Testbed and Experiments for High-Performance Networking

Nageswara S.V. Rao, Susan E. Hicks, Stephen W. Poole, and Paul Newman

Oak Ridge National Laboratory
raons@ornl.gov

Abstract. UltraScience Net is a network research testbed for supporting the development and testing of wide-area networking technologies for high-performance computing and storage systems. It provides dynamic, dedicated, high-bandwidth channels to support large data transfers, and also provides stable high-precision channels to support fine command and control operations. Its data-plane consists of 8,600 miles of cross-country dual OC192 backbone, which can be dynamically provisioned at different bandwidths. Its out-of-band control-plane is implemented using hardware Virtual Private Network (VPN) devices. In terms of the testbed infrastructure, it demonstrated the following capabilities: (i) ability to build and operate national-scale switched network testbeds, (ii) provisioning of suites of 1 and 10 Gbps connections of various lengths up to 70,000 and 8,600 miles, respectively, through automated scripts, (iii) secure control-plane for signaling and management operations, and (iv) bandwidth scheduler for in-advance connection reservation and provisioning. A number of structured and systematic experiments were conducted on this facility for the following tasks: (i) performance analysis and peering of layer 1-3 connections and their hybrid concatenations, (ii) scalability analysis of 8Gbps InfiniBand (IB) transport over wide-area connections of thousands of miles, (iii) diagnosis of TCP performance problems in using dedicated connections to supercomputers, (iv) detailed TCP performance analysis of wide-area application acceleration devices, and (v) TCP throughput improvements due to 10Gbps High Assurance Internet Protocol Encryptor (HAIPE) devices.

Keywords: Network testbed, dedicated channels, SONET, 10GigE WAN-PHY, control-plane, data-plane, bandwidth scheduler, WAN accelerators, HAIPE, InfiniBand.

1 Introduction

Large-scale computing and storage applications require high-performance networking capabilities of two broad classes: (a) high bandwidth connections, typically with multiples of 10Gbps, to support bulk data transfers, and (b) stable bandwidth connections, typically at much lower bandwidths such as 100s of Mbps, to support operations such as computational steering, remote visualization and remote control of instrumentation. These networking capabilities

T. Magedanz et al. (Eds.): TridentCom 2010, LNICST 46, pp. 19–34, 2011.
© Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2011

may be achieved by providing dedicated connections of the required bandwidths directly between the end users or applications. Such connections are needed only for certain durations between select sites, for example, for archiving at a remote storage facility a terabyte dataset produced on a supercomputer, or actively monitoring and steering a computation on a supercomputer from a remote user workstation. Current Internet technologies, however, are severely limited in meeting these demands because such bulk bandwidths are available only in the backbone, and stable control channels are hard to realize over shared network infrastructures. The design, building and operation of the needed network infrastructure with such capabilities require a number of technologies that are not readily available in Internet environments, which are typically based on shared, packet-switched frameworks. Furthermore, there have been very few network experimental facilities where such component technologies can be developed and robustly tested at the scale needed for large-scale computing and storage systems distributed across the country or around the globe.

The UltraScience Net (USN) was commissioned in 2004 by the U. S. Department of Energy (DOE) to facilitate the development of high-performance networking technologies needed for large-scale science applications, and has been supported by U. S. Department of Defense (DOD) since 2007. USN's footprint consists of dual dedicated OC192 connections, from Oak Ridge to Chicago to Seattle to Sunnyvale. It supports dynamic provisioning of dedicated 10Gbps channels as well as dedicated connections at 150Mbps resolution. There have been a number of testbeds such as UCLP [30], CHEETAH [6], DRAGON [15], HOPI [13] and others that provide dedicated dynamic channels, and in comparison, USN has larger backbone bandwidth and footprint. Compared to research initiatives such as GENI [12] in the U.S., FIRE [11] and FEDERICA [10] in Europe, AKARI [3] in Japan, and CNGI [7] in China, USN has a more focused goal of high-performance applications as in CARRIOCAS project [4] and ARRA ANI [28]. However, we note that USN has been operational for the past five years compared to CARRIOSCAS and ARRA ANI, which are currently being deployed.

The contributions of USN project are in two categories:

(a) **Infrastructure Technologies for Network Experimental Facility:** USN developed and/or demonstrated a number of infrastructure technologies needed for a national-scale network experimental facility. In terms of backbone connectivity at DWDM, USN's design and deployment is similar to the Internet. However, its data-plane is different in that it can be partitioned into isolated layer-1 or layer-2 connections. Its control-plane is quite different mainly due to the ability of users and applications to setup and tear down channels as needed as in [31,4,28]. In 2004, USN design required several new components including a Virtual Private Network (VPN) infrastructure, a bandwidth and channel scheduler, and a dynamic signaling daemon. The control-plane employs a centralized scheduler to compute the channel allocations and a signaling daemon to generate configuration signals to switches.

