8-2005

CHEETAH: Circuit-Switched High-Speed End-to-End Transport Architecture Testbed

Xuan Zheng
University of Virginia - Main Campus

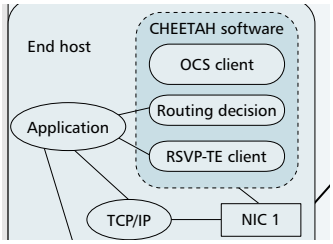
Malathi Veeraraghavan
University of Virginia - Main Campus

Nageswara S. V. Rao
Oak Ridge National Laboratory

Qishi Wu
Oak Ridge National Laboratory

Mengxia Zhu
Oak Ridge National Laboratory, mzhu@cs.siu.edu

Published in Zheng, X., Veeraraghavan, M., Rao, N.S.V., Wu, Q., & Zhu, M. (2005). CHEETAH: circuit-switched high-speed end-to-end transport architecture testbed. IEEE Communications Magazine, 43(8), s11-s17. doi: 10.1109/MCOM.2005.1497551



CHEETAH: CIRCUIT-SWITCHED HIGH-SPEED END-TO-END TRANSPORT ARCHITECTURE TESTBED

Xuan Zheng and Malathi Veeraraghavan, University of Virginia
Nageswara S. V. Rao, Qishi Wu, and Mengxia Zhu, Oak Ridge National Laboratory

ABSTRACT

We propose a circuit-switched high-speed end-to-end transport architecture (CHEETAH) as a networking solution to provide high-speed end-to-end circuit connectivity to end hosts on a dynamic call-by-call basis. Not only is it envisioned as a complementary service to the basic connectionless service provided by today's Internet; it also relies on and leverages the presence of this service. Noting the dominance of Ethernet in LANs and SONET/SDH in WANs, CHEETAH circuits will consist of Ethernet segments at the ends and Ethernet-over-SONET segments in the wide area. In this article we explain the CHEETAH concept and describe a wide-area experimental network testbed we have deployed based on this concept. The network testbed currently extends between Raleigh, North Carolina, Atlanta, Georgia, and Oak Ridge, Tennessee, and uses off-the-shelf switches. We have created CHEETAH software to run on end hosts to enable automated use of this network by applications. Our first users of this network testbed and software will be the Terascale Supernova Initiative (TSI) project researchers, who plan to use this network for large file transfers and remote visualizations.

INTRODUCTION

Next-generation supercomputers promise speeds approaching 100 Tflops within the next few years and petaflops further along. They hold enormous promise in meeting the demands of a number of large-scale e-science computations from fields as diverse as earth science, high energy and nuclear physics, astrophysics, fusion energy science, molecular dynamics, nanoscience, and genomics. These large-scale e-science applications require networking support for tasks such as transfers of large data sets, distributed collaborative visualization, remote computational steering, and/or remote instrument control. These tasks demand the following from the network:

- High throughputs for large data transfers
- Low-latency and low-jitter connections for remote visualization and computational steering
- Protocols/middleware to support collaborative work environments

Despite recent advances in middleware and grid technologies [1], these requirements of large-scale e-science applications are not met by current networks. In particular, neither the bandwidth nor delay/jitter-controlled transport needs are adequately met by the current Internet for different reasons. For the former, link bandwidths are simply not adequate, and furthermore current transport protocols (e.g., TCP) do not scale to hundreds of gigabits per second, even if such link capacities are available [2, 3]. For the latter, the connectionless nature of the Internet makes stable and controlled transport very difficult to achieve. As a result, the network capabilities required to support this scale and range of networking activities surpass, by several orders of magnitude, the capabilities of today's deployed high-speed networks. Indeed, the sheer data volumes together with the control requirements warrant quantum leaps in the functionality of current network infrastructures.

To meet the end-to-end bandwidth/delay/jitter requirements of large-scale e-science applications, circuit-switched networking is a natural choice because of its connection-oriented nature. Increasingly a number of optical circuit-oriented testbeds are being deployed under this paradigm, including CANARIE's CA*net 4, OMNInet, SURFnet, UKLight, DOE's UltraScience net, and the NSF DRAGON network. While these testbeds differ in details, they all offer end-to-end circuits of some sort. Some of these networks use all-optical switches to realize circuits with the granularity of a single wavelength, while others use hybrid electronic/optical switches that provide sub-lambda granularity.

Several research efforts in these optical circuit-oriented testbeds are focused on how to provision the circuits. For example, the User Controlled Lightpaths (UCLP) project supported by CANARIE aims to provide user-controlled end-to-end synchronous optical network (SONET) circuits. However, proposed applications for circuits in UCLP and other related projects focus on large data transfers and other e-science applications that require calls with long

