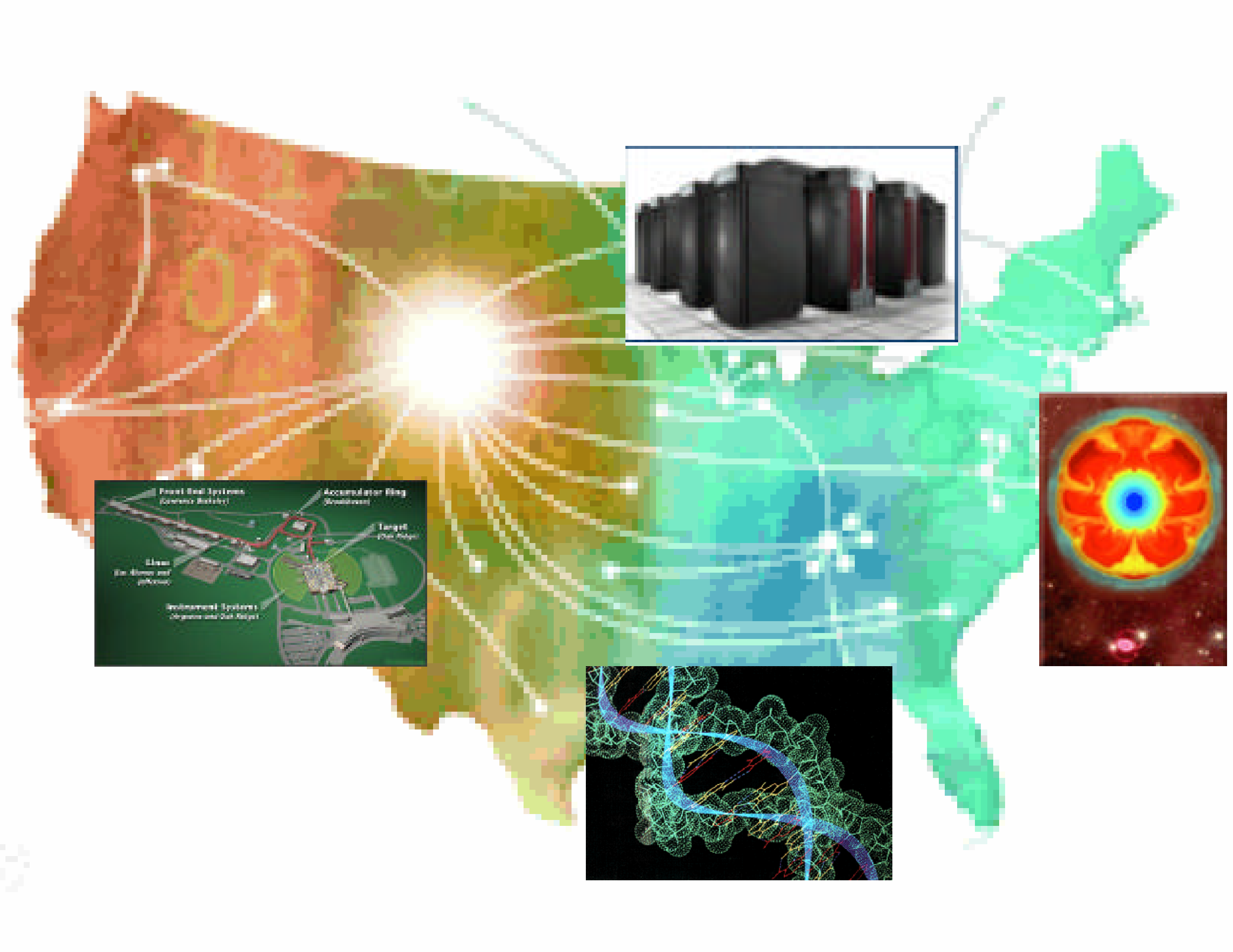
Network Provisioning and Protocols for DOE Large-Science Applications

October 27, 2003: Draft Copy: Not for circulation

Office of Science
U.S. DEPARTMENT OF ENERGY



Report of DOE Workshop on
Ultra High-Speed Transport Protocols and Dynamic Provisioning for
Large-Scale Science Applications
April 10-11, 2003

http://www.csm.ornl.gov/ghpn/wk2003.html

Network Provisioning and Protocols for DOE Large-Science Applications

Report of DOE Workshop on
Ultra High-Speed Transport Protocols and Dynamic Provisioning for Large-Scale Science Applications
April 10-11, 2003, Argonne, IL

Organizing Committee

Nageswara S. Rao
William R. Wing

Working Group Chairs

Transport Group: Wu Feng, Don Towsley
Provisioning Group: William R. Wing, Biswanath Mukherjee

Workshop Contributors

William E. Allcock
Ray Bair
Micah Beck
Gee-Kung Chang
Steve Cortez
Roger Cottrell
Tom DeFanti
Thomas H. Dunigan
Wu Feng
Dennis Ferguson
Mark Gardner
Robert Grossman
Glenn Heinle
Wesley K. Kaplow
James Leighton
Steven Low
Matthew Mathis
Mark Meiss
Biswanath Mukherjee
Bill Nickless
Nageswara S. Rao
Allyn Romanow
Daniel Stevenson
Raymond Struble
Don Towsley
Malathi Veeraraghavan
Jay Wiesenfeld
William R. Wing
Matthew Wolf

Sponsored by
High-Performance Networks Program
Division of Mathematics, Information and Computer Science
Advanced Scientific Computing Research
Office of Science
US Department of Energy

Table of Contents

Executive Summary

The next generation DOE scientific breakthroughs critically depend on large multi-disciplinary and geographically dispersed research teams, wherein the network has become an integral part of the science infrastructure much like the supercomputers and experimental facilities. Such DOE large-science projects span the disciplinary spectrum including high energy physics, climate computations, fusion energy, genomics, astrophysics, spallation neutron source, and others. These projects are inherently distributed in resources including the data, computations, personnel, or experimental facilities, and consequently, their effectiveness critically depends on a *seamless networking* of these resources. Such capability demands revolutionary advances in network technologies for tasks such as Petabyte data transfers at Terabps speeds, computational steering, interactive and collaborative visualization, and remote instrument control. Furthermore, these capabilities must be provided to the general science users and not just to the network experts with a privileged access to special networks. These requirements, however, will not be met by commercial and other agency networks because of the small user base and extremely high cost, and are also significantly beyond the evolutionary paths of current network technologies. But, if these network capabilities will not be available, several DOE large-science projects may fail to reach their potential, perhaps, with a negative impact on US science leadership.

Recent advances in the areas of network transport and provisioning hold an enormous promise in meeting these network requirements. In the network provisioning area, optical switching and communications technologies have seen significant advances, which provide several building blocks for flexible, agile and configurable bandwidth pipes or channels. Optical links capable of extremely high bandwidths over thousands of miles are becoming increasingly possible. To harvest these capabilities, however, the next generation provisioning technologies must be developed to realize ultra-high capacity switched channels with dynamically specified end-to-end requirements. These provisioning technologies must be integrated into a scalable architecture for a fast and on-demand setup of channels at various bandwidth resolutions between the application nodes. In the network transport area, the current methods are massively inadequate for sustaining throughputs at multi Gigabps levels or controlling the end-to-end delay dynamics. While the network experts can currently achieve several Gigabps for certain durations, such throughputs are mostly unavailable to the application users. Also, there are very few methods that provide stable low-jitter transport to support control operations needed for user facilities over wide-area networks. Focused efforts are needed to develop transport methods to exploit the underlying provisioning capabilities to meet the requirements of Terabps throughputs as well as stable and agile controls. To address the fundamental aspects of these challenges, a comprehensive and foundational theory of high-performance networking would be needed based on a synergy and extensions of several disciplines including stochastic control, non-linear control, statistics, optimization, and protocol engineering.

The developmental efforts of both provisioning and transport technologies require an application-centric test-bed with the capability for on-demand switched cross-country channels together with a testing environment for new protocols. The existing test-beds and simulators are inadequate to provide the realistic operating conditions required by these high performance networks. In addition to developing the individual high performance provisioning and transport technologies, it is equally important to integrate them with the applications, middleware and operating systems, and particularly with the legacy and evolutionary networks. Furthermore, the test-bed must facilitate a smooth transition of the provisioning and transport technologies into operational networks and application environments.

This workshop is one among a series of workshops conducted by DOE Office of Science in addressing the networking needs of large-science applications. The participants constituted a balanced mix of experts from universities, national laboratories and industries, representing the network provisioning and transport areas as well as large-science applications.

1. Introduction

The large-scale U. S. Department of Energy (DOE) science projects of the next generation will increasingly depend on close collaborations of multi-disciplinary researchers dispersed across the country or around the globe. Such collaborations collectively represent capabilities unavailable at any single national laboratory or university. Furthermore, these projects span a wide spectrum of disciplines including high energy physics, climate computations, fusion energy, genomics, astrophysics, and others, which are of large interest to DOE. These collaborations invariably involve geographically distributed resources such as supercomputers and clusters that offer massive computational speeds, user facilities that offer unique experimental capabilities, repositories of experimental and computational data, and human experts with deep and broad knowledge in technical areas. Of particular importance are the new experimental facilities coming on-line such as the spallation neutron source (SNS), and the relativistic heavy ion collider (RHIC), which present unprecedented opportunities and challenges for distributed and collaborative remote experimentation and data analysis. The ability to remotely perform the experiments and then transfer the large measurement datasets can significantly enhance the productivity of scientists and facilities. In general, a seamless access to the distributed resources by the researcher teams is essential to carry out the DOE large-scale science missions:

> *Indeed, the "network" has become a critical component of the modern infrastructure for large-scale science, much like the supercomputers or experimental facilities.*

The above networking capabilities add a whole new dimension to the access of these computers and user facilities, thereby eliminating the "single location, single time zone" bottlenecks that plague these valuable resources.

Advances in high-performance networks hold an unprecedented potential in realizing these network capabilities, thereby expanding the impact of a number of DOE large-science computations and experiments. Such networking opportunities together with the potential benefits to various science areas have been identified in the DOE network planning workshop that took place in August 2002 [1], and have been repeatedly highlighted in other DOE workshops and conferences [2,3]. In June 2003, a roadmap has been formulated for the DOE networks, which envisions a seamless, high-performance network infrastructure to facilitate collaborations among the researchers and their access to remote experimental and computational resources [2].

The next generation DOE large-scale science projects and programs have requirements that will drive extreme networking. Some of these requirements involve massive (Petabyte sized) data transfers across the country and around the world. In other cases they involve distributed collaborative visualization, remote computational steering, and remote instrument control. These requirements place different, possibly mutually exclusive, demands on the network. The network capabilities required to support this scale and range of networking activities surpass, by several orders of magnitude, the performances achieved by today's leading-edge high-bandwidth networks. In summary, a main conclusion of the workshop is that:

> *An ultra high-performance network with powerful and flexible provisioning and transport modalities is needed to meet the demands of the DOE large-scale science applications.*

The field of ultra high-speed networking is currently at a critical crossroads with no clear evolutionary path to eliminate the performance gap that exists between the link speeds and application throughputs. While the optical technologies promise links at Terabps (Tbps) the corresponding *provisioning* and *transport* technologies needed to deliver this performance on-demand to the applications are severely lacking. The widely deployed Transmission Control Protocol (TCP) transport mechanisms do not scale to these unprecedented optical bandwidths in terms of application throughputs. While the commercial demand for faster backbone networks will continue to improve the link speeds based on optical networking technologies, the lack of such demand at the application-level will prevent the development of the required mechanisms including protocols and components. Consequently, with the advent of multiple Gigabps (Gbps) routers and switches, the end-to-end bottleneck has moved from the core network to host systems and end components, which are often outside the priorities of service providers.