(b) **Structured Network Research Experiments:** A number of network research experiments have been conducted on USN. It settled an open matter by demonstrating that the bandwidth of switched connections and Multiple Protocol Label Switching (MPLS) tunnels over routed networks are comparable [22]. Furthermore, such connections can be easily peered, and the bandwidth stability of the resultant hybrid connections is still comparable to the constituent pure connections. USN experiments demonstrated that InfiniBand transport can be effectively extended to wide-area connections of thousands of miles, which opens up new opportunities for efficient bulk data transport [24,18]. USN provided dedicated connections to Cray X1 supercomputer and helped diagnose TCP performance problems which might have been otherwise incorrectly attributed to traffic on shared connections [21]. Also, experiments were conducted to assess the performance of application acceleration devices that employ flow optimization and data compression methods to improve TCP performance [19]. USN demonstrated file transfer rates exceeding 1Gbps over 1GigE connections of thousands of miles. Recently, experiments were conducted to assess the effect of 10Gbps High Assurance Internet Protocol Encryptor (HAIPE) devices on TCP throughput over wide-area connections. Somewhat surprisingly, these devices lead to improvements in TCP throughput over connections of several thousands of miles [17].

This paper is organized as follows. In Section 2, we describe USN technologies for data- and control-planes. The results from network experiments are described in Section 3. This paper provides an overview of these topics and details can be found in the references [23,20,22,19,21,25,24,17,18,16].

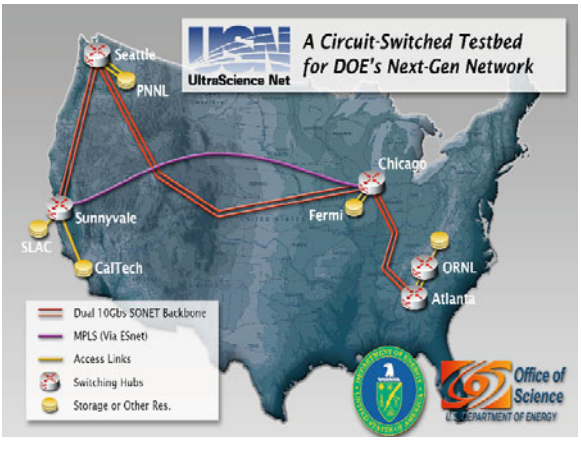


Fig. 1. UltraScience Net backbone consists of dual 10 Gbps lambdas from Oak Ridge to Chicago to Seattle to Sunnyvale

2 USN Infrastructure and Technologies

USN infrastructure is supported by co-location sites at Oak Ridge, Chicago, Seattle and Sunnyvale as shown in Figure 1. USN backbone utilizes ORNL network infrastructure to provide two OC192 SONET connections from Oak Ridge to Chicago, and two OC192 SONET connections from National Lambda Rail (NLR) between Chicago, Seattle and Sunnyvale. USN peered with ESnet [9] at both Chicago and Sunnyvale, and with Internet2 [14] in Chicago. USN architecture is based on out-of-band control-plane as shown in Figure 2, since lack of data-plane continuity makes in-band signaling infeasible.

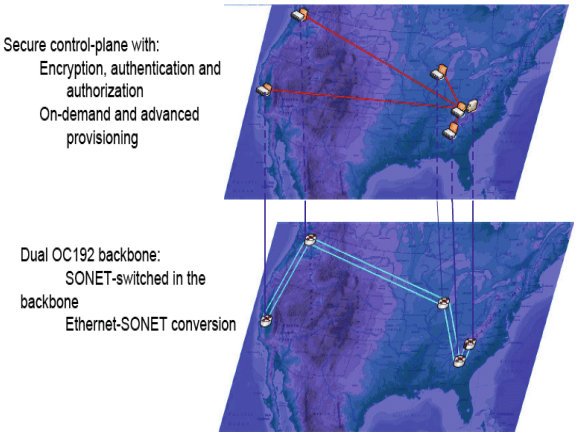


Fig. 2. USN architecture is based on separate data- and control-planes

2.1 Data-Plane

The data-plane of USN consists of two dedicated OC192 SONET (9.6 Gbps) connections as shown in Figure 3, which are terminated on Ciena CDCI core switches. At Oak Ridge, Chicago and Sunnyvale nodes, Force10 E300 Ethernet switches are connected to the corresponding core switches. At Oak Ridge and Chicago core switches also provide 1GigE ports. A variety of data-plane connections can be provisioned using combinations of core and edge switches. SONET connections with bandwidth in the range, 150Mbps - 9.6Gbps, at OC3 (150Mbps) resolution can be provisioned using the core switches. 1GigE connections can be provisioned using OC21 connections between the core switches and cross-connecting them to their 1GigE ports using General Framing Protocol (GFP). Wide-area OC192 connections are provisioned by switching entire lambdas exclusively at core switches. 10GigE WAN-PHY connections are provisioned by terminating OC192 connections on edge switches at the ends; also, intermediate WAN-PHY connections may be provisioned by utilizing E300 switches at

those nodes. Connections switched at 10GigE LAN-PHY are realized by utilizing E300 switches to terminate the WAN-PHY connections and then suitably cross-connecting them to LAN-PHY ports. Thus, USN provides dedicated channels of various resolutions at distances ranging from few hundred miles to thousands of miles, which may be terminated on third party routers or switches or hosts. USN also provides Linux hosts connected to edge switches as shown in Figure 3 to support the development and testing of protocols, middleware, and applications.

