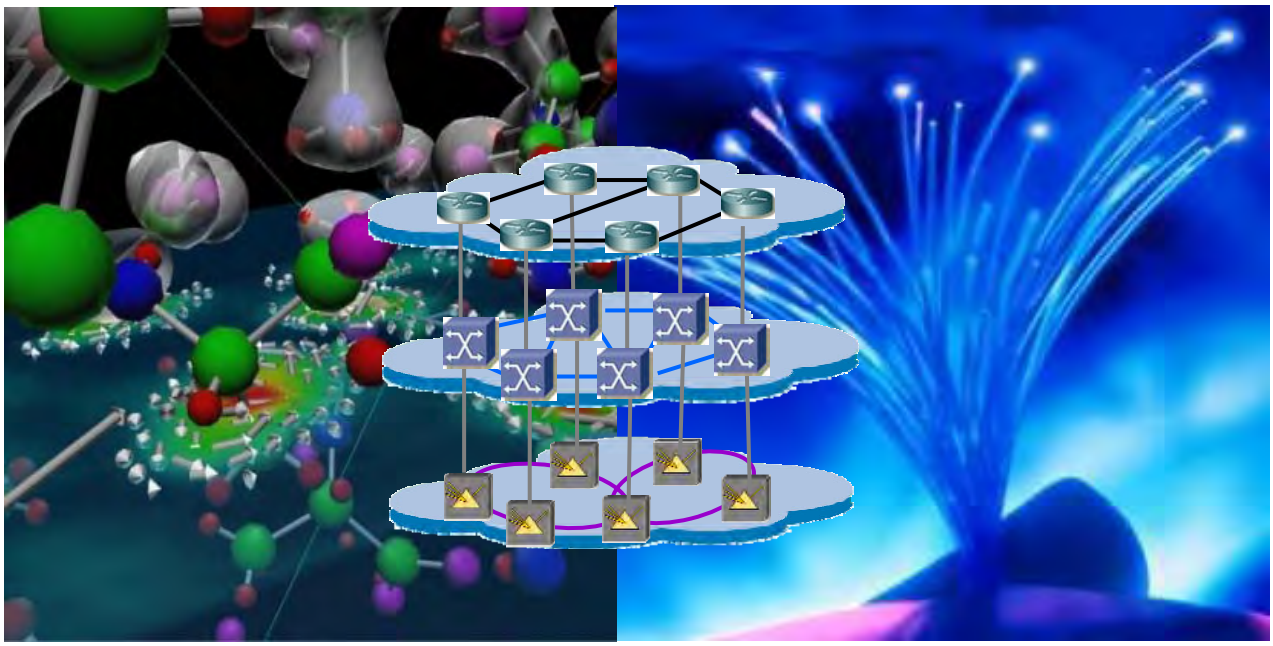
US Department of Energy
Office of Science

Workshop Report on
Advanced Networking for Distributed Petascale Science:
R&D Challenges and Opportunities

April 8–9, 2008



Workshop on:

Advanced Networking for Distributed Petascale Science:

R&D Challenges and Opportunities

April 8–9, 2008
Hilton Hotel
Gaithersburg, Maryland

General Chair:
Donald Petravick, Fermi National Accelerator Laboratory

Co-Chairs:
Nasir Ghani, University of New Mexico
Joe Mambretti, Northwestern University
Donald Petravick, Fermi National Accelerator Laboratory
Bill Wing, Nagi Rao, Oak Ridge National Laboratory
Dantong Yu, Brookhaven National Laboratory

Contents

Executive Summary

The first DOE Workshop on Advanced Networking for Distributed Petascale Science, R&D Challenges and Opportunities was held at the Hilton Hotel in Washington D.C., North/Gaithersburg, on April 8–9, 2008. The workshop brought together leading network researchers in optical transport, middleware, and high-performance protocols. Their charge was to develop a high-level roadmap for the network research and development (R&D) that will be required to support DOE's distributed Petascale science over the next decade.

The workshop was organized in the context of an impending data tsunami that will be produced by DOE's distributed Petascale computing at Leadership Computing Facilities (LCF), other agency facilities, and by large-scale DOE-supported science experiments such as the Large Hadron Collider (LHC), and the International Thermonuclear Experimental Reactor (ITER). The urgency to develop effective networking capabilities to manage and distribute these massive data sets has been emphasized in a series of workshop reports by DOE science programs and ESnet. A related context is the recognition that the performance of today's networks and middleware will have difficulty meeting the emerging needs of the distributed Petascale science that is expected to dominate the landscape of large scientific research in the next decade. An equally important trend is that middleware capabilities are becoming support-intensive, especially at the high end. Even the largest and best-funded research projects may have difficulty in sustaining these middleware support burdens. In addition, it is becoming clear that terabits/sec networks will be essential components of distributed Petascale science. These terabit networks will likely be based on advanced optical transport infrastructure, ultra high-speed protocols, and dynamically reconfigurable services. Drafting a concrete and robust roadmap to develop, test, and implement these advanced capabilities will be critical to DOE's science mission in the next decade. This situation is driven by the fact that key requirements for DOE Petascale science will emerge years before they are recognized by other communities. Many of these requirements will remain unique to DOE science programs for the foreseeable future. Such requirements can be met only through DOE network R&D directly targeted at addressing its key needs. DOE cannot compromise its mission-critical programs by assuming other communities will address its needs in a timely manner. It is clear that no other R&D community is positioned to address these needs within the required time frame. Experience has demonstrated that DOE can successfully provide for its needs through its own proactive network R&D efforts.

Workshop participants assessed current requirements, and made recommendations in the following four key areas where DOE high-performance network R&D programs are required to meet current, emerging, and future challenges:

First— As emerging distributed Petascale science increasingly requires networking capabilities beyond what is available today, and as standards for terabits networks are being formalized, DOE must be an active R&D participant by providing leadership in shaping the important technologies that are vital to its science mission.

Second— With 100G optical transport at hand and higher levels of capacity on the horizon, static provisioning of network resources and maintenance-intensive middleware systems must

give way to more agile networking capabilities that can meet the requirements of DOE's Petascale distributed science.

Third— To provide the Petabyte-scale data distribution expected in the near future, and the Exabyte-scale that will follow, a new family of ultra high-performance, secure, and compose-able protocols will need to be developed. These protocols will need to be self-adaptive to make efficient use multi-scale, dynamic, terabit networks.

Fourth—A particularly important resource for these R&D activities is a DOE experimental network research testbed to support and foster investigations of new innovative technologies that are unique to DOE Petascale science needs. Such a facility would serve as the primary resource for prototyping, testing, and deploying important novel methods and technologies, as well as transitioning them into production environments.

Networking R&D Requirements for DOE Petascale Science

DOE has a broad spectrum of network requirements ranging from routine production IP services to very specialized high-throughput capabilities that are required to support emerging distributed Petascale and Exascale science applications. A number of workshops have been and are being organized by various ESnet and DOE program offices to identify these requirements and to quantify them for ESnet. Reports from these workshops describe the spectrum of requirements in depth (listed in the appendix). Today, it is clear that commercially available networking services, even advanced services, fall short of meeting these requirements. While a considerable portion of DOE's networking requirements are being met by the combination of the current DOE efforts, Petascale and Exascale applications will require significant and unique *combinations of capacity and capabilities* beyond the evolutionary paths of these efforts, as well as those in other federal agencies and industry.

Petascale and Exascale applications will be characterized by unprecedented computational, communications, and experimental capabilities. These capabilities will be utilized by communities of researchers who will be geographically dispersed across national laboratories, universities and international research institutions. Supercomputers at DOE Leadership Computing Facilities (LCF's) are rapidly approaching Petaflop performance, and are expected to reach Exascale levels within a decade. Within the next few years DOE will field three Petascale computing facilities at Oak Ridge National Laboratory (ORNL), Argonne National Laboratory (ANL), and National Energy Research Scientific Computing Center (NERSC). These facilities will be interconnected by ESnet as a fundamental part of their mission to advance science. Like DOE's experimental facilities, such as Large Hadron Collider (LHC), International Thermonuclear Experimental Reactor (ITER), and Spallation Neutron Source (SNS), they are expected to generate datasets in the Petascale and Exascale range. However, simply creating data sets does not produce science. Scientific discovery is the result of analyzing, exploring, comparing, contrasting, and replicating results. For both experimental and computational Petascale facilities, this means that the data produced must be shared. Thus, advancing science in fields of interest to DOE including high energy and nuclear physics, combustion computations, astrophysics, climate modeling, nanoscale materials science, and genomics will require end-to-end network capabilities that match the Petascale and Exascale requirements. In addition, networking Petascale computing facilities offers cost savings: 1) it allows expensive resources such as storage systems and backup systems to be shared, 2) it promotes an efficient utilization of computing resources, and 3) it facilitates remote access to Petascale computing resources. Simply stated, the most important reason for interconnecting these facilities is the advancement of science.

