Terabit Networks for Extreme-Scale Science

February 16th-17th, 2011
Rockville, MD

U.S. DEPARTMENT OF ENERGY | Office of Science

ASCR

Workshop on:

Terabits Networks for Extreme Scale Science

February 16th-17th, 2011
Hilton Hotel
Rockville, MD

General Chair:

William E. Johnston, Energy Sciences Network

Co-Chairs

Nasir Ghani, University of New Mexico
Tom Lehman, University of Southern California
Inder Monga, Energy Sciences Network
Philip Demar, Fermi National Laboratory
Dantong Yu, Brookhaven National Laboratory
William Allcock, Argonne National Laboratory
Donald Petravick, National Center for Supercomputing Applications

Table of Contents

I. EXECUTIVE SUMMARY 4

II. R&D REQUIREMENTS TO SUPPORT DOE EXASCALE SCIENCE 7

III. FINDINGS AND RECOMMENDATIONS OF BREAK-OUT GROUPS 11

GROUP 1: ADVANCED USER LEVEL NETWORK-AWARE SERVICES 25
GROUP 2: TERABIT BACKBONE, MAN, AND CAMPUS NETWORKING 11
GROUP 3: TERABITS END SYSTEMS (LANS, STORAGE, FILE, AND HOST SYSTEMS) 33
GROUP 4: CYBERSECURITY IN THE REALM OF HPC AND DATA-INTENSIVE SCIENCE 43

IV. APPENDIX 48

WORKSHOP AGENDA 49

LIST OF WORKSHOP ATTENDEES 51

I. EXECUTIVE SUMMARY

The first DOE Workshop on Terabits Networks for Extreme Scale Science was held at the Hilton Hotel in Rockville, MD, from February 16th-17th, 2011. The venue brought together a strong and diverse team of researchers, practitioners, and experts across key areas such as backbone network design, campus/regional network communities, high-performance computing application. This group was tasked with identifying the core network-related challenges facing the DOE community in the area of extreme scale science, and also providing a set of recommendations and guidelines for future research and development (R&D) efforts.

Overall the workshop was organized in light of the emerging need to support distributed —exascale” applications within the DOE research community, e.g., simulation, experiment data management, data analysis, and remote visualization. This expansion is being fueled by the deluge of data being generated by researchers working across areas such as high energy physics, climate change, genomics, fusion, synchrotron light sources, etc. For example, the data repositories for the DOE-supported Large Hadron Collider project (LHC) are already multiple petabytes in size and growing at the rate of petabytes per year, and large climate modeling data repositories are projected to reach ten petabytes within the next few years. Many future projections are also calling for exabyte bulk transfers by the end of the decade, i.e., 1,000 petabytes. Indeed, this is an extremely large figure, representing an order or more increase from current levels. As a result there is a clear and burgeoning need to support the distribution, management, and analysis of this vast amount of data between dispersed DOE Leadership Computing Facilities (LCF), scientific instrument facilities, and university and international partner institutions. In fact these trends are being evidenced by the increased levels of archival data at key sites as well as usage levels on the DOE Energy Sciences Network (ESnet) backbone. Furthermore many of these growth projections (and networking requirements) have also been outlined in a series of focused DOE workshop reports for differing scientific fields.

Nevertheless the emergence of exascale data transfer and computing requirements poses many challenges for currently-deployed science facilities. For example, existing backbone networks have been built using 10 Gigabit technologies and will find it very challenging to support for genuine exascale transfer needs. As a result the DOE is actively working to deploy ultra-scalable 40/100 Gigabit line rate technologies in order to achieve high-terabit networking capabilities. However as these developments emerge in several years, new and improved network provisioning and user service models will be needed in order to achieve operations across federated infrastructures comprising of multiple domains and diverse technology layers. Meanwhile most existing end hosts and storage systems will face challenges when handling terabit-level data streams, posing a crucial gap in *end-to-end* exascale performance. Finally, available end system middleware capabilities are rather specialized and support-intensive, and hence pose further limitations for future DOE applications. The move to exascale will mandate critical advances and possibly paradigm shifts across several cross-cutting technologies to realize the goal of reliable and predicable end-to-end terabit service, i.e., network hardware, network resource management/provisioning, network services, advanced middleware and co-designed applications, end systems architectures, parallelized storage, diagnostics, performance monitoring, and end-to-end security.

In light of the above it is imperative to develop a clear research roadmap to develop, test, and deliver advanced exascale solutions to meet the DOE's needs. This requirement is given added impetus by the fact that most commercial applications will likely generate much smaller amounts of data for the foreseeable future, i.e., as compared to the levels generated by a single exascale DOE application. As a result the industrial R&D community is not expected to address such high-end transfer, storage, and computational concerns within the desired timeframe, i.e., next 5-10 years. Instead these challenges can only be met via transformational R&D efforts involving the DOE community and its academic and industrial partner institutions. To date this community has established a very strong track record of innovation and delivered many successful solutions to meet its computing needs. Indeed such a proactive effort will introduce many new research opportunities and will play an invaluable role in preserving the DOE's best-in-the-class position and maintaining its preeminence in key scientific fields.

In order to address these challenges, the workshop attendees evaluated the emerging requirements in the exascale application space and made some key recommendations for targeted future R&D efforts. Specifically, the following areas were identified for the DOE to meet the challenge of exascale science:

1) Multi-site exascale computing paradigms will require increasingly advanced federated networking, end host systems, and storage technologies that are beyond what is available today. Hence the DOE must play a leading and active R&D role to help shape and develop the critical technologies in these areas that are vital for achieving its science mission.

2) Intelligent and agile resource management/traffic engineering solutions will be critical in future terabit networks comprising of multiple layers and federating across multiple domains. These schemes must support end-to-end provisioning across terabit links with varying granularities and also provide inter-domain peering at all technology levels.

3) Network resource management capabilities must be tightly integrated with advanced middleware on end-host systems in order to deliver *end-to-end* throughputs at the exascale. These solutions must facilitate a full range of flexible services and achieve virtualization co-design across local/wide area networking, computing and storage resources for terabit-level data movement.

4) Sustainable solutions for end-to-end diagnostics and performance monitoring will play a key role in improving network usability across diverse, distributed exascale infrastructure components—networking, computation, and storage. In particular the development of a standardized set of real-time metrics/statistics will be highly beneficial for exascale application development and support, along with high-level visualization tools.

5) Improved data transfer services and protocol capabilities are needed to support the distribution requirements of exascale applications. These offerings must feature a high level of parallelization, composeability, and error-recovery. In addition, tight integration with underlying middleware and traffic shaping capabilities on end host systems is also required to provide guaranteed performances.

6) The availability of persistent terabit networking testbeds, along with federated ecosystems of such testbeds, will be vital for new R&D initiatives focusing on the development of exascale applications for the DOE community. Indeed, these facilities will play an invaluable role in prototyping, testing, and deploying novel technologies across extended multi-site locations and then transitioning them into ―live" production environments.

7) New best-practice security models should be investigated for data-intensive science environments. Related R&D efforts should focus on new architectures that separate exascale circuit-like data transfers from regular site/enterprise traffic and build appropriate security controls for them. In addition, there is also a pressing need for new research initiatives in hardware-based acceleration for intrusion detection at 100 Gbps speeds and beyond.

8) All exascale architectures are under a very tight energy budget. It will be imperative for end-to-end networking paradigms to provide terabit scalability without dramatically increasing the

energy profile of the network. Network research efforts should leverage the silicon and photonic integration being aggressively pursued for new low-power computing architectures and focus on seamlessly integrating them with campus local area network (LAN) and wide area network (WAN) terabit networks.

II. R&D Requirements To Support DOE Exascale Science

The DOE research community is heavily-dependent upon networking services to support its mission to advance science. These services range from regular best-effort IP packet forwarding all the way to high-capacity multi-ten gigabit circuit provisioning for bulk data transfers. However with ever-expanding research endeavors and collaborations, current trends are pointing to the emergence of new era of *exascale* science by the end of this decade. Indeed a number of DOE-organized workshops have already identified the key requirements for emerging exascale applications in a series of reports, see references in the Appendix. Nevertheless it is becoming increasingly evident that currently-deployed DOE science facilities will be unable to meet the sheer scale of these burgeoning demands. Although these existing deployments are well-suited for today’s user and application needs, they will require significant improvements in scalability to support emerging exascale needs. At the same time most commercial networking and information technology (IT) solutions are not expected to meet the needs of exascale research within the desired timeframe either.

Future exascale applications will be characterized by unprecedented storage, computational, and networking requirements, i.e., across areas such as high energy physics (HEP), climate modeling, genomics, fusion, light sources, astrophysics, etc. For example, HEP applications are expected to generate exabyte-level datasets within the next decade, e.g., projects such as the Large Hadron Collider (LHC), International Thermonuclear Experimental Reactor (ITER), etc. Owing to the highly-distributed nature of these scientific collaborations, the massive data volumes will have to be transported, and stored/replicated, to dispersed global high performance computing (HPC) facilities and/or individual user groups and communities for analysis and duration. In general these transfers will be achieved over a federated set of backbone network domains, built using optical dense wavelength division multiplexing (DWDM) technologies. In fact many fiber infrastructures are already in place and are being actively expanded to support 40/100 Gbps ports to achieve petabyte transport. Meanwhile, key supercomputing centers hosted at the DOE’s Leadership Computing Facilities (LCF's) will play an increasingly-important role in distributed simulation and data analysis, e.g., Oak Ridge National Laboratory (ORNL/OLCF), Argonne National Laboratory (ANL/ALCF), National Energy Research Scientific Computing Center (NERSC), etc. In fact many of these facilities have petaflops-scale systems deployed today. Moreover many DOE supercomputing facilities are further projected to deploy unprecedented *exaflops*-scale capability systems within a decade.

Given the ultra-fast transfer speeds of exascale applications, a high-level of *end-to-end* coordination and integration will be required across all components of the future DOE network infrastructures —including backbone networks, access networks, end host systems, storage systems, etc. However many these capabilities are lacking in current provisioning systems and tools. To effectively meet the challenges for exascale science, new capabilities and advances are necessary across a range of cross-cutting and inter-related areas, i.e., network resource management/provisioning, network services and virtualization, advanced middleware and co-designed applications, parallelized storage, diagnostics, performance monitoring, and security.

Along these lines, this workshop was divided into three key breakout sessions, i.e., terabits networking, advanced user level network-aware services, and terabit end systems (LAN, storage, file, host systems). Here each breakout session was designed to answer a series of questions that built upon each other, and an additional —virtual” track was also added to discuss issues relating to security in exascale computing environments. A high-level summary of the key findings is presented, and interested readers are referred to the individual breakout section reports for further details:

1) **Network layer:** Exascale computing will require continued scalability improvements at the base network layer. Several core backbones (including DOE's ESnet) are already starting to deploy equipment that will support 100 Gbps channel speeds (i.e., on a single DWDM wavelength). Albeit impressive, these speeds still pose excessively lengthy data transfer times for exabyte-sized datasets, e.g., for example, transferring 1 exabyte of data within a week requires a staggering 13.23 Tbps of capacity. Hence R&E networks are also considering terabit-level speeds via a combination of approaches, i.e., 10 x 100 Gbps, 3 x 400 Gbps, 1 Tbps. As these heterogeneous line rates emerge, there will be a crucial need to develop intelligent network resource management/traffic engineering solutions and network services to support data flow transfers across multiple technology layers and domains. In particular, survivability is a very critical concern here as the amount of data lost during failure switchovers at ultra-fast line speeds can be very large, significantly complicating higher-layer recovery procedures. It will therefore be important to pursue a combination of robust network (pre-) design strategies as well as survivable service provisioning. Furthermore, to deliver true exascale performance to the *application* layer, resource management capabilities will have to be extended across multiple federated domains, into LAN, and even up to the end host systems/storage systems. These ―edge" expansions will likely pose further challenges in terms of new standards (client-side port interfaces), reduced packaging/footprints overheads, and lower price-points for components. Overall these development challenges can only be addressed via targeted R&D efforts, many of which will be very disruptive of today's approaches. Further efforts to model/quantify network behaviors will also be very beneficial, as they will provide users and applications with detailed performance prediction capabilities.

2) **End-systems and storage/file systems:** End system hosts with multi-/many-core capabilities give scientists very powerful single-node systems. These systems will also have access to highly-evolved technologies such as non-volatile memories, e.g., flash, phase-change, memristor, solid state drives, etc. However the input/output (I/O) capabilities on these hosts are not expected to increase as fast until operating systems (OS) kernels are improved to fully leverage the new interfaces (e.g., photonics off/on the chip) and multiple cores/accelerators to enhance existing and emerging data transfer protocols. Meanwhile file systems also need to be investigated for terabit networking paradigms to help scale data transfers and allow scientists to interact with their data using the models natural to the problem at hand. In particular these R&D efforts must focus on building middleware and system software packages to optimize parallel file system performance in an end-to-end manner, i.e., optimization of single-node and/or coordinated multi-node parallel file system performance for bulk transfers and remote file system access. Further efforts should also address the design of an improved network/browser-based automated data transfer service with support for non-interactive third party transfers, retry for temporary data errors, and tie-in to network bandwidth reservation and authentication systems. Finally, in order to provide truly end-to-end paradigms, R&D efforts should also develop architectures, systems, services, interfaces, and protocols to simplify provisioning, reservation, and monitoring of every component in the end-to-end transfer, i.e., including WAN, LAN, storage, processors, memory, even internal buses. This mandates the ability to co-schedule every component along the way, features which are largely lacking today, particularly closer to the network edge, i.e., terabit-capable end systems, storage clusters, LAN domains, etc

3) **End-to-end co-design and virtualization:** Exascale applications will need tight coordination between different hardware/software components along an end-to-end path in order to achieve high throughputs, i.e., application-to-application. As a result, intelligent automated workflow management and co-design techniques are needed to effectively match the ―impedance" of end hosts with edge networks and wide-area network domains. These solutions must provide

complete co-scheduling across diverse resource types—computational, storage, and networking. Furthermore, advanced virtualization capabilities will also become increasingly important for end-to-end resource partitioning of computing/storage/networking infrastructures for individual communities. These virtualized environments will allow operators do dynamically construct ―dedicated" security enclaves for user groups, simplifying single-sign-on for federated resources and ensuring community-based access control. Indeed many of the above capabilities are very complex and extend well beyond basic end-to-end connection provisioning paradigms. However related provisioning solutions are lacking in currently-deployed operational and testbed networks, and these challenges are unlikely to be addressed within other communities. The DOE must initiate concerted efforts to support the R&D of end-to-end resource co-design/scheduling and virtualization technologies.

4) Performance monitoring: Effective monitoring of exascale data flows will provide invaluable benefits for scientific users and their applications, i.e., fault detection, diagnosis, performance analysis. However this need poses many challenges owing to the large number and high diversity of components involved in the transfer process, e.g., technologies, vendors, and the diversity of network operators. Here uniform and ubiquitous monitoring capabilities will be needed with a genuine *end-to-end* scope. Specifically these solutions must support a rich set of visualization tools and application program interfaces (API). Higher-level visualization tools will help provide a comprehensive, coherent representation of data movement status across all layers, interfaces, and software components. Additionally, intelligent diagnosis, troubleshooting, and performance analysis services (that encompass the end-to-end perspective) will also be needed by end-users and scientific applications to make efficient use of terabit network infrastructures. Therefore the DOE must take a leading R&D role to develop a new generation of proactive and highly-capable network monitoring capabilities. The further integration of these new features into existing solutions, e.g., such as PerfSonar, should also be given priority.

5) Experimental testbeds: Many of the technologies that need to be developed for exascale user/application support will likely be of a disruptive nature. It is important to test these capabilities in dedicated, i.e., persistent, experimental testbed facilities in order to avoid potential disruptions on live production facilities. Testbed facilities must incorporate complete end-to-end setups (i.e., application software, end hosts, storage systems, edge/core networking gear, etc) and allow their users to validate key performance aspects, particularly in the data plane. Further mechanisms and tools to dynamically interconnect different testbeds and build broader ecosystems are also needed to model more diverse real-world scenarios. Overall the DOE community has strong experience in building a range of such facilities and integrating them into its live infrastructure for testing validation.

