Project Description

1 Introduction

We begin this work by asking the question, "What are connection-oriented networks really good for?" Despite decades of tussle over connection-oriented vs connection-less services, the answer remains a subject of debate. A commonly accepted answer is that connection-oriented networks (aka circuits) are "good for applications that need delay, jitter and/or rate guarantees." To this, we suggest the additional condition that "the network is heavily loaded." As any VoIP user can attest, it works just fine as long as the network is lightly loaded. In lightly loaded (i.e. over-provisioned) networks, the overhead of circuits is not necessary and is in fact wasteful. The goal of this work is to focus the overhead of circuit management only on the specific portions of the network where the load can justify it.

This work combines large-scale measurement, analytical modeling and real-time control to deliver improved resource efficiency and better performance for applications that require moderate amounts of bandwidth. The proposed architecture is dubbed SOCRATES, a Series of Circuits Rapidly AllocaTed for Efficient Sharing.

2 Problem Statement and Background

In this section, we start by posing a question, examine related work for answers to this question, and then formulate our problem statement for this project.

2.1 Preliminary question

The two global switched networks, the telephone network and the Internet, use significantly different techniques to enable millions of users to share communication link bandwidth. On the one hand, the telephone network is circuit-switched, which by definition makes it connection-oriented. By "connection-oriented," we mean that an explicit signaling phase is used to reserve bandwidth on every link of the end-to-end path before users start speaking (exchanging data)[1]. In contrast, the Internet is a connectionless packet-switched network in which there is no bandwidth reservation prior to data transfer. Instead, bandwidth sharing is achieved by TCP senders automatically adjusting their sending rates based on indicators of traffic load.

Of late, there have been at least **two** seemingly **contradicting** developments of notable interest in the networking community. **First**, telephony traffic is rapidly being moved to the Internet making the need for a network that offers end-to-end reservation-based services (as is currently offered by the telephone network) perhaps questionable. **Second**, there seems to be an interest in adding control-plane protocols, such as signaling and routing protocols, to SONET/SDH and WDM optical circuit-switched networks (referred to as GMPLS[2] networks) to enable "bandwidth-on-demand" services. These high-speed optical circuit-switched networks have traditionally only been used to provide leased-line circuit service between IP routers of the Internet or DS0-based circuit switches of the telephone network. Leased-line circuits are typically long-held. In other words, these networks have seldom been considered for providing on-demand short-duration bandwidth-guaranteed connectivity between equipment of end users, such as computers.

Similarly **connection-oriented packet-switched networking** technologies, such as ATM and MPLS[3], have traditionally found usage only in the role of providing leased-line virtual-circuit (VC) services. For example, MPLS is increasingly being used to create Virtual Private Networks (VPNs) for large enterprises. These VPNs interconnect edge IP routers/Ethernet switches of geographically dispersed offices of an enterprise. In addition, control-plane protocols have been implemented in MPLS switches. These include

[1] This "connection-oriented networking" is not to be confused with the "connection-oriented transport-layer" service offered by TCP.

[2] GMPLS: Generalized MultiProtocol Label Switching

[3] ATM: Asynchronous Transfer Mode; MPLS: MultiProtocol Label Switching

both signaling protocols, such as Resource reSerVation Protocol with Traffic Engineering (RSVP-TE), and routing protocols, such as Open Shortest Path First with Traffic Engineering (OSPF-TE).

The networking community appears repeatedly to invest considerable resources to develop, standardize and implement signaling and routing protocols, similar to those used in the telephone network. Before this recent work on signaling and routing protocols for MPLS/GMPLS networks, the ATM Forum created PNNI[4] signaling and routing protocols to enable dynamic "switched virtual circuits" in contrast to "provisioned virtual circuits."

Based on these observations, the **preliminary question** we pose is *"do connection-oriented networks equipped with control-plane protocols that enable on-demand, short-duration calls have value?"* and *"If so, what applications are suitable for such networks?"* Our motivation for pursuing answers is three-fold. **First**, if we find that indeed there is value in such networks, new mass-appeal applications, which exploit the features of these networks, can be created (increasing revenues and promoting economic growth), or existing applications can be handled more efficiently (saving costs, which, in turn, impacts the economy). **Second**, if, on the other hand, we can prove a lack of value for such networks, the results of our work can be used to justify redirection of R&D investments from both industry and government sectors to other more promising technologies. **Third**, we expect our answers to these questions to **inform future network architectures**, an important goal of the NSF NeTS program.