In terms of end-to-end networking, Petascale and Exascale applications will require both *capacities* and *capabilities* unprecedented in currently envisaged network infrastructures and associated support technologies*:*

(a) Network Capacity: To support Petabyte and Exabyte-scale data distribution and other science applications, Terabits-capable networks will be needed. This implies protocols and services that can efficiently operate at ultra high speeds. DOE's supercomputing, storage, visualization, and experimental facilities will be required to have sufficient capacity to

handle Petabyte or larger data sets. The one hundred Gbps/lambda circuit technologies, currently emerging from development laboratories, barely meet this requirement. It still takes 24 hours to transfer a Petabyte of data at such rates. Nonetheless, this serves as a starting point in planning for Petascale- and Exascale science, since current experience shows that conducting large-scale data transfers using soft aggregation of circuits or data streams is difficult and support-intensive. While planning 100-1000 Gbps capacity requirements for the DOE core, it is critical that the needed network capacity be provisioned end-to-end: including metro, campus, edge and host. Technologies capable of providing connections with such capacities are necessarily disruptive and therefore require theoretical and experimental research. Hence, capacity solutions based on such technologies must be fostered and developed through highly focused efforts on experimental networks and systems, since the advanced technologies may prove to be too disruptive for production network infrastructures.

(b) End-to-End Capabilities*:* In addition to the transport network's path capacities, an extremely important and vital part of the solution consists of optimized systems of software and edge/host technologies that will enable users to achieve throughputs commensurate with the provisioned capacities. If, as expected, the post-Petascale generation of supercomputers is based on optical interconnects at the chip, module, and system level, it is extremely likely that these optical interconnects will form the basis of the I/O subsystems. It will be critical for the host system, protocol stack, and network developers to work together to make sure the required I/O capabilities can be realized in both the local and the wide area. This may require development of new or novel I/O subsystems (a future follow-on to InfiniBand for example) and the exploitation of low- or zero-latency switching in the wide area. As was pointed out in the ASCAC network Subcommittee report, these "systems of systems" considerations are falling into a void between the funded efforts in network infrastructure being built by ESnet, and the software development funded by SCIDAC. In particular, it is important to avoid cases such as the now-infamous Cray X1E supercomputer that only achieved 5Mbps TCP throughput over a dedicated 1Gbps connection, even with an optimized stack. We want to emphasize that the required solutions are likely beyond the purview of the wide-area network infrastructures such as ESnet and high-level middleware such as GridFTP. Indeed, they will probably require the deployment of novel edge solutions such as wide-area InfiniBand (IB) devices and hosts with high performance Network Interface Cards (NIC's) or Host Channel Adapters (HCA's). Furthermore, all these components must be co-scheduled along with the necessary network services to ensure effective application execution. To achieve such capabilities, systems and application software must be developed to match the impedance of these edges and host configurations with the wide-area connections, so that optimal throughputs may be achieved. We want to emphasize that these capabilities must still be developed through focused efforts. They are essential even if ESnet provides the needed wide-area network capacities, since ESnet only reaches the edges of participant sites. The development of these capabilities is very complex, and often requires non-traditional solutions to achieve the needed quantum leaps in the capabilities. Examples might include special interconnects to supercomputers, and direct wide-area IB interfaces to storage systems. Such developments would require specific combinations of specialized technologies in this area, and would be extremely unlikely to become available as incidental byproducts of other projects.

The challenges of developing the needed capacities and capabilities for Petascale and Exascale applications are multi-fold, spanning the wide-area connections, edge and hosts systems, systems and application software, and middleware tools. Several DOE efforts have so far contributed to a number of developments that constitute important steps towards meeting Petascale- and Exascale demands. We cite a few such instances to illustrate the potential nature of the non-conventional approaches that might be needed.

(a) *High-Speed On-demand Dedicated Networks:* The DOE Ultra Science Net (USN) demonstrated that a wide-area switched 10Gbps network can be built with production quality performance. This led directly to the switched infrastructures now used in LHCnet and Science Data Network. Furthermore, by analyzing the combination of switched infrastructure of USN and dedicated channels on the routed infrastructure of ESnet, it was established that they can provide comparable performance. This finding led to an extra degree of flexibility in designing high-performance networks, namely to use an existing routed infrastructure to a large extent and capitalize on a cheaper switched solution in newer areas. In addition, it demonstrated a method by which dedicated channels on Science Data Net can be extended to university collaborators over their existing networks.

(b) *Novel Wide-Area Transport Protocols*: Achieving multiple Gbps throughput over wide-area connections using conventional TCP traditionally require significant effort. USN experiments over wide-areas demonstrated that IB and Fiber Channel (FC) based solutions not only provide peak performance over thousands of miles but also offer drop-in solutions with very little per-connection optimization required. New approaches beyond traditional TCP and UDP such as these are among the most promising data transport solutions between supercomputers and high-performance storage systems.

The solutions needed for Petascale- and Exascale applications will require novel and non-conventional approaches along the lines described above. They may well be composed of newer and existing technologies, which must be combined and tested in non-production environments. In particular, they may transcend the conventional boundaries of network infrastructures, computing and storage site facilities, and middleware, requiring solutions that have not been tried or possibly those that appear to be natural extensions to production environments.

Findings and Recommendations of the Break-Out Groups

The following sections describe the individual networking R&D areas addressed by the individual working groups. The sections support the overall conclusions of the workshop and are designed to be read by subject-matter experts.

Group 1: Transport Protocols and Data Distribution Services

The throughput rates achievable over today's best-effort IP networks, regardless of the provisioned network capacity, is limited by the performance of TCP/IP protocol stacks on host systems. The TCP protocol, which provides reliable data communications, is the basis of the default transport protocol for File Transfer Protocol (FTP) services. Developed in the late 1970's for low-speeds networks, FTP and its variants are still widely used for data movement across the Internet. Over the years, TCP has been extended, modified, and parallelized to improve its performance. However, the fundamental limitations that make it sub-optimal for large-scale data movement over very high-speed networks remain in place. The search for better performing transport protocols has resulted in a plethora of TCP and UDP variants. Enhanced TCP variants include FastTCP, High-Speed TCP, Scalable TCP, Hamilton TCP, TCP –Vegas, Stream Control Transmission Control Protocol (SCTP), and eXplicit Congestion Control Protocol (XCP). UDP variants include Reliable Blast UDP (RBUDP), UDT, Hurricane, TSUNAMI, and Simple Available Bandwidth Utilization Library (SABUL). Despite the proliferation of these variants, existing transport protocols still have difficulty harnessing the abundant optical bandwidth made possible by the Dense Wave Division Multiplexing (DWDM) technology. Even the best tuned TCP/UDP protocols are rarely able to operate at 90-95% of the line rate of 10 Gbps a single DWDM Lambda. TCP/UDP and its variants also provide the core functions of GridFTP widely used in the grid community for large data transfers. Both FTP and GridFTP are based on TCP/UDP and as a consequence, the performance of these file transfer applications is bounded by the limitations of TCP/UDP. We find that transport protocols along with host software stacks and high-speed NICs will need to be suitably composed and optimized to address the Petascale- and Exascale data distribution needed in DOE in the next decade.

Finding 1.1

Petascale- and Exascale science applications require data sets to be movable over networks or be accessible across them. Such capability requires Terabit networks to deal with Petabyte datasets at bandwidths of 100Gbits/sec/lambda and higher.

Recommendation

DOE must provide leadership in the R&D of ultra high-speed transport protocols and data transfer services to address its data distribution needs. This next generation of terabits/sec transport protocols will need to offer performance capabilities far beyond the existing TCP and UDP stacks, if DOE is going to meet its needs for moving Petabyte to Exabyte data sets over very long distances. DOE also needs to participate in field trials being conducted by commercial vendors and other federal agencies, such as National Science Foundation (NSF) and Department of Defense (DOD), or else hold its own, since the implications of this capability on data center requirements and user expectation will be profound.