6) Security: The impact of exascale level data flows on network security paradigms is not at all obvious today. Since many of these flows will likely be transported using circuit-like connection services, their integrity can be provided by means other than traditional deep packet inspection for intrusion detection. In particular the concept of a ―Science DMZ" is one possible means to decouple/isolate large-scale scientific flows from regular site/enterprise traffic. The DOE should fund new research efforts to properly investigate and develop this best-practice approach. In addition, further R&D efforts should also address the scalability challenges for achieving genuine 100 Gbps (and beyond) line rate intrusion detection using a combination of hardware acceleration and massive parallelization designs.

Overall, achieving exascale science will require significant R&D innovations that transcend the conventional boundaries of existing technologies at the networking, computing, storage, and application layers. In many cases, researchers may have to develop altogether new approaches and methodologies that have not been tried before. Meanwhile in other instances, these new innovations may have to be coupled with natural evolutions of current production solutions.

III. Findings and Recommendations of Break-out Groups

The following sections describe the individual R&D areas identified and addressed by the various break-out working groups. These sections support the overall conclusions of the workshop effort and are designed to provide further details for subject matter experts and interested readers.

Group 1: Terabit Backbone, MAN, and Campus Networking

This breakout session focused on the challenges relating to the development and deployment of DOE terabit networking infrastructures within the next 4-8 years. The main objective was to consider terabit networking in the *end-to-end* context, which includes core, regional, campus, and site networks, as well as host end system interfaces. The topic areas here addressed the fundamental technical challenges of scaling existing networks by “100x” from the core to the edge and realizing federated end-to-end terabit capabilities. Another key objective was also to identify the technical hurdles, disruptive technologies, and roadblocks that are on the horizon as transition planning begins for terabit-capable infrastructures. The intent here was to identify these key issues so that the DOE can plan where to best apply its resources in order to address open technical issues and/or speed up timelines for critical capabilities. Along these lines the session was structured to address a series of discussion topics designed to highlight these key technical issues and challenges:

- Revisiting network fundamentals at the extreme scale
- Multi-layer, multi-domain network provisioning
- Federated network services, management, and performance
- Testbeds and experimental infrastructures
- Extension of core network capabilities to end systems for massive data movement

The overall outcome of this session resulted in multiple findings and associated recommendations. These are presented below and organized in the context of the above discussion topics.

Finding 1.1 Core transmission technologies: *The core transmission links in next-generation DOE networks will require an order of magnitude increase in capacity to support future science requirements. This will require a transition from the 10 Gbps per channel capacities commonly-found in today's networks to terabit (Tbps) channel capacities. From a technology perspective, the optical physical layer community (including commercial equipment vendors) is on track to develop and provide such core transmission technologies in the future, as dictated by economic and market factors. However there are still many variables in determining when the economic and market factors will tip in the favor of generally available terabit-speed core technologies. This timeline will be very difficult to predict.*

Significant parallelization will be needed both in the LAN and the WAN (from multiple parallel lanes at physical layer to multiple parallel streams at transport layer) even when taking into account anticipated increases in spectral efficiency of fiber optic transmission based upon current research looking out towards 2020. Current mechanisms to increase spectral efficiency to get to 400 Gbps and 1 Tbps are also feasible but challenging from a commercialization perspective.

> **Recommendation:** While commercial network equipment vendors are on track to develop and provide terabit per channel core transmission links in the future, the economic and market place factors which will drive the timelines are not clear. Likely, a clearer view will emerge in a few years after the introduction and deployment of 100 Gbps technology in the marketplace. The following recommendations are made:

- DOE should closely monitor the development of this technology and its associated market conditions over the next several years, i.e., so as to identify areas where targeted investments may be helpful. The DOE may also find it beneficial to accelerate the progress via targeted investments at specific timeline and technology inflection points.

- Transition scenarios are likely to result in the coexistence of 100 Gbps, 400 Gbps, and full 1 Tbps per channel core links for an extended period of time. This heterogeneous mix and parallelization of capacity and technologies will present several new challenges in terms of managing aggregation and data movement over networks that have diverse link channel granularities. In turn this will require much more sophisticated systems for network capacity planning, monitoring, provisioning, and real-time data flow correlation (than what are currently available). The DOE should conduct R&D efforts to develop technologies to specifically operate in this environment.

- Clever techniques in component integration will need to be explored to enable dense optical parallelization in the future with constrained space and power requirements.

Discussion: Overall there appears to be adequate technical vision, design, and prototype work to believe that 1 Tbps per channel core links will become practically feasible if economics justify. In fact, prototype development and testing efforts indicate that technology enhancements will allow 1 Tbps per channel links to be developed within the target timeframe of this workshop. In particular, these advances revolve around the three key technical approaches, i.e., multiple carriers, increased bits-per-symbol, and increased symbols-per-second. Namely, the use of multiple carriers provides the basis to send several channels over a single wavelength. Meanwhile increased symbols-per-second leverages the increased performance of complementary metal-oxide semiconductor (CMOS) devices following Moore's Law. Finally, increased bits-per-symbol schemes leverage improved modulation schemes. The evolution to 1 Tbps speeds may involve an intermediate development of 400 Gbps per channel transport links, i.e., akin to the recent transition from 40 to 100 Gbps. Carefully note that there are fundamental technical limits and challenges associated with going significantly beyond 1 Tbps speeds on a single wavelength. There are also other limitations when placing more than 20 Tbps of traffic over a single fiber in the currently-installed fiber plant. However, both of these limitations are beyond the projected 4-8 year timeline and exceed the technical objectives of this workshop (and hence should be addressed in future efforts).

Overall observations of the industry’s current transition from 10 to 100 Gbps technologies can provide some key insights into the future evolution for 1 Tbps systems. For example, today’s industry is making good progress in terms of developing commercial products with 100 Gbps per channel link technologies and is expected to have market-ready products deployed over the next several years. However the largest uncertainty for 100 Gbps technology revolves around its economics. Namely, it is not clear when the price points for 100 Gbps technology will justify deployment as compared to multiple 10 Gbps per channel links. This transition will result in network deployments containing a mixture of 10 Gbps and 100 Gbps link rates. Similarly, future 1 Tbps core evolutions will likely embody a hybrid mix of 100 Gbps, 400 Gbps, and full 1 Tbps per channel transport links as well, and these solutions will have to co-exist in the core for a long time. Here the economics and traffic profiles will determine when various network segments are updated, and a clearer timeline of this evolution will only emerge after the wider adoption of 100 Gbps technology. Overall this heterogeneous mix of technologies will present a new set of challenges for the deployment and management of next-generation networks. In particular this heterogeneity will complicate data movement and will require much more sophisticated systems for network capacity planning, monitoring, and provisioning. As a result the DOE will need to continuously monitor the evolutions in the 1 Tbps field and may find it beneficial to accelerate the progress via focused investments at specific timeline and technology inflection points.

It should be emphasized that this discussion is focused on the transport links of core networks, i.e., as opposed to client-side interfaces connecting to routers, switches, or end systems. This is an important distinction since core transport links are typically developed as vendor-proprietary systems and do not necessarily require standards. In turn this allows for much faster development timelines. Overall this flexibility provides added opportunities for the DOE to make an impact on the timeline and availability of technologies in this area. Namely, focused investments by the DOE can yield vendor solutions which can be deployed much faster when there is no need to wait for wider standards to mature. However at the same time this flexibility can cause inter-networking concerns and complications for the other components in the network which are required to be standards-compliant. This presents an overall tradeoff. For example, consider a typical scenario involving standards-based 100 Gbps Ethernet products interacting with new core transmission elements. Here it is generally expected that vendors will develop proprietary 400 Gbps or 1 Tbps core network transmission systems to multiplex and transport traffic from these multiple 100 Gbps Ethernet client-side interfaces. However this is not a major concern as it allows for core network capabilities to scale in a faster manner, i.e., as opposed to waiting for 1 Tbps Ethernet standards to finalize. Nevertheless the DOE will still have to decide when it is more desirable to encourage new standards developments or simply deploy faster proprietary core networking technologies (along with how to manage data flows across these types of infrastructures).

Note that a separate finding also addresses Ethernet standards development and associated data flow management challenges. It is noted that these broader issues will likely influence decisions and plans related to core transmission technologies and investments.

Finding 1.2 Client-side terabit technologies: *Improvements in client-side interface speeds are dependent upon the standards process. Although 100 Gbps Ethernet standards are in place today, there are no specific timelines or plans to develop faster 400 Gbps or 1 Tbps Ethernet standards. As a result, even though there may be environments in which the network core is terabits-capable, data movement across higher layers and/or inter-domain boundaries will require multiple parallel 100 Gbps interconnects.*

Recommendation: Significant optical and component integration will be needed to create client-side interfaces that are commercially-feasible by learning from multi-lane parallelization in 100 Gigabit Ethernet client interfaces today. On top of the lower layer parallelization, DOE should prepare for hybrid link/port speed environments with multiple, aggregated physical links to move a collection of end-to-end data flows across the cyber infrastructure.   These efforts should include managing transitions across layer and inter-domain boundaries and also address other concerns, i.e., basic performance, single-flow striping, reordering, packet loss/recovery, and load balancing.   Research should also be conducted to evaluate techniques to handle flows that may exceed the physical capacity of a single channel or link.

Discussion: While line-side terabit-speed technologies will likely progress on a schedule largely driven by economics and customer requirements, this will not be the case for client or edge interfaces, i.e., those interfaces connecting network elements from different layers or vendors.

In general these interfaces will largely be standards-based. A common example here would be the interfaces that a router uses to connect to underlying DWDM, optical transport network (OTN), or carrier Ethernet transport systems. In addition the interfaces on end systems, between routers, and at inter-domain boundaries will also fall into this category. Currently 100 Gbps Ethernet interface standards are complete, but there are no specific timelines or plans to develop faster 400 Gbps or 1 Tbps standards

In light of the above there will be class of client link speeds which will be limited to 100 Gbps for an extended period of time. It is estimated that faster client-side port speed standards will take at least 5 years to emerge after the standards process has been initiated, followed by another 2 years for price-points to decline and make it economically feasible. Overall this timeframe may not pose too much of a limitation for host interfaces, as there are many other unresolved challenges in terms of getting data in and out of a host at those speeds. However the following categories of connections will still face limitations and drive network architecture and management decisions for terabit networks:

- Router and switch connections to core transport systems (DWDM, carrier Ethernet, OTN)
- Router-to-router inter-domain connections
- Lower-layer inter-domain connections, e.g., multi-protocol label switching (MPLS) label switch routers (LSR), Ethernet switches, etc

It is expected that these connections will be restricted to 100 Gbps speeds (over a single physical link) well beyond the time that core transmission links scale to full 1 Tbps rates. However for peering networks using the same vendor equipment, workarounds may be developed to achieve lower-layer peering in excess of 100 Gbps.

Overall, the above factors will likely result in environments in which the network core is terabit-capable but data movement across higher layers and/or inter-domain boundaries will still depend on multiple parallel 100 Gbps interconnects. This interconnection will mandate more sophisticated data flow management and network control systems to intelligently control end-to-end data transfers across different network infrastructure elements, i.e., issues such as single flow striping, reordering, packet loss/recovery, and load balancing. Finally, the DOE may consider trying to accelerate the development of 1 Tbps standards, but this evolution may be subject to wider market forces.

Finding 1.3 Hybrid network traffic engineering: *The co-existence of packet and multi-layer circuit-switched services in next-generation networks will require an advanced class of higher-level functions to flexibly move data flows between these services, i.e., hybrid network traffic engineering. Specifically this class of functions will need to operate at the nexus of the advanced network services; the observation, monitoring, and prediction of user data flows: the knowledge of user requirements; and real-time feedback mechanisms regarding end-to-end performance.*

Recommendation: The DOE should invest in the development of "hybrid network traffic engineering" applications which can apply knowledge of data flows and network capabilities to flexibly move data between packet and circuit paths through the network. The typical use case is expected to be management of specific IP flows across lower-layer provisioned router bypass connections. However any application or client data flow could utilize these services.

Discussion: Previous findings have outlined the importance of developing mechanisms to enable the flexible movement of data flows across and between network technology layers on both an intra and inter-domain basis. These findings detailed the capabilities that next-generation networks will require in order enable the efficient use of network resources and manage data flows in accordance with a diverse set of user requirements. These functions will be tightly coupled to the network elements and control mechanisms. As a result this capability set will need to be exposed to higher-layer applications in the form of a Network Service Interface (NSI). This service interface will be a well-defined access point for higher level services to utilize these multi-layer services.

The concept of hybrid networking is identified here in the context of using this multi-layer network service interface to conduct sophisticated management and movement of data flows across the various layers in a flexible and dynamic fashion. Hybrid networking agents would be responsible for operating at the nexus of the advanced network services; the observation, monitoring, and prediction of user data flows: the knowledge of user requirements; and real-time feedback mechanisms regarding end-to-end performance. Hybrid network traffic engineering applications are also needed to combine this information and network capability sets into an intelligent data movement and layer mapping system.

Finding 1.4 Multi-layer network control and management: *Next-generation networks will continue to be built using multiple layers of protocols (point-to-point DWDM lightpaths, OTN switching, pseudo-wires, MPLS, IP routing, etc) along with a variety of physical transmission link, i.e., 100 Gbps, 400 Gbps, and 1 Tbps per channel links. The trend towards consolidation of layers into smaller integrated and more capable physical devices will also continue, thus providing a larger degree of control and flexibility. Furthermore, the 100 Gbps optical layer, currently statically provisioned, could become manageably dynamic as the optical technologies push the boundaries of ultra-long distance transmission from coast to coast. New techniques and intelligent control systems will need to be created to manage data flows across and within technology layers for cost-effective utilization of the network resources and improving the user experience.*

Recommendation: The DOE needs to research and explore new technologies that can intelligently manage resource provisioning across multiple network layers for both real-time (i.e., immediate) and scheduled traffic demands. A research agenda should be outlined to develop a capability for multi-layer traffic engineering which spans technology and vendor regions.

Discussion: Next-generation networks will continue to be built using multiple technology layers, with each technology layer becoming more specialized. In other words there will not likely be a trend towards a single network element supporting routing, packet-switching, wavelength-switching, and long haul transport functionalities within a common chassis. Instead economic and power budget issues will drive network elements to become more specialized and increasingly deployed in use-specific manners. Therefore the trend of moving data through a network using a variety of mechanisms (e.g., point-to-point DWDM lighpaths, OTN switching, pseudo-wires, MPLS, routing) will increase. This will push the need for intelligent systems which can move data flows to those technology layers which are most appropriate from a network resource utilization and user experience perspective. Namely, multi-layer network control and management systems will be required to enable the efficient use of network resources and manage data flows in accordance with a diverse set of user requirements, e.g., selecting which layer to provision services, providing mechanisms to dynamically move data across layers, etc. Moreover these capabilities will have to be multi-vendor. Overall such solutions are not available today, though technologies like Openflow that are gaining traction with vendors can be explored to enhance the current approaches.

Also note that the heterogeneity of future networks will be further compounded at the terabit level. For example, as described in an earlier finding, this evolution will likely result in a mixed environment consisting of 100 Gbps, 400 Gbps, and 1 Tbps per channel links for extended periods of time. The practical implication of this is that intelligent data flow management will be required not only across technology layers, but also *within* technology layers, i.e., to account for the diversity of links speeds and traffic demands. Finally, vendor-based GMPLS systems will continue be a key part of future product lines, but their functionalities will likely remain isolated to vendor and technology regions. Now network-to-network (NNI) GMPLS interoperability across technology layers and vendor regions is not expected. However, there are initial GMPLS user network interfaces (UNI) running across these boundaries and these are expected to grow.

