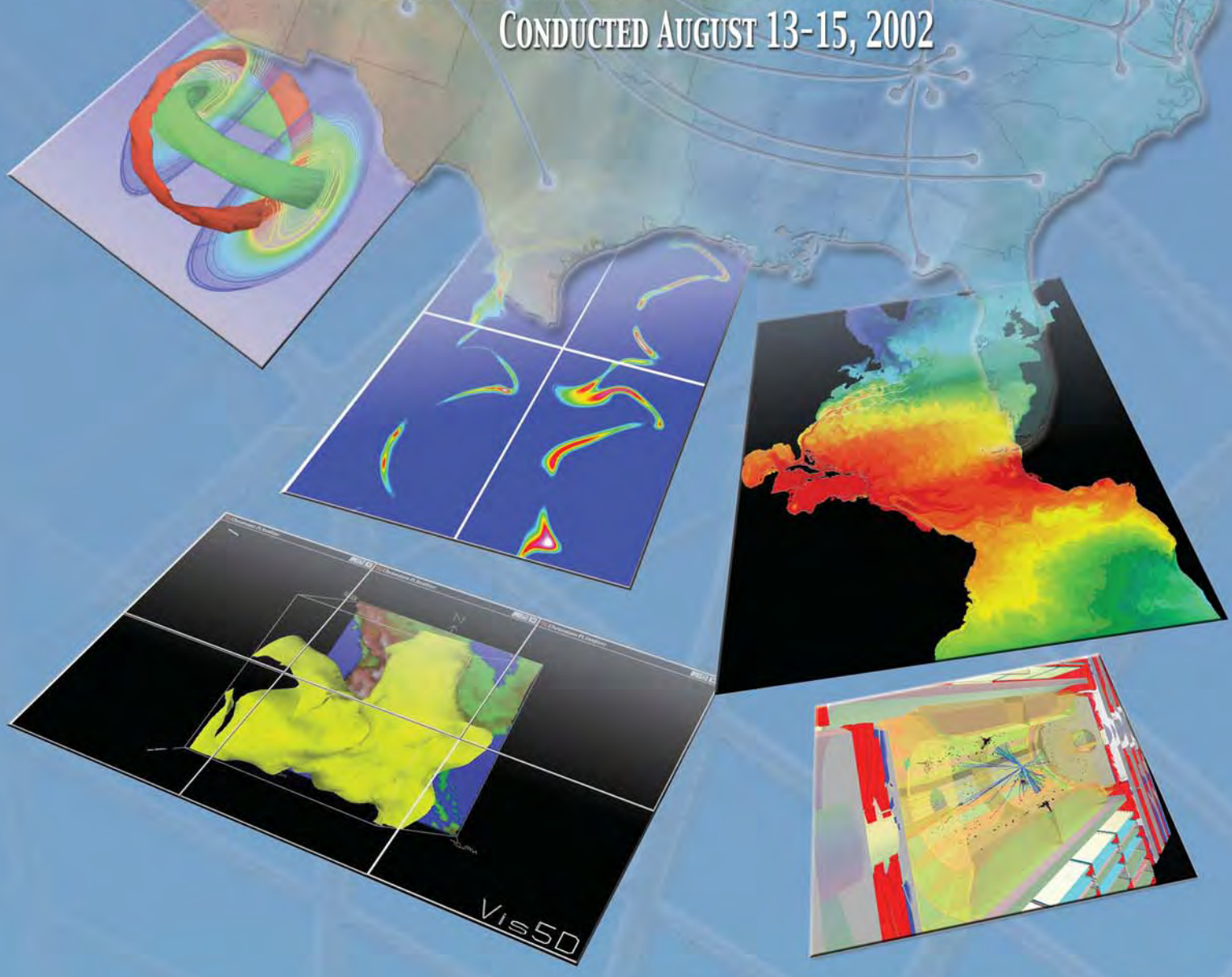
High-Performance Networks for High-Impact Science

Office of Science
U.S. Department of Energy

Report of the High-Performance Network Planning Workshop Conducted August 13-15, 2002



http://DOECollaboratory.pnl.gov/meetings/hpnpw/

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC06-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service,
U.S. Department of Commerce, 5285 Port Royal Rd., Springfield, VA 22161
ph: (800) 553-6847
fax: (703) 605-6900
email: orders@ntis.fedworld.gov
online ordering: http://www.ntis.gov/ordering.htm


This document was printed on recycled paper.
(8/00)

High-Performance Networks for High-Impact Science

Report of the August 13-15, 2002, Workshop
Conducted by the Office of Advanced Scientific Computing Research of the U.S. Department of Energy Office of Science

Report of the High-Performance Network Planning Workshop

DOE Organizing Committee

Mary Anne Scott, Chair
Dave Bader
Steve Eckstrand
Marvin Frazier
Dale Koelling
Vicky White

Contributors to the Report

Ray Bair, Editor

Deborah Agarwal	G. McDermott
Arthur S. Bland	Sandy Merola
Julian Bunn	Thomas Ndousse-Fetter
Charles Catlett	Harvey Newman
C.W. Cork	Larry Rahn
David Dixon	David Schissel
T.N. Earnest	Mary Anne Scott
Ian Foster	Gary Strand
Dennis Gannon	Rick Stevens
M.J. Greenwald	J.R. Taylor
Jason Hodges	Brian Tierney
William Johnston	James B. White III
William Kramer	Michael Wilde
James Leighton	Linda Winkler

Workshop Organization Support

Ray Bair
Charlie Catlett
Bill Johnston

Table of Contents

Figures

Tables

Executive Summary

The strategy for high-performance networking infrastructure in the U.S. Department of Energy (DOE) Office of Science is a corporate concern, and it is important at this time to reexamine that strategy. In the past decade, we have seen a revolution in telecommunications technology that has driven remarkable changes in the process of science. For example, massive datasets generated by experimental sciences that could not be shared a decade ago now are exchanged routinely and analyzed remotely among those institutions connected to the highest-speed backbone networks. But this is only a hint of the powerful changes in the scientific process that could occur.

The first step in developing a comprehensive networking infrastructure vision for the Office of Science is to understand the potential scientific impact of science unfettered by communication limitations. The process of defining a vision was initiated in advance of the workshop—through scenarios describing radical departures from how the science currently is done. With this vision in hand, a workshop was held August 13 through15, 2002, to examine what network provisioning, middleware service development and deployment, and network research need to be completed to enable these science scenarios over the next five to ten years. It brought together a selection of 55 end users, especially representing the emerging, high-visibility science initiatives, and network visionaries to identify opportunities and begin defining the path forward.

Advanced Infrastructure Enables DOE Science. In this workshop, researchers in a range of major Office of Science programs were asked to provide information on how they currently use networking and related services, and what they saw as the future process of their science that would require, or be enabled by, high-performance networks and advanced middleware services. Several general observations and conclusions may be made after analyzing these application scenarios.

- Increasingly, science depends critically on high-performance network infrastructure, where much of science already is a distributed endeavor or rapidly is becoming so.

- We can define a common "infrastructure" with advanced network and middleware capabilities needed for distributed science.

- Paradigm shifts resulting from increasing the scale and productivity of science depend on an integrated advanced infrastructure that is substantially beyond what we have today.

These paradigm shifts are not speculative. Several areas of DOE science already push the existing infrastructure to its limits as they implement elements of these approaches. Examples include high-energy physics with its world-wide collaborations distributing and analyzing petabytes[(a)] of data; systems biology access to hundreds of sequencing, annotation, proteome, and imaging databases that are growing rapidly in size and number; and the astronomy and astrophysics community that is federating huge observation databases so it can, for the first time, look all of the observations simultaneously. The clear message from the science application areas is that the *revolutionary shifts in the variety and effectiveness*

(a) 1 petabyte = 1,000 terabytes = 1,000,000 gigabytes = 10^9 megabytes = 10^{15} bytes.

of how science is done can only arise from a well integrated, widely deployed, and highly capable, distributed computing and data infrastructure, and not just any one element of it.

Enabling Middleware Research. Middleware is needed to translate the potential of fast, functional networks into actual scientific progress by enabling easier, faster access to, and integration of, remote information, computers, software, visualization and/or experimental devices—as well as interpersonal communication. Middleware makes it possible for an individual scientist or scientific community to address its application requirements, by

- facilitating the discovery and utilization of scientific data, computers, software, and instruments over the network in a controlled fashion
- integrating remote resources and collaboration capabilities into local experimental, computational, and visualization environments
- diagnosing (or averting) the cause of failures in these distributed systems
- managing, in a community setting, the authoring, publication, curation, and evolution of scientific data, programs, computations, and other products.

Grid middleware has shown considerable potential to provide much of the required integration and is an important element of the required science infrastructure. Grids currently are focused on resource access and management. This is a necessary first step but is not sufficient if we are to realize the potential of Grids for facilitating science and engineering. Grids are also evolving to incorporate web services for managing information.

This workshop identified six high-priority areas in which middleware research, development, deployment, and support are required to enable DOE science:

- *secure control over who does what,* where the challenging demands of DOE science lead to unique requirements
- *information integration and access,* to computers, storage, networks, code, services, instruments, and people
- *coscheduling and quality of service,* coordinating resources critical to many experimental and computational scenarios
- *effective network caching and computing,* to stage large datasets and rapidly access computing
- *services to support collaborative scientific work* among distant partners and collaborators
- *monitoring and problem diagnosis*—end-to-end and top-to-bottom for science applications and services.

Enabling Network Research. To enable the science applications and middleware capabilities, it is vital that "the network" be fast, dependable, predictable, and secure. Achieving those objectives in a high-performance, integrated science infrastructure raises a host of challenging research issues, including

- *network measurement and analysis*, a scalable infrastructure that provides end-to-end monitoring and diagnosis for both current capacities and future forecasts

- *high-performance transport protocols,* that significantly improve the end-to-end throughput of distributed science applications, e.g., ultra high-speed data transfer, remote visualization, and distributed supercomputing

- *multicast and secure group communication,* for real-time collaboration and control

- *service models,* to provide reliable network service guarantees for time-critical and scheduled activities.

Network Provisioning Model. To be responsive to applications' needs in a timely manner, the programs of the Office of Science would benefit from the formation of an integrated three-element network provisioning model that provides

1. *production level networking* in support of traditional program requirements

2. *network resources for high-impact DOE science programs* including science application and Grid research—This element provides a platform for deploying prototypes to those programs that require high-capability networking or advanced services that are not satisfied by production level networking.

3. *network resources for network research* that enable experimentation with new concepts.

An integrated network provisioning strategy would benefit from a process of planning, coordination, funding, and implementation that encompasses all three elements. Factors that should be taken into consideration include the following:

- *A shared vision of success must be motivated, where some measures of success are across all three elements.*

- *As new services are moved into production, some production support costs likely will increase.*

- *The network program must position itself to be agile and not rooted too firmly in any one provisioning model.*

Network Governance Model. The current governance model includes DOE program management components, national laboratory and university project management components, and forums for input (e.g., standing committees, workshops). An integrated networking provisioning strategy will require revisiting the current governance model.

- *DOE science networks currently have no network governance model overarching the three elements.*

- *The governance model for integrated network provisioning must allow for the management of each element's requirements in a context that is highly influenced by the opportunities and risks confronting the other two elements.*

- *The steering structure for an integrated high-performance network program should include a breadth of representation across the Office of Science.*

- *The networking program would benefit from an approach similar to that used to allocate computing resources.*

Path Forward. It is essential to develop a detailed implementation strategy that outlines what network services are required by the community in each of the three elements (production level networking, resources for high-impact science programs, and resources for network research), and what opportunities exist for providing these services. The detailed analysis needed is beyond the scope of this workshop and requires a team of experts from both within and outside the DOE community.

Multiple opportunities must be evaluated, some of which are time-critical, as they leverage efforts of other agencies and the academic community.

A key challenge for such an analysis is that the DOE strategy must be integrated, both at a high level between programs and at a technical level between components of the infrastructure. To be successful, an integrated program would require

- a *road map* that expresses the future of network elements, network research, and middleware research in the context of a shared networking vision across the networking program and the Office of Science

- a *network research program*, geared to address the issues of scale presented by emerging network requirements of DOE's high-impact science applications

- a *middleware research program*, geared to address the issues of complexity and diversity of DOE's widely distributed science collaborations, compute and data resources

- a *middleware deployment program*, to deliver the ubiquitous infrastructure for science

- a *network provisioning model* that provides a flexible and dynamic network and Grid infrastructure for all three elements

- a *network business model* that optimizes the services provided to applications and research, in the context of an evolving set of commercial services and opportunities, and the growing size and scope of the network

- a *network governance model* that features the participation of DOE science programs and program management in the planning, prioritization, and allocation of network offerings at all levels.

High-Capability Networking and Middleware Enable Advanced Science. A major network initiative is probably necessary to break the 'zero sum game' in networking that has confronted the community for many years, limiting what can be done and slowing the progress of the programs across the Office of Science. The Scientific Discovery through Advanced Computing (SciDAC) initiative model is one worth considering, as it champions ownership of the initiative and its efforts across the entire Office of Science.

Abbreviations Used in This Report

AIMD	additive increase multiplicative decrease
API	application program interface
ASM	Any-Source Multicast
CMS	Compact Muon Solenoid
CCSM	Community Climate System Model
DNS	domain name server
DOE	U.S. Department of Energy
e-commerce	electronic commerce
ECN	explicit congestion notification
FES	Fusion Energy Sciences
FTP	File Transfer Protocol
GGF	Global Grid Forum
GriPhyN	Grid Physics Network
HTML	hyper text markup language
IP	Internet Protocol
LHC	CERN's Large Hadron Collider
MICS	Mathematical, Information, and Computational Sciences
MPLS	Multiprotocol Label Switching
MTU	Maximum Transmission Unit
NACP	North American Carbon Project
NAT	network address translation
NCAR	National Center for Atmospheric Research
NERSC	National Energy Research Scientific Computing Center
NSF	National Science Foundation
NWS	Network Weather Service
ORNL	Oak Ridge National Laboratory
PCM	Parallel Climate Model
PKI	Public Key Infrastructure

QoS	quality of service
RHIC	Relativistic Heavy Ion Collider
SciDAC	Scientific Discovery through Advanced Computing
SCTP	Stream Control Transmission Protocol
SNMP	Simple Network Management Protocol
SNS	Spallation Neutron Source
SOAP	Simple Object Access Protocol
SSH	secure shell
SSL	Secure Sockets Layer
ST	scheduled transfer
TCP	Transmission Control Protocol
TE	Traffic Engineering
TLS	Transport Layer Security
UCAR	University Corporation for Atmospheric Research
UDP	User Datagram Protocol
XCP	eXplicit Control Protocol
XML	extensible markup language

Chapter 1 Introduction

In the past decade, we have seen a revolution in network and telecommunications technology that has driven some remarkable changes in the process of science. For example, the massive datasets generated by experimental sciences that were all but unsharable a decade ago now are exchanged routinely and analyzed remotely among those institutions that have appropriate connections to the very high-speed backbone networks. However, this provides only a hint of the potential changes in the scientific process when such bandwidth becomes fully deployed in the scientific community.

The strategy for high-performance networking infrastructure in the U.S. Department of Energy (DOE) Office of Science is a corporate concern. The vision of what is possible in the realm of science, together with a number of influences that are at work, have brought us to a juncture where it is important for the Office of Science to re-examine its strategy for high-performance scientific networking.

- The current approach for providing a high-quality production network backbone (i.e., ESnet) that is responsive to the bandwidth and connectivity requirements of the program offices is fast approaching the point at which resources are not sufficient to continue being responsive to all of the needs.

- There is an increasing awareness that to realize the vision, the end-to-end bandwidth problem must be solved, not just the backbone bandwidth problem.

- With the rapid development and deployment of Grid technologies in support of many applications, there is an increased need for advanced services to be provided along with the high-performance networking infrastructure.

- Advances in optical networks and rapid sweeping changes within the telecommunications industry are creating opportunities for fundamentally different business models and partnerships that have the potential to enable dramatic improvements in the price-performance ratio of wide-area networks.

1.1 Vision

Science applications and specialized experimental facilities are *n*-way interconnected to terascale computing, petascale storage, high-end visualization, and remote collaborators in a seamless environment that provides the performance levels required to move science, especially large-scale science, to a new regime. In this regime, seamless collaboration among scientists and between scientists and experimental and computational resources eliminates isolation, discourages redundant efforts, and promotes rapid scientific progress through the interplay of theory, simulation, and experiment.

1.2 Workshop Objectives

The first step in developing a new strategy for this vision is to understand the scientific potential of greatly increased network capabilities for the major applications drivers within the Office of Science. How would the process of science change if available bandwidth no longer was a limiting factor and middleware services were available to facilitate the routine construction and use of widely distributed

science environments? What level of connectivity and bandwidth does it take for these elements to cease to be a problem? The process of defining a vision of science unfettered by communication limitations was initiated in advance of the workshop through the development of a vision for where science applications would like to be in five to ten years; in some cases, this vision is a radical departure from how the science is currently done. With this vision in hand, we can identify what network provisioning, middleware service development and deployment, and network research needs to be completed to enable these scenarios over the next five to ten years.

The workshop was the first major activity in developing a strategic plan for high-performance networking in the Office of Science. Held August 13 through 15, 2002, it brought together a selection of end users, especially representing the emerging, high-visibility initiatives, and network visionaries to identify opportunities and begin defining the path forward.

This report documents the workshop.

In advance of the workshop, science scenarios were developed by discipline scientists, often working with networking and middleware researchers. These scientists developed a vision of where their high-impact science applications need to be in five to ten years; they also outlined the network capabilities and associated services required to carry out the science visions. The scenarios covered eight research domains: particle physics (several documents), magnetic fusion, chemical sciences, climate modeling, neutron-scattering sciences, macromolecular crystallography, systems biology, and astronomy and astrophysics. Although these domains are just a subset of the fundamental science activities in the Office of Science, they span the major patterns of network use for science in DOE.

During the workshop, these scenarios, along with other information offered by those present, were used to develop application-driven network infrastructure requirements. The resulting requirements are summarized in Chapter 2 of this report. The science scenarios also served as input to identify important middleware and network research areas and to evaluate provisioning strategies for future networks. Chapters 3 and 4 provide summaries of the middleware and network research requirements needed to support applications like these over the next five to ten years. Workshop participants also considered provisioning strategies and business models for providing network infrastructure to support DOE science; these are summarized in Chapter 5. Findings about the next steps needed to develop a road map for an integrated network provisioning program for the Office of Science programs also are presented in Chapter 5. A series of appendixes includes more information on application requirements and specifics of the workshop agenda and attendees. The full science scenarios themselves are reproduced in their entirety on the Internet at http://DOECollaboratory.pnl.gov/meetings/hpnpw/finalreport/.

Chapter 2 Advanced Infrastructure as an Enabler for Future Science

2.1 General Observations

In this workshop, representatives of a range of DOE science disciplines were asked to provide information on how they currently use networking and network-associated services and what they saw as the future process of their science that would require, or be enabled by, high-speed networks and advanced middleware support.

Several general observations and conclusions may be made after analyzing these application scenarios.

The first and perhaps most significant observation is that a lot of science already is, or rapidly is becoming, an inherently distributed endeavor. Science experiments involve a collection of collaborators who frequently are multi-institutional, where data and computing requirements are addressed routinely with compute and data resources that frequently are even more widely distributed than the collaborators. Further, as scientific instruments become more and more complex (and therefore more expensive), they frequently are used as shared facilities with remote users. Even numerical simulation—an endeavor previously centered on one, or a few, supercomputers—is becoming a distributed endeavor. Such simulations are increasingly producing data of sufficient fidelity that it is used in post-simulation situations—as input to other simulations, to guide laboratory experiments, or to validate or calibrate other approaches to the same problem. This sort of science depends critically on an infrastructure that supports the process of distributed science.

A second observation is that when asked what sort of services are needed to support distributed science, the answer always involves many significant middleware and collaboration services beyond just basic computing and networking capacity.

A third observation is that there is considerable commonality in the services needed by the various science disciplines. This means that we can define a common "infrastructure" for distributed science.

Fourth, every one of the science areas needs high-speed networks and advanced middleware to couple, manage, and access resources like the widely distributed, high-performance computing systems, the many medium-scale systems of the scientific collaborations, high data-rate instruments, and the massive data archives that, together, are critical to next-generation science and to support highly interactive, large-scale collaboration. That is, all of these elements are required to produce an advanced distributed computing, data, and collaboration infrastructure for science that will enable paradigm shifts in how science is conducted. Paradigm shifts resulting from increasing the scale and productivity of science depend completely on such an ***integrated advanced infrastructure*** that is substantially beyond what we have today. Further, these paradigm shifts are not speculative. Several areas of DOE science already are pushing the existing infrastructure to its limits while trying to move to the next generation of science. Examples include high-energy physics with its world-wide collaborations analyzing petabytes of data (described in Section 2.6) and the data-driven astronomy and astrophysics community that is trying to

federate the huge databases being generated by a new generation of observing instruments so that entirely new science can be done by looking at all of the observations simultaneously (e.g., the National Virtual Observatory [1] illustrates this point very well. Specifically see "New Science: Rare Object Searches" in [2].)

As indicated by Figure 2.1, this integrated advanced infrastructure

- Provides the DOE science community with advanced distributed computing infrastructure based on large-scale computing, high-speed networking, and Grid middleware.

- Enables the collaborative and interactive use of the next generation of massive data-producing scientific instruments.

- Facilitates large-scale scientific collaborations that integrate the DOE laboratories and universities.