Discussion

Science is advanced by the systematic process of replicating results, challenging conclusions, and iterating to consensus among researchers. In the case of the "third leg" of science (computational modeling), this requires the systematic study of large data sets or results by many groups using different tools and approaching the problem with different mindsets, *or* it requires parallel efforts at modeling the same problem (as in climate studies) by groups, which then compare their results. In either case, it requires that data sets be movable over networks or be accessible through them. As described in the previous section, such networks require minimum bandwidths of 100Gbits/sec/lambda. Such networks are currently being demonstrated in limited field trials by at least two vendors, and other vendors are expected to follow shortly as standards emerge to support the wide deployment of these technologies (IEEE 802.3ba).

Finding 1.2

Performance of current wide-area networks, including ESnet and Internet2, is being constrained by routine use of "available" techniques developed to deal with problems that are not germane to envisioned Petascale networks.

Recommendation

DOE should be prepared to take advantage of opportunities to leverage the low error rates of current optical networking technologies to develop wide-area data transport and file systems based on low- or zero-latency switches. File systems distributed across DOE Petascale systems should always maintain global consistency checks on stored data.

Discussion

Current optical networking technologies are able to operate essentially error free. Bit error rates of 10^{-15} or 10^{-16} are routine. There is reason to believe error rates are lower than this; however such low rates are difficult to measure accurately. Nonetheless, routers, switches, and many other network elements still use store and forward technology (accumulate a frame, compute a checksum, compare with the transmitted checksum, and forward the frame). This can create delays that are significantly greater than the propagation delay across the media. These delays impose significant limitations on (for example) our ability to implement wide-area file storage systems. Jumbo frames, while improving end-host performance, exacerbate this delay problem.

Errors can still be introduced within the host systems (disk errors, memory errors), so global checksums still need to be carried along with data; nonetheless, properly configured networks can be essentially ignored as sources of errors.

Finding 1.3

The ESnet architecture provides both its classic layer-3 routed network service and a circuit-based layer-2 network service (Science Data Net) for large-scale data movement. With such alternate services available, and the context of findings 1.1 and 1.2, DOE's high-performance

networking requirements can no longer be best served by simply building data transfer services such GridFTP services only on TCP, or parallel TCP streams.

Recommendation

DOE should consider funding the development of compose-able transport protocols that are able to leverage knowledge of the connection to select (possibly in dynamic fashion) and optimize a conditional transport protocol, based on such information as bandwidth, latency, jitter, probability of packet reordering, and whether the connection is shared or dedicated. Hybrid transport methods systems, based on optimal choices or combinations among these technologies, should be built. Based on them, new scalable data transfer services should be developed that will offer advanced functionalities and performance several orders of magnitude greater than current GridFTP.

Discussion

With both classic routed IP service and circuit-based network services available, applications can benefit from selecting among various available transmission protocols or technologies. In certain cases, exclusive access to the entire connection bandwidth could obviate the need for complex congestion control mechanisms. Multi-stream or multi-system TCP (e.g., Grid FTP, wherein a suitable number of flows must be selected and tuned to achieve optimal throughput) may impose a high maintenance and management load on sites with small or limited staff. Alternative transmission protocols, such as NACK-based UDP, that are able to make more efficient use of dedicated channels might provide a simpler, more efficient approach to data transport. Alternative transmission technologies, such as Infiniband or fiber channel, that are able to serve as building blocks for both supercomputers and wide-area networks, are projected to be important components of the next generation networks for petascale science in DOE. The ability of future networks to adapt to the needs of different DOE programs, as well as future technologies arising from inter-agency-roadmap cooperative development, will require the development of data services and solutions that adapt to future connection provisioning services, complex end-hosts, and edge devices. Furthermore, these services must optimize the compose-able protocols to account for the potential high complexity of the connections. When datasets need to be exchanged between a supercomputing center and remote data archival system, appropriate network connections must be setup, either on-demand or through advanced reservations. These connections may range from dedicated wide-area Infiniband connections, to end-to-end layer-2 circuit connections, or just default routed IP network service. In the first case, an appropriate data transport protocol such as IB-RDMA must be invoked to achieve data transport, whereas a suitable UDP-based transport might be chosen in the second case, and a suitable choice of dynamically loaded TCP transport in the third case. The variability and complexity of end systems bring an additional level of complexity to optimizing the data transport solutions. End systems that invoke and receive data transfers may range from supercomputers to computational clusters to hosts of varying capabilities, and even to experimental facilities. Coupled with various choices for data connections, including concurrent connections such as a dedicated Infiniband connection supported by a shared IP connection, the end system configurations lead to complex data paths. The overall data transport solution must be made

with appropriate choices for a suite of transport methods to match the end-to-end data connection.

Group 2: E2E Federated Network Measurement

The networking technologies needed for Petascale and Exascale science are necessarily leading-edge with first-time, complex deployments at least in some parts. These capabilities must not place undue burdens on application scientists by requiring them to diagnose network problems and tune the performance. It is essential that the network environments, including core and edge connections, be suitably instrumented and monitored to achieve and maintain peak network performance with minimal demands on network personnel and application scientists.

Finding 2.1

Scientists spend considerable time seeking explanations for poor network performance. This is in part due to a lack of secure and scalable tools to perform end-to-end network performance prediction, fault diagnosis, and network management over multiple domain networks.

Recommendation

Distributed monitoring must span multiple autonomous networks and implement diverse network technologies to address scaling, topology, discovery, and other issues. Distributed monitoring systems must provide information about components at all levels (backbone, edges, end points, layers, dedicated and shared circuits, etc.) on both production and testbed networks to meet the federated monitoring needs. Each autonomous system might adopt its own monitoring technology, and the federated monitoring system must ensure the interoperability among the separate monitoring frameworks. DOE needs to support participation in working groups in organizations such as DICE and the Open Grid Forum to define interoperable protocols.

Discussion

A distributed monitoring framework must scale to large volumes of monitoring data in terms of storage and data movement. Furthermore, discovery, topology, and federated trust support of such a framework must be designed to support not only the monitoring framework but also parallel control plane and session-application frameworks. Well-defined intra-network-monitoring-framework and component protocols as well as inter-system protocols need to be included.

Finding 2.2

Lack of ability to perform timely network troubleshooting across multiple hybrid network domains will lead to under performance of the needed terabit networks. Diagnostic and inference tools must be developed to correlate, extract, and display performance results from distributed monitoring data from multiple hybrid network domains for fault identification and recovery.

Recommendation

In current networks, the ability to diagnose problems and detect anomalous events on these new hybrid networks is just emerging. Further study is needed in this area. Monitoring data from a diverse collection of networks must also be made available to enable researchers to gather new data and correlate multiple metrics and events from multiple sources. Newly developed intrusive, active monitoring tools must be fully tested and simulated in a network testbed before being deployed into the production environment. Research is needed to support new methods to analyze such data and to bridge the gap between testbed networks and production network environments for network monitoring.

Discussion

A common mistake in federated environments is to expect that everything works all the time. In reality, autonomous systems can join or withdraw from the federation at any time. The monitoring system must be able to handle the possibility of incomplete information and inference techniques need to work in the absence of complete information. Data preprocessing techniques can be used for data reconstruction, and cleaning. Event translation between application, middleware, and network events is significantly more complicated given the increasing use of hybrid networks. The monitoring system also needs to provide high-level user-friendly reports. Such reports must contain information such as network availability, reliability, utilization, and reachability.

A corollary of the increasing complexity is that the monitoring data generated by the large number of applications using high-speed hybrid network keeps increasing. Large numbers of flows will lead to less time for data inspection, which raises big challenges for traffic monitoring, accounting, and intrusion detection. Therefore, data processing, data mining techniques, and associated federated network monitoring systems need to scale to large volumes of data, and a corresponding complexity of data event structure.

Finding 2.3

Scientific applications and their workflow are not currently "network aware." New simplified middleware is needed with the ability to react appropriately to changes in network behavior.

Recommendation

Research is needed to integrate smart performance monitoring and fault diagnosis into distributed high-end science application. This must be preceded by development of robust techniques for classifying network monitoring and fault diagnosis events to determine which ones should interrupt the normal execution of applications. Applications should be able take advantage of the new network monitoring capabilities to improve their performance. The collaboration between application development groups and monitoring groups should be formed to analyze the changing requirements over federated computing environment, and provide access to appropriate network events and measurements to improve the efficiency of applications. There is a need to design and deploy new applications using appropriate programming interfaces. This will allow comparison of the difference between the "network aware" applications, and the stand-alone applications, documentation of the improvement and evaluation of the impact of such techniques.

Discussion

Improving the resource utilization and efficiency of distributed applications is becoming increasingly difficult due to the lack of federated monitoring capability. Appropriate federated monitoring capability and interactions with applications should be developed and supported to bridge the gap between high layer application and control layer and the network monitoring at the lower layer. Application or control plane event triggering should lead to appropriate monitoring behavior. The federated network system provides a new suite of tools and capability to guarantee the performance.