This workshop is a foundational step in identifying the critical networking technologies for DOE large-scale science projects and programs. To keep it manageable, this workshop concentrated only on two key areas

? *Dynamic Provisioning of Ultra High-Speed Channels*

? *Protocols for Ultra High-Speed Transport*

which pose major challenges to realizing the required ultra high-performance networks. Two working groups consisting of experts from industry, national laboratories, and universities, identified the limitations of the current technologies, and formulated research and development activities in the respective areas. The overall conclusion is that the current technologies in both areas are significantly inadequate, and only through focused efforts in the design, analysis, testing and deployment, the required capabilities could be provided to DOE large-science communities. *And without the focused efforts to develop such network capabilities, the above DOE large-science needs will simply be not met*. Considering that a large number of current and planned large-scale scientific computations are DOE projects, it is appropriate and imperative for DOE to take a leadership role in developing such network capabilities.

The workshop is a follow-on effort to the planning workshop [1] with a specific technical agenda to investigate deeper into the aspects of provisioning and transport. Other follow-on workshops may be planned to cover other possible areas such as cyber security, optical network components, and wireless networks.

In this report we briefly discuss the networking requirements of DOE large-science applications in Section 2 to highlight their needs and scope. Section 3 discusses various details of the workshops including the composition of working groups. Sections 4 and 5 are devoted to the main technical topics of this workshop, namely provisioning and transport, respectively. In each section the problem space and basic issues are described briefly, followed by the recommendations of working groups in terms of topics of interest in respective areas. The development of the required network technologies warrants a science of high-performance networks described in Section 6.1. The transport and provisioning technologies must be transparently integrated into the applications, and such issues are described in Section 6.2. Furthermore, these technologies can be efficiently tested using powerful test-beds that support close interactions with the applications, and these aspects are discussed in Section 6.3. Although originally not intended, the cyber security aspects were discussed in Section 6.4 due to their increasing and often very intrusive impact on the provisioning and transport methods.

2. Advanced Networking Requirements for DOE Large-Scale Science

2.1 Ultra-Scale Science Environments

The DOE large-science applications are quite varied in terms of their network requirements in part due to their disciplinary origins, which are as diverse as earth science, high energy and nuclear physics, astrophysics, fusion energy science, molecular dynamics, nanoscience, and genomics. The networking requirements for some of these areas are listed in Table 1. In this section, we describe these needs only briefly to highlight their general nature, and a detailed account of a number of DOE large-science applications and their networking requirements can be found in [1].

Science Areas	Current End2End Throughput	5 years End2End Throughput	5-10 Years End2End Throughput	General Remarks
High Energy Physics	0.5 Gbps E2E	100 Gbps E2e	1.0 Tbps	high throughput
Climate Data & Computations	0.5 Gbps E2E	160-200 Gbps	*n* Tbps	high throughput
SNS NanoScience	does not exist	1.0 Gbps steady state	Tbps & control channels	remote control & high throughput
Fusion Energy	500MB/min (Burst)	500MB/20sec (burst)	*n* Tbps	time critical transport
Astrophysics	1TB/week	N*N multicast	1TB+ & stable streams	computational steering & collaborations
Genomics Data & Computations	1TB/day	100s users	Tbps & control channels	high throughput & steering

Table 1. Network requirements for DOE large-scale science applications.

Many DOE applications rely on high-performance heavy-lift data transport that requires an optimal combination of network provisioning and transport protocols. For example, the network requirements of High Energy Physics (HEP) data transport applications are unprecedented: they must deliver hundreds of Gbps throughputs between two applications in near future and several Tbps within the next decade. In contrast, some other applications could require several concurrent channels for tasks to be cooperatively performed over wide-area networks by experts distributed at various national laboratories and universities. These tasks could range from cooperative remote visualization of massive archival data through the distribution of large amounts of simulation data, to the interactive evolution of computations through computational steering. In the case of remote visualization, the data must be rendered and presented on-line to various participant sites with different end-devices ranging from visualization caves through high-end workstations to personal desktops. Furthermore, the control of such visualization streams may have to be handed back and forth among the sites, while maintaining a smooth response of the distributed rendering engine. Details of two specific example applications with their requirements are provided in Appendix A.

2.2 Supercomputing and High-Performance Networking

Supercomputers are among the most vital components of the infrastructure for large-scale science. The continual increases in the computational speeds of supercomputers enabled unprecedented simulations, computations and explorations in several areas such as climate, genomics and astrophysics. Currently, the Japanese earth simulator provides a peak execution speed of 37 teraflops. Within a few years supercomputers with speeds in excess of 100 teraflops are expected to be available. Such computational speeds present unprecedented challenges to the networks both in terms of moving the massive amounts of generated data as well as in interactively steering the ultra high-speed computations running on them.

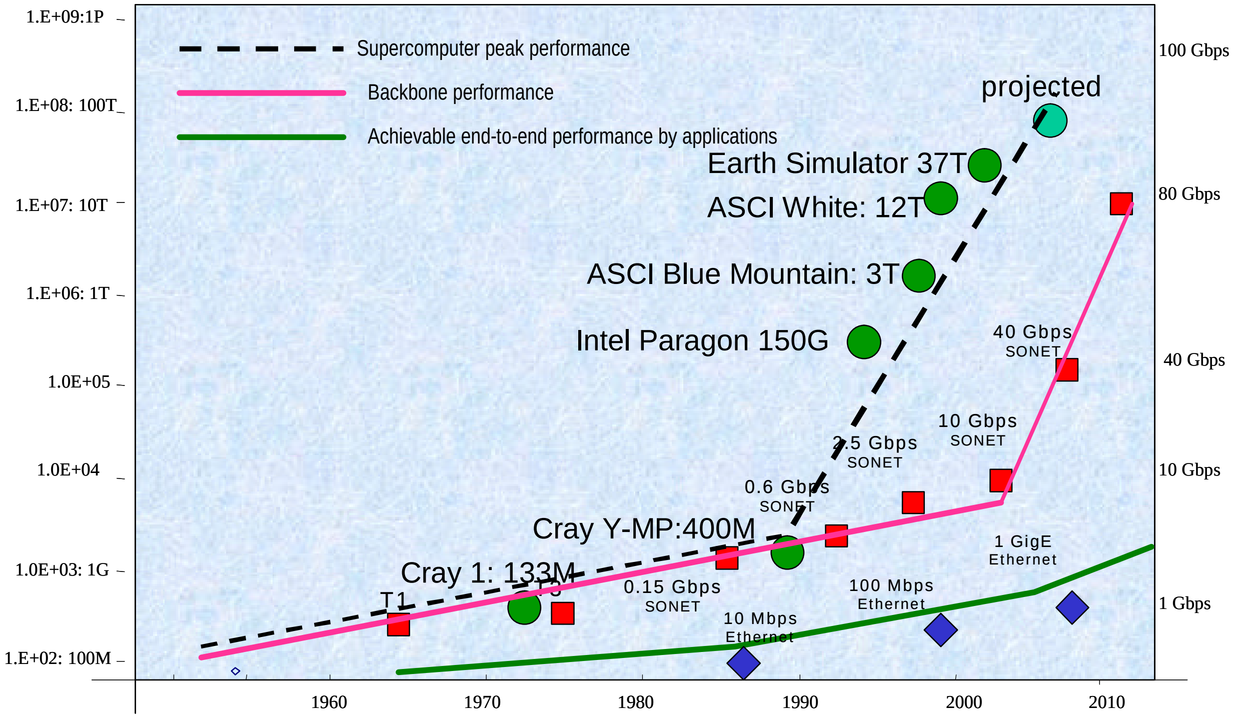


Figure 1. Supercomputer computational speeds consistently outpaced the network speeds.

Historically, the network speeds have been consistently outpaced by the computational speeds of supercomputers as shown in Fig. 1. Furthermore, the speed mismatch between the computational speeds of supercomputers and throughputs of their network connections continues to grow, often isolating the former from the wide-area remote access. Consequently, the supercomputers are often restricted to mainly local use and/or batch jobs. Note that in Fig 1 the computational speeds are on a log scale and networks speeds are on (almost) linear scale, which represents an ever widening performance gap. If this situation is not addressed, it is likely that:

(a) massive amounts of data generated and/or required by the computations will not be transported in a timely manner, thereby choking or idling the computations,

(b) batch executions could result in computations entering into undesired parameter domains due to the lack of active monitoring, thereby causing multiple reruns that waste the computational resources, and

(c) unstable control loops (typical of the current Internet) would result in “flops on the floor” phenomenon, wherein the supercomputers idle while waiting for control messages to arrive over the network.

Due to the critical role played by the supercomputers in several DOE large-scale science applications, it is particularly important to develop the network technologies capable of specifically addressing the supercomputing needs.

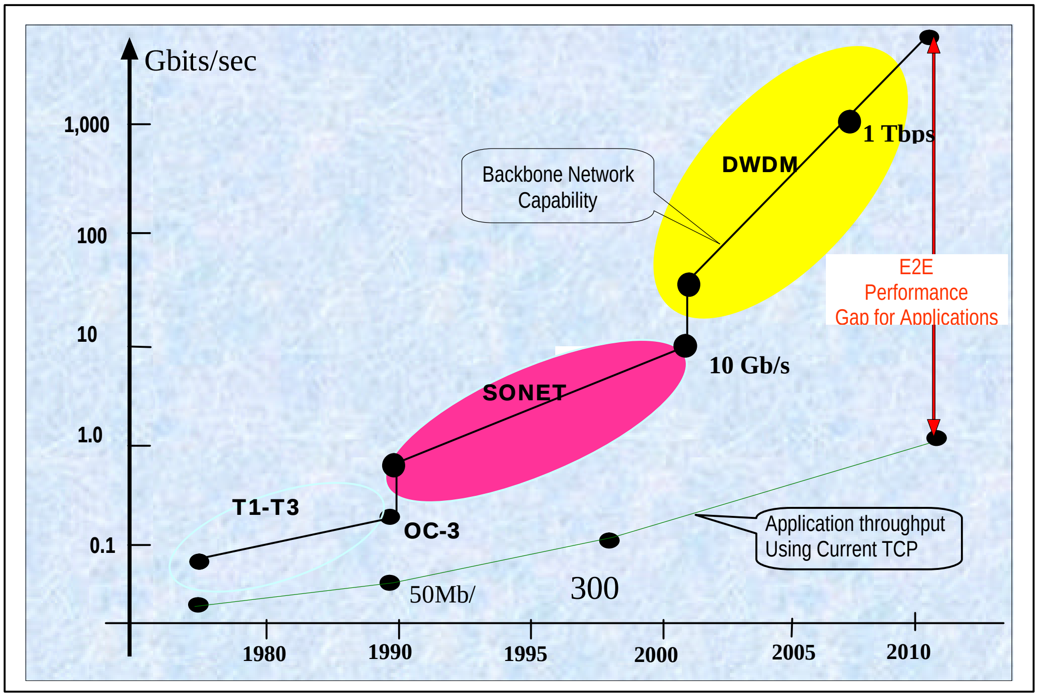


Figure 2. The end-to-end throughput continues to be small fraction the backbone speeds.

