center is the Alcator C-Mod tokamak. 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 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, with about 30 Ph.D. students carrying out thesis research at any given time.

The PSFC has ~1,500 network-attached devices, more than half associated with the C-Mod team and experiment. The infrastructure is switched 1 Gbps Ethernet, with Gb connectivity to all workstations and desktops. Deployment of a 10 Gbps backbone has begun. The C-Mod experiment is directly supported by ~10 multicore Linux servers for data acquisition, storage, and analysis and approximately 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 ~10 GB of data. All experimental data is maintained on disk, with approximately 40 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 Tivoli Storage Manager (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 the National Center for Computational Sciences (NCCS) as well as a local computer cluster with 600 cores (with various upgrades planned). PSFC researchers are actively involved in several SciDAC collaborations and in the FSP planning activity.

6.2.2 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 decision making between shots. Thus, a high degree of interactivity with the data archives and among members of the research team is required.

6.3 Key Remote Science Drivers

6.3.1 Instruments and Facilities

Activity focuses on a set of national and international experiments (the latter described in more detail in the international collaboration case study). 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 Oak Ridge National Laboratory (ORNL).

The PSFC WAN connection is through a local Gbps link to an off-campus interconnect in downtown Boston at which MITnet, ESnet, and Internet2 peer. The local fiber infrastructure allows this link speed to increase with only moderate effort and expense, if traffic warrants. The WAN link is shared with other DOE/SC-funded researchers at MIT, particularly the Laboratory for Nuclear Science (LNS).

6.3.2 Process of Science

A 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. (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 Lawrence Livermore National Laboratory [LLNL] ran the C-Mod experiment from a control room in California.) Fast and efficient data access is clearly a requirement for this mode of operation.

Multi-institutional collaborations are critical to 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 nationwide computational projects and rely on use of remote supercomputers. MIT also supports the MDSplus data system, which is installed at about 30 fusion facilities worldwide.

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

The PSFC makes significant use of ESnet-provided collaboration services. The H.323 videoconference facilities are used for both scheduled and ad hoc collaboration. Data/screen sharing is regularly used to broadcast visuals from presentations. Over the next 5 to 10 years, we would like to see expansion of these services in both technical and support areas (see below for details). Room- and person-based paradigms should be provided for, recognizing that the needs for these classes of users differ significantly.

MIT is 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 integrated new tools into the normal workflow. One aim is to improve ad hoc interpersonal communication, which, we believe, currently 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 offer a method for seamless integration of needed services.

6.4 Local Science Drivers – the Next 2-5 Years

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

6.4.1 Instruments and Facilities

To provide for the ongoing expansion in data rates, we have begun planning upgrades to the local area network. To improve performance, it will be segmented into several routing domains with a 10 Gbps backbone. We will provide 10 Gbps links to servers and switches. Workstation and desktop connectivity will remain at 1 Gbps 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.

6.4.2 Process of Science

We do 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 portion of the research team.

6.5 Remote Science Drivers – the Next 2-5 Years

6.5.1 Instruments and Facilities

A set of new international facilities (EAST, KSTAR, Steady State Superconducting Tokamak [SST-1]) has recently come into operation and they are entering a physics exploitation phase. Others will begin operation during this period (W7-AX in 2014, JT-60SA in 2016).

6.5.2 Process of Science

Collaboration with off-site researchers will continue to grow over the next five years. For example, the NSTX experiment will be offline for the next three years for a major upgrade, increasing PPPL collaborations at the other two FES facilities. Planning will also begin for the next large U.S. experiment, expected to be a national facility run by a broad consortium. Activities supporting ITER construction will be based around the U.S. ITER Program Office and will probably not drive much additional traffic to MIT in this time period.

6.6 Beyond 5 Years – Future Needs and Scientific Direction

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

Future collaboration tools need to include:

1. Global directory services
2. Centrally administered conferences (call out)
3. VoIP/S collaboration tools
4. Screen-sharing presentation tools
5. Recording and playback
6. Instant messaging — collaboration and meeting set-up
7. Higher-quality multipoint video
8. Presence / availability information
9. Better integration of the above elements
10. Integration with authorization tools

Support needs:

1. Quickly determine the operational state of collaboration services.
2. Communicate this state to meeting participants (e.g., via Instant Messaging or Web).

The 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 a meeting is taking place or is about to start with remote participants. In the case of technical difficulties, decisions to cancel a meeting or remote session must be made promptly.

6.7 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” certificate authority. The fusion collaboration also created a distributed authorization system – the Resource Oriented Authorization Manager (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 above, we make extensive use of the ESnet collaboration services.

There is a real need for better integration of remote participation tools. The integration mentioned above should encompass all modes of machine-mitigated interpersonal communication, enabling users to initiate conversations on the most appropriate media, then move from tool to tool (voice, video, IM, e-mail, screen sharing, data sharing, etc.) seamlessly and as needed. Tools should be standards-based, modular, role- and presence-aware, and Web friendly in a multiplatform environment.

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

6.8 Outstanding Issues

Issues related to existing collaboration services:

H.323

- Reliability improving, but still occasional glitches
- Still problems with global dialing. Users typically use international MCUs. ESnet MCUs aren't functional.
- Need functional and symmetric GDS

Support model

- Slight improvement with new vendor
- Still no real-time support
- Propose real-time "pay per call" service

ReadyTalk

- Generally works quite well. On-demand voice and screen conferencing is widely used and effective.
- Meetings with shared control can be fragile. It's easy for users to break off meetings by mistake.
- Need more functional integrated chat (IM). Any user should be able to talk to any other user or users. Right now only available to moderator.
- All users should be able to see a list of participants. Right now, only available to moderator.
- What is path for user requests to get to vendor?

Integration

- Integration of ReadyTalk and H.323
- Directory and presence support

- Some users are put off by the number of steps required to set up a remote conference

6.9 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	**Process of Science**	**Data Set Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
Near Term (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	• <3 minutes	• Bursty — small portion of data set transferred with no noticeable delay (e.g., 20 MB in 1 second) • Endpoint — see note 1
2-5 years				
• Prep work for ITER • Additional international collaborations	• Increased use of collaboration technologies including SIP/VoIP • Involvement in development of next major U.S. experiment	• Data volume 1 TB/day) • Data set composition 3,000 files — largest about 2 GB	• <3 minutes	• Bursty — small portion of data set transferred with no noticeable delay (e.g., 40 MB in 1 second) • Endpoint — see note 1
5+ years				
• Additional international collaborations • Possible new facilities for materials PWI, or FNS • Research on ITER	• 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	• <3 minutes	• Bursty — small portion of data set transferred with no noticeable delay (e.g., 80 MB in 1 second) • Endpoint — see note 1

1. Domestic endpoints for WAN traffic include GA (San Diego), PPPL, UT Austin, LLNL, LANL, ORNL, and NERSC. International endpoints include ASIPP (Hefei, China), IPP (Max-Planck — Garching, Max-Planck — Greifswald, Germany), Tore-Supra & ITER (Cadarache, France), JET-EFDA (Culham, U.K.), CRPP (Lausanne, Switzerland), NIFS (Toki, Japan), NFRI (Daejeon, South Korea), IPR (Gujarat, India)

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

7.1 Background

The DIII-D National Fusion Facility at General Atomics’ (GA’s) site in La Jolla, California, is the largest magnetic fusion research device in the United States. The research program on DIII-D is planned and conducted by a national (and international) research team. There are more than 500 users of the DIII-D facility from 92 worldwide institutions including 41 universities, 36 national laboratories, and 15 commercial companies. The mission of the 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 measurement system are unsurpassed. 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.