 0163-6804/05/$20.00

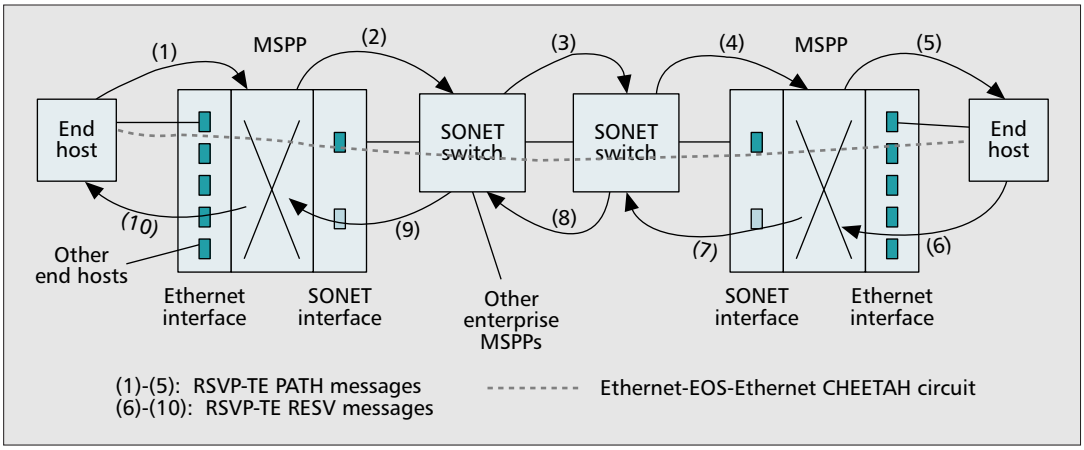


FIGURE 1. The CHEETAH setup and an example of hop-by-hop circuit setup.

holding times. Unlike commodity applications on the Internet, these long-duration calls are generated by a relatively small group of users and applications. To handle circuit provisioning for this set of users, UCLP uses a centralized approach, in which the network inventory, topology, and routing information are stored in a global database, and circuit setup requests are processed by a central management system. The complexity of such a centralized management system makes fast provisioning hard to implement and limits the scalability of networks.

We proposed a solution, the circuit-switched high-speed end-to-end transport architecture (CHEETAH), in [4] to address a wide range of applications that require rate-guaranteed service. The goal of CHEETAH is to enable dynamic sharing of bandwidth in optical circuit-switched networks. To enable large-scale network deployments based on the CHEETAH solution, it requires control-plane functionality to be distributed to network switches. Once a CHEETAH circuit is established, user data can simply be streamed unhindered, resulting in low transfer delays. Furthermore, there is almost no variation in the delays experienced by different data blocks sent over the circuit. Therefore, end-to-end latency and jitter guarantees are possible in this networking solution.

We are currently deploying a CHEETAH testbed primarily for use by the Terascale Supernova Initiative (TSI) e-science project. The two main applications that require network services in the TSI project are remote visualization and file transfer tasks. Thus, our additional objective is to support these two applications by optimizing them to circuits provided by the CHEETAH testbed.

The rest of the article is organized as follows. First, we describe the CHEETAH concept. We then describe the wide-area CHEETAH testbed being deployed between Raleigh, North Carolina, Atlanta, Georgia, and Oak Ridge, Tennessee. We present how we envision CHEETAH's usage for the remote visualization and file transfer tasks of TSI. Finally, we describe our ongoing work to extend the CHEETAH concept and network.

THE CHEETAH CONCEPT

The CHEETAH concept is based on the following observations:

- Ethernet is dominant in LANs. Gigabit Ethernet (GbE) and 10GbE technologies have been deployed at end hosts.
- The SONET-based optical network is dominant in WANs.
- A new system called the multiservice provisioning platform (MSPP) is being rapidly deployed at enterprises, which has the capability to map Ethernet frames onto equivalent Ethernet-over-SONET (EoS) signals
- Not only is the standardization of generalized multiprotocol label switching (GMPLS) rapidly reaching completion; more important, these protocols have already been implemented in vendors' switches.

Figure 1 illustrates the CHEETAH configuration. It consists of hosts equipped with two Ethernet network interface cards (NICs), in which the **secondary** Ethernet NICs are directly connected to the Ethernet ports of enterprise MSPPs. MSPPs, in turn, are connected to a SONET circuit-switched network, in which switches are equipped with GMPLS protocols to support call-by-call dynamic bandwidth sharing. GMPLS protocols, as a complete set of control-plane solutions, consist of the triple combination of Link Management Protocol (LMP), Open Shortest Path First — Traffic Engineering (OSPF-TE) routing protocol, and Resource Reservation Protocol — TE (RSVP-TE) signaling protocol. This set of protocols was designed to minimize manual support needed for normal operation of the network. Thus, they are critical to the deployment of large-scale connection-oriented networks.

An end-to-end circuit in a CHEETAH network consists of hybrid Ethernet signals from/to end hosts within LANs and equivalent-rate EoS circuits across the wide area. These end-to-end connections are indeed "circuits" because the MSPP behaves as a space switch, mapping all data coming from an Ethernet port into a set of time slots/outgoing port on the SONET side, and vice versa.

Figure 1 illustrates an example of circuit setup. End hosts (e.g., PCs) run RSVP-TE clients, allowing both human users and application software to trigger the setup and release of CHEETAH circuits. These circuits are set up dynamically between end hosts with RSVP-TE signaling messages being processed at each intermediate switch in a hop-by-hop manner. Once set up, the circuit is held and used for a short duration, and then released using a similar hop-by-hop circuit release procedure. Subsequently a different communication session can reuse the same resources. The network resources (i.e., link capacities) are used by customers in a complete sharing mode.