2.3 State of Networking for Large-Scale Science: As Assessment

There are several limitations of the current networks in meeting the requirements of DOE large-science applications in the areas identified in previous sections. While the needed transport speeds are currently available at the backbone links based on *dense wavelength division multiplexing* (DWDM) technologies, several architectural and design factors of provisioning, TCP stacks, network interface cards, and related software, currently limit the typical application throughputs to less than 1 Gbps as shown in Fig. 2. Experts in the field agree that sustaining multi-Gbps throughputs at the application level will not be achieved by simply replacing the existing links with ultra faster ones. For instance (to give a dated example), when the OC3 (150 Mbps) backbone was upgraded to OC12 (600 Mbps), the typical application throughput improved only marginally (25-50%) instead of the expected fourfold. Indeed, it took several years of protocol tuning and enhancements to reach 300Mbps throughput at the application-level. A similar fate awaits the simple-minded approach of just replacing the current links with OC-768 (40 Gbps) links or other high-speed optical links. In fact harnessing the abundant backbone bandwidth to provide it to the applications will require new advances in host system as well as network components, including transport protocols, network optimized system bus architectures, and dynamic provision of high-speed optical links. This last item may appear to be a non sequitur, but it follows directly from the fact that new transport protocols may demand segregated links on which they can run unimpeded, and this in turn, will indeed require the on-demand provisioning of those links.

To place current network limitations in perspective, consider a 4 Terabyte data transfer from North Carolina State University to ORNL, a typical daily output from a high-end supercomputer. Initial measurements show a bandwidth of 10 Mbps between the end nodes, which would thus require about 40 days to transfer the data *if TCP can be optimized and executed continuously with very low losses*. The infrastructure has been upgraded to equip the end hosts with Gigabit Ethernet (GigE) cards and connect them via Internet2 and ESnet via Gbps connections. In reality, today's TCP implementations deliver only about 400 Mbps to the user under good conditions over wide-area Gbps connections. Thus, this scheme has a best-case transfer time of roughly a day, assuming suitable un-congested bandwidth exists. If the end nodes could be upgraded to 10 GigE cards and links can be upgraded to OC192, the best possible transfer time is about one hour, but only if the entire bandwidth is provided for this transfer and the hosts are equipped with the required transport and middleware modules. Hence, to support such transfers we require: (a) an infrastructure that provides dedicated channels of 10Gbps or higher, and (b) network technologies that can provide link bandwidths at the application level. In general the underlying infrastructure must be upgraded to meet the requirement. But, simply upgrading the network links is not adequate since current off-the-shelf transport protocols are not adequate to achieve throughput that match link rates. To highlight the Issues, consider an old example in Figure 3 that shows TCP throughput of a large data transfer over an OC12 link between ORNL and NERSC. The link rate is about 620 Mbps but TCP achieved only 20 Mbps after 50 seconds: initial losses prematurely terminated the slow-start, and the subsequently TCP spent most of its time in recovering as per the Additive Increase Multiplicative Decrease (AIMD) scheme. While this limitation can be easily rectified by employing parallel-TCP or adapting the TCP dynamics, it is archetypical of the issues that should be paid close attention to when the infrastructure is upgraded. In particular, the transport methods may not scale with the link rates and it is very important to optimize them to the specific infrastructure.

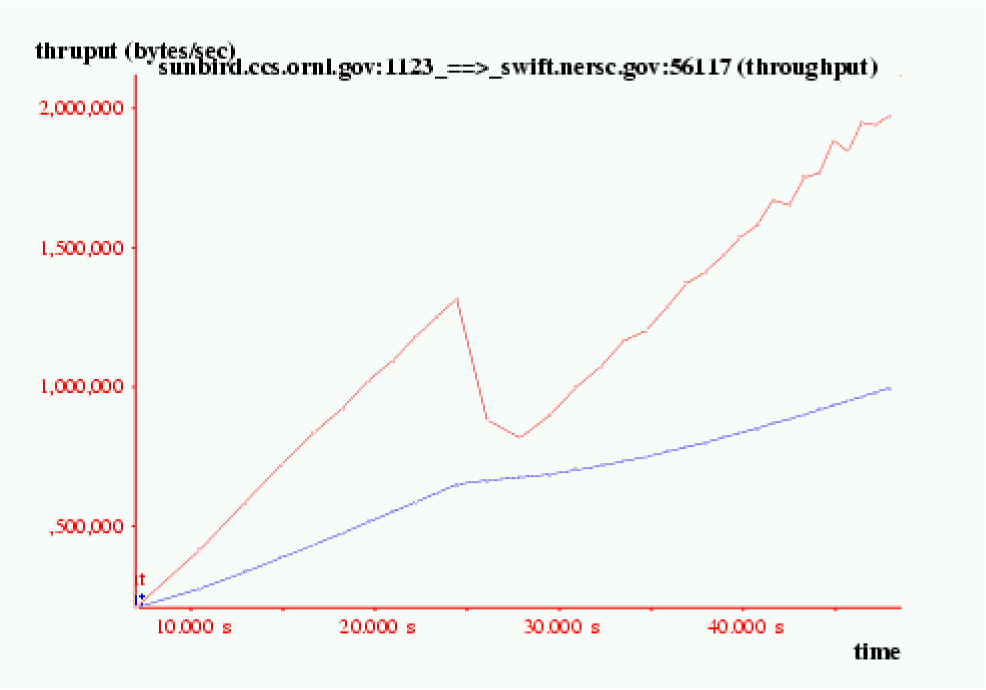


Figure 3. Premature termination of TCP slow-start severely limits the achievable throughput.

Collaborative visualization of dynamic objects does not need extraordinary amounts of bandwidth (30-50 Mbps is often adequate), but it does impose a different type of dynamic constraint on the throughput. That is, an interactive visualization stream can not wait through a normal "TCP slow-start" bandwidth ramp-up; it should be capable of starting and stopping in response to interactive requests in tens or at most hundreds of milliseconds, irrespective of the congestion levesl. In contrast, and as Figure 3 demonstrates, TCP can require tens or hundreds of seconds to achieve full speed. An interactive visualization that responded to requests for fast-forward, rewind, jog, and play with this sort of latency would be unacceptable. As another example, consider the remote control of an instrument (an electron microscope for example or neutron goniometer). Although it may seem obvious, remote instrument control requires a stable control loop. This in turn requires a tight control of the packet-arrival-time jitter, something TCP is famously unable to provide. To ensure smooth control of end devices, a computation or an experiment, it is important to send control messages quickly and without jitter. Jitter introduces high frequency components in the control signals that destabilize the control loops, and as a result, controlled objects (including devices, instruments, visualizations, and computations) may be damaged or driven into undesired regions. TCP is unsuited to support remote control loops due to its highly non-linear and abrupt dynamics in presence of even small amount of losses. In Figure 4, we show delay measurement of fixed-sized control messages sent at regular intervals from ORNL to University of Oklahoma. Indeed, TCP can be analytically shown to contain chaotic dynamics, which makes it very difficult to deploy it for supporting control-loops. Another complication arises over the Internet since the chaotic TCP dynamics are often mixed in with its response to the inherent randomness of traffic dynamics.

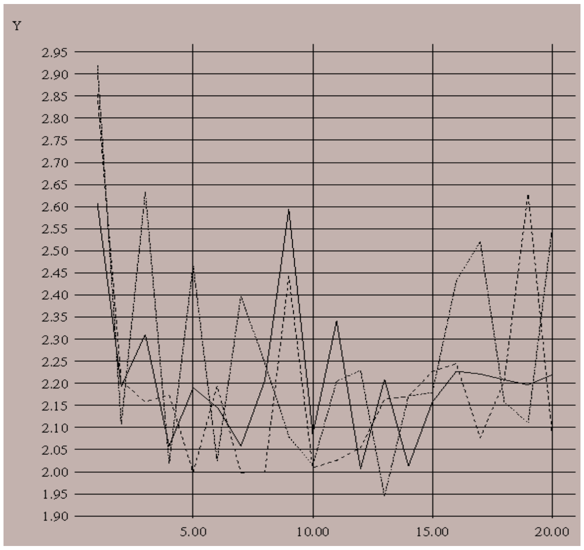


Figure 4. Delay measurements in seconds for fixed sized messages (10K) sent at regular intervals over the Internet

The main problems with TCP dynamics are due its congestion control part, which can be circumvented if the congestion is avoided altogether by using dedicated bandwidth pipes. Even so, TCP exhibits utilization problems due to the bandwidth unutilized within the “teeth” of its sawtooth profile for the congestion window; in such a case it will be more efficient to use a different class of protocols that incorporate certain TCP properties. Finally, even bandwidth requirements can be difficult to deal with when they are so large that it is not cost-effective to purchase and keep the bandwidth available 24 hours a day, seven days a week, 365 days a year. The approach of *dynamic provisioning* is an effective way of addressing this issue but such a capability requires newer ways of configuring the networks, arbitrating the bandwidth requests, setting up and tearing down of the dedicated channels, and matching the transport and middleware with the provisioned channels.

2.3 Networking Infrastructure for Large-Scale Science

In June 2003, a roadmap was formulated for DOE networks, which envisions a seamless, high-performance network infrastructure to facilitate the collaborations among researchers and their access to remote experimental and computational resources [2]. Such an infrastructure can eliminate resource isolation, discourage redundancy, and promote rapid scientific progress through the interplay of theory, simulation, and experiment. For example, a timely distribution of multi Petabytes of Hadron Collider data produced at CERN in Switzerland, can eliminate the bottleneck experienced by US physicists today due to inadequate bandwidth in the trans-Atlantic and US networks. Also the ability to access remote, complex, scientific instruments such as SNS or High Flux Isotope Reactor (HFIR) in real time will enable interactive collaborations among geographically dispersed researchers, without the need for coordinated travels and duplications of specialized experimental instruments.

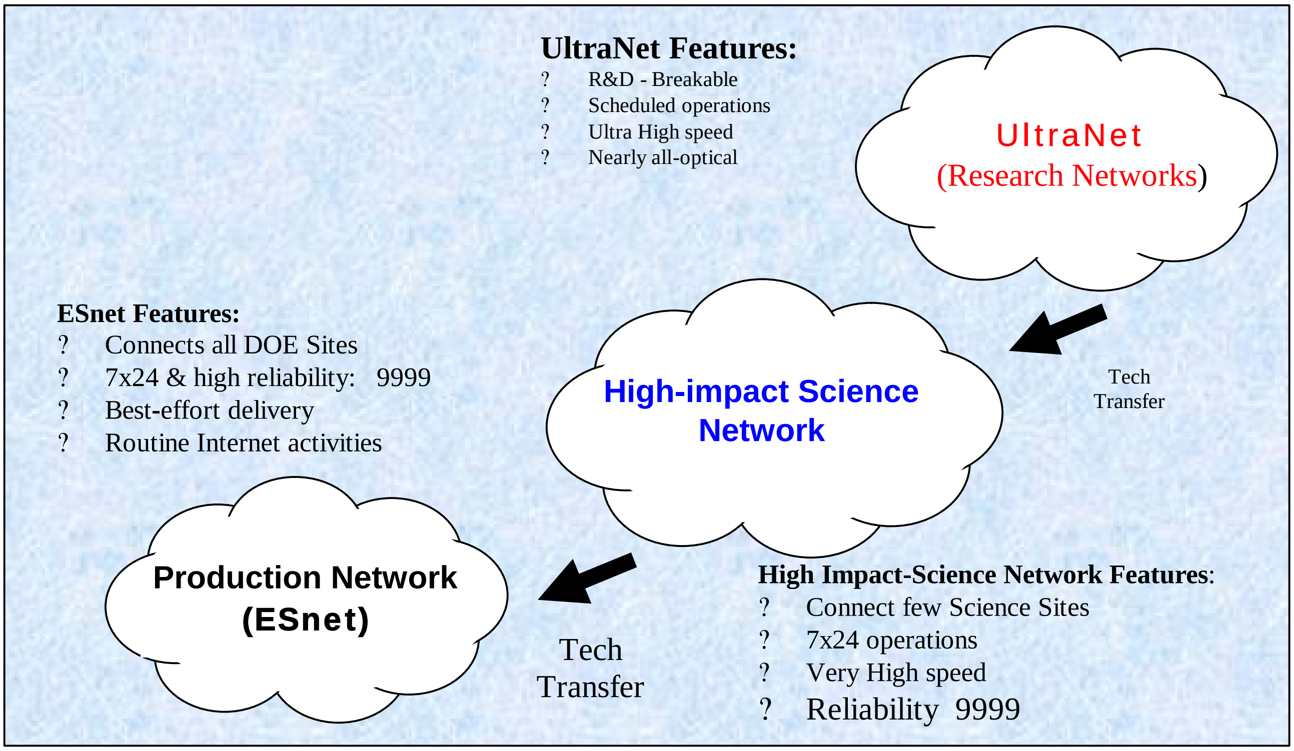


Figure 5. Paradigm for DOE networking for large-science applications and network research

Two important classes of high-performance networking capabilities are critical to a successful execution of the above tasks. First, large volumes of data must be *transported* to various end nodes over networks of disparate and varying capacities and traffic. Such data transports might be required in an off-line mode for data archival or post processing operations, or in an on-line mode for interactive visualization tasks. Second, the visualizations and computations must be *controlled* remotely over wide-area networks to ensure the responsiveness as well as the stability of control loops. This task requires that the higher-order moments of transport delays be kept suitably bounded: high levels of jitter in the control signals can destabilize and steer the remote process into unwanted regions. This problem is particularly acute when the computation is guided by a number of remote experts, each with a different process view, different parameters to control, and with different network connections.