GA also conducts research in theory and simulation of fusion plasmas in support of the Office of Fusion Energy Sciences’ (FES’) overarching goals of advancing fundamental understanding of plasmas, resolving outstanding scientific issues and establishing reduced-cost paths to more attractive fusion energy systems, and advancing understanding and innovation in high-performance plasmas including burning plasmas. The theory group works in close partnership with the DIII-D experiment 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. Members of the theory group are also active in several SciDAC Fusion Simulation Project (FSP) prototype centers and fusion SciDAC projects. Numerical simulations are conducted on multiple local Linux clusters (multiple configurations and sizes) as well as on computers at NERSC and NCCS.

7.2 Key Local Science Drivers

7.2.1 Instruments and Facilities

The GA connection to ESnet is at 1 Gbps, with major computing and storage devices connected by a switched 1 Gbps Ethernet LAN. The ESnet-to-GA connection will be upgraded to 10 Gbps. Network connectivity between the major computer building and the DIII-D facility is by dual 1 Gbps circuits. The major data repositories for DIII-D comprise approximately 110 TB of online storage with metadata catalogs stored in a relational database. Network connectivity to offices and conference rooms is at 100 Mbps on a switched Ethernet LAN. Approximately 2,000 devices are 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 multidimensional measurements are acquired and analyzed representing several Gigabytes of data.

The experimental data is accessed both locally and over the WAN. Access to the experimental data, data analysis tools, and audio/video-based collaboration tools creates significant network traffic during the experiment.

7.2.2 Process of Science

Throughout the experimental session, hardware/software plasma control adjustments are debated and discussed among the experimental team members 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.

7.3 Key Remote Science Drivers

7.3.1 Instruments and Facilities

The highly distributed nature of the DIII-D National Team requires the use of substantial remote communication and collaboration technology. Five conference rooms are equipped with Polycom H.323 videoconferencing hardware. The DIII-D control room has the ability to use Access Grid, Virtual Room Videoconferencing System/Enabling Virtual Organizations (VRVS/EVO), and dedicated H.323 hardware for remote video-conferencing. The remote control room at DIII-D contains two high-definition cameras and supports high-definition H.323. Additionally, scientists use various technologies to communicate with audio/video to the desktop. The DIII-D morning operations meeting is automatically recorded and published with podcasting capability.

7.3.2 Process of Science

The pulsed nature of the DIII-D experiment, combined with its highly distributed scientific team, results in cyclical WAN traffic. The additional constant demand of collaborative services is mostly associated with several types of videoconferencing, the majority H.323 based. With the increase of collaborative activities associated with DIII-D comes an increased use of collaborative visualization tools by offsite researchers, which requires efficient automatic data transfer between remote institutions.

The scientific staff members associated with DIII-D are very mobile in their working patterns. They travel to meetings and workshops, work actively on other fusion experiments around the world, and work from home. For those offsite individuals 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.

7.4 Local Science Drivers – the Next 2-5 Years

7.4.1 Instruments and Facilities

Although the operation of DIII-D will remain similar for the next five years, the rate of acquiring new data is expected to continue to increase. From 2008 to 2011, the total amount of data taken at DIII-D increased threefold. To keep up with this demand and the increased use of collaborative technologies, even within the local campus, discussion has begun about increasing the major LAN backbone to 10 Gbps and the connections to selected desktops to 1 Gbps.

7.4.2 Process of Science

While the operation of DIII-D is expected to hold steady for the next five years, scientists will increasingly focus on remote collaborations between DIII-D and other institutions.

7.5 Remote Science Drivers – the Next 2-5 Years

7.5.1 Instruments and Facilities

For DIII-D, the experimental team’s need for real-time interactions with one another and the requirement for interactive visualization and processing of very large simulation data sets will be particularly challenging. Some important components to help make this possible include user authentication and authorization frameworks that are easy to use and manage, global directory and naming services, distributed computing services for queuing and monitoring, parallel data transfer between remote institutions, and network quality of service (QoS) to provide guaranteed bandwidth at particular times or with particular characteristics.

7.5.2 Process of Science

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 five years, operating these tokamaks will become routine and the remote participation of DIII-D scientists is expected to 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 China and South Korea will become increasingly important.

7.6 Beyond 5 Years – Future Needs and Scientific Direction

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; ITER, located in France, will be close to coming online, 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 tokamaks simultaneously, placing a further strain on the WAN and creating a need for efficient peering to our Asian and European partners.

7.7 Middleware Tools and Services

A variety of database technologies are used at DIII-D, including MDSplus and PTDATA for storage of multidimensional data sets, and MySQL and Microsoft SQL Server for metadata.

GridFTP and FDT are used for automated parallel data transfer between remote institutions and DIII-D to facilitate remote collaborations.

Multiple videoconferencing technologies are used by the research staff, including H.323 and Skype.

FusionGrid Uses the DOEGrids Certificate service.

7.8 Outstanding Issues (if any)

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 network latency is also important. Currently, the amount of data needed to transfer between DIII-D and its partner sites is 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, a 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 paths and better peering helped with increased network throughput and decreased latency (e.g., ESnet peering with Global Ring Network for Advanced Applications Development [GLORIAD] on November 2010 decreased the latency about 25% between DIII-D and EAST).

Real-time support for collaboration services is needed. Almost all the collaboration tools (videoconferencing, screen sharing, etc.) rely on the network. Therefore, standardizing the tools and increasing ESnet support is the most effective option. Currently, ESnet collaboration services support a typical U.S. work schedule. However, collaboration in fusion is international and therefore around the clock; users need support 24 hours a day. Numerous examples exist where meetings have failed due to technical difficulties that could not be resolved because of inadequate real-time support.

Additionally, ESnet should consider standardizing collaboration tools and videoconferencing equipment proactively. Thus, hardware upgrades and software updates can be done in a coordinated manner. Previously unannounced updates have unintentionally left H.323 equipment not fully functioning, resulting in lost productivity.

7.9 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	**Process of Science**	**Data Set Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
Near Term (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 (2 TB/day) • Data set composition: TCP/IP-based client server data, audio/ video	• Consistent streaming 24x7	• Consistent data streaming 24x7 (1-2 minute 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 (5 TB/day) • TCP/IP-based client server data, audio/ video	• Consistent streaming 24x7	• Consistent data streaming 24x7 (1-2 minute 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 (20 TB day) • TCP/IP-based client server data, audio/ video • Possible file-based simulation	• Consistent streaming 24x7	• Consistent data streaming 24x7 (1-2 minute delay is tolerable)

8 Laboratory for Laser Energetics (LLE), Rochester, NY

8.1 Background

LLE is a unique national resource for research and education in science and technology located at the University of Rochester. Established in 1970 as a center for the investigation of the interaction of intense radiation with matter, LLE has a five-part mission: (1) to conduct laser-fusion implosion experiments in support of the National Inertial Confinement Fusion (ICF) program; (2) to develop new laser and materials technologies; (3) to provide education in electro-optics, high-power lasers, high-energy-density physics, plasma physics, and nuclear fusion technology; (4) to conduct research and development in advanced technology related to high-energy-density physics; and (5) to operate the National Laser Users' Facility (NLUF) laser fusion energy research. Experiments produce up to 20 GB of data a day stored on a local SAN. Local computer simulations produce up to 40 GB.