2.2 Related Work

Given the long-term investment made in these control-plane-protocol-equipped circuit/VC networks, we assumed that some significant value has already been identified for such networks. We found two groups of users for these networks: commercial service providers and high-end scientific-research networks. Commercial service providers cite operational-expense savings [35] for the use of control-plane protocols in their MPLS and GMPLS networks. High-end eScience networks use these technologies to meet their high-throughput requirements [1, 2, 3, 4, 54]. What is common in both forms of usage is that **bandwidth is reserved in advance**. We refer to calls that request bandwidth for future pre-specified call-initiation times as **book-ahead calls** in contrast to **immediate-request calls**, which request bandwidth for immediate usage. The book-ahead mode of bandwidth sharing is required when the amount of bandwidth being requested is high (relative to the link capacity) or if the bandwidth is being requested for a long duration. On the other hand, if the amount of bandwidth requested per call is low, the number of "circuits" available on the link is high. Classical loss models show us that under these conditions, high link utilization and low call blocking probabilities are possible. Similarly, when call durations are small, offered traffic load is low, leading to lower call blocking probabilities. In other words, the immediate-request mode of bandwidth sharing is most appropriate when per-call bandwidth is low relative to link capacity and/or call durations are short.

As eScience applications typically require high-bandwidth and provisioning services from commercial service providers are typically for long-duration leases, it is no surprise that book-ahead is the preferred mode of bandwidth sharing. But the RSVP-TE protocol, the control-plane used, only supports immediate-request calls. It has no parameter to carry information required for book-ahead calls, such as future call-initiation time and call duration. RSVP-TE engines built into switches maintain a record of only the *currently available* bandwidth. This has led to the implementation of centralized schedulers in both the eScience [24, 26] and commercial communities to accept book-ahead requests and manage future allocations of bandwidth. The RSVP-TE protocol is used in the provisioning phase allowing for a distributed handling of the procedures needed to stitch together the circuit/VC just prior to a call-initiation time. Thus, these above applications are not fully exploiting the distributed bandwidth-management capability of built-in RSVP-TE and OSPF-TE engines in MPLS and GMPLS switches.

Unlike these commercial leased-line networks or the high-end eScience networks that provide high-bandwidth connectivity between a few organizations, noting Metcalfe's observations [32] on the value of

[4]PNNI: Private Network-to-Network (Node) Interface

a network growing exponentially with the number of users, we are interested in understanding the value of connection-oriented technologies for low- to medium-bandwidth connectivity between millions of end users. Therefore, our interest is in immediate-request, short-duration, medium-bandwidth calls rather than high-bandwidth or long-duration, book-ahead calls[5].

Prior work undertaken by one of the PIs on this subject is a wide-area experimental network funded by the NSF called CHEETAH (Circuit-switched High-speed End-To-End ArcHitecture) [49, 54]. Leveraging the dominance of SONET in MANs/WANs, Ethernet in LANs and the availability of equipment capable of creating Ethernet-mapped-to-SONET circuits, the basic idea in CHEETAH is to create end-to-end, wide-area, medium-bandwidth (100Mbps, 1Gbps) Ethernet "circuits". Further, these Ethernet/SONET hybrid switches come equipped with built-in RSVP-TE and OSPF-TE engines, which allows the CHEETAH-network bandwidth to be shared in a classical telephone-network-like immediate-request short-duration call-by-call mode. The idea behind the choice of SONET in the CHEETAH network was that if successful, it would require a simple control-plane upgrade of already existing SONET switches and Ethernet switches. Ethernet switches offer VC service through their support for the IEEE 802.1q Virtual LAN (VLAN) services. However, we discovered that even with these advantages, a significant problem remains. This problem is the onerous burden of needing *every switch* on the end-to-end path to be upgraded with the control-plane software in order to achieve end-to-end resource reservation.