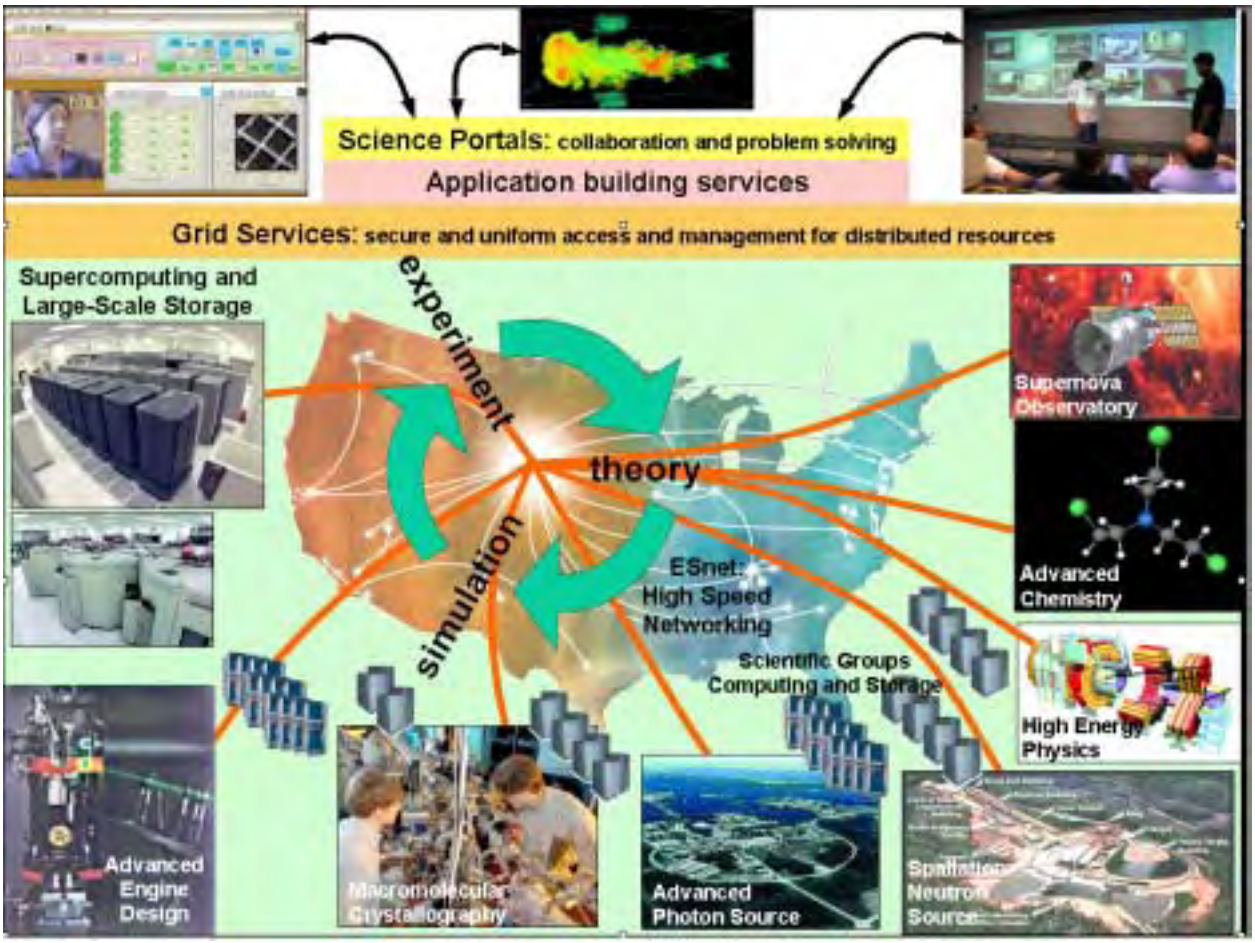


Figure 2.1. Integrated cyber-infrastructure enables advanced science.

There is a clear trend toward the need for services that allow distributed science activities to scale up in several ways—for example, in the number of participants in a distributed collaboration, the amount of data that can be managed, the diversity of the use of data, the number of people who can discover and use the data, the number of independent computational simulations that can be combined to represent a more realistic or complex phenomenon or physical system.

The challenging task of the integrated advanced infrastructure is to deliver an overall computing, data, and collaboration quality of service to scientific projects, with these key features:

- Computing capacity adequate for a science task is provided at the time it is needed.
- Data capacity sufficient for the science task is provided independent of location and in a transparently managed, global name space.
- Communication capacity sufficient to support all of the aforementioned is provided transparently to both systems and users.
- Software services support a rich environment that lets scientists collaboratively focus on the science simulation and analysis aspects of software and problem-solving systems rather than on the details of managing the underlying computing, data, and communication resources.

The clear message from all of the science application areas is that the paradigm shifts in how science is done will come about from a well integrated, widely deployed, highly capable distributed computing and data infrastructure and not just any one element of it.

The requirements for the highly capable distributed science environments needed to support the sorts of science described above include a range of technologies, all of which must be integrated and persistent. The technologies that we discuss here either are being deployed today or are in development. This is not a list of things that will require a decade of computer science research before we can deploy them. On the other hand, a good deal of development and deployment remains to be done to make these technologies into a highly capable infrastructure.

Two years ago, we did not have the systems, communications, tools, or experience to do this. Today, we are at a point where building and deploying this infrastructure is possible in the three- to five-year time frame in all of the technology areas, given adequate support.

2.2 Collaboration Capabilities and Facilities Access

Scientific collaborations require access to data storage facilities, information, scientific instruments, and other collaborators. Several different patterns of interaction emerge among the scientists, computer systems, instruments, and data repositories. The most common patterns divide into three general models. The interaction models, shown in Figures 2.2, 2.3, and 2.4, differ in ways that significantly affect the network capacities required.

Figure 2.2 illustrates tiered access to community resources, typically large data collections, in which specific portions of a many-terabyte to many-petabyte database are replicated strategically and/or cached at locations readily accessible to users at specific institutions. High-energy particle physics research is developing this type of distribution. To some extent, the genomics community also has adopted a similar pattern but with replication of entire repositories and creation of local augmented repositories. In each case, the primary data sources are at specific locations determined by facility siting decisions. Replication and update requirements drive network capacity needs between the primary repositories and mid-level repositories. There may, in fact, be multiple intermediate tiers. Typically, the end users need high-performance access to the mid-level centers; tools and predictable network capacity are needed there for distributing the data to the various centers that specialize in the data categories appropriate to their research.

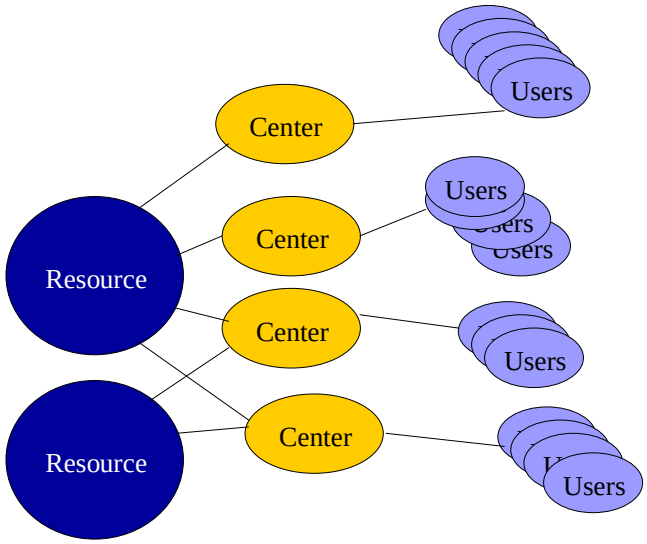


Figure 2.2. Many scientists share a small number of resources.

In other domains such as macromolecular crystallography, no single computer, instrument, or data resource is shared by a majority of the community. Instead, individual scientists or groups at an institution may utilize high-performance computers at different locations for specific projects or types of computations, access multiple instruments based on the earliest available, and retrieve data from multiple databases. This pattern, diagrammed in Figure 2.3, requires a generalized n-way interconnection, with performance determined by the most demanding types of interactions such as real-time coupling of instrument data and simulations.

In some domains, scientists or science groups interact with the different resources largely independently, working within a topical area (see Figure 2.4). However, they also collaborate across tasks to coordinate an attack on a large scientific question. This pattern includes the need for support of intense person-to-person collaboration on a demand basis. For example, some chemical sciences or computer science research has this characteristic.

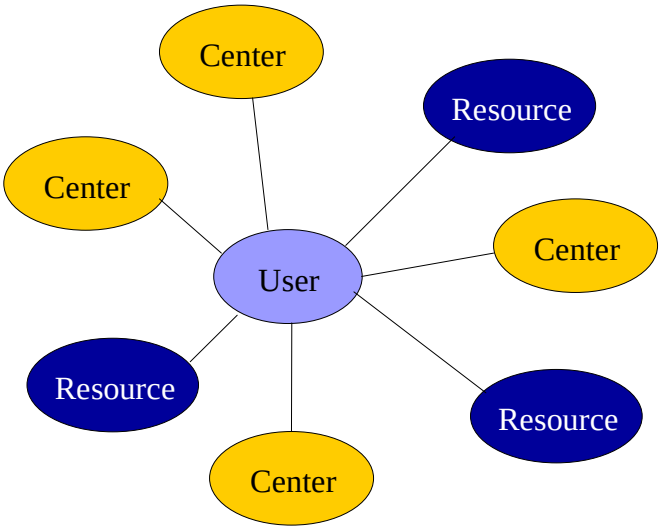


Figure 2.3. Individual scientists interact with many resources.

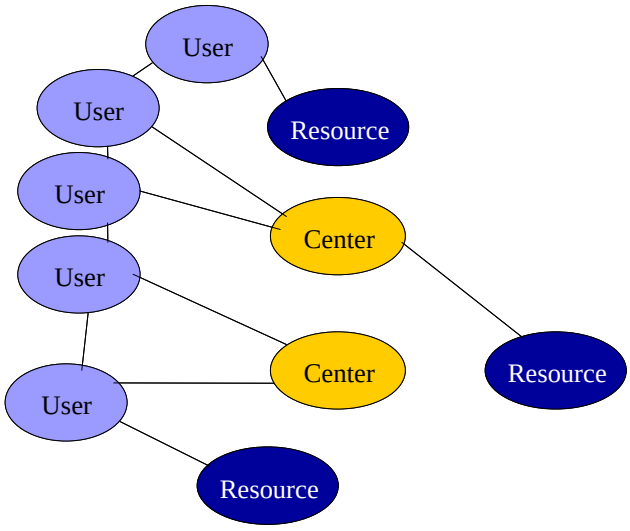


Figure 2.4. Scientists interact independently with resources and each other.

Overall, no particular scientific endeavor may fit any of these patterns precisely. However, examining the characteristics of such patterns sheds light on network capacity, services and connectivity required to address challenging science problems. To develop an informed road map, it is important to thoroughly understand how data moves in a community and how it is shared among and beyond the team that creates it.

Looking at the scientific processes that must be supported by high-performance networks for collaboration and facility access, several important themes are present. It is clear from examining the science

scenarios that network advances are critical to supporting the altered science processes dictated by high-throughput experiments, more detailed simulations, and integrated science communities being proposed by DOE scientists. The scientific processes that characterize such efforts will severely challenge traditional network provisioning strategies. It is clear that the distribution of experimental facilities, data storage facilities, and researchers is increasing along with the size of the datasets. In addition, there's a clear need to support much more tightly integrated research among ever larger sets of collaborators. Although past network service and provisioning approaches might have addressed the needs of a particular community, the aggregate needs of DOE science in the next five years call for a new strategy and a much more aggressive approach to providing high performance and reliable network capabilities.

Across the board, the application communities also require a secure, seamless, and ubiquitous authentication and authorization infrastructure. Without it, they cannot justify making ongoing investment in remote access to many of the advanced scientific processes that high-performance networks enable or, alternately, the costs of doing so becomes quite large. In many respects, we are beginning to treat DOE-funded laboratories and universities as well as other partner institutions as a highly connected scientific enterprise. However, to make this a reality, we need not only the middleware and services of the Grid but also higher-level services that facilitate collaboration among remote scientists and facilitated access to distributed data.

The following sections summarize the requirements of seven major programs or facilities in the Office of Science. Although this list of efforts is far from exhaustive, it covers the principal modes of high-performance network usage envisioned. Additional material is available in the appendixes, including an expanded version of these summaries and the science scenarios prepared in advance of the workshop.

2.3 Climate Modeling Requirements

To better understand climate change, we need better climate models. Climate models today are too low in resolution to get some important features of the climate right. To determine things like climate extremes (hurricanes [1], drought and precipitation pattern changes [2], heat waves and cold snaps) and other potential changes as a result of climate change [3], we need better analysis. Over the next five years, climate models will see an even greater increase in complexity than that seen in the last ten years. The North American Carbon Project (NACP), which endeavors to fully simulate the carbon cycle, is an example. Increases in resolution, both spatially and temporally, are in the plans for the next two to three years. The atmospheric component of the coupled system will have a horizontal resolution of approximately 150 km and 30 levels. A plan is being finalized for model simulations that will create about 30 terabytes of data in the next 18 months.

Table 2.1. Climate Modeling Requirements Summary

Feature	Characteristics That Motivate Advanced Infrastructure	Vision for the Future Process of Science	Anticipated Requirements	
Time Frame			**Network**	**Middleware**
Near-term	*A few data repositories, many distributed computing sites • NCAR[(a)] - 20 Tbytes • NERSC[(b)] - 40 Tbytes • ORNL[(c)] – 40 Tbytes		• Authenticated data streams for easier site access through firewalls	• Server side data processing (computing and data cache embedded in the net) • Information servers for global data catalogues
5 years	• Add many simulation elements/components as understanding increases • 100 Tbytes / 100 model yrs generated simulation data – 1-5 Pbytes / yr (at NCAR) • Distribute in large datasets to major users/collaborators for post-simulation analysis	• Enable the analysis of model data by all of the collaborating community (major US collaborators are a dozen universities, and several Federal Agencies)	• Robust access to large quantities of data (multiple paths)	• Reliable data/file transfer o Across system/ network failures
5+ years	• Add many diverse simulation elements/components, including from other disciplines - this must be done with distributed, multidisciplinary simulation as the many specialized sub-models will be managed by experts in those fields • 5-10 Pbytes/yr (at NCAR)	• Integrated climate simulation that includes all high-impact factors	• Robust networks supporting distributed simulation - adequate bandwidth and latency for remote analysis and visualization of massive datasets	• Quality of service guarantees for distributed, simulations • Server side computation for data extraction/ subsetting, reduction, etc., before moving across the network
	• Virtualized data to reduce storage load			• Virtual data catalogues for data generation descriptions, data regeneration planners, data naming and location transparency services for reconstituting data on demand

(a) NCAR = National Center for Atmospheric Research.
(b) NERSC = National Energy Research Scientific Computing Center
(c) ORNL = Oak Ridge National Laboratory

These studies will be driven by the need to determine future climate at both local and regional scales as well as changes in climate extremes—droughts, floods, severe storm events, and other phenomena. Over the next five years, climate models also will incorporate the vastly increased volume of observational data now available (and available in the future), both for hind casting and intercomparison purposes. The end result is that instead of tens of terabytes of data per model instantiation, hundreds of terabytes to a few petabytes (10^{15}) of data will be stored at multiple computing sites, to be analyzed by climate scientists worldwide. The Earth System Grid and its descendents will be fully utilized to disseminate model data and for scientific analysis. Additionally, these more sophisticated analyses and collaborations will demand much greater bandwidth and robustness from computer networks than is now available.

As climate models become more multidisciplinary, scientists from fields outside of climate studies, oceanography, and the atmospheric sciences will collaborate on the development and examination of climate models. Biologists, hydrologists, economists, and others will assist in the creation of additional components that represent important but as-yet poorly known influences on climate. These models, sophisticated in themselves, will likely be run at computing sites other than where the parent climate model was executed. To maintain efficiency, data flow to and from these collaboratory efforts will demand extremely robust and fast networks.

In the period five to ten years out, climate models will again increase in resolution, and many more fully interactive components will be integrated. At this time, the atmospheric component may become nearly mesoscale (commonly used for weather forecasting) in resolution, 30 km by 30 km, with 60 vertical levels. Climate models will be used to drive regional-scale climate and weather models, which require resolutions in the tens to hundreds of meters range, instead of the hundreds of kilometers resolution of today's Community Climate System Model (CCSM) and Parallel Climate Model (PCM). There will be a true carbon cycle component, where models of biological processes will be used, for example, to simulate marine biochemistry and fully dynamic vegetation. These scenarios will include human population change, growth, and econometric models to simulate the potential changes in natural resource usage and efficiency. Additionally, models representing solar processes, to better simulate the incoming solar radiation, will be integrated. Climate models at this level of sophistication will likely be run at more than one computing center in distributed fashion, which will demand extremely high speed and tremendously robust computer networks to interconnect them. Model data volumes could reach several petabytes, which is a conservative estimate.

2.4 Spallation Neutron Source Requirements

Six DOE laboratories are partners in the design and construction of the Spallation Neutron Source (SNS), a one-of-a-kind facility at Oak Ridge, Tennessee, that will provide the most intense pulsed neutron beams in the world for scientific research and industrial development. When completed in early 2006, the SNS will enable new levels of investigation into the properties of materials of interest to chemists, condensed matter physicists, biologists, pharmacologists, materials scientists, and engineers, in an ever-increasing range of applications.

The SNS supports multiple instruments that will offer users at least an order of magnitude performance enhancement over any of today's pulsed spallation neutron source instruments (Figure 2.5). This great increase in instrument performance is mirrored by an increase in data output from each instrument. In

Figure 2.5. Spallation Neutron Source Facility at Oak Ridge National Laboratory

fact, the use of high-resolution detector arrays and supermirror neutron guides in SNS instruments means that the data output rate for each instrument is likely to be close to two orders greater than a comparable U.S. instrument in use today. This, combined with increased collaboration among the several related U.S. facilities, will require a new approach to data handling, analysis, and sharing.

The high data rates and volumes from the new instruments will call for significant data analysis to be completed offsite on high-performance computing systems. High-performance network and distributed computer systems will handle all aspects of post-experiment data analysis and the approximate analysis that can be used to support near real-time interactions of scientists with their experiments.

Each user is given a specific amount of time (0.5 to 2 days) on an instrument. The close to real-time visualization and partial analysis capabilities, therefore, allow a user to refine the experiment during the allotted time. For the majority of SNS user experiments, the material or property being studied is novel, and this capability is essential for the experimentalist to focus in on the area of interest and maximize the science accomplished in the limited amount of beam time.

In this scenario, the combined data transfer between the 12 SNS instruments and a distributed computer network for real-time data mapping is estimated to be a constant 1 Gbits/sec (assuming 50% of users using real-time visualization). The return data stream to servers managing the visualization and analysis

tasks as well as communicating to the users across local area networks (LANs) and/or the Internet likely will be around 140 Mbits/sec (dominated by the four- and three-dimensional response maps). The servers (one for each instrument) would generate selected views of the response function as well as send the response function back out to the distributed computer network for quick/partial analysis.

It is anticipated that analysis of experimental data in the future may be achieved by incorporating a scattering law model within the iterative response function extraction procedure. These advanced analysis methods are expected to require the use of powerful offsite computing systems, and the data may transit the network several times as experiment/experimenter/simulation interaction converges to an accurate representation.

Table 2.2. Spallation Neutron Source Requirements Summary

Feature	Characteristics That Motivate Advanced Infrastructure	Vision for the Future Process of Science	Anticipated Requirements	
Time Frame			**Networking**	**Middleware**
Near term	(Facility comes on-line in 2006)			
5 years	• The 12 instruments at the SNS will operate about 200 days/year and generate an aggregate 80 Gbytes/day • The data analysis will be accomplished mostly on computing systems that are remote from the SNS		• 50-80 Mbits/sec sustained • 320 Mbits/sec peak	• Workflow management • Reliable data transfer
	• Neutron scattering instruments operate 24 hr 7 days a week during facility run periods, real time data visualization, some real time analysis capabilities, and security to modify experiment conditions by a user at his/her hotel via an internet browser will be required.	• Real-time data analysis and visualization will enhance the productivity of the science done at SNS, which runs 24 hr/day.	• 1 Gbits/sec sustained	• Security (authentication and access control) to permit direct interaction with the instrument remotely.
5-10 years	• Statistical scattering models will be incorporated into analysis code requiring supercomputer levels of remote computing.	• Iterative analysis of the data with the use of models running on supercomputing systems will produce much more accurate results and understanding.		

2.5 Macromolecular Crystallography Requirements

Macromolecular crystallography is an experimental technique that is used to solve structures of large biological molecules (such as proteins) and complexes of these molecules. The current state-of-the-art implementation of this technique requires the use of a source of very intense, tunable, x-rays that are produced only at large synchrotron radiation facilities. In the United States, 36 crystallography stations are distributed among the synchrotron facilities and dedicated to macromolecular crystallography[4]. The high operating cost of these facilities, coupled with the heavy demand for their use, has led to an emphasis on increased productivity and data quality that will need to be accompanied by increased network performance and functionality.

The data acquisition process involves several interactive online components, data archiving and storage components, and a compute-intensive offline component. Each component has associated networking requirements. Online process control and online data analysis are real-time, interactive activities that monitor and coordinate data collection. They require high-bandwidth access to images as they are acquired from the detector. Online data analysis now is limited primarily to sample quality assurance and to data collection strategy. There is increasing emphasis on expanding this role to include improved crystal scoring methods and real-time data processing to monitor sample degradation and data quality. Online access to the image datasets is collocated and could make good use of intelligent caching schemes. Datasets from previously exposed samples are not required during online processing.

High-performance networking can play several roles in online control and data processing. Bob Sweet at the Brookhaven National Laboratory National Synchrotron Light Source has outlined several approaches to remote, networked, collaboratory operation [5]. The datasets most often are transferred to private institutional storage. This requirement places a large burden on the data archiving process that transfers the data between online and offline storage units. Current requirements for average data transfer rate are 1 to 25 Mbytes/s per station; it is expected that in five to ten years, this will increase by an order of magnitude to 10 to 250 Mbytes/s per station. This is exacerbated further by the fact that most research facilities have from four to eight stations; this places a future requirement of 40 to 2000 Mbytes/s per facility. Advanced data compression schemes might be able to reduce these figures by a factor of 5 to 10.

In addition to increased raw network bandwidth, the next-generation high-performance networking infrastructure will need to provide tools and services that facilitate object discovery, security, and reliability. These tools are needed for low-latency applications such as remote control as well as high-throughput data transfer applications such as data replication or virtual storage systems.

2.6 High-Energy Physics Requirements

The major high-energy physics experiments of the next twenty years will break new ground in our understanding of the fundamental interactions, structures, and symmetries that govern the nature of matter and space-time. The largest collaborations today, such as CMS [6] and ATLAS [7], are building experiments for CERN's Large Hadron Collider (LHC) program [8] and encompass 2000 physicists from

150 institutions in more than 30 countries. Each of these collaborations includes 300 to 400 physicists in the United States, from more than 30 universities, as well as the major U.S. high-energy physics laboratories.

The high-energy physics problems are the most data-intensive known. The current generation of operational experiments at SLAC (BaBar [9]) and FermiLab (D0 [10] and CDF [11]), as well as the experiments at the Relativistic Heavy Ion Collider (RHIC) program at Brookhaven National Laboratory [12], face many data and collaboration challenges. BaBar in particular already has accumulated datasets approaching a petabyte (10^{15} bytes). These datasets will increase in size from petabytes to exabytes (1 exabyte = 10^{18} bytes) within the next decade. Hundreds to thousands of scientist-developers around the world continually develop software to better select candidate physics signals, better calibrate the detector, and better reconstruct the quantities of interest. The globally distributed ensemble of facilities, while large by any standard, is less than the physicists require to do work in an unbridled way. There is thus a need and a driver to solve the problem of managing global resources in an optimal way to maximize the potential of the major experiments for breakthrough discoveries.

Several important collaborations already are involved in the high-energy physics work to use Grids for distributed data processing. For example, the Grid Physics Network (GriPhyN) project (http://www.pgriphyn.org) is a collaboration of computer science and other information technology researchers and physicists from the ATLAS, CMS, LIGO, and SDSS experiments. The project is focused on the creation of petascale virtual data grids that meet the data-intensive computational needs of a diverse community of thousands of scientists spread across the globe (Figure 2.6).

Collaborations on this global scale would not have been attempted if the physicists could not plan on excellent networks—to interconnect the physics groups throughout the life cycle of the experiment and to make possible the construction of Data Grids capable of providing access, processing, and analysis of massive datasets. The physicists also must be able to count on excellent middleware to facilitate the management of worldwide computing and data resources that must all be brought to bear on the data analysis problem of high-energy physics.

