

DOE Office Of Science

High Performance Network Planning Workshop

Background:

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 un-sharable a decade ago, are now routinely exchanged and remotely analyzed 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 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.

- o The present 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 where resources are not sufficient to continue being responsive to all of the needs.
- o There is an increasing awareness that to realize the vision, the end-to-end bandwidth problem must be solved, not just the backbone bandwidth.
- o 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.
- o 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 of wide area networks.

The first step in developing a new strategy for the vision given below is to understand the scientific potential of greatly increased network capabilities for the major applications drivers within the Office of Science. How can the process of science change if available bandwidth is not a limiting factor and middleware services are available to facilitate the routine construction and use of widely distributed science environment? At what level of connectivity and bandwidth do these elements cease to be a problem? This is important to quantify these levels. Once this is done, then we can address the issue of what network provisioning, middleware service development and deployment, and network research needs to be done to enable these approaches over the next five to ten years. The process of defining a vision of science unfettered by communication limitations will be initiated in advance of the workshop through the development of scenarios depicting the vision of where science application would like to be in five to ten years in this circumstance. In some cases this will be radically different from how the science is currently done.

This workshop, scheduled for August 13-15, 2002, is the first major activity in developing a strategic plan for high performance networking in the Office of Science. It will bring together end users, especially representing the emerging, high visibility initiatives, and network visionaries to identify opportunities and begin defining the path forward.

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 that move science, especially large scale science, to a new regime— a regime in which seamless collaboration between 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.

Approach:

- Develop scenarios (using the HEP example from last year's interagency Large Scale Network workshop) for where the high impact science applications want to be in five to ten years – this may be done by the discipline scientists alone or working with networking/ middleware researchers to develop a vision.
 - HENP
 - Climate
 - Computational Biology
 - Nanoscience

- Chemical Sciences
 - Fusion
 - ...and others
- The scenarios are input to identifying research areas (network and middleware)—partially in advance, partially during the workshop—and to evaluate provisioning strategies.
 - Develop alternative business plans within the context of the scenarios and general needs —partially prior to workshop, complete afterward – possibly have three views presented at the workshop (from the most conservative to the most innovative)

Overall Workshop Purpose/Objective:

The provisional strategic approach contained in Appendix A will be used as a starting point. The workshop will examine its environment and dynamics, evaluate the strengths of the approach, the weaknesses of the approach, the external opportunities that might support the approach and should be taken advantage of, and the risks for the provisional strategy that should be taken into account.

Workshop Output:

A comprehensive report that discusses the Findings and Recommendations for the following.

- Science and high-speed networking scenarios
 - Characteristics of the science that motivate the need for high-speed networking
 - A vision of the process of the science over the next decade given high-speed networking
 - The anticipated requirements for networking and middleware services to achieve the vision
- A provisional strategy
 - Elements of the strategy
 - Evaluation of the ability of the strategy to address the vision of highly connected science
 - The risks and tradeoffs of the strategy

Report Outline

Chapter 1. High Impact Science--Drivers for the Future Network Infrastructure.

What do the high impact DOE/SC applications need over the next five to ten years? *(Include all scenarios and discuss derived requirements.)*

- a. Production requirements
- b. Requirements for Application Testbeds and Middleware

Chapter 2. Application-enabling Network Research

What are the network research requirements over the next five to ten years that are driven by high impact applications?

- a. Requirements driven by production
- b. Requirements driven by application testbeds and middleware

Chapter 3. Roadmap for Production, Experimental Application Testbeds, and R&D Network Infrastructure

High Level Strategy for providing Network Infrastructure to Support DOE/SC Science *(Consider the environment and dynamics for the provisional strategic approach, evaluate the strengths of the approach, the weaknesses of the approach, the external opportunities that might support the approach and should be taken advantage of, and the risks for the provisional strategy that should be taken into account.)*

- a. Multi-tier model
- b. Business models
- c. Governance model to prioritize needs

Potential invitees:

Total attendance target is about 50 people

- About 10-15 representatives of high impact applications
- About 8-10 network researchers, half from outside DOE labs
- About 8-10 individuals providing networking both in and outside the DOE complex
- About 8-10 middleware/grid researchers
- About 10 agency program managers

Appendix A: Provisional Strategy

The model for high performance networking infrastructure has the following components.

- A high bandwidth production network that provides quality networking and services to support distributed science.
- An experimental network that provides a pilot environment where new capabilities for high end science applications can be developed and emerging network technologies can be deployed.
- A research network where network technology five or more years out can be explored.

This three-pronged model for networking infrastructure is one that currently appears or is planned in a number of communities, mainly a result of externalities driving it in that direction. Two additional elements are important to realizing the vision.