The overall network requirements of DOE large-science applications range from the routine to extreme. The network capabilities to address the DOE large-science needs include the following:

1. Reliable and sustained transfers of terabyte scale data at Gbps to Tbps rates,
2. Remote interactive and collaborative visualization of large datasets of Petabyte scale,
3. Steering of computations on supercomputers and experiments at user facilities,
4. Interactive collaborative steering of computations visualized through multiple perspectives, and
5. Securing scientific cyber environments with minimal impact on applications.

In particular, it is *essential* that these capabilities be transparently available to the application scientists with little or no additional demands on their time and effort to utilize them. In particular, it is not very effective if these capabilities require sustained efforts from teams of network and application experts just to use them.

To adequately cover the broad spectrum of DOE large-science networking requirements, several network research areas have been identified at the workshop and are listed as follows:

? **High-Performance Data Transport**: For high performance data transfers there are two distinct approaches. At one extreme, TCP methods on shared Internet Protocol (IP) networks can be adapted and scaled to Gbps to Tbps rates. The challenges here include investigating various parts of TCP, such as sustained slow-start and robust congestion avoidance, to achieve the require throughputs. At the other extreme, one could provide dedicated high bandwidth pipes or channels from source to destination nodes wherein a suitable rate control method can be used for transport. In this method, both provisioning and transport methods must be developed (unlike the first method which can be executed on the current IP networks). Nevertheless, this approach circumvents the complicated problem of optimizing TCP congestion control by avoiding it altogether. Note that the network is still be shared (albeit not simultaneously) in this mode by allocating paths on-demand into time-slots for applications. In either case the networking modules must be suitably interfaced and integrated with the middleware and applications.

? **Stable Visualization Control Channels**: For supporting interactive visualizations over wide-area networks, two channels are needed: a *visual channel* transfers the image data from source to destination; and a *control channel* transfers the control information from user to visualization server. The former channel must provide appropriate sustained data rates to present adequate visual quality to the user, whereas the latter should provide low jitter to avoid destabilizing the control loop. There are several possibilities for implementing the visual channels: at one extreme transporting the geometry (for example as OpenGL codes) to be rendered at the user locations to rendering, and at the other extreme rendering at the host and just forwarding the visuals (for example using xforwarding). In a particular application, a combination might be required based on the bandwidths needed for data and visualization pipelines. In either case, the throughput should be sustained appropriately to maintain adequate visual quality. From the network transport viewpoint, both these channels require stable throughput, which can only be partially achieved over IP shared networks, that is in a probabilistic sense. On the other hand, they are easier to achieve if two dedicated channels can be provided on-demand. However, advances in both transport and provisioning methods would be required to achieve these capabilities.

? **Collaborative Steering and Control:** Agile transport protocols are needed to control remote computations. Typically a computation is monitored remotely, perhaps by visualizing a partial projection of the parameter space, and steered into regions of interest by interactively specifying the parameters. It is very important that the steering operations be supported on a robust channel to place the computation in an appropriate parameter region. Note that an inadequate control channel can result in undershoot or overshoot problems, wasting valuable computational resources, particularly on supercomputers. Also, undue delays in control messages to a waiting computation on a supercomputer could result in the "flops on the floor" phenomenon. The control problem is more acute in the remote control of experimental devices, where delays in control commands can result in severe damage. In an extreme case, high frequency components in jitter can result in resonance, which could lead to a complete loss of control. Furthermore, when the steering or control operations are to be performed by multiple users at geographically dispersed locations, the control channels must be suitably coordinated. Except for very simple steering and control operations, TCP on IP networks does not provide the desired stability levels. The approach based on dedicated channels together with associated transport methods must be investigated for this class of capabilities.

? **On-Demand Channel Allocation:** Provisioning of on-demand dedicated channels or bandwidth pipes requires allocation policies and implementations that are absent in packet switched IP networks. Requests for dedicated channels will be sent by the applications to *allocation servers*, which maintain the "state" of the network. Once the request is grated and accepted, *implementation servers* will setup the channels, maintain them for the allocated duration, and then tear them down. Such a capability does not exist over IP networks and must be developed for this class of DOE large-scale applications. Note that the allocation servers must be capable of implementing higher level policies for granting the requests as well as scheduling the channels by maintaining the state of available bandwidth levels of various network links. In addition, suitable routing and switching hardware and software must be in place to enable on-demand setup, maintenance and tear down of the various channels. In particular, it is important to be able to allocate the channels in groups and at various bandwidth resolutions, for example, a high bandwidth data channel together with a low bandwidth control channel.

? **Architecture and Infrastructure Issues:** Due to the highly exploratory nature of several components needed both in provisioning and transport areas, a number of new host and network architecture issues (that are not typical in Internet environments) must be investigated. To sustain ultra high data rates, OS-bypass, zero-copy, Remote Direct Memory Access (RDMA), and other non-conventional implementations for network technologies must be investigated to avoid the undue load on host processors just to support networking operations. Also, several computations and data generation operations might take place on clusters, and in some cases clusters might be used to generate aggregate streams at Tbps rates. To support such operations striping methods would be required to aggregate and separate the transport streams. It is also very important to provide ubiquitous monitoring and measurement capabilities as a part of the infrastructure to assist in diagnosis, debugging and performance optimization of various components.

3. Workshop Details

The stated goal of the workshop was to:

> "*address the research, design, development, testing and deployment aspects of transport protocols and network provisioning as well as the application-level capability needed to build operational ultra-speed networks to support emerging DOE distributed large-scale science applications over the next 10 years*".

It is to be emphasized that the required network capabilities must be available to the end-users most of whom are from various science areas. It is particularly desirable to minimize the demands on the users in utilizing the networking capabilities by providing suitable and transparent application interfaces. The following are the component tasks for the workshop in addressing the above goal:

1. Assess the viability of existing transport protocols in supporting ultra high-speed data transfers over very long distances for distributed Petabyte datasets.

2. Assess whether current core network technologies are adequate to meet the diverse network requirements of large-scale science applications.

3. Assess the role of industry and federally funded network research program in developing the advanced networking technologies that meet the needs of large-scale science applications.

4. Identify the major network technologies that need to be developed, enhanced or replaced in order to build and operate cost-effective networks capable of supporting distributed large-scale science applications

This workshop was not intended to be a general research workshop to address a wide class of Internet problems but only to specifically address the needs of DOE large-science applications. The participants were tasked, within their areas of expertise, to: (a) identify the major bottlenecks in meeting the large-science networking capabilities, (b) identify the critical technical topics and directions, (c) outline a roadmap to develop the required networking technologies to meet the performance requirements, and (c) identify critical interfaces and connections with other areas to deliver the end-to-end solutions to application users.

3.1 Network Provisioning and Transport Areas

Network provisioning and transport are two of the most critical areas representing the immediate and major bottlenecks in achieving the required networking capabilities. But at the same time they hold an enormous potential to contribute to these capabilities. A major objective of provisioning is to provide a lower layer capability to support high bandwidth on-demand and dedicated end-to-end channels. A major objective of the transport is to optimally utilize the provisioned channels to achieve stable and controlled ultra-high throughputs at the application-level. In order to tackle the above issues posed to the workshop attendees, two parallel groups were formed:

- ✍ **Dynamic Provisioning Technologies Group**: The objective is to develop recommendations in the provisioning area for dynamically reconfigured ultra high-speed channels to meet the diverse network requirements of large-scale science applications. This working group was chaired by Biswanath Mukherjee and William R. Wing. It consisted of fourteen members, 3 from national laboratories, 5 from universities, and 6 from industry.

✍ **Ultra High-Speed transport Protocols and Services Group**: The objective is to develop recommendations for protocols that can deliver and sustain multi-Gbps throughputs to the scientific applications. This working group was chaired by Wu Feng and Don Towsley. It consisted of fifteen members, 7 from national laboratories, 6 from universities, and 2 from industry.

3.2 Workshop Organization

This was a “working” workshop with focused discussions on very specific problems, methods, and potential solutions in the *transport* and *provisioning* areas. The workshop started with very short introductory presentations that identified the needs and the problem space. The rest of the workshop then consisted of meetings in two parallel tracks until the last joint session.

The participants provided a balance of expertise from universities, industry and national laboratories in representing the needs, technologies, research areas and business aspects. It was recognized that providing high-performance networking capabilities to the large-science application users requires the development of new and novel technologies. Furthermore, these technologies must be tested and deployed in production infrastructures. Hence it was essential that the participants as a whole represent a broad spectrum of research, academic and industrial viewpoints. There were altogether 32 participants.

? Ten from national laboratories:
  - o Oak Ridge National Laboratory (3), Argonne National Laboratory (2), Los Alamos National Laboratory (2), Pacific Northwest National Laboratory (1), Stanford Linear Accelerator Center (1), and ESnet (1).

? Eleven from universities:
  - o University of Massachusetts (1) , Georgia Institute of Technology (2), University of Virginia (1), University of Illinois at Chicago (2), Indiana University (1), University of Tennessee (1), University of California at Davis (1), Pittsburgh Supercomputer Center (1), and California Institute of Technology (1).

? Eight from Industry:
  - o Celion (1), Cienna (1), Cisco (1), Juniper (1), Level3 (1), Lightsand (1), MCNC (1), Qwest (1).

? Three from DOE Headquarters (two via access grid).

The provisioning working group consisted of 14 participants and the transport working group consisted of15 participants.

4. Workshop Findings in Dynamic Provisioning Area

Network provisioning generically refers to various aspects of a class of lower layer services of the (conventional) protocol stack to support applications. When dedicated channels are provided to the applications, however, the conventional view of the network protocol hierarchy is blurred, and additional middleware and/or transport modules are needed by the applications to utilize the provisioned channels. In general, the modes for the provisioned channels can range from the dark fiber at the lowest level through photonic switching, DWDM, Synchronous Optical Network (SONET), Generalized Multiprotocol Label Switching (GMPLS), to IP at the highest level. Depending on the mode of operation, the precise meaning of the provisioning varies; for example, it can stand for either a SONET link setup on-demand with a specified rate or a connection over conventional IP links with specified bandwidths. Based on the type of channel between two hosts, the provisioned path may be composed of different types of components such as switches, routers, service provisioning platforms, line cards, and Network Interface Cards (NICs). Typically, for IP networks such paths consist of Ethernet cards at the hosts connected to local hubs or switches which in turn are connected to router blades; the routers themselves can be connected through line cards to each other in various ways, for example via SONET or GigE links. Note that since channels can be provisioned under various modes, appropriate higher level mechanisms must be utilized to suitably expose and provide their functionalities to the transport, middleware and applications modules.

