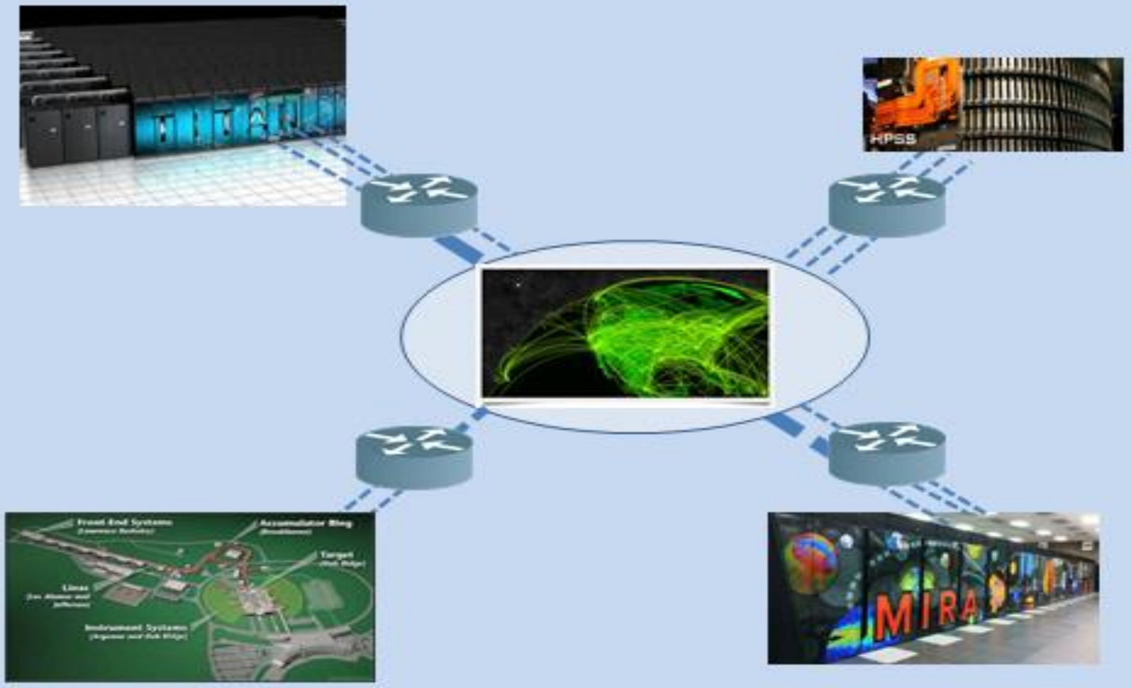
DOE Network 2025:
Network Research Problems and Challenges
for DOE Scientists

Workshop Report



DOE Network 2025: Network Research Problems and Challenges for DOE Scientists Workshop
February 1–2, 2016
Bethesda, Maryland

Organizing Committee

Lars Eggert, NetApp, Inc.
Christos Papadopoulos, Colorado State University
Nageswara Rao, Oak Ridge National Laboratory
Brian Tierney, Energy Sciences Network
Joe Touch, University of Southern California/Information Sciences Institute
Don Towsley, University of Massachusetts, Amherst
Lixia Zhang, University of California, Los Angeles

Workshop website: http://www.orau.gov/networkresearch2016

CONTENTS

ACRONYMS

API	application program interface
ASCR	Advanced Scientific Computing Research (program, DOE)
DMZ	demilitarized zone
DOE	US Department of Energy
DTN	data transfer node
DWDM	dense wavelength division multiplexing
ESnet	Energy Sciences Network (SC)
Gbps	gigabits per second
GPFS	General Parallel File System (IBM)
I/O	input or output
IB	InfiniBand
IP	Internet Protocol
IT	information technology
LAN	local area network
NFV	network functions virtualization
NIC	network interface card
OEO	optical-electrical-optical
PB	petabyte
Pbps	petabits per second
PCIe	Peripheral Component Interconnect Express or PCI Express
QoS	quality of service
REN	research and education network
RTT	round-trip time
SC	Office of Science (DOE)
SDN	software-defined network
TB	terabyte
Tbps	terabit per second
TCP	Transmission Control Protocol (usually with IP: TCP/IP)
UDP	User Datagram Protocol
UV	ultraviolet
VoIP	Voice over Internet Protocol
WAN	wide area network

EXECUTIVE SUMMARY

The growing investments in large science instruments and supercomputers by the US Department of Energy (DOE) hold enormous promise for accelerating the scientific discovery process. They facilitate unprecedented collaborations of geographically dispersed teams of scientists that use these resources. These collaborations critically depend on the production, sharing, moving, and management of, as well as interactive access to, large, complex data sets at sites dispersed across the country and around the globe. In particular, they call for significant enhancements in network capacities to sustain large data volumes and, equally important, the capabilities to collaboratively access the data across computing, storage, and instrument facilities by science users and automated scripts and systems.

Improvements in network backbone capacities of several orders of magnitude are essential to meet these challenges, in particular, to support exascale initiatives. Yet, raw network speed represents only a part of the solution. Indeed, the speed must be matched by network and transport layer protocols and higher layer tools that scale in ways that aggregate, compose, and integrate the disparate subsystems into a complete science ecosystem. Just as important, agile monitoring and management services need to be developed to operate the network at peak performance levels. Finally, these solutions must be made an integral part of the production facilities by using sound approaches to develop, deploy, diagnose, operate, and maintain them over the science infrastructure.

Guided by (1) a desire for an overall million-fold increase in bulk data transfer rates over the next decade and (2) a need to support a more complex and diverse mix of traffic flows over current and future networks, the DOE Network 2025 workshop, held in Bethesda, Maryland, February 1–2, 2016, articulated an integrated R&D portfolio to achieve these goals. The workshop identified short-, medium-, and long-term challenges to guide a robust network research program within the DOE Advanced Scientific Computing Research program. Integrated and agile protocols and related technologies must be developed and tailored to these science environments in the short- and medium-terms by leveraging current backbone capacities and their future enhancements to multiple 100 Gbps and Tbps rates. New capabilities and services need to be integrated into these networks to support complex and dynamic network traffic flows.

The long-term research challenges, with multiple Pbps data rate requirements, however, necessitate infrastructures with significantly larger and more complex capacities. They also require advanced integrated and agile protocols and higher level methods to match them. In all three time frames, research programs must be tailored to use and complement the science tools at instrument and computing sites. Furthermore, they must be integrated across the distributed science ecosystem to reduce the performance mismatches, align with security postures and measures, and support transparent use by science users.

A robust network research program will greatly enhance the productivity and usefulness of current and future science facilities managed by DOE.

1. INTRODUCTION

The US Department of Energy Office of Science (DOE SC) has a longstanding interest in advancing basic research in the physical sciences. To accomplish this goal DOE SC manages a large and diverse collection of unique scientific instruments and facilities. They include nanoscale research centers, atmospheric measurement facilities, genomic research facilities, fusion energy research centers, high energy and nuclear physics facilities, leadership class computing facilities, and international network research facilities. The scientists and engineers using these facilities come from academia, industry, and government, both nationally and internationally, and may work in small groups or large international collaborations.

Improvements in detector and sensor technologies have created exponential growth in both the experimental and observational data produced by some of these large science instruments. Concurrently, improvements in computing technologies have resulted in large high resolution simulations and complex analysis tools, which create their own tsunami of data.

Future science discoveries will increasingly rely on the production, storage, sharing, analysis, movement, and management of, as well as interactive access to, large, complex data sets distributed across sites around the country and the world. This drives the need for significant enhancements in the network capacities to carry such large data volumes. Equally important, it also calls for the integrated and agile transport capabilities to facilitate collective access to complex, multi-modal data across the computing, storage, and instrument facilities by scientists and automated scripts and systems.

The physical network infrastructure capacities have increased by several orders of magnitude over the past few decades, now achieving hundreds of gigabits per second. But the protocols and support tools to manage high performance, agile data flows across the science infrastructure have seen only incremental improvements over that time period. Furthermore, increases in the network backbone speeds, while essential, represent only a part of the solution for decreasing the time to science discovery. Indeed, they have to be matched with the transport methods and tools that scale, in volume and functionality, to aggregate, compose, and integrate several disparate systems, including instruments, supercomputers, and storage into a complete science ecosystem. Thus, the promise of future scientific discoveries, driven by remote monitoring and steering of simulations and experiments, can only be realized by networks that routinely support a rich set of dynamic and agile functions and capabilities.

To address these issues the DOE SC Advanced Scientific Computing Research (ASCR) program office sponsored the DOE Network 2025 workshop in Bethesda, Maryland, February 1–2, 2016. This workshop was driven by the need to develop R&D strategies in the network and transport protocol areas for science networking, which are not being addressed by industry due to the specialized needs of the DOE SC science community. This need can be addressed by short- (1–3 years), medium- (4–6 years) and long- (10–12 years) term strategies that set two major objectives.

1. Increase routine end-to-end bulk data transfer rates by 6 orders of magnitude, namely, from 1 terabyte/hour (2.4 Gbps) to 1 exabyte/hour (2.4 Pbps).

2. Improve network capabilities to simultaneously support a large number of data flows with a wide mix of complex and diverse performance characteristics.

Over the course of 2 days of plenary talks and breakout sessions, the DOE Network 2025 workshop developed an integrated R&D portfolio for the short-, medium-, and long-term time frames by examining science application requirements; assessing the current and projected technology trajectories; and identifying the needed technologies, protocols, and priorities. Integrated and agile transport solutions will be needed for both short and medium terms, and they are expected to be adequately supported by the backbone capacities of existing and planned future infrastructures. The long-term solutions with Pbps capacities, however, are much more challenging, requiring solutions significantly beyond the projected technology trajectories: in addition to the much higher capacity, the corresponding integrated and agile protocols will be needed to achieve these rates.

2. FINDINGS AND RECOMMENDATIONS

The DOE SC science ecosystem, consisting of instruments, supercomputers, storage, and other systems, presents unique networking challenges that are highly unlikely to be met as side effects of advances in industry. In particular, these solutions not only need to be customized, optimized, and integrated into the science ecosystem, but also need to be made transparently available to science users and automated workflows.