2.3 Problem statement

The lessons learned from the CHEETAH project led us to examine the question "is a reservation for bandwidth required on every link of an end-to-end path for a particular flow or are Partial-Path Circuits (PPCs) an acceptable alternative?" We emphasize "for a particular flow" because PPCs are already used on most end-to-end paths across the Internet, since many router-to-router links are realized with SONET circuits. However, the significant difference is that these router-to-router SONET circuits carry aggregated traffic from many flows. This means no single flow has a specific amount of bandwidth reserved for itself. In contrast, our concept of a PPC[6] is to reserve bandwidth for a particular flow on part of an end-to-end path.

We combine this notion of PPCs with our preliminary question from Section 2.1 to formulate our problem statement for this project: *"Do connection-oriented networks equipped with control-plane protocols enabling on-demand, short-duration calls have value when bandwidth is reserved on only portions of an end-to-end path rather than on every link? If so, what are suitable applications for such a network design?"*

Restating our problem statement in this way demonstrates that practical implementation is one of our important goals. If there is value only for end-to-end reservations on every link, we will conclude that the costs may be prohibitive for connection-oriented networking to catch on to the level at which the type of value noted by Metcalfe becomes significant. If we find value even for PPCs, then it allows for a gradual growth of connection-oriented networking making it much more feasible for deployment.

3 Preliminary Work

Given our experimental orientation, we first address the question of whether this concept of reserving bandwidth on one or more partial segments of an end-to-end path is feasible with existing switches. We then address the question of when and where PPCs would be beneficial.

3.1 Feasibility: Can PPCs be set up for individual flows?

The current Internet is dominated with IP routers and Ethernet switches that typically operate in connectionless mode. However, in recent years, both these types of equipment have been upgraded to include

[5] An important task of the proposed research is to develop quantitative bounds for what is meant by "short-duration" and "medium-bandwidth."

[6] We use the term PPC generically for a Parital-Path Circuit. In practice, the PPC could be implemented as a physical circuit or a virtual circuit (VC).

connection-oriented networking capabilities. IP routers have built-in MPLS switches and RSVP-TE/OSPF-TE control-plane capabilities. Ethernet switches have IEEE 802.1q Virtual LAN (VLAN) capabilities that can be combined with external control-plane engines, as has been done by the NSF-funded DRAGON project [3], for a connection-oriented mode. IP routers and Ethernet switches currently deployed in WANs and LANs have these upgraded capabilities. Hence, the key technological components needed to "peel off bandwidth" with a reservation on a part of an end-to-end segment, even on just one link, are currently available.

Partial-path VCs were attempted in IP-over-ATM efforts of the nineties. The thinking was that automatic flow classification techniques at IP routers could be used to redirect packets from long-lived flows to dynamically setup ATM VCs [19, 34]. The ATM VC would thus extend across only part of the end-to-end path. However, prediction of flow length is difficult, and hence automatic flow-length detection and flow redirection to ATM networks were not realized in practice.

Several analogies for PPCs exist in the transportation world. A typical airline traveler uses the "datagram" roadway system to get to an airport, but occupies a reserved seat "virtual-circuit" on a flight, before traveling across another "datagram" set of roads to reach the destination. In this analogy, we noticed that the *airline traveler is the one who makes the explicit reservation* on the partial segment of the end-to-end journey. An analogous solution would be to have *an application* running on a generic Internet-connected end host *send an RSVP-TE signaling request for bandwidth* to some intermediate switch to make a reservation on a segment of the end-to-end path, before it starts sending IP packets on a specific TCP or UDP flow. Although router vendors were not successful in developing good flow-length detection algorithms, they did design mechanisms for filtering out packets from one or more flows and directing them to specific MPLS VCs. This feature is called Policy-Based Routing (PBR) and is the technique exploited in both the BRUW and OSCARS schedulers [26, 24]. While we are pursuing a different sharing mode from the BRUW and OSCARS projects (i.e., immediate-request, short-duration calls), their usage of this PBR feature provides us a key starting point for how to create PPCs and direct a specific flow at some intermediate router onto a PPC.