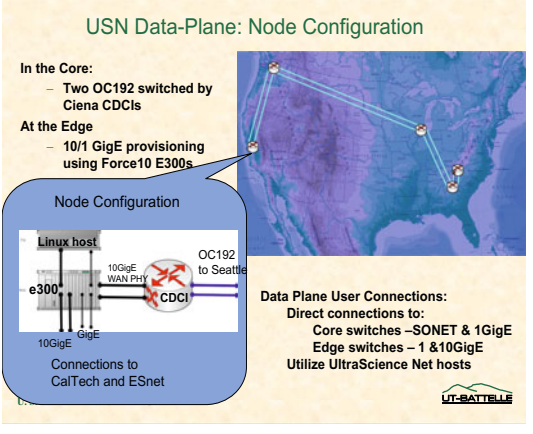
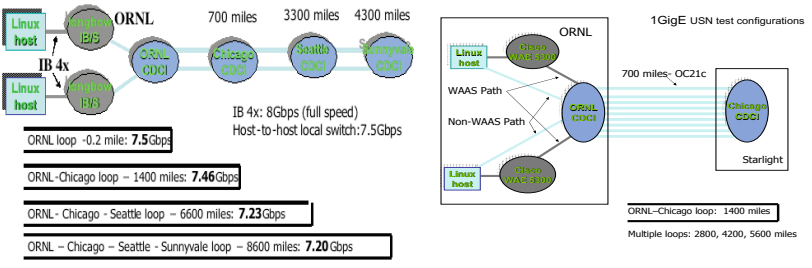


Fig. 3. USN data-plane

In conducting experiments to test the distance scalability of devices and protocols, suites of connections with varying lengths are provisioned between two fixed ports. A suite of 10Gbps connections of lengths 0, 1400, 6600 and 8600 miles are provisioned as shown in Figure 4(a) for InfiniBand experiments. By utilizing OC21 multiplexing we provision 1GigE non-interfering connections on a single OC192 connection, and by switching them at the ends realize several lengths. By using 700 mile dual OC192 connections between Oak Ridge and Chicago we create 1GigE connections with lengths from 0 to 12600 miles in increments of 1400 miles as shown in Figure 4(b). We developed automated scripts that dynamically cycle through all connections of a test suite by invoking a single script.

2.2 Control-Plane

USN control plane consists of the following components [23]: (a) client interface, (b) server front-end, (c) user management, (d) token management, (e) database management, (f) bandwidth scheduler, and (g) signaling daemon. USN control-plane software is implemented in C++ using CGI and PHP scripts for the server and JavaScript and HTML for user interface. It is deployed on a central management node on a Linux workstation at ORNL. The control-plane is implemented



(a) 10Gbps OC192 connections

(b) 1GigE GFP-mapped OC21 connections

Fig. 4. Suites of 1 and 10 Gbps connections provisioned on USN

using hardware-based VPN devices as shown in Fig. 5. Secure VPN tunnels are implemented using a main unit (Netscreen NS-50) at ORNL and secondary units (Netscreen NS-5) at each of the remote sites so that only authenticated and authorized traffic is allowed, and the traffic is encrypted. Each VPN tunnel carries three types of encrypted traffic flows: (i) user access to hosts, (ii) management access to hosts and switches, and (iii) the signaling messages to switches.

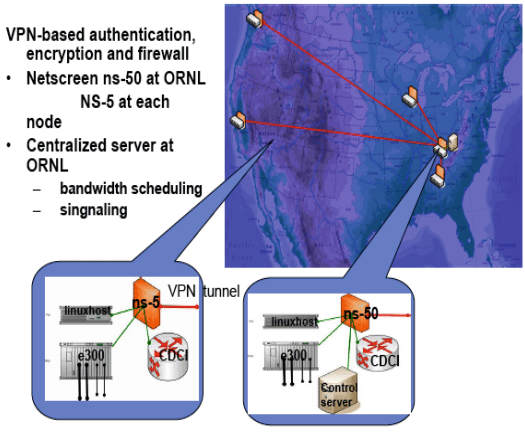


Fig. 5. USN control-plane

Based on network topology and bandwidth allocations, the scheduler computes a path as per user request. The signaling daemon executes scripts to set up or tear down the needed connections. The CDCI core switches are signaled using TL1 commands, and E300 edge switches are signaled using CLI commands. In both cases, EXPECT scripts are utilized by the signaling daemon to login via encrypted VPN tunnels to issue the commands. The bandwidth scheduling

algorithms are developed for: (i) a specified bandwidth in a specified time slot, (ii) earliest available time with a specified bandwidth and duration, (iii) highest available bandwidth in a specified time slot, and (iv) all available time slots with a specified bandwidth and duration. The first three algorithms are extensions of the classical Dijkstra's algorithm [8], and the last one is an extension of Bellman-Ford algorithm, which is an improvement over previous transitive closure algorithm [20]. USN control-plane has the first (2005) implementation of mathematically-validated advance scheduling capability, and more recently such capabilities have been developed in [4,31,2].

3 USN Network Experiments

In this section, we present a summary of experiments conducted on USN facility; their detailed accounts can be found in the references.