8.2 Key Local Science Drivers

8.2.1 Instruments and Facilities

The OMEGA Laser System and the OMEGA EP (Extended Performance) Laser System are the pre-eminent facilities within the inertial fusion and high-energy-density physics communities that support an important national mission. Extensive use of OMEGA is essential to the national program to achieve ignition, to provide laser facility time for national laboratory experiments, and to operate the NLUF. OMEGA is the staging and support facility for the National Ignition Facility (NIF). In support of the OMEGA Laser System and the general computing requirements of the staff scientists and engineers, the laboratory employs a variety of local networks (10 Gbps, 1 Gbps, 100 Mbps) comprising more than 1,000 PCs and more than 2,000 computing cores and 17 GPUs in 3 HPC clusters. We also use a 1,024-core cluster at LLNL remotely.

8.2.2 Process of Science

The laser installation produces bursts of data approximately every hour. The data are initially collected in various instruments and attached computers and subsequently transmitted to a central server and registered in a database. From the server, the data is available for processing via various interfaces. Extensive computer simulations are used for experiment planning and throughout all phases of the scientific process.

The two-dimensional, radiation-hydrodynamics code DRACO is the main computational workhorse in the laboratory. Consisting of roughly 2 million lines of code, DRACO is employed to provide design and predictive capability for all experimental target campaigns carried out on OMEGA, OMEGA EP, and LLE contributions to NIF. DRACO has over 10 users, each generating 1-2 GB/day. DRACO keeps our HPC cluster utilization at 99% over the year.

LLE is currently adding a network of digital GigE video cameras to its Laser Facility image acquisition system. The cameras are connected to processing servers with a 10 Gbps network.

8.3 Local Science Drivers – the Next 2-5 Years

8.3.1 Instruments and Facilities

We intend to continue augmenting the general computing resources available to staff scientists and engineers. This entails network additions/improvements, as well as doubling our HPC computing cores and, pending GPU code development, a tenfold increase in GPU deployment.

The digital camera network will grow to 100-150 units and gradually replace existing analog cameras.

8.4 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	**Process of Science**	**Daily Data Set Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
Near Term (0-2 years)				
• DRACO on LLE and LLNL clusters	• Large 2D Rad-hydro code, both local and remote sites	• 50 files, 100 GB	• 1 Gbps per user • 10 Gbps backbone	
• HYDRA on LLNL Clusters	• Large 3D rad-hydro code on remote sites	• 3,000 files, 6 GB		• 160 Mbps
• VISIT	• Graphical post-processing	• 100,000 pixels • 30 frames/s		
• Laser Facility digital camera network	• Image acquisition and processing	• 50 cameras, 20 viewers	• 1 Gbps per camera • 10 Gbps per server	
2-5 years				
• DRACO on LLE and LLNL clusters	• Large 2D Rad-hydro code both Local and remote sites	• 100 files 200 GB	• 1-10 Gbps user • 100 Gbps backbone	
• HYDRA on LLNL Clusters	• Large 3D rad-hydro on remote sites	• 12,000 files, 24 GB		• 200 Mbps
• VISIT	• Graphical post processing	• 1 M pixels • 30 frames/s		
• Laser Facility digital camera network	• Image acquisition and processing	• 30 viewers, 150 cameras	• 1 Gbps per camera, 10 Gbps per server	

9 Major International Collaborations

9.1 Background

International collaboration has been a key feature of magnetic fusion energy research since declassification in 1958. Over the past 30 years, formal multilateral and bilateral agreements have created, in effect, a single, loosely coordinated research enterprise. The fusion community, which traditionally included the United States, Western Europe (including Australia), Russia (previously the USSR), and Japan, has expanded in recent years to include Eastern Europe, Korea, China, and India. Planning and program advisory committees typically have cross membership, particularly among the most active nations (United States, European Union, Japan). Preparation for ITER has further strengthened cooperative research, especially through the International Tokamak Physics Activity (ITPA). Driven by improvements in 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 execution of experiments now routine. However, despite technological advances, challenges remain; the increase of multinational research teams has created ever-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.

9.2 Key Local Science Drivers

9.2.1 Instruments and Facilities

The United States runs three major experimental fusion facilities: the Alcator C-Mod device at MIT, DIII-D at GA, and NSTX at PPPL. All three have large, extended research teams and run, essentially, as national facilities. In addition to their collaborators, the facilities carry out coordinated joint research under specific DOE-SC targets and as part of the ITPA.

9.2.2 Process of Science

See Section 9.3.2

9.3 Key Remote Science Drivers

9.3.1 Instruments and Facilities

AUG: The Axially Symmetric Divertor Experiment (ASDEX) Upgrade (AUG) is a midsize divertor tokamak located at the Max Planck Institute for Plasma Physics (IPP) in Garching, Germany. The machine’s primary mission has been to support ITER design and operation, focusing 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 (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 resonance frequency (ECRF) heating and current drive and steady-state scenario development); and NSTX (diagnostics development and turbulence studies). Important collaborations on theory and modeling are also in place with many U.S. groups.

JET: The Joint European Torus that is under the EFDA is 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 Alven Eigenmode [TAE] physics, and disruption mitigation); DIII-D (H-mode pedestal physics, especially edge-localized mode [ELM] suppression, neoclassical tearing modes, resistive wall modes and rotation, steady-state scenario development); NSTX (Alfven eigenmodes physics, neoclassical tearing modes, and resistive wall mode research).

ITER: ITER is a collaboration among seven parties (Europe, Japan, United States, China, South Korea, Russia, and India) to build the world's first reactor-scale fusion device in Cadarache, France. The project expects to finish major construction in 2018 and to operate for 20 years. The current date for first plasma is November 2019. The project is scheduled to begin deuterium-tritium operation in March 2027. Collaboration during the construction phase is discussed in another chapter of this report; the research phase is discussed below under *Beyond 5 Years*.

KSTAR: Korea Superconducting Tokamak Advanced Research is an all-superconducting tokamak experiment located at Daejeon, Korea. It will operate with hydrogen and deuterium. KSTAR's size, operation capabilities, and mission objectives for the initial operating period will be comparable to the present DIII-D tokamak. The main research objective of KSTAR is to demonstrate steady-state high-performance advanced tokamak scenarios. Collaborators include PPPL (plasma control system [PCS], diagnostics, ICRF), ORNL (fueling), DIII-D (PCS, data analysis, electron cyclotron heating [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 past several experimental campaigns.

EAST: The Experimental Advanced Superconducting Tokamak, 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 DIII-D (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 EAST's successful first plasma in September 2006. Since then, collaborations have continued in every EAST experimental campaign. During the 2010 campaign, a dedicated MDSplus data server was deployed at GA to serve EAST data to U.S. scientists. An automated data-replication mechanism was created to automatically move experiment data after each plasma pulse. Fast data-copying tools, such as GridFTP and FDT, were used for fast data transfer between EAST and the United States.