CHEETAH is designed as an **add-on** service to Internet connectivity, **leveraging** the services of the latter. As shown in Fig. 2, the primary NICs (NIC 1) in the end hosts are connected to the public Internet through the usual LAN Ethernet switches and/or IP routers, while the secondary NICs (NIC 2) are connected to Ethernet ports on the enterprise MSPPs, allowing them to be crossconnected on demand to equivalent wide-area EoS circuits. This allows an end host engaged in a CHEETAH circuit to simultaneously communicate with other hosts using TCP/IP on its primary NIC. For communication between two CHEETAH end hosts that can be connected by a direct CHEETAH circuit, there is a choice of two paths:

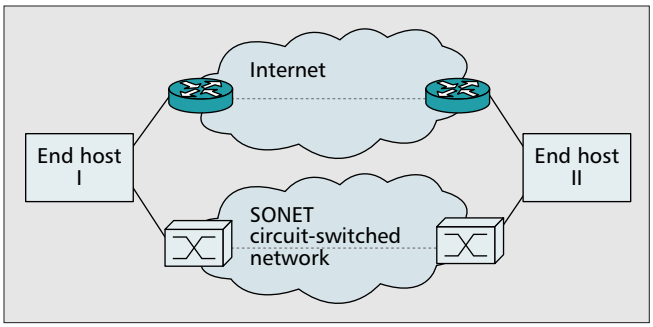


FIGURE 2. CHEETAH as an add-on service to the Internet.

- An Internet packet-switched path
- A dedicated end-to-end CHEETAH circuit

Since there is contention for resources on both paths, but the sharing mode on the Internet path allows for new transfers to simply join in (start sharing bandwidth without an admission gate), we included a fallback option in the CHEETAH design. Applications can fall back to the Internet path if the CHEETAH circuit setup attempt fails due to unavailability of resources on the CHEETAH network. This is another example of how we leverage the Internet path in the CHEETAH solution.

End hosts implement a **routing decision** function to determine whether or not to request setup of a CHEETAH circuit based on the data transfer size and the loading conditions on the two paths. For utilization reasons, not all data transfers should attempt to use CHEETAH circuits. For example, in file transfer applications, if the file transmit time is itself smaller than the time to set up a circuit, the application should choose the Internet path. We provide further details on this routing decision algorithm later.

Preserving Internet connectivity allows us to gradually upgrade end hosts (one at a time) with CHEETAH capability. But this results in scenarios in which a CHEETAH host may need to communicate with a non-CHEETAH host. To handle this condition, we add an *optical connectivity service* (OCS). OCS is used to determine whether or not the correspondent end host has access to the CHEETAH service, and furthermore whether or not it is connected via the same optical circuit-switched network.

A Wide-Area CHEETAH Testbed

Network Configuration

We selected the Sycamore SN16000 Intelligent Optical Switch to serve both roles, MSPPs and SONET switches, in our CHEETAH testbed. This choice was primarily driven by the advanced implementation of GMPLS protocols on the SN16000 platform. The SN16000 platform supports both SONET and Ethernet interface cards. Therefore, it can be configured as an MSPP with both Ethernet and SONET interface cards, or as a SONET switch with exclusively SONET interface cards.

Figure 3 shows the wide-area CHEETAH testbed topology. The CHEETAH network testbed currently extends between Raleigh, North Carolina, Atlanta, Georgia, and Oak Ridge, Tennessee, allowing enterprises such as North Carolina State University (NCSU) to be connected to Oak Ridge National Laboratory (ORNL). Three CHEETAH points of presence (PoPs) equipped with Sycamore SN16000 switches are located at MCNC[1] in Raleigh, Southern Crossroads (SOX)[2]/Southern Light Rail (SLR)[3] in Atlanta, and ORNL in Oak Ridge. The long-distance SONET circuits are leased from National Lambda Rail (NLR),[4] SLR, and ORNL, all of which support research-oriented lambda infrastructures.

[1] MCNC: MCNC provides access for North Carolina institutions to national research networks, including Internet2 and the NLR.

[2] SOX: SOX is a cooperative initiative by the members of the Southeastern Universities Research Association to provide next-generation Internet services for university researchers at institutions throughout the Southeast.

[3] SLR: SLR is a Georgia Tech non-profit corporation providing NLR access to universities in Georgia and other Southern regions of the United States, and governmental and private sector organizations involved in university research initiatives.