3.1 Hybrid Network Connections

Dedicated bandwidth connections may be provisioned at layers 1 through 3 or as combinations. For example, they can be MPLS tunnels over routed network as in ESnet [2], or Ethernet over SONET as in CHEETAH [33], or InfiniBand over SONET as in USN [5], or pure Ethernet paths [1]. An objective comparison of the characteristics of the connections using these technologies is important in making deployment decisions. Once deployed, the costs of replacing them could be very high, for example, replacing MPLS tunnels with SONET circuits entails replacing routers with switches. We collected measurements and compared the throughput and message delays over OC21C SONET connections, 1Gbps MPLS tunnels, and their concatenations over USN and ESnet.

For these experiments we utilized (a) OC21C connections of lengths 700, 1400, ..., 6300 miles on USN as described in the previous section, and (b) 1Gbps 3600 mile VLAN-tagged MPLS tunnel on ESnet between Chicago and Sunnyvale via Cisco and Juniper routers. USN peered with ESnet in Chicago as shown in Figure 6, and 1GigE USN and ESnet connections are cross-connected using E300 switch. This configuration provided hybrid dedicated channels of varying lengths, 4300, 5700, ... , 9900 miles, composed of Ethernet-mapped layer 1 and layer 3 connections. We collected throughput measurements using iperf and Peak Link Utilization Protocol (PLUT) over these connections.

For TCP, we varied the number of streams n from 1 and 10, and for UDP we varied the target rate as 100, 200, ..., 1000, 1100 Mbps; each set of measurements is repeated 100 times. First, we consider USN and ESnet connections of lengths 3500 and 3600 miles respectively and their concatenation. TCP throughput is maximized when n is around 7 or 8 and remained constant around 900, 840 and 840 Mbps for SONET, MPLS and hybrid connections, respectively. For UDP, the peak throughput is 957, 953 and 953 Mbps for SONET, MPLS and hybrid connections, respectively. Hence, there is difference of 60Mbps and 4Mbps between the TCP and UDP peak throughput, respectively, over SONET and

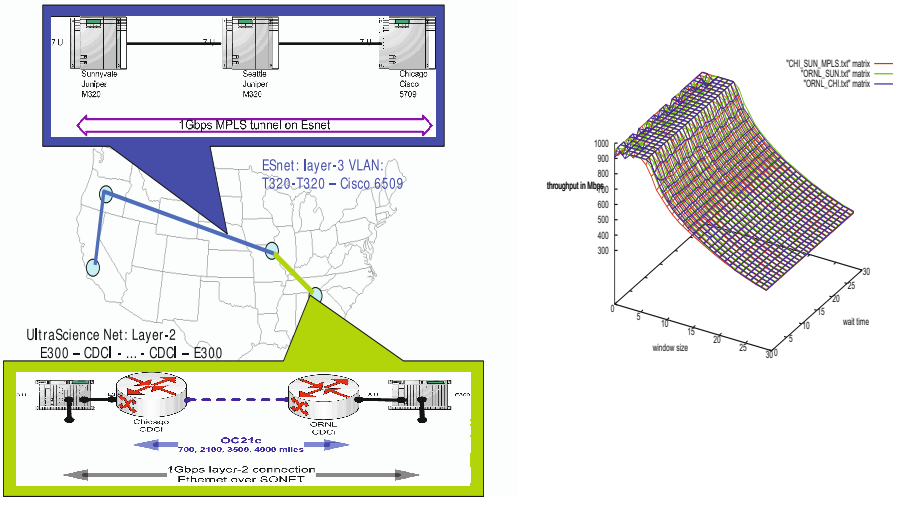


Fig. 6. USN switched OC21C connections and MPLS tunnels implemented using ESnet routers are peered using an Ethernet switch

MPLS connections. Also, there is a difference in peak throughput achieved by TCP and UDP in all cases, namely, 57 and 93 Mbps for SONET and MPLS connections, respectively. We measured file transfer rates over these connections using UDP-based PLUT, which achieved 955, 952 and 952 over SONET, MPLS and hybrid connections, respectively.

USN thus demonstrated that connections provisioned at layers 1-3 can be peered and carried across networks using VLAN technologies, and throughput and message delay measurements indicate comparable performance of layer 1, layer 3 and hybrid connections (detailed results can be found in [22]).

3.2 InfiniBand over Wide-Area

The data transport across wide-area networks has traditionally been based on 1/10GigE technologies combined with SONET or WAN-PHY technologies in the wide-area. InfiniBand was originally developed for data transport over enterprise-level interconnections for clusters, supercomputers and storage systems. It is quite common to achieve data transfer rates of 7.6 Gbps using commodity IB Host Channel Adapters (HCA) (SDR 4X, 8 Gbps peak) by simply connecting them to IB switches. However, geographically separated IB deployments still rely on transition to TCP/IP and its ability to sustain 7.0-8.0 Gbps rates for wide-area data transfers, which by itself requires significant per-connection optimization. Recently, there have been hardware implementations of InfiniBand over Wide-Area (IBoWA) devices, in particular Longbow XR and NX5010ae.