To address these concerns, the DOE needs to develop new technologies which can intelligently manage resource provisioning across multiple network layers for both real-time (i.e., immediate) and scheduled traffic demands. Therefore a research agenda should be outlined to develop a capability for multi-layer traffic engineering which spans technology and vendor regions. In particular, native vendor capabilities should be utilized to the maximum extent. However, other technologies should be developed to bridge the gap between technology layers and vendor regions. Multi-layer extensions to Openflow may provide a good base from which to build in this area.

Finding 1.5 Multi-domain networks peering and provisioning: *Dynamic management of multi-layer aspects will be a key necessity and central to network operation in emerging next-generation research infrastructures. As a result inter-domain peering is expected to evolve from current single (router) layer scenarios to broader multi-layer inter-domain peering constructs. Here the same reasons that motivate dynamic management of data flows across technology layers within a domain will also motivate the extension of this model to the inter-domain case.*

Recommendation: The DOE needs to develop new architectures and technologies to support inter-domain peering at multiple technology layers. These new capabilities must consider all technology layers and also address scalability concerns. This sort of thing tends not to be done by industry because competitive advantage and pressures prevent active sharing of resources. In addition, robust authentication/authorization mechanisms and service provisioning interfaces also need to be developed here.

Discussion: Many scientific applications require data transfers across inter-domain boundaries as a standard mode of operation. This is very common in larger research projects where data often originates or is destined for sites not directly connected to DOE research networks. For example, more than 85% of ESnet’s traffic is to or from sites – like the LHC at CERN – that are on other R&E networks. In fact the norm is that the sites are two or three network domains away, not just on adjacent networks. In addition such inter-domain boundary issues also arise when connecting core backbones (e.g., ESnet) to laboratory site networks. In each of these cases there is a requirement for peering across the domain boundaries. Most current inter-domain peering use IP routing setups, providing best-effort transfer of data packets. However in the last several years an additional solution has emerged in ESnet in the form of On-Demand Secure Circuits and Advance Reservation System (OSCARS) provisioned Layer 2 QoS protected paths. In particular, this offering allows an engineered Ethernet virtual LAN (VLAN) path to be constructed across the ESnet backbone and transition across an inter-domain boundary into another WAN network or a DOE site network. Note that these Layer 2 paths are constructed at the routing layer, i.e., by setting up Layer 2 virtual private networks (VPN) across Layer 3 devices (routers). There is also a basic level of Layer 2 inter-domain exchange which enables this type of service.

Overall, the evolution towards next-generation research networks will make the dynamic management of multi-layer aspects a key necessity. As a result inter-domain peering is expected to evolve from current single (router) layer scenarios to broader multi-layer inter-domain peering constructs. Here a typical inter-domain boundary configuration may include the standard router-to-router setup (BGP peering) as well as a direct connection at the DWDM, OTN, or Ethernet layers. However, achieving this functionality will require the ability to exchange lower-layer reachability and connectivity information, something that is not possible today. The OSCARS control plane has already provided a good start on this in terms of inter-domain peering for Layer 2 service provisioning. Leveraging this, a more generic set of inter-domain peering and provisioning capabilities will need to be developed for emerging terabit networks. In particular these new capabilities must consider all technology layers and address scalability concerns. Robust authentication/authorization mechanisms and service provisioning interfaces also need to be developed here. Along these lines OSCARS can interact with additional control systems like OpenFlow to realize these extended peering capabilities. The notion of a multi-layer exchange point may also be a concept worthy of further investigation.

Finding 1.6 Protection and restoration at terabits: *The potential amount of data that can be lost due to network failures is going to increase significantly with the transition to terabit networks. Current approaches to protection and restoration requires network designers to set aside unused network capacity, in some cases dedicated capacity, to be used in the event of failure. This unused capacity is not only expensive to deploy and operate, but also consumes energy as it idles.*

Recommendation: A clean-slate approach to network protection and restoration—one that reduces both capital and energy costs—will be very helpful for end-to-end deployment of such capabilities. The DOE has started to fund some research in this area and should continue to focus on ―usable” clean-slate solutions.

Discussion: Error-free and reliable networks are crucial necessities for many scientific experiments and data transfer applications. For example some instruments such as the LHC stream data 24/7 and in case of service outages, can cause receiving sites to fall behind and never catch up with newer data. The currently-deployed protection and restoration capabilities within DOE networks mostly center around physical interface and physical/virtual circuit protection. In particular this is done as part of the (static) network design process and hence does not necessarily reflect the importance of the ―data” or the flows traversing the links. More flexible protection and restoration paradigms are needed to provide higher reliability during critical data transfer periods. Meanwhile at other times, network capacities can be leveraged for other purposes, thereby improving resource efficiency. Indeed, such flexible ―flow-optimized” survivability architectures are one of key components of a clean-slate protection/restoration design. This topic can benefit from research funding support from the DOE, as long as practical solution implementations are also considered.

Finding 1.7 Network virtualization: *Virtualization technology is an important emerging capability which is primarily focused on processing resources today. Expanding the virtualization paradigm to include network resources could provide many benefits to the DOE user community. Namely, this would* *facilitate the communities need for "purpose built networks" which are tailored to a specific use case and sets of performance requirements.*

Recommendation: DOE should conduct research to enable network virtualization with the longer term goal of integrating processing, storage, and networking virtualized resources into domain-specific topologies and configurations. This capability should include the ability to virtualize across one or more of a networks technology layers, and should also include virtualization of the network control and management systems.

Discussion: An overarching theme expressed by the DOE advanced network user community is the desire to have multiple "purpose built networks" which are tailored to a specific use case and set of performance requirements. There are multiple motivations for these requests, including issues such as problematic coexistence of large and small packets inside the network, security concerns, and being constrained by transport protocols that have to share the network fairly with traffic using other protocols. The multi-layer construct of next-generation terabit networks lends itself to providing this type of virtualization capability. Specifically, dynamic management and control across multiple network technology layers would allow "virtual networks" to be created via the provision of topologies based upon dedicated resource provisioning. The long term goal should be the integration of virtualized network, storage, and compute resources. Research should be conducted into how best to architect and build a network virtualization capability.

Finding 1.8 "Service transparent" networks: *The "services" available across the WAN via current advanced networks are limited to IP routed services and in some cases Ethernet VLAN-based Layer2 services. These are the services as presented at the edge of the network, i.e., to the client attachment point. There is interest in expanding this set of available services to include capabilities such as InfiniBand, Small Computer System Interface (iSCSI), RDMA, domain-specific packet formats, and possibly others. Currently, there is no architecture or vision to provide a wide-ranging service portfolio such as this across WAN infrastructures.*

Recommendation: The DOE should investigate the depth and breadth of the need for these types of diverse service offerings. If a strong case is identified for a wider set of "network services", further research work should be conducted to determine feasibility and possible architectures to realize this capability.

Discussion: While internal network technologies and architectures can vary from network to network, the "services" available across the WAN via current advanced networks are limited to IP routed services and in some cases Ethernet VLAN based Layer2 services, i.e., as presented to the client attachment point. It should be emphasized that these edge services may be encapsulated or transported via a variety of mechanisms across different networks. For example, some networks may encapsulate Ethernet VLANs into router based MPLS LSPs, others may map VLANs directly onto DWDM-based wavelengths. The key point here is that the service offering at the network edge is presented as an Ethernet VLAN in this case, independent of the specific networks transport technology. Emerging terabit networks are expected to retain their overall multi-layer, multi-technology architectures. The opportunities to utilize a variety of transport technologies to encapsulate a multiple common network service offerings will remain. In addition, the emergence of OTN is specifically designed to provide a set of multi-service features for next-generation WAN transport. However, there is currently no architecture or vision to provide a wide-ranging service portfolio such as this across DOE WAN infrastructures.

As a result further investigations should be conducted to determine what additional "network services" are needed by the DOE science community. Examples may include a requirement to transport native InfiniBand frames across the wide area between two edge-connected datacenters. Another possibility may be to transport block data (i.e., iSCSI), RDMA based-transfers, or other domain-specific formats as well. Based upon a more detailed requirement analysis, specific network architectures and technologies can be evaluated. It is expected that R&D beyond what will be readily available from commercial providers will be needed, particularly in the area of adapting data transfers at the network and client attachment points.

Finding 1.9 Next-generation network architectures to bridge the end-to-end gap between campus and core network capabilities: *Incompatible network architectures, protocols, and capabilities in the campus, regional and core networks significantly hamper the performance of end-to-end data flows. New and innovative network paradigms and architectures are needed to improve the management and optimization of end-to-end data flows in the next-generation terabit networks.*

Recommendation: The DOE should continue R&D efforts to look at alternate network architectures to bridge the gap between the network and traffic management capabilities in the campus and core network domains. These efforts should consider local site networks and focus on solutions to specifically tune and focus resources for efficient and high performance data movement to/from the WAN core. The few proposals include architectures that have dedicated data centers directly connected to the WAN at campus edge or flow control mechanisms like Openflow that provide granular flow management on devices in the campus. This expands on the ESnet concept of dedicated data transfer nodes (DTN) to a more capable dedicated data transfer facility (DTF) concept.

Discussion: While next-generation WAN setups are expected to follow multi-layer constructs, i.e., switching and data movement across several technology layers, a different architecture may make sense for sites or edge-connected data centers. Namely, with the advent of 100 Gbps Ethernet technologies, a data center (or set of dedicated end system resources) sitting at the edge of the WAN may represent a good model for moving data onto and off of the network core. This edge-connected facility can be modeled using a fully-meshed data center architecture and can be further tuned to optimize end-to-end data flows.

Finding 1.10 Federated network services, management, and performance: *There are many challenges in terms of developing new network management/performance tools/services (or scaling existing ones) to work efficiently in end-to-end federated hybrid terabits networks with cross-layer support. Moreover current concerns relating to federated monitoring data exchange, data collection and abstraction, and real-time troubleshooting will increase exponentially with the transition to terabit networks. Namely, these challenges will result from both the speed increase and added complexity of emerging multi-layer data flows.*

Recommendation: The DOE should conduct R&D to build technologies which enable current federated monitoring and measurement systems to scale to future (federated) terabit networks. A special emphasis should be placed on integrating the native operations and management (OAM) features across vendor and technology layers, in order to greatly increase the real-time nature of the information and system response. A goal should be set to support real-time "network situational awareness" spanning across network layers and federated domains.

Discussion: DOE network infrastructures embody a complex mix of autonomous local, national, and international network systems. These setups collaborate to deliver end-to-end performance that is several orders of magnitude higher than what the commercial best-effort IP Internet offers. However there are many challenges in terms of developing new network management/performance tools/services (or scaling existing ones) to work efficiently in end-to-end federated hybrid terabits networks with cross-layer support. In particular, some of the key provisions that need to considered here include 1) multi-layer capabilities that are accessible by scientists at the application level and by network engineers at all network layers, 2) inter-domain state exchange, 3) access control, 4) flow manipulation and control, and 5) policy coordination. In addition DOE settings impose added stringencies as applications have to have predictable, repeatable performance as well as real-time debugging/troubleshooting support during live experiments. The combination of operational environments, user requirements, and federated networking infrastructures presents some serious technical and policy challenges for terabit networks. These challenges are further compounded by the sheer scale of data being transferred end-to-end across multiple network layers.

Many vendor-based OAM solutions are starting to introduce some new capabilities, and these will be very helpful in this area. In particular new carrier-grade OAM systems already support critical features such as multi-layer correlation of errors and faults, multi-domain circuit error correlation (tandem connection monitoring), etc. In addition, further integration into the network elements provides many benefits in terms of achieving real-time monitoring and debugging. These types of technologies are mainly present in carrier class Ethernet, OTN, and DWDM systems. Although these new features do not solve all of the problems faced by the DOE in terms of federated service management and performance monitoring, they still provide some key additions which can greatly-enhance the real-time capabilities.

Overall, concerted R&D efforts are needed to address all of the issues relating to terabit networking monitoring. Specifically, these initiatives must study the integration/correlation of monitoring data across multiple network layers, federated monitoring data exchange, as well as user interfaces to relate monitoring data to application-specific flows. Existing federated monitoring systems, like PerfSonar, provide a strong base from which to build such capabilities. However these settings will still demand a much higher level of integration with vendor supplied monitoring/fault data along with rapid correlation of this data across multiple vendors/technologies. The further abstraction of this data to user-level interfaces is also vital in order to enable real-time network situational awareness at the application layer.

Finding 1.11 Testbeds and experimental infrastructure: *Testbeds are an important component of the overall innovation cycle. Given the many technical challenges associated with the migration to terabit networks, testbeds are expected to play a critical role in developing the required technologies. It is very difficult to expect one testbed to meet all research needs, even as discussed within the workshop. Thus, purpose-built, persistent testbed setups which provide flexible mechanisms for users to reconfigure and connect in external facilities and equipment will give the largest benefit to the users and technical development efforts.*

Recommendation: The DOE should maintain a persistent testbed effort which is focused on addressing the key technical challenges associated with moving to the terabit level (several of these challenges are listed in the discussion section below). In addition, testbed mechanisms and polices should be developed to allow other research and commercial laboratories to federate and connect in for short-term specially-focused testing runs. Federating with other research testbeds, e.g., such as the NSF Global Environment for Network Innovations (GENI) setup, should also be evaluated.

Discussion: Testbeds are identified as an important component of the innovation cycle. A key point noted here is that simulation and testbed experimentation needs to be done jointly as each has its own specific strengths. Namely, testbeds allow for the evaluation of physical features, control/provision of real equipment, and resolution of issues across multiple administrative boundaries. These are details that simulations may simply gloss over. Meanwhile simulation allows testing at scales which may not be possible with physical testbeds. The consensus is that testbeds should be an important component of the transition to terabit networks.

In addition it is also observed that many companies, universities, and research laboratories already operate testbeds of varying sizes and capabilities. The availability of mechanisms to dynamically inter-connect a set of testbeds for specific testing purposes would greatly facilitate interaction and development. Namely developing an ecosystem in which self-forming groups of testbeds could come together for a specific testing event and then disband would help to encourage wider collaborations. Furthermore two testbed models are identified, 1) a testbed which is always available and maintained by dedicated staff, and 2) a testbed that is setup for a specific use case by joining a federated testbed group (and disconnected after accomplishing the testing objectives). In an ideal world, both of these models could co-exist, i.e., a persistent testbed can be set up to focus on key technical challenges and a loosely-organized group of other testbeds used to federate with this persistent testbed (or self-organize and federate with other testbeds) for a specific purpose. In particular, persistent testbeds can focus on addressing the most difficult technical issues, but can still be designed with enough flexibility to be reconfigurable over time, i.e., address ever-changing problems. Overall the following terabit network focus areas are identified for the persistent testbed component:

- Host system data transmit/receive/processing: Testing and development to allow hosts to receive, process, and transmit at 100 Gbps single streams and beyond (multi-stream).
- End-to-end performance: Testing and development for protocols and end systems which can perform at terabit speeds.
- Multi-layer network control and provisioning: Testing and development of protocols and systems to enable multi-layer control.
- Multi-domain network control and provisioning: Testing and development of protocols and systems to enable multi-domain, multi-layer control.
- User applications: Testing and development of user applications with a special focus on those which integrate dynamic network services into their workflows.

Note that joint testing initiatives between the DOE and vendor communities can also be of much benefit to both sides. However the concerns with pursuing this have less to do with technical challenges than they have to do with issues relating to proprietary information and confidentiality of test results.

Finding 1.12 Extension of core network capabilities to end systems for massive data movement: *End system limitations continue to be a bottleneck in terms of moving data on and off the network. This problem will be an increasing concern as the evolution to 100 Gbps Ethernet tributaries unfolds. However many techniques can be utilized to circumvent these issues, i.e., such as parallelized data transfers, careful mapping of network flows to dedicated processor cores, and minimizing movement of data to disk.*

Recommendation: The DOE should invest in disruptive "outside-the-box" paradigms in order to scale up data flow speeds in and out of end systems. These efforts can consider many possible approaches, several of which are listed in the discussion section below. The overall end goal here should be to develop new line rate mechanisms for moving data between the network and end system memory or disk storage.