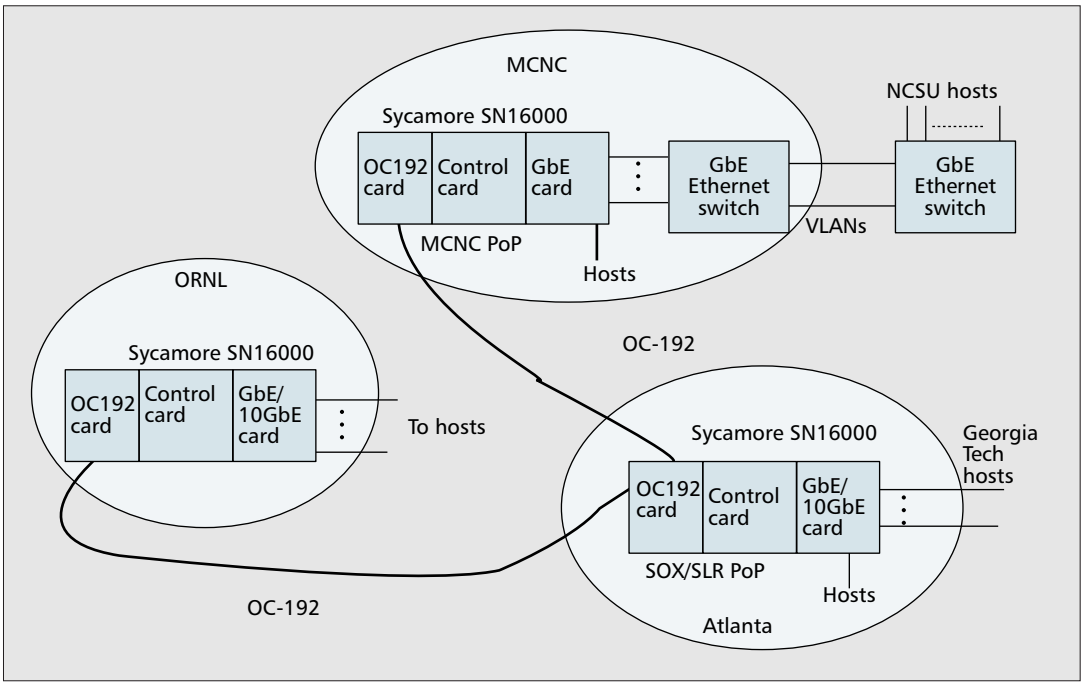


FIGURE 3. Wide-area CHEETAH testbed (data plane).

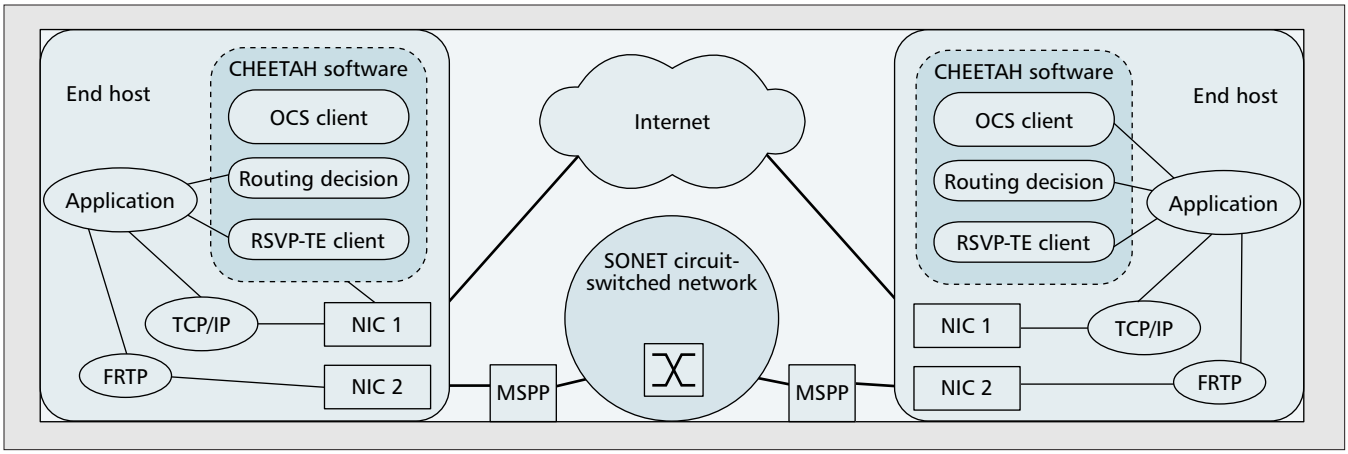


FIGURE 4. CHEETAH end-host software.

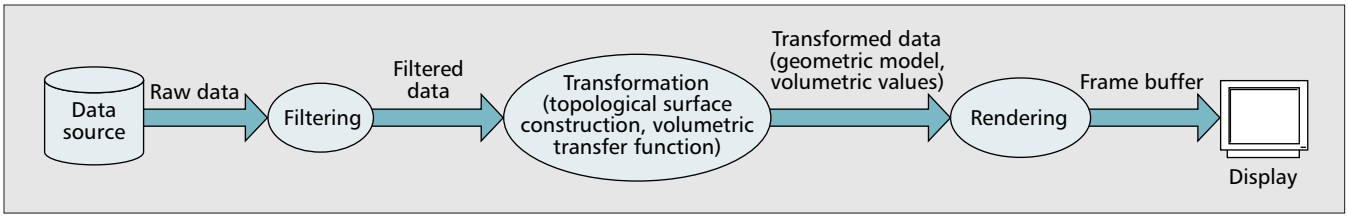


FIGURE 5. A general visualization pipeline.

In this CHEETAH testbed, some end hosts, physically located at the CHEETAH PoPs (set up for experimental research), are connected to the SN16000 switches with direct cables/fibers. However, since enterprises are geographically distant from the CHEETAH PoPs, it is not feasible to connect enterprise end hosts to the SN16000 switches via direct physical connections. Therefore, we connect such end hosts to an SN16000 switch through provisioned virtual LAN (VLAN) or MPLS connections. For example, in Fig. 3 multiple GbE VLANs are set up through GbE Ethernet switches at NCSU and MCNC. Each VLAN connects one end host at NCSU (via the secondary NIC in the end host) to one GbE port on the SN16000 switch in MCNC. The SN16000 switch then maps the GbE signal from each host onto the equivalent EoS signal on the wide-area SONET circuit dynamically using GMPLS protocols.