Finding 2.4

Currently there is a lack of consensus on what monitoring data should be made public and why. Thus, there is a need to develop community consensus around best practices for deployment of monitoring resources and best practices for policy decisions regarding the sharing of monitoring data.

Recommendation

There is a need for a working group to draft and continually refine "best practices" guides on publishing monitoring data. The work of this group includes gaining practical experience on a sufficient number of test bed and production networks, since prototypical best practices need to be studied in the wild. We need to express the results of such studies into RFC-like documents that can lead to standards and influence the future implementations.

Discussion

Current distributed monitoring infrastructures are not widely deployed across multiple, diverse (international, national, regional, campus, virtual organization) network domains, and therefore there is little experience with the issues involved in getting network providers and sites to "buy into" such a system. Also, current inter-network Memorandums of Understanding (MoUs) do not cover policy and deployment expectations for network monitoring partnerships. A large obstacle to deployment will be the lack of a widely deployed federated trust infrastructure on which to build distributed monitoring framework partnerships.

Finding 2.5

Current security mechanisms and associated access policies for federated network monitoring data and monitoring devices do not adequately protect data integrity and privacy. Also, current intrusion detection systems cannot adequately secure federated network monitoring data and its associated systems.

Recommendation

Federated network measurement framework must ensure the confidentiality, authenticity, correctness, and integrity of measurement data across multiple network domains. There is a need to improve/redesign (as appropriate) scalable access control mechanisms that

provide multiple levels of security granularity for the monitoring dataset and monitoring devices. We need to provide a flexible access policy system to ensure data privacy, which varies from one network domain to another. A flexible framework needs to be provided to allow individual participating network domains to specify their own access control lists (ACLs) and conform to their security requirements. An automated data publishing mechanism can ensure the open nature in a federated network environment without violating individual network domain security requirements. In addition clear documentation of the measurement details and limitations are needed to enable clear interpretation of results.

The monitoring system must provide necessary distributed data analysis and data mining functionality for various monitoring data applications, such as an intrusion detection system for the federated network environment. DOE should consider funding computer scientists familiar with data mining to develop tools which harvest useful information from data monitoring streams generated at various monitoring sources.

Furthermore the security of the monitoring systems is critical because the monitoring archive itself could be a valuable target for malicious hackers. Good security engineering practices must be included in the software development process.

To ensure better coordination and oversight among several network research areas regarding security, a security and enabling policy working group should become the focal point for coordinating the security and policy framework R&D.

Discussion

A well-known problem with many monitoring systems is that they provide a simple binary access control of either granting or denying clients access to the whole monitoring data. Uses of network measurement systems range from applications, middleware, to network stacks. This necessitates that access policies support finer access control granularities for different customers. The federated network environment has more security vulnerabilities than a stand-alone network due to the loose coupling of participating domains, each of them adopting its own security policy. Under this network environment, the monitoring system itself can be attacked and compromised. The resulting security framework must be compatible with inter-operating instantiations in multiple domains.

To provide security for the monitoring federation, monitoring systems must also provide, with appropriate privacy, intelligent data mining, and event correlation to enable rapid detection of anomalies that suggest attacks or compromises on the system.

Group 3: Multi-Layer Federated Network Provisioning

The emerging environment for large-scale sciences consists of researchers, computing resources, and experimental facilities distributed around the globe. Networks which link these major subcomponents of the global scientific complex involve multiple heterogeneous autonomous

domains. Many technical and policy challenges not encountered in traditional best-effort IP network will have to be resolved before new services such as on-demand dynamic circuits, cyber security, and end-to-end monitoring systems are provisioned over multi-domain networks. Federating and controlling heterogeneous networks developed with different types of technologies in each layer to deliver common end-to-end services with strict performance requirements is very challenging. Technical issues such as inter-domain signaling and control, inter-domain bandwidth reservation and traffic engineering, data plane bandwidth mismatch, and inter-domain cyber security which are absent in best-effort IP network emerge as major technical obstacles. These are important issues that have not been fully explored in theory or through prototyping. These functionalities are significantly beyond those of current networks, and their realization requires focused R&D efforts.

Finding 3.1

Dynamic provisioning of end-to-end network services across multi-domain federated network infrastructure is a significant challenge and has yet to be accomplished in general production environments. This capability is particularly important in Petascale and Exascale science applications requiring expensive resources interconnected over terabit networks, which may have to be co-scheduled.

Recommendations

There is an urgent need for architectures delivering guaranteed network services across federated multi-domain infrastructures. In particular, more work is required on flexible provisioning solutions that integrate "hybrid" circuit and packet-switched networks. These solutions must satisfy the need of individual users yet optimize the utilization of global network resources.

Discussion

Scientific research organizations worldwide are actively deploying a wide range of advanced high-speed networking infrastructures to support large-scale data distribution. For example DOE operates its production ESnet network, which implements guaranteed service delivery using IP/MPLS (Layer 3) packet-switching technology, and the Science Data Network, based on a Layer 2 service. Meanwhile the pan-European scientific community operates the GEANT2 backbone to interconnect 30 national Research and Education (R&E) networks. By and large, each of these networks has developed its own set of control plane provisioning solutions.

However, as scientific datasets scale towards petabyte-scale and beyond, many organizations are migrating to more scalable optical DWDM-based infrastructures. The DOE/Internet2 partnership for ESNet4 and DOE's Ultra Science Net testbed are prime examples. This evolution is introducing new circuit-switching network layers with very different provisioning policies and procedures. Given the expanding scope of global research collaborations, scientific users are demanding data delivery across similarly expanding, dispersed, heterogeneous "circuit-packet" network infrastructures. In turn, this is driving the need for integrated control plane provisioning across multiple network domains. Due to the

federated nature of multi-layer networks, these solutions will represent new challenges for control plane authentication, authorization, accounting (AAA), and security. There are few working solutions or standards that address this issue to a satisfactory level. Given the diverse network technologies (Layers 1-3) and control plane methodologies involved, this is a very challenging problem.

Finding 3.2

Current multi-layer networks are largely based upon "best-effort" IP connectivity models. The current network architectures do not support more capable network service models such as dedicated circuits with deterministic user guarantees. Moreover, existing multi-layer networks are also unable adapt to implicit user needs.

Recommendation

Federated multi-layer control plane solutions must define and support differing service models for widely varying higher-layer data transfer needs, such as short- or long-term bulk transfers and high-priority services. In particular, scientific users should be able to specify both best-effort services as well as more deterministic and guaranteed services with measurable and enforced parameters, including bandwidth, loss rates, latency, jitter, survivability options, recovery times, and other user-specific constraints. In addition, service models must evolve beyond "point-to-point" paradigms and support connection groups in order to build "application-specific" topologies to interconnect dispersed collaborations.

Discussion

Scientific users today are seeking to dynamically interconnect numerous globally dispersed, high-performance computing resources, such as large computing farms, super-computing facilities, storage systems, visualization facilities, remote sensors, and other instruments. Increasingly many of these interconnection scenarios require new services with defined, guaranteed properties such as bandwidth, delay, loss rates, jitter, etc. Furthermore, many scientific applications today are leveraging complex work flow process setups distributed across multiple locations. The High-Energy Physics community, for example, uses globally distributing models for storage and analysis of LHC experiment data. Applications like these will require carefully coordinated network path setups between multiple users in order to integrate resources, and support a sequenced set of tasks. Therefore, underlying network service models will have to extend beyond existing "point-to-point" models, and support broader "application-specific topologies" that can interconnect multiple user sites. Given the multi-layer scope of e-science applications, these services will have to be provisioned across heterogeneous network technology domains. This poses many new, unresolved challenges for service parameter translations/mapping across domain boundaries.

In some cases, end users may not be sufficiently knowledgeable or simply unwilling to directly specify "lower-layer" service parameters. It is desirable to develop "intelligent" services that transparently adapt to users needs. For instance, an intelligent service could dynamically detect and provision a specified amount of bandwidth for a certain user's data transfer, based on the history of previous allocations and actual usage.

Finding 3.3

Most of today's control plane solutions are capable of "on-demand" bandwidth services provisioning across single network domains. Furthermore, some offerings also support provisioning based on advance reservation as well. However, there are no working solutions capable of extending secure, scheduled provisioning across multiple federated network domains.

Recommendation

New control plane solutions are needed to support dynamic scheduling of user reservations across distributed multi-layer federated networks. A key requirement here is the development of new theories that address dynamic optimization of multi-domain network resources for scheduled demands. Traditional circuit-switching theory does not address those challenges. From a broader perspective, the joint scheduling of end-system resources (e.g., storage, CPU) in conjunction with network-layer resources should also be considered.