- Advanced services, a suite of shared middleware services, must be available for enabling distributed science applications.
- DOE science programs must identify the high impact projects that set the scientific networking priorities and there must be a program-based mechanism for prioritization.

Production Network

The strategy for providing the production network has evolved over the past two decades as more and more DOE researchers are effectively served by non-DOE networks. As we move forward, it is essential that we continue to examine strategies for focusing resources on network needs that are unique to DOE, partnering opportunistically with other network providers where there are common interests (for example, connecting a given university).

Objectives:

- Production network
- Advanced middleware services for distributed applications and collaborations

Drivers:

- Science
- End-to-end performance

Characteristics:

- Rich connectivity to DOE collaborators
- Highly reliable and available
- High performance

Experimental Network

Objectives:

- Integrate very high performance distributed science applications with very high bandwidth networks and advanced services
- Decrease the time to move successes in network R&D to support for bandwidth-intensive science projects

Drivers:

- Very high bandwidth, distributed science applications
- Distributed systems middleware R&D
- Network R&D
- Network security R&D

Characteristics:

- Network and application innovation are encouraged
- Ties to academia and the commercial sector enhanced
- Stable, but production quality

Research Network

Objectives:

- Explore and influence future network technology

Drivers:

- Capacity and capability increases needed by next-generation science applications and facilities

Characteristics:

- Opportunistically determined
- Experimental applications
- Limited reliability

Appendix B: High Energy Physics Scenario generated for the Workshop on New Visions for Large Scale Networks¹

Never before has the scientific mission of particle physics research been so dependent on state-of-the-art information technology. Collaborations of hundreds to thousands of physicists and engineers are formed to create accelerators, detectors, and analysis systems with a productive life of tens of years. These analysis systems form a complex and widely distributed “fabric” of computing and storage resources.

The non-deterministic nature of quantum physics, uneasily understood during the last century, inevitably requires the measurement and analysis of billions of particle interactions to observe and understand fundamental processes. Particle physics experiments have pushed against the limits of technology, electronics, computing, and networking for decades. Detectors with millions of channels, each recording precise amplitudes with a resolution of picoseconds, have in the course of 40 years succeeded detectors with a few single-bit “yes/no” measuring devices. Information flows from such a detector at up to a terabit per second and must be drastically filtered in real time because of limited storage, analysis, and networking facilities.

The Large Hadron Collider (LHC) experiments at the European Organization for Nuclear Research (CERN) will rapidly reach tens of petabytes of stored data under intense analysis. The design, construction, and data analysis for an experiment require the combined intellect and dedicated work of international collaborations. However, technological limitations on the storage, transmission, and analysis of data impose difficult, even dangerous choices. For example, the LHC experiments expect to be able to record and share over networks less than one millionth of the collisions they observe. This draconian real-time selection will necessarily have to be optimized for “somewhat expected” new discoveries rather than the “totally unexpected” ones that are the dream of every scientist.

Even after the draconian selection, LHC collaborations will face the challenge of empowering thousands of geographically distributed physicists to use their intellect and wisdom to derive physics insight from tens of petabytes of data. Although the raw cost of bandwidth is no longer a crippling impediment, the end-to-end performance of applications is often unacceptable. Success will rely on middleware research to support data-intensive, worldwide collaborative science that is only beginning. A minimum requirement is the location-independent ability to analyze data to empower all of an experiment’s physicists to work collaboratively on databases, growing to tens of petabytes in 2005-2010, using all computing resources to which they have access.

However a qualitative change in the way research is performed would be enabled if we could free the real-time selection of data from the constraint of a single filter system with

¹ “Workshop on New Visions for Large-Scale Networks: Research and Applications”, Large Scale Networking (LSN) Coordinating Group Of the Interagency Working Group (IWG) for Information Technology Research and Development (IT R&D), March 12-14, 2001, Vienna, Virginia

selection algorithms decided by committees. The availability of networks with end-to-end terabit performance could make this possible, but speed alone is not enough.

The data acquisition and filtering systems might profitably become geographically distributed and operate as highly parallel, largely asynchronous data flow systems. The terabit systems that will become operational in 2005 for the LHC will include a multi-terabit capacity switching fabric, but individual data acquisition nodes and filtering nodes will communicate at gigabit speeds. In addition to the substantial bandwidth requirement, challenges include:

- The multicast service required when more than one remote filtering center is available.
- Achieving adequate error rate and robustness without *ever* allowing the implementation of the “wild idea” to impact the detector-site data acquisition system

Based on this scenario, the participants in the workshop generated the requirements given in Appendix C.