This also addresses another oft-quoted "drawback" of IntServ [10], which is that the idea of setting up reservations per-flow is unscalable. Indeed it would be so if a reservation was needed for each and every flow. Clearly for short-lived flows, a reservation would represent an unnecessary overhead. Since requests are explicitly generated by end-user applications, presumably they would only be issued for flows that require reserved bandwidth. As with CHEETAH, our approach is to use PPCs only when required, and default instead, for many applications, to the connectionless IP service.

3.2 Where would a PPC be required?

Ask any networking researcher familiar with connection-oriented networking the question "when is connection-oriented networking useful?" and a typical response is "when quality-of-service (QoS) guarantees are required." For example, in a telephone call, 150ms is cited as a requirement for one-way end-to-end delay for a user to perceive "excellent-quality" service. However, earlier in Section 2.1 we noted that telephony is rapidly being moved to the connectionless (reservationless) Internet. So how is this possible? Our intuition is that the "QoS" answer needs to be qualified with an AND clause, which is "links are heavily loaded." If all the links on an end-to-end path are lightly loaded (relative to a single call), then not only would a low-bandwidth VOIP call be handled easily, but even a low-compression, high-quality, interactive-video call (as in video-telephony and video-conferencing applications) may not require an explicit bandwidth reservation. This gives us an insight that a PPC is perhaps most useful on heavily loaded links.

We also consider applications that do not have an explicit QoS requirement, such as file transfers, the dominant application on the Internet. Is the familiar answer that CO networking is useful only when QoS guarantees are required imply that it has no value for file transfers? What about its potential benefit for improving the overall efficiency of the link? As noted in [13], packet switching has an efficiency advantage under light loads while circuit switching has the advantage under heavy loads. Consider the use of TCP for

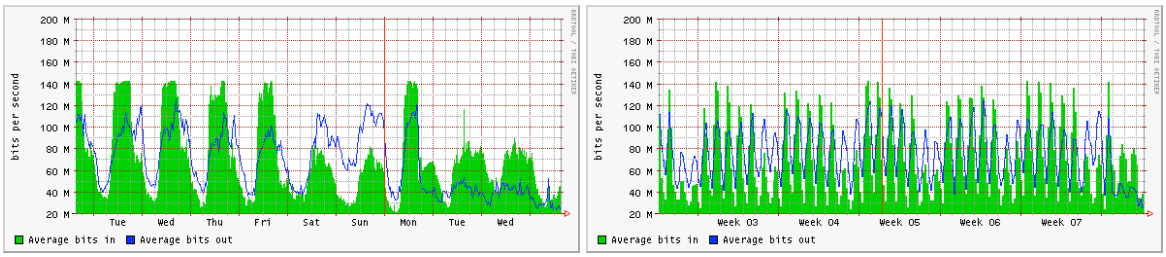

Figure 1: Traffic for UVA OC3 link depicting regular cycles of loaded/unloaded periods

bandwidth sharing in connectionless packet-switched networks. Under heavy loads, either losses are incurred which cause retransmissions, or rate reductions may be over-aggressive causing momentary underutilization of links. Prior partitioning of bandwidth to circuits/VCs under heavy loads could therefore lead to lower delays and better utilization. On the other hand, on lightly loaded links, partitioning off a small amount of bandwidth for a particular flow will lead to longer transfer delays than if that flow was allowed to run freely and enjoy as much of the link bandwidth as possible. Our proposed work includes **simulation and analytical studies** to gain a quantitative understanding of the advantages and disadvantages of these two bandwidth sharing techniques under light and heavy loads.

If we find that indeed connection-oriented networking is more efficient for file transfers under heavy loads, confirming Modiano's findings [13], then it would seem that the answer to *"when is connection-oriented networking useful?"* may be *"when quality-of- service (QoS) guarantees are required* **AND/OR** *when links are heavily loaded."*

Thus for both applications with QoS requirements and those without (i.e., file transfers), an answer to the question of "where would a PPC be required," seems to be "on heavily loaded links." We note a direct analogy of reserving bandwidth on heavily loaded links to the growing use of reserved HOV lanes or Toll-based express lanes in some metropolitan areas. In these cases, a commuter does not reserve the entire path for a journey but instead has privileged access to a less congested lane that is allocated through only the most congested areas [48]. For roadways that are not congested or for time periods that are not congested, the enhanced lane is not necessary and the driver will likely not benefit from using it. This analogy matches our approach both in the focus on congested links and times and also on the possible incentive (billing) model that might be applied to this approach.