CHEETAH End Host Software

To support the functionalities described earlier, we are developing CHEETAH end host software for Linux PCs. We are implementing four basic modules as part of the CHEETAH end host software: an OCS client module, a routing decision module, an RSVP-TE client module, and a transport protocol module called Fixed-Rate Transport Protocol (FRTP). An overview of the CHEETAH end host software architecture is shown in Fig. 4.

The **OCS client** module is used to determine whether or not the correspondent end host has access to the same CHEETAH network. The **routing decision** module can answer queries from other software modules on whether or not to attempt CHEETAH circuit setup for a given data transfer. The **RSVP-TE client** module initiates circuit setup by issuing an RSVP-TE PATH message to the enterprise MSPP, which is a Sycamore SN16000 in our network. The RSVP-TE signaling messages are transmitted through the primary NIC and the Internet (out-of-band signaling). The final component of the CHEETAH end host software is a transport protocol module, **FRTP**. The reason we developed a new transport protocol is that TCP and other IP-related transport protocols were developed for connectionless packet-switched networks, requiring them to implement rate-altering congestion control schemes, which are counterproductive to utilization of dedicated end-to-end circuits. The goal of a transport protocol for an end-to-end dedicated circuit should be to keep the pipe full by continuously sending data at a rate equal to the circuit rate; hence, the phrase *fixed-rate* in the name of our transport protocol. The reader is referred to [5] for further details on FRTP design, implementation, and experimental results.

[4] *NLR: NLR is a major initiative of U.S. research universities and private sector technology companies to provide a national-scale infrastructure for research and experimentation in networking technologies and applications.*

Applications on the CHEETAH Network

In this section we describe how two applications in the TSI project, remote visualization and file transfers, exploit the rate-guaranteed nature of circuits.

Remote Visualization for TSI Large-Scale E-Science Applications

A remote visualization system enables an end user equipped with a simple display device and network access to visualize large volumes of scientific data stored and/or rendered at remote sites. Visualization process involves several steps that form the so-called visualization pipeline as shown in Fig. 5. The filtering module extracts the information of interest from the raw data and performs the necessary preprocessing. The transformation module typically uses a surface fitting technique to derive 3D geometries. The rendering module converts the transformed 3D geometric to a pixel-based 2D image. The performance of such a system critically depends

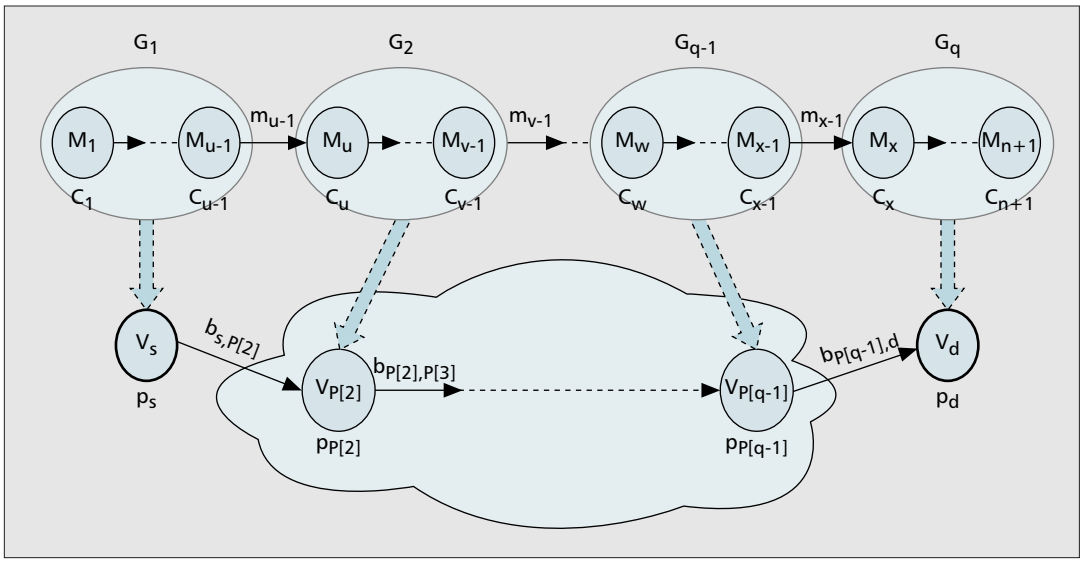


FIGURE 6. Pipeline partitioning and mapping.

on how efficiently its visualization pipeline is mapped onto the network nodes.

Many existing remote visualization systems employ a predetermined partition of the visualization pipeline and typically send fixed-type data streams such as raw data, geometric primitives, or images to remote clients. While such schemes are common, they are not always optimal for high-performance visualizations that typically deal with large data sets for various visualization purposes. Over Internet connections, this problem is further compounded by the limited bandwidths and time-varying network dynamics. The CHEETAH circuits solve both problems by providing dedicated circuits with large bandwidths as well as low losses and jitter due to the absence of competing traffic.