Appendix C: Excerpt from “Workshop on New Visions for Large-Scale Networks”²

2.6 High-Energy Physics

High-energy physics (HEP) has pushed against the limits of networking and computing technologies for decades. Twenty years ago, the largest HEP experiment involved 100 physicists from many nations and acquired tens of thousands of magnetic tapes of data per year; graduate students spent months reading those tapes to perform data queries. Life is not so different for today’s physicists. The new BaBar detector at the Stanford Linear Accelerator (SLAC) was designed by a large international collaboration of physicists at institutions. The BaBar collaboration enables hundreds of physicists worldwide to query its 300-terabyte and rapidly growing database in hours or days rather than months. In the next 10 years, the Large Hadron Collider (LHC) experiments at CERN, the European Physics Laboratory, where some 600 U.S. physicists form the largest national group, will face the challenge of distributed analysis of hundreds of petabytes of data.

2.6.1 High-Energy Physics Scenario

The physics community greatly values being able to distribute digitized data electronically at the rate at which it is produced from the site of an experiment to collaborators worldwide who can analyze them. The HEP community has the goal of using affordable network and computational resources to provide physicists with transparent access to a distributed data analysis system that uses all available resources as efficiently as possible. By 2005 to 2010, HEP computing will involve queries on databases containing exabytes (10^{18} bytes) of data structured as up to 10^{16} individually addressable objects. These massive amounts of data will require the distribution of terabits per second of real-time data to major HEP data analysis centers.

Challenging networking and other information technology research needed to enable distribution of data, analysis, and collaboration includes:

- Multicast service delivered to multiple remote centers with diverse firewall filters
- Network error rate and robustness control *without* impacting the experiment’s data acquisition system
- Massive applications software – e.g., 3 million lines of BaBar C++ code
- Commercial object database management software
- Interfaces of the database with the network and storage
- Technology improvements including:
 - Computing technologies
 - Computer science
 - Networking

² “Workshop on New Visions for Large-Scale Networks: Research and Applications”, Large Scale Networking (LSN) Coordinating Group Of the Interagency Working Group (IWG) for Information Technology Research and Development (IT R&D), March 12-14, 2001, Vienna, Virginia

- Computing system-to-network interfaces
- Fiber technologies
- Data storage

Improvements in HEP applications must be accomplished at minimal incremental costs. To help contain costs, network engineering labor, required to configure, optimize, and maintain networks, should be minimized by developing automated network engineering and management.

HEP collaborations are increasingly international in composition. It is difficult to adopt standards across the resulting international boundaries, so that the implementation of uniform, collaboration-wide middleware, security, or hardware technologies is almost always unrealistic. The best that can be achieved is the adoption of a set of protocols and interfaces to link components that will almost certainly be implemented in different ways.

The international HEP research community is increasingly using Grid technologies, an integrated suite of services developed with Federal IT R&D funding. The Grid is a set of middleware tools and capabilities that enable seamless end user access to applications, data storage, and compute resources to support high-end modeling. The Globus project (<http://www.globus.org/>) is one state-of-the-art example of Grid development. Grid middleware faces many hard computer science problems. Vertical integration of existing components to provide Grid services to demanding, well-defined communities is essential to progress on Grid architecture and technologies.

2.6.2 High-Energy Physics Networking and Networking Research Needs

Networking underlies many of the services and applications being developed to support HEP. Progress in networking is expected to be evolutionary over the next five years, with revolutionary capabilities being developed over the longer term. The following table presents the current state of the art in various networking areas supporting HEP, what could evolve by around 2006, and the requirements to approach meeting the HEP goals. The current HEP capabilities are what is affordable, not what could be obtained with unlimited funding.