Optical networking technologies have seen significant advances recently both in terms of routers, switches and provisioning platforms as well as high bandwidth long-haul links. Because of the Internet demands, a good number of these components are targeted towards supporting the IP networks, in particular by providing high capacity backbones and faster connectivity to the end users. Several of these technologies can also facilitate the provisioning of dedicated channels at various levels. A number of flexible, agile and configurable routers, switches, and provisioning platforms are also becoming available both commercially as well as for conducting research and development activities.

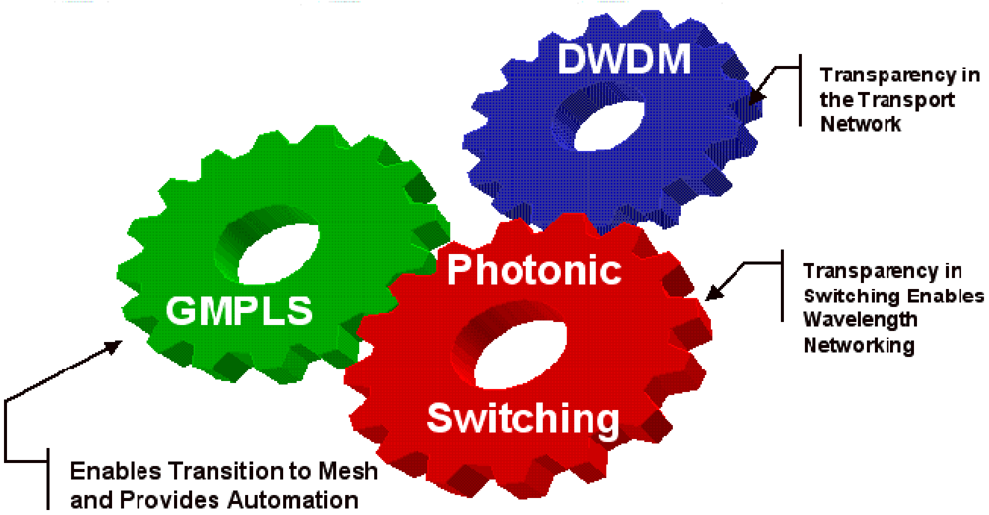

DWDM
Transparency in the Transport Network
GMPLS
Photonic Switching
Transparency in Switching Enables Wavelength Networking
Enables Transition to Mesh and Provides Automation

Figure 6. Lower level provisioning modes for dedicated channels.

Most current production networks, if not all, that support DOE large-science applications are based on IP. Since IP networks utilize packet switching over shared links wherein the packets are "queued" at the routers, there are certain inherent end-to-end characteristics of such connections. The queue occupancy levels depend on the competing traffic streams, and hence the "available" bandwidth levels and the packet delays. As a result, source nodes have very limited control either on throughput or packet delay, particularly under high traffic conditions. The implications are two-fold: (a) transport methods must account for the congestion levels and resultant losses to ensure reliable delivery, and (b) there are fundamental limitations on the packet delays and hence the end-to-end stability of the control channels implemented on such networks. In particular, large jitter levels can introduce high frequency components into control loops, which can seriously damage instruments at user facilities. Note that a number of DOE user facilities such as SNS are extremely expensive and should be safeguarded against such problems.

The dedicated channels provisioned between the hosts provide capabilities that are not normally possible in IP networks, such as the absence of competing traffic which altogether avoids the difficult congestion control problems. But to share the underlying network between the nodes, it is essential to circuit-switch the paths, preferably dynamically at various bandwidth resolutions and on-demand. Such channels can potentially simplify several design aspects of transport protocols since congestion control is no longer needed and the channel bandwidths are known. But channel utilization is a primary goal and must be explicitly incorporated into the protocols. Furthermore, they can also provide low jitter control channels since variations due to competing traffic streams are no longer present. Note however, that the delay measurements between the application modules that utilize such paths may still experience certain levels (albeit lower than Internet levels) of jitter due to the dynamics of other components such as NICs, provisioning platforms, SONET multiplexing and application modules. Furthermore, the provisioning of dedicated channels necessitates scheduling mechanisms to allocate the paths, and signaling mechanisms to setup and teardown the paths. While such circuit-switched networks may not necessarily be suitable for deployment of the scale of the Internet, they are still viable candidates for specialized deployments for connecting a small number of DOE large-scale science nodes.

4.1 Recommendations in Provisioning

To support DOE large-science applications, there is a need for next generation provisioning methods with the following capabilities:

Recommendation 1: *Agile Optical Network infrastructure*:

> A scalable architecture is needed that enables fast provisioning of circuit switched dedicated channels specified on-demand by the applications. This system accepts application requests and optimally allocates the bandwidths on various links. It then prepares and maintains a schedule of allocations. Then it utilizes a signaling mechanism to setup and teardown the paths as per the schedule.

Recommendation 2: *Hybrid Switched Networks*:

> High capacity (Tbps) switchable channels are needed to support Petabyte data transport, for example, in climate modeling applications. In particular, various provisioning modes must be supported so that applications requiring multiple channels with a combination of requirements can be effectively supported. Also, the capabilities of provisioned channels must accommodate burst, real-time streams as well as lower priority traffic. The channels must also provide for multi-point or shared use, for large file and data transfers, and for low latency and low jitter.

Recommendation 3: *Dynamically Reconfigured Channels*:

Provisioning of dynamically specified end-to-end quality paths must be supported so that channels can be dynamically reconfigured. In support of operations such as computational steering (for example, in genomics applications) and time-constrained experimental data analysis (for example, in fusion energy applications), multiple traffic streams may have to be supported for different periods of execution.

Recommendation 4: *Multi-Resolution Quality of Service:*

Channels with various types of Quality of Service (QoS) parameters must be supported at various resolutions using GMPLS, service provisioning and channel sharing technologies. The resolution levels could be quite varied and qualitatively different such as lambda, sub-lambda, various levels of OC-X, and IP-shared channels with specified total data rates. For example, pools of dynamically provisioned channels might be needed to support the collaborative visualization and steering operations in reconfigurable logical topologies.

Recommendation 5: *Experimental Test-Beds:*

Experimental research networks are needed to validate new ultra high-speed protocols and dynamic provisioning technologies. These aspects are discussed more in greater detail in Section 6.3.

Due to the leading edge nature of several dynamic provisioning components (as opposed to IP infrastructure components), the above recommendations are valid for the short term of next five years and a long term of ten years.

Typically, the application users will send their channel requests along with the performance parameters. These requests will be granted by a resource scheduler, which hands over the schedule to a signaling system. This system will then setup the paths as per the schedule by sending the signals to various switches and provisioning platforms, and will tear them down at the end of the allocated periods.

The provisioning technologies must be developed in close coordination with the developers of transport methods, middleware, applications and operating systems. They must be gracefully integrated with applications and middleware, and interoperate with legacy and evolutionary networks in appropriates cases; the latter is particularly important in IP connections.

The required provisioning technologies must be developed under realistic test conditions and in close interaction with applications developers. Such efforts must be supported by a developer-scale test-bed which is application-centric and is capable of dedicated cross country bandwidth pipes. A similar test-bed has also been recommended by the protocols working group, and hence is discussed separately in Section 6.3, where the ideas of both groups are integrated into one recommendation.

4.2 Barriers to Provisioning

In addition to the technological issues, there are financial and organizational barriers to the development and deployment of provisioning technologies. These are listed as follows:

- Limited deployment of ultra-long haul DWDM links;

- Lack of support for striped/parallel transport (that utilize multiple data streams to fill the available link bandwidths) both at the core and application levels;
- Lack of high-speed circuit-switched infrastructure with network control-plane design and synchronous NICs with high-speed and on-demand reconfigurability; and
- Lack of well-developed methods and application interfaces for scheduling/reserving, allocation, initiation.

In connection with the on-demand end-to-end provisioned channels, there are additional challenges as described below:

- DOE applications do not follow the commercial scaling model of large number of users each with smaller bandwidth requirements;
- Lack of a security model for dedicated paths and the infrastructure that to manage them;
- Lack of a robust multi-cast solution efficiently supported on dedicated channels;
- High cost of equipment, including the costs of links, routers/switches and other equipment as well as deployment and maintenance;
- Lack of field-hardening of optical components such as memory/buffer, high-speed switches, Reamplification, Reshaping and Retiming (RRR) equipment, and lambda conversion gear;
- Lack of effective contention resolution methods for the allocation of channel pools; and
- Limited interoperability with other data networks, particularly legacy networks.

5. Workshop Findings in Network Transport Protocols Area

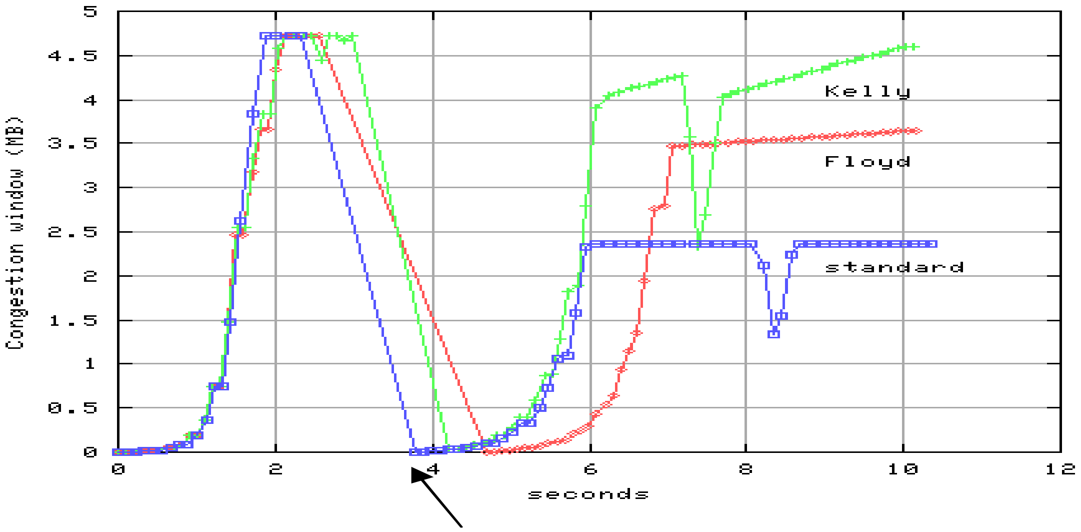
The current dominant Internet transport protocol, TCP, was originally designed and optimized for what we now consider low-speed data transfers over low bandwidth connections. It lacks the performance and scalability to meet the challenges of DOE large-science applications in terms of large throughputs as well as agile and stable dynamics. To achieve high throughput data transfers, TCP methods on shared IP networks can be adapted and scaled to Gbps and Tbps rates. But, this is a challenging task that requires investigations into various parts of TCP, including sustained slow-start and robust congestion avoidance, to achieve the require throughput levels. At the other extreme, one could provide dedicated high bandwidth channels from source to destination nodes wherein a suitable rate control method can be used for transport. This approach avoids the complicated problem of optimizing TCP by avoiding congestion altogether, but still requires mechanisms to account for non-congestive packet losses and suitable flow control to optimally utilize the provisioned bandwidth. Recently there have been several UDP-based methods that attempt to fill the available bandwidth but such methods can have very negative effects on the TCP transfers simultaneously taking place on shared links.