SHORT-TERM RECOMMENDATIONS

S-1: ASCR should continue funding research to provide scientists with the tools and services that routinely use a reasonable portion of the end-to-end network capacity.

S-2: ASCR should continue funding research into the development of monitoring, analysis, and predictive tools to improve the performance of science applications.

S-3: ASCR should continue to fund end-to-end transport solutions that integrate network, storage, and Input/Output (I/O) systems.

S-4: ASCR should continue funding research that exploits Software-Defined Networks (SDNs), Network Functions Virtualization (NFV), and other virtualization technologies for science.

S-5: ASCR should continue funding network research in support of integrity and chain-of-trust of science data sets.

S-6: ASCR should continue funding research into the development of tools and services that aid in network monitoring, troubleshooting, and debugging.

S-7: ASCR should establish a mechanism to prioritize and fund start-up short term activities that will lead to medium- and long-term research programs.

MEDIUM-TERM RECOMMENDATIONS

M-1: ASCR should fund research into transport layer protocols that can work effectively over high speed links with typical uncorrected loss rates.

M-2: ASCR should fund research into new transport layer protocols that can work well over multiple parallel links and over extremely high speed links.

M-3: ASCR should fund research into network layer protocols that can efficiently support a wide range of service requirements.

M-4: ASCR should support research in the creation of effective management tools and services that match the complexity and dynamic nature of the network infrastructure.

M-5: ASCR should conduct research into better tools and methods for actively and passively monitoring the performance of large, complex network infrastructures.

M-6: ASCR should fund research into better, more intelligent debugging tools that will support the expected network infrastructures.

M-7: ASCR should develop the complex models and simulations needed to create high-fidelity predictions that match observed workflow behaviors.

M-8: ASCR should invest in the development and deployment of hybrid systems that effectively support the end-to-end movement of data across parallel and multi-domain infrastructures.

M-9: ASCR should fund research into mechanisms that allow the network to simultaneously carry multiple types of traffic flows with minimal cross-flow impact on the performance.

LONG-TERM RECOMMENDATIONS

L-1: ASCR should invest in research to manage and control massively parallel data flows in a multi-domain network environment.

L-2: ASCR should develop management and troubleshooting tools that can make massively parallel network infrastructures operationally reliable.

L-3: ASCR should develop a wide array of protocols and services to ensure the management of multisite data transfers and interactive flows.

L-4: ASCR should explore new architectures that support the rapid growth of parallelism, high performance protocols, and management services.

L-5: ASCR should fund research into advanced cyber security mechanisms that protect both data and infrastructures from harm and exploitation.

L-6: ASCR should investigate multiple Pbps networking areas in more detail to better articulate the problems and identify potential solutions that can be economically deployed.

There are four major themes that cut across various recommendations that drive future work:

1. predictability, usability, and user experience (S-1, S-2; M-4, M-5, M-6, M-7; L-2);
2. integrated storage, file, computing, and instrument end systems (S-3, S-4; M-2; L-1, L-3, L-4);
3. chain-of-trust, integrity, and access control for data (S-5; M-3; L-5); and
4. deployment and diagnoses for operations and performance (S-6; M-1; L-6).

2.1 Known Knowns

The short-term data rate of 1-terabyte/hour (2.4 Gbps) is achievable today using host-to-host memory transfers and disk-to-disk data and file transfers over the current infrastructures. However, routine bulk data transfers at these rates are not that common. In addition, streaming data at these rates and combining large and small data flows (e.g., elephant vs. mice flows) continues to be problematic. It should also be noted that bulk data transfers typically rely on parallel and/or concurrent TCP/IP sessions to achieve high throughput rates.

The medium-term goals will require targeted R&D efforts in both hardware and protocol upgrades to achieve the required rates. In most cases the solutions can be developed based on robust, successive refinements of the existing technologies and strategies.

2.2 Known Unknowns

There does not appear to be a current path for achieving the long-term goal of routine Pbps bulk data transfers. Current backbone technologies and protocols would require massive numbers of fiber optic cables all operating at Tbps rates. It is unclear whether existing physical infrastructures will support these rates, or the optical modulation schemes that would need to be developed and widely deployed currently exist. It is also unclear what interconnect technologies hosts will use to drive individual data streams and the amount of parallelism that will be needed to reach these Pbps rates. Furthermore, much uncertainly exists about development of the integrated and agile network and transport protocols needed to support science discovery at these rates.

3. Workshop Overview

In April of 2015 the DOE SC ASCR program office convened a small group of network researchers and tasked them with evaluating the current state of network research. This group determined that future DOE scientific research programs would be hampered without a significant effort to improve the DOE science network infrastructures and services. The group recommended that ASCR convene a larger community workshop to identify the specific research challenges that needed to be addressed.

During the summer of 2015, the organizing committee for DOE Network 2025: Network Research Problems and Challenges for DOE Scientists Workshop was formed and generated and issued a call to the open community for position papers in September 2015. The submissions were reviewed, and 19 were selected and posted on the conference website before the workshop (Appendix A). The authors of selected papers were invited to the workshop as were other members of the network research community. A number of panelists and invited speakers were also identified, and they were provided with a list of questions and topics to be addressed (listed in Appendix B).

On February 1–2, 2016, ASCR brought together about 40 experts from academia, industry, and national laboratories to conduct this workshop. The workshop was organized around a series of small breakout sessions that each tackled the same set of questions (Appendix B). Each breakout session focused on research challenges associated with the time horizons for short, medium, and long terms (as shown in the agenda included in Appendix C). This report synthesizes those breakout group discussions and presents a set of findings and recommendations that can guide ASCR in the creation of a decadal long basic network research program.

The workshop opened with a presentation by the DOE program manager describing how DOE SC programs increasingly rely on global networks to reach their science objectives. It was also noted that it can take a decade or more before basic research activities produce widely deployable tools or services; yet this research should start now if these services are to be produced. Finally, it was noted that the network infrastructure is becoming more complex with more parallel and redundant paths, higher speed links, more complex encoding schemes, and an increasing need to ensure the reliability and integrity of data in flight. The conclusion was that research into new/revised network and transport layer protocols is needed to keep up with the demands of both science use cases and infrastructure upgrades.

This talk was followed by a panel session where members of the DOE and academic science communities described the network needs of their particular science communities. Each panelist gave a brief presentation that highlighted the exponential growth in the amount of experimental and observational data that his/her science community was generating. They also highlighted the global nature of science, with hundreds to thousands of scientists remotely using data captured at some, possibly remote, site. Data volumes ranging from petabytes to exabytes are forecast to routinely come from experimental facilities and simulations over the next decade. These presentations were followed by a detailed question and answer period that allowed workshop attendees to better understand the issues raised in these presentations.

Following the panel session, the major work began. The workshop organizers identified three major time scales—short -term (1–3 years), medium-term (4–6 years) and long- term (10–12 years), each requiring a targeted R&D strategy.

A major emergent theme was that short-term research should focus on tasks where the fundamental issues are known, but specific solutions are still being developed. Put another way, the workshop was tasked to identify major problems facing the science communities today. As a strawman objective the organizers noted that today it is possible to achieve a 1 terabyte-per-hour (2.4 Gbps) bulk data transfer rate over most research and education networks (RENs). However, routinely obtaining these rates is extremely challenging and difficult for individual scientists. Achieving these rates for other types of traffic flows to support computational steering and instrument control is even more difficult. Research that provides tools and services that routinely achieve these rates will greatly improve scientists’ abilities to conduct their research and contribute to the scientific discovery process.

The medium-term research theme should focus on the challenges associated with the network infrastructures required for the near future. As noted during the introductory presentation, future network infrastructures will be highly parallel and redundant, resulting in increased complexity and the need for advances in tools and support for automation. It was also noted that individual host interface speeds are not increasing as fast as they have in the past. Although 10 Gbps is common and 40 Gbps network interface cards (NICs) are deployed in high performance servers today, 100 Gbps NICs are just starting to hit the market, largely for network interconnects. Issues that must be addressed before NIC speeds increase further include increased internal bus interconnect speeds/protocol capacity and increased computational parallelism. Thus, there appears to be an increasing need for network/transport protocols to handle these node-specific performance and parallelism issues. As a strawman objective the organizing committee recommended that aggregate bulk data transfers routinely achieve petabyte-per-hour (2.4 Tbps) rates over transcontinental distances.

Long-term research should focus on identifying and exploring fundamental challenges that arise with the growing demands of science. With exabyte data sets, exascale supercomputers, and massive arrays of sensors on the horizon, ASCR needs to begin networking research now if there is any chance of having solutions ready in the next decade. As a strawman the organizing committee recommended that aggregate bulk data transfers routinely achieve Exabyte-per-hour (2.4 Pbps) rates over global distances. Workshop attendees concluded that creating protocols that can manage this type of complexity poses a significant challenge but one that if solved would have a significant impact on day-to-day science activities.

To encourage and support a diverse and detailed set of discussions the organizing committee established four breakout groups and assigned roughly 10 attendees to each group. Each group was then tasked to address each short-, medium-, and long-term research challenge. A unique set of questions was developed for each theme as was a set of overarching questions that covered all three themes (see Appendix B). The breakout groups were monitored by members of the organizing committee, who also identified a scribe and lead to report the discussion results back during plenary sessions. Following the workshop, select groups of scribes, leads, and organizing committee members processed the notes and produced this workshop report.

4. SHORT TERM

Short-term goals generally comprise areas of immediate need and importance to be developed within the next few years. The participants felt that the short-term challenges could be classified into two categories: (a) areas with immediate impact and (b) essential building blocks for longer term impacts (i.e., areas that should be started on now to meet the future networking research challenges both in the medium and long terms).

Application drivers: Application drivers in the short term are dominated by data transfers, in particular, wide area file transfers involving disparate file systems such as Lustre and the General Parallel File System (GPFS). Moving large amounts of data between globally distributed sites is a common occurrence in the DOE SC laboratory complex. Unique science instruments are typically accessed remotely, and globally distributed data analysis services are common in several science communities. National and international RENs built to handle these bulk data flows provide the backbone for these activities.