As shown in Fig. 6, the visualization pipeline consists of $n + 1$ sequential modules denoted by $M_1, M_2, \ldots, M_{n+1}$. Module $M_j, j = 2, \ldots, n + 1$ performs a computational task of complexity cj on data of size m_{j-1} from module M_{j-1} and generates data of size m_j, which is then sent over the network link to module M_{j+1}. An underlying network consists of k geographically distributed computing nodes denoted by $v_1, v_2, \ldots, v_{k+1}, v_k$. Node $v_i, i = 1, 2, \ldots, k$ has a normalized computing power p_i and is connected to its neighbor node $v_j, j = 1, 2, \ldots, k, j \neq i$ via an edge or link $L_{i,j}$ of bandwidth $b_{i,j}$ and link delay $d_{i,j}$. The transport network is represented by a graph $G = (V, E)$, $|V| = k$, where V and E are sets of nodes and edges, respectively. In CHEETAH both the bandwidths and links delays are quite accurately known, unlike in the Internet where they have to be estimated using complex techniques [6].

We consider a path P of q nodes from a source node v_s to a destination node v_d in the transport network, where $q \in [2, \min(k, n + 1)]$ and path P consists of nodes $v_{P[1]} = v_s, v_{P[2]}, \ldots, v_{P[q-1]}, v_{P[q]} = v_d$. Each edge of this path corresponds to a CHEETAH circuit with known bandwidth. The pipeline is decomposed into q *visualization groups* denoted by $G_1, G_2, \ldots, G_{q-1}, G_q$, which are mapped one-to-one to the nodes of P. The data flow between two adjacent groups originates at the last module in the preceding group such that we have $m(G_1) = m_{u-1}, m(G_2) = m_{v-1}, \ldots, m(G_{q-1}), G_q$. The client at last node v_d sends control messages to one or more preceding visualization groups to support interactive operations. The transport time for a control message is assumed to be negligible due to its small size.

Our goal is to minimize the total delay on computing and transport incurred from the source node to the destination. We assume the transport time between modules of a group assigned to the same node to be negligible. We now describe a dynamic programming method to compute the mapping to minimize the total delay. Let $T^m(v_i)$ denote the minimal total delay with the first m messages mapped from source node v_s to end node v_i.

$T^m(v_i)$ is computed from the minimal of the two following scenarios based on adding the module M_{m+1} to the partial pipeline. In scenario 1 we execute M_{m+1} at node v_i itself, and add its computing time to $T^{m-1}(v_i)$, a subproblem of v_i of size $m - 1$. In scenario 2 we map link of M_{m+1} to a network link from among all links incident to node v_i, and choose the minimum. Thus, in each iteration $T^m(v_i)$ either inherits the mapping scheme from $T^{m-1}(v_i)$ by simply adding module M_{m+1} to the last group, or just starts a separate group with module M_{m+1} to the mapping scheme of $T^{m-1}(u), u \in adj(v_i)$. The complexity of this algorithm is $O(n \times |E|)$. Details of this derivation can be found in [6].

We implemented and deployed our visualization system at three nodes located at NCSU, ORNL, and Louisiana State University (LSU) over Internet connections. The parallel computation is implemented on the Orbitty cluster at NCSU, which consists of 23 nodes (total of 92 CPUs each at 2.4 GHz, and total of 441.6 Gflops). Linux workstations with 3 GHz CPUs are used as hosts at ORNL and LSU. Our system provides functionalities such as iso-surface extraction, ray casting, streamline, and linear integral convolution (LIC). The server estimates the delay time based on the entity being visualized and bandwidth. The best routing table and pipeline partition scheme for each task is then calculated. Appropriate visualization modules are assigned to corresponding nodes for minimal total delay.

File Transfer Applications

We identify two reasons end-to-end high-speed circuits are suitable for file transfers. First, file transfers can take advantage of any data rate, however high, provided the end hosts can keep up with the rate offered by the network connection. Second, from a utilization perspective, a file transfer, which is a streaming of data from one storage disk or memory location to another, can keep a circuit fully utilized. This can be accomplished by matching the sending rate to the circuit rate (or, more correctly, having the network set up a circuit at a rate that can be matched by the sender and receiver of the file transfer).

Prior to actual data transfer, a routing decision module determines whether to attempt a CHEETAH circuit setup or

Measure of loading on Ckt. Sw. network / TCP/IP path	Number of switches on the circuit k = 4			Number of switches on the circuit k = 20		
	P_b			P_b		
	0.01	0.1	0.3	0.01	0.1	0.3
p_{loss} = 0.0001	610 Kbytes	640 Kbytes	840 Kbytes	2.4 Mbytes	2.65 Mbytes	3.4 Mbytes
p_{loss} = 0.001	490 Kbytes	550 Kbytes	730 Kbytes	2 Mbytes	2.2 Mbytes	2.8 Mbytes
p_{loss} = 0.01	120 Kbytes	140 Kbytes	180 Kbytes	500 Kbytes	550 Kbytes	650 Kbytes

TABLE 1. Threshold file sizes when $r_{cheetah} = r_{primary}$ = 100 Mb/s and T_{prop} = 0.1ms.