USN was among the first to conduct experiments that showed that these technologies could provide throughput far superior to TCP for dedicated high

bandwidth data transfers [24]. We utilize 0.2, 1,400, 6,600 and 8,600 mile connections and Longbow IBoWA devices in configurations shown in Figure 4 (a). Our results indicate that IB throughput of 7.6Gbps (4x) scales well (with 5% degradation) with no customization to 8600 mile 10GigE and OC192 connections as shown below:

connection length (miles)	0.2	1400	6600	8600	average
average throughput (Gbps)	7.48	7.47	7.37	7.34	7.47
std, dev (Mbps)	45.27	0.07	0.09	0.07	11.40
decrease per mile (Mbps/mile)	0	0.012	0.017	0.016	0.015

In contrast, various TCP (BIC, HTCP, HSTCP and CUBIC) achieved only a few Gbps on this connection. An additional benefit of IBoWA solution is that one could utilize native IB solutions to access remote files systems. However, this solution performed poorly under cross-traffic and dynamic bandwidth conditions, wherein cross-traffic levels of above 2Gbps degraded IB throughputs to about 1Gbps over 8600 mile connection. Thus this approach is mainly suited for dedicated high-bandwidth connections. This work illustrates that transport solutions for high-performance applications could be radically different from current TCP/IP based solutions. More details on this work can be found in [5,24,18].

3.3 Dedicated Connections to Supercomputers

The shared Internet connection from Cray X1 supercomputer at ORNL to North Carolina State University (NCSU) is provisioned at a peak rate of 1Gbps in 2006. But, the data path is complicated. Data from a Cray node traverses System Port Channel (SPC) channel and then transits to FiberChannel (FC) connection to CNS (Cray Network Subsystem) as shown in Figure 7. Then CNS converts FC frames to Ethernet LAN segments and sends them onto GigE NIC. These Ethernet frames are then mapped at ORNL router onto SONET long-haul connection to NCSU; then they transit to Ethernet LAN and arrive at the cluster node via GigE NIC. Thus the data path consists of a sequence of different segments: SPC, FC, Ethernet LAN, SONET long-haul, and Ethernet LAN. The default TCP over this connection achieved throughputs of the order 50 Mbps. Then bbpc protocol adapted for Cray X1 achieved throughputs in the range 200-300Mbps using multiple TCP streams. This low throughput is thought to have been the result of traffic congestion on the shared connection.

Since the capacity of Cray X1's NIC is limited to shared 1Gbps, we developed a dedicated interconnection configuration with 1Gbps capacity by using USN host and direct FC connections from Cray. Hurricane protocol that achieved 90% utilization on other 1Gbps connections was tuned for this configuration. But it only achieved throughputs of the order 400Mps when no jobs are running on Cray X1. Furthermore, its throughput degraded to 200Mbps as jobs are brought online. The throughput problem was diagnosed to the inadequate CPU cycles being allocated to TCP stack, and the network connection had not been the main bottleneck from the start.

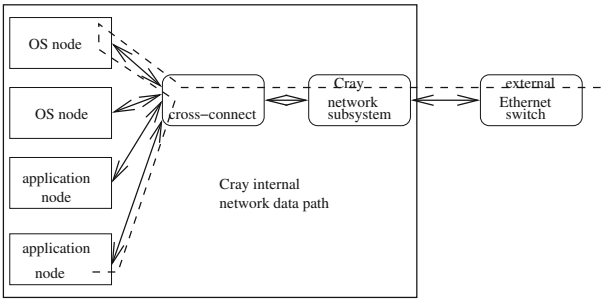
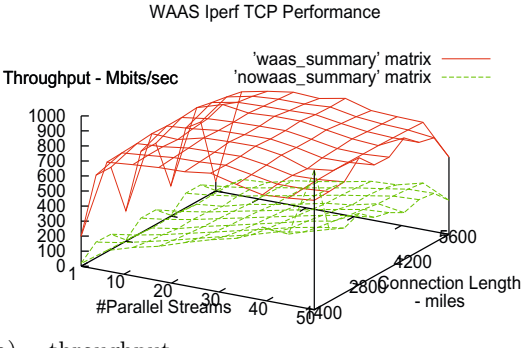


Fig. 7. Dedicated channel to Cray X1 supercomputer

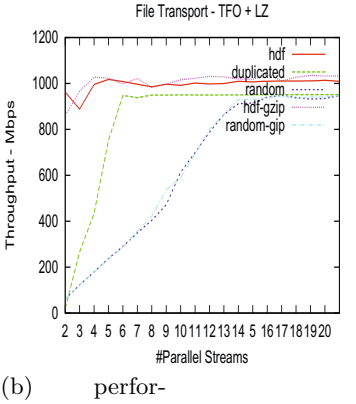
3.4 Wide-Area Application Accelerators

To address the challenges of optimizing the wide-area transport performance, new generation devices are being developed that can simply be "dropped-in" at the network edges. These devices transparently optimize the network flow performance using a combination of Data Redundancy Elimination (DRE) [27], TCP Flow Optimization (TFO) [26] and network data compression [29]. Cisco Wide Area Application Services (WAAS) products are examples of these new technologies. We conducted experiments to quantify the performance of Cisco WAE-500 and WAE-7300 series devices over 1Gbps connections.

We measured iperf TCP throughputs on paths with and without WAAS devices at the ends, called WAAS and non-WAAS paths respectively. We used various connection lengths shown in Figure 4(a) and varied the number of parallel streams from 1 to 50. The relative performance of WAAS and non-WAAS



(a) throughput profiles of WAAS (top) and non-WAAS (bottom)



(b) performance of TFO and LZ on 1400 mile connection

Fig. 8. Performance of WAAS devices

paths for default TCP are summarized in Figure 8. For single streams, WAAS throughput is higher by at least factor of 10 and was as high as 40 in some instances as shown in 8(a). For most multiple streams, WAAS paths throughputs are at least three times higher than non-WAAS paths. The two distinguishing features of WAAS paths are: (a) throughputs reached their maximum with about 5-10 streams, where as non-WAAS paths needed 40 or more streams to reach same levels, (b) WAAS path throughputs were not monotonic in the number of streams unlike the non-WAAS paths.