In addition, there has been a steady increase in other traffic flows (e.g., interactive exploration of simulations, large data analytics jobs, and remote monitoring of code executions on supercomputers) that have different performance needs. While these applications are in limited use currently, they are expected to be some of the key drivers in the medium and long term. Workshop attendees noted that some of the problems associated with newer applications and increased traffic flows will begin to appear in the medium term, and the data rates, volumes, and varieties will continue to grow.

The short-term data requirement of 1 TB/h or 2.4 Gbps stated in the introduction is achievable today at the network level over widely deployed 10 Gbps connections using multi-core hosts with 10 GigE (Gigabit Ethernet) NICs and production TCP/IP protocol stacks. In principle, such network throughputs, by themselves, do not currently pose a challenge when operating over dedicated or lightly loaded 10 Gbps end-to-end paths. However, not all science flows today routinely achieve 2.4 Gbps TCP throughput [e.g., bulk data transfers using default TCP over end-to-end paths with 200 ms round-trip times (RTTs) are typically under 1 Gbps without manual tuning]. The challenge is getting 99% of the bulk data transfer traffic to routinely achieve the 2.4 Gbps throughput the network supports [13].

Recommendation S-1: ASCR should continue funding research to provide scientists with the tools and services that routinely use a reasonable portion of the end-to-end network capacity.

Understanding the operational state of the network (e.g., the load on individual links, the queue lengths on router or switch ports, the impact of failed or failing components) is a significant challenge. Management tools routinely capture port and link use data and monitor the health of routers and switches throughout the network. However, traffic rates can fluctuate widely as new flows begin and existing flows terminate, and thus capturing the instantaneous changes in the network can be difficult. Another significant challenge is identifying soft errors in network paths. A router with an optical transmitter that is slowly degrading is much harder to identify than one that has failed completely. Troubleshooting is such a manual process that even an experienced network engineer can spend days or weeks trying to find intermittent faults or problems. Customized analysis tools that can capture and automate the engineer's actions would greatly improve the operator's ability to maintain the high performance aspects of the network.

The remainder of this section proposes short-term goals that are intended to serve as building blocks for future research programs and activities. These goals are organized into four themes: (1) predictability, usability, and user experience; (2) integrated storage, file, computing, and instrument end systems; (3) chain-of-trust, integrity, and access control for data; and (4) deployment and diagnoses for operations and performance. In each case, the underlying needs and challenges are described first, followed by a brief overview of the state of the art and a list of specific tasks that should be explored.

4.1 Predictability, Usability, and User Experience

Challenges: A continuing issue in the area of user experience is the challenge that networks present to nonexpert users in terms of efficient use and application performance optimization. Domain experts typically design applications that are expected to operate in a networked environment. These applications seldom operate at the expected performance level without extensive fine-tuning and troubleshooting activities. Application developers should not need to become network experts to maximize the performance of their applications; nor should they be required to debug the network for correctness or performance.

There is a lack of monitoring infrastructure and instrumentation for determining causality and thus explaining an application's performance in any given system and network. Currently domain experts and network experts have limited means to accurately predict an application's performance on a given network. Having such a capability would benefit both scientists and network operators. Predictability does not have to be perfect to be useful at preliminary stages; it need only be accurate enough to calculate expected performance with high confidence and low effort. Thus a major challenge is development of the infrastructure and instrumentation needed to provide a rapid prediction capability that could improve performance engineering.

In addition, it can be difficult for application developers and users to access and control the network configuration parameters that are relevant for their applications. Part of this challenge, including predictability of network performance, is due to the complexity of today's networks and the interactions with numerous other applications sharing resources. Even if these challenges could be resolved, designing next-generation networks will still be challenging because future applications are not known, notably their expected use of network aggregate load and load imbalance. Many of these challenges are made more difficult by the multi-domain nature of modern large-scale networks. In such settings, common tools are needed that span very different domains for a common purpose. This requires cross-collaboration for the development of common protocols and infrastructure.

State of the Art: Today's systems are hard to predict and analyze. Domain experts are often unable to explain why their applications do not make full use of network capacity [13]. Some simulation tools exist, but they suffer from precision loss, a lack of scalability to realistic system sizes, unreliability, or difficulty of use. Some analytical models also exist, but typically only for simplified versions of the network or simplified traffic [7]. In addition, today's systems have not been configured transparently and do not provide Application Programming Interfaces (APIs) that allow a user to extract configuration parameters. Configuration includes not only network topology and overall network capacity, but also link bandwidths, buffer space, quality of service parameters, physical and virtual channels, etc. This challenge is more severe when trying to find relevant information about end nodes, such as determining offered services and nodes configurable tuning parameters. The community has identified a small set of applications that it believes to be characteristic of certain domains, but no longer term application property projections exist that suffice as design and research guides. Finally, crossing domain boundaries in networks is a challenge due to the lack of common tools and protocols.

To address these challenges, the workshop attendees identified five research tasks that should be started or expanded.

Research Tasks

1. *Identify representative applications*. A significant contribution would be to identify the set of applications that represent future traffic patterns and to predict how their demands will increase at larger computation scales and in future systems.

2. *Develop monitoring and analysis tools*. These tools should automatically generate feedback on what parts of the network are suffering from contention, poorly reserved resources, interaction with other applications, or any other factor that degrades performance. These tools should provide recommendations on how to modify an application (such as the application's traffic pattern) to increase performance as well as provide debugging information.

3. *Use tools to quantify the impacts of changes in network configuration, soft failures, or permanent faults*. This information is relevant to network administrators, designers, and application developers alike. This information can also be used to debug network functionality, especially if the network provides relevant feedback to the application. It is even possible to envision two-way interactions between domain scientists and the network such that the former notify the latter in advance of the nature of their expected communication.

4. *Develop prediction tools*. Given an expected network configuration and traffic pattern, these tools should be able to predict with high confidence network performance in terms of throughput, latency, and energy. Taking this one step further, computational models can be developed that predict computation and stall times for processing nodes and use them to predict the entire application's execution time. These models can also inform the domain expert developing the application with recommendations on how to optimize the application for a specific network, such as matching the traffic pattern to the network topology.

5. *Readily expose network and end node configuration parameters to users*. This will enable better optimization and transparency.

Recommendation S-2: ASCR should continue funding research into the development of monitoring, analysis, and predictive tools to improve the performance of science applications.

4.2 Integrated Storage, File, Computing, and Instrument End Systems

Challenges: Data transfers between high performance computing and science experimental facility sites often consist of different types of data being transferred between two remote sites. Files may be transferred from supercomputers to remote distributed file systems such as Lustre and GPFS, and measurement streams may be transported from science instruments to supercomputers. The physical paths over which these transfers take place can vary considerably; they may be composed of traditional Ethernet segments, InfiniBand (IB) connections to file systems, or custom cross-connect segments within a supercomputer. End systems themselves are complex, with multiple tunable parameters, and their participation in end-to-end flows can lead to impedance mismatches that can severely degrade the quality of the transfer. Distributed site file systems are mounted on disk complexes with parallel data servers and multiple metadata servers, which have to be aligned and matched with transport methods to ensure high throughput. Failure to match disk read/write performance with network performance leads to poor overall performance and low system utilization.

Historically, supercomputers have used batch-oriented operating systems to run large complex computationally intensive applications and simulations. In this mode, data are transferred to/from the local storage system. An emerging trend is to support near-real-time applications where data from scientific instruments are quickly analyzed to determine the quality of the collected data [8, 10, 12]. This may require streaming data directly into the supercomputer, bypassing the local storage system. This use of supercomputers presents another class of challenges because data may traverse nontraditional physical paths from compute nodes to LANs and data transfer nodes (DTNs) across internal interconnects.

Another set of challenges arises when end systems are science instruments, some of which generate data at very high rates (e.g., terabits per second in Large Hadron Collider and Spallation Neutron Source experiments). These instruments typically stream data from the detector to a local or remote storage system for post analysis. As noted previously, an emerging trend is to stream data directly to the supercomputer for a near-real-time analysis. Multiple parallel links are needed to support these high data rates, and the challenge is efficiently supporting this inherent parallelism across WANs.

Finally, scientists have expressed an interest in supporting more interactive uses of supercomputers and data intensive instruments. This requires support for the concurrent data and control flows needed for computational steering and computation-driven experimentation over globally distributed science communities.

State-of-the-Art: Currently traditional transmission control and user datagram protocols and services have been extended or adapted for file systems, instruments, and supercomputers, often in point and ad hoc solutions. There are two fundamental limitations to this approach. First, it creates impedance mismatches between subsystems and does not adequately exploit the inherent data and executional parallelism. For example, multiple I/O streams are needed for file systems, which must be appropriately matched with TCP parallel streams to ensure high throughput. Second, data transfer protocols do not provide the stable dynamics needed for control channels for computational monitoring and steering. For example, TCP is designed to reduce its sending rate (“back-off”) in the presence of losses, whereas the opposite (that is increased sending rate) may be needed in certain environments. Currently, science data transfers are supported by parallel TCP streams or a combination of TCP and UDP streams, which are limited in supporting disparate requirements such as stable control flows coexisting with large bulk flows. For example, there is no simple way to separate the contents of a TCP flow into data and control flows with different requirements. In terms of end-to-end performance, current approaches are limited to simple and ad hoc solutions that attempt to minimize the impedance mismatches and provide robust performance monitoring and prediction.

To address these challenges, the workshop attendees identified seven research tasks that should be started or expanded.

RESEARCH TASKS

1. *Parallel data transport*. Transport protocols and middleware are needed for transport between parallel file systems to match striped files and distributed metadata servers that use parallel flows with paced I/O.

2. *Transport for monitoring and control*. A new class of protocols is needed to provide stable dynamics for monitoring and control channels, especially supporting low latency and jitter to ensure rapid response.

3. *Transport over multimodal paths*. Methods are needed for smooth and efficient data transport over complex paths from supercomputer nodes and interconnects (e.g., over a concatenation of IB and Ethernet paths).

4. *Parallel and heterogeneous data flows*. Protocols and architectures are needed that enable the coexistence of multiple high-bandwidth data flows with low-bandwidth control channel flows.

5. *Impedance matching*. Systematic methods are needed to align various subsystems to ensure efficient data flows by co-optimizing their parameters to achieve high and stable throughputs and to ensure efficient monitoring and diagnosing of performance problems.