Discussion: End systems limitations continue to be a bottleneck in terms of moving data on and off the network, and this problem is expected to worsen with the move to faster 100 Gbps Ethernet speeds. As a result there is a clear need to develop techniques that can boost the speed at which data can be transferred between the end system and network. Along these lines many different strategies can be considered, e.g., including parallelized data transfers, careful mapping of network flows to dedicated processor cores, and minimizing movement of data to disk. Additionally, further adjustments to existing interconnect technologies (such as 100 Gigabit Ethernet PCIe and Infiniband) can also be considered to improve performance, e.g., varying Ethernet maximum transmission unit (MTU) sizes. However these strategies would obviously require further compatibility changes on the network side. Note that it is crucial to consider all host-level improvements within the wider context of WAN data transfers. In turn this requires any new or modified network services to be factored into emerging R&D plans.

Finally, it may timely for the DOE to pursue some R&D activities in more disruptive —outside-the-box” solution paradigms, i.e., such as rapidly moving data directly from network interfaces to user memory, using integrated photonics approaches, developing modified block data movement techniques over the WAN core, etc.

Finding 1.13 Multicast and broadcast service models: *The era of exabyte-level data requirements and terabit networks gives reason to revisit one-to-many (broadcast) and many-to-many (multicast) service models. Current network deployments generally do not support broadcast and multicast, i.e., except for limited instances where IP multicast is utilized for non-reliable, non-critical purposes. As a result the potential application uses for multicast and broadcast go largely unserved. The uses could include one-to-many large file transfers, content distribution systems, real-time data distribution, amongst others.*

Recommendation: DOE should evaluate the potential application use cases and available technologies to support multicast and broadcast service models in next-generation networks. These service offerings should go beyond the standard IP multicast services available today and should address scalability, reliability, as well as integration and co-design with special-purpose applications and dedicated resource data transfers. This may require evaluating options for multicast/broadcast at technology layers others then IP.

Discussion: Current network deployments generally do not support broadcast and multicast except for limited instances where IP multicast is utilized for non-reliable, non-critical purposes. As a result, the potential application uses for multicast and broadcast go largely unserved and it is important to revisit one-to-many and many-to-many service models, i.e., beyond what is typically provided by IP-based multicast. This may require evaluating other options for multicast/broadcast at technology layers others then IP. For instance DWDM, synchronous optical network (SONET), and OTN systems often support a "drop-and-forward" features which, with extensions, could provide the basis for a DOE-specific one-to-many communication infrastructure. In addition, there are other technologies utilized in the consumer broadband industry, such as passive optical networks (PON), which may offer some approaches which are applicable to DOE environments. Emerging IP-based reliable multicast technologies such as pragmatic general multicast (PGM) should also be evaluated for applicability in these scenarios.

Finding 1.14 Large packet sizes: *The design decisions on the current Internet were based upon hardware and transmission constraints that existed a few decades ago. New packet and network paradigms need to be investigated in order to transport large, long-term flows that are typical in DOE science environments.*

> **Recommendation:** The current paradigm of processing and switching packets can be a bottleneck, requiring specially-built application specific integrated circuits (ASIC) and memory in network devices. However for large long-term flows, these architectural choices may not be the most ideal from a cost, performance, and efficiency perspective. Therefore new paradigms must be explored that are targeted towards improving the end-to-end large-flows paradigm, i.e., including the possibility of supporting very large packet sizes and corresponding network system architectures looking beyond conventional routers with store-and-forward packet processing.
>
> **Discussion:** Packet header processing introduces large overheads on existing network hardware like routers and switches. Moreover, for 1 Tbps rates and large end-to-end flows, processing each packet header mandates the development of custom, power-hungry ASICs and expensive heat-generating high-speed memories, i.e., to store and process packets at high-speeds. The current constraints on network design are driven by historical limitations and the fact that the same network equipment delivers all types of Internet services, i.e., from real-time voice to large data transfers. However purpose-built networks also may finally be possible with new advances in network virtualization and slicing. Therefore new paradigms for supporting very large packet sizes (over 64 kbytes) for large flows should be investigated given the current advances in technologies. In addition, new system design mechanisms should be studied to evolve buffering and packet header processing architectures to move beyond existing store-and-forward designs.

Finding 1.15 Network co-design: *As the Internet grows to serve the diverse communication needs of billions of people, purpose-built networks and architectures may be needed to serve specialized communities. The need for these purpose-built networks (either virtual or physical) to realize optimal end-to-end performance can only be achieved by adopting the concept of co-design between the application, computing, networking, and storage communities.*

> **Recommendation:** The DOE should encourage and fund co-design activities among the various science disciplines (i.e., users, applications), computing/HPC, storage, and networking research community. This co-design approach is a great methodology to meet the end-to-end-user and application requirements in the most optimal manner.
>
> **Discussion:** The current set of Internet protocols have allowed the architecture to scale cost-effectively. However, all new innovation and services have been offered using the same base infrastructure. Even though the Internet meets the need of the majority of the users, certain specialized communities can do much better with purpose-built networks. This requirement has caused the creation of ad-hoc network federations like Dante Internet2 CANARIE and ESnet (DICE) and the Global Lambda Interchange Facility (GLIF) to better serve the specialized needs of the R&E community. With advances in network virtualization techniques, these purpose-built networks do not have to follow the fairness or design constraints of the wider Internet. Moreover these specialized designs can also be built as well as deployed cost-effectively and dynamically over existing physical infrastructure. Leveraging co-design techniques and the benefits of network virtualization, one can imagine building virtual network capabilities that meet the specific application requirements effectively without increasing the cost dramatically.

Finding 1.16 Network simulations to scale testbeds: *Physical testbeds supported by commercial or research-funded equipment and software may not scale to test all the scenarios at the exascale or Internet scale, i.e., especially failure scenarios.*

> **Recommendation**: An intelligent network simulation capability (i.e., one that leverages computer virtualization capabilities and can interface into the testbed as a virtual device) can greatly improve the effectiveness of network research. This is particularly amenable for testing certain control plane algorithms at scale. However this simulation methodology will not be suitable for physical layer or data plane throughput testing. Overall, the DOE should interact with industrial partners and build an advanced network simulation platform that allows multiple device ―clones” to be built and deployed easily, i.e., very similar to the cloud paradigm.

Discussion: Current network simulators used by universities do not accurately reflect the behavior of a real network. Conversely, real-world testbeds are very effective in testing protocols and physical layer interactions, but are limited in complexity and scale by the availability of resources, both capital and operational. An effective simulation platform that can augment a physical testbed can be a huge asset to prove network research at scale before it is deployed. Namely, this platform should interact intelligently with the deployed testbed and support a configurable number of virtual nodes, i.e., providing a near-replica of the hardware being deployed. Elements of such simulation platforms are already being used by various commercial vendors to augment lab testing of their products. A similar approach, if funded, can provide an excellent resource for the DOE networking community, i.e., particularly for operators like ESnet, to help test their ideas out at scale.

Group 2: Advanced User Level Network-Aware Services

This breakout session focused on the *end-to-end* perspective for users leveraging terabit networking capabilities, and the challenges and opportunities those capabilities would provide. A key goal here was to address the interactions of users and applications with underlying hosts and networking systems (issues addressed in the other two breakout sessions) in order to migrate to terabit/second data flow environments. Along these lines the session was structured to answer a series of questions that built upon each other to address how users might make effective use of DOE's emerging terabit network capabilities. Namely, these enquiries addressed the fundamental challenges in developing network services for terabit network infrastructure, what and how terabit network resources need to be exposed to make efficient use of them, and what higher-level services were needed to develop a productive, understandable user experience environment. Here the actual definition of a —user" was recognized as being sufficiently undefined so as to potentially cause confusion. Two categories of users were identified, i.e., scientific applications and the people who make use of them. In particular user experience issues and questions were recognized as applying to the latter, whereas integration with terabit network services and middleware more broadly grouped with the former.

Overall, the following topics were discussed:

- Network users and applications' experience at extreme scale
- Exposing network capabilities, resources, and security policies to end users and applications
- Enhancing network users' experience with network monitoring and performance diagnosis services
- Automated and intelligent workflows
- Community-based advanced network services
- Access control and authorization in advanced network services

The outcome of this breakout session focused on these 6 topics and resulted in the following findings and associated recommendations.

Network Users and Applications' Experience at Extreme Scale

Finding 2.1: *Terabit networks will be far richer and more complex than today's networks. This complexity will be due in part to end-to-end paradigms, increasing use of parallelization, wider reliance on multi-layer technologies, and increased data component counts.*

> **Recommendation:** It is crucial to develop metrics, measurement tools, and methods to quantify the complexity of networks and end-to-end services, i.e., so that different architectures and systems can be effectively compared and contrasted in terms of complexity. Designing and integrating terabit services and tools will be challenging. To mediate this complexity, coordinated design (co-design) methods need to be applied to new network software, hardware, and services in conjunction with all components in the end-to-end paradigm.
>
> **Discussion:** New measurement, troubleshooting, and diagnostic tools for all layers of the terabit network infrastructure are needed to deal with this complexity. Automated monitoring and diagnosis systems that make use of these tools are also required to capture what is happening across terabit infrastructures. It is often claimed that the complexity of networks and services increases with time. Although this is likely true, one still needs to develop a more systematic and quantitative framework to evaluate the complexity of networks and end-to-end services. For instance is it always true that the complexity of a terabit network will be higher than that of a

gigabit network? Alternatively, given two architectures or systems that offer the same functionality, how can one effectively compare and contrast their complexity? Clearly it is important to develop metrics and methods to answer these questions in a scientific and quantitative manner. Furthermore the DOE also needs to develop advanced simulation tools to analyze different network technologies in term of their performance and complexity. In turn these simulation tools can be used to verify the accuracy of those developed metrics and methods, and tune and improve them in an iterative matter. Overall a better understanding of the sources of network complexity may provide a significant advantage in design choices moving forward, i.e., as well improving the overall reliability and usability of terabit networks.

Given the complexity of terabit networking layers and the requirement that all components must work together in order to deliver network service guarantees, it is very desirable to co-design new network transport protocols along with the user services that make use of these protocols. Here trade-offs between software on-load and hardware offload, i.e., via remote direct memory access (RDMA), need to be carefully balanced within the recommended co-design. Finally, it is highly desirable that the co-design of network services take into account all of the end-to-end components, i.e., including terabit backbones, campus LAN's, end hosts, storage systems, and high level scientific applications.

Finding 2.2: *DOE's data-intensive science applications (i.e., bio-energy combustion, high-energy and nuclear physics, earth/climate science, visualization tools), combined with exaflop computing systems, are expected to generate, analyze, and distribute data volumes on the exascale level. These applications will be critically dependent upon advanced terabit networks to move their enormous datasets between local and remote computing and storage facilities.*

Recommendation: It is highly recommended that the DOE networking community assist in the co-design of exascale systems and applications. The scope of this effort should include enhancing exascale configurations, debugging, management, and also tuning capabilities. There is also a further need to design and develop highly-automated/integrated network services and high-level data transport middleware in order to make effective use of terabit network capabilities. Indeed, easy-to-understand user interfaces and integrated user environments can significantly improve scientific productivity with exascale machines and terabit networks.

Discussion: The data volume from DOE-funded applications, experiments, and simulations will continue to grow into the exabyte range in the next few years. In addition, exascale supercomputers will generate large amounts of data to be analyzed, and will require commensurate network services to move this data in/out. The performance of network I/O for supercomputers is a major challenge, particularly in the era of exascale OSs. Therefore, associated exascale OS solutions need to be more efficient and support faster network I/O mechanisms in order to match the very high bandwidths sourced by terabit networks. Due to memory- and heat-wall problems, computing has shifted towards a multi-processing approach. Namely, multi-core and many-core technologies have been widely applied in the computing industry. As a result it is also crucial to re-design and optimize exascale network I/O subsystems with these emerging technologies. Furthermore network I/O subsystems need to provide an efficient means of data assembly, scheduling, and validation in order to support data movement between storage systems and terabit networks. Therefore it is vital for exascale software and hardware designers, application software developers, and network infrastructure implementers to work together to ensure that the required I/O capacity can be realized for data movement, i.e., for local file systems and remote machines.

Large-scale scientific collaborations will require highly-automated/integrated network user services to support their applications. These services must effectively hide the complexity of the underlying network layer and also offer accurate and predictable network performance. As such, new services must provide automated monitoring and problem diagnosis, as well as the ability to react to and remedy problems within the network infrastructure. Again it is not necessary to expose users to these advanced service capabilities. Meanwhile transport middleware (such as Globus-online) will need to be built on top of the advanced network services in order to ease the complexity of exascale data management and make effective use of network core capacities. Finally, easy-to-understand user interfaces and integrated user environments need to be developed to allow users/applications to leverage a simple set of parameters to configure complex networks for large-scale data transfers that will all result in increased scientific productivity.

<u>Exposing Network Capabilities, Resources, and Security Policies to End Users and Applications</u>

<u>Finding 2.3:</u> *Open Science Grid (OSG) and Earth System Grid (ESG) are two grid middleware and software stacks used by DOE data-intensive science to interact with high performance networks. The OSG and ESG will be the interface layers between a range of users (i.e., in high-energy physics, nuclear physics, and climate science) and the underlying terabit network.*

<u>Recommendation</u>: The networking community needs to provide clean and well-defined API and services that will expose the underlying network capabilities to OSG and ESG users and bind these middleware stacks (OSG and ESG) to the underlying network capabilities and resources.

<u>Discussion</u>: OSG and ESG are funded to advance science through the concept of open distributed computing. Namely these initiatives are based upon partnerships which federate local, regional, and national science facilities to meet the needs of the research and academic communities at all scales. Furthermore they are built on top of a networking infrastructure capable of effectively interconnecting locations with sufficient capabilities and capacity to meet the needs of their supported academic communities. To make this federated model work, a high performance and reliable network becomes critical because either users explicitly require significant inter-site bandwidth to move/manage datasets in the terabytes-petabytes ranges, or systems dynamically identify opportunistic resources and dispatch jobs, send input data, and collect output results. However both of these projects have focused on building the middleware needed for their members to participate and also provide security, while networking-related developments were left to ―satellite” projects rather than being their core activity. Focused network research and development are needed to fill this gap and meet the requirements of OSG and ESG. Under these scenarios it is important to now provide transparent and effective networking capabilities for such middleware users, i.e., without requiring them to become seasoned network engineers. It is also necessary to provide a well-defined set of user interfaces and services to be used by the OSG/ESG in order to discover what is available, make reservations, bind available networks into their own capabilities in OSG and ESG, and provide applications with services that OSG/ESG cannot provide alone.

<u>Finding 2.4:</u> *Data-intensive applications within the DOE have a wide range of networking requirements, e.g., bulk data transfers (high bandwidth, loss-less), climate visualization (large bandwidth, less jitter), and real-time data analysis and decision making (high bandwidth, low latency, and loss-less). However the user's or application's knowledge of the available set of underlying network service capabilities to satisfy those requirements is typically minimal, and often completely lacking.*

Even if the user or application has adequate knowledge of such advanced network services, the absence of simple, uniform, and dynamic authentication and authorization (AA) services can make accessing those services difficult.

Recommendation: Resource discovery services are needed to bridge the gap between what the typical users' applications are aware of, and what services the DOE's advanced network infrastructures can actually deliver. These services need to include not only basic network capabilities, such as bandwidth, quality of service (QoS) guarantees, and schedule availability, but also policy and access constraints as well. The scope of these resource discovery services needs to be end-to-end, presumably as part of federated service offerings extending beyond the DOE network domain. These services would optimally contain as much ―intelligence” as feasible, but limit the details exposed to the application or user to the minimum level practical. The services should be complemented by standard, ubiquitous AA services that control access to the resources they serve.

