Please summarize the conclusions that have emerged from your activities. Later screens will invite you to identify publications and other concrete products (collections, databases, software, inventions, and so on) and to explain the significance and implications of both findings and products for your field, for other fields, and even beyond science and engineering.

- If you have no findings to report, at least for now, please click the corresponding button. We anticipate that as the project progresses your emphasis in reporting will shift from activities to findings and products, and ultimately to contributions.

Our key conclusions:

1. Control-plane: In an April 2007 IEEE JSAC publication, we (i) described our experiences in implementing the CHEETAH experimental wide-area GMPLS network, (ii) explained our design approaches to scalability (public IP addressing), control-plane security, and end-host signaling client usage by integration into applications for user-transparent usage of circuits, and (iii) presented measurements for typical end-to-end circuit setup delays across this network (an end-to-end OC1 circuit setup delay from a Linux end host in Raleigh to a host in Atlanta, traversing two Sycamore SN16000 switches, is 166ms). Usage of ssh and CLI commands to program the Force10s on the HOPI testbed adds to call setup delay. Our control-plane findings with respect to the HOPI testbed are that the CHEETAH distributed control-plane solution can be applied to a network of Ethernet switches with rate-limited VLANs used to create virtual circuits.

2. Transport protocols: Our main finding is that given there is a call setup phase in circuit/virtual-circuit (VC) networks, in which the committed rate, peak rate and buffer size\(^1\) are configured at each switch on the end-to-end circuit/VC and are hence known, these values should be used to determine and set appropriate values for various parameters of a CTCP/BIC-TCP connection at the end hosts (such as send and receive buffer sizes, txqueuelen, etc.). We studied the differences between CTCP and BIC-TCP behavior over the three types of circuits/virtual circuits described in the Item 3 of the Activities

---

\(^1\) for virtual circuits only.
document. These differences are most interesting when the SONET VCAT circuit is of lower rate than the GbE interface from the sending host to the switch. When the congestion window at a TCP sender increases above the total allocated buffer space at the sending end host and the GbE interface card of the Sycamore switch to which the host is connected, Pause frames are sent by the switch card to the TCP sending host’s NIC if Layer-1 circuits are configured. This prevents packet loss. On the other hand, with the Layer-2 circuits, if a burst is sent at a Gb/s speed from the TCP sending host, the Sycamore switch’s GbE interface card will drop packets that are not served out quickly enough by the lower-rate SONET VCAT circuit, when buffer space runs out. With the multiplexed Layer-2 circuit, BIC-TCP’s constant probing for available bandwidth causes interference between the multiple flows routed to VLANs sharing a single SONET VCAT circuit. On the HOPI testbed, determining and setting the correct parameter values for the “rate police” command on the Force10 E600 switches was important to achieving high throughput with CTCP. We are currently documenting these findings in a journal submission. In terms of performance, CTCP has an advantage over TCP for small file transfers that occur over an idle circuit since Slow Start is disabled in CTCP, and TCP resets its congestion window if a connection is idle for some preset time period. However, for large file transfers, both perform equally well on a dedicated Layer-1 circuit since the initial Slow Start period has a negligible effect. With congestion control no longer a limitation, the limitations of other factors such as disk I/O and Ethernet device drivers has become clearer.

3. Our BWdetail and CTCP (congestion control module insertion) implementation experiences show that increasing levels of transparency offered in the newer releases of the Linux kernel are most useful to networking researchers.