6. *Profile composition and performance prediction*. Analytical and experimental methods are needed to generate performance profiles of individual subsystems and to compose them for end-to-end performance prediction.

7. *Integration with SDN, NFV, and other virtualization technologies*. Solutions are needed to integrate and take advantage of emerging virtualization technologies (SDN, NFV, Docker, Vagrant, etc.) in the network and end subsystems.

Recommendation S-3: ASCR should continue to fund end-to-end transport solutions that integrate network, storage, and I/O systems.

Recommendation S-4: ASCR should continue funding research that exploits SDNs, NFV, and other virtualization technologies for science.

4.3 Chain-of-Trust, Integrity, and Access Control for Data

Challenges: A gap exists between the security, privacy, and provenance features that exist in corporate and mission-critical national laboratory networks and between features used for the science end of DOE networks. Too often security, reliability, and resilience mechanisms are not supported over the complete data life cycle (i.e., for the transition of data from generation at the data sources, to data being preprocessed at the data stores, to the processing of data to obtain domain-specific results, to sharing of data among researchers). Once generated, most data reside in data clusters behind secure firewalls, but there are seldom mechanisms to (1) enable secure data generation at the data sources (e.g., from the myriad of sensors used to capture data in remote experimental locations); (2) enable encrypted data transfer from data sources to the data stores that exist behind firewalls; (3) enable intrinsic embedding of provenance information in the data, such as a chain of provenance reflecting the individuals and entities that have processed the data over their life cycles; (4) identify inaccuracies in the data resulting from malfunctioning (or compromised) data sources and subsequently identify the sources; (5) track changes in the processed data shared between researchers that work independently to verify/dismiss each other’s research conclusions; and (6) provide domain users with a seamless experience to confirm data validity, accuracy, and integrity, even when the data are shared among several users spread across the globe, any of whom can modify the data. These challenges are not currently addressed in most DOE SC networks and applications.

State of the Art: This subsection presents the state of the art in DOE SC data from the data life cycle viewpoint. Data are generated from a large number of sensors and other data sources that have been in place for several years, most of which are in remote fields or in facilities where they are not secured behind strong firewalls. The data can be easily corrupted in transit; there is no support for fast and compute-efficient encryption. The sensors are prone to malfunctions (especially as they age) and, with more of them available directly on the Internet, are prone to compromise as well. These sensors could become break-in points through which malicious users could access the protected networks—these data sources will become the first line of contact.

Typically the created data are transmitted to data stores that reside behind firewalls as part of secured and monitored clusters and preprocessed by an IT group or a group of domain experts. However, frequently there is no information about who within the group created the resultant data. The data are accompanied by message digests to indicate whether they have changed or were corrupted in transmission; however, there are no mechanisms for guaranteeing message authenticity and provenance—the data and message digest pairs can easily be changed together, leaving no method for identifying such tampering. Scientists operate under the trust model that anything they access from a secure data store (typically through an SSL tunnel) has not been tampered with. That is, the assumption is that the data are authentic (not corrupted due to errors or malicious intent). In the event of an error, there is no way to identify the point of error creation (i.e., by the data source, during transmission, or during storage).

Further, currently there is no mechanism for identifying ownership of data or defining a chain-of-provenance, which would help identify the entities or processes that have worked on the data—the data go through several domains on the journey from generation at the source to consumption by clients (scientists running models, visualizing, etc.). A mechanism is needed to define and implement a per-application security level in which the user or system architect decides the level of security essential for the application and the network. Mechanisms are also needed to watermark the data. There has been an attempt to implement science Demilitarized Zones (DMZs) and move them closer to the end users in university networks across the globe, but progress on this has been slow. New technologies such as software-defined networking and network virtualization and networking paradigms such as information-centric networking are being explored, and the community needs to determine which of these can be leveraged for improving end-to-end data security.

To address these challenges, the workshop attendees identified four research tasks that could be started or expanded.

Research Tasks (in decreasing order of short-term priority)

1. *Secure the growing number of sensors in the field.* Given the large number of data sources that will continue to grow, with more coming online on the Internet, there is an urgent need to ensure they do not become points of entry into secured networks.

2. *Create a chain of provenance for data that changes many hands.* Guaranteeing provenance is important and especially so when data changes many hands during its life cycle.

3. *Design mechanisms for identifying data corruption and its source.* Identification of malformed, inaccurate, or malicious data and data corruption or modification is also important as data are becoming "big" and are being shared more widely.

4. *Integrate new technologies and paradigms.* New technologies such as software defined networking and network function virtualization and networking paradigms such as information-centric networking may have the potential to be leveraged to enable security and resilience of data, depending on their ability to scale as the volume of data, the number of users sharing/accessing data, and the required speed at which data are transmitted to end users all increase simultaneously.

Recommendation S-5: ASCR should continue funding network research in support of integrity of, chain-of-trust of, and access to science data sets.

4.4 Deployment and Diagnoses for Operations and Performance

Challenges: An important goal of the DOE scientific network is to enable innovations in protocols, software tools and other services that support research. However, network operators and researchers today face a challenge in testing and deploying these new capabilities without disrupting existing traffic on production networks. They also have difficulty understanding and diagnosing problems in production systems that affect application performance.

State of the Art: Researchers and network operators currently test protocols by running them on end systems, but they do not have tools to ensure that the test traffic does not interfere with other ongoing activities in the network. Moreover, there is no easy way to test and deploy new functionality inside production networks.

Although flows in controlled environments have achieved 100 Gbps, the median end-to-end performance over the Energy Sciences network (ESnet) is still poor. A major obstacle to achieving better end-to-end performance is that routers do not collect the information needed to support the monitoring and diagnosis of application performance problems by network operators. Moreover, protocols running at the routers do not contain feedback mechanisms to detect problems in real time.

To address these challenges, the workshop attendees identified four research tasks that could be started or expanded.

Research Tasks

1. *Support research to enable testing of new software* in current production networks but in an isolated (e.g., "test") mode such that network functionality and traffic are not affected. This isolation needs to extend from end nodes into the network to enable testing of potentially end-to-end protocols.

2. *Ease deployment of new protocols and tools into production networks* without halting existing activity on the network.

3. *Incorporate measurement capability* into network protocols and router software to support the monitoring and diagnosis of performance problems.

4. *Build real-time failure detection and handling mechanisms* into networked systems to deliver predictable and high end-to-end performance that meets application needs.

Recommendation S-6: ASCR should continue funding research into the development of tools and services that aid in monitoring, troubleshooting, and debugging networks.

In general terms, across all the medium- and long-term areas discussed next, there is a need to identify and start certain basic R&D activities in the short term to lay the necessary foundations.

Recommendation S-7: ASCR should establish a mechanism to prioritize and fund start-up activities that will lead to medium- and long-term research programs.

5. MEDIUM TERM

The goal of this breakout session was to look at the research challenges that will become relevant in the medium term (4–6 years). The organizers proposed two challenges and a list of questions that each breakout group was asked to address. The first challenge was to increase the bulk data transfer rate by 3 orders of magnitude and have 1 PB/h transfers (2.4 Tbps) become a common activity. The second challenge was to support a wider variety of traffic flows—beyond simple bulk data flows across the multi-domain RENs the DOE SC community uses.

There was general agreement that petabyte-per-hour transfer rates are achievable goals in the 2021 time frame. The predicted deployment of 400 Gbps interfaces and flexible wavelength optical transport systems is expected to provide the basic building blocks for such infrastructures. Core networks will have multiple 400 Gbps lambdas available, which can be bonded to reach these terabit/sec rates. Furthermore, it is also possible that 1 Tbps optical technologies using single lambdas will also be available in the 2021 time frame. Faster, large capacity storage systems (e.g., solid-state drive and NVM Express) will be able to source/sink data at the rates necessary to keep these links full. Multihomed hosts, or small clusters, will be able to physically combine the five to six 400 Gbps paths needed to achieve petabyte-per-hour transfer rates.

Although the physical infrastructure is expected to be in place to support these bulk data transfers, it is less clear whether they will be available to all science users. Hosts currently rely on Peripheral Component Interconnect Express (PCI Express or PCIe) bus technologies to transfer data between storage/memory and the network. Ethernet NIC's are the primary network interfaces, while IB and Ethernet both support network-attached storage devices. The major impediment to increased end system networking speed is the PCIe bus technology. Version 4.0 (PCIe 4), due out in 2018, will provide 512 Gbps transfer capabilities on each 16-lane connector. This essentially limits the number of very high speed NICs an individual host can support.

An additional concern is the number of lambdas available at the edge of networks. Although dense wavelength division multiplexing (DWDM) systems are already widely deployed in network cores, they are not widely used in regional or campus networks. In addition, even core networks do not typically use every DWDM channel, further limiting the end-to-end capability of networks. Workshop attendees assumed that the physical network infrastructure will be adequately provisioned to meet the needs of the DOE science community.

Workshop participants spent less time discussing how networks will handle other types of science traffic (e.g., that needed for monitoring and steering of codes running on supercomputers and automated analysis codes and remote users driving experiments). DOE science communities are showing increasing interest in streaming data from large science instruments and computations and carrying out analyses on remote computers to prepare for the next cycles of science experiments. This includes widely distributed sensor arrays and extremely dense charge-coupled device–based detectors. In addition to the increased number of sensors, their refresh rates are also increasing, allowing detectors and sensors to collect more data in the same amount of time.

The traditional method for handling different traffic types is to create multiple independent or virtual networks that reserve specific amounts of capacity for each traffic type. SDN technologies can play a role here, it is unclear whether network operators will release operational control of networks to individual scientists or science communities. DOE needs to fully explore this issue to understand the implications of virtualized networks and the mechanisms needed to implement and operate this functionality.

Thus, although it is probable that the physical infrastructure will be available to support end-to-end petabyte-per-hour bulk data transfers, the protocols for doing so and the ability to simultaneously handle both bulk data and other classes of traffic flows are open research question.

Workshop participants were tasked with addressing three general questions and six breakout session–specific questions. Each group was asked to rank order the breakout questions to provide guidance on the priority that it placed on developing the research program needed to answer these questions.