To meet these technical goals, priorities have to be set, the system has to be managed and monitored globally end-to-end, and a new mode of “human-Grid” interactions has to be developed and deployed so that the physicists, as well as the Grid system itself, can learn to operate optimally to maximize the workflow through the system. Developing an effective set of trade-offs between high levels of resource utilization and rapid turnaround time, plus matching resource usage profiles to the policy of each scientific collaboration over the long term, present new challenges (new in scale and complexity) for distributed systems.

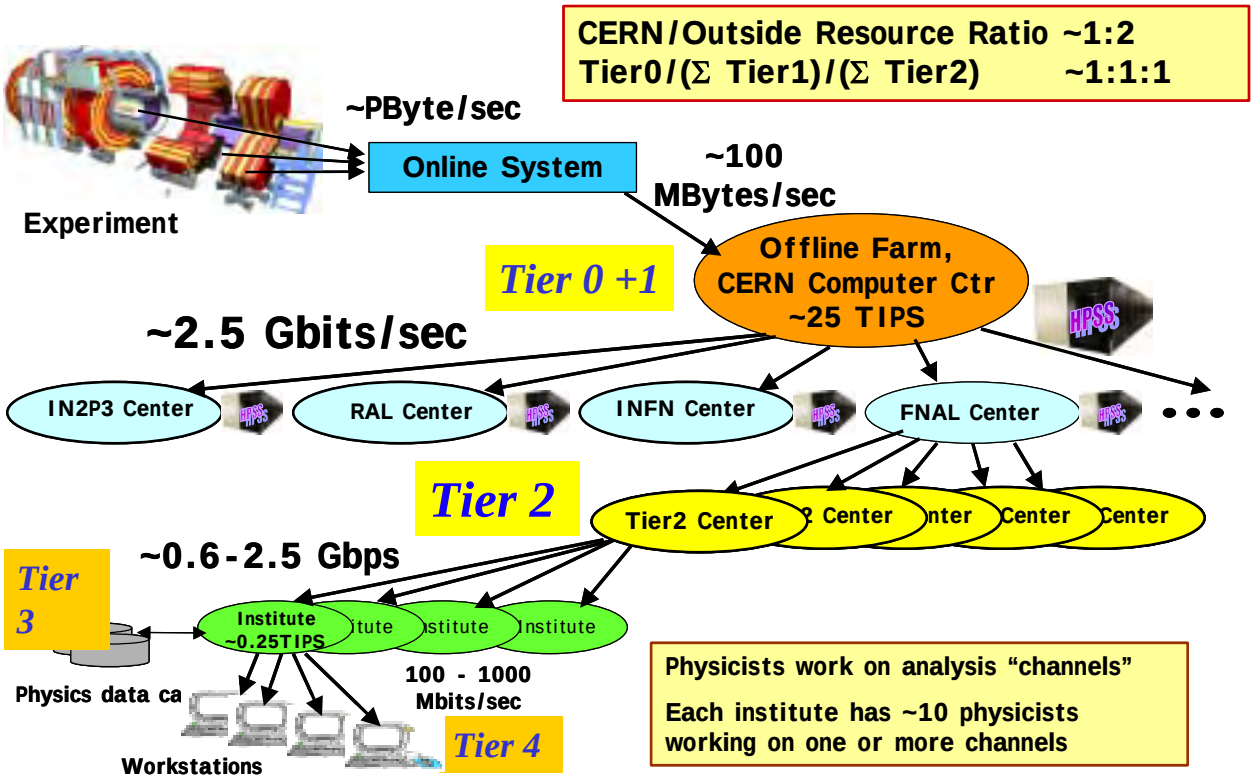


Figure 2.6. A Hierarchical Data Grid as Envisioned for the Compact Muon Solenoid Collaboration. The grid features generation, storage, computing, and network facilities, together with grid tools for scheduling, management, and security.

Table 2.3. High-Energy Physics Requirements Summary

Feature	Characteristics that Motivate Advanced Infrastructure	Vision for the Future Process of Science	Anticipated Requirements	
Time Frame			Networking	Middleware
Near-term	• Instrument based data sources • Hierarchical data repositories • Hundreds of analysis sites • 100 gigabytes of data extracted from a 100 terabyte data store and transmitted to the analysis site in 10 minutes in order not to destabilize the distributed processing system with too many outstanding data requests • Improved quality of videoconferencing capabilities • Cross-site authentication/authorization	• The ability to analyze the data that comes out of the current experiment • Remote collaborative experiment control	• gigabit/sec • end-to-end quality of service	• Secure access to world-wide resources • Data migration in response to usage patterns and network performance o naming and location transparency • Deadline scheduling for bulk transfers • Policy based scheduling / brokering for the ensemble of resources needed for a task • Automated planning and prediction to minimized time to complete task
5 years	• 100 terabytes of data extracted from a 100 petabyte data store and transmitted to the analysis site in 10 minutes in order not to destabilize the distributed processing system with too many outstanding data requests • Global collaboration • Compute and storage requirements will be satisfied by optimal use of all available resources	• Worldwide collaboration will cooperatively analyze data and contribute to a common knowledge base • Discovery of published (structured) data and its provenance	• 100 gigabit/sec o lambda based point-to-point for single high-bandwidth flows o capacity planning • Network monitoring	• Track world-wide resource usage patterns to maximize utilization • Direct network access to data management systems • Monitoring to enable optimized use of network, compute, and storage resources • Publish / subscribe and global discovery
5+ years	• 1000s of petabytes of data		• 1000 gigabit/sec	

2.7 Magnetic Fusion Energy Sciences Requirements

The long-term goal of magnetic fusion research is to develop a reliable energy system that is environmentally and economically sustainable. To achieve this goal, it has been necessary to develop the science of plasma physics, a field with close links to fluid mechanics, electromagnetism, and nonequilibrium

statistical mechanics. The highly collaborative nature of the Fusion Energy Sciences (FES) is due to a small number of unique experimental facilities and a computationally intensive theoretical program that are creating new and unique challenges for computer networking and middleware.

In the United States, experimental magnetic fusion research is centered at three large facilities (Alcator C–Mod [13], DIII–D [14], and NSTX [15]) with a present-day replacement value of over $1 billion. Magnetic fusion experiments at these facilities operate in a pulsed mode producing plasmas of up to 10 seconds duration every 10 to 20 minutes, with 25 to 35 pulses per day. For each plasma pulse, up to 10,000 separate measurements versus time are acquired, representing several hundreds of megabytes of data. Throughout the experimental session, hardware/software plasma control adjustments are debated and discussed amongst the experimental team and made as required by the experimental science. The experimental team is typically 20 to 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.

The computational emphasis in the experimental science area is to perform more, and more complex, data analysis between plasma pulses. Five years from now, analysis that is today performed overnight should be completed between pulses. It is anticipated that the data available between pulses will exceed the Gbyte level within the next five years. During an experimental day, anywhere from 5 to 10 remote sites can be participating. Datasets generated by these simulation codes will exceed the Tbyte level within the next three to five years. Additionally, these datasets will be analyzed in a manner analogous to experimental plasmas to which comprehensive comparisons will need to be made.

Enhanced visualization tools now being developed will allow this order of magnitude increase to be effectively used for decision making by the experimental team. Clearly, the movement of this quantity of data in a 15- to 20-minute time window to computational clusters, to data servers, and to visualization tools used by an experimental team distributed across the United States and the sharing of remote visualizations back into the control room will place a severe burden on present-day network technology.

In fusion, the need for real-time interactions among large experimental teams and the requirement for interactive visualization and processing of very large simulation datasets are particularly challenging. Some important components that will help to make this possible include easy-to-use and easy-to-manage user authentication and authorization framework, global directory and naming services, distributed computing services for queuing and monitoring, and network quality of service (QoS) in order to provide guaranteed bandwidth at particular times or with particular characteristics.

Table 2.4. Magnetic Fusion Energy Requirements Summary

Feature	Characteristics That Motivate Advanced Infrastructure	Vision for the Future Process of Science	Anticipated Requirements	
Time Frame			Network	Middleware
Near-term	• Each experiment only gets a few days per year - high productivity is critical • Experiment episodes ("shots") generate 200-500 Mbytes every 15 minutes, which has to be delivered to the remote analysis sites in two minutes in order to analyze before next shot • Highly collaborative experiment and analysis environment	• Real-time data access and analysis for experiment steering (the more that you can analyze between shots the more effective you can make the next shot) • Shared visualization capabilities		• PKI certificate authorities that enable strong authentication of the community members and the use of Grid security tools and services. • Directory services that can be used to provide the naming root and high-level (community-wide) indexing of shared, persistent data that transforms into community information and knowledge • Efficient means to sift through large data repositories to extract meaningful information from unstructured data.
5 years	• Gbytes generated by experiment every 15 minutes (time between shots) to be delivered in two minutes • Gbyte subsets of much larger simulation datasets to be delivered in two minutes for comparison with experiment • Simulation data scattered across United States • Transparent security • Global directory and naming services needed to anchor all of the distributed metadata • Support for "smooth" collaboration in a high-stress environments	• Real-time data analysis for experiment steering combined with simulation interaction = big productivity increase • Real-time visualization and interaction among collaborators across United States • Integrated simulation of the several distinct regions of the reactor will produce a much more realistic model of the fusion process	• Network bandwidth and data analysis computing capacity guarantees (quality of service) for inter-shot data analysis o 500 Mbits/sec for 20 seconds out of 15 minutes, guaranteed • 5 to 10 remote sites involved for data analysis and visualization	• Parallel network I/O between simulations, data archives, experiments, and visualization • High quality, 7x24 PKI identity authentication infrastructure • End-to-end quality of service and quality of service management • Secure/authenticated transport to ease access through firewalls • Reliable data transfer • Transient and transparent data replication for real-time reliability • Support for human collaboration tools
5+ years	• Simulations generate 100s of Tbytes • Next generation experiment: Burning Plasma	• Real-time remote operation of the experiment • Comprehensive integrated simulation	• Quality of service for network latency and reliability, and for co-scheduling computing resources	• Management functions for network quality of service that provides the request and access mechanisms for the experiment run time, periodic traffic noted above.

2.8 Chemical Sciences Requirements

The chemistry community is extensive and incorporates a wide range of experimental, computational, and theoretical approaches to the study of problems, including advanced, efficient engine design; cleanup of the environment in the ground, water, and atmosphere; the development of new green processes for the manufacture of products that improve the quality of life; and biochemistry for biotechnology applications including improving human health. The advanced computing infrastructure that is being developed will revolutionize the practice of chemistry by allowing us to link high-throughput experiments with the most advanced simulations.

To overcome current barriers to collaboration and knowledge transfer among researchers working at different scales, a number of enhancements must be made to the information technology infrastructure of the community:

- A collaboration infrastructure is required to enable real-time and asynchronous collaborative development of data and publication standards, formation and communication of interscale scientific collaborations, geographically distributed disciplinary collaboration, and project management.

- Advanced features of network middleware are needed to enable management of metadata, user-friendly work flow for web-enabled applications, high levels of security especially with respect to the integrity of the data with minimal barriers to new users, customizable notification, and web publication services.

- Repositories are required to store chemical sciences data and metadata in a way that preserves data integrity and enables web access to data and information across scales and disciplines.

- Either tools now used to generate and analyze data at each scale must be modified or new translation/ metadata tools must be created to enable the generation and storage of the required metadata in a format that allows interoperable workflow with other tools and web-based functions. These tools also must be made available for use by geographically distributed collaborators.

- New tools are required to search and query metadata in a timely fashion and to retrieve data across all scales, disciplines, and locations.

The advanced computing infrastructure that is being developed will revolutionize the practice of chemistry by allowing us to link high-throughput experiments with the most advanced simulations. Chemical simulations taking advantage of the soon-to-come petaflop architectures will enable us to guide the choice of expensive experiments and reliably extend the experimental data into other regimes of interest. The simulations will enable us to bridge the temporal and spatial scales from the molecular up to the macroscopic and to gain novel insights into the behavior of complex systems at the most fundamental level. For this to happen, we will need to have an integrated infrastructure including high-speed networks, vast amounts of data storage, new tools for data mining and visualization, modern problem-solving environments to enable a broad range of scientists to use these tools, and, of course, the highest-speed computers with software that runs efficiently on such architectures at the highest percentages of peak performance possible.

Table 2.5. Chemical Sciences Requirements Summary

Feature / Time Frame	Characteristics That Motivate Advanced Infrastructure	Vision for the Future Process of Science	Anticipated Requirements	
			Network	Middleware
Near-term	• High data-rate instruments running for long times producing large data sets • Greatly increased simulation resolution-data sets ~10–30 terabytes • Geographically separated resources (compute, viz, storage, instmts) & people • Numerical fidelity and repeatability • Cataloguing of data from a large number of instruments • Large scale quantum and molecular dynamics simulations	• Distributed multi-disciplinary collaboration • Remote instrument operation / steering • Remote visualization • Sharing of data and metadata using web-based data services • Computing on the net by linking large scale computers	• Robust connectivity • Reliable data transfer • High data-rate, reliable multicast • Quality of service • International interoperability for namespace, security • Large-scale data storage needed both for permanent and temporary data sets. Can the network serve as a large scale data cache?	• Collaboration infrastructure • Management of metadata • High data integrity • Global event services • Cross discipline repositories • Network caching • Server side data processing • Virtual production to improve traceability of data • Data Grid broker / planner • Cataloguing as a service
5 years	• 3D Simulation data sets 30–100 terabytes • Coupling of MPP quantum chemistry and molecular dynamics simulations for large scale simulations in chemistry, combustion, geochemistry, biochemistry, environmental studies, catalysis • Validation using large experimental data sets • Analysis of large scale experimental data sets including visualization and data mining	• Remote steering of simulation, e.g., control of the time step, convergence of the SCF, introducing a perturbation in an MD simulation • Remote data sub-setting, mining, and visualization • Shared data/ metadata with annotation evolves to knowledge base	• 10s of gigabits for collaborative visualization and mining of large data sets	• Remote I/O • Collaborative use of common, shared data sets – version control on the fly • International interoperability for collaboratory infrastructure, repositories, search, and notification • Archival publication
5+ years	• Accumulation of archived simulation feature data and simulation data sets • Multi-physics and soot simulation data sets ~1 petabyte • Large-scale MD simula-tions – 100s of terabyte to petabyte datasets	• Internationally collaborative knowledge base • Remote collaborative simulation steering, mining, visualization	• 100+ gigabit for distributed simulations – computational quantum chemistry, molecular dynamics, CFD combustion simulations	• Remote collaborative simulation steering, mining, visualization

2.9 Bioinformatics Requirements

The field of computational biology, in particular that of bioinformatics, has undergone explosive growth since the first gene-sequencing work emerged in the mid 1980s. Our understanding of biological processes, our ability to model them, and our ability to organize information and develop algorithms, also have progressed rapidly. The field is now transitioning to a stage where algorithmic progress has outpaced computing capabilities in terms of raw compute cycles, storage, and especially fast, secure, and usable information discovery and sharing techniques. These factors limit progress in the field.

Applications that dominate today's computing requirements in bioinformatics include genome sequence analysis, pairwise alignment, computational phylogenetics, coupling of multiple model levels to determine metabolic pathways, and secondary database searching. On the more distant research horizon, research areas include sequence-structure-function prediction, computation of the genotype-phenotype map [16], protein folding [17, 18], molecular computing [16], genetic algorithms [16], and artificial intelligence solutions that will require real-time harnessing of Grid resources for large-scale parallel computation.

Although the networking requirements of computational biology have much in common with other areas of computational science, they differ substantially in the aspects described in the remainder of this section. We note that some of these differences are of a quantitative nature, while others are qualitatively unique to the characteristics of the information bases and algorithms that make up the field.

The growth of the number of researchers involved in computational biology is outpacing that of almost any other biomedical science. This necessitates highly effective solutions to authentication and authorization for Grid access; policy-based control and sharing of Grid resources; and automated management of individual logins at large numbers of Grid sites. National and international research communities will also need to construct virtual organizations, resource allocation policies, and charging mechanisms that span Grid providers, because bioinformatics Grids have different funding sources (ranging from state funds in North Carolina and Michigan, to federal R&D programs, to foreign funds in the European Union and Japan).

Bioinformatics has a large component of symbolic data, which requires highly diverse data models, and makes far heavier use of large-scale relational databases than most other sciences. This necessitates high-quality end-to-end solutions for database integration and federation, an issue of data type and identifier standards, coupled search and analysis tools, and interoperable security rules and models.

While genomic databases of the past decade were sized in gigabytes, today's databases are pushing terabytes and growing roughly according to Moore's law—doubling approximately every 18 months [16], with petabyte applications well within view. Performing Grid computation on relational data will require the integration of heterogeneous databases to form worldwide federations of unprecedented scale. In addition, database replicas will need to be maintained accurately and synchronized with high integrity as huge amounts of data are exchanged. Significant research will be required in distributed database replication and Grid-wide database mining applications to meet the federation and performance requirements of bioinformatics.

One of the most important collaborative activities in bioinformatics today is that of annotation, which would be greatly enhanced by the integration of multiparty messaging technologies with database versioning techniques, possibly augmented by multicast with closely integrated file transport and visualization. This requires enhancements to network data transport protocols and QoS mechanisms. Collaborative imaging systems for use in the life sciences will involve both shared exploration and annotation of ultra-high-resolution images by multiple distant collaborators, coupled with computational intensive pattern recognition, that require real-time transport of large image data.

Chapter 3 Middleware Research Enabling Advanced Science

As illustrated in Chapter 2, DOE Office of Science laboratories operate a wide range of unique resources, from light sources to supercomputers and petabyte storage systems, intended to be used by a large distributed user community. The laboratories' geographically distributed staff frequently are faced with scientific and engineering problems of great complexity, requiring the creation and effective operation of large multidisciplinary teams. The problems to be addressed are large and challenging, often greatly exceeding the limits of traditional computing and information systems approaches.

These application requirements demand an *infrastructure for distributed science* capable of overcoming the obstacles that distance and distribution pose to scientists whose work requires access to remote information, resources, and people. This infrastructure must include not only a fast, functional network but also a broad spectrum of middleware services. In fact, *the vast majority of "network" requirements as expressed by DOE application groups are concerned with middleware rather than connectivity.*

Working from application requirements, we identify six high-priority areas in which middleware research, development, deployment, and support are required in order to enable DOE science. These areas, and their benefits, are as follows:

- *secure control over who does what*—This capability is a fundamental prerequisite for essentially any distributed science scenario, beyond the most basic "put public data on a web server." While certainly not a new problem, the challenging demands of DOE science applications and the distributed, multi-institutional nature of the DOE laboratory system leads to unique requirements.

- *information integration and access*—The ability to discover and access networked scientific information as well as information about information, or about other resources such as computers, storage, networks, code, services, instruments, and people is a second fundamental prerequisite. It enables, among other things, "data-mining-based" science. Again, not a new problem but one with some particularly challenging requirements.

- *coscheduling and quality of service*—The ability to coordinate multiple distributed resources (whether computers, storage systems, services, networks, or other assets) in order to provide the required level of performance guarantees is critical to a range of application scenarios, including coupling of experiments with computation (e.g., fusion), remote visualization, data analysis pipelines, and collaboration. Despite pioneering DOE research, this capability does not yet exist in any general sense.

- *effective network caching and computing*—Science scenarios often depend upon the ability to stage large quantities of data to intermediate locations and to obtain rapid access to computing for purposes such as data filtering or experiment decision-making. An attractive alternative to using dedicated supercomputers or other site resources would be to integrate into the network infrastructure storage and computing resources designed for on-demand use by network services. This concept is new within the science Grid context, and its realization will require substantial work.

- *services to support collaborative work*—Even a brief review of the science application requirements identified in Chapter 2 leads us to identify a need for a wide variety of "community services" designed to facilitate collaborative work. Although many such services exist, many others must be created. Their design, deployment, and operation raise challenging technical and policy issues.

- *monitoring and problem diagnosis*—Distributed resources cannot be used effectively if the reasons for failures cannot easily be diagnosed and corrected. Thus, an indispensable prerequisite for essentially all distributed science applications is end-to-end, top-to-bottom monitoring and diagnosis capabilities. DOE researchers have led the way in this area, but no comprehensive solution exists.

In this chapter, we identify requirements and priorities for this *network middleware*. First, we discuss briefly the nature of middleware and the rationale for including middleware in an infrastructure to support DOE distributed science. Then, building on and integrating across the application requirements identified in Chapter 2, we identify high-priority middleware requirements and, for each, specify the research, development, deployment, and support activities required to address these requirements within the DOE laboratory context.

The application descriptions listed above have made the case for distributed science. Here, we discuss the vital enabling role played by middleware.

3.1 Middleware Infrastructure for Distributed Science

The unique and varied characteristics and large information scale of the DOE scientific environment lead to demanding requirements for an *infrastructure for distributed science* capable of overcoming the obstacles that distance and distribution pose to scientists whose work requires access to remote information, resources, and people.

The foundation for such an infrastructure must necessarily be a fast, functional network. As discussed in Chapters 4 and 5, this network must provide high-performance links, Internet services, some sort of quality of service (QoS) support, and instrumentation. However, as discussed in Chapter 2, DOE science applications require much more than simple connectivity.

3.2 The Role of Middleware

The purpose of middleware is to translate the *potential* of fast, functional networks into *functionality that facilitates new science paradigms* by enabling easier, faster access to, and integration of, remote information, computers, software, and/or experimental devices—as well as interpersonal communication. It is middleware that makes it possible for an individual scientist or scientific community to address the six application requirements described above, by providing the services that enable

- making data, computers, software, and instruments available over the network in a controlled fashion, so they can be used by remote users without fear of damage or access by unauthorized users

- discovering available information resources and engaging in data-mining science based on the integration and synthesis of information from multiple sources

- integrating remote resources into local experimental and computational environments while meeting performance constraints, whether expressed in terms of time-to-completion (for a transfer or computation), frame rate (for video), or other metrics

- manipulating, analyzing, and visualizing datasets that are too large to hold in local storage

- managing, in a community setting, the authoring, publication, curation, and evolution of scientific data, products, programs, and associated computations

- diagnosing the cause of failures in distributed computations—or, even better, having those problems corrected before they become apparent to users.

3.2.1 What Is Middleware?

What exactly do we mean by "middleware"? This term is used to refer to many different technologies, ranging from network services (e.g., certificate authorities, reliable multicast) to component technologies (e.g., CORBA®). Although all of these technologies can play an important role in scientific computing, we focus our attention here on middleware components that are concerned with enabling distributed science and that can reasonably be considered infrastructure. In particular, we consider

- global services designed to support all users of a network regardless of discipline (e.g., root directory services, certificate authorities)

- community services designed to support members of specific communities (e.g., membership services)

- resource or site services deployed on a resource or at a site to enable the participation of that resource or site in the larger network (e.g., storage and compute system access services).

Although out of scope for this workshop, we note here the importance of the middleware components and application-specific tools required to build distributed applications.