Discussion: Users are typically unaware of network capabilities, such as bandwidth, QoS guarantees, circuit services, and reservation availability schedules. Moreover, even if users are aware that these advanced services exist, enabling their applications to make use of them can be extremely challenging. As a consequence, there is a gap between what an application typically achieves in terms of network performance, and what the application could achieve if it were able to take advantage of the advanced services that the network infrastructure may be capable of offering. The implications of this gap are expected to become more pronounced in terabit network environments. Therefore increased network performance across terabit networks will almost certainly require the type of advanced network services that applications currently have little awareness of. As a result, resource discovery services are needed to make applications aware of these services. Such services should have the dual task of making applications aware of available advanced network services, but also hiding their complexity. For example, multiple network paths with differing levels of bandwidth, QoS guarantees, and availability might be some options for an application. Here the application does not need to know the topology configuration of each path, but rather only a select set of service primitives. Resource discovery service hides those details, but preserves them for resource reservation purposes. Resource discovery services also would need to be policy and access aware. Moreover, underlying that policy and access awareness is the presumption that uniform and ubiquitous AA services exist to leverage resources required for advanced network services. However such AA infrastructures do not exist today, but will be necessary for effective use of advanced terabit network services.

Finding 2.5: *Multi-/many- core processors are widely used across many applications domains including general purpose computing, embedded, network, and exascale computing design.*

Recommendation: The network community should leverage the proliferation of multi-core technologies to design network protocols and services to improve performance and provide a high-speed transport substrate and service layer for DOE's exascale programs. The technology of protocol and service on-load to multi- and many-core architectures can be complimentary to hardware offloading and can lead to cost-effective implementations.

Discussion: High-end computing systems utilize specialized offload engines to accelerate data transfer, checksum validation, and high-level network protocol processing. Some key examples include RDMA, InfiniBand, and transmission control protocol (TCP) off-loading. These offload technologies have already demonstrated superior performances as compared to their software counterparts. However these hardware acceleration approaches also have several pitfalls. Most notably, the hardware units are either expensive or program development on these hardware unit is proprietary and complex. Furthermore, the manufacturing complexity increases exponentially depending upon the number and complexity of tasks handled. Moreover once off-load hardware is implemented, it becomes very difficult customize it for various environments. For example, RDMA-enabled network interface cards (R-NIC) with Internet wide-area RDMA protocol (iWARP) features demonstrate significant performance degradations for larger round trip delays. Although iWARP can give better performance in LAN settings, it still lags behind its software alternative (i.e., TCP) in WAN because the network card does not have the flexibility to choose different TCP types and dynamically adjust send/receive parameters and buffers.

Moving forward, multi-/many-core technologies will become increasingly cost-effective owing to economies of scales. In order to leverage this when designing network services and protocols, a smaller number of available cores can be used to —on-load” complex tasks that need more dynamic adjustments. Meanwhile low-level tasks can also be off-loaded to hardware, e.g., such as checksum support, direct memory access, and network frame generation. Currently the DOE Advanced Scientific Computing Research (ASCR) Office is already supporting research in threading technology to help scale applications to leverage a large number of available cores. These multi-threading technologies can also be leveraged for exascale network R&D efforts to improve parallelism at the application layer, i.e., underlying libraries for performing task mapping onto cores, load-balancing across cores, etc.

Enhancing Network Users’ Experience With Network Monitoring and Performance Diagnosis Services

Finding 2.6: *The current set of data and metrics used to characterize the performance within components of a scientific application’s end-to-end data movement path are disjoint. Furthermore, tools to represent and analyze this disjoint data and metrics in a comprehensive manner are also lacking. The overall implications of this disjoint monitoring data will become much more severe at terabit data rates.*

Recommendations: New tools and services need to be developed to provide a uniform and ubiquitous monitoring framework with an end-to-end scope. The existing PerfSonar platform provides a foundation on which to implement this framework. However this solution must be adapted to incorporate lower-layer networking technologies (as well as end system and application resource monitoring) to provide a true end-to-end perspective. Additionally this end-to-end monitoring capability needs to be complemented by a rich set of visualization tools and APIs. Such higher-level visualization tools will be needed to provide a comprehensive, coherent representation of data movement status across all layers, interfaces, and software components. Finally, intelligent diagnosis, troubleshooting, and performance analysis services (that encompass the end-to-end perspective) will also be needed by end-users and scientific applications to make efficient use of terabit network infrastructures.

Discussion: Large-scale scientific research projects built using distributed computing models are dependent upon efficient, predictable data movement. However data transfer paths typically traverse a highly-complex set of layers and domains, including application software, middleware, OS software, system hardware, and network software/hardware. Understanding the particulars of this process at any particular time within any particular component is very challenging. Furthermore, understanding what is concurrently happening within all of these components is well beyond the capability of current tool sets. This makes detection, identification, and correction of data movement problems quite difficult. Furthermore, as large-scale projects become increasingly-dependent upon terabit networks and exascale computing facilities, the impact of performance problems will be much more severe. For example, consider that one —bad” second within a terabit data flow will affect over a 100 gigabytes of data. The monitoring tools must capture network misbehaviors and respond to any potential anomalies in an even shorter timeframe to minimize the impact and data loss. However the current set of monitoring data and metrics used to characterize the performance are disjoint and distributed in end-to-end network paths, and it takes a certain amount of latency to collect, propagate, assemble, and analyze these monitoring data to diagnose potential problems. Therefore the impacts of this disjoint monitoring data will become much more severe at terabit data rates because it is difficult to continuously minimize the latency to react to network issues.

Therefore if the efficiency and predictability of data movement within distributed exascale computing systems is to be sustained, real-time performance monitoring capabilities with stringent latency requirements will be needed, including problem detection, diagnosis, localization, and correction. To accomplish this, a coherent monitoring capability of all components (and services) needs to be available. As a result the DOE must develop the set of tools and services needed to provide a uniform/ubiquitous monitoring capability with an end-to-end scope.

Along these lines the existing PerfSonar platform should be leveraged to establish this end-to-end monitoring framework. However this will require enhancement of existing PerfSonar software implementations to incorporate multi-layer network technologies, not just the current focus on Layer 3 measurements and counters. Additionally this framework needs to include the end systems and the scientific applications running on them. Adaptation of the PerfSonar platform to provide comprehensive system parameters (including processor, memory, disk I/O, network interfaces, etc) as well as application and middleware logs will be necessary to extend the monitoring framework to a true end-to-end scope. Active measurements currently supported by PerfSonar servers within ESnet and other R&E networks should also be extended to end systems, at least on an on-demand basis. Optimally, the network infrastructure itself would support these types of active measurements as well.

Note that the simple collection of the various data flow monitoring measurements and metrics will not be enough. The DOE needs to develop a new analysis model as well as visualization tools to make the collection of componentized views of network and host data/metrics coherent and understandable to users, i.e., within the context of end-to-end data movement. Such visualization tools will require APIs to enable the applications themselves to react to unusual or unexpected monitoring results within any component of the data movement. Finally, the analysis and diagnosis aspect of end-to-end data flow movement at terabit rates must also be addressed. Here the DOE should develop intelligent, automated analysis services that provide expected and realized performance feedback to users and their scientific applications. These intelligent analysis services should lay the foundation for remedy or self-correction of sub-optimal data movement by the applications themselves.

<u>Automated and Intelligent Workflows</u>

<u>Finding 2.7:</u> *Applications and users have been migrating from an individual file movement environment towards a large-scale federated content delivery environment. This migration will become more pronounced as terabit networks and exascale computing paradigms start to emerge. Moreover, these* *new environments will encompass a broad set of components and tasks (authenticate, schedule, reserve, monitor, and commit).*

Recommendations: Network-aware, automated, and intelligent workflow systems that can offer resource co-scheduling solutions for data movement and processing applications need to be developed.

Discussion: The DOE needs to develop automated and intelligent workflow systems for data movement applications. These systems should be capable of integrating individual resource scheduling components to coordinate network and storage resources and create appropriate schedules, reservations, and performance predictions. In particular more R&D efforts are needed to build *co-scheduling* frameworks for network and storage resources that not only satisfy resource requirements for data movement, but also ensure efficient resource utilization. Furthermore, such frameworks should facilitate resource discovery by providing the means for individual scheduling components to register the availability of the resources they are brokering for use by co-schedulers. Overall, a successful workflow system should provide scientific applications with a ―fire-and-forget” solution that can initiate complex data movement workflows and relieve users from worrying about what needs to happen next. Finally, to ensure wide-scale adoption of this solution amongst the user community, an integrated user environment (such as Facebook or Web 2.0) would also be needed to enable collaborators to track the status of various tasks and pool their expertise/knowledge.

Community-Based Advanced Network Services

Finding 2.8: *Network virtualization is a potentially transformative methodology. Terabit networks, in conjunction with successful network R&D activities, can therefore bring about new opportunities for scientific communities to make computing and data processing more efficient and secure.*

Recommendations: R&D efforts are needed to build new virtualization capabilities that extend beyond current end-to-end networking paradigms which focus on a virtual connection between a pair of source/destination nodes. In particular new virtualized computing paradigms must allow the co-design and co-scheduling of computer, storage, and network resources of multiple, geographically-distributed end-sites with the goal of satisfying the needs of a scientific community. Such paradigms should facilitate community use by providing single-sign-on capabilities encompassing authentication and authorization control, resource allocation, and network security enforcement. Here a user should not need to provide their credentials more than once in order to be allowed to utilize the entire set of virtualized resources allocated to a specific community.

Discussion: In cloud/grid environments, resources could be federated dynamically for serving large tasks more efficiently. Distributed users and/or small communities with common goals could be grouped into larger virtual organizations as needed. At the same time, in-house data processing could be off-loaded to remote grids/clouds to help reduce computing costs. In fact, it is even feasible for some types of applications to form exascale supercomputers through the federation of computing resources. However such resource federations require reliable virtual networks to ensure the QoS within a federation and reduce interference between multiple federations. While host and storage virtualization is already being widely supported by both vendors and open resource communities, network virtualization remains a key challenge due to device complexity/heterogeneity issues. Typically, interconnecting multiple end-sites involves the coordination of multiple administrative domains and configuration of multi-vendor networking devices. As a result the DOE needs to sponsor R&D of network virtualization middleware to provide users and applications the requisite support to establish virtual environments that allow the dedication of resources to tasks, selection of efficient transport protocols and software, and prevention of interference between other users and applications.

Access Control and Authorization in Advanced Network Services

Finding 2.9: *While the best-effort Internet requires no AA, this access control model does not work in terabit networks supporting advanced network services. As a result, heterogeneous access control policies and AA mechanisms have been adopted by today's existing network service providers, but their optimality and efficiency have not been well studied.*

Recommendations: R&D efforts are needed to develop AA platforms to support various access control policies required by network service providers. A holistic approach is needed here across various types of resources (computing, storage, and networking) that are co-scheduled for one science task.

Discussion: Best-effort networks require no AA for end users sending and receiving packets to and from the network. This model works well because it greatly simplifies application design and it has effectively ensured the success of the wider Internet in the past thirty years. However this model does not work well in terabit network environments since the volume of data traffic can flood the best-effort Internet and greedy users can potentially abuse systems. For example, consider the scenario for data replication by an OSG user. In order to accomplish this goal, the resources first needed to be reserved in end storage systems (by a storage resource manager), LAN (by LAN control system), and WAN (by ESnet) before Globus-Online can be triggered to do actual data transfers. Each of these service providers can have their own AA policies, i.e., such as Grid Security Infrastructure (GSI), Shibboleth, plain passwords, etc. Interoperability and mutual trust are typically established off-line manually. Moving ahead, a well-defined system is needed to ensure that service providers can satisfy users across local, distributed, and grid/HPC environments. Namely, this solution must define and implement protocols that will allow resource providers to define the access and security policies governing their resources to be shared by authorized users.

Group 3: Terabits End Systems (LANs, storage, file, and host systems)

This breakout session focused on exascale end use cases and understanding the type of research needed to prepare end systems (such as LANs, storage, file, and host systems) to be elements of a system satisfying these use cases. The session also addressed the need for test stands (testbeds) to support this research. The overall outcome of this breakout session resulted in multiple findings and associated recommendations. These are now presented.

High-Level Use Case Development

Finding 3.1: *The growth in the DOE's scientific research activities point to an increased need for large-scale data transfers. These requirements include bulk transfers of data from site to site and as well as specialized use cases with stringent performance constraints.*

> **Recommendation:** The DOE needs to develop ultra-high capacity multi-site connectivity to support the emerging needs for exascale bulk data transfers and remote visualization capabilities.
>
> **Discussion:** DOE scientists need to work efficiently and experience has shown that many want to process data sets on time scales that match their ability to think. The rule of thumb here is ―whatcan be transferred in 8 hours is desired by the scientists." In addition the very scale of scientific collaborations required for fundamental science is growing. For example, simulation science has grown to the point where uncertainty quantification (and exploratory science through ensemble analysis) requires whole families of simulations to be produced and analyzed. In turn, this expansion requires larger communities of scientists to construct, produce, and interpret the data. Additionally, larger groups of scientists are also needed to refine and analyze data generated by the massive instruments at DOE and overseas facilities, e.g., LHC, ITER, etc.
>
> There are several ubiquitous scenarios where scientists need to work remotely from their data, or, conversely need to have data imported to a local computing system. Here remote analysis describes the situation where scientists located anywhere can analyze resident data at exascale or other data centers. Now DOE networks require high-bandwidth dedicated-QoS channels to support such remote visualization and other forms of remote presence. Examples here can include very demanding applications such as lossless high-definition video. The other scenario is distributed analysis, where scientists are using an increasing number of specialized computational tools, e.g., such as map-reduce file systems and database like file stores. Overall, the size of these distributed collaborations, and the number of data instruments and data sources they use, continues to grow. As a result the networking layer must provide high-performance multi-site connectivity, i.e., as the number of coupled workflows increases and scientists leverage diverse resources at diverse sites for a single task.
>
> Note that several analysis models indicate that bulk data movement in DOE will increasingly leverage sophisticated systems for distributed data management, thereby replacing site-to-site file transfers. These setups include caching-based data transfer schemes (similar to commercial content distribution networks but far exceeding them in scale) as well as remote I/O.
>
> **Finding 3.2:** *Real-time selection of resources and emergency re-purposing of DOE facilities requires networking support.*

Recommendation: Emerging exascale computing facilities must provide resource provisioning paradigms and services to support rapid, real-time re-purposing and planning capabilities for large scientific projects.

Discussion: DOE science (as well as DOE facilities) are becoming increasingly responsive to real-time needs. For example, consider the plan to use DOE LCF facilities to support fusion research via participation in shot planning for the ITER tokamak (in France). Such planning is critical as the number of shots at this facility is limited and these also need to occur at a defined operational cadence. Moreover this planning has to be done independent of the availability of any one DOE computing facility. As a result there is a pressing need to rapidly move data to DOE facilities in the USA and also select several available facilities in near real-time. A related but more general goal for DOE facilities is to also support emergency re-purposing, i.e., in order to respond to national emergencies, e.g., such as oil spill modeling, etc.

Finding 3.3: *The historical growth of DOE archive capacity compares favorably to the historical growth of its network capacity. Overall, terabit networking would support such a plan.*

Recommendation: The research agenda should address the design of distributed data archives to improve distributed analysis performance and as well as the overall robustness/recovery of critical experimental and analytical data.

Discussion: Another important programmatic is the increased standard of care needed for scientific data. For example a basic disaster recovery strategy is to keep data in diverse places so that it is protected from local stressor events—such as earthquakes, fires, tornadoes, or national emergencies. An obvious strategy is to take advantage of the geographic diversity of DOE labs and implement redundant archives for significant data. This is a reasonable approach, since historically the growth in networking capacity has tracked growth in data volumes in DOE tape archives quite well. Overall, a distributed archival strategy is also synergistic with the other desired elements of a DOE enterprise architecture for scientific computing, i.e., such as distributed analysis, as mentioned above.

Finding 3.4: *Models that establish baselines or key performance indicators (KPI) for performance and usability can play an important role in establishing expectations and evaluating the performance of research networks.*

Recommendation: Research work focusing on the identification and production of a set of KPI's which are reflective of DOE use cases—and also meaningful to a representative body of scientists, network providers and those in oversight roles—is highly warranted.