Current transport methods are massively inadequate to meet the multitude of DOE large-science networking requirements. The required throughput levels are unattainable except with significant efforts from teams of technical experts, that too often only in demonstration scenarios and typically for small periods of time. But such throughputs are needed at the application level on a daily basis. Furthermore, it is also important to sustain the throughputs during the entire execution of the application rather than ephemerally achieving the peak bandwidth. Currently, TCP methods are not able to provide sustained and stable streams for control operations particularly in networks with heavy traffic loads. Since TCP has provably complicated dynamics, it might be difficult to use it for control operations. The transport protocols needed to support control operations on dedicated channels must be developed particularly for long haul connections. Another important consideration is that the time-to-solution in the protocols area is currently too high; for example TCP tuning for Gbps throughputs took several years. In view of impending DOE needs, it is important to develop the needed protocols in a much more timely manner. It is also important to develop and integrate the functionalities between middleware and transport, and these aspects are discussed in the next section.

In meeting the DOE large-science requirements, it is instructive to note that the end users view the network as a tool or a resource much like a computer. Their main goal is to conduct scientific activities in their areas with minimal demands for using the network. Over the past years, however, several users have become (not always willingly) network experts in their attempts to scale TCP or other protocols to the required throughput levels using methods such as parallel streams and buffer tuning. But such improvements require in-depth knowledge of the protocols and are achieved by significant efforts by groups of experts. Such a “wizard gap” exists at all levels and the expertise needed for such efforts is beyond a typical science user. In a nutshell, the “gray matter tax” for such efforts is undesirably high. Thus, one of the considerations in meeting the DOE large-science needs is to advance the state of network protocols to make them plug-and-play for the application users. In particular, the use of protocols must be transparent to the users; they just specify the performance requirements for connections and all the other details such as the underlying provisioning or the corresponding parameter values must be hidden from them. It is to be noted that achieving such a level of transparency in the presence of a multitude of provisioning modes and matching protocols represents a significant challenge.

5.1 *Existing Transport Protocols*

The limitations of current transport methods, particularly TCP, in addressing high performance transport applications, have prompted a number of solutions with varying degrees of successes. To a large extent these efforts are focused on IP-based protocols such as the various TCP enhancements, net100, HSTCP, STCP, FAST, and UDP-based methods such as tsunami and SABUL. There are also efforts to adapt the protocols designed for Storage Area Networks (SAN), such as Fiber Channel, to the wide-area networks. Protocols that are specifically optimized to exploit the properties of dedicated channels are quite limited. Since a major consideration of IP-based protocols is the impact on other traffic streams, a significant effort has been extended to ensure their “gracefulness” or “fairness”. Lack of this consideration in dedicated channels opens up a vast potential for customization and optimization of the transport protocols, if not, motivating a whole new approach to their design.



UDP burst

Figure 7. Response of TCP variants to a UDP burst.

The problem of optimizing TCP to achieve ultra high throughputs is extremely complicated due the high non-linearity of its dynamics. By suitably controlling the slow-start phase and AIMD parameters, it is quite possible overcome some TCP limitations. But optimizing certain measures of performance might result in degrading the others; for example, as shown in Figure 7 avoidance of overshoot during slow start might result in slower recovery. Among the TCP based methods presently under investigation are various versions of TCP (Reno, Vegas), HSTCP, STCP, XCP, net100, and FAST. Protocols that optimize the flow rates of UDP streams achieve high utilization of the connection bandwidths, for example, tsunami, SABUL, IQ-RUDP, hurricane, RBUDP, FOBS, and IUDP, but they often significantly degrade the performance of competing TCP traffic. The class of protocols that adapt methods used in SANs, such as STP and Fiber Channel, typically achieve much higher data rates since the distance involved are much smaller and there is limited or no competing traffic. Recent efforts in Fiber Channel over SONET focus on utilizing SONET links as carriers for Fiber Channel streams.

In addition to the protocol designs, their actual implementations have a large impact on the performance. Since link bandwidths are several Gbps or 10s of Gbps, and current off-the shelf

NICs are typically operate at 1Gbps, striping methods are needed to utilize multiple NICs to generate aggregate throughputs commensurate with the link speeds. Also, these data rates are significantly higher than processor speeds, and hence methods are needed to minimize the impact on the processors by utilizing the OS-bypass methods such as RDMA.

5.2 Recommendations in Transport

The recommendations in the transport area are provided for two time-frames: a short time-frame of the next five years and a longer time-frame spanning five to ten years. In the short time-frame the following items are to be addressed:

Recommendation 1: *Transport Protocols and Implementations*

Transport methods for dedicated channels and IP networks must be designed for achieving high throughput, steering and control. The transport methods include TCP-, UDP- and SAN-based methods together with newer approaches. To utilize extremely high bandwidth channels, striping methods must be developed for multiple hosts and/or interfaces. Technical topics include TCP auto-tuning, large MTU, scavenger TCP, RDMA, and OS bypass methods.

Recommendation 2: *Transport Customization and Interfacing*

Transport methods must be customized to optimally match single and multiple hosts as well as channels of different modes. Transport methods must be suitably interfaced with storage methods to avoid impedance mismatches that could degrade the end-to-end transport performance. Interaction of sharing, resource allocation, and session control must be handled by providing interfaces from transport modules to provisioning infrastructure.

Recommendation 3: *Stochastic Control Methods*

Stochastic control theoretic methods must be developed to design protocols with well-understood and/or provable stability properties. These methods can also be utilized in analyzing the other transport protocols for their properties.

Recommendation 4: *Monitoring and Estimation Methods*

Monitoring and statistical estimation techniques must be designed to monitor the critical transport variables and dynamically adjust them to ensure transport stability and efficiency.

Recommendation 5: *Experimental Test-Beds:*

Experimental research networks are needed to validate new ultra high-speed protocols and dynamic provisioning technologies. These aspects are discussed in greater detail in Section 6.3.

Over the longer time-frame, the following items are to be addressed:

Recommendation 1: *Modular Adaptive Composable and Optimized Transport Modules:*

Highly dynamic and adaptive methods must be developed with the capabilities to statically and dynamically compose transport methods (including splicing transport methods) to match the application requirements and the underlying provisioning. Advanced transport methods must be developed to optimally exploit the dedicated links for achieving ultra high throughputs, and precise steering and control operations.

Recommendation 2: *Stochastic and Control Theoretic Design and Analysis:*

Stochastic control theoretic methods must be developed for the composable transport methods to analyze them as well as to guide their design to ensure stability and effectiveness

Recommendation 3: *Graceful Integration with Middleware and Applications:*

Application data and application semantics must be mapped into transport methods to optimally meet application requirements; the boundary between middleware and transport could be made transparent to the applications so that can operate without being aware of it.

Recommendation 4: *Vertical Integration of Applications, transport and Provisioning:*

Vertical integration of resource allocation policies (cost and utility) with transport methods must be carried out to present a unified view and interface to the applications.

There is a need to support the design and testing activities of the protocols under conditions that closely match the real operating environments, particularly those deploying newer provisioning methods. While the overall need for a test-bed is quite similar (in terms bandwidths and distances) to both provisioning and transport areas, their specific requirements are different. Nevertheless, due to the overall significance to both the areas, the issues of test-bed are discussed separately in Section 6.3.

6. General Recommendations

6.1. Network Science for High-Performance Networks

There is a need for systematic scientific approaches to the design, analysis and implementation of the transport methods and to network provisioning. This calls for a new “science” of high-performance networking. In fact such a need has been recognized for future networks in a more general sense, wherein it is no longer sufficient to adopt ad hoc methods for their design and analysis. Several recent workshops [10-12] identified such fundamental and scientific aspects of future network research from a more general perspective (see Appendix B). While it is not the main goal of this workshop to identify the need for a science of networks in general, such need in the context of high performance networking has been strongly felt by both groups. With regard to DOE large-scale science applications, there are several essential components to such a science.

? **On-Demand Bandwidth and Circuit Optimization:** Dynamic optimization and scheduling methods are needed to allocate the bandwidth pipes to various application requests. A comprehensive approach is needed for on-line estimation of the “bandwidths” of various network links and for their allocation on-demand to applications. It is likely that many of these allocation/scheduling problems are computationally hard, and efficient on-line methods must be designed to efficiently handle the allocations. Since the channel configurations are needed continuously across the network, the signaling aspects must be closely investigated to provide the required timeliness and reliability of the allocated channels. Once an understanding of signaling has been achieved, methods for both in-band and out-of-band signaling can be developed for dynamically switching the channels as per the allocations. There is a need for a scientific, systematic understanding of how to integrate the components for bandwidth allocation, channel scheduling, channel setup and teardown, and performance monitoring. At a higher level, a systematic framework is needed within which to analyze and classify application requirements and to design network configurations tailored to the resulting application classes.

? **Comprehensive Theory of Transport**: A comprehensive theory of transport is needed to rigorously design transport methods tailored to the underlying provisioning modes. Such a theory should enable the rigorous evaluation and sound statistical analysis of the resulting transport methods. Such a theory would require a synergy and extensions of a number of traditional disciplines. Since delays and losses experienced by packets depend on the competing traffic, they exhibit apparent randomness. Such effects are particularly pronounced in heavily loaded networks. These effects are compounded by the complicated dynamics of TCP, which are apparently quite complicated even under very simple conditions. New stochastic control methods may be required to design suitable transport control methods. The resultant controllers are likely to be non-linear with delayed feedback, and new ideas in non-linear control theory may be required to analyze them. The feedback delays could themselves be a source of chaotic dynamics in non-linear systems. The area of statistics can play several roles in the theory by providing methods for designing rigorous measurements and tests. Optimization theory can potentially provide ways to obtain suitable parameters for tuning the various protocols.

? **Strict Algorithmic Design and Implementation**: Strict algorithmic design methods must be developed to efficiently implement the designed protocols. In particular, the implementations must be modular, autonomic, adaptive, and composable. Considering that a single host might be connected via channels that are provisioned under different modes, it is important that transport stack includes different modules to match different provisioning modes at the same time. Such modules may be composed on demand to match the application at hand. Such efforts require strict algorithmic designs and software engineering practices to ensure the quality of implementations.

? **Statistical Inference and Optimized Data Collection**: Due to the sheer data volumes, it is inefficient to collect measurements from all nodes all the time for the purposes of diagnosis, optimization and performance tuning. The measurement and data collected must be aided by systematic inferencing methods to identify the critical and canonical sets of measurements needed. Statistical methods such as involved in the design of experiments, are required to ensure that the measurements are strategic and optimal.

6.2. Integration and Interactions

The network functionalities must be made available to the end users in a transparent manner independent of the underlying provisioning modes and the associated transport methods. Such levels of transparency can only be achieved by focused integration efforts, which have been often ignored by technology developers. As result, the application performance could be suboptimal and solutions could be hard to use. Both provisioning and transport methods must be developed to gracefully interface and integrate with each other as well as with the other components such as the operating system, middleware, and applications. In particular, to realize the require throughputs between the applications at the channel end-points, it is essential that the "impedance matches" be achieved at all the interfaces. Furthermore, these technologies must smoothly interact and co-exist with legacy networks, and provide a smooth transition to newer networks in the appropriate cases.

? **Middleware and Application Interfaces:** In IP networks it is common for the applications to interface with the transport modules which in turn communicate with the lower level services. The situation of on-demand provisioning is somewhat different. The applications themselves can request and be granted the dedicated channels to be used exclusively. Both the middleware and applications must be provided the needed interfaces to the provisioning and transport modules. Indeed, it would be most desirable if the transport and provisioning methods are developed in close association with the middleware, applications and end users. Test-beds outlined in the next section could facilitate such activities.