3.2.2 Middleware and the End-to-End Problem

It is by now customary to talk about the end-to-end problem in networks, normally meaning *How do I achieve performance not just between two sites, but from application to application?* Addressing this problem requires careful attention to both the engineering and configuration of both site networks and applications.

In fact, while performance has, to date, received the most attention, the end-to-end problem is far larger and more significant than simple performance difficulties. Essentially every middleware capability

® CORBA is a registered trademark of the Object Management Group, Needham, Massachusetts.

required by DOE science applications has an end-to-end component to it, and associated requirements for site and application engineering and application. Examples of end-to-end requirements include the following:

- *security*—Authentication and authorization decisions have to be delivered end-to-end, taking into account—and mapping to—the identity, authentication, and authorization mechanisms used by participating sites.

- *policy*—Sites have to describe the policies and capabilities in a way that allows others to discover them and respond to them. Those same policies and capabilities may need to be adapted to meet requirements of distributed execution.

- *scheduling*—Application demands for coscheduling and end-to-end performance guarantees can require support for reservation, pre-emption, and other specialized scheduling support at sites.

- *transport*—In many DOE science applications, "transport" is not simply the movement of data from one computer to another (which can already involve challenging end-to-end issues) but rather involves the movement of data from one complex end-system device (e.g., scientific instrument, parallel file system) to another.

The impact of these issues on middleware research, development, deployment, and support is frequently underestimated.

3.3 Grid Middleware

The evolution of middleware and distributed systems in the scientific computing environment is currently embodied in the endeavor called computing and data Grids [3, 19-20]. The role of Grid middleware is to greatly simplify the construction and use of widely distributed and/or large-scale collaborative problem solving systems. Grid-managed resources are the geographically distributed, architecturally and administratively heterogeneous, computing, data, and instrument systems of the scientific milieu.

Grid middleware provides services for uniform access, management, control, monitoring, communication, and security to application developers using these distributed resources. The international group working on defining and standardizing Grid middleware is the Global Grid Forum (GGF [21]) that now consists of about 700 people from some 130 academic, scientific, and commercial organizations in about 30 countries. GGF involves both scientific and commercial computing interests. It also involves an evolving understanding of the issues that must be addressed in order to facilitate the expeditious construction of the complex distributed systems of science from a very dynamic pool of resources.

There is now enough experience in building Grids that the basic access and management functions noted above are fairly well understood, and reference implementations are available for most of these through the Globus toolkit [22]. However, as our experience with Grids grows, more issues arise that must be addressed to meet the goals of easily building effective distributed science systems.

To be effective, the Grid middleware must be deployed widely. This involves two things:

1. recognition on the part of the funding agencies that Grids represent an essential new aspect of the infrastructure of science and thus must be supported as persistent infrastructure

2. an educational process that addresses the critical sociological issues involved in changing operational procedures, intersite cooperation and sharing, homogenizing security policy, and other related issues. Many of these issues have been addressed in the building and operation of networks, and now must be addressed in the operation of computing, data storage, and instrumentation facilities.

The type of Grid middleware described thus far provides the essential and basic functions for resource access and management. As we deploy these services and gain experience with them, it has become clear that higher-level services also are required, to make effective use of distributed resources. One such higher-level service is the brokering functionality to automate building application-specific virtual systems from large pools of resources. Another example is collective scheduling of resources so that they may operate in a coordinated fashion. This is needed to allow a scientist to use a high-performance computing system to do real-time data analysis while interacting with experiments involving on-line instruments. It can also allow simulations from several different disciplines to be run concurrently, exchange data, and cooperate to complete a composite system simulation, as is increasingly needed to study complex physical and biological systems. Such services currently are being developed and/or designed.

Higher-level services also provide functionality that aids in componentizing and composing different software functions so that complex software systems may be built in a plug-and-play fashion. The current approach to these services leverages large industry efforts in web services based on extensible markup language (XML) to integrate web services and Grid services. This will allow the use of commercial and public domain tools such as web interface builders and problem-solving environment framework builders to build the complex application systems that provide the rich functionality needed for maximizing human productivity in the practice of science. Much work remains, but the potential payoff for science is considerable.

To provide the advanced infrastructure that will facilitate the next generation of science, Grids must be fully developed and widely deployed and combined with the next generation of ultra-scale computers, ultra-scale storage systems, and very-high-bandwidth networks to knit together the many physical resources.

3.4 Platform Services

Another aspect of the middleware requirements is the support that is needed on the resource platforms themselves.

Computing systems must have schedulers that enable coscheduling with other, independent resources. Data archive systems must have access servers that allow for reliable, high-speed, wide-area network data transfer. Networks must provide capabilities for QoS (usually in the form of bandwidth guarantees) that

let distributed resources communicate at high speeds during critical times in coupled simulation or on-line instrument data analysis. All of the storage, computing, and network resources must have support for the detailed monitoring that is essential for debugging, fault detection, and recovery in widely distributed systems.

These services must be developed, installed, and integrated into the operational environments of all of the individual systems that make up the resource pools of science.

3.5 Middleware Research Priorities

The application requirements developed in Chapter 2 provide a detailed statement of infrastructure requirements as defined by a set of important DOE applications. That material should be consulted for a comprehensive list of important middleware needs.

In this section, we synthesize from these requirements a set of *six priority areas* in which we believe work is required most urgently to advance DOE science. In brief, these areas—and their connections to DOE application requirements—are

- secure control over who does what
- information integration and access
- coscheduling and quality of service
- effective network caching and computing
- services to support collaborative work
- monitoring and problem diagnosis.

3.5.1 Secure Control over Who Does What

Fundamental to essentially every DOE science application requirement identified above is the need to be able to share and access remote resources. However, turning a group of scientific collaborators into a functioning virtual organization within which network, computing, and data resources can be shared and managed effectively is not a trivial task. Historically, each institution, and often each virtual organization, has its own mechanisms for establishing identity, for determining authorization, for enforcing policy, and so forth. Comprehensive middleware solutions are required for mapping between different treatments of identity, authentication, authorization, policy, accounting, auditing, and other functions, so that individual users, communities, and sites can individually and collectively establish policies, negotiate access, monitor activities, and in general engage effectively in their desired tasks, without compromising site or application security.

At a lower level, many scientific collaborations confront the problem of access limitations caused by firewalls and network address translation (NAT) units that exist at the network boundaries of our laboratories. These firewalls may not only prohibit some legitimate interactions, but also introduce serious performance problems for allowed interactions. Research is urgently needed on the construction of a new generation of institutional protection that is dynamically driven by well-vetted organizational policy rather than being enforced via blanket restrictions based on Internet addresses.

These problems are not unique to DOE or to science; they are, for example, fundamental to many electronic-commerce (e-commerce) scenarios. However, the particular combinations of sophisticated applications, high performance, and dynamic collaboration scenarios make DOE requirements especially demanding. DOE scientists have taken a lead in recent years in the development of Grid security solutions. Hence, it is important that this lead now be translated into deployable security solutions able to support the full range of DOE science activities. Achieving this goal will require not only substantial further research and development but also investment in site deployment and work aimed at defining broadly acceptable site security policies.

3.5.2 Information Integration and Access

The problem of sharing access to information is common to essentially all DOE science applications. In general, the problem becomes increasingly difficult as collaborations become larger and more loosely coupled, which is the trend in most DOE science areas. In fact, in some large communities (e.g., biology) the problem of discovering what data and other resources are known to the community can be one of the biggest obstacles to scientific progress.

The challenge, therefore, is to provide middleware services that make it possible for one group to easily discover and use relevant scientific results generated by another, or for a user to learn how a particular piece of scientific data was generated and how can it be computationally or experimentally reproduced. This implies that shared data needs to be well described by metadata that provides not only a description of the data's contents but also its provenance. Middleware must be provided that can help locate the services that allow users to publish metadata. And, just as DNS allows us to discover the binding of a network host name to an Internet protocol (IP) address and the Google search engine allows us to search for web links based on hyper text markup language (HTML) text references, scientists need access to services that discover, catalog, and mine scientific metadata.

Accessing the actual data objects is also a serious problem. A global name space is required to make it possible to provide a handle that can be passed from one place on the network to another. Registries and other information services must be deployed to support discovery of global names and the binding of names to object metadata. Because a global network file system is not a scalable solution, one needs other mechanisms to resolve global names into access mechanisms to reach very large data objects. Once located, advanced file transfer protocols are needed that can move large files without data corruption. As described in Section 3.5.4, this is related also to network caching, where middleware services insulate the application from the complexities of selecting caches of data replicas and staging the movement of information from one cache to another.

3.5.3 Coscheduling and Quality of Service

Several of the science applications need to coschedule computations at different points in the network. For example, computational preprocessing at a data source may be required to deliver a stream of data used as an input to a remote simulation. If the data source is an instrument, there may also be real-time constraints imposed on the scheduled post-processing or filtering of the output. This class of distributed computation requires a special class of network middleware services and brokers able to reserve key resources, whether computers, storage systems, or network resources. In some cases, the reservation

requirements are fixed—the data source must start at a specified time. In other cases, the experiment will take place within a specified window, but the exact time may not be known. In still other cases, the required QoS is expressed in terms of a required completion time for a transfer or computation.

Meeting end-to-end application QoS requirements requires end-to-end middleware support for performance monitoring and tuning. QoS for network bandwidth and computational resources should be expressed in terms of high likelihood of worst-case performance; for example, with 95% likelihood, I am assured of at least 10Gbps bandwidth, 1 TB storage at location X, and 100 Tflops of computing on computers at Y and Z at any time on date D. In all but the simplest cases, the current state of the art requires that coscheduling involve the active participation of humans at every stage of the resource negotiation process. A great deal of experimentation and research, as well as work on service deployment, will be needed to truly automate these processes. Achieving this may require more than just automated allocation of network bandwidth. Middleware services may need to monitor bulk data transfers and increase the service level if it looks like the transfer will not finish in time. For example, a researcher knows that a large dataset needs to be copied to a given location by 8:00 a.m. the next day. Middleware could monitor the transfer, starting out with best-effort service, and then increase the service level as the deadline approached. *The delivery of such capabilities in a production setting is vital to DOE science, and its successful realization will represent a major advance in the state of the art.*

3.5.4 Network Caching and Computing

A common crosscutting DOE science application requirement is for on-demand access to storage and/or computing. Storage may be required for the staging of a large dataset near its eventual user(s), for purposes of reducing network bandwidth demands via caching or as an impedance-matching mechanism when the user requires high-speed interactive access, as in visualization.

When access to large datasets also involves large-scale computation, it often is critical to minimize the data access time via caching data locally. Caching is also a basis for specialized services such as reliable multicast or format conversion in collaboration environments. Caching and computing can be used together to cache the results of intermediate computations for subsequent reuse or to repeat a computation locally when access to previous data is not available. Data caches can also be used to facilitate large remote file transfers, by providing a reliable staging point intermediate between a source and the eventual destination.

The operating envelope of most existing DOE systems may not be well suited to the on-demand access required for the scenarios laid out in Section 3.5.3. Thus, we believe that it is important to explore new approaches based on the integration “into the network” of storage and computing resources that can be managed by network services and allocated in an on-demand fashion to network applications.

This concept is not entirely new, of course; it is fundamental to web caches and distributed content distribution infrastructures operated by e-business infrastructure service companies. However, it is clear that the peculiar demands of DOE science applications will require new approaches to configuration, operation, and use, for example, in the science Grid context. *Research and development on these new concepts represents a major opportunity for DOE computer scientists to contribute to both DOE science and the state of the art in Grid computing.*

3.5.5 Services to Support Collaboration

Many DOE science collaboration scenarios depend on the existence of persistent community services, ranging from e-mail list servers, to web servers, code repositories, persistent data archives, authoring services, identity certificate authorities, file servers, replica managers, registries of various sorts, application servers, interaction spaces, and portals. In some cases, the establishment of these services is primarily a policy question of whether it is most cost-effective to centralize their operation (e.g., within ESnet or some comparable entity). In other cases, the technologies required to achieve secure, robust operation of a service do not yet exist.

Collaboration support frequently requires support for many services that must span multiple campuses and autonomous systems. For example, Access Grid technology, which has become an essential part of DOE collaboration, requires administrative support and care for the underlying multicast support at the router level. Future middleware will need to provide scaleable, easy-to-use primitives to support multiple modes of secure, wide-area collaboration. It will not be possible to support every new type of collaboration tool if each requires the network to employ a different underlying synchronization or security or distribution model. Middleware collaboration primitives need to be accessible to the application design community and easily supported by all the stakeholders who must manage the resources and networks.

3.5.6 End-to-End Monitoring and Diagnosis

Although not explicitly stated by many science application groups, an essential requirement for any substantial progress in distributed science is technologies and tools capable of diagnosing problems that arise in a distributed setting. Debugging and tuning widely distributed applications is substantially more difficult than applications on a single system or even a local area network. End-to-end application behavior in distributed applications is affected by a host of unpredictable and difficult to reproduce network-related faults and synchronization anomalies. Hence, the optimization of network-wide resource usage of distributed scientific applications requires specialized tools and careful design.

Top-to-bottom monitoring and diagnosis is required if users (or, better still, automated tools) are to identify the source of problems in distributed applications. This requires both new middleware and new approaches to data collection described in Chapter 4. Middleware research is required with the goal of automating the process of application instrumentation so that performance faults can be easily diagnosed and fixed. To accomplish this task, we must engage the network research community and middleware developers to provide tools and deploy the services that application programmers can easily use. This will require a substantially greater involvement of all three groups in understanding end-to-end performance and application fault tolerance. *Success in this area will represent a significant contribution to knowledge and advance over the current state of the art.*

Chapter 4 Network Research Enabling Advanced Science

In more than 25 years of DOE networks for science research, both science and networks have changed dramatically. Over that time, the routine objects transported by ESnet have grown from kilobytes to terabytes, and the range of science activities supported by the network has grown to include many nationwide and worldwide resources and with large, close-knit collaborations. The underlying communications technologies of the network infrastructure have undergone a number of technological revolutions, from copper telephone lines to multiple light paths on fiber optics, from switched circuits to packet switching and soon to allocation of wavelengths. Handling these massive changes in scale, function, and capability has depended upon the fruits of network research, where DOE has made a number of significant contributions.

Today, we are at the brink of another massive change in high-performance network capabilities and science applications. Achieving the best performance in a dependable, predictable, and secure manner is a major challenge. Simply adapting or scaling up today's network protocols and management software is not an option; they cannot do the job. Providing the network performance, allocation, management, and security capabilities required by science applications (and other high-performance network uses) requires advances in both middleware and network services, fully informed by science application needs.

In terms of network functionality, analyzing the applications indicates that there is a clear progression from needing more bandwidth, to needing robust bandwidth, to needing the ability to manage data within the network and even transform that data in the network [23]. This progression (illustrated in Figure 4.1) is evident—although on different timelines—in all of the science scenarios summarized in Chapter 2 and detailed at http://www.DOECollaboratory.pnl.gov/meetings/hpnpw. What we see in the near term is a need for more bandwidth. This is followed by a need for network quality of service, usually bandwidth guarantees in which operating experiments are connected to large-scale computing and data, and sometimes for bounded message latency to do real-time remote control. In the five-year time frame, several science application areas indicated they expect modeling will be sufficiently advanced that it will not be practical to generate and store all possible simulations, and that virtual data catalogues should provide such data on demand.

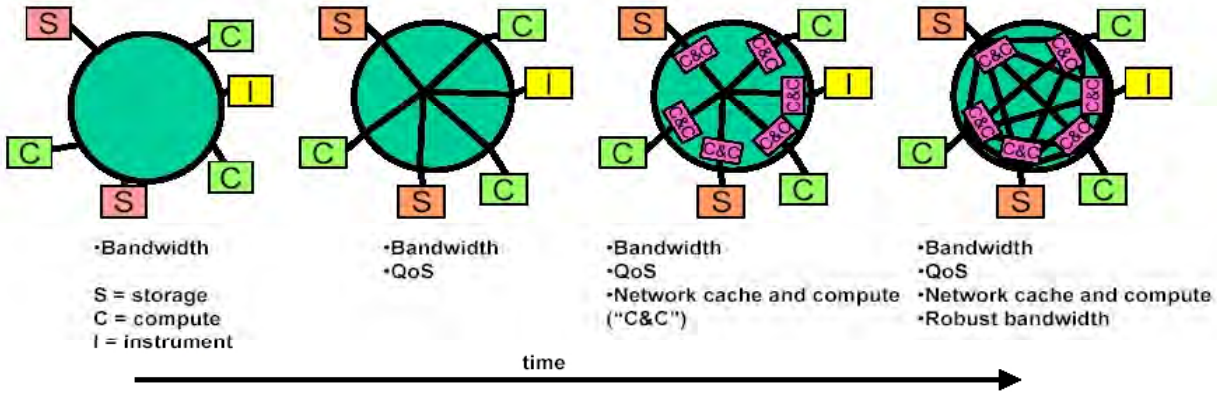


Figure 4.1. Evolution of Network Services Requirements over Time

This, together with massive data analysis scenarios of, for example, high-energy physics, that involve moving large quantities of data to many remote sites, leads to the requirement for "network" caches (these appear as the first **C** in the **C&C** box of Figure 4.1. These caches may literally be in the network itself (already some of the large commercial telecommunication carriers are investigating this as a potential service with the disk nodes located in their central offices and points of presence) in a general science service version of the sort of web caching provided by the Akamai Content Delivery Service.

The second **C** refers to a need for computing in the network. Even with very high-speed networks, there will always be circumstances when data should be filtered or reduced before it leaves the systems or site of the storage system. These computing nodes are probably not general service but instead provide a fairly stylized service to transform data streams (see, for example, DataCutter [24]).

Finally, as the scientific community inevitably reaches the point where their science is completely reliant on an integrated computing-middleware-network infrastructure, there is a clear requirement for redundancy in the communication paths. This requirement is indicated schematically by the all-to-all connectivity of the right-most icon in Figure 4.1. It is primarily a network engineering requirement for the communication service providers but marks a clear transition to a fully integrated environment.

Although much work is under way in the commercial sector as well as in National Science Foundation-funded programs, workshop participants felt that DOE has some unique requirements for future networks. Only DOE has the very large distributed applications that generate many petabytes of data, and only DOE has collaborations that include hundreds of researchers at tens of sites.

From the point of view of the science application user, it is vital that "the network" is fast, dependable, predictable, and secure. Under those simply stated network requirements are a host of challenging research issues. We have identified a set of research topics required to enable the types of middleware described in the previous chapter. These research topics are discussed in detail in the following sections.

- *ubiquitous monitoring and measurement infrastructure*—Much of the middleware described in the previous chapter depends on some understanding of the underlying network to base its decisions on what to do next. For example, current network conditions may determine where and when to use network data caches. Therefore, networks must be monitored, and the results of this monitoring must be published in a format that is understandable by the middleware.

- *high-performance transport protocols*—The current predominant data transport protocol, Transmission Control Protocol (TCP), has some well-known performance limitations. DOE applications will require optimally efficient use of the networks. Therefore, research into both improving TCP and looking into new protocols is necessary.

- *multicast*—Large, distributed collaborative projects are becoming increasingly common within the DOE scientific community, and Access Grid-like technologies will become essential. However, the Access Grid depends on IP multicast, which has proven to be an incredibly fragile technology. Research is needed into mechanisms to make IP multicast more robust.

- *guaranteed performance and delivery*—Some DOE applications have strict performance requirements and demand some form of deployable network QoS. Research is needed to determine what network service model will satisfy the needs of DOE scientists and will work across a wide variety of sites and networks. There is a trade-off between predictability and reliability, and new approaches to network management will be needed to provide this. For example, emerging technology that allows 0.5-nanosecond optical switching invites new approaches to this problem, like dynamic allocation of wavelengths.

- *intrusion detection*—Intrusion detection, as a primary line of defense in environments that are open enough to support large-scale science collaboration, is an important component of any network monitoring infrastructure today. The main unsolved problem of the intrusion detection world is predictive analysis. In other words, based on what happened in the recent past, one can get an indication and warning of what attack is about to occur.

- *distributed systems versus firewalls*—The problem of vetting traffic through firewalls is increasingly difficult because of the increase in user-application traffic, much of which is encrypted. Research is needed into mechanisms to integrate Grid security middleware with firewalls, so that the firewall can efficiently allow the transmission of authorized streams.

The research environment needed to carry out key aspects of this work requires isolated testbed networks for running controlled experiments.

In addition, it is important to address the communications gap that exists between network engineers and application and middleware developers. To design and build successful science applications of the type described in the science scenarios (http://DOECollaboratory.pnl.gov/meetings/hpnpw/finalreport/) science software developers need more information from network engineers to better understand what is feasible on the network. Conversely, network engineers need more information from the applications and middleware community as to what is required by the applications. Hence, establishing the forums for these discussions is a vital element of next-generation network research programs.

4.1 Network Research Priorities

4.1.1 Network Monitoring, Measurement and Analysis

Detailed network monitoring information is required by much of the middleware described in Chapter 3. Network performance characteristics must be monitored, and the results of this monitoring must be published in a format that is understandable by the middleware.

As an example of how network measurements would be used in a Grid environment, we use the case of a Grid file transfer service. Assume that a *Grid Scheduler* determines that a copy of a given file needs to be copied to site A before a job can be run. Several copies of this file are registered in a *Data Grid Replica Catalogue*, so there is a choice of location from which to copy the file. The Grid Scheduler needs to determine the optimal method to create this new file copy and to estimate how long this file creation will take. To make this selection, the scheduler must have the ability to answer these questions:

- What is the best source (or sources) from which to copy the data?
- Should parallel streams be used and, if so, how many?
- Which TCP window and buffer size should be used?

Selecting the best source from which to copy the data requires a prediction of future end-to-end path characteristics between the destination and each possible source. Accurate prediction of the performance obtainable from each source requires measurement of available bandwidth (both end-to-end and hop-by-hop), latency, loss, and other characteristics important to file transfer performance.

Determining whether there would be an advantage in splitting up the copy and, for example, copying the first half of the file from site B while in parallel copying the second half of the file from site C requires hop-by-hop link availability information for each network path. If the bottleneck hop is a hop that is shared by multiple paths, then there is no advantage to splitting up the file copy in this way.

Parallel data streams will usually increase the total throughput on uncongested paths. However, on congested links, using parallel streams may just make the problem worse. Therefore, measurements such as delay and loss are needed to determine how many parallel streams to use.

For this reason, one of the most important new services the network should provide is a service that indicates how much bandwidth is available at a given point in time between any specified set of end points. With this information, applications and middleware would have the ability to adapt to current and future network conditions. Ideally, this network service would provide both end-to-end and hop-by-hop information and would include information on network capacity, available bandwidth, delay, loss, and jitter. A mechanism for layer 2 network topology discovery also would be useful for network engineers to better understand and debug network problems and for middleware services to efficiently utilize the network.

However, simply scaling up today's approaches will not suffice. A monitoring and measurement infrastructure is needed to avoid too much measurement traffic. Although some monitoring can be done passively, other information can be collected only by using active probes—but too much active probing is intrusive to the network. In certain cases, a short real-time probe may be required. However, in other cases, the best solution would be to look up measurement data in a distributed measurement database, similar to the way that hostnames are resolved using the domain name server.