| Current HEP Capabilities | Evolution to 2006 | HEP Goals |
|--|--|---|
| <u>Links Between Major Centers</u> • 1 or 2 x 1.55 Mbps | • 10 Gbps | • 1 Tbps |
| <u>Bulk Transfer Protocol</u> • TCP/IP + fixes | • TCP/IP + more fixes | • New, <i>widely adopted</i> , transport protocol |
| <u>Differentiated Services (CoS, QoS, Mixture of Packet and Circuit Switching, etc.)</u> • Provide ~1.1 differentiated services (best effort + some Voice over IP (VOIP)) | • Provide ~2 differentiated services | • Provide ~6 differentiated services that are application-negotiated, on-demand, and responsive to cost and policy |
| <u>Network Measurement, Analysis, Interpretation, and Action and Network Modeling</u> • Limited measurement, analysis, and modeling | • More/better measurement and analysis, and some interpretation • Models begin to predict non-obvious failure modes | • Automated measurement, analysis, and interpretation • Automated action based on measured and modeled information |
| <u>Support for Collaboration</u> • Some proof-of-concept (PoC) prototypes • Some commercial tools | • New PoC prototypes • Some mature components • Still incomplete | • Collaborations form via the Internet • Real sense of working together |
| <u>Data-Grid: Authentication and Authorization</u> • Local and manual | • Cross-authentication via proxies | • New approaches to regulating access to resources |
| <u>Data-Grid: Information Infrastructure (Replica Catalog, Resource Catalog, Software Catalog, Operation/Task Catalog, etc.)</u> • Manual and local • Limited <i>ad hoc</i> automation | • Evolution of Globus by 2+ generations | • Efficient distributed information management for more than 10^{16} virtual objects using millions of operations each using millions of lines of code (MLoC) |
| <u>Data-Grid: Data Payload Infrastructure (Exabyte Databases, Reliable Replication, Storage Management, etc.)</u> • Few x 100-Tbyte databases • PoC replication prototypes • PoC storage management | • Bleeding-edge 10^{19} Byte databases • Grid replica management • Grid storage management | • Industry-standard exabyte databases, replication, and storage management |
| <u>Data-Grid: Resource Discovery</u> • Telephone, e-mail | • Telephone, e-mail, partial automation | • Automated discovery • Standardized information models |
| <u>Data-Grid: Distributed Resource Management, Distributed Job (Task, Operation) Management</u> • Local batch systems • Prototype systems | • Grid job management • Early distributed resource management | • New approaches to regulating access to resources |
| <u>Data-Grid: Virtual Data</u> • Conceptual phase | • Starting to work for cutting-edge HEP experiments | • A generally accepted and implemented paradigm |
| <u>The Grid an Integrated "Network" Service</u> • Manually integrated services have been in use for more than 10 years | • Vertical integration of fabric and data payload services • Incomplete information services • Incomplete resource management services | • Easy creation of vertically integrated, worldwide information management and processing systems from standard industry components |

Notes:

1. HEP technologies that work well locally but do not become widely adopted and supported may inhibit collaboration and prove costly. Qualifiers like "widely adopted, industry-standard," and "generally accepted" are vitally important.
2. Elegant approaches to authentication and authorization appear to be available for organizations that are part of a single administrative structure. Worldwide collaborations seem unlikely ever to fit this model. Discussion identified that a totally new approach to regulating access to resources might foster more open scientific research.

The Role of Industry in HEP Networking R&D

Wherever possible, high-end science takes advantage of capabilities that are developed and commercialized by industry. For example, the HEP community has benefited from cost reductions and reliability increases provided by industrial commercialization of individual middleware components, such as databases and well-defined information systems. Also, the HEP community has benefited from the availability of commercial high-end computing systems, high-bandwidth networks, and extensive middleware. It is likely that higher bandwidth will be more affordable in the future due to economies of scale, greater supply, and competition among providers. Carriers are beginning to make individual wavelengths available to major customers. Affordable links between major HEP computer centers should exceed 10 Gbps within five years and may approach 1 Tbps in less than a decade. However, it is likely to be difficult to exploit the available bandwidth using industry-standard transport protocols. TCP/IP requires fixes such as multiple streams to use today's affordable bandwidth. Additional fixes will be needed to accommodate the expected increases in numbers of users, number of nodes, and network traffic. It is possible to develop a new protocol or to extend TCP to work over dedicated links, but the extensive investment of industry and users in the current protocols would likely hinder acceptance of alternatives.

Workshop participants identified a need for a vertically integrated HEP solution for managing and processing the massive amounts of data expected from HEP experiments. Networking research, development of faster computing systems and more capable computational algorithms, and commercial development and marketing (productization) together deliver components that provide part of this vertically integrated HEP solution. New component technologies emerging from networking research and computer science are funded normally only to the proof-of-concept stage and fall short of the level of product hardening and support needed to provide technologies that can be reliably integrated into a complex operational system. Collaboration by network researchers, computer scientists, and application scientists required to provide vertical integration of the component capabilities are also research and development and, in the view of the workshop participants, should be funded by the Federal IT R&D funding agencies.

The HEP community is rapidly taking advantage of the Grid infrastructure to enable transparent, distributed, and international collaborations, resulting in improvements in the ability to cooperatively carry out science and to analyze increasingly large volumes of HEP data. However, the Grid primarily has been developed in universities and industry is currently largely decoupled from development of an integrated Grid capability. Thus, Grid software and infrastructure have not benefited from the standardization, cost reductions, and increased reliability often provided by commercial productization. This productization will take place only if industry perceives the potential for profitably marketing the technologies. Federal funding could help bridge the gap between the proof-of-concept prototype and the point at which successful vertical integration has demonstrated commercial viability.