SST-1: The Steady State Superconducting Tokamak is located at the Institute for Plasma Research (IPR), in Gujarat, 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 is anticipated in 2009. The main objective of SST-1 is to study the steady-state operation of advanced physics plasmas. At this time, facilities collaborations are with DIII-D in the areas of physics, plasma operation, theory, and electron cyclotron emission (ECE) diagnostics. This collaboration is expected to grow to encompass other groups within the United States.

LHD: The Large Helical Device 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 for Fusion Science in Toki, Japan. There are active U.S. collaborations on this device.

A number of additional facilities are targets of somewhat less intense collaboration, including Tore Supra in Cadarache, France; Tokamak à Configuration Variable (TCV) at the Plasma Physics Research Center (CRPP) in Lausanne, Switzerland; and the Mega Ampere Spherical Tokamak (MAST) at the Culham Science Centre in the United Kingdom.

9.3.2 Process of Science

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

Experimental planning typically involves data access, visualization, data analysis, and interactive discussions among members of the distributed scientific team. Skype and H.323 videoconferencing have been used for such discussions. For some foreign collaborations (e.g., EAST), conversing via traditional phone lines is not an option due to prohibitive expense. The technology used often depends on the technical capability and experience of scientists at each end. The recent trend is to use H.323 for more formal, larger meetings; these are facilitated by multipoint control units (MCUs) to connect numerous participants. Data analysis and visualization is typically done in one of two ways: Either the scientist logs on to a remote machine and uses the foreign laboratory's existing tools or the scientist uses his or her own machine and tools to remotely retrieve the data. The widespread use of MDSplus makes the latter technique easier and more time efficient, yet this is not possible at all locations.

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

Despite improvements in intercontinental links and the development of national networks, collaborators still report problems with link speed to sites in China, Korea, and Japan. This information, which is anecdotal rather than systematic, is usually brought to our attention when U.S. scientists travel abroad. This suggests that expectations by researchers at some foreign laboratories are still relatively low. It is not clear if the problem is with the connection from laboratory to national backbone or with the LAN at these laboratories.

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

1. **Federated security.** Technical and policy advancements to allow sharing of authentication credentials and authorization rights would ease burdens on individual collaborating scientists. This is crucial for more complex interactions, such as when a researcher at one site accesses data from a second site and runs a computational analysis on the data at a third site. (The 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 must be hidden from end users.
3. **Document and application sharing.** Improved tools for sharing displays, documents, and applications are urgently needed. Cognizance of different technology standards and policies will be important.
4. **Network monitoring.** The network backbone should be monitored, as well as end-to-end connections. Tools for testing and visualizing the state and performance of the network should be readily accessible.

9.4 Local Science Drivers — the Next 2-5 Years

9.4.1 Instruments and Facilities

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

9.4.2 Process of Science

See Section 9.3.2

9.5 Remote Science Drivers – the Next 2-5 Years

9.5.1 Instruments and Facilities

EAST. Over the next several years, 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 the machine is operated. The EAST tokamak's superconducting nature allows for 24-hour-a-day operation for weeks at a time. The United States and China have discussed having U.S. scientists becoming actively involved in EAST's third-shift operation (daytime in the United States). If this is pursued, there is the possibility of a greater increase in the breadth and scope of this collaboration and an increase in network traffic.

KSTAR. Physics research on KSTAR is expected to begin in the middle of 2008. In a similar fashion to EAST, as KSTAR continues to operate over the next five years, more data will be available to remote participants, with greater opportunity to participate in experiments. In contrast to EAST, discussion of third-shift operation of KSTAR has not taken place. Therefore, for the time being, the network requirements from the United States to EAST will exceed those of the United States to KSTAR.

W7X. Located at the Max Planck Institute for Plasma Physics in Greifswald, Germany, W7X is a large (R = 5.5 m, a = 0.53 m, B = 3 T) superconducting modular stellarator device scheduled to begin operating in 2014. W7X will test the principle of "quasi-omnigeneity" for 3-D-shaped plasmas. The United States has an active stellarator program centered at PPPL and ORNL. The recent cancellation of a major U.S. stellarator will likely increase the importance of collaborations on this device.

International Fusion Energy Research Centre (IFERC). As part of ITER's Broader Approach, the IFERC is being built in Rokkasho, Japan. The center's purpose is to complement the ITER project through R&D in nuclear fusion; it will perform 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 10 times more powerful than any existing system in Europe and Japan dedicated to simulations in fusion. 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. Full U.S. exploitation of this computer will require fast and reliable network connectivity between Rokkasho and U.S. fusion facilities.

9.5.2 Process of Science

See Section 9.3.2

9.6 Beyond 5 Years — Future Needs and Scientific Direction

ITER research phase. Though the ITER experiment is not scheduled to start up for roughly eight years, detailed planning has begun for the research program and for data and communications systems to support that program. Estimates on data volume are based on extrapolation from the current generation of experiments. A (we hope) more accurate bottom-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/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/day during operation, with data rates in the neighborhood of 0.3-50.0 GB/sec. 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 10 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 essential to 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.

Next-generation experiment. With the construction of ITER well under way, the United States has begun to look into new facilities to operate contemporaneously with ITER and fill knowledge gaps left by that project and the balance of the world program. Presently, the U.S community is investigating research areas in materials science and technology needed to fill gaps to create the basis for a Demonstration Power Plant (DEMO) facility.

JT-60SA. 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 EFDA and Japan Atomic Energy Agency (JAEA). While there is a rich history of U.S.-Japan collaboration, the extent of U.S. involvement in this experiment is not clear at the present time.

9.7 Middleware Tools and Services

- GridFTP and FDT have been used in an effort to 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/videoconferencing services are used to conduct remote meetings as well as participate in remote experiments. Today, H.323 is the most commonly used service.
- QoS for scheduled time period is not used presently but is desired for getting data in real time from remote experiments.
- Multicast is used for some video services, where allowed, to reduce the strain on the network.

9.8 Outstanding Issues

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

Lowering network latency is also important. Currently, the amount of data needed to transfer between international and domestic sites is 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 kind 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 GLORIAD on November 2010 decreased the latency about 25% between DIII-D and EAST).

Real-time support for collaboration services is required. Almost all the collaboration tools (videoconferencing, screen sharing, etc.) rely on the network. Therefore, standardizing the tools and increasing the support by ESnet is the most effective option. Currently, ESnet collaboration services support a typical U.S. work schedule. However, collaboration in fusion is international and therefore around the clock, and users must be supported 24 hours a day. In several instances, meetings have failed due to technical difficulties that could not be resolved because of inadequate real-time support.

Additionally, ESnet should consider proactively standardizing collaboration tools and videoconferencing equipment. Thus, hardware upgrades and software updates can be done in a coordinated manner. Previously unannounced updates have unintentionally left H.323 equipment not fully functioning, resulting in lost productivity.

9.9 Summary Table

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 and audio/video streaming is critical. When ITER comes online (5+ years), data traffic, both domestically and between Europe and the United States, will immediately increase.