Application drivers

The breakout groups were asked to comment on application drivers that will justify building the proposed infrastructure. We identified a generic set of application drivers that were also discussed in the short-term session and more generally at previous workshops [1, 4, 9, 11]:

- bulk data transfers,
- computational monitoring and steering, and
- computation-driven experimentation.

Because this question was discussed during the first breakout session, little additional discussion was conducted in this session and no new applications or application requirements were noted.

What basic services/functions are needed/expected from networks?

Workshop participants expressed an interest in improving the network monitoring environment, scaling existing monitoring and management tools to match the speeds and complexities of the new infrastructure. They also expressed interest in identifying ways to secure network switches/routers and to ensure that authorization and authentication systems improve and scale appropriately. These security and trust issues need to scale to peer networks operated by multiple independent administrative network domains (e.g., backbone, regional, campus).

HOW WILL NETWORKS BE USED, AND HOW DOES THAT CHANGE OR STAY THE SAME?

The participants discussed how to deal with a more heavily loaded network infrastructure and what mechanisms operators will need/have to manage this limited resource in providing the required data rates to applications. As noted previously, the physical infrastructure needed to support petabyte-per-hour bulk data transfers requires multiple 400 Gbps links at each end of the path and multiple lambdas through the core. Participants also expect other types of traffic flows, from long-lived streaming flows to content-based delivery type flows, to emerge in the next 4–6 years. This diversity led some participants to consider the need for tools and incentives that will encourage users to shift load from peak to off-peak times. This approach will allow more efficient use of this limited infrastructure capacity to more efficiently support the science users.

RANKING OF MEDIUM-TERM RESEARCH NEEDS

Participants were asked to discuss and rank six areas ripe for research in the network and transport protocol space. Those research areas and the ranking they came up with are as follows.

1. Support of bulk data and more interactive flows
2. Provision of more predictable transport performance
3. Provision of increased monitoring and management of the network
4. Support of improved debugging and troubleshooting
5. Provision of routing control over this infrastructure
6. Support of active queue management in routers/switches (e.g., congestion notification—helps with predictable transport performance)

5.1 EFFECTIVE BULK DATA TRANSFERS AND MORE INTERACTIVE FLOWS

For the past decade, bulk data transfers have been the cornerstone of scientific networking. In the early days it was a single host transferring data to a remote host. This evolved to a small cluster of nodes acting in concert to transfer data and now collections of specially tuned hosts acting together to transfer data between sites. Bulk data represents the majority of data transferred over ESnet—more than 100 PBs per month at the time of the workshop. It has also been shown that these dedicated transfer nodes can achieve 40 Gpbs NIC rates using the existing TCP/IP suite of protocols. However, to do so requires the network to be nearly loss free. For some protocol implementations, the loss of one packet can cause throughput to collapse followed by a large number of round-trip exchanges to recover from the loss, resulting in very low effective path loss rates

The problem is well known, TCP's throughput is a function of bandwidth, delay, and packet loss. The larger the bandwidth-delay product, the lower the loss rate must be to maintain an adequate throughput. Over the past two decades the community has focused mostly on the delay portion of this equation. The longer the distance between nodes is, the greater the impact on throughput. However, at Tbps link rates, the distances over which problems occur become smaller. To illustrate this, a 10 Gbps link over continental distances (100 ms RTT) has the same bandwidth-delay product as a 1 Tbps link over a campus distance (1 ms RTT). Each of these paths can see a significant drop in throughput by losing just one packet (for some variants of TCP), which can take more than an hour to recover, requiring path loss rates below 10^{-14} bit error rate.

The other issue that needs to be considered is that no system has an absolute zero loss rate. Loss rates for bit error rate–corrected optical systems can be exceedingly small—as low as 10^{-15}—but this is not zero. As link rates increase, bit error probability increases. For example, 1Tbps (10^{12} bps) link running over an optical system with a 10^{-15} loss rate would be expected to experience a 1-bit loss every 17 minutes. Increased error correction encoding can reduce these rates to any desired (but nonzero) level, but with increased link encoding comes complexity and delay.

Recommendation M-1: ASCR should fund research into transport layer protocols that can work effectively over high speed links at typical uncorrected loss rates.

As noted previously, the PCIe 4 protocol will support up to 512 Gbps connections. This will allow 400 Gbps NICs to be inserted into a single multi-core host. At least two NICs will be present in most high performance DTNs. One will connect to a high speed storage system and the other to an external network. This will result in large bulk data transfers relying on 5–6 DTNs all working together to satisfy the transfer request. This will require the use of multiple 400 Gbps lambdas across the entire end-to-end path or a larger collection of slower rate lambdas bundled together to form this end-to-end path.

Mapping a single 400 Gbps NIC onto a single 400 Gbps lambda appears to be a logical way to implement a transfer, but it may not always be feasible to find such a resource along the entire path. Multiplexing lower rate links together to carry this high speed traffic is an alternative approach. Unfortunately, splitting flows across multiple parallel links has proven to be very difficult. Some current TCP implementations require in-order packet delivery to maintain high performance. Hardware vendors have taken great pains to ensure that only a single link in any multi-link path is used for an individual flow. Although this prevents the network from reordering packets, it also limits throughput to the speed of that individual link. An effective means of using multiple parallel links (anti-multipath as one participant called it) is a requirement of future networks.[*]

Recommendation M-2: ASCR should fund research into new transport layer protocols that can work well over multiple parallel links and over extremely high speed links.

Although bulk data transfers have been the data driver for several decades, networks have always carried a mix of traffic types. Sometimes grouped only as "Elephant" and "Mice," these flows can exhibit a wide

[*] Recently a multipath TCP (MPTCP) has been introduced that deals with reordering that occurs due to flows being split across multiple paths. A promising research direction is to investigate its use and adaptation to the very high bandwidth, highly parallel link settings envisioned in the midterm.

variety of characteristics. It has proven extremely difficult to categorize these different types of flows. Previous attempts have identified loss, throughput, delay, and jitter as fundamental path characteristics and metrics that should be considered when looking at the requirements of an individual flow. However, it is not clear whether these characteristics can be independently managed (e.g., high throughput requires low loss and low delay).

Current support for such flows attempts to separate only large bulk data flows (Elephants) from low-speed flows (Mice). This results in two levels of service to better manage network resources. It is unclear how (or whether) this partitioned approach can be extended to support additional flow types. Using just the four characteristics listed above and having three levels of granularity (high, medium, and low) for each results in 81 service levels. Increasing the resolution of these levels would further increase the number of service levels.

Networks need to offer multiple service levels to applications in a fair and obvious manner. Applications need to choose the appropriate service level and have a disincentive (including monitoring and enforcement) to using an incorrect or excessive service.

Recommendation M-3: ASCR should fund research into network layer protocols that can efficiently support a wide range of service requirements.

5.2 MONITORING AND MANAGEMENT:

As also identified in short-term discussions, managing today's networks requires a technically skilled staff with a suite of sophisticated tools. Routers, switches, and servers all require configuring and customization to work in the deployed environment. Maintaining and updating these configurations pose a significant challenge. This is just the basic level of management; beyond this is the work required to identify and respond to faults, misconfigurations, cyber attacks, and numerous other events that impact the stability and usefulness of the network.

Furthermore, the distributed nature of the network means no single entity owns or controls the entire end-to-end network path. This means that faults or problems anywhere in the network can cause noticeable effects that cannot be easily corrected.

To manage networks it is essential that operational data (flow records, interface counters, health monitoring, error logs, etc.) be extracted and composed in such a manner that operations staff make use of them. As network complexity increases, with bonded links and SDN-controlled flows, so does the necessity of creating better and more intelligent monitoring and management tools and services.

It is expected that future network infrastructures will offer more services than simple packet forwarding. Content delivery networks or logistical networks insert storage and possibly compute hardware and services into the network. SDNs also offer the opportunity for network operators to provide more service levels on a time scale faster than current network vendors can offer. All of this complexity needs to be managed and controlled by the network operations staff.

Recommendation M-4: ASCR should support research in the creation of better management tools and services that match the complexity and dynamic nature of the network infrastructure.

Monitoring the network requires that operations staff perform two different functions focused on answering the following: (1) is the network up and transporting data and (2) is the network performing up to specifications? These two tasks require different levels of data and may also require different mechanisms. Basic health monitoring (is a link up) can be achieved with relatively small amounts of data and a few simple tools. The data will scale linearly with the number of links or device interfaces. Every device interface is typically designed to report its link status (up, down, administratively disabled), and devices report changes in the link state.

Monitoring network performance requires a different suite of tools and the transfer of more intensive amounts of data over the network. In some cases active monitoring is used where monitoring hosts send data probes between two points in the network and record the observed throughput, delay, loss, and jitter. Such tests increase the load on the network. Passive monitoring is also used to observe existing flows and report their performance without adding additional load to the network. Determining the proper mix of active and passive data is a research challenge that needs to be resolved.

As the number of links and amount of parallelism increase, the ability to actively monitor network performance becomes more complex. Tools are needed to select specific paths through the network and maintain them throughout the monitoring period. Increasing the number of links also increases the number of tests that need to be performed. It is current practice to create a mesh of tests to actively monitor the network; adding more links increases the mesh size and increases the amount of measurement data being transmitted over the network.

Recommendation M-5: ASCR should conduct research into better tools and methods for actively and passively monitoring the performance of large, complex network infrastructures.

5.3 Debugging and Troubleshooting

As also discussed for the short term, while managing and monitoring the network is important, any complex system will fail and debugging or troubleshooting activities will be required. These activities also fall into two broad categories: (1) hard failures, where something stops working, and (2) soft failures, where things work, but not at the expected performance level. An additional difficulty is that different parts of the end-to-end path reside in different administrative domains, and no single network operator has the capacity to find/fix every problem that can occur.

The attendees agreed that currently this issue is addressed by a few highly skilled network engineers spending hours to months analyzing each problem as a one-off incident. The problem with this is that, unfortunately, this solution does not scale. There aren't enough highly skilled engineers, and there aren't enough good tools to assist the engineering staff in locating where performance problems exist. It is also understood that scientists and users have few, if any, tools to effectively report problems when they do occur. This means that once a problem is reported, the network engineer begins the troubleshooting task at step zero, trying to understand exactly what the problem is.