Discussion

Scientists are using a wide range of shared research facilities today, including large computing farms, super-computing facilities, storage systems, instruments and sensors. It is common practice for multiple project teams to be sharing a resource or facility. Given the time-sensitive nature of many of the experiments and work flows, it is becoming evident that network resource scheduling (dedicated service agility) will be required to coordinate access to these devices. For example, a domain-specific user may require dedicated, high-bandwidth to access a data repository for a particular time period. Alternatively, high-energy physics collaborations may want to pre-schedule and coordinate connection setups to achieve timely transfers with minimal loss. Although some control planes can support scheduled demands (such as ESnet OSCARS and the centralized Ultra Science Net control plane), no such capability exists across multi-layer federated domains. Advance bandwidth reservation capabilities will enable significant performance improvements over standard "on-demand" provisioning, in which a request is either served or rejected immediately. For instance, these mechanisms could offer the flexibility of specifying the starting time and duration of a connection so that an "optimized" route can be selected. Similarly, route selection could be deferred until the corresponding connection is set up. Hence there is an urgent need to develop theoretical and engineering approaches for advance reservation across multi-layer networks. The theoretical underpinnings here are fundamentally different and more complex than those in well-studied packet and circuit-switching networks. In turn, these challenges will require enhancements to existing algorithms and theories, such as queuing and distributed routing, to support real-time and advance reservation capabilities. Specifically, one must investigate how to efficiently conduct scheduling and routing in a distributed manner without excessively loading network elements with future state.

Finding 3.4

There are no well-developed or standardized techniques for discovery and exchange of information on network state between multiple federated network domains operating at different layers.

Recommendation

There is a need to develop new algorithms and information models to condense domain-internal state in order to facilitate traffic engineering across federated domains with differing levels of trust and inter-domain policies. These solutions must be scalable and work across multiple network layers (i.e., vertical-horizontal integration) in order to achieve a proper balance between aggregation accuracy and scalability (i.e., partial visibility, inexact state). These models must also consider the time dimension so that resource scheduling for advance reservation can be supported.

Discussion

Network state discovery is a crucial step in the overall services provisioning framework. This function is typically achieved via either distributed means using routing protocols such as OSPF, or centralized means using a network operational support system. For example the DOE ESnet control plane uses OSPF-TE routing to implement resource discovery. Conversely, the Ultra Science Net relies upon centralized routing to maintain network-wide resource information. However, there are few working solutions for information state exchange across multi-layer federated domains. This is a very challenging problem, given the differing network layer technologies involved. The need for variable and flexible trust levels/policies between domains adds to the complexity of the challenge. Although the IP framework supports routing between autonomous system domains, using variants of the *border gateway protocol* (BGP), these offerings do not suffice for emerging needs. Specifically, BGP protocols are focused on end-point address reachability exchange, and cannot provide the detailed level of link state required for guaranteed services provisioning. Ongoing efforts within the OIF to define an E-NNI routing protocol seek to address this problem, albeit only for optical layers. Moreover, existing routing protocols do not provide features for capturing the time dimension, a requirement for advance reservation scheduling. Hence there is a critical need to develop new information models and routing exchange mechanisms for achieving state discovery across federated domain boundaries. These solutions must coalesce existing inter-domain routing frameworks, and inter-operate with diverse intra-domain resource discovery methodologies (e.g., centralized, distributed routing). An added challenge is to incorporate augmented capabilities for propagating time-usage information.

Finding 3.5

Today's data transfer applications such GridFTP are not seamlessly integrated with network provisioning software (control plane software) and thus cannot effectively compute and configure data transfer paths across multiple federated domains. Scientists often spend considerable time assembling different end-to-end technologies to perform data transfers. This requirement will become increasingly critical, given the growing scale and footprint of global scientific research partnerships.

Recommendation

There is a need to integrate network technologies at different layers so that they can automatically perform the path computation (and scheduling) needed to achieve desired end-to-end throughputs over multi-domain heterogeneous networks. These solutions must scale across multiple layers and leverage hierarchical distributed approaches. In addition, signaling techniques are needed to expand routes and normalize request parameters across multiple network layers. These solutions should follow, as closely possible, open standards-based architectures.

Discussion

The ability to compute and configure data transfer paths across a network is essential for moving beyond legacy best-effort services and achieving genuine guaranteed services support. Within current single network domains, many of these capabilities are already well in place. For example, the ESnet OSCARS framework provides a centralized path computation engine driven by link-state routing databases. Subsequent path setup/takedown procedures are done using RSVP-TE a distributed signaling protocol. Similarly, Ultra Science Net uses centralized path computation along with centralized signaling setup, using TL1-based messaging. USN also provides a path scheduling capability, allowing users to schedule and setup guaranteed data paths. Nevertheless, these path computation and setup solutions are largely homogeneous in nature, designed to support a specific technology layer. The extension of these frameworks across multiple federated domains is not entirely straightforward. It is very difficult and impractical to implement any strict form of centralized path computation across such domains, owing to policy restrictions and scalability limitations. Decentralized or hierarchical setups are more feasible, in which domains rely upon partial information state about the global network to compute paths. Another challenge arises from the differing resource granularities that exist across multiple domains. For example, an underlying SONET/SDH or DWDM circuit connection can be treated as a traffic-engineered link at the IP/MPLS level, introducing an inherent grooming dimension. Finally, the need for advance reservation capabilities across multiple domains further complicates the problem. In all, path computation and setup across federated domains is a very challenging and largely open area. Although some standards are emerging, most notably the IETF *path computation element* (PCE) framework, much work needs to be done to adapt, extend, and deploy solutions for production networks.

Finding 3.6

Service reliability is a crucial issue for scientific users given the sheer scale and stringent transfer requirements of the data being generated and transported. However, the needed failure detection, localization, and recovery mechanisms to facilitate reliable services across federated multi-layer networks are generally lacking today.

Recommendation

Reliable data and control plane recovery mechanisms must be developed to support guaranteed end-to-end services across federated networks. In particular, this

includes schemes for rapid fault detection and localization at layer boundaries as well as robust service recovery strategies.

Discussion

Over the years, scientific researchers have become increasingly sensitive to network service reliability. For example, high-energy physics experiments constantly streaming raw experiment data for remote storage require highly reliable transfers to prevent costly additional data handling. Many remote instrumentation/steering applications have requirements to minimize data loss and prevent possible equipment damage. Various service reliability provisions are already in place. For example, the ESnet provides fast re-route capabilities for connectionless data paths. Underlying fiber trunks can also be protected using a variety of monitoring and recovery schemes. However, these offerings are only suited for a given network domain and rely upon highly specialized, technology–dependent mechanisms for fault detection. As a consequence, data-plane service reliability across multiple traversed domains becomes much more challenging. For example, monitoring and fault notification procedures must be harmonized at domain boundaries. In addition, end-to-end data plane service recovery mechanisms must be devised to handle multiple resource granularity levels. These schemes may follow either localized or "end-to-end" strategies, and can further be based upon pre-fault (protection) or post-fault (restoration) methodologies. It is important to emphasize that the above discussions are focused upon data plane recovery to user service. Additionally, reliable control-plane mechanisms must also be developed to ensure continued federated network provisioning during node failures. The main requirement here is to effectively distribute/replicate control provisioning functions, allowing nodes to reconstitute lost connection and resource state information after failure recovery.

Group 4: High-Performance End System/Middleware

Middleware and distributed systems that bind users and science applications to the underlying network infrastructure are a critical component of the DOE networking and distributed science fabric. They provide secure interfaces for accessing science facilities, computing resources, data archives, and virtual organizations anywhere and anytime. In the past, DOE has relied on commercially available components to provide middleware services. However, as distributed Petascale science facilities supported by terabits/sec networks are emerging, questions have been raised about the viability of commercially available middleware products in the Petascale era. Experiments such as LHC, ITER, SNS, and distributed Petascale – Exascale computing in general are anticipated to generate data sets 100TB-EB in size. These data sets must be manipulated, stored, visualized and compared on a timescale conducive to scientific productivity. The data handling, storage, distribution, collaboration policies, and scientific workflow for each of these experiments will be unique. There is a clear need for development resources to build flexible scientific tools that can be easily adapted to meet the unique needs of each experiment, at scale and capabilities commensurate with the requirements of the collaboration involved, as well as the capabilities of the underlying network infrastructure.

Finding 4.1

Scientists with distributed Petascale requirements are currently being forced to use middleware and distributed system software adapted from its original use over low-speed networks.