There are many open research issues in designing such a monitoring system. One of the hardest problems is to separate network issues from host and application issues. It is also difficult to separate physical layer issues from protocol layer issues in the network.

A large amount of passive monitoring information now is collected from ESnet routers using the Simple Network Management Protocol (SNMP) but not published like Internet2, which makes this information available (see [25-26]). Making this network performance information available via a middleware API (e.g., using web services and a SOAP API) would provide applications and middleware developers insight into the network behavior. Archiving measurement data also is important. Doing so allows one to compare current and previous performance and to determine what has changed.

To make this network measurement data more accessible, work also is required to enhance or replace the SNMP. Shortcomings of the SNMP include the lack of privacy, authentication, and access control, which limit the protocol's usefulness across domain boundaries. A reliable transport option also is needed.

A network bandwidth prediction service is essential, too. The Network Weather Service at the University of California, Santa Barbara [27] is being used by some groups to make short-term predictions (e.g., what will the network be like in 5 minutes). However, DOE science applications will require the ability to know what to expect from the network farther into the future, say at 2:00 p.m. tomorrow when a particular experiment will be running. This requires a system that can do long-term forecasts based on both historical data and anticipated usage. Grid scheduling systems and any large users of the network will need to notify this service of planned network usage.

4.1.2 High-Performance Transport Protocols

Most distributed applications use TCP as a transport protocol. While a large amount of work has gone into TCP over the years, there is general consensus that there are still some well-known performance limitations using TCP over high-speed, high-latency networks which severely limit the performance of many large-scale science applications. DOE applications will push the limits of the network, and so we must try to ensure the protocols used are as high-performance and efficient as possible.

Transmission Control Protocol Issues

A large number of transport protocol issues have been identified related to bulk data transport over high-speed networks. These issues will become critical as transport of multi-terabyte to petabyte datasets becomes widespread. TCP uses what is called an *additive increase multiplicative decrease* (AIMD) algorithm to respond to network loss. This algorithm assumes that a loss is due to congestion and backs off transmission. However, it has been shown that much of the loss experienced is not due to congestion, so the AIMD parameters are far too conservative for high-speed networks. Work is currently being done to correct this (see [28]); however, meeting the requirements of science applications for transport protocols in the 10-100 gigabits/s range will require additional research

In the interim, while a truly high-performance protocol is developed, there are some straightforward changes to TCP that would buy some time.

- Maximum Transmission Unit (MTU) size has not changed in 30 years. This means that the protocol has to work much harder, and that the control feedback loop is effectively much smaller than it was 30 years ago when the size was specified. Much larger MTUs could make a huge difference in TCP performance.

- Currently the TCP checksum is only 16 bits. This could result in corrupt data going undetected on very large transfers. A 32-bit checksum has been proposed, but issues exist about how to deploy this.

- The 32-bit sequence number is inadequate for very large transfers.

Some experts advocate that TCP should not be used by applications with large bulk data transfer requirements, and that a reliable User Datagram Protocol (UDP)-based protocol should be used instead. However, there is consensus in the network community that, while new protocols research should be done, TCP will certainly be with us for a long time, and TCP performance can be improved dramatically by adapting its congestion control algorithms for this type of environment.

Alternative Protocol Investigations

Alternative protocols also are worth exploring to determine if they meet the needs of the DOE community. These include Stream Control Transmission Protocol (SCTP, see [29]) and Scheduled Transfer (ST, see [30]). A more radical new protocol worth considering is the eXplicit Control Protocol (XCP, see [31-32]). XCP rapidly converges on the optimal congestion window using a completely new router paradigm. This makes it very difficult to deploy and test this new protocol on a large scale, because all new routers are required.

Host Performance Issues

Throughput of any protocol is affected by a number of host issues, and the host system plays a critical role in end-to-end performance measured at the application layer. Potential research areas to reduce host congestion include

- host system architecture for network-intensive applications
- very high speed network interface cards
- operating system bypasses to reduce operating system network activities
- a very efficient transport protocol stack to increase the throughput to applications
- high-speed system input/output
- network-aware operating systems
- middleware and APIs that efficiently couple applications to the network.

4.1.3 Multicast

Multicast support has been identified as an important capability for DOE science. Many projects are distributed and want to use multipoint distribution tools for collaboration (e.g., Access Grid technology to hold meetings). However, problems with traditional Internet Protocol (IP) multicast make it difficult to deploy and support. IP multicast has scaling properties that are different from the Internet in general, and, as the Internet grows, it gets progressively harder to make IP multicast work. Therefore, we suggest that in addition to the current solution, Any-Source Multicast (ASM) alternative service models should be explored, including Unicast relay (i.e., VRVS, IRQ, IM), peer-to-peer technology, and overlay networks.

4.1.4 Advanced Service Models

Another very important network topic to DOE science is network QoS. As mentioned in the preceding chapters, applications require bandwidth guarantees and sometimes bounded message latency in order to do real-time remote control. For example, computational preprocessing at a data source may be required

to deliver a stream of data used as an input to a remote simulation. If the data source is an instrument, there may also be real-time constraints imposed on the scheduled post-processing or filtering of the output.

In the past several years, a great deal of work has gone into a QoS model based on DIFFSERV. However, we feel that the DIFFSERV model will not solve the needs of DOE scientists, as it is extremely difficult to make it work in an interdomain environment (as is typically the case with laboratory-university or U.S.-international interactions), and it typically does not address the local area networks at a site. Applications need to have QoS end-to-end, not just on the wide-area network. End-to-end includes hosts, disks, and local-area networks. Guaranteed QoS on a host or disk can conceivably be achieved by just denying access to other users of the host during critical time; however, LAN issues are much harder to address.

Instead of the DIFFSERV model, researchers need to be able to frame their request for network services within a more concise service-level framework. Therefore, we suggest research into the following as an alternative to bandwidth partitioning:

- on-demand reconfiguration of network paths (i.e., MPLS/TE, lambda switching)
- research into control, management and measurement of switched paths
- active queue management
- network resource and capability discovery tools that operate securely at high speed.

4.1.5 Intrusion Detection

Intrusion detection is a key component of any network monitoring infrastructure. Intrusion detection systems scan packets for known malicious patterns of behavior and then block connections from that host. The best intrusion detection systems today, such as Bro [33], use a fiber tap to duplicate traffic on a fiber and send it to a monitoring host for real-time analysis.

The main unsolved problem in intrusion detection is predictive analysis. In other words, based on what happened in the recent past, can one get an indication and warning of what attack is about to occur? Other issues include how to scan packets at speeds greater than 1000 Mbits/sec, including approaches to parallel packet processing, and working with router vendors to be able to perform some types of filtering directly in the router. One approach is to put the fiber tap functionality into the router. For example, the router could be configured to send only non-file transfer protocol (FTP) and non-secure shell (SSH) traffic to the scanning host.

4.1.6 High-Speed Firewall Systems

The problem of vetting traffic through firewalls is increasingly hard because of the increase in user application traffic, much of which is encrypted. Expansion of the site authorization and authentication infrastructure and its integration with firewalls has the possibility of vetting data streams based on the rights of the entity initiating the stream rather than stream content. Grid middleware uses Transport Layer Security (TLS), Secure Sockets Layer (SSL) secure transport protocol; the information needed to

authenticate and authorize can be presented at the firewall. This would both simplify the task of firewalls and allow users much more freedom of access when that access is authorized.

4.2 Network Testbeds

In many cases, network research issues require an isolated network for running controlled experiments. Researchers need to be able to test new protocols, QoS mechanisms, and middleware without having to worry about affecting the production network. Testbeds serve as an environment to develop and test technology for the next-generation production networks. For more on why testbeds are important, see *A Vision for DOE Scientific Networking Driven by High Impact Science* [34].

Chapter 5 Road Map for Production, Testbed, and Research and Development Network Infrastructure

The network requirements of the Office of Science range from routine to extremely demanding and complex. A production network (ESnet) has traditionally provided the bulk of network services needed in the DOE science community. Increasingly, however, science activities such as ultra-high-speed data transfers, advanced visualization, and remote steering are demanding advanced networking capabilities that cannot be cost-effectively supported on a production network. The programs of the Office of Science would benefit from the formation of an integrated network provisioning model with three key elements—production-level networking addressing traditional program requirements, advanced networking to support high-impact DOE science applications, and easily separable experimental networking for research and development of advanced services and capabilities to meet future needs.

Successful integration requires

1. a road map that expresses the future of all three elements in the context of a networking vision shared across all DOE Mathematics, Information, and Computational Sciences (MICS) programs—Currently, a coherent overall set of architecture-level requirements for the network environment does not exist.

2. the fostering of a federal DOE Networking Initiative that provides funding to carry out the road map, using the SciDAC Initiative as a model, for example

3. a governance model that allows

    a. for the planning, management, resource allocation, and support across its elements in the context of the integrated program, with particular emphasis on resources needed for integrating these efforts across the elements

    b. for the participation of the DOE Office of Science programs in the planning and prioritization of network offerings with sufficient regard for the mission of all three elements.

All of the above requires a high-priority examination of possible business models, in which the overall goal is to provide a flexible and dynamic network infrastructure for all three elements.

These requirements are discussed in this chapter.

5.1 Three-Element Network Provisioning Model

The three-element network provisioning model would consist of

1. *production-level networking* in support of traditional program requirements—This element provides the capabilities and capacity required by existing DOE applications and research teams.

2. *network resources for high-impact DOE science programs*, including science application and Grid research, especially in areas that require capability networking or advanced services—The nature of the research served by this element might include distributed large-scale experiments, a distributed high-performance computing environment, or application/tools/middleware development requirements that cannot be satisfied by the bandwidth or services of the production-level networking element.

3. *network resources for networking research* that are easily separable in support of needed networking research.

5.1.1 Observations

An integrated network provisioning strategy would benefit from planning, coordination, funding, and implementation that encompass all three elements—a process in which each of these elements generates a requirement for a suite of needed capabilities and enabling networking services that must be met by one or more network providers.

An integrated strategy also could help overcome a number of challenges:

- Both technical and resource-related barriers to migration exist at the boundaries between each of the elements. For example, how do we migrate an application and supporting services from Element 2 to Element 1? This is a complicated question in the current environment because the roles and responsibilities associated with addressing the technical and support challenges required to move development efforts into full production support are not clear.

- Advancements moved from development into production require one-time funding for the migration, and also ongoing funding to support their production. However, no program or funding model exists today to permit this migration.

- Moving funding across the boundaries of the elements is currently somewhat difficult, thereby making it difficult for the networking requirements of one element to be satisfied by the provisioning of another element.

- A shared vision of success must be motivated, where some measures of success extend across all three elements.

5.1.2 Findings

A combination of network visionaries, managers, and technologists—some of whose perspective is based on long-term interactions with the MICS suite of network-related programs and others familiar with a breadth of agency and university networking programs—found the following with respect to the three-element model:

- A shared networking provisioning effort would benefit the MICS networking programs elements by motivating increased interaction among the elements, improved responsiveness of each element to

the other elements, identification of potential integration of efforts currently contained with an individual element, and development of a shared focus for all elements. This ideal result would be a high-level strategy that guides how these programs are integrated into a coherent road map, and that would motivate changes in governance.

- Funding is needed to acquire, deploy, and operate the additional network resources needed for high-impact DOE science programs and for network research.

- Specific funding is needed to bridge the elements—to help move R&D and advanced applications and technologies into common usage (infrastructure) in support of science.

- Additional funding is needed to move prioritization decisions currently based on "either/or thinking" toward allowing alternative approaches to be investigated across all three elements. Networking technologies and approaches will continue to change radically, and the networking program must position itself to be agile and not too firmly rooted in any one networking provisioning model.

- Additional funding is needed to support the resulting growth of production services that have migrated into production from the R&D community. This is the cost of success.

- As Grids and applications are moved into production, the model of providing service, inherent in supporting a growing production infrastructure, needs to be revisited. That is, as time goes by, more and more services will move into production, and the service model will need to become increasingly distributed as more production services feature end-to-end support.

- The business model may affect the manner in which the network is provisioned to the program elements.

The articulation of three sets of capability requirements has traditionally suggested three networks or three types of networks, and that approach may be overly constraining. One can imagine (but not yet safely forecast) that in a flexible network infrastructure (e.g., lambda-based), all three elements could receive their networking infrastructure from a common resource that can be reallocated easily. This should be investigated as part of the business model effort.

5.2 Business Models

Just as the multiple elements of the DOE networking strategy are driven by different sets of requirements, and just as they require different management approaches, the implementation of these capabilities may require different business models. The selection of a business model (or multiple models) involves a complex set of trade-offs among factors including (but not limited to)

- specific services required by the target customers
- available commercial services and opportunities
- size and scope of the network (number of sites, distances involved, size of user community)
- time frame in which the network must be deployed, and over which it will be operated.

In this section, we first look at three examples roughly corresponding to the three classes of networks ("elements") outlined earlier as part of the DOE networking strategy. Next, we summarize options for moving forward toward understanding and recommending business models to provide DOE networking infrastructure.

5.2.1 Three Types of Infrastructure, Three Business Models

To illustrate the interrelationship of various factors and the network business plans, we consider three examples, one from each of the three general classes of infrastructure in the DOE strategy:

- a "production" Internet service—DOE ESnet production services network [35]
- a special-purpose applications- and middleware-oriented network with capabilities not easily provided through production Internet services—NSF TeraGrid advanced applications network [36]
- a flexible network infrastructure intended to support multiple networking, applications, and middleware research projects—State of Illinois I-WIRE optical network infrastructure [37].

The business models for these networks are contrasted in Table 5.1, which illustrates a number of the factors that influence the business model.

We discuss two decision axes with respect to business models for providing network services and infrastructure. The first axis relates to what portions of the service or infrastructure are provided in-house and what portions are provided by outsourcing. The second axis relates to the type of financial and contractual vehicles (contracted services, leased facilities, purchased assets) used to obtain those portions of the infrastructure that are outsourced and the time horizons used to evaluate various options (one-year cost year-by-year, costs of n-years, and so on).

Table 5.2 shows a simplified layered view of the infrastructure required to provide networking (and middleware) services for scientific computing enterprises such as those that support DOE science. For comparison, the information in Table 5.2 also illustrates the way in which the three types of networks have approached this layering from the standpoint of their business models.

A central factor of these business models is the set of services typically provided to the network's target audience. Related to this is the domain of responsibility of the network provider. Above this region are services that users provide for themselves (although the network may provide some value-added services above this layer). Below this region are services that are contracted out in some fashion. Note that other service entry points also may be arranged in some cases, but the *primary* entry point (shown in Table 5.2) is what most customers use. The shaded layer in each case represents the layer that is generally opaque to the user community and the layer that is the *primary* service provided by the network.

Table 5.1. Comparison of Three Types of Networks

Network	Number of Sites	Geography	Reliability Requirements	Size of User Community	Deployment Time Frame	Business Model
ESnet	Dozens	National	99.9%	10s of thousands	12-18 months	Multiyear service contract
TeraGrid	5-10	National	99%	Thousands	12-18 months	Multiyear leased wavelengths
I-WIRE	10	Regional	99%	Thousands	36 months	Purchased fiber and equipment, managed in-house

Table 5.2. Cost Factors for Major Network and Middleware Infrastructure Layers (left) and Business Models of Three Example Types of Networks (right). Bold-outlined features indicate the primary domain of responsibility of the providers of the network, with the primary service entry point (what most users see) shaded.

Layer	Capital Investment	Facilities	ESnet	TeraGrid	I-WIRE
Embedded Services (middleware, video conferencing, instrumentation, etc.)	Servers	Site space/power	Some in house	Some in house	External projects
IP Network Services	IP Routers	Site space/power	All in house	All in house	Some in house
Wavelength Services	DWDM, Optical Amplifiers	Hut space/power	Contracted Services	Contracted Assets	All in house
Dark Fiber	Fiber	Route maintenance			Purchased Assets

ESnet

ESnet as an example of production network infrastructure provides, as its primary service, the IP network services layer (see Table 5.2). That is, ESnet operates IP routers to provide IP connectivity as a service to a collection of sites. The IP network services layer is done in-house, and layers below this are provided via a service contract. ESnet provides some services at the embedded services layer (video conferencing, for example) but the primary service entry point is above the IP network services layer that is, most users interact with ESnet by exchanging IP packets (generated by typical applications such as web browsers, data transfer programs, and e-mail programs). ESnet customers also can (and do) provide services at the embedded services layer and can do so without intervention by ESnet—simply by using the primary

services (IP network service). ESnet's reliability requirements (99.9%) are such that redundancy is mandatory to avoid single points of failure and allow for network equipment upgrades and maintenance without incurring outages. Use of a contracted commercial service that has this level of assurance increases the costs of the services but also allows ESnet to leverage the economy of scale associated with a commercial provider's infrastructure (i.e., the vendor can rely on statistics to provide redundancy for many customers using shared rather than dedicated resources). Simply put, the cost of providing dedicated redundancy (as in a purchased network) is higher than the cost of a commercial redundant service that amortizes the infrastructure costs across many customers.

TeraGrid

TeraGrid also provides the IP network services layer (see Table 5.2) in-house and also has an arrangement with Qwest for the bottom two layers. However, the TeraGrid contract is optimized for capability at the expense of reliability assurances, opting not to require redundancy. Because of the cost of redundancy (spare wavelengths, in effect), greater capacity per unit cost is possible. The TeraGrid contract also involves an up-front payment rather than a monthly payment, changing the financing terms to further reduce capacity costs.

I-WIRE

I-WIRE's business model is to do all layers (see Table 5.2) in-house while contracting out the facilities portion of the lowest layer (route maintenance, which includes restoration of any fiber cable outages). There are several factors that influenced this business model. First, the I-WIRE service model is intended to provide primarily wavelength services to a variety of network and middleware research projects, including the flexibility to serve projects that require access to dark fiber. Second, of the approximately 10 sites involved, only one site (Urbana) is outside a 30-mile radius. This meant that only one span on the network required equipment along the route (optical amplifiers), whereas all of the other spans only required equipment at the endpoints. This makes equipment and ongoing support costs lower. Finally, I-WIRE was designed to provide for both low-level network research that might require access to dark fiber and networking applications, and middleware research projects that require access to wavelengths or groups of wavelengths. I-WIRE chose to obtain fiber and wavelength equipment assets so that a common low-level infrastructure could be leveraged to create multiple types of networks. This approach has the added benefit that I-WIRE could carry production-quality Internet traffic by simply creating a set of wavelengths that are isolated from the research networks.

5.2.2 Considerations for Development of a DOE Networking and Middleware Infrastructure Business Model

Considerable progress was made at this workshop in clarifying requirements for providing networking and middleware infrastructure to support DOE science. Prior to the workshop, it was already clear that in addition to production Internet services, there are needs for both advanced networking capabilities (such as wavelengths to create test networks for applications and/or middleware) and for a growing set of middleware capabilities (as is being done, for example, with the Public Key Infrastructure project the ESnet team is undertaking). A central issue that this workshop began to address is the question of how DOE might provide the networking and middleware infrastructure required to support an increasing

number of advanced applications and middleware projects, the capabilities of which also are increasing rapidly. The strategy of providing three classes of infrastructure (outlined in Section 5.1), attempts to identify the types of needs that are present.

DOE must rapidly develop a detailed implementation strategy that outlines what services are required by the community in each of these three classes, as well as what opportunities exist for providing these services. A specific business model, and plan, will require this input. There are multiple opportunities that must be evaluated, some of which are time-critical because they leverage efforts of other agencies and the academic community. For example, the National Science Foundation recently created the TeraGrid network infrastructure (one of the examples above) to provide large-scale application and middleware projects with capabilities between 10 and 100 times those that can be provided by production networks such as Abilene or ESnet. Another example is the National Light Rail project involving collaboration among several communities (including California's CENIC project, the Pacific Northwest GigaPOP, multiple NSF sites, Starlight, and UCAID/Abilene). This project aims to create a network similar to I-WIRE but with a national footprint.

The detailed analysis needed is beyond the scope of this workshop, and requires a team of experts both from both within and outside the DOE community, including those from industry.[(a)]

5.3 Governance Model

An integrated networking provisioning strategy that attempts to meet all the needs of the existing programs will require revisiting the governance model. The governance model includes DOE program management components, laboratory-university project management components, and forums for input (e.g., standing steering committees, workshops).

5.3.1 Observations

The following observations were made:

- Currently, DOE has no comprehensive network governance model covering all network elements. However the governance of ESnet provides some good examples.

- For production level services, ESnet has a governance structure that, in addition to the MICS Program Office, relies on a group of representatives from its Office of Science constituencies (ESnet Steering Committee). The ESnet Steering Committee has an admirable long-term history as a champion and forum representing a community with increasing network demands and a single network provider.

(a) We note that there is an existing ESnet Research Subcommittee that includes a significant fraction of the expertise needed, including representation from within and outside of the DOE community. This committee's charter was revised in March 2002 to evaluate opportunities and business plans, and may be an important resource for such a study.

- In addition, members of the ESnet Coordinating Committee—a group of technologists organized by the ESnet Steering Committee—have served well to ensure needed coordination of production networking across the national laboratories and to identify and investigate technological opportunities and issues that cut across that community. This committee is more a part of the service provisioning model than the governance structure.

To manage an integrated high-performance networking program, complete with production, testbeds, and R&D, an integrated governance model makes the most sense. An integrated model can balance the input from existing users, who tend to be conservative about changes, with the need to take risks when necessary (e.g., in support of disruptive technology development and evaluation). We believe that a steering structure should be an inherent part of the governance and that it must include a breadth of representation across the DOE Office of Science, encompassing all three elements of networking requirements, including using principal investigators, network project managers, and technical staff from the national laboratories. This will generate a productive friction between traditional network usage and advanced requirements. This friction will result in an encompassing overview of requirements for the network provisioning efforts as well as a forum for prioritizing and resolving those requirements.

We have observed an Office of Science-wide dilemma in regard to network resources. A vision exists that DOE-funded laboratories and principal investigators at universities are an enterprise furthering the DOE mission. However, this vision currently does not benefit from an enterprise perspective that guides decisions concerning the enterprise-wide components needed to realize this vision (e.g., should networking requirements and associated funding requests be included in network-intensive proposals initiated by Office of Science program offices?).

The formalized approach for generating network requirements has not matured to the level of computing requirements. DOE investigators who require a significant amount of a computing resource are familiar with the allocation grant requesting process where they must scope and predict their computing requirements. We believe that the networking program (especially elements 2 and 3) would benefit from a similar approach. A more rigorous set of user requirements would benefit

- long-term network planning
- short-term network resource allocation
- Office of Science programs by ensuring the understanding of network demands and benefits, and motivating corresponding support of the MICS networking program across the breadth of the Office of Science programs.

5.3.2 Findings

The governance model for integrated network provisioning must allow for the management of each element's requirements in a context that is highly influenced by the opportunities and risks that face the other two elements. For example, as networking advances are developed and become critical to support

scientific processes, there must be a prioritized and managed transfer of these network advances into the production network environment (e.g., global discovery and scheduling, uniform computer data access, authentication, collaboration support).