Discussion: Many user and applications scenarios point to the use of high-performance (terabit) networks by communities of increasing size. As a result DOE computing paradigms have to become increasingly responsive to real-time needs and provide more flexibility and multi-site (i.e., networked) support. Nevertheless, many scientists today still routinely underestimate the performance and capacity of their research networks. Therefore it is important to give systematic program level attention to study the utility delivered to DOE science activities. Such an activity requires the support of reasonable, high-level KPIs or other metrics that are understandable by a range of personnel, i.e., program people, networking people, and scientists.

Test Stands/Testbeds

Finding 3.5: *Researchers need access to end host and LAN testbeds in order to investigate end system issues. These testbed systems also need governance mechanisms to ensure that their usage is not skewed towards testbed host sites, especially if the number of testbeds is limited. In particular testbeds that can support short bursts of terabit data will likely provide the greatest benefit at the least cost.*

> **Recommendations**: Testbeds, as described in this report, need to be furnished by the late 2011-early 2012 timeframe, and a technical refresh will be necessary in the 2014-2016 timeframe.
>
> **Discussion:** The WAN requirements of emerging DOE exascale computing applications are well beyond those for many current commercial applications. Consequently as hardware transfer rates increase by a factor of ten, terabit networks will require adaptations to local system architectures along with resolution of problems within existing systems. Experience shows there is substantial lead-time in adapting designs and resolving these types of problems. Therefore there is a clear need to deploy end system testbeds ahead of terabit network buildouts. In particular, two focus areas are envisioned here. One is on the performance of individual system components, i.e., such as server computers. Meanwhile the other is on system elements, mostly LAN equipment, operating between end systems and WAN devices. These testbeds should be supported by personnel with networking and computer systems expertise and initial offerings should carry one or more 100 Gbps flows. Future plans should also call for terabit flows as soon as possible, although support for long-lived flows of such scale may not be possible in testbeds.

Host/End systems

Finding 3.6: *As exascale science starts to emerge and systems on a chip (SOC) platforms become a reality, it is likely that commercial networks will not provide sufficient network capability.*

> **Recommendation:** The research agenda should allow investigators to address the needs of scientists possessing the smallest computing system which can meaningfully participate in a terabit communication.
>
> **Discussion:** End system hosts will continue to grow in terms of CPU capabilities, giving scientists very powerful single-node systems by 2015, e.g., X86 processors with accelerators, GPUs, MIC/Intel, or FPGA-based systems. However the input/output (I/O) capabilities on these hosts are not expected to increase as fast, and hence it is likely that a fixed number of PCIe Gen 3 lanes will be used to connect to I/O, e.g., a 16x PCIe gen 3 slot is capable of 128 Gbps. With multiple slots on a motherboard, this will allow for a couple of 100 Gigabit Ethernet cards and access to some type of storage. However for a system to source/sink data, a fast interconnect will be required to connect to a large amount of storage, along with multiple 100 Gigabit Ethernet links to the remote site. However, the PCI-SIG (special interest group) committee has not yet started to work on next generation PCIe bus standards, and this needs to be taken into account.
>
> Furthermore by about 2018 host end systems will likely be comprised of extremely powerful computational SOCs which directly host networking and interconnection capabilities and interfaces. These chips will likely also be integrated with memory and CPU devices, and will be placed on very high-speed buses to enable terabit-level data movement. In particular, these systems will likely have access to highly-evolved technologies such as non-volatile memories (i.e., flash, phase-change, memristor, solid state drives, etc). Considering the fact that future

exaflop systems may host many tens of thousands or even millions of such SOCs, external networks must be capable of effectively moving terabytes-petabytes worth of data. Therefore it is plausible that new data transfer protocols will have to be developed, or at a minimum, the workhorse TCP protocol parallelized. In addition improvements to the Linux OS kernel will also be needed in order to take full advantage of the new interfaces and multitude of cores/accelerators on the SOC.

Finding 3.7: *Commonly-available edge architectures are generally not designed to facilitate the interconnection of terabit networks to LANs, direct end host-systems, or large computing or storage systems. In addition, effective diagnostics tools for end-to-end monitoring and debugging are not readily available today.*

Recommendation: There needs to be a concerted effort to look at interconnection issues between terabit core networks and edge LAN and host end systems. In particular, end systems have to be designed for local and remote access and end-to-end ―impedance matching” capabilities have to be developed to maximize throughput performance. Sustainable solutions for end-to-end diagnostics and performance monitoring must also be developed as they are crucial for improving network usability.

Discussion: Most high-speed networking systems are designed to meet the requirements for long-distance data transfers, and these can be quite different from those developed for the LAN environment. For remote sites to achieve full exascale performance, WAN circuits need to be landed in local networks (or end systems) that can effectively handle the transfer of terabyte-level datasets. In addition, new LAN switches also have to host WAN-compatible interfaces and support the requisite interface port counts to connect to internal host and storage systems that source/sink large data transfers.

File Systems and Metadata

Finding 3.8: *While current-generation parallel file systems are able to fully utilize the available single-node bandwidth on HPC systems, achieving similar levels of performance across the WAN has proved challenging. With the advent of terabit WAN, the gap between single-node performance across the WAN and within the HPC environment is likely to widen dramatically, with single-node WAN performance achieving only a fraction of the available bandwidth.*

Recommendation: Additional R&D efforts are needed to develop middleware technologies and system software packages that optimize parallel file system performance in an end-to-end manner. These efforts should focus on optimization of single-node and/or coordinated multi-node parallel file system performance across the WAN for bulk transfers and remote file system access. In addition, the developed technologies should be portable across a variety of parallel file system environments.

Discussion: HPC facilities such as the ALCF, NERSC, and the OLCF currently utilize dedicated DTNs for bulk data movement over the ESNet backbone. While the use of these setups has substantially improved WAN data-transfer rates between sites, a substantial gap still remains between maximum potential and realized performance. In particular, a number of bottlenecks currently exist, most notably with regards to single-node parallel file system performance. While some work is required within the parallel file system and is likely vendor specific, other efforts to overcome these bottlenecks should be considered within the realm of middleware and portable system software. This will also help improve portability across a wide variety of environments.

Finding 3.9: *While the aggregate amount of data stored across many DOE facilities continues to grow—some at exponential rates—the average file sizes remain small, i.e., particularly files stored on parallel file systems. This trend is likely to persist as applications continue to utilize file-per-process techniques to avoid the overheads associated with single-shared files. Achieving terabit/sec WAN data access and bulk data movement when accessing large-numbers of small files will result in metadata intensive workloads, often from a very small number of nodes. Metadata overheads will quickly dominate these WAN oriented workloads, resulting in substantially degraded bandwidth over WANs.*

> **Recommendation:** Research initiatives are needed to develop middleware technologies and system software packages to alleviate these metadata overheads in WAN-oriented workloads. These efforts should address a variety of use cases that go beyond simple bulk data transfers, e.g., such as remote analysis and visualization in across terabit-speed WAN settings. The developed technologies should also be portable across a variety of parallel file system environments.
>
> **Discussion:** While file sizes on archival storage systems are generally large, parallel file systems tend to have a much larger number of relatively small files. Namely, even though application data sets continue to increase, these may include tens, or even hundreds of thousands of individual files, that in aggregate make up the entire dataset for a single application invocation. For example, the average file size on the OLCF parallel file systems is only 14.8 megabytes, indicating that the median file size is much smaller. As a result current bulk data movement tools are unable to saturate high-performance WAN links with such small file sizes. Given that terabit networking enables the widespread adoption of cutting edge use cases, e.g., such as remote analysis and visualization, optimized access to smaller files will be critical. Therefore single- and multi-node metadata performance will need to be substantially improved in order to support these workloads over terabit WAN infrastructures.

Finding 3.10: *Scientific data models are almost universally used with leadership computing applications. The developers of these applications think in terms of saving rectilinear, curvilinear, or unstructured 2-D and 2-D grids, amongst other advanced data-structures, rather than saving a stream of bytes to a file in a POSIX file system. Meanwhile, the users of these applications then run a variety of tools to manipulate (i.e., summarize, subset, aggregate, etc) these data structures once they have been saved to the file system. As aggregate dataset sizes continue to increase, directly supporting (and manipulating) these scientific data models over the WAN will become increasingly important. Not only will this allow scientists to think about their data in a more natural manner, increasing their productivity and reducing time to discovery, it can also hide the complexity of manipulating the many small files mentioned above, or ideally, optimally balance the tradeoffs between I/O bandwidth, metadata performance, and WAN data transfer speeds and file sizes.*

> **Recommendation:** Research efforts are needed to develop middleware technologies and system software packages to support the access and manipulation of scientific data models over the terabit WAN. In particular these solutions must provide efficient end-to-end support all the way from middleware to backing parallel file system environments. In addition these research efforts should also investigate the development of new scientific data models that more efficiently support remote manipulation and access use cases.
>
> **Discussion:** Evolving exascale leadership computing platforms will likely provide more direct support for scientific data models within the storage system. Rather than relying on POSIX-based file system semantics for persistent storage, these systems may provide alternative semantics that map more directly to these scientific data models. In order to support this

transition, a variety of WAN oriented use cases, most notably remote manipulation of these datasets, will need to be supported. While this work is forward looking towards the exascale I/O environment, it also has nearer term implications since scientific data models have been adopted across most related disciplines. For example, advanced use cases running distributed analysis tasks across major DOE HPC facilities(i.e., where computational resources are co-resident with the data of interest) can benefit from such capabilities.

Finding 3.11: *It is likely that HPC sites will not converge to a single file system on the same timescales as the emergence of exascale applications.. In other words, no one existing file system will be able to meet the needs of all computational sites for the foreseeable future.*

Recommendation: Significant research efforts should be directed towards improving/expanding data access and management paradigms that allow scientists to interact with their data using the data models natural to the problem at hand. These solutions should be able to seamlessly translate scientists' interactions with their data models regardless of the underlying file/storage system being used.

Discussion: In general a wide range of file systems are in use today, and the average time needed to develop a usable HPC file system is on the order of a 3-5 years. There is a consensus that current data management stacks are somehow more unwieldy then they ought to be, and that some kind of systematic simplification is possible if the whole problem is addressed in a different way. For example, using POSIX-based I/O and dealing with files is not necessary when dealing with data. Alternatively many scientific domains already have their own de-facto standardized I/O libraries, and associated APIs capture most of the semantics of how scientists interact with their data, e.g., NetCDF, ROOT, etc. Studying these higher-level interfaces can yield strong insights into scientific use cases and distributed data management. These findings can then be leveraged to build more useable data systems that are better suited for terabit networking support. Additionally, while perhaps not directly applicable to DOE environments, relational databases are also examples of systems where file management concerns are factored away from the end-users interacting with data.

It is also noted that most file sizes on DOE networks are still small with respect to (increasing) bandwidth-delay products, and this is likely to remain true in the coming years. As such these smaller file sizes pose further issues in terms of data movement and management (essential functions for distributed storage systems running at terabit speeds). Future research efforts need to look at effective storage resource management and efficient scheduling for data movement/aggregation.

Finding 3.12: *It is likely that new non-POSIX semantics for data access will yield key breakthroughs in terms of application performance, application simplicity, and data management.*

Recommendation: The POSIX file system standard should be evolved or displaced to make it better suited for scientific data management tasks. For instance some key features include metadata-only space allocation, exposure of data locality to optimize analysis (like Hadoop), and standardized mechanisms to allow other processes to interrogate the state of the file system.

Discussion: Overall many HPC file system designs are quite old and make use of POSIX-based semantics which are rather difficult to work with and are generally unwieldy for scientific needs. Alternatively, there are many clear cases of non-POSIX file system innovations being used in industry, e.g., such as Google Hadoop etc. Nevertheless, many of these commercial file systems do not provide a perfect fit for all scientific computing needs either, and may prove difficult to interoperate between DOE facilities due to the lack of standards. Note that other practical difficulties also remain in exposing file system properties to the WAN and also data management systems, e.g., such as storage system capacity.

<u>Data Centers, Data Grid, Data Cluster, Host Provisioning, Performance Monitoring for High-Speed Data Transfer</u>

<u>Finding 3.13:</u> *As the size and amount of scientific data increases exponentially over time, terabit networks will serve a critical need by moving data between systems within each DOE facility and also between DOE facilities. However higher capacity and more capable terabit networks are likely to exacerbate the difficulty in seeking optimal performance for long duration bulk data transfers.*

<u>Recommendation:</u> Research into end-to-end performance monitoring for inter- or intra site data transfers is critical for understanding component failures and conducting root cause analysis of low performance and/or failed data transfers. In addition, performance monitoring capabilities that provide a standard set of statistics that can be used to determine the health and real-time capability of a number of different components (e.g., storage and network devices, servers, WAN/LAN devices) would be ideal for terabit network settings. These solutions should also interwork across a variety of vendors.

<u>Discussion:</u> In general it is quite challenging to determine component failure in large, multi-provider networked environments, as often is the case with WAN transfers. Moreover the bandwidth and performance of WAN networks is also more robust than that in —last mile" networks connecting to the actual end system. In particular, LAN settings are often the most complicated and difficult environments to understand and maintain for high performance transfers. Therefore experience has shown that positioning data transfer systems as close as possible to the site —demilitarized zone" (DMZ) border often gives the greatest success in high performance data transfers. Furthermore users in current network environments are commonly the ones who first notify network operators about poor and/or failed data transfers. Subsequently it normally takes manual intervention by a storage and network administrator to reproduce and understand the exact problem. Therefore providing *per-transfer* monitoring capabilities between key data transfer systems or components (as opposed to everywhere), can also provide invaluable information and help to drastically lower the lead times in diagnosing/fixing problems.

<u>Finding 3.14:</u> *Data transfers are typically per-file interactive processes at DOE facilities. Given the increased size and amount of data expected in terabit environments, the average amount of data moved over the network will make per-file and interactive data transfers unproductive for both scientific users and network operators.*

<u>Recommendation:</u> Research on developing a network/browser-based automated data transfer service with support for non-interactive third party transfers and retry for temporary data errors would significantly improve productivity for scientists using terabit networks. The service should include a wide variety of features or options that are also useful for bulk data transfers (i.e., reservations or guarantee of network transfer bandwidth between sites, authentication assistance for multiple sites, estimates for completion of time to last byte, etc).

<u>Discussion:</u> The need for a robust browser-based file transfer capability has evolved from lessons learned from the cloud computing and as well as commercial (i.e., data center) communities. In addition the DOE facilities Data Transfer Working Group has also expressed interest in such a solution, i.e., as it has been specializing in optimizing and simplifying WAN data transfers between sites such as the ALCF, OLCF, and NERSC since 2008.

Finding 3.15: *Supporting both IP routing and circuit-switched capabilities in terabit networking environments is still advantageous for helping separate traditional "fair-share" type data applications from competing high-bandwidth streaming/bulk transfers. However effective solutions for delivering circuit-based network capabilities closer to terabit-capable end systems or storage clusters inside a LAN are still lacking. This is deemed as a key enabling technology.*

Recommendation: It is crucial to develop circuit-based network protocols or devices that can be deployed inside a LAN, i.e., as close as possible to terabit networking capable end systems. Such solutions will play a vital role in providing guaranteed bandwidth for end-user transfers.

Discussion: The successful demonstration of high-performance data transfers typically requires networks with circuit-based reservation capabilities. In addition, specially-tuned end systems and software are also necessary here. There is a general consensus for supporting a few systems with high-speed network connections (and tuned software) close to the border site of several DOE facilities, i.e., to help support consistent high-performance transfers. It is also vital to understand how introducing circuits into the LAN environment can help improve data transfer capabilities, i.e., to interconnect with a limited number of site systems that are capable of provisioning circuits across the WAN.

TCP/UDP Transport Protocols and Viable Protocols for Massive Data Transfers

Finding 3.16: *The TCP transport protocol is a poor choice for high-bandwidth streaming/bulk data transfers over dedicated circuits*

Recommendation: When considering exascale transfers, existing assumptions need to be set aside and each feature of the TCP protocol re-addressed to determine which will be most effective for end-to-end transfers. In particular, targeted research efforts should address issues such as out-of-order packets, loss recovery, and QoS support. Furthermore, the likelihood and impact of having many tens, hundreds, or even thousands of on-chip network interfaces should also be factored into these investigations.