As new functions and services are introduced, and the physical infrastructure becomes more parallel and complex, these performance problems will become even more difficult to debug and fix. Having multiple parallel links means that problems may appear more random and chaotic as flows transparently alternate between different links. As the number of service levels increases, flows will have different characteristics and needs, and they will experience different impacts even over identical paths.

Recommendation M-6: ASCR should fund research into better, more intelligent debugging tools that will support the expected uses of future network infrastructures.

5.4 PREDICTABLE TRANSPORT

In overall agreement with the short-term discussions, participants agreed that current TCP/IP and UDP/IP protocols do not provide appropriate service guarantees. There was general agreement that providing strict guarantees requires significant changes in the way networks are managed and built. Yet, network operations and science user communities both continue to talk about the need for mechanisms that provide quality of service (QoS) capabilities.

Workshop participants were asked to consider a different approach, one focused more on creating predictive services instead of guaranteed services to provide this QoS capability. It was noted that predictive services do not guarantee service levels, just that the probability of meeting the expected service level is high.

Discussions focused on two major elements. The first is that traffic flows are end-to-end and that more than the network impacts the achieved performance. Other factors include CPU load, amount of cross traffic transiting shared links in the network path, file system performance, and storage area network performance. The second element is that TCP is adaptive. It attempts to fill the path with traffic, assuming that the sender has enough data to do so. TCP variants use different congestion control and congestion avoidance algorithms that determine how TCP performs over different paths.

Creating models and simulating the behaviors of every part of such systems with enough fidelity to make predictions is a challenging task. Doing so can provide the network operations community with a better understanding of how the infrastructure is operating. It can also allow science communities to better plan and use the networks.

Recommendation M-7: ASCR should develop the complex models and simulations needed to create high fidelity predictions that match observed workflow behaviors.

5.5 ROUTING

Attendees discussed the issue of routing and path building over multi-domain network infrastructures. There was general agreement that at the speeds being discussed (400+ Gbps) making complex, context-based packet forwarding decisions on a per-packet basis is not practical. However, there was agreement that simple packet switching decisions will continue to be practical. The challenge is how to get data through the network infrastructure in an effective manner.

Attendees put forward several ideas. One is to move to a more circuit-based network where optical links or lambdas are connected together to form a virtual or physical end-to-end circuit. Intermediate devices (i.e., optical add-drop multiplexers, optical switches, electrical switches) are programmed to transit arriving packets along this predetermined path. Another idea is to continue statistical multiplexing and allow intermediate switches to make forwarding decisions based on fixed, context-free packet header information.

Attendees agreed that the critical problem is that when the path or route changes, forwarding tables in intermediate routers and switches need to be reprogrammed with new information. Doing this in a timely manner is extremely challenging. Current link state routing protocols require significant amounts of time to detect and recover from a failure along this end-to-end path.

Although circuit-based paths may offer fast restoration, there is a cost to creating and holding these supplemental circuits. One challenge is that the setup time for circuits can be quite long O(millisecond) making them unsuitable for use on a per-flow basis for short-lived flows. Another challenge is that a circuit requires centralized control of the end-to-end path, at least inside a single domain. A control node is required to keep track of all current and pending requests to determine the optimal condition of the network. Doing this over a multi-domain path is extremely challenging.

Finally attendees noted that current routing and path generation controllers often create a single path between any two end points. Although some work has been done in multipath routing and packet forwarding, it is not clear how these techniques will scale to meet the demands of future DOE science data flows.

It is possible that no single solution, circuits or statistical routing, will meet every need. Thus a more complex and hybrid suite of solutions is required for REN networks.

Recommendation M-8: ASCR should invest in the development and deployment of hybrid systems that effectively support the end-to-end movement of data across parallel and multi-domain infrastructures.

5.6 ACTIVE QUEUE MANAGEMENT

Attendees discussed finer grained packet forwarding management techniques. The current Internet supports a single best effort class of service. Routers and switches treat every packet the same, typically hosting a single first in, first out queue to handle packet multiplexing. Although there have been several efforts to create more complex packet forwarding behaviors (i.e., type of service markings, differentiated services markings, flow header, random early detection, and controlled delay), nothing has been widely deployed over the global Internet or by the REN community.

Yet there continues to be an understanding that different traffic flows have different characteristics and requirements, and routers/switches should make more intelligent decisions when forwarding packets. For example, packets belonging to a Voice over Internet Protocol (VoIP) call have strict delay and jitter requirements beyond which the call becomes unusable. It is also understood that voice streams compress very well and thus the bandwidth needs of a VOIP call can be quite low. In contrast a bulk data transfer flow wants as much bandwidth as possible but is comparatively delay and jitter insensitive.

Historically the challenge has not been in creating mechanisms that can handle these different traffic flows. Instead the challenge has been in implementing, deploying, and enforcing the correct use of these services due to the design of network and transport protocols. Packets are marked by the source node and injected into the network. There is no mechanism that allows the network to tell the source node what services it can provide.

This lack of a network signaling mechanism makes it difficult to deploy any effective QoS forwarding mechanisms. In addition, application users and developers typically do not know what type of service they need to ask for. Finally, because the source node creates the original packet there is nothing to prevent the user from marking every packet as high priority, even if that is not appropriate, and efficient monitoring and enforcement remain an open issue.

What is needed is a mechanism that will allow scientists, application developers, and network operators to mark and send data flows that have different service needs. This may require the ability of source nodes to request virtual channels that have a specific set of characteristics. It may require that intermediate switches, routers, and end nodes police and shape traffic to conform to different needs. It will require some mechanism to enforce policies and penalize flows that inadvertently or actively choose the wrong service level.

Recommendation M-9: ASCR should fund research into mechanisms that allow networks to simultaneously carry multiple types of traffic flows without having flows from one type impact the performance of other types.

6. LONG TERM

The goal of this breakout session was to look at the relevant long-term research challenges (10–12 year time horizon). The organizers proposed two challenges and a list of questions that each breakout group needed to address. The first challenge was to increase the bulk data transfer rate by another 3 orders of magnitude and have 1 exabyte-per-hour transfers (2.4 Pbps) become a common activity. The second challenge was to consider how the need for parallelism will impact both the design and operation of network protocols. This issue reflects the common perception that transmission speeds on individual lambdas will stabilize in Tbps range, and higher throughputs will come from using multiple lambdas in parallel.

There was general agreement that Exabyte-per-hour transfer rates will be difficult to achieve in the 2025 time frame. This is not because demand will decrease but because of the currently perceived limits of optical fiber capacity. Although optical fibers can carry a wide range of optical channels, not all of these channels are concurrently usable. Issues of dispersion and absorption make today’s deployed fibers usable over a narrow range of frequencies, often separated by large guard bands that reduce the aggregate fiber capacity. Redeploying physical fibers with different basic characteristics would be a major undertaking, and that topic is beyond the scope of this workshop.

Fiber deployed in the ground today can support 10 to 50 Tbps [1, 2] using existing optical windows and ITU grid spacing. Using current technologies, achieving 2.4 Pbps will require 44 fiber pairs between source and destination nodes. If each source or destination node deploys 400 Gbps NICs, 125 nodes will be required per fiber for a total of 5,500 nodes at each end to source/sink this 2.4 Pbps data flow. The number of nodes, NICs, and fibers needed to achieve this goal is prohibitive.

Attendees identified several potential hardware and software approaches that may help to simplify the task of developing networks that operate at these speeds. Hardware approaches include deploying multi-core fibers, expanding optical frequencies into the UV and x-ray ranges, and buffering and switching of optical signals. Software approaches included advanced modulation schemes and better methods for managing data parallelism.

Although the technology path is not clear, it is understood that multiple DOE science communities are evolving along a path in which exabytes and zettabytes of data will be generated by their experimental facilities. Making this data useful and available to globally distributed science communities will require novel and imaginative solutions.

The workshop participants were tasked with addressing three general questions and three breakout session–specific questions. They were also asked to rank order the issues derived from the three specific questions according to their priority in a research program needed to answer these questions.

APPLICATION DRIVERS

The breakout groups were asked to comment on the application drivers that will justify the building of the proposed infrastructure. The general consensus was that there would be much more interactive data as scientists explore the results of large scale simulations and large data analytics jobs. Scientists will remotely access supercomputers and advanced analysis systems to interactively explore the data, focusing in on regions of interest and looking for ways to extract knowledge out of the data. These real-time streams will place a variable load on the network that must be anticipated and accommodated. Science experiments may be driven by remote computations that analyze the data and provide information needed for next cycle of experiments.

WHAT ARE THE BASIC SERVICES/FUNCTIONS YOU NEED/EXPECT FROM THE NETWORK?

Workshop participants agreed the network should move data from one location to another as quickly and efficiently as possible. There was also agreement that some type of optical network will be required to support Exabyte-per-hour data flows. It was also generally agreed that existing infrastructure will not easily support these data rates without significant changes. It was unclear what role new technologies deployed directly at the chip level might play in this new network environment.

HOW DO YOU EXPECT THE NETWORK TO BE USED, AND HOW DOES THAT CHANGE OR STAY THE SAME?

The participants did not expect to see large changes in how networks are used. What is expected is that a growing number of science communities will place a larger demand on the networks, both with more experiments streaming data and each experiment generating larger amounts of data. Examples include

- experimental data from remote sensors, including experimental facilities and instruments such as the Large Hadron Collider;

- simulation data from exascale computing on supercomputers; and
- a large number of aggregated small flows from billions and even trillions of nanosensors.

Attendees planned for a wide mix of traffic types and production-oriented network infrastructures managed by a professional network operations community.

Attendees noted that some of these problems appear in the medium term (2021 time frame). However the data rates, volumes, and varieties will continue to grow. This means that scalability needs to be a consideration in the program developed for the medium-term horizon.

Ranking of long-term research needs

Participants were asked to discuss what revolutionary changes would be required to prevent the networks of 2025 from becoming bottlenecks to science. Attendees were then asked what technologies or work must begin now and what can be delayed without negatively impacting the ability to deploy these technologies in the 2025 time frame. The following are the research areas they identified in their order of importance.