3.3 Analysis of heavily loaded links

As an example, we look at the current network for University of Virginia. The University of Virginia (UVA) campus LAN consists almost entirely of Gigabit Ethernet switches with lightly loaded links. Similarly, the traffic weather maps of the Abilene connector and backbone links routinely show moderate loads since the upgrade to OC192 (2.5 Gbps) [47]. The UVA campus is however connected to Abilene via an OC3 (155 Mbps) link, which is heavily loaded during workday hours (See Figure 1).

For users accessing the Internet from the UVA network, the loading is highly variable along the end-to-end path, with some links lightly loaded and others heavily loaded. This suggests that perhaps it would be sufficient to reserve bandwidth for specific communication sessions only on the heavily loaded links. Further, it may not be necessary to reserve capacity for every call since there are many periods where even this bottleneck link is not congested.

To obtain an idea of the impact of UVA's congested link, we conducted several measurement tests using Iperf [46] and other TCP-based applications. On a 20ms round-trip time path between hosts with Fast Ethernet NICs (i.e., the bottleneck link was the NIC card rather than the UVA OC3), we found that the throughput leveled off at 70Mbps for transfer sizes larger than 40MB in off-peak times. During peak weekday hours, throughput dropped to 7Mbps, a factor of 10 reduction.

These experiments provide us some insights. **First**, for calls with QoS requirements, such as a 10Mbps motion-JPEG video call, measurements such as that shown in Figure 1 be used to indicate that the user is better off setting up a PPC across the loaded OC3 link prior to data transfer to keep losses at an unacceptable level. But what about a 128kbps VoIP call? Would that require a reservation? Clearly the decision of whether to peel off bandwidth for a PPC depends on the amount of bandwidth required.

Second, for file transfers, there is no pre-specified bandwidth amount required for a flow. It is clear from the experimental results that even if a link is saturated, a new flow is able to obtain some level of bandwidth at the expense of other flows. The amount of bandwidth that a TCP flow can acquire depends on its RTT and the RTTs of all other ongoing flows. With newer TCP variants RTT dependence will be lower [60]. Nevertheless, to determine whether a PPC is beneficial for a file transfer, we must estimate the amount of bandwidth the file transfer can obtain on the congested link and compare this to the bandwidth that could be allocated to it in a PPC.

3.4 Pricing

As with the analogous HOV or Toll-lane, the PPC concept includes an incentive notion based on pricing. The assumption is that the improved performance would be available to users at some additional cost beyond the best-effort, datagram Service Level Agreement (SLA).

Efficient pricing will be required to answer the question of how much bandwidth to allocate to a PPC on a congested link for a file transfer. In general, connection-oriented networking appears to be better suited for per-flow pricing models since the call setup & release messages offer an opportunity to collect flow related data for billing purposes. While we intend to consider this aspect in the SOCRATES work, the precise pricing model is an area of significant future study outside of the proposed effort.

4 Research Challenges

Our preliminary work has identified many interesting problems to be addressed in this project. We describe a few of these here.

- **Identification of congested links:** While identifying a congested link is a straightforward monitoring task, we find that it is less important whether a link is congested than whether it has the **potential to be a congestion point** for a given flow. A definition of a *congestion-potential link* in our context is a link with the following property: the probability that the QoS available on the link, *if it were not reserved for the call*, falls below the required levels at some time during the call, exceeds some probability-threshold value (say 0.001). In other words, a link is a congestion-potential link if it is likely to violate the call QoS requirements. We start by limiting QoS to just bandwidth.