To study the effects of file contents, we measured file transfer throughputs using iperf with -F option for three different types of files over 1400 mile 1Gbps connection: (a) file with repeated bytes, (b) file with uniformly randomly generated bytes, and (c) supernova simulation files in hdf format. We also gziped these files and utilized them in iperf measurements; gzip implements Limpel-Ziv (LZ) compression on the entire file unlike the incremental implementation on WAE devices. In case (a), gziped file is highly compressed to about 1030 times smaller than the original size, and in case (c) compressed file size is about 0.6831 times the original size. In case (b), however, the gziped file is larger by 0.01% since the file contents were not compressible and the header added to the size.

TFO achieved the best performance for hdf files with throughputs exceeding 1Gbps with 3-6 streams. Least performance improvements are observed for files with repeated contents and random contents. TFO combined with LZ achieved the best performance for hdf files with throughputs exceeding 1Gbps (1017Mbps) with 4-5 streams as shown in Figure 8(b). This performance is about the same order as TFO alone. Least performance improvements are observed for files with random contents, but the performance is much higher than using TFO alone but somewhat lower than using TFO and DRE. In particular, throughputs of 900Mbps were achieved with more than 13 streams, whereas TFO-DRE achieved 950 Mbps with 7 streams. But TFO-LZ performed better than using TFO alone which could sustain 852Mbps with 13 or more streams. USN experimental results can be summarized as follows: (i) highest and lowest throughputs are achieved for hdf and random data files, respectively; (ii) most throughputs were maximized by utilizing 5-10 parallel TCP streams; and (iii) pre-compression of files using gzip did not have a significant effect. In all cases, throughput measurements varied when experiments were repeated.

3.5 IP Encryption Devices

Recent High Assurance Internet Protocol Encryptor (HAIPE) devices are designed to provide 10 Gbps encrypted IP traffic flows. HAIPE is Type 1 encryption device that utilizes cryptography Suites A and B for encrypting IP packet flows. HAIPE's interoperability specification is based on IPsec with additional restrictions and enhancements. They act as gateways between two enclaves to exchange data over an untrusted or lower-classification network. HAIPE device looks up the destination IP address of a packet in its internal Security Association Database (SAD) and picks up the encrypted tunnel based on the appropriate entry. They use internal Security Policy Database (SPD) to set up

tunnels with appropriate algorithms and settings. We generate the performance profiles for TCP and UDP with and without encryption devices. We enabled jumbograms consistently at all devices on the encrypted path, which achieved better performance compared to unencrypted path: (a) for connections 1400 miles and shorter, same throughput levels as unencrypted case were achieved with less number of parallel streams, and (b) throughput improved by more than 50% for longer connections. Thus for TCP, the encryption devices have an effect equivalent to reducing the end-to-end latency which in turn increases the throughput.

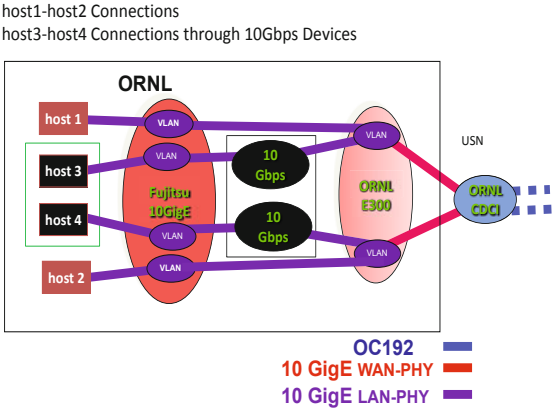


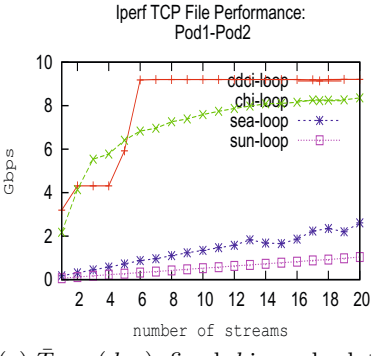
Fig. 9. Wide-area connections with and without HAIPE devices

The hosts are connected to Fujitsu 10 GigE switch which provides multiple connections to E300 switch. First pair of connections are via HAIPE devices as shown in Figure 4, and the second pair are direct connections. By appropriately utilizing VLANS as shown in Figure 9 we realize wide-area connections between pairs of hosts with and without HAIPE devices to realize encrypted and plain connections, respectively. Thus this setup enables side-by-side comparison of the performance of encrypted and plain connections. The same measurement software is used in both cases, and IP address scheme is used to direct the flows onto the encrypted and plain connections as needed. The same wide-area connection is used for both types of traffic, and the experiments are mutually exclusively scheduled so that only one type of traffic is allowed during each test.