? **Hardware and Operating Systems:** Due to the sheer speeds of the network connections needed in large-scale applications, the usual host controlled transport methods may not be adequate. For example, a low-end host might not be fast enough to produce and/or consume the data that is arriving at multiple tens of Gbps. The network interfaces and the end hosts must be designed to operate at network speeds. Special modules may be necessary to implement OS-bypass and RDMA methods to relieve the host CPUs from being exclusively consumed by data transfers. Also host clusters might be required to operate in striped mode to sustain data rates that are commensurate with the channels and end-systems such as High Performance Storage Systems (HPSS).

? **Legacy Networks:** Since the DOE large-science projects are carried out by teams of geographically dispersed scientists, it is reasonable to expect that not all networks will be endowed with the newer provisioning and transport technologies. It would be more efficient to phase in newer methods into the production environments, and thus it is important to support the co-existence of various operations with the legacy networks (at least during the transition period). Such gradual transition to newer technologies can help the user adoption and faster integration into production applications.

? **Instrumentation and Diagnostic Tools:** Considering that several of the required technologies are at the forefront of provisioning and transport areas, it is important to provide measurement and diagnostic tools both during the development and deployment phases. Such capabilities can potentially make the development processes more efficient, and can make it easier to diagnose the operational problems. For example, by utilizing the web100/net100 instruments as an integral part of TCP based methods, it would be possible to easily diagnose the problems and tune the protocol parameters. Similar instruments could be developed for other transport methods as well as provisioning methods.

6.3. Research Test-Beds

Due to the extreme demands imposed by DOE large-science applications on networking, existing test-beds and simulation tools are inadequate to provide sufficiently detailed operating conditions such as physical layer losses over long distances, background traffic levels at tens of Gbps, or realistic switching times of optical equipment. More generally, there have been two major shortcomings in previous efforts to develop high-performance network capabilities.

? First, there have been no test-beds to provide adequate operating conditions in terms of bandwidths, distances and traffic levels. Most simulators are not capable of supporting data rates of the order of Tbps, particularly over dedicated channels. Most existing test-beds do not provide tens of Gbps speeds with on-demand provisioned paths between application nodes that are separated by thousands of miles. Historically, methods based on simulations and small scale test-beds often resulted in technologies which fell short of the needs. In particular, simulations enable a detailed study of transport methods but mostly under small network configurations and IP connections. But such results are not extendable to high performance networks (particularly with dedicated channels), and furthermore they hide some of the subtle performance issues. The small scale test-beds are not able to accurately represent the physical losses typical of long haul high bandwidth channels and subtle timing effects of control and signaling channels.

? Second, the adoption of research tools by the users has been highly limited due to the lack of natural transition paths. Tools developed by network researchers often require a significant amount of integration before they can be used by non-experts, and consequently are not deployed extensively in the field. This is particularly true of the protocols developed using special purpose simulators or test-beds, since they often have to be re-implemented from scratch in application and production environments.

The provisioning and transport technologies described in the previous sections can only be adequately developed on powerful research networks or test-bed capable of providing real operating conditions. Thus, there is a need for a network test-bed that provides the following functionalities:

State-of-the-art Components: The test-bed must provide high-performance network links in the form of a dynamic combination of research and production links to support research and development of various networking components. Also, the test-bed must incorporate the required software and hardware networking components, including routers/switches, high bandwidth long-haul links, protocols and application interface modules.

Integrated Development Environments: The test-bed must provide mechanisms to integrate a wide spectrum of network technologies including high throughput protocol, dynamic provisioning, interactive visualization and steering, and high performance cyber security measures. In particular, the test-bed must provide an environment wherein these technologies can be developed through a close interaction with the application users in gradual evolutionary stages.

Smooth Technology Transition: The test-bed must provide a real application environment. It must support the transition of the network technologies from research stages to production stages by allowing them to mature in such an environment. Furthermore, the application users as well as system and network administrators should be involved during the testing and maturation process. Such an approach will potentially enable an easier adoption by users and facilitate the installation of new network technologies on production systems.

In addition to the development of provisioning and transport technologies outlined in the previous section, the test-beds must support the following operational activities:

- ? Transfer of network technologies to science applications through joint projects involving network research and applications users;
- ? Transfer of complete application solutions to production networks through projects that utilize the network research in combination with the production networks;
- ? Development of technology for end-to-end solutions for applications using teams of network researchers, network operations personnel and application users; and
- ? Experimental network testing activities involving researchers from across the country and the continents

Furthermore, as envisioned in the roadmap of June 2003 workshop [2], such a test-bed can be a valuable augmentation to the next generation ESnet, thereby providing the vital connectivity between DOE locations as well as with other organizations to enable the execution of large-science applications.

In summary, the characteristics of an experimental ultra high-speed network test-bed are as follows:

- ✍ Interconnection of at least three science facilities with large-scale science applications is needed to validate the performance of ultra high-speed network technologies;

- ✍ Geographical coverage must be adequate to capture optical characteristics (such as physical losses), transport protocols dynamics, and application behaviors comparable to that of real-word distributed large-scale science applications;

- ✍ Integration with appropriate middleware (GridFTP) is needed to effectively and efficiently couple large-scale science applications to ultra high-speed networks;

- ✍ Scalable network measurement tools must be provided to calibrate the performance of newly developed ultra high-speed transport protocols and dynamic provisioning network technology; and

- ✍ Well-defined technology transfer plan is needed to transition the mature network technologies from experimental to production networks, in particular ESnet.

6.4. Network Security Issues

While the network security was not an explicit item on the workshop agenda, it has come to play an important role in network performance over the past few years, particularly in operational networks connecting DOE sites. Hence it is important to be cognizant of network security considerations and implications in developing various provisioning and transport methods discussed in the previous sections. The network performance of the applications could be significantly affected by the cyber security measures that include security policies, firewalls and authentication methods. There are three important aspects to consider.

Securing Operational and Development Environments: First, network environments under which various transport and provisioning methods are implemented must be made secure through the use of proper authentication, validation and access controls. Considering the data speeds of multiple tens of Gbps (or higher), it is of particular importance to deploy intrusion detection and firewall systems capable of operating at these line rates. Furthermore, most intrusion detection and filtering methods have been developed for packet switched networks. A new class of methods may be needed to secure the networks that provide on-demand end-to-end dedicated channels, for example to protect against attacks that will lead to channels being allocated to attacker's traffic or denying them to legitimate users.

Effects of Security Measures on Performance: An important issue is the impact of security measures on application performance. As recently evidenced, the proliferation of strict firewalls, particularly at DOE sites, rendered several network-based applications inoperable. In particular, several legacy applications that relied on open socket communications simply stopped working since firewalls by default denied the communications on general ports. While this problem can be temporarily fixed by port exceptions or moving hosts into open portions of the networks, it leaves them vulnerable to attacks (defeating the very purpose of firewalls in the first place). More systematic efforts are needed to provide graceful interoperation of science applications under secured network environments. Obviously, today's crude packet filters and firewalls have limiting effects on the data transmission rates, which in turn limit the application throughputs.

Proactive Countermeasures: The provisioning technologies outlined in previous sections involve running services such as bandwidth allocation, and signaling to setup and tear down the paths over the networks. These services could be the target of newer attacks, particularly of denial-of-service type, which are not anticipated and handled in current IP networks. Similarly, the newer versions of transport protocols might be vulnerable to certain attacks as some of the current high-performance protocols. Such considerations might be taken into account in developing the provisioning and transport technologies as described in the previous sections.

References

1. High-Performance Network Planning Workshop, August 13-15, 2002, Report: High-Performance Networks for High-Impact Science, http://DOECollaboratory.pnl.gov/meetings/hpnpw
2. DOE Science Networking Workshop, June 3-5, 2003, Report: DOE Science Networking Challenge; Roadmap to 2008, http://www.osti.doe.gov/bridge
3. DOE Science Computing Conference: The Future of High Performance Computing and Communications, June 19-20, 2003, http://www.doe-sci-comp.info
4. NSF Workshop on Ultra-High Capacity Optical Communications and Networking, October 21-22, 2002
5. NSF Workshop on Network Research Testbeds, October 17-18, 2002, http://gaia.cs.umass.edu/testbed_workshop
6. NSF ANIR Workshop on Experimental Infostructure Networks, May 20-21, 2002, http://www.calit2.net/events/2002/nsf/index.html
7. NSF CISE Grand Challenges in e-Science Work, December 5-6, 2001, http://www.evl.uic.edu/activity/NSF/index.html
8. Network Modeling and Simulation Program, DARPA, http://www.darpa.mil/ipto/research/nms
9. First International Workshop on Protocols for Fast Long-Distance Networks, February 3-4, 2003, http://datatag.web.cern.ch/datatag/pfldnet2003
10. NSF Workshop on Fundamental Research in Networking, April 22-23, 2003 http://www.cs.virginia.edu/~jorg/workshop1/NSF-Workshop-Reportv10.doc
11. Workshop on Network Research: Exploration of Dimensions and Scope (NREDS) August 25, 2003, http://www.acm.org/sigs/sigcomm/sigcomm2003/workshop/nreds/
12. NSF/COST Workshop on Exchanges and Trends in Networking (NeXtworking '03), June 2003. http://cgi.di.uoa.gr/~istavrak/costnsf/Welcome.html

Appendix A: Requirements of Two DOE Large-Science Applications

To discuss the network requirements in concrete terms, we now consider two specific applications, High Energy Nuclear Physics (HENP) and Terascale Supernova Initiative (TSI); the former mostly deals with high-performance heavy-lift data transport and the latter highlights the breadth of networking functionalities needed. The HENP tasks predominantly require the transport of terabytes of data across the nation and the Atlantic at data rates matching the available link rates (OC48 and OC192). The network requirements of HENP data transport shown in Table A.1 are unprecedented: they must deliver hundreds of Gbps throughputs to the applications in near future and several Terabits/sec within the next decade.

Type of Interaction	Sources (Storage)	Bandwidth Requirements	Current TCP Performance
Data transfer: HENP – Tier 1	1 Pbytes	100 Gbps	300 Mbps
Data transfer: HENP – Tier 2	100 Gbytes	30 Gbps	300 Mbps
Data transfer: HENP – Tier 3	30 Gbytes	10 Gbps	300 Mbps
Data transfer: HENP – Tier 4	5 Gbytes	1 Gbps	300 Mbps
Computations	1 - 3 streams	100 Gbps	300 Mbps
Real-time steering	2-10 streams	10 Gbps	300 Mbps
Remote Visualization	2-10 streams	10 Gbps	300 Mbps

Table A1. Network requirements for high-energy and nuclear physics applications.

In contrast, the TSI involves a wide spectrum of tasks to be cooperatively performed over wide-area networks by a group of domain experts distributed at various national laboratories and universities. The tasks range from cooperative remote visualization of massive archival data through the distribution of large amounts of simulation data, to the interactive evolution of a supernova computation (computational steering). In the case of remote visualization, the data must be rendered and presented on-line to various participant sites with different end-devices ranging from visualization caves through high-end workstations to personal desktops. The control of this visualization will be handed back and forth among the sites, and the response of the distributed rendering engine should feel instantaneous. Total aggregate data rates needed are of the order of several hundred Gbps, although the local rate to a cave won't exceed a few hundred Mbps. The challenge is that TCP can easily take tens or hundreds of seconds to respond to a control or congestion event and to be interactive, the visualization data stream must respond in a few tens or at most a hundred milliseconds. TSI places a very similar, although not quite so stringent requirement on remote computational steering. However, the most challenging network task of TSI combines both elements. The computation itself is to be interactively visualized and at the same time controlled over the network by experts in various fields such as hydrodynamics, radiation transport, and nuclear physics, who are geographically dispersed across the country. The networks over which such collaborations will be carried out could be quite varied, with national laboratories connected over the ESnet, and the universities connected via Internet2.