  Clearly algorithms to determine whether a link is a congestion-potential link will be challenging to develop. This is especially true for file transfers, where even in the case of a heavily loaded link, a new file transfer session would in fact obtain some bandwidth if added to the offered load. The question is, how much would it obtain given the already present application mix and would it be a better choice to specifically reserve (and pay for) capacity.
- **Applications:** For what application profiles (e.g., large file transfers, voice/video calls) is the PPC most beneficial? To what degree can applicability be determined in advance?
- **CL vs CO capacity:** How much bandwidth should be set aside for CL services so that as bandwidth is peeled off for PPCs, the best-effort traffic still enjoys some level of service. The determination of this value will include both the performance gain for the priority application, the overall efficiency of the link as well as the pricing model for the link provider. We refer to this as the *PPC-bandwidth threshold*.
- **Call duration:** The above question brings us to the issue of fairness. For how long should a particular flow be allocated bandwidth on a PPC? Should there be a maximum holding-time limit for improved fairness? What about for file transfers? Is it better to specify a Maximum File Transfer Size (MFTS) corresponding

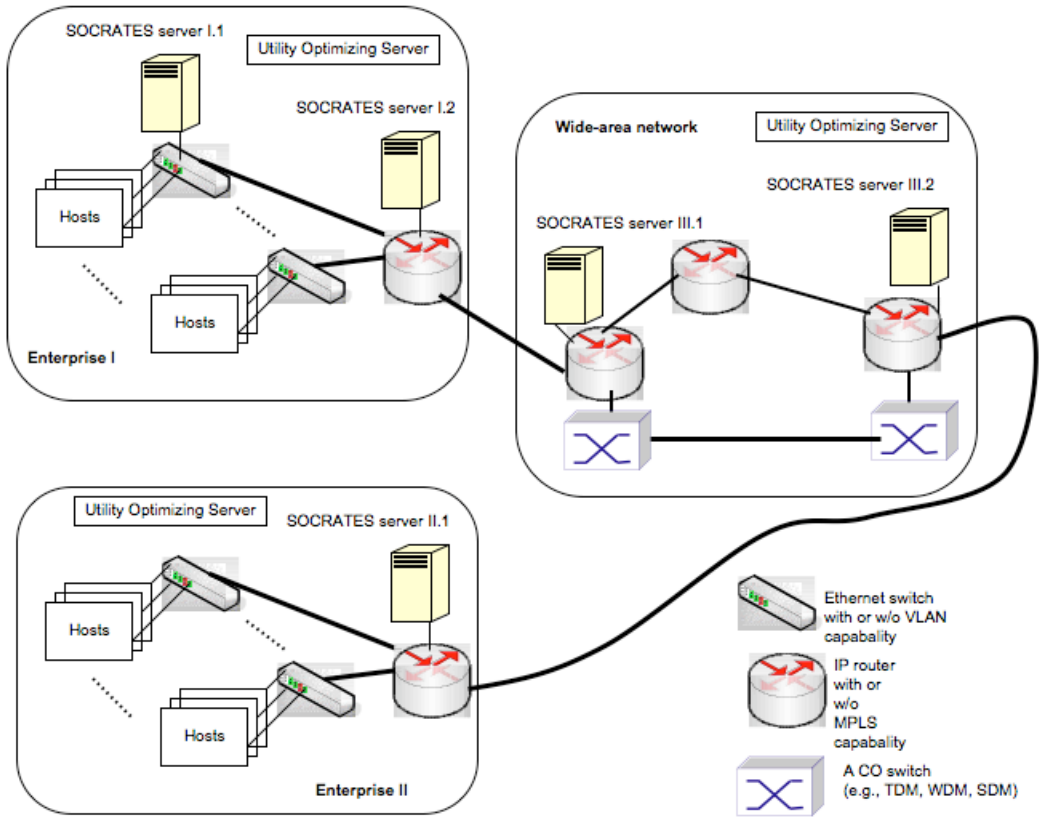

Figure 2: An example SOCRATES architecture with two enterprise networks connected to a WAN

to Maximum Transmission Unit (MTU) in packet-switched networks? If the network provides a low-bandwidth PPC for a particular file transfer, it can adjust the maximum call duration based on the MFTS value.

- **Security:** A major concern of network providers to allow such immediate-requests for bandwidth is security. With applications, running RSVP-TE clients that send in requests for bandwidth unbeknownst to users, security must be addressed. What are good solutions for this problem that do not slow down call setup to the point where PPCs make sense only if the call durations are long (e.g., hours)?