Shared network provisioning will encourage increased interaction throughout the governance model. We believe the resulting benefits will extend well beyond shared networking provisioning. Ideally the governance model for all networking programs (both at DOE and in the field) will foster the creation of

- a well articulated vision for all the elements
- an integrated goal-setting process
- strategies across the programs that are mutually understood and shared.

We suggest that a major Network Initiative is probably necessary to break the "zero sum game" in networking that has faced the community for many years, limiting what can be done and slowing the progress of programs across the Office of Science. We believe that the SciDAC Initiative model is worth considering, allowing the ownership of the initiative and its resulting efforts across the Office of Science.

References

1 National Virtual Observatory, NVO. http://www.us-vo.org/

2 Astro-IT challenges and big UK surveys. A. Lawrence. In *Virtual Observatories of the Future*. 2000. Caltech. http://www.roe.ac.uk/wfau/nvo/index.htm

3 Computational and Data Grids in Large-Scale Science and Engineering, W. Johnston. Future Generation Computer Systems, 2002.

4 BIOSYNC - Structural Biology Synchrotron Users Organization. http://biosync.sdsu.edu/

5 R. M. Sweet, M.Becker, J. M.Skinner, "Collaboratory Tools for Macromolecular Crystallography at the NSLS", private communication.

6 The Compact Muon Solenoid Technical Proposal, CMS. http://cmsdoc.cern.ch/

7 The ATLAS Technical Proposal. http://atlasinfo.cern.ch/ATLAS/TP/NEW/HTML/tp9new/tp9.html

8 The Large Hadron Collider Project, LHC. http://lhc.web.cern.ch/lhc/general/gen_info.htm

9 The BaBar Experiment at SLAC. http://www-public.slac.stanford.edu/babar/

10 The D0 Experiment at Fermilab. http://www-d0.fnal.gov/

11 The CDF Experiment at Fermilab. http://www-cdf.fnal.gov/

12 The Relativistic Heavy Ion Collider at BNL, RHIC. http://www.bnl.gov/RHIC/

13 The Alcator C-Mod Tokamak Fusion Research Project. http://www.psfc.mit.edu/cmod/

14 DIII–D National Fusion Facility at General Atomics. http://web.gat.com/diii-d/

15 The National Spherical Torus Experiment, NSTX. http://nstx.pppl.gov/

16 "Computers are from Mars, organisms are from Venus." J. Kim. *Computer* 35(7):25–32. July 2002.

17 "A High Performance Computing Network for Protein Conformation Simulations." M. Pellegrini. *ERCIM News* No.43 - October 2000. http://www.ercim.org/publication/Ercim_News/enw43/pellegrini.html

18 "Ab Initio Methods for Protein Structure Prediction: A New Technique based on Ramachandran Plots." A. Bernascone and A. M. Segre. *ERCIM News* No.43 - October 2000. http://www.ercim.org/publication/Ercim_News/enw43/bernasconi.html

19 The Anatomy of the Grid: Enabling Scalable Virtual Organizations. I. Foster, C. Kesselman, and S. Tuecke. *International J. Supercomputer Applications* 15(3). 2001. http://www.globus.org/research/papers.html - anatomy

20 *The Grid: Blueprint for a New Computing Infrastructure*. I. Foster and C. Kesselman, eds. 1998, Morgan Kaufmann. http://www.mkp.com/books_catalog/1-55860-475-8.asp

21 Global Grid Forum, GGF. http://www.gridforum.org/

22 The Globus Project. http://www.globus.org/

23 The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration. I. Foster, C. Kesselman, J. Nick, and S. Tuecke. http://www.globus.org/research/papers.html - OGSA

24 Middleware for Filtering Large Archival Scientific Datasets in a Grid Environment, DataCutter. http://www.cs.umd.edu/projects/hpsl/ResearchAreas/DataCutter.htm

25 Abilene Core Node Router Proxy, Abilene Network Operations Center, Global Network Operations Center at Indiana University. http://loadrunner.uits.iu.edu/%7Erouterproxy/abilene

26 Internet2 NetFlow Statistics, Abilene Network Operations Center, Global Network Operations Center at Indiana University. http://netflow.internet2.edu/

27 Network Weather Service, University of California, Santa Barbara. http://nws.cs.ucsb.edu/

28 Internet Engineering Task Force, ICSI Center for Internet Research, Berkeley, California. http://www.icir.org/floyd/papers/draft-floyd-tcp-highspeed-01.txt

29 Stream Control Transmission Protocol (SCTP). http://www.sctp.de/

30 Scheduled Transfer Protocol (ST), High-Performance Parallel Interface Standards Group. http://www.hippi.org/cST.html

31 "Congestion Control for High BandwidthDelay Product Networks." D. Katabi, M. Handley, and C Rohrs. *SIGCOMM'02*, August 19-23, 2002, Pittsburgh, Pennsylvania. http://www.acm.org/sigs/sigcomm/sigcomm2002/papers/xcp.pdf

32 ICSI Center for Internet Research, Berkeley, California. http://www.icir.org/mjh/xcp.ps

33 Bro: A System for Detecting Network Intruders in Real-Time. V. Paxson, Lawrence Berkeley National Laboratory Network Research Group, Berkeley, California. http://www-nrg.ee.lbl.gov/bro-info.html

34 *A Vision for DOE Scientific Networking Driven by High Impact Science.* W.E. Johnston, W.T.C. Kramer, J. F. Leighton, and C. Catlette. March 15, 2002. U.S. Department of Energy, Washington, D.C. http://www.lbl.gov/CS/Network_Vision_Whitepaper.pdf

35 Energy Sciences Network, Lawrence Berkeley National Laboratory, Berkeley, California. http://www.es.net/

36 TeraGrid, National Science Foundation. http://www.teragrid.org/

37 Illinois Wired/Wireless Infrastructure for Research and Education (I-WIRE). http://www.i-wire.org/

Appendix A

Climate

Appendix A Climate

Gary Strand, National Center for Atmospheric Research

A.1 Introduction

To better understand climate change, we need better climate models – and to get those, we need to exhaustively analyze what's incorrect about today's models in order to improve them. The cycle of analysis → improved model → analysis is typical of climate model work generally. One thing we do know is that climate models today are too low in resolution to get some important features of the climate right. Generally, the computing power will be there over the next 5-10 years, but to determine things like climate extremes (hurricanes,[(a)] drought and precipitation pattern changes,[(b)] heat waves and cold snaps) and other potential changes as a result of climate change,[(c)] we need better analysis. Currently, analysis is accomplished by transferring the data of interest from the computing site to the climate scientist's institution. This can be inefficient if the data volume is large, and several strategies to reduce the data volume before transfer have been developed. However, these processes are often ad hoc and need to be improved or rendered moot.

That means faster nets to access more climate model data more efficiently, and faster nets to do nifty things like visualizations and collaboratories to assist climate scientists in understanding climate models and climate change. Since climate models require large computing resources, there are only a few sites in the U.S. and worldwide that are suitable for executing these models at this time. In addition, for model efficiency reasons, the data produced by these integrations are stored at the same sites - however, climate scientists are scattered all over the globe, which means that data distribution for analysis is critical.

A.2 The Next Five Years

Over the next five years, climate models will see an even greater increase in complexity than that seen in the last ten years. Influences on climate will no longer be approximated by essentially fixed quantities, but will become interactive components in and of themselves. The North American Carbon Project (NACP), which endeavors to fully simulate the carbon cycle, is an example. Increases in resolution, both

(a) Hurricane Andrew was almost exactly 10 years ago and cost many lives and about $20 billion damage. Current climate models aren't quite good enough to resolve hurricanes, but research models driven by reasonably realistic future climate scenarios imply that Andrew-strength hurricanes striking the US will become more common. That implies many more billions in damage and more deaths.

(b) Likewise, the drought the Western US is currently facing could become the typical climate pattern, with millions of acres of forests burning in wildfires, and things like the cost of supplying water to the burgeoning populations of the Western U.S. Changes in precipitation location may also make agriculture in the Midwest U.S more problematic – either extended dry periods or floods like those that plagued the upper Midwest in the early 1990s.

(c) Here I'm talking about changes in disease patterns, for example. It's possible that climate change will make the US more susceptible to the spread of diseases found today mostly in the tropics. The West Nile virus is relatively innocuous compared to malaria.

spatially and temporally, are in the plans for the next two to three years. The atmospheric component of the coupled system will have a horizontal resolution of approximately 150 km and 30 levels. A plan is being finalized for model simulations that will create about 30 terabytes of data in the next 18 months, which is double the rate of model data generation of the Parallel Climate Model, PCM.

These much finer resolution models, as well as the distributed nature of computing resources, will demand much greater bandwidth and robustness from computer networks than is presently available. These studies will be driven by the need to determine future climate at both local and regional scales as well as changes in climate extremes - droughts, floods, severe storm events, and other phenomena. Climate models will also incorporate the vastly increased volume of observational data now available (and that available in the future), both for hind casting and intercomparison purposes. The end result is that instead of tens of terabytes of data per model instantiation, hundreds of terabytes to a few petabytes (10^{15}) of data will be stored at multiple computing sites, to be analyzed by climate scientists worldwide. The Earth System Grid and its descendents will be fully utilized to disseminate model data and for scientific analysis. Additionally, these more sophisticated analyses and collaborations will increase the needed network resources and infrastructure. It's expected that multiple climate scientists will examine the model data – more than today. PCM data has been analyzed by scientists at UCSD, the University of Colorado at Boulder, NOAA, NERSC, PNNL, and overseas, including in Sweden, Germany and Japan. Bulk data transfer will be necessary, as well as tools like Access Grids and personal Grids.

As climate models become more multidisciplinary, scientists from fields outside of climate, oceanography and the atmospheric sciences will collaborate on the development and examination of climate models. Biologists, hydrologists, economists and others will assist in the creation of additional components that represent important but as-yet poorly known influences on climate. These models, sophisticated themselves, will likely be utilized at computing sites other than where the climate model is executed. In order to maintain efficiency, dataflow to and from these collaboratory efforts will demand extremely robust and fast networks.

A.3 2007 and Beyond

In the following five years, climate models will again increase in resolution, and many more fully interactive components will be integrated. At this time, the atmospheric component may become nearly mesoscale (commonly used for weather forecasting) in resolution, 30 km by 30 km, with 60 vertical levels. Climate models will be used to drive regional scale climate and weather models, which require resolutions in the tens to hundreds of meters range, instead of the typical hundreds of kilometers resolution of the CCSM and PCM. There will be a true carbon cycle component, models of biological processes will be used, for example, simulations of marine biochemistry (which affects the interchange of greenhouse gases like methane and carbon dioxide with the atmosphere), and fully dynamic vegetation. These scenarios will include human population change and growth (which effect land usage and rainfall patterns) and econometric models, to simulate the potential changes in natural resource usage and efficiency. Additionally, models representing solar processes, to better simulate the incoming solar radiation, will be integrated. Climate models at this level of sophistication will likely be run at more than one computing center in distributed fashion, which will demand extremely high speed and tremendously robust computer networks to interconnect them. Data volumes could reach several petabytes, which is a conservative estimate.

Table A.1. Climate Requirements Summary

Feature	Characteristics That Motivate Advanced Infrastructure	Vision for the Future Process of Science	Anticipated Requirements	
Time Frame			**Network**	**Middleware**
Near-term	*A few data repositories, many distributed computing sites • NCAR[(a)] - 20 Tbytes • NERSC[(b)] - 40 Tbytes • ORNL[(c)] – 40 Tbytes		• Authenticated data streams for easier site access through firewalls	• Server side data processing (computing and data cache embedded in the net) • Information servers for global data catalogues
5 years	• Add many simulation elements/components as understanding increases • 100 Tbytes / 100 model yrs generated simulation data – 1-5 Pbytes / yr (at NCAR) • Distribute in large datasets to major users/collaborators for post-simulation analysis	• Enable the analysis of model data by all of the collaborating community (major US collaborators are a dozen universities, and several Federal Agencies)	• Robust access to large quantities of data (multiple paths)	• Reliable data/file transfer o Across system/ network failures
5+ years	• Add many diverse simulation elements/components, including from other disciplines - this must be done with distributed, multidisciplinary simulation as the many specialized sub-models will be managed by experts in those fields • 5-10 Pbytes/yr (at NCAR)	• Integrated climate simulation that includes all high-impact factors	• Robust networks supporting distributed simulation - adequate bandwidth and latency for remote analysis and visualization of massive datasets	• Quality of service guarantees for distributed, simulations • Server side computation for data extraction/ subsetting, reduction, etc., before moving across the network
	• Virtualized data to reduce storage load			• Virtual data catalogues for data generation descriptions, data regeneration planners, data naming and location transparency services for reconstituting data on demand

(a) NCAR = National Center for Atmospheric Research.
(b) NERSC = National Energy Research Scientific Computing Center
(c) ORNL = Oak Ridge National Laboratory

Appendix B

Spallation Neutron Source

Appendix B Spallation Neutron Source

J. P. Hodges, SNS Division, Oak Ridge National Laboratory

B.1 Introduction

Neutron scattering is a unique and powerful tool for studying the structure and dynamics of materials at the atomic, molecular, and macromolecular levels. Six U.S. Department of Energy (DOE) laboratories (Argonne, Brookhaven, Lawrence Berkeley, Los Alamos, Oak Ridge, and Jefferson Lab) are partners in the design and construction of the Spallation Neutron Source (SNS), a one-of-a-kind facility at Oak Ridge, Tennessee, that will provide the most intense pulsed neutron beams in the world for scientific research and industrial development.

Figure B.1. Spallation Neutron Source Facility at ORNL

When completed, the SNS will enable new levels of investigation into the properties of materials of interest to chemists, condensed matter physicists, biologists, pharmacologists, materials scientists, and engineers, in an ever-increasing range of applications.

Completion of the Spallation Neutron Source (SNS) facility in early 2006 heralds a new era for neutron scattering sciences in the U.S. SNS supports multiple instruments that will offer users at least an order of magnitude performance enhancement over any of today's pulsed spallation neutron source instruments. This great increase in instrument performance is mirrored by an increase in data output from each instrument. In fact, the use of high resolution detector arrays and supermirror neutron guides in SNS instruments means that the data output rate for each instrument is likely to be close to two orders greater than a comparable U.S. instrument in use today. This, combined with increased collaboration among the several related US facilities, will require a new approach to data handling, analysis and sharing.

The high data rates and volumes from the new instruments will call for significant data analysis to be completed offsite on high-performance computing systems. High-performance network and distributed computer systems will handle all aspects of post-experiment data analysis (few approximations and CPU intensive) and the approximate analysis that can be used to support near real-time interactions of scientists with their experiments.

Neutron scattering experiments are small affairs, and typically access to the data may be required for perhaps five people distributed between the neutron facility and the principal investigators home institution. However, since, neutron scattering instruments operate 24 hrs 7 days a week during facility

run periods, real time data visualization, some real time analysis capabilities, and security to modify experiment conditions by a user at his/her hotel via an internet browser is desired.

Users are given a specific amount of time (0.5 to 2 days) on an instrument. The close to real-time visualization and partial analysis capabilities, therefore, allow a user to refine the experiment during the allotted time. For the majority of SNS user experiments, the material or property being studied is novel and this capability is essential for the experimentalist to focus in on the area of interest and maximize the science accomplished in the limited amount of beam time.

In this scenario, the combined data transfer between the twelve SNS instruments and a distributed computer network for real time data mapping is estimated to be a constant 1 Gbit/sec (assuming 50% of users using real time visualization). The return data stream to servers managing the visualization and analysis tasks as well as communicating to the users across LAN and/or internet would be around 140 Mbit/sec (dominated by the 4-D and 3-D response maps). The servers (one for each instrument) would generate selected views of the response function as well as send (if requested by the user) the response function back out to the distributed computer network for quick/partial analysis.

B.2 Five to Ten Years Out

It is anticipated that analysis of experimental data in the future may be achieved by incorporating a scattering law model within the iterative response function extraction procedure. These advanced analysis methods are expected to require the use of powerful offsite computing systems, and the data may transit the network several times as experiment / experimenter / simulation interaction converges to an accurate representation.

Table B.1. Spallation Neutron Source (SNS) Requirements Summary

Feature	Characteristics That Motivate Advanced Infrastructure	Vision for the Future Process of Science	Anticipated Requirements	
Time Frame			**Networking**	**Middleware**
Near-term	(Facility comes on-line in 2006)			
5 years	• The 12 instruments at the SNS will operate about 200 days/year and generate an aggregate 80 Gbytes/day • The data analysis will be accomplished mostly on computing systems that are remote from the SNS	•	• 50-80 Mbits/sec sustained • 320 Mbits/sec peak	• Workflow management • Reliable data transfer
	• Neutron scattering instruments operate 24 hr 7 days a week during facility run periods, real time data visualization, some real time analysis capabilities, and security to modify experiment conditions by a user at his/her hotel via an internet browser will be required.	• Real-time data analysis and visualization will enhance the productivity of the science done at SNS, which runs 24 hr/day.	• 1 Gbits/sec sustained	• Security (authentication and access control) to permit direct interaction with the instrument remotely.
5-10 years	• Statistical scattering models will be incorporated into analysis code requiring supercomputer levels of remote computing.	• Iterative analysis of the data with the use of models running on supercomputing systems will produce much more accurate results and understanding.		

Appendix C

Macromolecular Crystallography

Appendix C Macromolecular Crystallography

T. N. Earnest, C. W. Cork, G. McDermott, J. R. Taylor, Lawrence Berkeley National Laboratory

C.1 Introduction

Macromolecular crystallography is an experimental technique that is used to solve structures of large biological molecules (such as proteins) and complexes of these molecules. The current state-of-the-art implementation of this technique requires the use of a source of very intense, tunable, x-rays which are only produced at large synchrotron radiation facilities. There are approximately 18 synchrotron radiation facilities in operation or under construction worldwide: 6 in the United States, 6 in Europe, 2 in Japan, and one each in Brazil, Canada, China, and Taiwan. In the United States alone, there are 36 crystallography stations which are distributed among the synchrotron facilities and dedicated to macromolecular crystallography [1]. Operating costs for each of these crystallography stations is estimated to be approximately $2K - $3K/hr. These stations are also responding to an increasing demand to solve new structures arising from both the national genomics research programs and from commercial drug development R&D. The high operating cost of these facilities, coupled with the heavy demand for their use, has led to an emphasis on increased productivity and data quality which will need to be accompanied by increased network performance and functionality.

Data acquisition for macromolecular crystallography typically involves repeated exposure and readout of imaging detectors while rotating the sample in the x-ray beam. Current systems can produce 10 - 100 Mbyte images at a maximum rate of 0.5 image/sec. Future systems which should become available within the next 5 - 10 years are expected to reach peak data rates of 50 - 500 Mbyte/sec. Average data rates are somewhat less due to issues with sample handling and instrument setup; however, new developments in automated sample handling and parallel data processing are promising average data rates which approach 50% of peak rates. A typical dataset consists of 500 - 1000 images which are usually stored uncompressed for fast online analysis (requiring 5 - 100 Gbyte disk storage).

C.2 Networking Requirements

As illustrated in Figure C.1, the data acquisition process involves several interactive online components, data archiving and storage components, and a compute-intensive offline component. Each component has associated networking requirements.

Online process control and online data analysis are real-time, interactive, activities which monitor and coordinate data collection. They require high-bandwidth access to images as they are acquired from the detector. Online data analysis is currently limited primarily to sample quality assurance and to data collection strategy. There is increasing emphasis on expanding this role to include improved crystal scoring methods and real-time data processing to monitor sample degradation and data quality. Online access to the image datasets is collocated and could make good use of intelligent caching schemes. Datasets from previously exposed samples are not required during online processing.

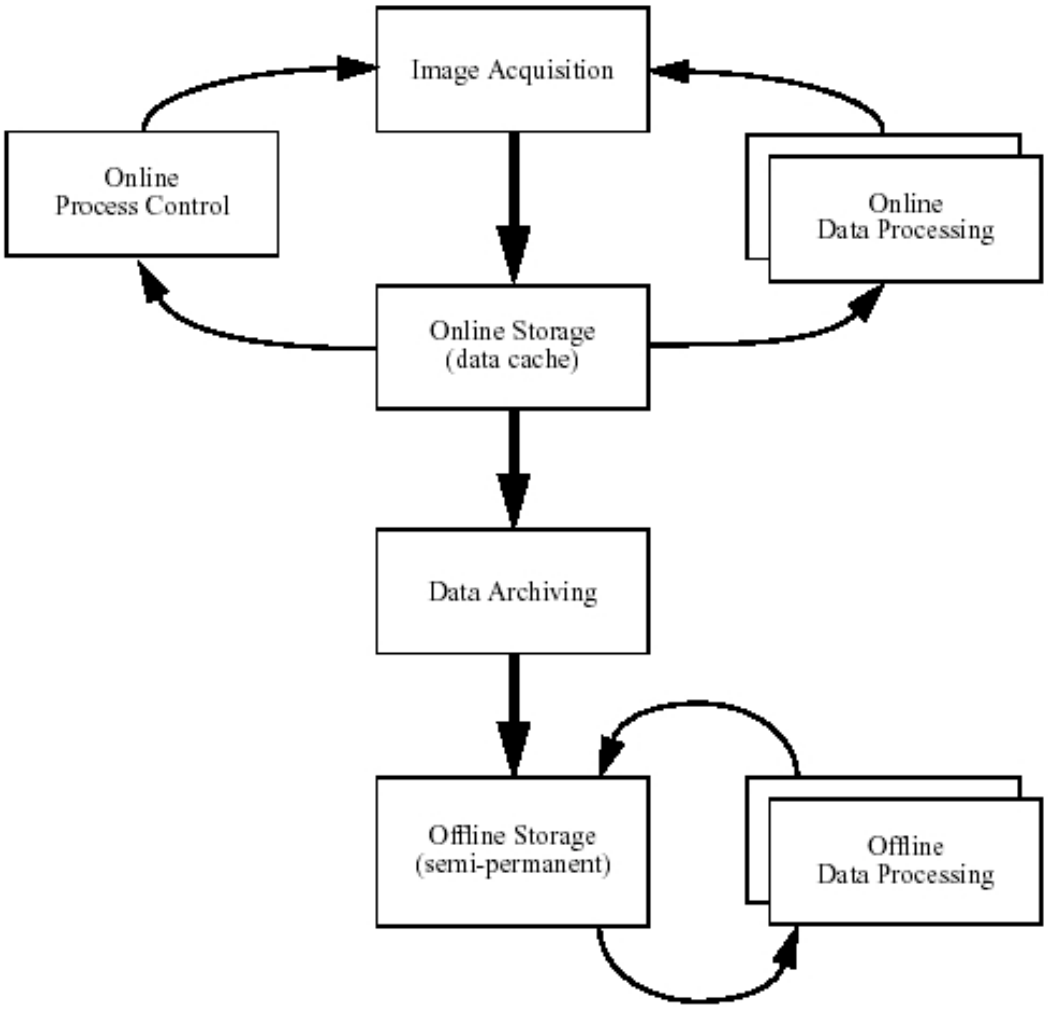


Figure C.1. A Simplified View of the Macromolecular Crystallography Data Handling Process

End-to-end data transfer rates have been the main limitation in network transfer mechanisms. In the following sections we will briefly discuss possible networking scenarios for the online and offline data acquisition components.