1. System management issues associated with massive data transfer parallelism, including signaling and control.
2. Reliability issues for data sets that are much larger than the underlying hardware (network, processor bus and memory, or disk) errored data rate.
3. System architectures, and associated power and cooling, for storage and network interconnect associated with Pbps parallelism.
4. Security models for application data transactions which span multiple locations and which run on very highly parallel infrastructures.
5. Next generation fiber and optoelectronics technology for Pbps on a single fiber [5, 6].

6.1 System Management Issues Associated with Massive Data Transfer Parallelism

Attendees noted that in the medium-term breakout session there was talk about making networks more intelligent. This would be accomplished by adding better control software, inserting storage and/or compute functions directly into the network infrastructures, and virtualizing parts of the networks. In contrast, the long-term breakout session discussions seemed to focus on less intelligent purely optical networks. This is due to the lack of good hardware and software mechanisms that provide basic network functionality, such as packet buffering at the optical layer, as described below on electronic-optical-electronic and optical-electrical-optical (OEO) conversions.

The key finding of these discussions was that parallelism at the physical network layer will become commonplace and that network and transport protocols need to evolve to deal with this reality. The current view that individual data transfer sessions must use a single physical path to prevent packet reordering must give way if applications and workflows are to take complete advantage of the physical infrastructure.

Massive parallelism will force resource constraints to be considered for every transaction. Data location and the availability of communications capacity and computational resources will combine with cost, security, and power considerations in planning exascale jobs. It is impossible to plan the networking components without also planning the storage and processing components.

High degrees of parallelism will result in partial failures, that is, failure of one fiber or one lambda. The management system must be able to provide so-called “hitless” or graceful reconfiguration in the face of failure (i.e., supporting configuration changes without adversely affecting live traffic already inside the network). For data streaming off of instruments, failure should not result in damaged data sets.

The inter-domain nature of DOE science will make the management of parallel flows much more difficult. Sometimes traffic will cross inter-domain boundaries at multiple peering points, potentially creating large variability in delay, jitter, and bandwidth capacity. At other times the cross-domain traffic will use a single peering point, potentially driving it near saturation and causing other traffic to find alternate routes. Dealing with these issues will involve communications between both network operations staffs and network management tools and device controllers.

In addition to deploying physical network devices, network operations staffs will need advanced modeling and simulation tools capable of accurately describing the networks and the behavior of protocols, applications, and workflows.

The performance of massively parallel network transport protocols at Tbps link speeds is poorly characterized or understood. It is not known whether new network mechanisms will be required to support such transfers.

The move to connecting instruments to distributed computation or storage will result in more interactive workflows. This may mean that traffic will change and become more “bursty” as a result. Techniques for transport parallelism may include making the network stack more aware of the structure of the data being transported.

Recommendation L-1: ASCR should invest in research to manage and control massively parallel data flows in a multi-domain network environment.

Recommendation L-2: ASCR should develop management and troubleshooting tools that can make massively parallel network infrastructures operationally reliable.

6.2 Reliability Issues for Data Sets That Are Much Larger Than the Underlying Hardware

Attendees recognized that with exabytes of data not everything will be located at a single geographic location. Transfers of data or subsets of data and analysis jobs will be common activities. Hundreds of transfers from potentially disparate locations must be started in parallel and monitored through completion. When failures occur, alternative resources need to be found and the transfers restarted. With transfers of 10^{18} bytes, the bit error rates of the underlying fiber, computer memory, busses, and storage units cannot be ignored, nor can the impact of error-correcting protocols and encodings

Managing this complexity requires more than simply developing new transport protocols. It requires a coordinated set of activities across multiple layers of the network stack. At the physical layer, fast connection setup and teardown protocols are needed to create optical links that interconnect the ingress and egress points of the administrative domain. An alternate approach would be to develop and refine optical switching capabilities to the point that they could replace current electrical switches. Keeping traffic in the optical domain may be an essential first step in achieving high end-to-end performance. At the network layer, multipath routing algorithms are needed to ensure that no single path becomes a bottleneck. In addition, fast reroute services are needed to deal with both hard and soft failures. At the transport layer, congestion-control algorithms are required to maintain high performance even in the face of failure, loss, and infrastructure changes. Finally, an overarching management function is required to deal with the multisite issue of geographically distributed data and compute resources.

Recommendation L-3: ASCR should develop a wide array of protocols and services to ensure the management of multisite data transfers and interactive flows.

6.3 System Architectures

Based on cost and technology, trade-offs must be made between the size of transmission units (lambdas in today’s systems) and the number of such units. Even with Tbps lambdas, system architectures for the networking equipment and its interface to storage systems will include hundreds of NICs and associated equipment for muxing and de-muxing of signals onto the fiber. An understanding of the architectures and trade-offs for the end-to-end combination of transmission technologies and computer system architectures at 2.4 Pbps does not currently exist [6].

The state of the art in NICs is 100 Gbps. Tbps speeds will exceed the capacity of currently available commercial bus and memory systems. Attendees agreed that 400 Gbps NICs will probably be available in the 2021 time frame, but significant increases in the degree of parallelism and the associated management issues will still exist. Attendees expressed considerable uncertainty about the level of parallelism that the combination of 2025 NIC technology, end-system processor and storage capabilities, and fiber-transport architecture considerations will dictate.

The IEEE PCIe 4 road map calls for 16-lane 512 Gbps buses to be deployed by the 2020 time frame. This will allow support of 400 Gbps NICs, but nothing faster is on the industry horizon. Pushing optics closer to the chip edge allows for all-optical end-to-end connections, but there are no industry road maps that show the use of this technology in LAN or WAN environments.

As noted previously, deployment of architectural innovations such as embedding transient storage (“burst buffers”) in the network may be possible in the next 5–6 years. What is less clear is how these embedded devices will be used in a future optical and highly parallel network. What are the costs for power, cooling, and space in network co-location space? How will these storage devices be managed within the context of

DOE science community needs? Will the OEO conversion costs make these intermediate devices performance bottlenecks or enhancers?

Recommendation L-4: ASCR should explore new architectures that support the rapid growth of parallelism, high performance protocols, and management services.

6.4 Security Models for Application Data Transactions

The attendees recognized that massive parallelism, ultrahigh speed, and physically dispersed data sets may break current security mechanisms. Traditional security mechanisms such as firewalls already represent significant impediments to performance. At 2.4 Pbps and with hundreds of parallel flows, the firewall mechanism no longer works. Mechanisms that authorize circuits dedicated to specific tasks exist but have not been carefully explored in the context of high performance scientific computation.

Recommendation L-5: ASCR should fund research into advanced cyber security mechanisms that protect both data and the infrastructure from harm and exploitation.

6.5 Next-Generation Fiber and Optoelectronics Technology

There was general consensus that network infrastructures will initially evolve to include more parallel links. This will be a gradual or incremental growth strategy for ESnet and other REN network operators. However, as noted in the introduction to this section, scaling to thousands of nodes and dozens of fiber pairs to support Pbps data rates may not be a realistic solution. There was widespread agreement that the networking community in general had not given enough thought to the problems posed by Exabyte-per-hour data transfer rates and potential solutions.

New technologies and ideas need to be explored to increase the capacities of nodes, links, switches, and routers. Research in the following areas holds the potential for novel and innovative technologies and potential solutions to exascale data transfer problems.

- Specialized network hardware
- Advanced optical modulation techniques
- Game-changing physical layer transport (e.g., laser signaling)
- UV or non-optical technologies that can support x-rays

Recommendation L-6: ASCR should investigate these Pbps networking areas in more detail to better articulate the problems and identify potential solutions that can be economically deployed.

7. Discussion/Analysis

The technical areas identified at the workshop are driven by a common set of science application drivers with increasingly complex requirements on the data rates over time. Several classes of the required network capabilities have been identified in previous workshops based on science drivers [1, 3, 4]. In the short term, the applications drivers are bulk data transfers, in particular, the file transfers between DOE sites and with REN peers. Emerging drivers such as computational monitoring and steering scenarios and remote facility control are expected to become more dominant in the medium term. Scenarios involving tightly coupled instruments and remote computations are expected to lead to more complex network capabilities in the medium and long term. In addition to the application drivers and transport requirements, the following technical areas are common to all three terms.

- **Transport Protocols.** High performance, agile transport protocols must be developed in stages to match the individual data rates of the short-, medium-, and long-terms, improving on those of each previous term along the way.

- **Integrated Transport.** Transport solutions must be integrated into the science ecosystem by reducing the impedance mismatches in compositions at all rates.

- **Performance Optimization.** Custom solutions must be developed to identify configurations and parameters for optimal performance at different data rates over time.

- **Operations and Security.** Transport protocols and architectures must be aligned with the security postures of the various sites and provide security attributes of trust and controlled access commensurate with data rates over time.

8. Conclusions

Science drivers will continue to push the boundaries of network functionalities well into the future by requiring higher data rates and advanced capabilities across the science infrastructure of instruments, supercomputers, massive storage, and other subsystems. Certain future science workflows that require collaborations involving geographically distributed systems may be negatively impacted without these networking capabilities. These capabilities will not be simple by-products from industry but will require sustained, focused R&D investments and strategies as identified at this workshop.

9. References

1. ASCR, *Software Defined Networking for Extreme-Scale Science: Data, Compute, and Instrument Facilities*, Report of the DOE ASCR Intelligent Network Infrastructure Workshop, August 5–6, 2014; http://www.orau.gov/ioninfrastructure2014/default.htm.

2. President's Council of Advisors on Science and Technology, *Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology*, Executive Office of the President, 2010; http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-nitrd-report-2010.pdf.

3. J. Ahrens et al., "Data-Intensive Science in the US DOE: Case Studies and Future Challenges," *Computing in Science and Engineering*, Vol. 13, Issue 6, pp. 14–23, 2011; http://ieeexplore.ieee.org.proxy.osti.gov/stamp/stamp.jsp?tp=&arnumber=5999634.

4. *Data and Communications in Basic Energy Sciences: Creating a Pathway for Scientific Discovery*, report of a US Department of Energy workshop linking experimental user facility needs with

advances in data analysis and communications, 2012; http://science.energy.gov/~/media/ascr/pdf/research/scidac/ASCR_BES_Data_Report.pdf.