These research challenges will be explored through data analysis from our proposed measurement work. We will extend loss models for some of the circuit network analyses using Markov modeling techniques. A combination of simulations and analysis is required to answer these questions. We plan to develop a large scale simulation using GTNets [37, 38] to look at these issues.

5 Research Approach

5.1 Architecture

The proposed SOCRATES architecture provides for the deployment of a number of *SOCRATES servers* which perform the measurement, analysis and control functions for PPCs. A possible deployment scenario is depicted in Figure 2.

In most enterprises, SOCRATES servers are likely to be required only at the WAN access router as shown in Enterprise II of Figure 2. However, for generality, we show that it could be placed even at an Ethernet switch as in Enterprise I of Figure 2. Within WANs, SOCRATES servers can be associated with any router

that has links that could potentially become congested. For example, if a router has an OC192 interface that is far from being loaded (e.g., some of the Abilene links [47]), then a SOCRATES server is not required for this interface. Links could be shared between connectionless (CL) and connection-oriented (CO) modes or a separate CO network could be available as shown in the example WAN of Figure 2. IP routers equipped with MPLS engines effectively have links that can be shared between the CL and CO modes.
A SOCRATES server performs the following functions:

- *Measurement processing:* process the measurement data (see Section 5.2 to derive metrics useful to the congestion-potential determination process,
- *Congestion-potential determination:* perform the analysis discussed in Section 5.3,
- *RSVP signaling:* parses and constructs RSVP messages with the additional SOCRATES parameters,
- *Router interface:* interfaces with the IP routers using CLI to trigger PPC setup/release, map particular flows to PPCs, set PPC-bandwidth threshold, read Netflow data and SNMP MIB data,
- *Neighbor SOCRATES server discovery process:* performs auto-discovery of peer SOCRATES servers,
- *PPC-triggering process:* generates the RSVP-TE messages to the external CO networks if such connectivity is available (e.g., in the WAN in Figure 2)

Note that SOCRATES servers do not perform any bandwidth management functions. They only receive and parse RSVP messages to extract destination, source and QoS parameters, such as bandwidth. Bandwidth management is strictly performed by the RSVP-TE engines running at the MPLS/GMPLS switches.

The Utility Optimizing Server (UOS) is used to set PPC-bandwidth threshold parameters on the links that offer both CO and CL services. In practice, we plan to exploit the label stacking feature of MPLS to implement this threshold. An MPLS VC with a bandwidth equal to the PPC-bandwidth threshold is set on interfaces that participate in the SOCRATES architecture. This PPC can then be configured as an "interface" at the router allowing for inner MPLS VCs to be set within this outer VC. Algorithms run at the UOS are domain-wide to optimize utility across the network. Details are described in Section 5.3.

5.2 Measurement

Measurements are key in our proposed approach. In this research area, we design and implement a large-scale measurement and monitoring infrastructure.

5.2.1 Monitoring and Measurement Infrastructure for Large-Scale Network Modeling

The SOCRATES project will extend and leverage existing and emerging measurement projects. We expect to gather and use data from the following sources.

- The IEPM-BW toolkit, developed at SLAC, is currently deployed at monitoring hosts at about 40 sites around the world, including major measurement hosts at SLAC, CERN, FNAL, BNL and Caltech. These measurement hosts run active end-to-end light-weight measurement tools, such as ping, traceroute, pathchirp [36] and pathload [27], and heavy-weight measurement tools, such as thrulay [43], iperf [46], and GridFTP [6] at regular intervals. The light-weight, more frequent measurements, will be used to assist in interpolating the less frequent, more heavy-weight measurements. The type of data collected by these measurement tools includes round-trip-time, hop-by-hop router response, capacity and available bandwidth, achievable throughput and file transfer rates. We also plan to evaluate the effectiveness of the pathNeck tool [25] and if successful integrate it into the IEPM-BW monitoring suite and develop algorithms to detect congestion points across the network.
- A second component of the monitoring infrastructure will be from core routers and switches which form the backbone of the Academic and Energy Science Networks around the USA (and in Europe). The Abilene Measurement Infrastructure (AMI) and the perfSONAR projects will provide router interface utilization and capacity data using standardized schemas and web service facilities. Data from these projects is already publically available.