Key Science Drivers			**Anticipated Network Needs**	
Science Instruments and Facilities	**Process of Science**	**Data Set Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
Near Term (0-2 years)				
• Multiple remote experiment facilities	• Real-time data access and analysis. Team-based collaboration with data sharing, screen sharing, and multi-user high-definition video conferences.	• Data volume (2 TB/day) • Data set composition: o TCP/IP-based client server data, audio/ video	• Consistent streaming 24x7	• Consistent data streaming 24x7 (1-2 minutes' delay is tolerable)
2-5 years				
• Multiple experiment facilities (new facilities will be added)	• Real-time data access and analysis. Team-based collaboration with data sharing, screen sharing, and multi-user high-definition video conferences.	• Data volume (5 TB/day) • TCP/IP-based client server data, audio/ video	• Consistent streaming 24x7	• Consistent data streaming 24x7 (1-2 minutes' delay is tolerable)
5+ years				
• Multiple experiment facilities (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 video conferences.	• Data set (20 TB/day) • TCP/IP-based client server data, audio/ video • Possible file-based simulation	• Consistent streaming 24x7	• Consistent data streaming 24x7 (1-2 minutes' delay is tolerable)

10 PPPL Computational Science Networking Requirements

10.1 Background

PPPL physicists develop and run 8-10 major massively parallel physics codes, mostly at NERSC (Lawrence Berkeley National Laboratory [LBNL]), the Oak Ridge National Laboratory (ORNL) Leadership Computing Facility, or at the Argonne Blue Gene/P supercomputer. PPPL scientists have collaborated on many recent SciDAC projects, including the Center for Simulation of Plasma Microturbulence (CSPM), the Center for Gyrokinetic Particle Simulations of Turbulent Transport in Burning Plasmas (GPS-TTBP), the Center for Simulation of Wave-Plasma Interactions (CSWPI), the Center for Extended Magnetohydrodynamic Modeling (CEMM), and the Center for Nonlinear Simulation of Energetic Particles in Burning Plasmas (CSEP). PPPL scientists also participated in the Proto Fusion Simulation Projects (Proto-FSPs) including the Center for Simulation of Wave Interaction with MHD (CSWIM), the Center for Plasma Edge Simulation (CPES), and the Framework Application for Core-Edge Transport Simulations (FACETS) and are participants in many of the SCIDAC3 proposals (awards yet to be announced). There are also typically several INCITE awards given to PPPL scientists each year.

The codes study different aspects of physical phenomena that occur in fusion confinement configurations, and are 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-C^1. (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-C^1.) We focus here on one code of each category: GTS, XGC1, and M3D-C^1.

In addition to the massively parallel physics codes 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 analysis package. TRANSP is used by tokamak physicists worldwide to interpret experimental data from experiments and to predict the operation of future experiments. In FY 2011 there were 6,604 TRANSP runs that accounted for about 0.5 M hours of CPU time. About half 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 8-16 processors.

10.2 Key Local Science Drivers

Most of the massively parallel jobs are run at NERSC, ORNL, or Argonne, and the data is stored and analyzed local to where the job is run via remote connections to scientists at their desks. Smaller jobs are typically run and analyzed at PPPL, as are all TRANSP jobs.

10.2.1 Instruments and Facilities

Approximately 3,000 processing cores and 450 TB of storage are available locally at PPPL. This provides local computing resources and storage for small simulations. While processing ~200,000 jobs per year, 40% are single CPU jobs, 55% utilize 2-32 CPUs, and the remaining 5% use between 32 and 512 CPUs.

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 off site at collaborative institutions, both within the United States and overseas. Off-site 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 10 Gbps 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 moved core services like e-mail to the Internet "cloud." PPPL also has a backup connection to ESnet's New York router, at 1Gbps, which is automatically utilized if the main 10 Gbps link to Washington, D.C., is interrupted. However, this backup link is shared by PPPL, the Geophysical Fluid Dynamics Laboratory (GFDL), Princeton High Energy Physics, and Ithaka Harbors, which saturates the link. *PPPL thus requests a dedicated backup circuit for PPPL's exclusive use to NYC/32 AOA if the main circuit is down.*

With its current ESnet bandwidth of 10 Gbps, PPPL is comfortably meeting the current needs of its research mission. PPPL needs to extend that high-speed capability further into PPPL so that the transfer of large data sets to local storage can be accomplished in a timely manner, which is not the case today. This will require a rework of the data transfer systems currently in use, as well as the tools used to accomplish those transfers (e.g., SCP, FTP, etc.)

10.2.2 Process of Science

A typical GTS simulation 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 are dominated by the particle data: 4×10^{10} particles x 8 bytes x 12 variables = 4 TB (Terabytes). If particle data from every time step is saved, it would require 40 PB (Petabytes) of storage. However, normally only the mesh data are saved for post-processing. If mesh data are saved every 10 time steps, this would require 4 x 10^8 (mesh size) x 8 (bytes) x 4 (variables) x 10^3 (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, a mesh of about 5 million node points, and 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 steps and its file size is (10^{11} particles) x (9 variables) x (8 byte) + field data = about 10 TB at every 1,000 time steps, or 100 TB total. If particle data from every time step is saved, it would require 100 PB 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 10^7 data points) x (5 variables) x (8 byte) = 2 GB. The total file size of the grid field data (coming from 10,000 time steps of simulation) is 2 TB. Wall clock time for one simulation is about one day on Jaguar and two days on Hopper. Particle data are 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 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 filesystem 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 LAN speed requirement is about 16 Gbps.

The M3D-C^1 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 10^5 high-order finite element nodes to represent a tokamak. A large job will run for 750 hours on 1,536 processors for a total of 1.1 M processor-hours. Each node requires 12 numbers to represent a single scalar variable, and there are typically eight scalar variables resulting in 30 x 10^6 words or 2 GB 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 200 GB. The restart files are written with ADIOS and the graphics files are written with parallel HDF5.

10.3 Key Remote Science Drivers

PPPL utilizes ESnet to access data at the supercomputer centers and other off-site locations and to provide access to the PPPL computing facilities for TRANSP users worldwide.

10.3.1 Instruments and Facilities

PPPL researchers are heavily involved with off-site fusion projects within the United States and overseas, particularly the C-Mod and D3D experiments in the United States, the JET experiment in England, KSTAR in Korea, and EAST in China. 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, as the newest reactors are those built overseas.

Access to the TRANSP analysis tools is provided to physicists worldwide via the FusionGrid.

The current 10Gbs connection enjoyed by PPPL is sufficient to support this research. PPPL has deployed the popular Starnet product, which optimizes XWindows traffic over long hauls. This product is equivalent to the NX or NoMachine protocol in use within the DOE research community.

10.3.2 Process of Science

Mostly, data analysis for the massively parallel codes is performed on the remote supercomputing site where the data are generated. However, for advanced analysis and visualization, the whole physics-study file needs 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 12 TB for GTS, 2 TB for XGC1, and 200 GB for M3D-C^1. For a data transfer in 10 hours, the corresponding network requirement would be about 2.5 Gbps for GTS.