directly choose the Internet path. The routing decision algorithm uses the data transfer size and loading conditions on the two paths. The basic idea is to estimate and compare the file transfer delays expected if a CHEETAH circuit setup is attempted with the expected delay incurred if the application resorts directly to the Internet path. If, in a successful circuit setup, the delay is much lower on the circuit, this difference will offset the probability of call setup failure, making the attempt worthwhile. The mean transfer delay on the TCP/IP path is primarily influenced by three factors: the bottleneck link rate on the end-to-end Internet path, $r_{primary}$, the packet loss rate on the end-to-end Internet path, p_{loss}, and the end-to-end propagation delay,T_{prop} [7]. The mean transfer delay expected if a CHEETAH circuit setup is attempted depends on four factors: call blocking probability on the CHEETAH circuit path, P_b, the data rate of the CHEETAH circuit, $r_{cheetah}$, the number of switches on the end-to-end CHEETAH circuit, k, and the end-to-end propagation delay, T_{prop} (we assume that the Internet path and CHEETAH circuit have the same end-to-end propagation delay).

We computed the threshold file sizes for various combinations of P_b, $r_{cheetah}$, $r_{primary}$, p_{loss}, k, and T_{prop}. For transfers of sizes below the threshold size, the end host software should resort directly to using the primary path, and for file sizes larger than the threshold size, the software should attempt a CHEETAH circuit setup. Due to space considerations, we include only one set of numerical results as an example of the routing decision results. For a complete description of the algorithm and detailed numerical results, the reader is referred to [4, 8].

Table 1 shows the threshold file sizes when $r_{cheetah} = r_{primary}$ = 100 Mb/s and T_{prop} = 0.1 ms. As an example, if the call-blocking probability P_b on the optical circuit-switched network is 30 percent and the packet loss rate on the TCP/IP path p_{loss} is 1 percent, for calls in which the CHEETAH circuit traverses 20 switches, 650 kbytes is the threshold file size.

SUMMARY AND FUTURE WORK

Unlike comparable testbeds that focus exclusively on e-science applications, the CHEETAH solution is aimed at creating a scalable connection-oriented network for a wide range of commodity as well as e-science applications. It achieves this goal by using switches that have built-in control plane engines to handle bandwidth management in a completely sharing, distributed, and dynamic manner. This makes it more scalable than solutions with centralized bandwidth management. Furthermore, its feature of leveraging Internet connectivity (e.g., the fallback option if a circuit attempt is rejected) is key to gradual growth of CHEETAH. It allows service providers to operate networks based on the CHEETAH concept at a high call blocking probability to achieve a targeted network utilization. This is very important in the growth phase of CHEETAH service, allowing gradual addition of endpoints into the CHEETAH network.

Commodity applications targeted for CHEETAH circuits include storage area networks (SANs), interactive gaming, video telephony/conferencing, file transfers as in video file distribution, and peer-to-peer (P2P) file sharing. Some of these applications have relatively low bandwidth requirements when compared to e-science applications, but instead have strict delay/jitter requirements. All can take advantage of the bandwidth/delay/jitter guaranteed service provided by CHEETAH networks.

While the CHEETAH concept was originally developed for end-to-end Ethernet-EoS-Ethernet circuits, it can be applied to other types of connections. Therefore, we are broadening the scope of this work to the creation of a *heterogeneous connection-oriented Internet* that would offer connection-oriented (bandwidth reservation) services to complement the connectionless service of today's Internet. The possibility of creating such an internetwork is based on our observation that many deployed switches, both in local area and metro/wide area networks, while originally designed for connectionless service, have now been upgraded with hardware and software to support connection-oriented services. For example, IP routers from major vendors now include MPLS as well as corresponding RSVP-TE signaling software. Similarly, Ethernet switches from many vendors now include implementations of the IEEE 802.11q VLAN standard, combined with features such as ingress rate shaping. While Ethernet switches currently do not include control plane signaling software, such software can be run at external hosts connected to switches' control ports. External control plane signaling software could process RSVP-TE messages on behalf of these switches, perform bandwidth management through connection admission control algorithms, and then issue SNMP commands to configure VLANs through the switch fabric. Thus, MPLS switches and VLAN-based Ethernet switches can be used in a network to offer dynamically controlled connections. In addition to these solutions, WDM-based switches are becoming available to construct yet another kind of connection-oriented network and to offer corresponding services.

The CHEETAH concept can be readily applied to any of these connection-oriented networks. There can however be subtle details in this extension. For example, the relatively longer switch fabric configuration time in optical WDM switches and hence longer end-to-end circuit setup delay will change the quantitative analysis results described earlier. However, this difference only changes the values of crossover file sizes in the routing decision module of CHEETAH software (Table 1); it does not as such alter the CHEETAH concept.

With this multitude of connection-oriented technologies, it has become clear we need to design and create standards that implement "glue" functionality to enable the dynamic setup/release of heterogeneous connections (i.e., ones that

traverse different types of connection-oriented networks). Thus, a heterogeneous connection could consist of one or more SONET/SDH circuit segments, WDM lightpath segments, VLAN Ethernet segments, and MPLS label switched path segments. Additionally, interesting bandwidth sharing algorithms will be needed to set limits on how much of a link's capacity can be peeled off for connections on links connected to switches that support both connectionless and connection-oriented service capabilities (e.g., IP routers with MPLS/RSVP-TE capabilities).