4. Circuit-aware application gateway (CAG): We ran experiments in which a Web client was configured to use one CHEETAH-connected CAG as its proxy server with the Web server that it was accessing being closer to a distant CAG. The wide-area path consisted of two metro-area TCP connections over (connectionless) regional/enterprise segments to connect the client and the server to their respective CHEETAH-connected CAGs, and one wide-area CTCP connection between the two CHEETAH-connected CAGs over a dedicated high-speed circuit. Our experimental results demonstrate performance benefits. For example, in one experimental setup, we showed that for a 10-MB file, the Web throughput improves from 17Mb/s to 61Mb/s by using this overlay solution relative to the direct client-to-server path over the Internet. This gain was not strictly due to a reduction in Round-Trip Time (RTT) caused by a splitting of the end-to-end TCP connection; instead owing to PlanetLab policy limitations, we could not increase the receive buffer size on the Web client from the default 64KB to the required 125KB needed for the direct Internet path’s bandwidth-delay product. In general, if the dominant fac-
tor is RTT, using the Internet core subnetwork between the CAGs will impact the delay only for small transfers. The TCP sender resets the congestion window if the connection is idle for more than one retransmission time-out. But for large transfers, there will be not a significant difference in delay if both paths are of the same data rate. Given our argument is based on observations that the cost of higher-speed interfaces are lower for circuit switches than IP routers, potentially the Internet path could be of a lower rate than the path through the circuit-switched subnetwork. Therefore, even without the Planetlab limitations, we expect our solution of introducing circuit-switched subnetworks in the core to provide performance-to-cost ratio benefits. A paper on our implementation of the CAG and experimental results appeared in the Proc. of IEEE ICCCN2007. New experimental results gathered in Fall 07 include measurements of per-packet processing delay within the CAGs. We found that per-packet processing delay in CAGs is on the order of 0.1ms. We also captured traffic traces and learned that packet flow across the three connections, client-to-CAG, CAG-to-CAG, and CAG-to-server, are pipelined. Further, we experimented with different values for parameters in the CAG to learn the influence of these parameters on CAG performance.

5. On the virtualizer software, developed for the HOPI testbed, our preliminary finding is that it is feasible to virtualize a switch such as the Force10 E600, to allow multiple networking researchers to experiment with their own software to configure their virtual switch partitions as needed. We are currently writing a proposal for the NSF GENI solicitation based on this work.

6. Network administration of an experimental wide-area network is complex and time-consuming. Network administrators must be prepared for different kinds of problems, such as switch hardware failures, hardware incompatibilities, cable failures, misconfigurations and other procedural errors, and security problems. These problems can be more severe in an experimental network than in a production network for various reasons. First, experimental network administrators should allow researchers to test out new ideas that could require actions not commonly allowed in production networks, which could lead to unanticipated problems. Second, a lack of experience, which is to be expected, in graduate and undergraduate research students often cause problems. Finally, experiments sometimes require the configuration of network switches and routers in ways that are not commonly seen in production networks. A careful balance is required between providing individuals flexibility (to support experimentation) while simultaneously protecting the network to avoid problems for all. Our recommendations include having remote access cards on distant PCs for an automatic reboot from the CD-ROM drive (we have installed these cards on the CHEETAH hosts and left operating-system CDs in the drives), and redundant hardware in switches (such as second SMCs in the Sycamore switches), which are usually avoided initially to save costs (however, such purchases will save time in the long run).
7. CHEETAH network extension: We were interested in extending CHEETAH to support sub-Gbps VLAN circuits. For hybrid Ethernet-SONET circuits, Sycamore’s software release 7.6.2.1 only supports signaling for port-mapped GbE circuits (mapped to OC3c-7v VCAT signals). As standardization is still pending for hybrid Ethernet VLAN/SONET virtual circuits, the Sycamore control-plane implementation does not support these types of sub-Gbps virtual circuits. Therefore, we tried several network designs that would allow us to continue using the built-in GMPLS switch controller of the Sycamore SN16000, with its support for lower-rate SONET circuits (as low as OC1). We first researched the availability of IP routers or MSPPs with SONET interfaces and UNI-C implementations with the idea that it would allow us to connect end host GbE NICs to these systems, which in turn would be connected to the Sycamore switches via SONET interfaces. There were two problems. First, channelized OC SONET interface cards in routers are very expensive. Second, only high-end (expensive) IP routers, such as Cisco’s 12000XR and the CSR router, implement UNI-C. Both these issues made IP routers unsuitable for the purpose of serving as gateways between Ethernet NICs in hosts and a SONET network with a vendor-built distributed control-plane solution for SONET circuits. We then explored the use of Cisco 15454 MSPPs. These are low-cost systems that support both Ethernet and channelized SONET interface cards. While UNI-C was available in earlier versions of the 15454 software, its use had been discontinued in the current release. We next considered the use of these devices as multiplexers, through which sub-Gbps VLANs on Ethernet ports could be hard-wired to equivalent-rate SONET circuits. RSVP-TE SONET signaling could then be used from end-host clients to signal through directly to the Sycamore switches “pretending” to set up end-to-end SONET circuits. This solution was feasible for OC1-rate circuits, but owing to design of the 15454 Ethernet cards, with their support for only a limited number of VCGs (Virtual Concatenation Groups) per card, this design was unscalable for higher-speed circuits (e.g., 300Mbps, which is ideal for file transfers given disk-access limits). Our conclusion was that the most viable solution is to implement an external control-plane software module for the Sycamore switch to handle sub-Gbps hybrid VLAN-SONET circuits, or await Sycamore’s implementation while experimenting on the HOPI testbed. We chose the latter option. We describe these and related findings in a report titled “Extension options for the CHEETAH network” posted on our project web site.