Recommendation

A new generation of middleware and distributed system software is needed to support emerging distributed Petascale DOE science. This effort must be properly coordinated across various groups and autonomous domains that are involved in the development of end-to-end solutions for DOE. Middleware component architectures that seamlessly and securely bind scientists to science applications and ultimately to the underlying network infrastructure should be investigated. Coordinating bodies are needed to eliminate systematic problems that arise when attempting to match policies across the DOE complex and, more broadly, across federated systems of interest to the DOE.

Discussion

Effective use of networks requires effective distributed system middleware. We believe that the use of networks is impeded by a large number of simple items. These items may be as trivial as determining remaining quota on file systems. There is no simple canonical and accepted way of presenting an abstract computing system to a collaboration or user.

DOE requires general solutions that are usable in the context of federated systems. The majority of data flows into or out of DOE open science sites are with non-DOE sites. Therefore, middleware components will need to extend across both DOE and non-DOE systems in the general case.

Investigation into increasing the flexibility, ease of use, and reliability of middleware frameworks is needed. Web services technologies provide for structured interfaces, and have appropriate support for federated operations in the wide area. However, per-call and installation overhead remains high. The web services framework may not be appropriate for connecting small bits of logic into services.

Middleware needs to be able to test functionality in order to facilitate early detection of problems. The entire software stack must be instrumented sufficiently to allow triage in reasonable time.

All distributed systems need to be capable of functioning within a federated environment. This means they must run on computers that are provisioned and operated by different organizations. In addition to the problem of technical diversity, distributed systems must work across administrative boundaries that have differing management policies. With proper support and encouragement, many disruptive aspects of policies can be reduced. Grid organizations function in this area. Since policy harmonization can not remove all defects, there is a need for middleware to mitigate the effects of conflicting policy. For example, security controls are seen as hindrance to high-performance data movement and management. In the Petascale era, with its requirements for data streaming and service-oriented networking, the impediments to scientific productivity resulting from non-interoperating security controls will become even more significant.

Distributed operations need to be recognized as an essential element of support at DOE open science sites. Middleware needs to be recognized as providing security for these essential operations, not detracting from security. For this to happen, reasonable models of site security policies need to be constructed and recognized as good practice.

Finding 4.2

Policy and middleware-constrained end systems could supersede the core network as the major bottleneck in the end-to-end performance of distributed applications. These problems will extend into distributed Petascale applications, and will present a major obstacle to Petascale science activities in the next decade.

Recommendation

Resolving performance issues of end systems within the end-to-end loop of distributed Petascale science applications should be a high priority for DOE network research activities. A new generation of Petascale middleware software (host protocols stacks, networked-applications interfaces, operating system support, Web services, etc.) should be developed to deliver performance commensurate with distributed Petascale computing, terabits/sec networking, and the requirements of large-scale national and international collaborations.

Discussion

Further investigation of problems using currently deployed IP stacks is warranted. It is appropriate to conduct stack research in context of improving the usability of specific software implementations that are of interest to DOE. The research may involve liaison with proprietary systems, or contributions to open systems (e.g., Linux stack).

Given the amount of inertia in the deployed IP software base that scientists currently rely on, it is also useful to study alternate transports, such as Infiniband, including understanding useful deployment scenarios.

Finding 4.3

The networking interface seen by applications is becoming richer and more complex. How an application can detect, select, and have its problems diagnosed within such a rich network environment needs to be addressed. The emergence of circuits as a resource that can be scheduled for applications increases the desirability for scheduled data transfers.

Recommendation

Integrated smart network interfaces and advance network services should be developed on host systems to dynamically optimize network stacks such that they can make effective use of the terabits/sec networks. Increasing the service orientation of the

network infrastructure will correspondingly enhance the richness of interfaces to middleware, and the usability of distributed systems.

Discussion

Networking monitoring frameworks, such as Perfsonar, are increasingly service-oriented. Such architectures have the potential to minimally couple middleware to evolving network features. These architectures also have the potential for transparently integrating support and diagnostics by expert groups. Over the longer term, automated troubleshooting and performance enhancement for distributed applications should be achievable.

The services approach will be especially useful for implementing and deploying emerging network features. However, creation of a service-oriented system that can make use of current networking, and is still flexible enough to adapt to its future evolution toward terabit networking is preferred. The service must be increasingly reliable, robust, and stable as a prerequisite for benefitting the anticipated communities.

Service orientated methodologies are often accompanied by processes to manage incidents, availability, change, and releases. These processes will make networks more useable.

Finding 4.4

Middle-boxes, or gateway-boxes, have been developed to improve wide-area network throughput for high performance distributed systems across lower speed, longer latency networks. The terabit/sec networks required for Petascale science will provide a new and very different network environment for middle-box technology.

Recommendation

A viable middle-box research program currently exists; however, the fruits of this research program have not yet been fully incorporated in production. The current round of development should go through a vetting, test, and deployment cycle before additional research programs are started. We should pay particular attention to those aspects of the middle-boxes that affect their acceptance by sites and network operators, and their practical deployment. This experience is necessary to inform future research.

Discussion

Middle boxes which are inserted into the data path are primarily useful to the "small" class of users. That is, they are typically useful to users whose sole problem is TCP or other protocol tuning.

Middle-boxes must be ubiquitously deployed to be useful. These boxes need to be "close" to both end systems involved in a transfer. This means that, at a minimum, a successful middle-box solution would need to see wide deployment in both ESnet and the

university community. Therefore, middle-boxes must be shareable by different science activities, and capable of being provisioned by networks or facilities. Middle-boxes must have appropriate usability, including acceptable manageability, maintainability, and security.

The current generation of middle-boxes aims to assist high performance host systems in making more effectively use of network paths, particularly longer latency ones. In addition to questions about scaling and deployment, the role of middle-boxes within terabit/sec network infrastructures, where end systems are the bottlenecks, is less clear. The potential usefulness of middle-box architectures for the emerging Petascale distributed computing environment warrants investigation.

Finding 4.5

Large scale storage systems have begun to be accepted and treated as entities distinctly different from file servers. As storage and caching systems grow beyond the Petabyte scale, their manner of attachment to advanced networks will grow more complex, and they will demand a richer network interface.

Recommendation
An advanced model for presentation of network resources to complex systems, such as distributed storage systems, should be developed, with participation from the system, network, and protocol experts. Recent progress in provision and exploitation of dynamic circuits should inform this effort, although dynamic circuit provisioning may be only one facet of several in an advanced network resource.

Discussion

Storage systems manage their internal state in a somewhat autonomous manner. Neither the client (nor peer) nor the system itself can generally predict the endpoints or overall performance for a single flow, although aggregate peak performance may be well understood. Successful dynamic provisioning for Petascale transfers may be possible only through close cooperation between the storage system and the network.

Group 5: Experimental Networking and Testbeds

A particularly important resource for DOE R&D activities is an experimental network research testbed that will support investigations of innovative technologies, particularly those unique to DOE Petascale science needs. Such a facility would serve as a primary resource for prototyping, testing, and deploying important novel methods and technologies, as well as transitioning them to production environments. Current network research testbeds lack the capacity, scale, capabilities, and end-to-end scope that will be required for DOE Petascale and Exascale science.

Finding 5.1

DOE Petascale science creates multiple significant challenges in terms of network services, technology and infrastructure, which can only be addressed through R&D rounded solidly on experimental testbeds.

Recommendation

Fulfilling the DOE science mission requires the management, transport, exchange, analysis, and sharing of extremely large amounts of scientific data. Many important networking capabilities for Petascale science are not available today. These capabilities will need to be developed through DOE R&D efforts. To ensure that mission science requirements are met, DOE must increase its investment in research and development related to advanced, high-performance communications infrastructure. These investments should be directed at the specific services and technologies required by the DOE Petascale and Exascale science, which are not being addressed elsewhere. Many DOE mission science projects require special capabilities that are unique to its research programs. Typically, these capabilities are required by the DOE science community years before they are recognized as necessities in other environments. A few such capabilities may never be addressed by other initiatives. Progress on developing new methods and technology that are essential to DOE's Petascale and Exascale science requires establishing an experimental network testbed. Traditionally, DOE has led the way in overcoming Internet limitations in order to accomplish its science mission. As a consequence, not only has DOE science advanced, but the wider networking community has also benefited. DOE must maintain its leadership in advanced networking innovations, which is only possible by investigative experimentation on large-scale testbeds.