ACKNOWLEDGMENTS

We thank John Blondin, John Moore, and Elliot Peele from NCSU, Ibrahim Habib, Song Qiang, and Zhaoming Li from the City University of New York, Steven Carter and Bill Wing from ORNL, Mark Johnson and Ray Suitte from MCNC, Brian Savory and Cas D'Angelo from Georgia Institute of Technology, Jerry Sobieski, Tom Lehman, Aihua Guo, Chris Tracy, and Xi Yang from the NSF Dragon project, Vijay Pandian and Tom DiMicelli from Sycamore Networks, and Xiuduan Fang, Zhanxiang Huang, Anant Padmanath Mudambi, and Xiangfei Zhu from the University of Virginia for their help in creating the CHEETAH testbed.

This work was carried out under the sponsorship of NSF ITR-0312376, NSF ANI-0335190, and DOE DE-FG02-04ER25640 grants at the University of Virginia, and also NSF_ ANI-0229969 and NSF ANI-0335185 grants at ORNL.

REFERENCES

[1] "High-Performance Networks for High-Impact Science," Rep. of the Aug. 13–15, 2002 Wksp. conducted by the Office of Science, U.S. Department of Energy.
[2] I. Foster *et al.*, "Grid Services for Distributed Systems Integration," *IEEE Comp.*, June 2002, pp. 37–46.
[3] W. Feng and P. Tinnakornsrisuphapá, "The Failure of TCP in High-Performance Computational Grids," *Proc. SC2000: High-Perf. Network and Comp. Conf.*, Dallas, TX, Nov. 2000, http://citeseer.nj.nec.com/386136.html
[4] M. Veeraraghavan *et al.*, "CHEETAH: Circuit-switched High-speed End-to-End Transport ArcHitecture," *Proc. Opticomm 2003*, Dallas, TX, Oct. 13–17, 2003.
[5] M. Zhu *et al.*, "Adaptive Visualization Pipeline Decomposition and Mapping onto Computer Networks," *Int'l. Conf. Image and Graphics*, 2004.
[6] X. Zheng, A. P. Mudambi, and M. Veeraraghavan, "FRTP: Fixed Rate Transport Protocol — A Modified Version of SABUL for End-to-End Circuits," *Pathnets2004 Wksp. Broadnet2004 Conf.*, San Jose, CA, Sept. 2004.
[7] N. Cardwell, S. Savage, and T. Anderson, "Modeling TCP Latency," *Proc. IEEE INFOCOM*, vol. 3, Tel-Aviv, Israel, Mar. 2000, pp. 1742–51.
[8] M. Veeraraghavan and X. Zheng, "A Reconfigurable Ethernet/SONET Circuit Based Metro Network Architecture," *IEEE JSAC*, vol. 22, no. 8, Oct. 2004, pp. 1406–18.

BIOGRAPHIES

XUAN ZHENG (xz3y@virginia.edu) is a research associate in the Department of Electrical and Computer Engineering at the University of Virginia. He received his B.E. and M.S. degrees from Xian Jiaotong University, Xian, China, in 1997 and 2000, respectively, and his Ph.D. degree from the University of Virginia in 2004. His current research work on optical networks is supported by NSF and DOE. He received a best student paper award at Opticomm '03.

MALATHI VEERARAGHAVAN (mv@cs.virginia.edu) is an associate professor in the Departments of Computer Science and Electrical and Computer Engineering at the University of Virginia. She also serves as director of computer engineering. She received her B.Tech. degree from Indian Institute of Technology, Madras, in 1984, and M.S. and Ph.D. degrees from Duke University in 1985 and 1988, respectively. After a 10-year career at Bell Laboratories, she served on the faculty at Polytechnic University, Brooklyn, New York, 1999–2002. Her current research work on optical networks is supported by NSF and DOE. She holds 22 patents and has received five best paper awards. She served as Technical Program Committee Chair for IEEE ICC 2002.

NAGESWARA S. V. RAO (raons@ornl.gov) received a B.Tech. in electronics engineering from Regional Engineering College, Warangal, India, in 1982, an M.E. from the School of Automation, Indian Institute of Science, Bangalore, in 1984, and a Ph.D. in computer science from Louisiana State University in 1988. He is currently a distinguished research staff member at Oak Ridge National Laboratory, which he joined in 1993. His research interests include network transport dynamics, statistical approaches to transport control, and information and sensor fusion.

QISHI WU (wuqn@ornl.gov) received a B.S. in remote sensing from Zhejiang University, China, in 1995 and an M.S. in geomatics from Purdue University in 2000. He received an M.S. in systems science in 2002 and a Ph.D. in computer science in 2003, both from Louisiana State University. He has been a post-doctoral fellow at Oak Ridge National Laboratory since 2003. His research interests include high-performance transport protocols, stochastic approximation methods for network transport, and distributed sensor networks.

MENGXIA ZHU (mzhu@bit.csc.lsu.edu) earned her Bachelor of Engineering in biomedical engineering from Zhejiang University, China, in 1996, and her Master of Science in system science from Louisiana State University in 2002. She is now a Ph.D. student in the Computer Science Department at Louisiana State University. She also works as research scholar in the Computer Science and Mathematics Division at Oak Ridge National Laboratory. Her primary research interests include remote visualization and distributed sensor networks.