8. Basic research findings (Xiangfei Zhu): Our models for book-ahead (BA) schemes provide network designers a means to select the size of the reservation window, $K$, in time-interval units, corresponding to a desired set of values for the output metrics, for a given value of $m$, the link capacity expressed in channels. For example, when $m$ is 8, increasing the reservation window beyond 4 intervals does not affect the call congestion or system utilization, but causes the mean scheduling delay to increase. We
showed with comparative simulations that our analytical model can be used as a solution for an M/D/m/p queueing system at moderate-to-high loads, a queueing system for which, to our knowledge, no analytical solution exists. From our extensions to the simulations, we found that call blocking probabilities of BA schemes with bell-shaped distributions for call-initiation time options are close to those of BA schemes with uniformly distributed call-initiation time options when the standard deviation of the bell-shaped distribution is large and the “bell” is flat. However, the call blocking probability increases as the standard deviation decreases, because the call-initiation time options are likely to congregate in a small interval when the standard deviation is small. We also understood the fundamental difference between the BA and immediate-request (IR) bandwidth-sharing modes. Book-ahead mechanisms lower call blocking probability relative to IR schemes because the advance-reservation horizon effectively serves as a buffer for calls that cannot be accommodated immediately. This reduction in call blocking probability is more significant when the number of channels is small.

9. Basic research findings (Xiuduan Fang): Mean waiting time for applications in the first category (holding time independent of per-call circuit rate) will be intolerably high when both the per-call circuit rate and mean call holding time are high, and hence we do not recommend the use of IR mode for such applications. At the other extreme, we found that the high level of aggregation and multiplexing needed when the per-call circuit rate is a small fraction of link capacity (e.g., \( \frac{1}{1000} \)) makes the use of IR mode prohibitively expensive. For file-transfers, our primary conclusion is that to justify the use of a circuit, the per-call circuit rate should be larger than the expected effective throughput on the IP-routed path. This rate is chosen to maximize the average reduction in delay. As the traffic load increases, the ideal per-call circuit rate decreases. Moreover, we observed that the minimum file size (above which a circuit setup is attempted) is typically determined by utilization considerations rather than delay considerations due to high call setup delays. This finding simplifies the computation of the minimum file size. We are preparing a revised version of our paper for submission to the IEEE Transactions on Communications.