We generate throughput profiles for TCP and UDP between hosts connected over connections of different lengths to characterize the achievable throughputs and the corresponding configuration parameters. For TCP, let $T_{TCP}(d, n)$, denote the throughput measurement for files transfers using iperf with -F option over connection of length d using n TCP streams. Let $\bar{T}_{TCP}(d, n)$ denote the average TCP throughput over 10 repeated measurements. TCP distance-profiles are generated by measuring throughputs $T_{TCP}(d, n)$ for the number of parallel

streams n from 1 to 20, for connection length $d = 0.2, 1400, 6600, 8600$ miles. To collect the *stability-profile* for fixed connection length d, we repeat throughput measurements 10 times, for $n = 1, 2, \ldots 20$ for TCP, and $r = 4, 5, 6, 10$, 11 Gbps.

We show the performance profiles without encryptors in Figure 10. For local connections, TCP throughput of 9.17 Gbps was achieved with 6 parallel streams, and it was 8.10 Gbps over 1400 mile connection with 15 streams. However, the throughput was about 1Gbps for 8600 mile connection even with 20 streams. These results were produced with BIC congestion control [32], and similar results with other congestion control modules are described in [24].



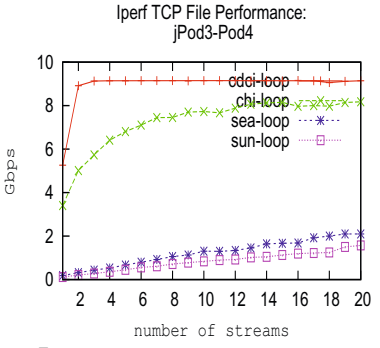
(a) $\bar{T}_{TCP}(d, n)$, fixed d in each plot

miles	0.2	1400	6600	8600
throughput (Gbps)	9.17	8.10	2.61	1.04
# flows	6	15	20	20

(b) TCP file transfer throughputs

Fig. 10. TCP performance over unencrypted connections

We collected iperf measurements over when encryptors are included, and the results are shown in Figure 11. There are several interesting observations. (i) for local connection 9Gbps throughputs were achieved with 2-3 parallel streams compared to 6 streams needed for unencrypted connection; (ii) for 1400 mile connection, 8Gbps throughput was achieved with about 12 parallel streams compared to about 20 for unencrypted case; and (ii) for 8600 mile connection, 1.57 Gbps throughput was achieved with 20 streams compared to about 1Gbps for the unencrypted connection. In summary, the TCP file transfer throughputs were higher when encryptors were employed for the same number of parallel streams as shown in Figure 12. Such throughput improvement is consistently present in all performance profiles at all tested connection lengths. This performance improvement is attributed to the availability of buffers on HAIPE devices which smoothens TCP dynamics and has an effect similar to shortened RTT.



(a) $\bar{T}_{TCP}(d, n)$, fixed d in each plot

miles	0.2	1400	6600	8600
throughput (Gbps)	9.12	8.06	3.11	1.57
# flows	3	13	29	20

(b) TCP file transfer throughputs

Fig. 11. TCP performance over unencrypted connections

files	0.2	1400	6600	8600
encrypted (Gbps)	9.12	8.06	3.11	1.57
plain (Gbps)	9.17	8.10	2.61	1.04

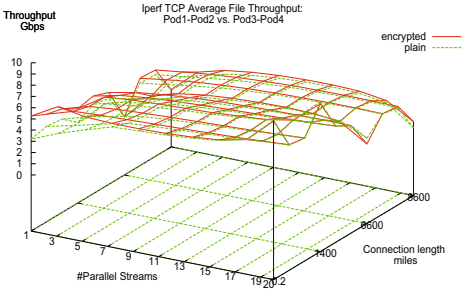


Fig. 12. Comparison of TCP over encrypted and plain connections

4 Conclusions

UltraScience Net is a national-scale testbed conceived and built in support of the development of high-performance networking technologies. It developed and/or tested infrastructure technologies needed for national-scale network experimental facilities, and also supported a number of high-performance research network experiments.

It would of interest to complement USN's experimental facilities with capabilities to (a) realize user-specified network footprints, and (b) analyze, interpolate and extrapolate the collected measurements. It would be beneficial to develop methods to map and realize a user-specified target network on USN infrastructure as closely as possible; such capability enables the testing of national-scale networks which are more general than the connection suites described here.

Despite the availability of connection suites, there are limits on the lengths of connections that can be physically realized on USN. It would be interesting to develop the theory and tools for the design of network experiments to: (i) identify a set of probe connections to optimize the cost and information from the measurements collected, and (ii) suitably interpolate or extrapolate the measurements to predict the performance at desired connection lengths and also derive qualitative performance profiles.

Acknowledgments

This work is currently sponsored by U. S. Department of Defense and is carried out at Extreme Scale Systems Center, Oak Ridge National Laboratory managed by UT-Battelle, LLC for U. S. Department of Energy under contract No. DE-AC05-00OR22725.