C.2.1 Online Control and Data Processing

High performance networking can play several roles in online control and data processing. Bob Sweet at BNLs National Synchrotron Light Source has outlined several approaches to remote, networked, collaboratory operation [2]:

- **Remote Observer**. This involves remote monitoring of data acquisition progress and snapshots of the online data processing status. Data acquisition is coordinated exclusively by local personnel. A limited set of agent programs will provide access and presentation services to the remote client.

Network bandwidth and QOS requirements are minimal. This scenario of operation would benefit from middleware utilities which facilitate agent discovery, authentication, and connection.

- **Collaboratory**. This is a hybrid remote/local control scenario where local operators communicate with remote collaborators to coordinate data acquisition. This scenario will require telepresence and network conferencing software, in addition to the facilities required for the 'Remote Observer' approach described above. Network bandwidth requirements are approximately the same as for 'Remote Observer', but QOS requirements are increased to guarantee low-latency communications.

- **Remote Control**. This scenario extends some or all of the online process control function to the remote site. Several variations of this approach are being developed, ranging from remote 'recipe' prescription to total remote instrument control. Most of the telepresence facilities required for the 'Collaboratory' approach will also be required here for interaction with local support personnel. Network bandwidth requirements are somewhat greater than for the 'Remote Observer' as update rates will need to be increased to improve interactive feedback. QOS requirements are also increased to include additional security and transaction features. For all of the above scenarios, high bandwidth is not so important as QOS and management services. Especially valuable would be middleware tools which facilitate the initiation, configuration, and monitoring of these services.

C.2.2 Offline Storage and Data Archiving

As mentioned above, offline storage is generally remote from the local data collection station. The datasets are occasionally stored at shared processing facilities (such as Brookhaven Lab's ASDP facility [3]). However, the datasets are most often transferred to private institutional storage. This requirement places a large burden on the data archiving process which transfers the data between online and offline storage units. Current requirements for average data transfer rate are 1 - 25 Mbyte/s per station; it is expected that in 5 - 10 years this will increase by an order of magnitude to 10 - 250 Mbyte/s per station. This is further exacerbated by the fact that most research facilities have from 4 - 8 stations; this places a future requirement of 40 - 2000 Mbyte/s per facility. Advanced data compression schemes might be able to reduce these figures by a factor of 5 - 10. Within this data rate performance envelope there are several scenarios that might be implemented in a high performance networking environment:

- **Data replication**. This involves variations on data copying between local and remote storage sites. It could be as simple as an automated remote backup tool to a full-fledged data mirroring service. Middleware services which support location, authentication, and data replication will be essential for the success of this scenario.

- **Virtual storage**. This involves variations of network filesystems. In this scheme online and offline data storage are merged into a single virtual storage facility. If network performance is sufficient, it could involve a scheme as simple as NFS or AFS; however, the heavy random access use by local online data processing applications probably dictates some sort of transparent local cache mechanism. Middleware should implement this virtual storage mechanism and provide simple management tools and API's for storage location, authentication, and attachment.

C.3 Summary

In addition to increased raw network bandwidth, the next generation high performance networking infrastructure will need to provide tools and services which facilitate object discovery, security, and reliability. These tools are needed for both low-latency applications such as remote control, as well as high throughput data transfer applications such as data replication or virtual storage systems.

C.4 References

1. BIOSYNC - Structural Biology Synchrotron Users Organization (http://biosync.sdsu.edu).

2. R. M. Sweet, M. Becker, J. M. Skinner, “Collaboratory Tools for Macromolecular Crystallography at the NSLS”, private communication.

3. NSLS Automated Structure Determination Platform (http://asdp.bnl.gov).

Appendix D

High-Energy Physics: Scientific Exploration at the High-Energy Frontier

Appendix D High-Energy Physics: Scientific Exploration at the High-Energy Frontier

Julian J. Bunn, Center for Advanced Computing Research California Institute of Technology
Harvey B. Newman, Physics Department, California Institute of Technology

The major high energy physics experiments of the next twenty years will break new ground in our understanding of the fundamental interactions, structures and symmetries that govern the nature of matter and space-time. Among the principal goals are to find the mechanism responsible for mass in the universe, and the "Higgs" particles associated with mass generation, as well as the fundamental mechanism that led to the predominance of matter over antimatter in the observable cosmos.

The largest collaborations today, such as CMS and ATLAS who are building experiments for CERN's Large Hadron Collider (LHC) program, each encompass 2000 physicists from 150 institutions in more than 30 countries. Each of these collaborations includes 300-400 physicists in the U.S., from more than 30 universities, as well as the major US HEP laboratories. The current generation of operational experiments at SLAC (BaBar) and FermiLab (D0 and CDF), as well as the experiments at the Relativistic Heavy Ion Collider (RHIC) program at BNL, face similar challenges. BaBar in particular has already accumulated datasets approaching a petabyte (10^{15} bytes).

The HEP (or HENP, for high energy and nuclear physics) problems are the most data-intensive known. Hundreds to thousands of scientist-developers around the world continually develop software to better select candidate physics signals, better calibrate the detector and better reconstruct the quantities of interest (energies and decay vertices of particles such as electrons, photons and muons, as well as jets of particles from quarks and gluons). The globally distributed ensemble of facilities, while large by any standard, is less than the physicists require to do their work in an unbridled way. There is thus a need, and a drive to solve the problem of managing global resources in an optimal way, in order to maximize the potential of the major experiments for breakthrough discoveries.

In order to meet these technical goals, priorities have to be set, the system has to managed and monitored globally end-to-end, and a new mode of "human-Grid" interactions has to be developed and deployed so that the physicists, as well as the Grid system itself, can learn to operate optimally to maximize the workflow through the system. Developing an effective set of tradeoffs between high levels of resource utilization, rapid turnaround time, and matching resource usage profiles to the policy of each scientific collaboration over the long term presents new challenges (new in scale and complexity) for distributed systems.

Collaborations on this global scale would not have been attempted if the physicists could not plan on excellent networks: to interconnect the physics groups throughout the lifecycle of the experiment, and to make possible the construction of Data Grids capable of providing access, processing and analysis of massive datasets. These datasets will increase in size from petabytes to exabytes (1 exabyte = 10^{18} bytes) within the next decade. As well, excellent middleware to facilitate the management of worldwide computing and data resources must be brought to bear on the data analysis problem of HEP.

Successful construction of network and Grid systems able to serve the global HEP and other scientific communities with data-intensive needs could have wide-ranging effects: on research, industrial and commercial operations. The key is intelligent, resilient, self-aware, and self-forming systems able to support a large volume of robust terabyte and larger transactions, able to adapt to a changing workload, and capable of matching the use of distributed resources to policies. These systems could provide a strong foundation for managing the large-scale data-intensive operations processes of the largest research organizations, as well as the distributed business processes of multinational corporations in the future.

Several important collaborations are involved in the HEP work to use Grids for distributed data processing. The DOE Science Grid [1] is working on identifying and resolving the issues for building production Grids for the DOE Office of Science [2]. The Particle Physics Data Grid (PPDG), is working on Grid middleware and systems for distributed analysis of HEP experiment data. These two Grid projects collaborate on several aspects of HEP Grids. The DOE Science Grid is funded by the DOE/MICS Office [3] and PPDG is funded by the DOE HENP Office [4].

To cite one example of the technology issues being addressed in HEP, we consider the development of virtualized data coupled with what the commercial sector calls Content Delivery Networks.

The GriPhyN (Grid Physics Network – http://www.pgriphyn.org) project is a collaboration of computer science and other IT researchers and physicists from the ATLAS, CMS, LIGO and SDSS experiments. The project is focused on the creation of petascale Virtual Data Grids that meet the data-intensive computational needs of a diverse community of thousands of scientists spread across the globe. The concept of Virtual Data encompasses the definition and delivery to a large community of a (potentially unlimited) virtual space of data products derived from experimental data. In this virtual data space, requests can be satisfied via direct access and/or computation, with local and global resource management, policy, and security constraints determining the strategy used. Overcoming this challenge and realizing the Virtual Data concept requires advances in three major areas:

- Virtual data technologies. Advances are required in information models and in new methods of cataloging, characterizing, validating, and archiving software components to implement virtual data manipulations.

- Policy-driven request planning and scheduling of networked data and computational resources. Mechanisms are required for representing and enforcing both local and global policy constraints and new policy-aware resource discovery techniques.

- Management of transactions and task-execution across national-scale and worldwide virtual organizations. New mechanisms are needed to meet user requirements for performance, reliability, and cost. Agent computing will be important to permit the Grid to balance user requirements and Grid throughput, with fault tolerance.

The GriPhyN project is primarily focused on achieving the fundamental IT advances required to create petascale Virtual Data Grids, but is also working on creating software systems for community use, and applying the technology to enable distributed, collaborative analysis of data.

A multi-faceted, domain-independent Virtual Data Toolkit is being created and used to prototype the virtual data Grids, and to support the CMS, ATLAS, LIGO, and SDSS analysis tasks.

Figure D.1 shows a production Grid, as envisaged by GriPhyN, showing the strong integration of data generation, storage, computing and network facilities, together with tools for scheduling, management and security.

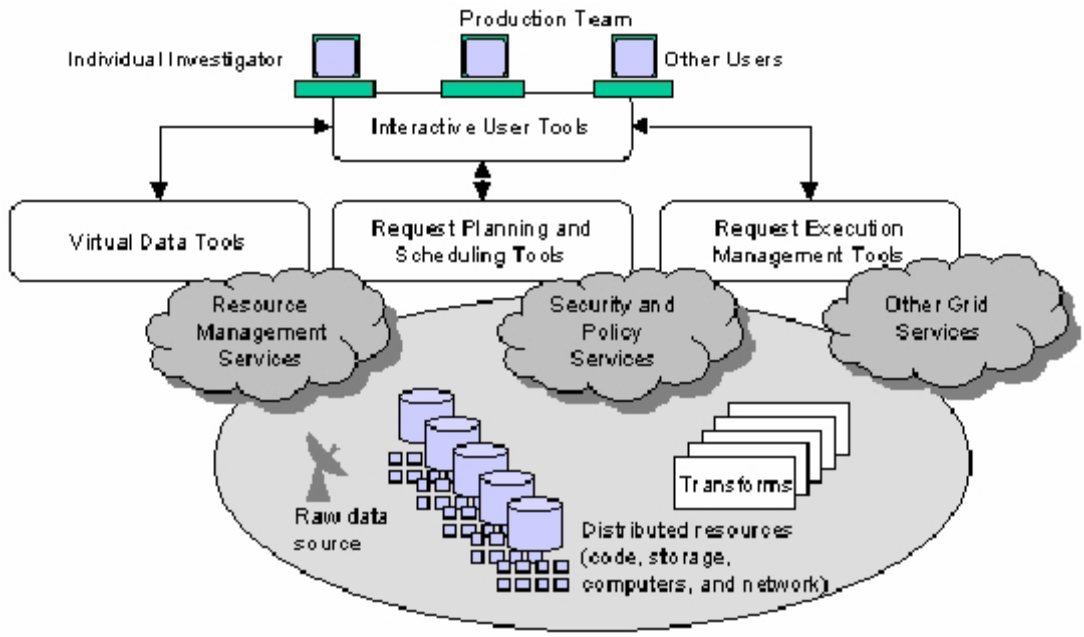


Figure D.1. A Production Data Grid as Envisioned by GriPhyN, Showing the Strong Interaction of Data Generation, Storage, Computing, and Network Facilities, Together with Grid Tools for Scheduling, Management, and Security

D.1 References

1. DOE Science Grid. http://www.doesciencegrid.org

2. Office of Science, U.S. Department of Energy. http://www.er.doe.gov/

3. Mathematical, Information, and Computational Sciences (MICS), Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy. http://www.sc.doe.gov/ascr/mics/

4. Office of High Energy and Nuclear Physics (HENP), Office of Science, U.S. Department of Energy. http://www.er.doe.gov/henp/

Table D.1. High-Energy Physics Requirements Summary

Feature Time Frame	Characteristics that Motivate Advanced Infrastructure	Vision for the Future Process of Science	Anticipated Requirements	
			Networking	Middleware
Near-term	• Instrument based data sources • Hierarchical data repositories • Hundreds of analysis sites • 100 gigabytes of data extracted from a 100 terabyte data store and transmitted to the analysis site in 10 minutes in order not to destabilize the distributed processing system with too many outstanding data requests • Improved quality of videoconferencing capabilities • Cross-site authentication/ authorization	• The ability to analyze the data that comes out of the current experiment • Remote collaborative experiment control	• gigabit/sec • end-to-end QoS	• Secure access to world-wide resources • Data migration in response to usage patterns and network performance o naming and location transparency • Deadline scheduling for bulk transfers • Policy based scheduling / brokering for the ensemble of resources needed for a task • Automated planning and prediction to minimized time to complete task
5 years	• 100 terabytes of data extracted from a 100 petabyte data store and transmitted to the analysis site in 10 minutes in order not to destabilize the distributed processing system with too many outstanding data requests • Global collaboration • Compute and storage requirements will be satisfied by optimal use of all available resources	• Worldwide collaboration will cooperatively analyze data and contribute to a common knowledge base • Discovery of published (structured) data and its provenance	• 100 gigabit/sec • lambda based point-to-point for single high-bandwidth flows • capacity planning • Network monitoring	• Track world-wide resource usage patterns to maximize utilization • Direct network access to data management systems • Monitoring to enable optimized use of network, compute, and storage resources • Publish / subscribe and global discovery
5+ years	• 1000s of petabytes of data		• 1000 gigabit/sec	

Appendix E

Magnetic Fusion Energy Sciences

Appendix E Magnetic Fusion Energy Sciences

D. P. Schissel, General Atomics Fusion Group
M. J. Greenwald, MIT Plasma Science and Fusion Center
W. E. Johnston, Lawrence Berkeley National Laboratory

The long–term goal of magnetic fusion research is to develop a reliable energy system that is environmentally and economically sustainable. To achieve this goal, it has been necessary to develop the science of plasma physics, a field with close links to fluid mechanics, electromagnetism, and nonequilibrium statistical mechanics. The highly collaborative nature of the Fusion Energy Sciences (FES) due to the small number of experimental facilities and a computationally intensive theoretical program are creating new and unique challenges for computer networking and middleware.

In the United States, experimental magnetic fusion research is centered at three large facilities (Alcator C–Mod [22], DIII–D [23], NSTX [24]) with a present day replacement value of over $1B; clearly too expensive to duplicate. As these experiments have increased in size and complexity, there has been concurrent growth in the number and importance of collaborations between large groups at the experimental sites and smaller groups located at universities, industry sites, and national laboratories.

Teaming with the experimental community is a theoretical and simulation community whose efforts range from the very applied analysis of experimental data to much more fundamental theory like the creation of realistic non–linear 3D plasma models.

The three main magnetic fusion experimental sites operate in a similar manner. The gross tokamak machine hardware parameters are configured before the start of the experimental day. Magnetic fusion experiments operate in a pulsed mode producing plasmas of up to 10 seconds duration every 10 to 20 minutes, with 25–35 pulses per day. For each plasma pulse up to 10,000 separate measurements versus time are acquired at sample rates from kHz to MHz, representing hundreds of megabytes of data. Throughout the experimental session, hardware/software plasma control adjustments are made as required by the experimental science. These adjustments are debated and discussed amongst the experimental team (typically 20–40 people) with most working on site in the control room but 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 places a large premium on rapid data analysis that can be assimilated in near–real–time by a geographically dispersed research team.

The computational emphasis in the experimental science area is to perform more, and more complex, data analysis between plasma pulses. For example, today a complete time–history of the plasma magnetic structure is available between pulses by using parallel processing on Linux clusters. Five years ago, only selected times were analyzed between pulses with the entire time–history completed overnight. Five years from now, analysis that is today performed overnight should be completed between pulses. Such enhanced between-pulse data analysis will include more advanced simulations. The ability to more accurately compare experiment and theory between pulses will greatly enhance the value of experimental operations. Today, these comparisons are done after experimental operations have concluded when it is

too late to adjust experimental conditions, and this is very limiting for the experimentalists who typically only get a few days a year on a fusion device to test out their theories.

It is anticipated that the data available between pulses will approach the 1 Gbyte level within the next 5 years. Overall data transfer rates must be fast enough to allow time for detailed analysis and subsequent examination by the scientific staff within the 20 minutes between plasmas. Peak network rates on the order of 500 Mbits/sec are required if a third of a minute is allowed to transfer the entire repository. This peak rate is required intermittently over the course of a year. Typically, experimental operation on one of the three main machines is 8 hours or 30 pulses a day, 5 days a week for approximately 20 weeks a year (two or more machines can be operating at the same time). During an experimental day anywhere from 5 to 10 remote sites can be participating. Although the entire repository is not transferred to each remote site, subsets of the data are transferred for visualization and analysis with results being written back into the main data storage system. It is the dynamic nature of the data repository combined with the large number of distributed users that makes replication at remote sites more than just a simple task.

With the creation of more data between pulses there exists an increasing burden to assimilate all of the data. Enhanced visualization tools are presently being developed that will allow this order of magnitude increase to be effectively used for decision making by the experimental team. Clearly, the movement of this quantity of data in a 15–20 minute time window to computational clusters, to data servers, and to visualization tools used by an experimental team distributed across the United States and the sharing of remote visualizations back into the control room will place a severe burden on present day network technology.

Although the fundamental laws that determine the behavior of fusion plasmas are well known, obtaining their solution under realistic conditions is a computational science problem of enormous complexity.

Datasets generated by these simulation codes will approach the 1 TB level within the next three to five years. Additionally, these datasets will be analyzed like experimental plasmas are analyzed to extract further information. Therefore, the data repository for simulations will be dynamically evolving rather than a write–once type scenario. Network rates of 500 Mbits/sec to 1000 Mbits/sec are required, similar to the above, if we assume that subsets of a simulation code run (~ 1 GB) are visualized and analyzed interactively (~ 20 s). These large datasets will most likely be dispersed across the United States and will be made available using a client/server interface. To facilitate efficient data transfer, parallel I/O will need to be investigated and made routine between computational computers, data repositories, and visualization systems.

Simulation data sharing will include new complex visualization capability that is presently being developed. As stated above, the desire to perform more complex simulation between experimental plasmas is strong and it will be desirable for data from these simulation codes to be available in the 15-20 minute time window of the experiment. Eventually, there will be an integrated simulation of the plasma that can be compared to the plasma itself all between plasma pulses. Such large–scale simulations using computer resources and data repositories shared across the United States, combined with the eventual compression of this analysis into a 20 minute time window will place a severe strain on existing network capability.

In addition to the network bandwidth requirements described above, the nature of FES research also leads to requirements for advanced network services. As in other sciences, valuable resources such as computers, data, instruments and people are distributed geographically and must be shared for successful collaboration. In fusion, the need for real-time interactions among large experimental teams and the requirement for interactive visualization and processing of very large simulation data sets are particularly challenging. Shared tools and solutions are especially valuable – reducing problems with N × M or N! interactions to more tractable scales.

The apparently conflicting requirements for transparency and security in a widely distributed environment point up the need for efficient and effective services in this area. Central management of authentications (PKI or equivalent technologies) using "best practices" and providing 24 **x** 7 support is essential. Further, it is essential that the user authentication framework and operational environments are such that common policy may be negotiated among international collaborators in order to enable collaborations to span international boundaries and between application development and site security groups. Development of mutually agreed upon tools and protocols for resource authorization is equally important.

As fusion collaboratory activities grow, the needs for global directory and naming services will expand as well. A hierarchical infrastructure with well–managed "roots" can provide the necessary glue for many collaborative activities. Analogous to the Internet's domain name services, this infrastructure would give local resource managers needed flexibility while maintaining global connectivity and persistence. A global name service could even solve the longstanding problem in the field of variable name translation between codes or experiments. Distributed computing services for queuing and monitoring are also needed. These must be easy to configure and deploy and robust in operation.

The fusion applications described above will also require network quality of service (QoS) in order to provide guaranteed bandwidth at particular times or with particular characteristics. Such QoS will be required to tailor the network to match the time dependent demand requirements rather than maintaining sustained bandwidth. For experimental collaborations, low network latency with minimum jitter and very low packet loss are essential if instruments and experiments are to be controlled remotely. It is anticipated that the next generation of fusion experiments will be routinely operated by remote teams. Relatively small quantities of monitoring data must be passed back reliably and quickly. Moderately large quantities of intermediate data are produced by simulations in burst mode before a code's conclusion and should be made available as quickly as possible or even during the code run for computational steering. End to end performance is crucial and must include partners at universities, private companies or international sites. Real-time network performance monitoring and problem resolution tools are essential.

For example, the fusion community anticipates a computing–for–data–analysis model that involves moving data at data rates of approximately 500 Mbit/sec between 40 sites. This network traffic is periodic with large bursts (1–2 seconds out of 20 minutes, 8 hours a day, 5 days a week, 20 weeks a year). When the traffic appears on the network it requires guaranteed bandwidth, however the network could be used for other purposes during the remaining ~90% of the time. This traffic will flow between the sites of the major experimental participants and their many collaborating institutions.

Table E.1. Magnetic Fusion Energy Requirements Summary

Feature Time Frame	Characteristics That Motivate Advanced Infrastructure	Vision for the Future Process of Science	Anticipated Requirements	
			Network	Middleware
Near-term	• Each experiment only gets a few days per year - high productivity is critical • Experiment episodes ("shots") generate 200-500 Mbytes every 15 minutes, which has to be delivered to the remote analysis sites in two minutes in order to analyze before next shot • Highly collaborative experiment and analysis environment	• Real-time data access and analysis for experiment steering (the more that you can analyze between shots the more effective you can make the next shot) • Shared visualization capabilities		• PKI Certificate Authorities that enable strong authentication of the community members and the use of Grid security tools and services. • Directory services that can be used to provide the naming root and high-level (community-wide) indexing of shared, persistent data that transforms into community information and knowledge • Efficient means to sift through large data repositories to extract meaningful information from unstructured data.
5 years	• Gbytes generated by experiment every 15 minutes (time between shots) to be delivered in two minutes • Gbyte subsets of much larger simulation datasets to be delivered in two minutes for comparison with experiment • Simulation data scattered across US • Transparent security • Global directory and naming services needed to anchor all of the distributed metadata • Support for "smooth" collaboration in a high-stress environments	• Real-time data analysis for experiment steering combined with simulation interaction = big productivity increase • Real-time visualization and interaction among collaborators across US • Integrated simulation of the several distinct regions of the reactor will produce a much more realistic model of the fusion process	• Network bandwidth and data analysis computing capacity guarantees (Quality of Service) for inter-shot data analysis o 500 Mbits/sec for 20 seconds out of 15 min-utes, guaranteed • 5 to 10 remote sites involved for data analysis and visualization	• Parallel network I/O between simulations, data archives, experiments, and visualization • High quality, 7x24 PKI identity authentication infrastructure • end-to-end QoS and QoS management • Secure / authenticated transport to ease access through firewalls • Reliable data transfer • Transient and transparent data replication for real-time reliability • Support for human collaboration tools
5+ years	• Simulations generate 100's of Tbytes • Next generation experiment: Burning Plasma	• Real-time remote operation of the experiment • Comprehensive integrated simulation	• QoS for network latency and reliability, and for co-scheduling computing resources	• Management functions for network Quality–of–Service that provides the request and access mechanisms for the experiment run time, periodic traffic noted above.