10. Basic research findings (Tao Li): We compared in-band and out-of-band signaling transport options under assumptions of the GMPLS switches having hardware-accelerated signaling engines or software signaling protocol processors. With hardware signaling engines, if the load per in-band signaling channel can be kept low by using many of the DCC channels (one per OC1), in-band signaling is the preferred option. Since the network delay for carrying the messages through IP routers can be significantly more in the out-of-band option, in-band signaling is better in spite of the per-channel transmitter rate being lower. With software-based signaling protocol processors, queueing delays are
likely to build up for the processor, making this component the most significant part of end-to-end call setup delay. In this case, the lower network delay for in-band signaling makes it a better option than the out-of-band signaling channel. The difference in transmitter rates does not matter since the message transmission service time is much smaller than the message processing service time, which means there is no queueing delay at the transmission servers for both in-band and out-of-band options. A paper on these results appeared in the Proc. of IEEE ICCCN07 conference.

11. We found that it is feasible to have applications use circuits without modifying the application source code through the use of the ptrace Linux system call.

12. Our business-development discussions with MCNC and UTK yielded a classification of services shown below. In coarse-grained sharing, the ratio of the link capacity to per-call circuit rate is small (around 10), and call holding time is long. In fine-grained sharing, either the ratio is large or the call holding time is short, or both. We considered three possible cases for the deployment scope of the Dynamic Circuit Services (DCS) network, which correspond to the columns in the table above. From the applications listed in the body of the table, we explored two further, business interconnect and Internet access private line (PL) for fine-grained sharing, and CDN web services. These are described in the next two items.

13. One of our key findings relates to the use of high-speed circuits. In today’s Internet architecture, high-speed circuits (which are effectively links between IP routers or Ethernet switches) are regarded as essential on only portions of the network where large volumes of traffic are aggregated. Our finding is that a high-speed circuit offers delay benefits if made available even for single file transfers. This finding has impact when viewed together with the observation that many leased lines are realized on rate-limited high-speed interfaces. For example, UVa’s WAN-access leased line is a 10GigE (Gbps), but three VLANs, each rate-limited to about 1Gbps, have been provisioned across the UVA WAN-access
link, with the first leading to an NLR Packetnet IP router, the second to an Internet2 connector IP router, and the third to a commercial service provider’s IP router. Taken together, we explore an application in which these rate limits can be increased or decreased dynamically for single file transfers that traverse such a rate-limited link. We call this application, “dynamic adjustment of PL bandwidth.” Consider for example a business-interconnect PL from a branch location to headquarters. The PL will typically be of modest bandwidth owing to the high prices of leased lines. However, if the line is realized with higher-rate interfaces, then, with dynamic call control signaling, the rate can be increased for single storage backups. This application is described in further detail in the presentation made to Sycamore on Nov. 20, 2007, which is available through our project web site.

14. Dynamic CDN: Based on the data presented in Chris Anderson’s 2006 best-selling book, “The Long Tail: Why the Future of Business is Selling Less of More,” we observed that Content Delivery Network (CDN) service could have greater value if it can be offered in a “dynamic” mode. We explored two example scenarios. The first example occurs when a company is featured in a popular magazine such as slashdot.org, which results in a surge of hits to that company’s web server. Dynamically recruiting CDN servers and pushing copies of the web pages to these servers would be most beneficial to handle the sudden surge. A second example is of a user in a remote location trying to access a web page at some distant server. In this case, even if the distant web server is that of a Fortune-500 company, the latter may not have contracted the CDN service provider to host a mirror close to this remote client’s location. A dynamic recruiting and data push would lower the client’s user response time. This dynamic CDN service requires high-speed connectivity just between PoPs where CDN servers are typically located. With GMPLS control, a high-speed optical circuit seems most appropriate for this application since lowering delay is important. Unlike with ordinary CDN service, where web page images and other data can be pushed at low speed during off hours, the dynamic CDN service requires high-speed connectivity between CDN servers after users start accessing web pages. This application is described in further detail in the presentation made to Sycamore on Nov. 2, 2007, which is posted on the project web site.