Discussion: When using dedicated high-bandwidth circuits for bulk data transfers, dropped packets most likely will not signal congestion events. By virtue of the underlying circuit's service level agreement (SLA), the transport layer protocol should not have to reduce its transmissions as it is not competing with other fair share traffic. Tighter coupling will be required to limit unnecessary throttling behaviors, and this work can make use of underlying circuit SLA/QoS information.

Community-Host Systems Provisioning, Virtualization, and Performance Monitoring

Finding 3.17: *End-to-end provisioning and monitoring will be required in order for scientists to take full advantage of terabit networking capabilities.*

Recommendation: New R&D efforts are needed to develop architectures, systems, services, interfaces, and protocols that will enable trivial provisioning, reservation, and monitoring of every component in the end-to-end transfer, i.e., including WAN, LAN, storage, processors, memory, even internal buses.

Discussion: Resource contention issues in WAN backbones are well-known and have in turn driven the adoption of dynamic circuit switching. However contention still remains a key concern in many of the components involved in data transfer, and this has forced scientists to resort to ―hero” efforts to achieve high performances for very specialized case scenarios. Therefore for a terabit-speed transfer to be truly *end-to-end*, the receiving hosts must have sufficient resources, i.e., processing power to execute the transfer, random access memory (RAM) for buffers, disk space, and perhaps the biggest bottleneck of all, storage bandwidth. This mandates the ability to co-schedule every component along the way, thereby ensuring guaranteed QoS even for advanced reservations, i.e., to close the ―wizard gap”. However this can become a two way street, i.e., as resource providers must insure that they provide the guaranteed level of service, but consumers must also guarantee that they do not consume more than their allocated resources. These guarantees must happen *automatically* and cannot assume a single security domain or trust between the domains. Similar requirements also arise for performance monitoring and will be vital for troubleshooting problems and failures.

Other Concerns and Issues

Finding 3.18: *A systems-level view of data management practices seems likely to produce breakthroughs in data management, thereby enabling better exploitation of WAN infrastructures.*

Recommendation: Research efforts should be launched to assess data management from an overall systems point of view.

Discussion: In light of the way scientists use files today, the number of elements present in current science facilities is indeed quite large. For example, many parallel file systems host a large number of relatively small files. Meanwhile a significant proportion of storage capacity is also taken up by larger data sets comprising of proportionally larger files. For example, the average file size on the OLCF's parallel file systems is only 14.8 Megabytes, indicating that median file size is much smaller. Broader data from the outside community also shows no effective difference in the distribution of file sizes on most Unix systems over a period of about 20 years[1].

Today's standard bulk data movement tools are essentially oriented towards file transfer, and underlying protocol message exchange patterns have per-file latencies. However, given that there has been no exponential growth in file size, it will become very difficult to saturate a modern ultra-high speed WAN without tedious extra work: Namely, practicing scientists must explicitly bundle their data to ship it over the WAN, perhaps bundling using several different ways for different steps in a scientific workflow. Nevertheless, as noted earlier, multi-site workflows are becoming an increasingly important use case, and will thus face increasingly high barriers to use.

Further consider the computing system architectures needed to provide high-end data transfers. Here, HPC facilities (such as the ALCF, OLCF, and NERSC) currently utilize dedicated DTNs for bulk data movement over the ESnet backbone. While the use of DTNs has substantially improved WAN data transfer rates between sites, there exists a substantial gap between

[1] Tannebaum, *et al*, *File Size Distribution on UNIX Systems—Then and Now*, ACM SIGOPS Operating Systems Review, January 2006

maximum and realized performance. Specifically, a number of bottlenecks currently exist, most notably, single-node parallel file system performance. While some work is required within the parallel file system area and is likely vendor specific, these bottlenecks can also be addressed within middleware and portable system software designs. These latter approaches would also improve portability across a wide variety of environments.

Some have stated that a fresh new systems-oriented approach for the whole data management problem is warranted. For instance data-intensive science seems to have file sizes that are effectively falling exponentially when compared to storage and network capacity. Hence selecting, serializing, and de-serializing these files is tedious ―pre-application” work that distracts from overall scientific productivity. More importantly, it also presents a drain on science budgets and is often done poorly. For example, the lead-time to construct the data management system for the LHC is comparable to the planned time between LHC accelerator upgrades. However the discussion in did not have sufficient time to consider altogether new systems-based strategies. Instead it is noted that there may be alternate system models that have good potential. For example, relational databases represent a more comprehensive solution for data management than the current file systems (and file manipulation tools) used by scientists. Namely, these systems naturally separate the roles of data producer, data manager, and data consumer. Data serialization/de-serialization is also supported here, along with many other desirable attributes.

In all, it data management is a complex concern and there remains much research work to be done. Moreover most scientific applications do not directly interface to the POSIX API. Instead, scientists have converged to a small number of APIs to write to standard format files, and many of these APIs support alternate I/O interfaces, e.g., such as the Tfile class of the ROOT package, C subroutine I/O libraries for the flexible image transport system (CFITSIO), etc. As a result it may be worthwhile to develop solutions to incorporate these interfaces into a comprehensive system that can be readily accepted by scientific users, helping mitigate complexity barriers for effective distributed system operation.

.

Group 4: Cybersecurity in the Realm of HPC and Data-Intensive Science

This cybersecurity session was not conducted at the workshop itself, but rather done as a virtual on-line discussion session after the main workshop was complete. The discussion here is intended to be a prelude to a more extended examination of the issues at a future time.

The volume of data involved in many modern science regimes is causing many disciplines to re-think their process-of-science strategies, their collaboration structures, how resources are shared, etc. This is because unique scientific instruments and facilities are now routinely accessed and used remotely by researchers from many institutions. Some of the new paradigms here include remote instrument control, real-time analysis of remote sensors for experiment health-check, remote data archiving and analysis, remote collaboration, etc.

Large-scale collaboration is common in today's science environment. This is driven by the fact that leading-edge science (observational, experimental, and computational) has fewer ―instruments" available due to the cost of the ever larger and more complex devices needed to explore new and ever more remote realms (in the sense of far from what can be done today). This results in more scientists needing to access any given instrument. Such collaborations imply increased need for collaboration tools, data sharing, supercomputer intercommunication, etc, over intercontinental distances. Many instruments create massive datasets that have to be archived, catalogued, and analyzed by the distributed collaborations. The LHC, for example, routinely distributes, archives, and analyzes petabytes of data per year per experiment. For example, between June 2009 and June 2010 the US-based ATLAS data center (at Brookhaven National Laboratory) transferred some 2.5 petabytes of data to 8 university sites for analysis. The analysis of such datasets is accomplished by using the approach of grid-managed resources that are world-wide in scope. As such, networks are the fabric that underpin every aspect of this type of scientific collaboration.

In all the above requirements have resulted in the baseline performance requirements for WAN infrastructures to increase dramatically, i.e., indeed scientific productivity in many disciplines heavily depends upon this (see networking requirement reports in Appendix for several key science fields as well as www.es.net/requirements.html). However if one looks at many of the DOE Office of Science disciplines other than HEP, then it is noted that today's networks, even though high bandwidth in nature, serve many scientists poorly or not at all due to:

1) Poorly-implemented local foundations
    - Many LAN networks still exhibit notable packet loss (due to incorrect configurations, cheap hardware, poor design, etc)
    - Most system defaults are incorrectly set for the systems that must communicate over the WAN at high-speed, i.e., hosts still need to be tuned for high-speed WAN data transfers

2) Wrong tools for the job
    - Use of SSH-based tools is common, although SSH has built-in performance limitations

3) Security issues tend to block scientists at every turn

All of these points have implications, directly or indirectly, for security in terabit networks supporting exascale science. The key findings from this discussion are now presented.

Finding 4.1 Security at terabits per second: *The impact of terabits/sec networks on security is not at all obvious and probably cannot be extrapolated from today's practices for 10 Gbps networks.*

Research: As 100 Gbps networks are deployed, it is already clear that the architectures of these networks is changing considerably, and therefore the security approaches will also have to change. Will another increase of ―10x” to terabits bring further changes? If so, what kinds? Are distributed denial of service (DDOS) threats at 1 Tbps different than those at 10 Gbps? Will the nature of intrusion detection have to change? How will the science community cope with the lack of commercially-available security devices (e.g., firewalls, intrusion detection systems) that are capable of supporting terabit-scale networks?

Discussion: These and other issues must be examined individually and in interaction.

Finding 4.2 LAN network architecture specialized for science: *LAN environments are rarely configured to support high-speed WAN data transfers. Namely, some important tools for high-speed data movement are blocked at the network layer or disallowed by policy. Firewalls also cause poor performance and high-performance tools are often incompatible with system access technologies, e.g., secure shell (SSH), etc. Furthermore 100 Gbps stateful firewalls are unlikely to be mature by the time terabit-scale networking must be deployed to support science (firewalls are currently a factor of 10 behind the curve for 100 Gbps networks, and this trend is not likely to change with emerging terabit/sec networks). Other security enforcement mechanisms must be examined. While this is primarily an issue in the LAN environment, LAN security devices such as firewalls can impact the scientific utility of the entire network infrastructure, both LAN and WAN.*

Recommendation: Decouple science and business environments in the site LAN so that security policy and control points may be defined specifically for the different environments. One site architecture that facilitates such decoupling is a ―Science DMZ” (demilitarized zone) that both physically separates science and enterprise traffic and builds a science environment that can be optimized for WAN data transfers (with security controls appropriate to the data and traffic of that environment). Move large-scale science assets toward a security model that does not rely upon stateful firewalls or push stateful firewalls to have some protection at the required line rates. Routers can typically filter packets at line rate, and are therefore able to provide significant protection without causing the performance problems typical of firewalls.

Research: The various LAN approaches that will support high-speed, long distance WAN data transfers must be examined. At the same time the nature of security controls probably have to be simplified so that they can be implemented with hardware systems that operate at terabit/sec speeds. In particular, adequate security can be implemented in router filters/access control lists (ACL) which can operate at 100 Gbps interface speeds today and will likely be available for 1 Tbps interfaces in the next 5 years as well.

Most large-data science traffic is moving to circuit-like services such as the virtual circuits and pseudowires provided by ESnet OSCARS, Internet2 Interoperable On-Demand Network (ION), GÉANT's AutoBAHN, etc. The integrity of these (mostly Layer 2 Ethernet VLAN) circuits is provided in ways that can preclude ―inspection” of the circuit contents (even for circuit health monitoring, much less deep packet inspection for intrusion detection). New security models, or at least new approaches, are also needed in end-to-end virtual circuit environments.

Discussion: A network DMZ is a network segment near the site perimeter with different security policy than the rest of the site. It is a commonly-used architectural element for deploying site WAN-facing services, simple mail transfer protocol (SMTP), domain name service, web, etc. A ―Science DMZ" is a term that can be used to describe a LAN network segment that is well-configured for high-performance WAN-facing science services. For example, this is where systems dedicated and tailored for WAN transfers, e.g., high-performance data movers running GridFTP, would be located. This network segment would be provisioned with highly capable network devices—wire-speed capable interfaces and internal data paths, and deep queues for managing short congestion periods without packet loss. In fact the whole focus of the architecture is to eliminate all places where packet loss might occur. The importance of this is that only with end-to-end error free networks can high throughput be achieved on high-bandwidth paths that are thousands of kilometers long, e.g., San Francisco Bay Area to CERN (in Geneva, Switzerland) is about 10,000 km. There are multiple reasons for this. First, the high latency of long-distance network paths dramatically increases the time to recover after loss events for TCP connections (TCP is used for the vast majority of data transfer applications). Second, the nature of TCP's sliding window method of ensuring reliable data transfer and effective utilization of the available bandwidth results in very large packet bursts on long-distance paths. This behavior makes TCP vulnerable to loss on long-distance paths where the same equipment and configuration would result in loss-free operation on shorter paths.

A Science DMZ environment must support a dynamic/virtual circuit infrastructure such as ESnet's OSCARS, Internet 2's ION, GÉANT's AutoBAHN, etc. This network segment would be fully-monitored and equipped with diagnostic tools such as the PerfSonar performance monitoring suite. The concept and elements are represented in the figure below.

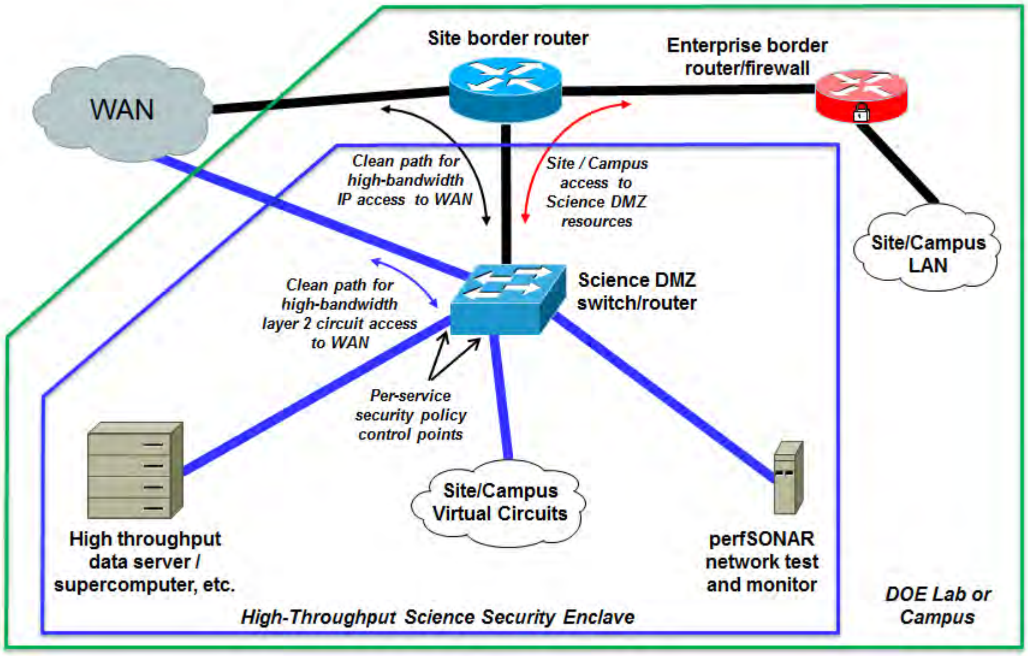
Site border router
Enterprise border router/firewall
WAN
Clean path for high-bandwidth IP access to WAN
Site / Campus access to Science DMZ resources
Site/Campus LAN
Clean path for high-bandwidth layer 2 circuit access to WAN
Science DMZ switch/router
Per-service security policy control points
Site/Campus Virtual Circuits
High throughput data server / supercomputer, etc.
perfSONAR network test and monitor
High-Throughput Science Security Enclave
DOE Lab or Campus

Figure: Prototype site network separation of high-throughput science and other lab network traffic

A Science DMZ would facilitate defining a science-specific ―cybersecurity enclave" that contains well-defined systems, specialized network architecture, and associated security policy. Separating the high-performance science systems from the rest of the network has two benefits. First, the stateful firewall that protects the business systems (e.g., desktops, servers, printers, etc) need not support the traffic volumes required on the Science DMZ. Second, the Science DMZ systems are not encumbered by security policy control mechanisms that exist to protect the business desktops and other lower-speed systems.

By virtue of the specialized nature of the Science DMZ as a high throughput environment (i.e., the systems that live in this environment are specific to high data throughput and are not general user workstations, PC's, printers, etc), security policy here can typically be implemented with router filters/ACLs. Modern routers can typically filter packets at wire speed—currently 10 Gbps and 100 Gbps. However routers do not rewrite packet headers and track state like firewalls do, which eliminates entire classes of bugs and network performance problems (firewalls have a long history of causing significant performance problems in high-performance networking environments). Moreover security policy for the Science DMZ does not affect business systems because enterprise systems (desktops, printers, corporate database servers, etc) are not on the Science DMZ. Therefore the Science DMZ systems are not subject to business security policies or policy enforcement devices, and the enterprise systems are not subject to Science DMZ policies (e.g., inbound open ports for support of parallel data transfers). This is a sane approach as appropriate policies and controls are applied in both cases.