Discussion

DOE is facing major networking challenges that directly arise from its core Petascale and Exascale science mission. Advanced science innovation and discovery require working directly with the world's largest data resources. Many of these data structures and file sizes are unique to large scale science, and require specialized capabilities. Also, science research requirements are based on heterogeneous resources that are nationally or globally distributed. Many DOE projects require collaborative data sharing among scientists around the globe. Currently, there are performance issues with transferring this data, restricting its availability to the scientific community. As a consequence, the potential benefits of this data may not be fully realized. In addition, the DOE must address multiple new data transport requirements in order to support its emerging Petascale facilities and related resources. These mission-critical requirements are not being addressed by industry. These requirements need to be addressed within the DOE R&D community.

Finding 5.2

A continuation of today's investment priorities and levels in communication services, processes, and technologies is not economically sustainable due to scaling issues. Science missions are becoming more diverse, larger, and more distributed. Domain science research requires the management of many different types of data streams with diverse characteristics.

Recommendation

A new investment approach to the design, implementation, and operation of science-driven communication services and infrastructure is required.

Discussion

Additional investment in attempts to scale and only incrementally improve existing communication services technology is of dubious value. New investment would result in multiple, orders of magnitude benefits. It would meet known and clearly defined needs of the research science communities. It would provide solutions that are directly responsive to the requirements of these communities. It would allow for the design of networks that fluidly scale and adapt to Petascale and Exascale science applications. It would provide for significantly more capacity at less cost, and it would enable automation of what are now human-intensive tasks. It would reduce severe power-consumption issues related to both equipment requirements and environmental conditioning.

Finding 5.3

Now is an optimal time for moving forward on this testbed initiative.

Recommendation

DOE should immediately design and implement a network research testbed, centered on enabling capabilities for its Petascale and Exascale science programs.

Discussion

As noted, the DOE has led the way in advanced networking research and development. The Ultra Science Net was the first national scale network testbed to pioneer capabilities such as advanced bandwidth reservation and wide-area Infiniband testing. Other agencies, including the DoD and NSF, are now establishing major testbeds. Similar testbed network infrastructures are being deployed within other nations. Canada, Europe, and Japan may well be more advanced than the United States in certain network R&D areas. The existence of such testbed projects provides major opportunities for resource leveraging and collaboration, such as cooperative efforts in design, federation, and research program structure. In addition, a new DOE testbed would allow advanced developments in methods, core technologies, and components currently in labs to be made accessible to external communities. The downward trend in cost curves should enable previously intractable problems to be addressed with these innovations.

Finding 5.4

The proposed network R&D agenda must address the specific capabilities required by future DOE mission-oriented science programs, especially those that will support Petascale and Exascale applications.

Recommendation

This initiative must ensure that the testbed will support a research agenda directly related to the requirements of large-scale distributed Petascale and Exascale science experiments and that the agenda be focused on major transformational advances, instead of incremental changes.

Discussion

This testbed will investigate several key topics required by DOE science. For example, very-large-scale, high-end application requirements may only be met through bandwidth that is supported with spectral efficiency of at least 1 Terabits/sec. However, more than bandwidth capacity may be required by these applications; capabilities for dynamic reconfiguration may also be critically important. These special capabilities for dynamic adjustment are essential to science research productivity. Petascale science will require more fluid environments than those commonly implemented. Networks services and infrastructure must be designed to serve applications, which are being significantly constrained by traditional implementations. This requirement leads to a need for new control and management planes, especially those based on distributed vs centralized models. DOE science requires capabilities for close application/workflow/dataflow/infrastructure integration with all resources, including service configuration capabilities within the core network. Data communication services are required for extremely large-scale interactive science experiments, including simulations and modeling. These applications require networks that can respond to multiple requests, within a policy-governed resource allocation framework. It is also essential that these capabilities be distributed among many remote geographic sites, world-wide. For example, science applications must access and efficiently utilize remote, globally distributed science instruments. Capabilities must exist that allow for integration among streams with multiple characteristics, including sizes, times, durations, service levels, and provisioning parameters. These capabilities must also support real-time streams, events, and processes. Network services must be designed and implemented as a major resource within a larger application/infrastructure ecosystem. Therefore, it must be possible to integrate addressable, configurable flows directly with multiple system components, including applications.

Finding 5.5

Although the testbed will support an R&D agenda, it should be designed, operated, and used as a large-scale distributed research instrument.

Recommendation

The testbed should be designed, implemented, operated, and managed as a research facility, and not just a research project. It must serve as a large-scale, highly distributed instrument to support experimental research, including network breakable experiments.

Discussion

The testbed must be able to match aggressive computational advances with corresponding communications performance. It should be designed with capabilities for partitioning resources among multiple simultaneous large-scale experiments and research communities.

The testbed must have capabilities for multi-layer service integration, large-scale multi-service channels, and multiple L1, L2, and L3 channels. It also must be scalable at all required levels. For example, it must be able to support future required capacity levels, including multiple 10 Gbps, 40 Gbps, 100 Gbps, 160 Gbps, 320 Gbps, 1 Tbps, and higher. At each level, it must support per stream bonded and unbonded channels. All channels should be directly addressable and configurable, controlled by edge processes. Such channels must support large-scale dynamic live data streams. The testbed must also provide experimental flexibility, including dynamic provisioning, with extremely short process timings. It must address the need for tactical and strategic timescales for dynamic changes, potentially at ms and ns granularity. It must support high-performance, adaptable protocols and middleware. The testbed should be able to incorporate state-of-the-art components, including those from advanced research labs. The testbed must make possible the integration of network services directly with high-performance edge interfaces, such as instrumentation, storage systems, and compute clusters. The testbed must be operationally manageable and secure at both the experimental level facilities level. It must provide a high degree of virtualization, and must have capabilities for monitoring, analysis, pre-fault diagnostics, and instrumentation, all supported with time synchronization. The testbed must also provide capabilities for federation with other major experimental testbeds, including GENI, CORONET, and international experimental research testbeds.

Finding 5.6

To ensure the success of this initiative, there must be ongoing communications and cooperative activities among all testbed stakeholder communities.

Recommendation

Formal processes should be established to provide for ongoing interactivity among the testbed infrastructure communities and other key stakeholders in this initiative. For example, secure access to the testbed is required by all communities who will participate in R&D programs that will utilize the facility.

Discussion

Formal processes are required to ensure the appropriate partnerships among all the communities that will participate in testbed activities. The primary constituency will consist of the research communities that will use the testbed for their experimental investigations, including academic researchers. This initiative should also establish formal processes to ensure interaction among communities of researchers, who usually focus exclusively on narrow sets of research topics. Advice from DOE application communities is required, including guidance on research directions related to application requirements, and opportunities for these communities to test and model their DOE applications on prototype network infrastructure. In addition, formal processes should be established to ensure the transition of concepts, services, designs, etc., to deployment environments, such as DOE production networks and specialized facilities. Provision should also be made for such production communities to interact with the testbed research community. Formal project engagements with industry may foster development and should be encouraged, as long as the commercial participants fund the cost of their activities. Operations should be oriented

toward providing services, facilities, and related resources to the network R&D community, including support for the design of experimental projects, structured experimentation, monitoring, and analysis of results. The design of the testbed should be based on sharable, open architecture and open source components. The design should be cost, energy, space, and environmentally efficient, using substantially less resources than are consumed today by communications equipment. The testbed should include capabilities for experimentation with edge and host technologies, such as special connections to supercomputers and high performance storage systems.