References

1. Dynamic resource allocation via GMPLS optical networks, `http://dragon.maxgigapop.net`
2. On-demand secure circuits and advance reservation system, `http://www.es.net/oscars`
3. Akari architecture design project for new generation network, `http://akari-project.nict.go.jp`
4. Audouin, O., Cavalli, A., Chiosi, A., Leclere, O., Mouton, C., Oksman, J., Pasin, M., Rodrigues, D., Thual, L.: CARRIOCAS Project: an experimental high bit rate optical network tailored for computing and data intensive distributed applications. In: Proc. Tridentcom (2009)
5. Carter, S.M., Minich, M., Rao, N.S.V.: Experimental evaluation of infiniband transport over local and wide-area networks. In: Proceedings of High Performance Computing Conference (2007)
6. End-To-End Provisioned Optical Network Testbed for Large-Scale eScience Application, `http://www.ece.virginia.edu/mv/html-files/ein-home.html`
7. China next generation network, `http://www.cernet2.edu.cn/en/bg.htm`
8. Cormen, T.H., Leiserson, C.E., Rivest, R.L.: Introduction to Algorithms. Hill Book Co. McGraw-Hill, New York (1990)
9. Energy Sciences Network, `http://www.es.net`
10. Federated e-infrastructure dedicated to european researchers innovating in computing network architecture, `http://www.fp7-federica.eu`
11. Future internet research and experimentation, `http://cordis.europa.eu/fp7/ict/fire`
12. GENI: Global environment for network innivations, `http://www.geni.net`
13. Hybrid Optical and Packet Infrastructure, `http://networks.internet2.edu/hopi`
14. Internet2, `http://www.internet2.edu`
15. NSF Shared Cyberinfrastructure Division PI Meeting, February 18-20 (2004), `http://hpn.east.isi.edu/nsf-sci`

16. Rao, N.S.V., Wing, W.R., Hicks, S.E., Poole, S.W., Denap, F.A., Carter, S.M., Wu, Q.: Ultrascience net: High-performance network research test-bed. In: International Symposium on on Computer and Sensor Network Systems (2008)
17. Rao, N.S.V., Poole, S.W., Hicks, S.E., Kemper, C., Hodson, S., Hinkel, G., Lothian, J.: Experimental study of wide-area 10gbps ip transport technologies. In: Proc. of Milcom (2009)
18. Rao, N.S.V., Poole, S.W., Newman, P., Hicks, S.E.: Wide-area Infiniband RDMA: Experimenal evaluation. In: Workshop on High-Speed Interconnects for Distributed Computing (2009)
19. Rao, N.S.V., Poole, S.W., Wing, W.R., Carter, S.M.: Experimental analysis of flow optimization and data compression for tcp enhancement. In: INFOCOM 2009 Workshop on Terabits Networks (2009)
20. 20. N. S. V. Rao, W. R. Wing,, S. M. Carter, and Q. Wu. Ultrascience net: Network testbed for large-scale science applications. IEEE Communications Magazine, expanded version (2005) (in press), `http://www.csm.ornl.gov/ultranet`
21. Rao, N.S.V., Wing, W.R., Carter, S.M., Wu, Q.: High-speed dedicated channels and experimental results with hurricane protocol. Annals of Telecommunications 61(1-2), 21–45 (2006)
22. Rao, N.S.V., Wing, W.R., Wu, Q., Ghani, N., Lehman, T., Dart, E., Guok, C.P.: Measurements on hybrid dedicated bandwidth connections. In: INFOCOM 2007 Workshop on Terabits Networks (2007)
23. Rao, N.S.V., Wu, Q., Carter, S.M., Wing, W.R., Ghosal, D., Banerjee, A., Mukherjee, B.: Control plane for advance bandwidth scheduling in ultra high-speed networks. In: INFOCOM 2006 Workshop on Terabits Networks (2006)
24. Rao, N.S.V., Yu, W., Wing, W.R., Poole, S.W., Vetter, J.S.: Wide-area performance profiling of 10gige and infiniband technologies. In: Pautasso, C., Tanter, É. (eds.) SC 2008. LNCS, vol. 4954. Springer, Heidelberg (2008)
25. Sahni, S., Rao, N.S.V., Li, Y., Jung, K., Ranka, S., Kamath, N.: Bandwidth scheduling and path computation algorithms for connection-oriented networks. In: Proceedings of International Conference on Networking (2007)
26. Semke, J., Madhavi, J., Mathis, M.: Automatic TCP buffer tuning. In: Proc. ACM SIGCOMM 1998 (1998)
27. Spring, N.T., Wetherall, D.: A protocol independent technique for eliminating redundant network traffic. In: Proceedings of the 2000 ACM SIGCOMM Conference (2000)
28. Tierney, B.L.: The ARRA ANI Network Testbed Project. In: JointTechs Meeting (2010)
29. Tye, C., Fairhurt, G.: A review of ip packet compression techniques. In: Proc.GNet (2003)
30. User Controlled LightPath Provisioning, `http://phi.badlab.crc.ca/uclp`
31. Varvarigos, E., Sourlas, V., Christodoulopoulos, K.: Routing and scheduling connections in networks that support advanced reservations. Computer Networks 52, 2988–3006 (2008)
32. Xu, L., Harfoush, K., Rhee, I.: Binary increase congestion control (bic) for fast long-distance networks. In: INFOCOM (2004)
33. Zheng, X., Veeraraghavan, M., Rao, N.S.V., Wu, Q., Zhu, M.: CHEETAH: Circuit-switched high-speed end-to-end transport architecture testbed. IEEE Communications Magazine (2005)