In some situations, transferring the restart file (4 TB for GTS, 10 TB for XGC1, and 2 GB for M3D-C^1) between NCCS and NERSC is required. The largest of these, XGC1, assuming four hours of transfer time, requires 5 Gbps of network speed.

10.4 Key Local Science Drivers – the Next 2-5 Years

10.4.1 Instruments and Facilities

PPPL’s local computing and storage resources will continue to grow to meet the need for small-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, and will reach approximately 750 TB in two years, and well over 1 PB in five years.

10.4.2 Process of Science

In five years, the number of particles and mesh points used by each of the GTS, XGC1, and M3D-C^1 codes will increase by an order of magnitude. Also, new physics and new variables will be added. 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.

10.5 Key Remote Science Drivers – the Next 2-5 Years

10.5.1 Instruments and Facilities

While the growth of network traffic has not historically increased at the same rate as storage, within two to five years PPPL’s current 10 Gbps network link may approach its limit, as fusion codes take advantage of much larger supercomputers at the leadership computing sites. Thus, it may be prudent to plan for an upgrade of ESnet’s current connection to 100 Gbps within the two-to-five-year time frame.

10.5.2 Process of Science

Due to a factor-of-10 increase in the size of the restart file (for XGC1), if we still require four hours to transfer the restart file between supercomputer centers, it will require 50 Gbps network speed. Similarly, if we require 128 TB of field data to be transferred to the local server in 10 hours, the required transfer rate would be 25 Gbps.

10.6 Beyond 5 years – Future Needs and Scientific Direction

Beyond five years, we see data sizes continuing to grow and the transfer times remaining the same or decreasing, implying increased network speeds. The growth in data sizes will come both from increased resolution and from new physics phenomena being included in the calculations that imply an increased number of variables. Network speeds may not need to increase as rapidly as the growth rates of the data because new techniques in intelligent data reduction, feature extraction, and other compression techniques will be used to reduce the data.

10.7 Middleware Tools and Services

Several of the codes use ADIOS for transferring data from memory to disk while running. For transferring data over the network, scientists use BBCP and GridFTP as well as SCP and FTP. Scientists at PPPL use VISIT, IDL, and Matlab for visualizing data remotely (for example at NERSC). An NX connection to NERSC greatly facilitates the response time when viewing data. The FusionGrid is used to provide access to TRANSP for remote users.

10.8 Outstanding Issues

As discussed in Section 10.2.1, PPPL requests a dedicated backup circuit for its exclusive use to NYC/32 AOA if the main circuit is down

10.9 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	**Process of Science**	**Data Set Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
Near Term (0-2 years)				
• GTS • XGC 1 • M3D-C^1	• Core microstability • Edge microstability • Global stability	• 4 TB (restart) • TB (graphics) • TB (restart) • 2 TB (graphics) • 2 GB (restart) • 200 GB (graphics)	• Transfer restart files from scratch to archive disk in 1 hour	• Transfer graphics files to local host in 10 hours • Transfer restart files between supercomputer centers in 4 hours
2-5 years				
• GTS • XGC 1 • M3D-C^1	• Core microstability • Edge microstability • Global stability	• 40 TB (restart) • 120 TB (graphics) • 100 TB (restart) • 20 TB (graphics) • 20 GB (restart) • 2 TB (graphics)	• Transfer restart files from scratch to archive disk in 1 hour	• Transfer graphics files to local host in 10 hours • Transfer restart files between supercomputer centers in 4 hours
5+ years				
• Existing codes and new codes	• Higher resolution and new physics couplings	• 1 PB restart files • 1 PB graphics	• Transfer restart files from scratch to archive disk in 1 hour	• Transfer graphics files to local host in 10 hours • Transfer restart files between supercomputer centers in 4 hours

11 ORNL Computational Science Networking Requirements

11.1 Background

The Oak Ridge Leadership Computing Facility (OLCF) manages the computing program at ORNL for the Department of Energy (DOE) while the National Institute for Computational Sciences (NICS) runs the computing facility for the National Science Foundation (NSF). Each has a professional, experienced operational and engineering staff composed of groups in HPC operations, technology integration, user services, scientific computing, and application performance tools. The ORNL computer facility staff provides continuous operation of the center and immediate problem resolution. On evenings and weekends, operators provide first-line problem resolution for users; additional user support and system administrators are available on-call for more difficult problems. The primary systems used by fusion researchers include the following:

Jaguar is a Cray XT5 system consisting of 37,376 AMD six-core Opteron processors providing a peak performance of over 2.3 PF and 300 TB of memory. Access to our 10 PB Spider parallel file system is provided by 192 service I/O (SIO) nodes at over 240 GB/sec. External log-in nodes (decoupled from the XT5 system) provide a powerful compilation and interactive environment using dual-socket quad core AMD Opteron processors and 64 GB of memory. Jaguar is the world's most powerful computer system and is available to the international science community through the DOE Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program. Jaguar is currently being upgraded to contain more than 300,000 cores, with additional GPUs to provide from 10–20 PF of computing power. Additionally, fusion users have the option to use the Lens cluster and the Sith cluster to perform pre- and post-processing of data-intensive tasks, and both are connected to the Spider file system. ORNL also has a high-performance storage system (HPSS) capable of archiving hundreds of petabytes of data and can be accessed by all major leadership computing platforms. Incoming data are written to disk and later migrated to tape for long-term archival. This hierarchical infrastructure provides high-performance data transfers while leveraging cost-effective tape technologies. Tape storage is provided by robotic tape libraries. The center has three SL8500 tape libraries holding up to 10,000 cartridges each, and is in the process of deploying a fourth SL8500 this year. The libraries house 24 T10K-A tape drives (500 GB cartridges, uncompressed) and 32 T-10K-B tape drives (1 TB cartridges, uncompressed). Each drive has a bandwidth of 120 MB/sec. ORNL's HPSS disk storage is provided by DDN storage arrays with nearly a petabyte of capacity and over 12 GB/sec of bandwidth.

ORNL is connected to every major research network at rates of 10 Gbps or greater. Connectivity to these networks is provided via optical networking equipment owned and operated by UT-Battelle that runs over leased fiber-optic cable. This equipment can simultaneously carry either 192 10 Gbps circuits or 96 40 Gbps circuits and connects the OLCF to major networking hubs in Atlanta and Chicago. Currently, 16 of the 10 Gbps circuits are committed to various purposes, allowing for virtually unlimited expansion of networking capability. At present, the connections into ORNL include ESnet's Science Data Network, TeraGrid, Internet2, and Cheetah at 10 Gbps as well as ESnet's routed IP network at 20 Gbps and National Lambda Rail at 40 Gbps. ORNL operates the Cheetah research network for NSF. To meet the increasingly demanding needs of data transfers between major facilities, ORNL is participating in the Advanced Networking Initiative (ANI) that will provide a native 100 Gbps optical network in a loop that includes ORNL, Argonne National Laboratory, NERSC, and the MANLAN exchange in New York.