Appendix

Workshop Agenda

Workshop on Advanced Networking for Distributed Petascale Science: R&D Challenges and Opportunities

Hilton Washington, D.C., Gaithersburg, Maryland

April 8, 2008

7:30–8:30 AM	**Continental Breakfast**	
8:20–10:00 AM	**DOE Network Requirements and Planning**	
8:20 AM	Opening Remarks	Fred Johnson
8:30 AM	Workshop Scope and Charge	Thomas Ndousse
8:40 AM	Distributed Petascale Science and Facilities	Dan Hitchcock
9:00 AM	Office of Science Networking Requirement Summary	Eli Dart
9:20 AM	ESnet4: Next-generation Network for Science	Joe Burrescia
10:00–10:40 AM	**Testbeds and Experimental Networking**	
10:00 AM	OSCARS	Chin Guok
10:08 AM	TeraPaths	Dantong Yu
10:16 AM	LambdaS tation	Don Petravick
10:24 AM	Multi-Layer, Multi-Domain Control Planes	Tom Lehman
10:32 AM	Ultra Science Net	Bill Wing
10:40–11:00 AM	**Break**	
11:00 AM	High-Performance Middleware/Workflow	Scott Klaski
11:20 AM	NSF-GENI Project Update	Chip Elliott
11:40 AM	High-Performance Networking for DOD	Hank Dardy
12:00–1:00 PM	**Working Lunch**	
1:00–1:20 PM	Breakout group overview	
1:20–3:00 PM	E2E Federated Network Measurement	
1:20–3:00 PM	Multi-Layer Federated Network Provisioning	
1:20–3:00 PM	High-Performance End System/Middleware	
1:20–3:00 PM	Transport Protocols and Data Distribution Services	
1:20–3:00 PM	Experimental Networking and Testbeds	
3:00–3:15 PM	**Break**	
3:15–4:30 PM	E2E Federated Network Measurement	
3:15–4:30 PM	Multi-Layer Federated Network Provisioning	
3:15–4:30 PM	High-Performance End System/Middleware	
3:15–4:30 PM	Transport Protocols and Data Distribution Services	
3:15–4:30 PM	Experimental Networking and Testbeds	
4:30–5:30 PM	Breakout Group Preliminary Presentations	
5:30 PM	**Brief logistical remarks. Adjourn**	

April 9, 2008

7:30–8:30 AM	**Continental Breakfast**	
8:30–10:00 AM	Transport Protocols and Data Distribution Services	
8:30–10:00 AM	E2E Federated Network Measurement	
8:30–10:00 AM	Multi-Layer Federated Network Provisioning	
8:30–10:00 AM	High-Performance End System/Middleware	
8:30–10:00 AM	Experimental Networking and Testbeds	
10:00–10:15 AM	**Break**	
10:15–11:00 AM	**Breakout Group Continuation; Draft of Recommendations**	
11:00–12:15 PM	**Presentation of Recommendations**	
11:00 AM	Transport Protocols and Data Distribution Services	Nagi Rao
11:15 AM	E2E Federated Network Measurement	Dantong Yu
11:30 AM	Multi-Layer Federated Network Provisioning	Nasir Ghani
11:45 AM	High-Performance End System/Middleware	Don Petravick
12:00 Noon	Experimental Networking and Testbeds	Joel Mambretti
12:00–1:00 PM	Open Discussion	
1:30–2:30 PM	**Working Lunch. Adjourn.**	
2:30–3:30 PM	Co-Chairs Work on Draft Report	

List of Workshop Attendees

Name	Affiliation
Keren Bergman	Columbia University
Eric Boyd	Internet2
Scott Bradley	Brookhaven National Laboratory
Joe Burrescia	ESnet, Lawrence Berkeley National Laboratory
Yan Chen	Northwestern University
Les Cottrell	Stanford Linear Accelerator Center
Henry Dardy	Naval Research Laboratory
Eli Dart	ESnet
Vince Dattoria	Department of Energy, Advanced Scientific Computing Research
Chip Elliot	BBN Technologies
Wu Feng	Virginia Polytechnic Institute
Nasir Ghani	University of New Mexico
Chin Guok	ESnet
Rick Hafey	Infinera Corporation
Dan Hitchcock	Department of Energy
Fred Johnson	Department of Energy
Kevin Jones	NASA
Rajkumar Kettimuthu	Argonne National Laboratory
Scott Klasky	Oak Ridge National Laboratory
Larry Landweber	National Science Foundation
Tom Lehman	University of Southern California, Information Sciences Institute
Joe Mambretti	iCAIR, Northwestern University
Shawn McKee	University of Michigan
Grant Miller	NCO-Corp.
Paul Morton	National Science Foundation
Thomas Ndousse	Department of Energy
Harvey Newman	California Institute of Technology
Donald Petravick	Fermi National Accelerator Laboratory
Byrav Ramamurthy	University of Nebraska–Lincoln
Yukiko Sekine	Department of Energy
Jerry Sobieski	Sobieski Systems and Service
David Starobinski	Boston University and EPFL
Deborah Stevens	Oak Ridge National Laboratory
Rick Summerhill	Internet2
Brian Tierney	Lawrence Berkeley National Laboratory
Malathi Veeraraghavan	University of Virginia
Jesse Wen	Arcadia Optronics
Bill Wing	Oak Ridge National Laboratory
Dantong Yu	Brookhaven National Laboratory

List of Supporting Documents

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

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

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

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

Network R&D Requirements of ESnet, April 2007.

DOE 20 Years Facilities Plan—Report (pdf)

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

NSF IIS-GENI Workshop Report: First Edition—Report

DOE ASCR 2008 budget—Report (pdf)

Presentations of ESnet Future Technology Assessment and Requirements

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

Science-Driven Network Requirements for ESnet, February 2006, Update of Office of Science Networking Requirements Workshop, 2002

High-Performance Network Planning Workshop, August 13-15, 2002, http://DOECollaboratory.pnl.gov/meetings/hpnpw.

DOE Workshop on Ultra High-Speed Transport Protocols and Network Provisioning for Large-Scale Science Applications, Argonne National Laboratory, April 2003.

DOE Science Networking: Roadmap to 2008 Workshop, June 3-5, 2003, Jefferson Laboratory.

High-Performance Network Planning Workshop, August 13-15, 2002, http://DOECollaboratory.pnl.gov/meetings/hpnpw.

Glossary

ACL	Access Control List
ASCAC	Advanced Scientific Computing Advisory Committee
BER	Office of Biological and Environmental Research
BES	Office of Basic Energy Sciences
BGP	Border Gateway Protocol
CANARIE	Canada's Advanced Internet Development Organization
CA*net	Canada's Advanced National R&E Network
CORONET	DARPA Research Program: Dynamic Multi-Terabit Core Optical Networks: Architecture, Protocols, Control and Management
DANTE	EU R&E Advanced Network Organization (Delivery of Advanced Network Technology to Europe)
DARPA	Defense Advanced Research Projects Agency
DICE	Middleware Project, DOE, I2, CANARIE, GEANT
DOD	Department of Defense
Petascale and Exascale	DOE Petascale Science Projects
DWDM	Dense Wavelength Division Multiplexing
End to End (E2E)	Consistent Service Among All Points On Paths
E-NNI	External Network–Network Interface
ESnet	High-Speed Network Serving Thousands of DOE Scientists and Collaborators Worldwide
FIND	NSF's Future Internet Network Design
FTP	File Transfer Protocol
Gbps	Gigabits Per Second
GEANT	European Multi-Gigabit Network Interconnecting NRNs
GENI	Global Environment for Network Innovations
GMPLS	Generalized Multi-Protocol Label Switching
GridFTP	Grid File Transfer Protocol
HCA	Host Channel Adapter
IB	InfiniBand, an I/O technology that provides high-speed data transfers and ultra low latencies for computing and storage

IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
Internet2	Advanced Networking Consortium Comprised of Universities, Corporations, Government Agencies, Laboratories, Other Institutions of Higher Learning, International Partners
IP	Internet Protocol
ITER	International Thermonuclear Experimental Reactor
LCF	Leadership Computing Facilities
LHC	Large Hadron Collider
LHCnet	DOE Network Supporting LHC Experiments
MoU	Memorandum of Understanding
MPLS	Microprotocol Label Switching
NASA	National Aeronautics and Space Administration
NIC	Network Interface Card
NIH	National Institutes of Health
NIST	National Institute of Standards and Technology
NITRD	Networking and Information Technology Research and Development
NLR	National Lambda Rail US National Optical Fiber Based R&E Network
NOAA	National Oceanic and Atmospheric Administration
NRN	National Research Network
NSF	National Science Foundation
OIF	Optical Internetworking Forum
ONT	Optical Network Testbed
OPN	Optical Private Network
OSCAR	Open Source Cluster Application Resources
OSCARS	ESnet Virtual Circuit Service (On-Demand Secured Circuits and Advanced Reservation System
OSPF	Open Shortest Path First

PCE	Path Computation Element
Petascale	Computers Able To Execute More Than 10^15 Floating Point Instructions Per Second, Communications Capable of Petabits Per Second, Storage Capable of Petabytes
RDMA	Remote Direct Memory Access
RFC	Request for Comments
RON	R&E Regional Optical Network
RSVP-TE	Resource Reservation Protocol–Traffic Engineering
SC	Office of Science
SDH	Synchronous Digital Hierarchy
SDN	ESnet's Science Data Network
SIP	Session Initiation Protocol
SNS	Spallation Neutron Source
SONET	Synchronous Optical Network
Tbps	Terabits Per Second
TCP	Transmission Control Protocol
TL1	Transaction Language 1
UPD	User Datagram Protocol
USN	UltraScience Net—Experimental Research Testbed to Enabling the Development of Hybrid Optical Networking and Associated Technologies