Appendix F

Chemical Sciences

Appendix F Chemical Sciences

David A. Dixon, Pacific Northwest National Laboratory
Larry A. Rahn, Sandia National Laboratories

The chemistry community is extensive and incorporates a wide range of experimental, computational, and theoretical approaches into the study of chemical problems. Chemistry is one of the fundamental sciences on which many applications are built. There is extensive use of basic chemical measurement techniques in a wide range of areas including atmospheric measurements, geochemical measurements, combustion and chemical process measurements, and cellular observations. Computational chemistry covers a wide range of areas ranging from accurate calculations on small molecules/processes such as heats of formation of radicals and electron scattering to intermediate accuracy calculations for the study of large molecules, separation systems and catalysts, and ultimately to molecular dynamics simulations of complex systems such as biomolecules and materials. There is also an extensive effort to discover the details of chemical processes as they interact with the unsteady, and often, turbulent fluids that transport and mix the reacting species. Such studies are key to developing understanding that will enable predictive design of complex chemical processes such as combustion or chemical processing in industry. This research includes the production and mining of extensive databases from direct simulations of detailed reacting flow processes.

The scientific process described above leads to a data- and model-centric view of the communications between sub-disciplines working at different time and size scales. Data at one level is analyzed to develop a model that produces data used in turn by another, repeatedly across the range of scales and types of chemical information required. However, in this process more than just the raw data values need to be communicated. Confidence in a value's accuracy, its uncertainty, dependencies on other data, etc. must all be considered when using it in further computational and experimental research. Enabling the rich bi-directional exchange of both data and metadata between scales is a critical issue in making progress.

To overcome current barriers to collaboration and knowledge transfer among researchers working at different scales, a number of enhancements must be made to the information technology infrastructure of the community:

- A collaboration infrastructure is required to enable real-time and asynchronous collaborative development of data and publication standards, formation and communication of inter-scale scientific collaborations, geographically distributed disciplinary collaboration, and project management.

- Advanced features of network middleware are needed to enable management of metadata, user-friendly work flow for web-enabled applications, high levels of security especially with respect to the integrity of the data with minimal barriers to new users, customizable notification, and web publication services.

- Repositories are required to store chemical sciences data and metadata in a way that preserves data integrity and enables web access to data and information across scales and disciplines.

- Tools now used to generate and analyze data at each scale must either be modified or new translation/metadata tools must be created to enable the generation and storage of the required metadata in a format that allows interoperable work flow with other tools and web-based functions, and must be made available for use by geographically distributed collaborators.

- New tools are required to search and query metadata in a timely fashion, and to retrieve data across all scales, disciplines, and locations.

The complexities of managing information within such an infrastructure are daunting and the creation, communication and use of the additional information could quickly become unwieldy. However, recent technological advances in middleware (such as DAV [1]) and the development of the extensible markup language (XML) for defining machine and human readable metadata based on standard schema, have significantly reduced the barriers to creating such a comprehensive informatics environment.

The chemistry community is extensive and incorporates a wide range of experimental, computational, and theoretical approaches to the study of problems including advanced, efficient engine design; cleanup of the environment in the ground, water, and atmosphere; the development of new Green processes for the manufacture of products that improve the quality of life; biochemistry for biotechnology applications including improving human health and cleanup; and the use of all of these to improve Homeland Security. The advanced computing infrastructure that is being developed will revolutionize the practice of chemistry by allowing us to link high throughput experiments with the most advanced simulations. Chemical simulations taking advantage of the soon-to-come petaflop architectures will enable us to guide the choice of expensive experiments and reliably extend the experimental data into other regimes of interest. The simulations will enable us to bridge the temporal and spatial scales from the molecular up to the macroscopic and to gain novel insights into the behavior of complex systems at the most fundamental level. In order for this to happen, we will need to have an integrated infrastructure including high speed networks, vast amounts of data storage, new tools for data mining and visualization, modern problem solving environments to enable a broad range of scientists to use these tools, and, of course, the highest speed computers with software that runs efficiently on such architectures at the highest percentages of peak performance possible.

F.1 Reference

1. Web-based Distributed Authoring and Versioning, WebDAV. http://webdav/org

Table F.1. Chemical Sciences Requirements Summary

Feature / Time Frame	Characteristics That Motivate Advanced Infrastructure	Vision for the Future Process of Science	Anticipated Requirements	
			Network	Middleware
Near-term	• High data-rate instruments running for long times producing large data sets • Greatly increased simulation resolution- data sets ~10–30 terabytes • Geographically separated resources (compute, viz, storage, instmts) & people • Numerical fidelity and repeatability • Cataloguing of data from a large number of instruments • Large scale quantum and molecular dynamics simulations	• Distributed multi-disciplinary collaboration • Remote instrument operation / steering • Remote visualization • Sharing of data and metadata using web-based data services • Computing on the net by linking large scale computers	• Robust connectivity • Reliable data transfer • High data-rate, reliable multicast • QoS • International interoperability for namespace, security • Large scale data storage needed both for permanent and temporary data sets. Can the network serve as a large scale data cache?	• Collaboration infrastructure • Management of metadata • High data integrity • Global event services • Cross discipline repositories • Network caching • Server side data processing • Virtual production to improve traceability of data • Data Grid broker / planner • Cataloguing as a service
5 years	• 3D Simulation data sets 30–100 terabytes • Coupling of MPP quantum chemistry and molecular dynamics simulations for large scale simulations in chemistry, combustion, geochemistry, biochemistry, environmental studies, catalysis • Validation using large experimental data sets • Analysis of large scale experimental data sets including visualization and data mining	• Remote steering of simulation, e.g., control of the time step, convergence of the SCF, introducing a perturbation in an MD simulation • Remote data sub-setting, mining, and visualization • Shared data/metadata w annotation evolves to knowledge base	• 10s of gigabits for collaborative visualization and mining of large data sets	• Remote I/O • Collaborative use of common, shared data sets – version control on the fly • International interoperability for collaboratory infrastructure, repositories, search, and notification • Archival publication
5+ years	• Accumulation of archived simulation feature data and simulation data sets • Multi-physics and soot simulation data sets ~1 petabyte • Large-scale MD simulations – 100s of terabyte to petabyte datasets	• Internationally collaborative knowledge base • Remote collaborative simulation steering, mining, visualization	• 100+ gigabit for distributed simulations – computational quantum chemistry, molecular dynamics, CFD combustion simulations	• Remote collaborative simulation steering, mining, visualization

Appendix G

Bioinformatics

Appendix G Bioinformatics

Michael Wilde, Argonne National Laboratory

The field of computational biology, in particular that of bioinformatics, has undergone explosive and exponential growth since the first gene sequencing work emerged in the mid 1980's. This field offers immense opportunities to improve the quality of human life, notable in disease prevention and cure; in the global distribution of the food supply; and in the remediation of environmental hazards. Of late, the field has become painfully vital to national security.

While progress in computational biology has for the most part, been limited till recently by our understanding of biological processes, our ability to model them, and our ability to organize information and develop algorithms, progress in this area has advanced with explosive rapidity in recent years. The field is now transitioning to a stage where algorithmic progress has outpaced computing capabilities in terms of raw compute cycles, storage, and in particular, fast, secure and usable information discovery and sharing techniques. These have now increasingly become the factors that limit progress in the field.

As biological computing pushes the envelope in these areas, the resulting enhancements that this pressure will drive have the potential to pay dividends in many commercial areas, both within the life sciences and in the much broader field of business informatics, intelligence, and workflow automation. This makes investment in tailoring research networks to the needs of bioinformatics both necessary, and one of immense potential benefits to the nation's productivity and economy.

Near-term applications that dominate today's computing requirements in bioinformatics include: Genome Sequence analysis, pairwise alignment, computational phyolgenetics; coupling of multiple model levels to determine metabolic pathways, secondary database searching, etc.

On the more distant research horizon, research areas include: sequence-structure-function prediction, computation of the genotype-phenotype map, protein folding; molecular computing; genetic algorithms; artificial intelligence solutions that will require real-time harnessing of Grid resources for large-scale parallel computation.

While the field and hence the networking requirements of computational biology obviously have much in common with other areas of computational science discussed in this report, they differ and stand out from the other application areas in the aspects described in the remainder of this section. We note that some of these differences are of a quantitative nature, while others are qualitatively unique to the characteristics of the information bases and algorithms that make up the field.

Size of the community. The large number of researchers involved in computational biology is out pacing that of almost any other biomedical science: The Higher Education Funding Council for England (HEFCE) shows (for 1997/98) 4000 faculty members in the life sciences, approaching that of clinical medicine and dentistry (4500) and far exceeding that of physics (1500) [http://www.ja.net/conferences/research_workshop/JGoodfellow.pdf]. This suggests that the user community sizes of successful computational biology Grids may exceed even that of the HEP community

as Grids are increasingly leveraged in this field. This necessitates highly effective solutions to authentication and authorization for Grid access; policy-based control and sharing of Grid resources; and automated management of individual logins at large numbers of Grid sites. The community needs to accelerate research and development of solutions like the Globus Community Authorization Service and virtual organization (VO) membership management and dynamic account creation and mapping facilities.

Several national and international Grid communities are forming specifically to support bioinformatics activities. To name a few: the North Carolina BioGrid [http://www.ncbiogrid.org/], the Michigan Center for BiologicalInformation [http://www.ctaalliance.org/MCBI/BioinfoArch.html] and the Japanese J-grid [http://www.gridforum.org/Meetings/GGF4/Speaker_Pres/ggf4-jpgrid-present-020219.pdf]. That national and international research communities will need to construct virtual organizations that span Grid providers; since these Grids have vastly different funding sources (ranging from state funds in NC and Michigan to foreign funds in EU and Japan), resource allocation policies, and charging mechanisms. Grid resource and data sharing mechanisms that span these boundaries will need to be created. Accounting, sharing, and charging mechanisms will need to be developed, not only for traditional resources like CPU and storage space, but, increasingly, but for network bandwidth as it starts to be allocated for very wide area data transport in units of transiently dedicated lambdas.

Nature of the data. More so than in other fields, bioinformatics in particular is dominated by heavily symbolic rather than quantitative data, which requires highly diverse data models, and makes far heavier use of large scale relational databases than most other sciences (although high energy physics is rapidly approaching this solution as well). This necessitates high quality end-to-end solutions for database integration and federation, an issue of datatype and identifier standards; security rules and models. A powerful example of this approach is SRS, the Sequence Retrieval System from the European Bioinformatics Institute [http://srs.ebi.ac.uk/]. Such federation activities are both focused on the provision of multi-database integration for search, matching, etc. (like SRS) or on specific multi-level systems model integration (such as in more speculative ventures such as the Digital Human or Tree of Life projects). These latter projects require not only sophisticated database federation and integration, but also a tight coupling of search and analysis tools and multi-level simulations and models.

Location of the data. The requirement to leverage large clusters and Grids of powerful computing resources is common to all of the applications areas that we discuss in this report. The need for a level of computing resources by a diverse and distributed user community that can only be satisfied by Grid solutions directly motivates the need for high performance transfer and replication of the data needed to sustain such parallel computations across wide-area resources. In the field of bioinformatics, this data is increasingly resident in relational databases, which requires the high performance replication of all or portions of databases whose size is now growing roughly according to Moore's law – doubling approximately every 18 months. While genomic databases of the past decade were sized in gigabytes, todays databases are now pushing terabytes, with petabyte applications well within view. This is a daunting challenge to for the management and in particular for the transport of relational data.

Performing Grid computation on relational data will not only require the integration of heterogeneous databases to form world-wide federations of unprecedented scale, but for database replicas to be accurately maintained and synchronized with high integrity as huge amounts of data is exchanged.

Databases have traditionally suffered from slower network transfer rates than that of flat files, due to their structured nature. End-to-end performance of the import-export and interchange of relational data needs to be addressed, as well as support for data replication models with varying degrees of performance in integrity assurance (single master, multi-master, etc.). Thus, significant research will be required in distributed database replication, potentially supported by high speed multi-cast; this overlaps and is served by research into Grid-wide data caches. Grid-wide database mining applications, as well as powerful search and discovery engines that operate over the worldwide federation of public genome information may necessitate research into other aspects of the data access patterns required for such applications.

To fully support the use of open source components across the community, development will be needed to make such high-speed transfer and replication mechanisms available to, and interoperable with, both public domain databases like PostgresQL and MySQL, as well as proprietary databases such as Oracle, Informix, and DB2.

We close this topic by noting that today, a large amount of bioinformatics data is shared over the web. As this approach grows in usage, the community will need to adopt website scaling, content caching, and load balancing techniques that begin to resemble those of the largest e-commerce sites, but with the difference that what is being shared is both database and tool access. Network research will be required to bring techniques popularized by commercial tools fully into the open source domain, and to adapt those techniques to the database and tool interfaces of the bioinformatics community.

Tools for analysis and integration. Scripting languages such as PERL have proven to be powerful integrators of distributed data and tools within the Bioinformatics community. [L.Stein; Bioinformatics Ch 17]. These tools need to be augmented by library modules that make it easy to allocate and connect Grid resources and distributed databases, and achieve high-speed connectivity and multicast communications with relative ease between Grid compute sites and databases. Seamless and easy integration of such scripts with tools, data sources and storage services based on the Open Grid Services Architecture is essential in order to take the next steps to further application integration and increase analysis capabilities. This approach has the potential of enabling scientists to create tools of immense power, which grows rapidly as the lower layers of functionality are successively set in place with easy to use abstractions that encapsulate powerful capabilities. This level of powerful composition of new capabilities from existing services can only be achieved in a Grid environment if platforms such as OGSA are used to provide universal discovery and interoperability of these components.

Increasingly, the techniques employed in computational biology have reached the point of sophistication and computational intensity that a large amount of effort will need to be expended on algorithmic engineering [see, e.g., *Towards new Software for Computational Phylogenetics*, Moret et al.; Computer 35, 7 pg 55-64]. While current efforts are applied mainly to computational efficiency on a single computer, in the future similar efforts will be applied in a systems context to both local clusters and computational Grids. Such efforts on a Grid will increasingly involve wide-area parallel libraries like MPICH-G [http://www3.niu.edu/mpi]. A fruitful area of network research involves the efficient tuning of ultra-high speed connections such as those of TeraGrid and iWire to the needs of such applications for both short and long messages, and for both dynamic short-duration lambda allocation, and for predictable and efficient packet-based QoS mechanisms.

Collaboration support. One of the most important collaborative activities in Bioinformatics today is that of annotation. This is typically done in very localized and individualized settings, both manual and automated. Manual annotation would be greatly enhanced by the integration of multi-party messaging interaction technologies with database versioning techniques, possibly augmented by multicast with closely integrated file transport and visualization. To accomplish this, further development and integration work is needed in end-to-end file transport and database replication and synchronization, which in turn requires enhancements to network data transport protocols and QOS mechanisms.

Shared visualization and collaborative assessment of large-multi-tiled displays of metabolic pathways are being rendered by the Computational Biology group at ANL; Imaging systems are being used to study phenomics using the BioSig system from LBNL [Parvin et al.; Computer 35, 7 pp. 65-71]. The BioSig architecture involves the integration of computation, storage, and networking using a CORBA backplane; systems such as this would benefit immensely in terms of public accessibility and collaboration if they can be adapted to Grid resource sharing architectures such as OGSA. Future imaging systems for use in the life sciences will involve both shared exploration and annotation of ultra–high-resolution images by multiple distant collaborators, as well as high volume computationally intense pattern recognition, harnessing in real time the ability to transport large image data to computational sites for analysis and correlation with and update of distributed database federations.

Appendix H

Workshop Agenda

Appendix H Workshop Agenda

The Vision - Tuesday, August 13, 2002 Morning Plenary Session, Lake Fairfax		
Time	**Location**	**Event**
7:30 am - 8:45 am		**Continental Breakfast**
8:45 am - 9:30 am		**Welcome, Workshop Objectives, Provisional Strategy Overview** Networks for Science - W. Polansky Networks for Science Provisional Strategy - M. Scott
9:30 am - 11:00 am		**High Impact Applications** (presentations in three areas – 20 minutes each, break from 10:10-10:30 am) **HENP - LHC experiments and other collaborations** High Energy and Nuclear Physics - C. Young **BES - Advancing Chemical Sciences** Advancing Chemical Science: Future Networking Requirements - D.A. Dixon and L.A. Rahn MSCF: HP & OBER's Flagship High Performance Computing Facility - D.A. Dixon **BER - Genomes to Life** BioGrid Models for Genomes to Life and beyond. - P. LoCascio and E. Uberbacher
11:00 am - 11:35 am		The Future of Advanced and Innovative Networks--What is possible? Bill St. Arnaud
11:35 am - 12:10 pm		Vision for a Science Environment Rick Stevens
12:10 pm - 1:15 pm		**Lunch**
Afternoon Parallel Breakout Sessions 1:15 pm - 5:00 pm (30 minute break around 2:30 pm) *Breakout groups to consider requirements which come from high impact applications that drive network capabilities including middleware and network research--output is a set of tables to be used by breakouts set for the second day. Topics to be addressed include network capacity, middleware, major connections to non-DOE facilities, etc.*		
Time	**Location**	**Event**
	North Point	**Breakout Group 1: Facilities Access and Collaboratories** Ray Bair, facilitator (with Deb Agarwal)
	Tall Oaks	**Breakout Group 2: Data Intensive Applications** Bill Johnston, facilitator (with Mike Wilde)
	Hunter Woods	**Breakout Group 3: Advanced Scientific Computing Facilities** Rick Stevens, facilitator

The Reality - Wednesday, August 14, 2002 Morning Plenary Session, Lake Fairfax		
Time	**Location**	**Event**
7:30 am - 8:45 am		**Continental Breakfast**
8:45 am - 9:30 am		Status of ESnet as a Model for the Future Jim Leighton
9:30 am - 10:05 am		TeraGrid and I-WIRE: Two Network Models Linda Winkler
10:05 am - 10:25 am		**Break**
10:25 am - 11:00 am		Optical Fiber Options: Creative Business Models for Networking Ron Johnson
11:00 am - 11:35 am		Network Research: What is Possible in the Near Term, Mid-Term and Long-Term Mari Maeda
11:35 am - 12:10 pm		Reality Talk on Grid Middleware - Challenges and State of the Art, What is Routine, What is Possible Now and Within 1-2 Years Dennis Gannon
12:10 pm - 1:15 pm		**Lunch**
1:15 pm - 2:00 pm		**Summarize Day 1 Breakout Discussions** Report from Breakout 1 Report from Breakout 2 Report from Breakout 3
Afternoon Parallel Breakout Sessions 1:15 pm - 5:00 pm (30 minute break around 2:30 pm)		
Time	**Location**	**Event**
	North Point	**Breakout Group 4: Turn Applications Requirements into Middleware Capabilities/Research Needs** Ian Foster and Dennis Gannon, facilitators
	Tall Oaks	**Breakout Group 5: Turn Applications Requirements into Network Capabilities/Research Needs** Linda Winkler and Brian Tierney, facilitators
	Hunters Woods	**Breakout Group 6: Critique the High Level Strategy, Governance Model for the Decision/Prioritization Process, and Business Models in Context of the Strategy** Sandy Merola and Charlie Catlett, Facilitators

Path Forward - Thursday, August 15, 2002 Morning Plenary Session, Lake Fairfax		
Time	**Location**	**Event**
7:30 am - 8:45 am		**Continental Breakfast**
8:45 am - 9:15 am		Report out of Breakout Group 4
9:15 am - 9:45 am		Report out of Breakout Group 5
9:45 am - 10:15 am		**Break**
10:15 am - 11:00 am		Report out of Breakout Group 6
11:00 am - 12:00 noon		**Closeout Discussion, next steps**
Afternoon		
		Leaders Complete Draft Report

Appendix I

Workshop Participants

Appendix I Workshop Participants

Deb Agarwal
Lawrence Berkeley National Laboratory

Guy Almes
Internet2

Bill St. Arnaund
Canarie Inc.

Ray Bair
Pacific Northwest National Laboratory

Arthur Bland
Oak Ridge National Laboratory

Javad Boroumand
Cisco Systems, Inc.

William Bradley
Brookhaven National Laboratory

James Bury
AT&T

Charlie Catlett
Argonne National Laboratory

Daniel Ciarlette
Oak Ridge National Laboratory

Tim Clifford
Level 3 Communications, Inc.

Carl W. Cork
Lawrence Berkeley National Laboratory

Les Cottrell
Stanford Linear Accelerator Center

David Dixon
Pacific Northwest National Laboratory

Tom Dunigan
Oak Ridge National Laboratory

Aaron Falk
University of Southern California
Information Sciences Institute

Ian Foster
Argonne National Laboratory

Dennis Gannon
Indiana University

Jason Hodges
Oak Ridge National Laboratory

Gerald Johnson
Pacific Northwest National Laboratory

Ron Johnson
University of Washington

Bill Johnston
Lawrence Berkeley National Laboratory

Wesley Kaplow
Qwest

Dale Koelling
U.S. Department of Energy

Bill Kramer
Lawrence Berkeley National Laboratory
National Energy Research Scientific Computing Center

Maxim Kowalski
Thomas Jefferson National Accelerator Facility

Jim Leighton
Lawrence Berkeley National Laboratory/ESnet

Phil Locascio
Oak Ridge National Laboratory

Mari Maeda
National Science Foundation

Mathew Mathis
Pittsburgh Supercomputing Center

William (Buff) Miner
U.S. Department of Energy

Sandy Merola
Lawrence Berkeley National Laboratory

Thomas Ndousse-Fetter
U.S. Department of Energy

Harvey Newman
California Institute of Technology

Peter O'Neil
National Center for Atmospheric Research

James Pepin
University of Southern California
Information Sciences Institute

Arnold Peskin
Brookhaven National Laboratory

Walter Polansky
U.S. Department of Energy

Larry Rahn
Sandia National Laboratories

Anne Richeson
Qwest

Corby Schmitz
Argonne National Laboratory

Thomas Schulthess
Oak Ridge National Laboratory

George Seweryniak
U.S. Department of Energy

David Schissel
General Atomics

Mary Anne Scott
U.S. Department of Energy

Karen Sollins
Massachusetts Institute of Technology

Warren Strand
University Corporation for Atmospheric Research

Brian Tierney
Lawrence Berkeley National Laboratory

Steven Wallace
Indiana University

James B. White III
Oak Ridge National Laboratory

Vicky White
U.S. Department of Energy

Michael Wilde
Argonne National Laboratory

Bill Wing
Oak Ridge National Laboratory

Linda Winkler
Argonne National Laboratory

Wu-Chun Feng
Los Alamos National Laboratory

Charles C. Young
Stanford Linear Accelerator Center

Prepared for the Office of Advanced Scientific Computing Research
of the U.S. Department of Energy Office of Science

http://www.sc.doe.gov/ascr/