11.2 Simulations, LAN Requirements, and Application Requirements for the Current Generation of Simulations

Many applications from the fusion domain run at the OLCF, and some of the largest data-producing simulations are from the GTC, GTS, and XGC1 simulation codes. Typically we have seen restart files producing over 2 TB of data per restart time step, and analysis data from these simulations typically ranges anywhere from 100 GB to 2 TB per time output. Some of this is documented in the PPPL report, where the I/O utilizes the ADIOS framework, and much of the I/O is undergoing a transformation to a Service Oriented Architecture. Currently the simulations create up to 150 TB of data in a one-week simulation, and this data needs to be archived in HPSS. The file system and archival system both suffer from interference from other users and the scheduling system. A major challenge is to archive data to HPSS before it is removed from the file system. Similarly fairly aggressive optimizations have been taken in the ADIOS framework to remove as much of the I/O variability as possible, since I/O times can vary by over 10X from one write to another.

11.3 Current WAN Requirements

Figure 11.1 shows current network usage from a daily graph with a one-minute average, and a monthly graph with a one-hour average, for all incoming and outgoing traffic to/from ORNL. Much of the data for fusion that is moved over the WAN seems to be from ORNL to other places, rather than having a large amount of fusion data transferred into ORNL from the outside. This is different from many combustion simulations, which have a large amount of data moved into ORNL for post-processing. Part of the issue with large data analytics run at ORNL is the challenge of getting "real-time" analytics on demand. If data are to be moved outside of ORNL for analytics/visualization, then we expect that the scalar field data and a large portion of the particle data would be transferred. This could easily add up to 100 TB of data per simulation over a one-week time. Our fusion users' habit is to only move data when they want to do analytics/visualization on their local resources; this places a large demand on our network requirements, since moving 100 TB from ORNL to some lower-performing sites (e.g. <50 MB/sec) would take almost a month to move with no failures.

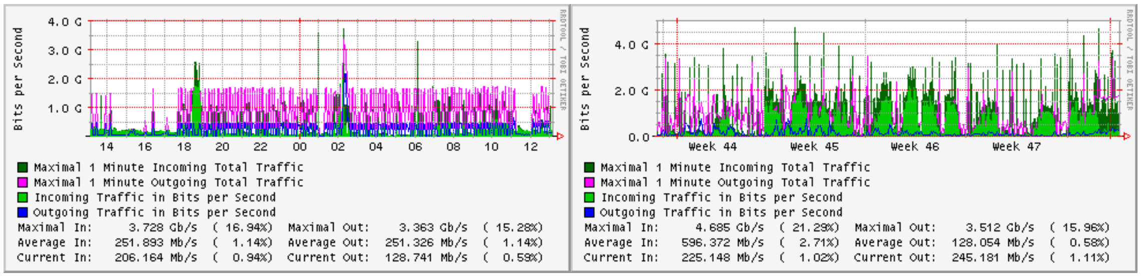


Figure 11.1. Usage for all incoming and outgoing traffic at ORNL. Left: Current daily network usage with a one-minute average. Right: Monthly graph with a one-hour average.

11.4 LAN and WAN Requirements over the Next 5 Years

We estimate that simulations such as XGC1 will start to run on the 10-20 Pflop machine at ORNL and these runs will initially be for C-Mod an will later transition to simulations for ITER, which will be much larger than the DIII-D simulations today. This gives us a restart file of over 10 TB for C-Mod and 100 TB for ITER per restart step. To cope with such demanding I/O, ADIOS is being outfitted with very advanced scheduling for staging, and (lossless) data compression to reduce the demands on the I/O system and we envision up to 10 PB of data written over a one-week period. Analysis and visualization data will be with the smaller-scale TEM turbulence at least, which translates into about 1 PB for ITER just for the field data over a one-day period.

11.5 LAN and WAN Requirements beyond 5 Years

Computing at ORNL will most likely be in the exaflop range, with much of the work focused on ITER. This will require 5 times the amount of data; it will also allow us to finish simulations in one day. This will have data for realistic collisions, and data size for ITER TEM turbulence will be about 10 times larger than the C-Mod runs. Analysis data will also be about 10 times larger, generating 1 PB of analysis data in one day. We anticipate that the data sets for visualization will be larger, with the increase due to the inclusion of particle data. We also envision looking at ETG turbulence in ITER, with data sizes up to 20 PB if we also store the neutrals.

11.6 Major Concerns

11.6.1 Why Move Data?

In fusion, we move data from experiments and simulations to serve as input for simulations. To date, this data movement has been very small, and typically we pre-stage data movement before the simulation. The implications of some future "coupled" runs are unclear, as they require much more data to be input for simulations; this may be an issue for future simulations. For output data, it is clear that restarts will affect I/O performance for many simulations. However, the eventual inclusion of NVRAM into compute nodes and the use of "I/O pipelines" — where data are preprocessed and reduced before writing to disk — will reduce the stress on data movement in the LAN and over the WAN. We can clearly see that we want “data-on-demand” when we process data on our local systems. We need to understand the algorithms used for analysis, and how to operate on “chunks” of data as it is moved, to ultimately deal with extremely large amounts of data. We still need fast and reliable data movement, and to have the system tell us when the data will be moved to us, and have some sort of guarantee on the arrival of the data.

11.6.2 Moving Code to Data

Ultimately we must begin to move to a model where the location of the data is not important – rather, there are guarantees for the bandwidth and latency of data access. Sometimes it is better to actually move code and processing to the data. Much of the data that is moved is moved only for analysis and visualization, and not checkpoint-restart.

11.6.3 Other Thoughts

Most data movement has no journaling that is known to the user or collaborators. At some point, data movers should understand and record (in a format that fits with the user’s data) information so that users can work with others to understand what happened during data movement, without “experts” in the loop. We would like the capture of this data to be automated, similar to the automated capture of performance data, and to have it placed in our metadata automatically.

[no table provided]

12 XGC Project

12.1 Background

The XGC project contains two kinetic particle codes — XGC0 and XGC1 — simulating tokamak plasmas in realistic diverted magnetic-field geometry. XGC codes simulate the background plasma physics together with the perturbed physics in multiscale by using the full-function (full-f) particle technique rather than the popular perturbed function (delta-f) technique. Hence, a full-function kinetic code contains more physics than a delta-f code, and is more expensive computationally. XGC0 is a drift-kinetic code and XGC1 is a gyrokinetic code. Both XGC0 and XGC1 scale efficiently to petascale on Jaguar and Hopper.

XGC0 has normally been running on Hopper at NERSC, using up to 80,000 cores for a one-day wall-clock job. XGC1 has mainly been running on Jaguar at OLCF, using up to 170,000 cores for a one-day wall-clock job. XGC1 has also been running on Hopper, using up to 120,000 cores for a two-day wall-clock job with restart submission. XGC1 produces large-size data, which has been handled by ADIOS technology. The data size is increasing rapidly as the physics capability of the code develops rapidly, from ion-scale turbulence (as described below) to the electron-scale turbulence. The network requirement, as described below, is based on XGC1 because it produces larger-size data than XGC0.