5. D. J. Richardson, "New Optical Fibres for High-Capacity Optical Communications," *Philosophical Transactions of the Royal Society A*, Vol. 374, Issue 2062, 2016; http://rsta.royalsocietypublishing.org/content/374/2062/20140441.

6. K. Bergman et al., *Scaling Terabit Networks: Breaking Through Capacity Barriers and Lowering Cost with New Architectures and Technologies*, report based on findings from a workshop at Optical Society of America Headquarters, Washington, DC, 19–20 September 2013; http://lightwave.ee.columbia.edu/files/STNFinalReport2014.pdf.

7. US Department of Energy, *Computational Modeling of Big Networks (COMBINE)*, report of a workshop held at the American Geophysical Union, Washington, DC, 11–12 September 2012; https://indico.fnal.gov/materialDisplay.py?materialId=0&confId=5397.

8. *Accelerating Scientific Knowledge Discovery (ASKD)*, working group report, 2013; http://science.energy.gov/~/media/ascr/pdf/program-documents/docs/ASKD_Report_V1_0.pdf.

9. US Department of Energy, *Advanced Networking for Distributed Petascale Science: R&D Challenges and Opportunities*, report on a workshop held in Gaithersburg, Maryland, 8–9 April 2008; http://science.energy.gov/~/media/ascr/pdf/program-documents/docs/Network_research_workshop_report_08.pdf.

10. US Department of Energy, *Scientific Collaborations for Extreme-Scale Science*, report on a workshop held in Gaithersburg Maryland, 6–7 December 2011; https://indico.bnl.gov/materialDisplay.py?materialId=1&confId=403.

11. US Department of Energy, *Terabit Networks for Extreme Scale Science*, report on a workshop held in Rockville, Maryland, 16–17 February, 2011; https://indico.bnl.gov/getFile.py/access?resId=1&materialId=5&confId=319.

12. US Department of Energy, *Data Crosscutting Requirements Review*, report from a workshop held in Germantown, Maryland, 4–5 April 2013; http://science.energy.gov/~/media/ascr/pdf/program-documents/docs/ASCR_DataCrosscutting2_8_28_13.pdf.

13. Energy Sciences Network, ESnet Requirements Review Reports (2002–2015); http://www.es.net/science-engagement/science-requirements-reviews/requirements-review-reports/.

APPENDIX A. LIST OF WHITE PAPERS

1. George Michelogiannakis, Key DOE network/transport challenge

2. M. Veeraraghavan, Position Paper for the DOE Network 2025 Challenges Workshop

3. Yves Roblin, Michael Spata, and Anthony Cuffe, Key DOE network/transport challenge Foster, R. Kettimuthu, D. Katramatos, D. Yu, B. Settlemyer, Q. Liu, S. Sen, Co-Design of Software Defined Network and Exascale Science Flows

4. Rajkumar Kettimuthu, Eun-Sung Jung, Sven Leyffer, Ian Foster, How can peak load in DOE's science network be contained?

5. Inder Monga, Towards network predictability – a missing ingredient

6. Alex Sim and John Wu, Challenges to Support Real-time Applications: a science version of Internet of Things

7. Alex Sim, John Wu, Jinoh Kim, Know Thyself: Monitoring Wide-Area Network for Science Applications

8. Lan Wang, Challenges in Designing a Network Service Model for Efficient Data Sharing

9. Satyajayant Misra and Mai Zheng, Rethinking Networking in a Non-volatile, Heterogenous World

10. Gagan Agrawal, Rajkumar Kettimuthu, Ian Foster, Providing Real-time Data Transfer Functionality Without Bandwidth Reservations

11. Armando Caro, Fluid Multipath Transport for Big Data Transfers

12. Chase Wu and Raj Kettimuthu, Distance-Agnostic, Application- and Resource-Aware Transport for Next-Generation Networks

13. Prasad Calyam and Saptarshi Debroy, "Research-Defined Networks" for Big Data Science Applications

14. Christian Tschudim, Results as a (Network) Currency

15. Satyajayant Misra and Beichuan Zhang, Integrating Storage and Data Transport for Furthering Science

16. Nathan Hanford, Christopher Nitta, Dipak Ghosal, Matthew K. Farrens, Venkatesh Akella, End-System Awareness and Transport-Layer Convergence

17. Phil DeMar, Wenji Wu, Liang Zhang, Simplifying Data Management Middleware Through Content-Based Network Services

18. Wenji Wu, Phil DeMar, Liang Zhang, The "Fast-network, slow-host" challenge

19. Yan Luo, Measurement Challenges and Opportunities for Future Science Networks

These white papers are available from the workshop website http://www.orau.gov/networkresearch2016.

APPENDIX B. PANEL AND BREAKOUT SESSION QUESTIONS

QUESTIONS FOR ALL PANELISTS:

From the list of breakout session questions, what should be removed from the list?
From the list of breakout session questions, what should be added to the list?
What are the cross-cutting issues that keep you up at night?
What does the future network look like to your science community?
What changes are necessary and what stays the same?
What must remain constant and what would be nice to leave alone?
What are the foundational research issues you want to see addressed?

OVERARCHING QUESTIONS FOR ALL THREE BREAKOUT SESSIONS

What is the mix of traffic and how will it change over time?
What are the application drivers for your science?
What are the basic services/functions you need/expect from the network?
How do you expect the network to be used, and how does that change or stay the same?

BREAKOUT SESSION 1: SHORT, 1-3 YEARS (TERABYTE/HOUR BULK TRANSFER RATES)

What is the most pressing problem for scientists?
What is the most pressing problem for network operators?
Rank order and discuss these issues

- status of analysis tools
- status of simulations and models
- status of management tools
- multi-domain troubleshooting
- protocol performance and implementations
- traffic growth and projections
- testing and deploying new tools and services

BREAKOUT SESSION 2: MEDIUM, 4-6 YEARS (PETABYTE/HOUR TRANSFER RATES)

Assume that the network infrastructure consists of highly parallel and redundant physical links.
Assume that new protocols can be deployed over the global Research and Education Network infrastructure.
Rank order and discuss the following problems

- Effectively support for bulk data and more interactive flows
- Routing over this infrastructure
- Active queue management in routers/switches
- monitoring and managing the network
- debugging and troubleshooting

predictable transport performance
When must work begin on solving each specific problem (now, or delay 3 years)

BREAKOUT SESSION 3: LONG, 10-12 YEARS (EXABYTE/HOUR TRANSFER RATES)

Assume that the network infrastructure consists of massively parallel links with multiple DWDM channels per link.

What revolutionary changes are required (e.g.; parallel bit streams)?
What work must start now to be ready in time?
What work can be delayed 3-6 years?

APPENDIX C. WORKSHOP AGENDA

MONDAY, FEBRUARY 1, 2016

7:30am - 8:30am: Continental Breakfast and Registration
8:30am - 9:00am: Welcome and Introduction, Richard Carlson, U.S. Department of Energy
9:00am - 10:00am: Panel Presentations: Network Frontiers for DOE
A View from the Cosmology Frontier, Don Petravick
Next Generation Networks and Systems, Harvey Newman
Entering A New Era, Ian Foster
Challenges for DOE Networking in 2025, Phil DeMar
10:00am - 10:30am: Break
10:30am - 11:45am: Panel Q&A Session
11:45am - 12 Noon: Breakout Session Change and Process
12 Noon - 1:00pm: Lunch
1:00pm - 2:30 pm: Breakout Session 1: Discussions - Short Term [Terabyte/hour data Transfer]
2:30pm - 3:00pm: Breakout Session 1: Report Out
3:00pm - 3:30pm: Break
3:30pm - 5:00pm Breakout Session 2: Discussion - Medium Term [Petabyte/hour data transfer]

TUESDAY, FEBRUARY 2, 2016

7:30am - 8:30am: Continental Breakfast
8:30am - 9:00am: Breakout Session 2: Report Out
9:00am - 10:30am: Breakout Session 3: Discussion - Long Term [exabyte/h data transfer]
10:30am - 11:00am: Break
11:00am - 11:30am: Breakout Session 3: Report Out
11:30am - 12 Noon: Conclusions and Next Steps
12 Noon - 1:00pm: Lunch
1:00pm - 4:30 pm: Report Writing

APPENDIX D. LIST OF ATTENDEES

Gagan Agrawal, Ohio State University

Prasad Calyam, University of Missouri-Columbia

Rich Carlson, Department of Energy

Armando Caro, BBN Technologies

Anthony Cuffe, Jefferson Science Associates

Phil Demar, Fermi National Laboratory

Ian Foster, Argonne National Laboratory

Chin Guok, ESnet/Lawrence Berkeley National Laboratory

Eun-Sung Jung, Argonne National Laboratory

Nathan Hanfordnate, University of California Davis

Raj Kettimuthu, Argonne National Laboratory

Sven Leyffer, Argonne National Laboratory

Yan Luo, University of Massachusetts Lowell

Bryan Lyles, Oak Ridge National Laboratory

George Michelogiannakis, Lawrence Berkeley National Laboratory

Satyajayant Misra, New Mexico State University

Inder Monga, ESnet/Lawrence Berkeley National Laboratory

Biswanath Mukherjee, University of California Davis

Harvey Newman, Caltech

Christos Papadopoulos, Colorado State University

Don Petravick, University of Illinois at Urbana-Champaign

Nageswara Rao, Oak Ridge National Laboratory

Yves Roblin, Jefferson Science Associates

Satyabrata Sen, Oak Ridge National Laboratory

Bradley Settlemyer, Los Alamos National Laboratory

Alex Sim, Lawrence Berkeley National Laboratory

James Sterbenz, University of Kansas

Brian Tierney, ESnet

Joe Touch, Information Sciences Institute

Don Towsley, University of Massachusetts, Amherst

Christian Tschudin, University of Basel

Malathi Veeraraghavan, University of Virginia

Lan Wang, University of Memphis

Chase Wu, New Jersey Institute of Technology

K. John Wu, Lawrence Berkeley National Laboratory

Wenji Wu, Fermi National Laboratory

Dantong Yu, Brookhaven National Laboratory

Beichuan Zhang, University of Arizona

Liang Zhang, Fermi National Laboratory

Lixia Zhang, University of California at Los Angeles

Mai Zheng, New Mexico State University