Finding 4.3 Security model: *In order to realize optimal environments for high-speed WAN data transfer, there is a need for a best-practice security model for data-intensive science.*

Research: Campus network architectures to support data-intensive science appear to be converging on a Science DMZ-like approach. This is true not only in DOE environments, but also in the campus environment as well with projects like Dynamic Network Systems (DYNES), see http://www.internet2.edu/ion/dynes.html. Developing a security model as mentioned above is more than just a technical issue, i.e., since the model must also be consistent with institutional and (in the case of the national laboratories) DOE policies. This represents a set of constraints that must guide the development of a security model. The environment, as mentioned, is already quite different from the current one and adding a set of additional constraints makes the issue even more complex. Hence the key research problem here is to develop a security model that is both compliant with a DOE/National Institute of Standards (NIST) policy framework and that can also be implemented in a terabit/sec Science DMZ environment. A similar problem exists for the tools that scientists use to accomplish data movement.

Discussion: A data-intensive science environment security model is currently lacking. The entire science environment (i.e., sites, networks, and scientists) stands to benefit significantly if this can be fixed, as it will yield increased productivity, new science paradigms, etc. Another important element of this is the notion of an approved toolset for data-intensive science tasks, e.g., such as bulk data movement. An approved toolset with best practices for deployment will go a long way toward giving scientists bargaining power with the security compliance personnel at their sites.

Note that a security model that protects the assets that generate the high-bandwidth flows without burdening other security infrastructure (i.e., that protects financial desktops, databases, human resources, and other sources of PII) is highly desirable. A model that allows those responsible for site security to separate the two classes of systems allows them to do their jobs better and with fewer engineering compromises. Note also that Science DMZ-like architectures significantly narrow the scope of the special permissions needed to support high-bandwidth science flows. There is no danger of compromising multi-purpose and/or user-managed desktops through the ports opened for GridFTP as the two operate in separate infrastructures.

<u>Finding 4.4 Intrusion detection:</u> *Intrusion detection at 100 Gbps and terabit speeds will be difficult or impossible without further development and hardware support.*

<u>Research:</u> Further R&D efforts are needed for hardware assists at 100 Gbps and 1 Tbps. Protocol analysis systems typically only parallelize at the flow level, which works today because individual flows are of manageable bandwidth due to the use of parallel data servers like GridFTP. Then each flow is passed to a separate computational node in a cluster that does the protocol consistency checking, packet inspection, etc. As individual flow bandwidths grow substantially, as they are likely to when supercomputer intercommunication is involved, then the analysis algorithms themselves need to be parallelized, which is hard. The same is also true for stateful firewalls, i.e., most so-called 10 Gbps firewalls can only deal with 1 Gbps flows. Further R&D efforts are needed in this area.

<u>Discussion:</u> Although some technologies are available now, they involve tradeoffs (e.g., filter-based port mirroring on some routers). Line rate intrusion detection is also desirable, if not essential, for 100 Gbps and terabit networks, and approaches for this exist. One specific example (because it is an open, research-orientated platform that has had success in addressing the issues of a protocol analyzer approach to IDS) is the Bro system, see reference in Appendix. This solution has been developed at UC Berkeley (Prof. Vern Paxson's group) and also implemented in a production-prototype form at LBNL. Overall Bro embodies approaches that point the way to 100 Gbps and terabit intrusion detection systems (IDS). In particular, Bro is currently integrated with a hardware, e.g., field-programmable gate array (FPGA) based, flow tagger that examines 10 Gbps data streams and assigns unique tags to each flow. These tags are then used by an Ethernet switch to direct each flow to a different Bro instance in a cluster of Bro systems, which also manages global state for alert messaging, etc.

Overall the Bro approach has worked well, and is very likely scalable to 100 Gbps with suitable flow tagging frontend hardware. Namely, in high-speed data movers in current environments are mostly using a GridFTP-like program that does massive transfers by breaking the flow up in to a relatively large number (10's to 100's) of moderate bandwidth flows. Each of these flows can be IDS-analyzed by a single Bro instance in a cluster.

However, in the future as more contiguous high-speed flows show up (e.g., from instruments to supercomputers) then the current Bro cluster approach breaks down because the entire high-speed flow is directed to one Bro instance. There has been some preliminary work by Prof. Paxson and his team on parallelizing the Bro analysis algorithms, but much more work remains to be done here before a practical implementation is possible. Note that there are, of course, other examples of IDS systems, but the vast majority of these are either proprietary or classified, so not much can be said about their approaches.

Overall, high-speed IDS is a topic that deserves sustained research and development because it is very likely that IDS will be at the heart of the security approach in LAN environments that run at terabit speeds and are tailored for WAN data transfers. Moreover the capabilities DOE needs for high-speed IDS solutions are likely to be greater than the capabilities of commercially-available IDS solutions for the foreseeable future.

IV. APPENDIX

WORKSHOP AGENDA

Terabits Networks for Extreme Scale Science
Hilton Rockville Hotel & Executive Meeting Center
Wednesday, February 16, 2011

8:00 - 10:00 a.m.	**Opening Welcome**	
8:00 - 8:20 a.m.	Advanced Science Computing Research Office	
8:20 - 8:35a.m.	Workshop Goals & Objectives	Bill Johnston
8:35 - 9:10a.m.	Science Drivers for DOE Office of Science Networking	Walt Polansky
9:10 - 9:40 a.m.	Energy Sciences Network: An Update	Steve Cotter
9:40 - 10:00 a.m.	Energy Science Network: Usage Patterns & Lessons Learned	Eli Dart
10:00 - 10:25 a.m.	**Break**	
10:25 - 10:50 a.m.	DOE/ASCR Network Research Program Overview	T. Ndousse-Fetter
10:50 - 11:20 a.m.	Extreme Scale Computing and Networking Environment	John Shalf
11:20 - 11:45 a.m.	State of High Capacity Optical Network Technologies	Kim Roberts
11:45 - 12:05 p.m.	Optical Packet Internetworking: An Assessment	David Ward
12:05 - 1:20 p.m.	**Lunch**	
1:20 – 2:50 p.m.	**Parallel Breakout Sessions**	
	Advanced User Level Network-Aware Services	Breakout 1
	Terabits Backbone, MAN, and Campus Networking	Breakout 2
	Terabits End Systems (LANs, Storage, File, Host Systems)	Breakout 3
2:50 - 3:05 p.m.	**Break**	
3:05 – 4:35 p.m.	**Parallel Breakout Sessions**	
	Advanced User Level Network-Aware Services	Breakout 1
	Terabits Backbone, MAN, and Campus Networking	Breakout 2
	Terabits End Systems (LANs, Storage, File, Host Systems)	Breakout 3
4:35 - 4:50 p.m.	**Break**	
4:50 - 6:20 p.m.	**Readout** (All Sessions – General Discussion)	

Terabits Networks for Extreme Scale Science

Hilton Rockville Hotel & Executive Meeting Center
Thursday, February 17, 2011

8:30 - 9:00 a.m.	**Plenary Session**	All
9:00 - 10:30 a.m.	**Parallel Breakout Sessions**	
	Advanced User Level Network-Aware Services	Breakout 1
	Terabits Backbone, MAN, and Campus Networking	Breakout 2
	Terabits End Systems (LAN, Storage, File, Host Systems)	Breakout 3
10:30 - 10:45 a.m.	**Break**	
10:45 - 12:15 p.m.	**Parallel Breakout Sessions**	
	Advanced User Level Network-Aware Services	Breakout 1
	Terabits Backbone, MAN, and Campus Networking	Breakout 2
	Terabits End Systems (LAN, Storage, File, Host Systems)	Breakout 3
12:15 - 1:00 p.m.	**Lunch**	
1:00 - 2:30 p.m.	**Readout** (All Sessions – General Discussion)	
2:30 - 5:00 p.m.	**Adjourn**	
2:30 - 5:00 p.m.	**Workshop Report Preparation**	Committee

LIST OF WORKSHOP ATTENDEES

Name	**Affiliation**
William Allcock	Argonne National Laboratory
Guy Almes	Texas A&M University
Ilia Baldine	Renaissance Computing Institute
Artur Barczyk	California Institute of Technology
Eric Boyd	Internet2
Prasad Calyam	OSC/OARnet, Ohio State University
Rich Carlson	Office of Science, Department of Energy
Steve Cotter	Lawrence Berkeley National Laboratory, Energy Sciences Network
John D'Ambrosia	Force10 Networks
Eli Dart	Lawrence Berkeley National Laboratory, Energy Sciences Network
Philip Demar	Fermi National Laboratory
Constantine Dovrolis	Georgia Institute of Technology
Brent Draney	National Energy Research Scientific Computing Center
Aaron Falk	Raytheon BBN Technologies
Parks Fields	Los Alamos National Laboratory
Ian Foster	Argonne National Laboratory & University of Chicago
Michael Frey	Bucknell University
Pat Gary	Goddard Space Flight Center, NASA
Nasir Ghani	University of New Mexico
Chin Guok	Lawrence Berkeley National Laboratory, Energy Sciences Network
Jason Hick	National Energy Research Scientific Computing Center
Craig Hill	Cisco Systems Inc.
William Johnston	Lawrence Berkeley National Lab, Energy Sciences Network
Admela Jukan	Technical University Braunschweig
Dimitrios Katramatos	Computational Science Center, Brookhaven National Laboratory
Rajkumar Kettimuthu	Argonne National Laboratory
Tom Lehman	Universeity of Southern California, Information Sciences Institute
Yee-Ting Li	National Accelerator Laboratory (SLAC)
Mike Loomis	Alcatel-Lucent
Joe Mambretti	International Center for Advanced Internet Research, Northwestern University
Shawn McKee	University of Michigan
Linden Mercer	Naval Research Laboratory
Grant Miller	National Coordination Office/QinetiQ
Gary Minden	Kansas University
Inder Monga	Lawrence Berkeley National Laboratory, Energy Sciences Network
Thomas Ndousee	Department of Energy, Office of Science
Dhabaleswar Panda	Ohio State University
Drew Perkins	Infinera Corporation

Donald Petravick	National Center for Supercomputing Applications (NCSA)
Walt Polansky	Department of Energy, Office of Science
Nageswara Rao	Oak Ridge National Laboratory
Kristin Rauschenbach	Raytheon BBN Technologies
Kim Roberts	Ciena Corporation
John Shalf	National Energy Research Scientific Computing Center
Robert Sherwood	T-Labs, Deutsche Telekom Inc
Galen Shipman	Oak Ridge National Laboratory
Alex Sim	Lawrence Berkeley National Laboratory
Martin Swany	University of Delaware
Brian Tierney	Lawrence Berkeley National Laboratory, Energy Sciences Network
Chris Tracy	Lawrence Berkeley National Laboratory, Energy Sciences Network
Malathi Veeraraghavan	University of Virginia
Vinod Vokkrane	University of Massachusetts
David Ward	Juniper Networks
Glenn Wellrock	Verizon Inc
Jesse Wen	Acadia Optronics LLC
Dean Williams	Lawrence Livermore National Laboratory
Peter Winzer	Alcatel-Lucent
Dantong Yu	Computational Science Center, Brookhaven National Laboratory
Jason Zurawski	Internet2

References: List of Supporting Documents

Science Driven R&D Requirements for ESnet Workshop, April 23–24, 2007—Report (pdf)

Networking Requirements Workshop Report: Office of Biological and Environmental Research, July 26, 2007—Report (pdf)

Networking Requirements Workshop Report: Office of Basic Energy Sciences, June 2007 —Report (pdf)

Networking Requirements Workshop Report: Office of Fusion Energy—Report (pdf)

Science-Driven Network Requirements for ESnet, February 2006, Update of Office of Science Networking Requirements Workshop, 2002—Report (pdf)

DOE 20 Years Facilities Plan—Report (pdf)

Federal Plan for Advanced Networking Research and Development—Presentation (ppt)

NSF IIS-GENI Workshop Report: First Edition—Report

DOE ASCR 2008 budget—Report (pdf)

Presentations of ESnet Future Technology Assessment and Requirements

- Dynamic Bandwidth and Circuits Provisioning (pdf)
- Federation Network Monitoring (ppt)

High-Performance Networks for High-Impact Science, August 13-15, 2002—Report (pdf)

DOE Workshop on Ultra High-Speed Transport Protocols and Network Provisioning for Large-Scale Science Applications, Argonne National Laboratory, April 2003—Presentation (ppt)

DOE Science Networking: Roadmap to 2008 Workshop, June 3-5, 2003, Jefferson Laboratory—Report (pdf).

High-Performance Network Planning Workshop, August 13-15, 2002—Report (html)

Bro Intrusion Detections System-Report (html)

Glossary

AA	Authentication and authorization
ACL	Access control list
ALCF	Argonne Leadership Computing Facility
ANL	Argonne National Laboratory
ASCR	Advanced Scientific Computing Research
ASIC	Application-specific integrated circuit
CFITSIO	C subroutine I/O libraries for the flexible image transport system
CMOS	Complimentary metal-oxide semiconductor
CPU	Central processor unit
Data plane	Set of network elements used to receive, send, and switch network data
DICE	Dante Internet2 CANARIE and GEANT2
DDOS	Distributed denial of service
DNS	Domain name service
DTN	Data transfer node
DMZ	De-militarized zone
DOE	Department of Energy
DWDM	Dense wavelength division multiplexing
DYNES	Dynamic Network System
ESG	Earth System Grid
ESnet	Energy Sciences Network
Exascale	Computers executing over 10^{18} instructions/second, networks sending over 10^{18} bits/sec, and storage systems holding over 10^{18} bytes
Exchange point	Common meeting point for network peering, usually includes policy-free or policy-neutral use policy. Used as a flexible peering point for multiple services such as Layer 3 routing or Layer 2 services.

FITS	Flexible image transport system
GENI	Global Environment for Network Innovations
GLIF	Global Lambda Interchange Facility
GMPLS	Generalized multi-protocol label switching
GSI	Grid Security Infrastructure
Hadoop	File system for data-intensive distributed applications built by Google
HEP	High-energy physics
HPC	High performance computing
IDS	Intrusion detection system
IP	Internet Protocol
IT	Information technology
ION	Interoperable On-Demand Network (for Internet2)
I/O	Input/output
ITER	International Thermonuclear Experimental Reactor
KPI	Key performance indicators
LAN	Local area network
LCF	Leadership Computing Facilities
LHC	Large Hadron Collider
LSR	Label switch router
MPLS	Multi-protocol label switching
MTU	Maximum transmission unit
NDA	Non-disclosure agreement
NERSC	National Energy Research Scientific Computing Center
NetCDF	Network common data format
NIST	National Institute of Standards

NSF	National Science Foundation
NNI	Network-to-network interface
NSI	Network service interface
OAM	Operations and management
OLCF	ORNL Leadership Computing Facility
ORNL	Oak Ridge National Laboratory
OS	Operating system
OSCARS	On-Demand Secured Circuits and Advanced Reservation System (for ESnet)
OSG	Open Science Grid
OTN	Optical transport network
PC	Personal computer
PerfSonar	Performance Service Oriented Network monitoring Architecture
Petascale	Computers executing over 10^{15} instructions/second, networks sending over 10^{15} bits/sec, and storage systems holding over 10^{15} bytes
PGM	Pragmatic general multicast
PON	Passive optical network
POSIX	Portable Operating System Interface (for Unix)
QoS	Quality of service
RAM	Random access memory
ROOT	Rapid Object-Oriented Technology
R&E	Research and education (networks)
SCSI	Small Computer System Interface (iSCSI)
SLA	Service level agreement
SMTP	Simple mail transfer protocol
SNS	Spallation Neutron Source

SOC	Systems on a chip
SONET	Synchronous optical network
SSH	Secure shell
Tbps	Terabits/sec
TCP	Transport control protocol
UDP	User datagram protocol
UNI	User network interface
VLAN	Virtual LAN
VPN	Virtual private network
WAN	Wide area network