TSI is particularly demanding because of the breadth of its networking requirements. Other DOE large-science applications typically require a subset of its capabilities. The Genomes to Life project, for example, requires similar capabilities only for handling the experimental and computational genomics data distributed at various sites across the nation. Other projects require visualization and steering of molecular dynamics computations on remote supercomputers. TSI requires both. Another important class of large-science applications deal with DOE experimental facilities such as neutron sources including SNS and HFIR. Currently scientists using such facilities these travel to their locations to conduct experiments and take back the data with them, often as stacks of CDs or data DVDs. The ability to perform experiments over the network and transfer the data could eliminate most of the need for travel thereby improving the flexibility of access and productivity of the users. Such network access has been attempted previously as part of the “instruments over the web” initiatives, and these technologies provided the functionalities needed at a high-level. But the underlying network performance was not adequate; for example, control loops were not stable and user commands (e.g., stop the source) weren't received in time. Also, the data sets in HFIR experiments are several tens of Gigabytes, which could not be reliably sent to remote locations over the current networks.

Appendix B: Related Workshops

There have been a series of workshops and other activities both within and outside DOE to identify the needs and plans for the next generation wide-area networks for scientific applications. While the other workshops are either more general in scope such as addressing the Internet environments in general or more focused on specific technologies such as optical components, the DOE series of workshops is focused heavily on large-science applications. DOE planning workshop [1] of August 2002 identified a number of science areas with the high-performance networking needs, and prepared a comprehensive list of requirements. A follow-on workshop in June 2003 [2] developed a roadmap for the network infrastructure to address the DOE needs with a special attention paid to the large-science networking needs.

Analogous activities but with typically different focus have been taking place within National Science Foundation (NSF) and other agencies over the past few years. The 2001 workshop [7] on e-Science grand challenges identified the cyber-infrastructure requirements, which included networking technologies, to address the nation's science and engineering needs. The 2003 workshop [10] was comprehensive in addressing several fundamental research aspects of future networks. The Association of Computing Machinery (ACM) workshop [11] dealt with exploration of dimensions and scope of network research. The joint European Union and NSF workshop in June 2003 [12] discussed the key new networking technologies and fundamental aspects. The scope of these workshops is broader than the current one both in terms of class of applications as well as infrastructure areas. The two NSF workshops on testbeds [5] and infostructure [6] specifically dealt with developing networks with capabilities beyond the current ones. Both these workshops focused on issues that are much broader than the current workshop but not specific enough to address the DOE large-science needs. Several of the high-performance network capabilities could be enabled by optical networking technologies, and the NSF workshop [4] on this topic is narrower in terms of the technologies considered but is broader in terms of the network capabilities. Similarly, the CERN workshop [9] concentrated on protocol issues and did not include dynamic provisioning aspects.

The Network Modeling and Simulation [8] program of Defense Advanced Research Projects Agency (DARPA) addresses simulation and emulation technologies of large-scale wireless and wireline networks. Its focus is on general aspects of networking for DoD and is not specific to large-scale scientific needs that are typical of DOE science projects.

Appendix C: List of Attendees

William E. Allcock, Argonne National Laboratory
Ray Bair, Pacific Northwest National Laboratory
Micah Beck, University of Tennessee
Gee-Kung Chang, Georgia Institute of Technology
Steve Cortez, CIENA Corporation
Roger Cottrell, Stanford Linear Accelerator Center
Tom DeFanti, University of Illinois-Chicago
Thomas H. Dunigan, Oak Ridge National Laboratory
Wu Feng, Los Alamos National Laboratory
Dennis Ferguson, Juniper Networks
Mark Gardner, Los Alamos National Laboratory
Robert Grossman, University of Illinois-Chicago
Glenn Heinle, LightSand Communications, Inc.
Wesley K. Kaplow, Qwest Government Services
James Leighton, Lawrence Berkeley National Laboratory
Steven Low, Caltech
Matthew Mathis, Pittsburgh Supercomputing Center
Mark Meiss, Indiana University
Biswanath Mukherjee, University of California-Davis
Thomas Ndousse, U.S. Department of Energy
Bill Nickless, Argonne National Laboratory
Walter Polansky, U.S. Department of Energy
Nageswara Rao, Oak Ridge National Laboratory
Allyn Romanow, Cisco Systems
George Seweryniak, U.S. Department of Energy
Daniel Stevenson, MCNC Research and Development Institute
Raymond Struble, Level 3 Communications, LLC
Don Towsley, University of Massachusetts
Malathi Veeraraghavan, University of Virginia
Jay Wiesenfeld, Celion Networks
William Wing, Oak Ridge National Laboratory
Matthew Wolf, Georgia Tech

Appendix D: Workshop Agenda

Thursday, April 10, 2003

7:30 a.m. *Registration*

8:30 a.m. Welcome: Nagi Rao and Bill Wing
8:35 a.m. High-Performance Networks and DOE's Science Mission:
Walter M. Polansky
9:25 a.m. Terabit Networking R&D for Petascale Sciences: Thomas Ndousse
9:50 a.m. A Vision for Energy Sciences Network: George Seweryniak
10:05 a.m. Challenges of Transport Protocols and Network Provisioning
Bill Wing and Nagi Rao

10:35 a.m. *Coffee Break*

10:50 a.m. DOE Network Planning Workshop Summary - Ray Bair
11:30 a.m. **Parallel Sessions**
Group A: Transport Protocols Wu Feng and Don Towsley
Group B: Network Provisioning Bill Wing and Biswanath Mukherjee

12:15 p.m. *Lunch*

1:30 PM **Parallel Sessions Continued**
Group A: Network Transport Wu Feng and Don Towsley
Group B: Network Provisioning Bill Wing and Biswanath Mukherjee

3:00 p.m. *Coffee Break*

3:15 p.m. **Parallel Sessions Continued**
Group A: Network Transport Wu Feng and Don Towsley
Group B: Network Provisioning Bill Wing and Biswanath Mukherjee
4:45 p.m. Joint Session - Group Summaries to all Participants

6:30 p.m. *Reception and Dinner*

Friday, April 11, 2003

7:30 a.m. *Coffee and Muffins*

8:00 a.m. Parallel Sessions Continued
Group A. Transport Protocols - Wu Feng and Don Towsley
Group C: Network Provisioning - Bill Wing and Biswanath Mukherjee

9:45 a.m. *Coffee Break*

10:00 a.m. Written Summary and Follow-on Assignments
10:00 a.m. Closing Remarks and Task Assignments – Thomas Ndousse
12:00 noon Workshop Adjourn
1:00 p.m. Meeting of Organizers and Group Chairs

Appendix C: Lists of Working Group Members

Transport Working Group:

Chairs: Wu Feng (LANL) and Don Towsley (U. Mass)

Tom Dunigan/Nagi Rao	ORNL
Don Towsley	UMASS
Matt Mattis	PSC
Mark Meiss	Indiana Uni
Wu Feng/Mark Gardner	LANL
Stephen Low	CalTech
Bob Grossman	UIC
Bill Allcock/Bill Nickless	ANL
Thomas DeFonte	UIC
Les Cottrell	SLAC
Ally Romanow	Cisco
Glenn Heinle	LightSand

Provisioning Working Group:

Chairs: Bill Wing (ORNL) and Biswanath Mukherjee (UC Davis)

Dan Stevenson	MCMC
GK Chang/Matt Wolf	Gatech
Bill Wing	ORNL
Jim Leighton	ESNET/LBNL
Wes Kaplow	Qwest
Steve Cortez	Cienna
Micah Beck	U. Tennessee
Malathi Veeraraghavan	U. Virginia
Jay Wiesenfeld	Celion
Raymond Struble	Level3
Dennis Ferguson	Juniper
Ray Bair	PNNL

DOE Headquarters

Thomas Ndousse	DOE
George Seweryniak	DOE
Walter Polansky	DOE

Appendix E: Guidelines to Participants

This workshop serves a very important role in helping to chart the networking directions of Mathematics, Information and Computer Science Division of DOE's Office of Science. This area represents one of the most crucial enabling technologies to support several DOE's large-science applications in radically new and effective ways. The objective is to develop a comprehensive understanding of the DLSA Networking Needs (DLSANN) to identify those that are met by (a) current technologies with small-scale incremental efforts, and (b) focused DOE efforts in the near term (around two years) and further along. The outcome of this workshop could provide an important advice in shaping the networking priorities of MICS and DOE for DLSA efforts. Some of the most important questions concern the *deployable solutions* in the near term, that is technologies leading to operational solutions that can be used by non-experts in ESnet and other environments.

There are three important parts to the workshop activities:

1. **Pre-Workshop Activities**: Participants are from federal agencies, national laboratories, universities and industry. Due to their diverse backgrounds and interests, they are requested to look at the DOE networking needs for high-impact science prior to the workshop. A comprehensive report on these matters was produced as a result of a DOE workshop. Please refer to "Background Information" on our workshop website. Participants are requested to identify the networking technologies in their individual areas that can (partially) meet these DOE networking challenges currently and in future if suitable projects are supported. Ideally, participants will prepare *individual lists* of DOE applications and the technologies from their areas before the workshop.

2. **Workshop Activities**: The workshop contains a series of group sessions interspersed with joint sessions. In the earlier group sessions (Day 1 morning session), the participants will make brief presentations of their lists to communicate their views to the other participants (please note the diversity of backgrounds of the participants). Then these lists are consolidated through open discussions to identify the currently available technologies together with their strengths and limitations (Day 1, afternoon sessions 1 and 2). Also, perhaps more importantly, the technologies that are to be developed by DOE to meet the challenges will be identified through open discussions. These results will be presented (in a preliminary form) by the group leaders to all participants in a joint session during the first day. At this time the connections between topics of various groups will be identified and discussed. In the second day, these group lists will be further refined with details and their connections with topics from other groups. These group lists will be combined and eventually converted into a workshop report. Work assignments will be identified to generate the final workshop report.

3. **Post-Workshop Activities:** Based on the workshop activities, a detailed report will be produced after the workshop. Each group will contribute to their individual area as well to the overall report. This report consists of three important parts: (a) explicit transport, provisioning and functional capabilities needed for DOE networks; (b) existing networking technologies that meet some of the needs; and (c) future directions to develop the technologies that are needed to meet the needs within the timeframe of few years.

Answers to the following questions could be useful for the participants in their contributions to this workshop.

General questions to all groups:

1. What are the current technologies in your expertise areas that can contribute to DLSA?
2. What DLSA networking capabilities are not achieved by current network technologies in your area?
3. What networking technologies can be developed and deployed in operational networks for DLSA within next two years or so?
    a. With explicit DOE efforts
    b. With industry and other agency efforts alone
4. What DLSA networking requirements are not likely to be met if we rely on
    a. With explicit DOE efforts
    b. With industry and other agency efforts alone
5. What DLSA requirements are too challenging to be completely met within next two-years in spite of various DOE and other efforts

The following questions for individual groups will be refined and expanded within next few days.

Transport Group

Assessment of TCP and non-TCP protocols for

a. high-performance throughput
b. visualization of large data sets over wide-area networks
c. computational steering
d. closed-loop control of remote devices and computations

Provisioning Group

What type of network levels provisioning are appropriate for immediate attention for DLSANN?

Routed IP networks vs.All Optical Networks vs. mixed networks

What is the role in DLSANN of

1. Optical Burst switching
2. Fiber channel over SONET

Please note that we explicitly deal with the networking requirements of DOE's large-science applications and do not target areas that are not relevant to DLSANN. Interesting network research areas that do not address DLSANN are not appropriate for this workshop – for example, simply coming up with a list of interesting topics to be studied over next several years is not very useful to this workshop. Also, our focus extends beyond the research part in that we like the capabilities to be deployed and tested so that they can be provided as production modules to application users.