12.2 Current LAN Requirements and Science Process

The XGC1 simulation for the C-Mod tokamak uses about 100 billion particles and a mesh of about 5 million node points, and runs for 10,000 time steps. The restart file writes out the whole particle data every 1,000 time steps and its file size is (10^{11} particles) x (9 variables) x (8 byte) + field data = about 10 TB. The physics-study files for spatial field variable from grid nodes is written at every 10 time steps, and the file size of each time step is ($5*10^{7}$ data points) x (5 variables) x (8 byte) = 2 GB. The total file size of the grid field data (out of 10,000 time steps of simulation) is 2 TB. One simulation wall time is about 20 hours. Particle data are also needed for a more complete understanding of underlying physics, such as wave-particle interaction in phase-space. However, it is prohibitively expensive at present to write out the 10 TB particle data at every 10 time steps, totaling 10 PB.

The I/O of XGC1 is utilizing the ADIOS library, which enables parallel I/O of more than 25 GB/sec. 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 LAN speed requirement is about 2 GB/sec.

12.3 Current WAN Requirements and Science Process

Mostly, data analysis is performed on the remote supercomputing site. 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 is about 2 TB. For a data transfer in one hour, the network requirement is about 550 MB/sec.

In some situations, transferring the restart file (~10 TB) between NCCS and NERSC is required and, assuming four hours of transfer time, requires ~700 MB/sec of network speed.

12.4 LAN Requirements — the Next 5 Years

In five years, the number of particles and mesh points used by XGC1 will increase by an order of magnitude. Also, the v-space grid data will be added. The data size is anticipated to be about 10 TB at a time step. This requires a factor-of-10 increase in network speed (which means 250 GB/sec I/O speed) to the network file system for five minutes' writing of a restart file. If the grid data is written on every 100 time steps, the whole data set size for a 10,000 time-step simulation will be 1 PB. Four hours of data transfer to a local server requires ~70 GB/sec.

12.5 WAN Requirements – the Next 5 Years

Because of a factor-of-10 increase in the restart file size, four hours of restart file transfer will require 7 GB/sec network speed.

Grid data of a single time step (~10 TB) or whole-field data (~20 TB) are anticipated to be transferred to a scientist's local server, and considering four hours for the transfer time, the required network speed is about 700 MB/sec to 1.4 GB/sec.

[no table provided]

13 Glossary

GB/sec: Gigabytes per second — a measure of network bandwidth or data throughput

Gbps: Gigabits per second — a measure of network bandwidth or data throughput

MB/sec: Megabytes per second — a measure of network bandwidth or data throughput

Mbps: Megabits per second — a measure of network bandwidth or data throughput

PB/sec: Petabytes per second — a measure of network bandwidth or data throughput

Pbps: Petabits per second — a measure of network bandwidth or data throughput

TB/sec: Terabytes per second — a measure of network bandwidth or data throughput

Tbps: Terabits per second — a measure of network bandwidth or data throughput

ALCF	Argonne Leadership Computing Facility
AMD	Advanced Micro Devices
ANI	Advanced Networking Initiative
ASCR	Office of Advanced Scientific Computing Research
ASDEX	Axially Symmetric Divertor Experiment
ASIPP	Chinese Academy of Sciences Institute of Plasma Physics
CEMM	Center for Extended Magnetohydrodynamic Modeling
CPES	Center for Plasma Edge Simulation
CRPP	Plasma Physics Research Center, Switzerland
CSPM	Center for Simulation of Plasma Microturbulence
CSWIM	Center for Simulation of Wave Interaction with MHD
CSWPI	Center for Simulation of Wave-Plasma Interactions
delta-f	perturbed function
DEMO	Demonstration Power Plant
EAST	Experimental Advanced Superconducting Tokamak
ECE	electron cyclotron emission
ECH	electron cyclotron heating
ECRF	electron cyclotron resonance frequency
ECS	ESnet Collaboration Service
EFDA	European Fusion Development Agreement
ELM	edge-localized mode
EP	Extended Performance
ESnet	Energy Sciences Network
EVO	Enabling Virtual Organizations
FACETS	Framework Application for Core-Edge Transport Simulations
FES	Fusion Energy Sciences
FDT	Fast Data Transfer
FSP	Fusion Simulation Prototype; Fusion Simulation Project
FTP	File Transfer Protocol

full-f	full-function
GA	General Atomics
GDS	Global Dialing Scheme
GFDL	Geophysical Fluid Dynamics Laboratory
GLORIAD	Global Ring Network for Advanced Applications Development
GPS-TTBP	Center for Gyrokinetic Particle Simulations of Turbulent Transport in Burning Plasmas
GPU	graphics processing unit
HEDP	high energy density physics
HPC	high-performance computing
HPSS	high-performance storage system
ICF	Inertial Confinement Fusion
ICRF	ion cyclotron range of frequencies
IFERC	International Fusion Energy Research Centre
INCITE	Innovative and Novel Computational Impact on Theory and Experiment
I/O	input/output
IPP	Max Planck Institute of Plasma Physics
IPR	Institute for Plasma Research
ITPA	International Tokamak Physics Activity
JAEA	Japan Atomic Energy Agency
JET	Joint European Torus
KSTAR	Korea Superconducting Tokamak Advanced Research
LAN	local area network
LANL	Los Alamos National Laboratory
LBNL	Lawrence Berkeley National Laboratory
LHC	Large Hadron Collider
LHD	Large Helical Device
LLE	Laboratory for Laser Energetics
LNS	Laboratory for Nuclear Science
MAST	Mega Ampere Spherical Tokamak
MCU	multipoint control unit
MHD	magnetohydrodynamics
MIT	Massachusetts Institute of Technology
MPI	Message Passing Interface
NCCS	National Center for Computational Sciences
NERSC	National Energy Research Scientific Computing Center
NFRI	National Fusion Research Institute
NICS	National Institute for Computational Sciences
NIF	National Ignition Facility
NIFS	National Institute for Fusion Science
NLUF	National Laser Users' Facility
NSF	National Science Foundation
NSTX	National Spherical Torus Experiment

NVRAM	Non-Volatile Random Access Memory
OLCF	Oak Ridge Leadership Computing Facility
ORNL	Oak Ridge National Laboratory
OSCARS	On-demand Secure Circuits and Advance Reservation System
PCS	plasma control system
perfSONAR	Performance Service Oriented Network monitoring Architecture
PF	Petaflop
PPPL	Princeton Plasma Physics Laboratory
PSFC	Plasma Science & Fusion Center
QoS	quality of service
RAID	redundant array of independent disks
RF	radio frequency
ROAM	Resource Oriented Authorization Manager
SAN	storage area network
SC	Office of Science
SciDAC	Scientific Discovery through Advanced Computing
SCP	Secure Copy Program
SIO	Service I/O
SIP	Session Initiation Protocol
SOL	scrape-off layer
SST-1	Steady State Superconducting Tokamak
TCV	Tokamak à Configuration Variable
TAE	Toroidal Alven Eigenmode
TSM	Tivoli Storage Manager
UT	University of Texas
VRVS	Virtual Room Videoconferencing System
VTF	Versatile Toroidal Facility
VoIP	Voice Over Internet Protocol
WAN	Wide Area Network

14 Acknowledgements

This work would not have been possible without the contributions and participation of those who provided information and attended the workshop. ESnet would also like to thank the Fusion Energy Sciences program office for its help in organizing the workshop and providing insight into the facilities supported by the Fusion Energy Sciences program. The Oak Ridge Institute for Science and Education (ORISE) conference support and logistics staff was also very helpful.

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-XXXX