- A third source of measured data from existing networks is Netflow [44] passive measurement data that can be obtained from select routers, in particular, border routers at collaborating sites. The Netflow records from a given router will be collected by a host co-located in the AS of the router. This host will suitably anonymize and select relevant records (e.g. long lived flows, selected ports/applications, etc.) and make them available. A goal of these measurements is to obtain data on the start times, transfer sizes, end times and characterizations (e.g. top talkers, transfer rates, arrival rates and call holding times) of long-lived flows and Real Time Protocol (RTP) [42] flows. Since dynamic call-by-call sharing services are not used in existing networks (enterprise networks, Internet2, ESnet, etc.), we have no way of estimating the potential size of connection-oriented traffic. As a crude model, we propose using flow data to gather statistics on long-lived flows for file transfer applications and RTP flows because these are the most likely candidates for connection-oriented service.

The host machines running the application software will also be utilized to provide a subset of the IEPM-BW measurements. The data obtained will be uploaded to the IEPM-BW hosts and added to the available pool of data.

To provide uniform access to all network performance related data, all the data assembled as part of the SOCRATES measurement work will be served in the web services format used by the Global Grid Forum (GGF) Network Monitoring Working Group (NMWG) and additional recommendations resulting from the emerging global measurement infrastructure. This enables us to leverage existing technologies whilst federating data access such that it will be scalable to current and future best practice methods. Synergistic collaboration and development of numerous components to aid federation of data is expected. In particular, discovery mechanisms to find, gather and analyze relevant 3rd party measurements from measurement infrastructures such as NLANR/AMP, and MonALISA will be incorporated.

5.2.2 Utilization of Measurement Data

The relevant data collected from the various monitoring and measurement hosts will be used as input to our large-scale parallel simulator. The purpose of this simulation is to evaluate quantitatively the benefits and costs of offering CO service on the communication link infrastructure of today's datagram networks.

The AMI/perfSONAR projects will provide performance measurements at the backbone and edge router interface level in a multi-domain environment initially including Abilene, GÉANT and ESnet. It will not tell us what is happening at the end-sites or through other links than through Abilene, GÉANT or ESnet routers, nor will it tell us how the applications will perform (both of these issues are addressed below). This latter feature will be provided through the active IEPM-BW tests and NetFlow passive monitoring results.

SOCRATES servers at each hop in a network can gather and quickly provide the link utilization for its interfaces, and provide current congestion information for each interface/link. For links with no SOCRATES but an AMI/perfSONAR service we will provide a proxy SOCRATES. For non perfSONAR links (especially at the end sites), we will coordinate with perfSONAR and the sites and assist in deploying tools like AMI/perfSONAR to add performance measurements for those links too. This may be problematic due to security and administrative concerns, especially in commercial networks. For such cases we will explore the applicability of deploying IEPM-BW and using tools such as pathneck to provide bandwidth estimates at the various hops.

Though our main focus is not directly concerned with measuring how applications perform, this is the window through which the user perceives network performance. We envisage launching an application as a multi-step process, e.g. as a preliminary step the user/application will look to see (using end-to-end forecasts related to the application and provided for example by IEPM-BW) whether there is a suspected problem for their application. If so the application/user will then move forward and request improved SOCRATES type services to meet the need, otherwise the application will run normally using best-effort.

The Netflow data will be mined for flows related to specific applications on particular paths, and if there are enough relevant flows then forecasts can be made. The accuracy of the forecasts will depend on the

periodicity/seasonality of the data (e.g. if the links are well provisioned then there will be little change in behavior with time so a few measurements can enable a forecast for a long time). It is possible to detect the use of multiple parallel streams (the flow records all start at the same time within some window, between the same sce/dst host to the same port) which is valuable for some applications such as parallel FTP. The Netflow records will also be valuable for billing and for the characterization of networks for simulation. Getting the raw Netflow records however will be problematic in many cases due to privacy. As necessary, the data will be privatized, however this reduces the value (e.g. one does not know the application or sce/dst).

5.2.3 Measurement servers for the proof-of-concept testbed

The SOCRATES architecture depicted in Figure 2 shows how measurement servers should be deployed within ASs. The purpose of these measurement servers is to collect data on both CO and CL traffic. First, we will design and implement a data-gathering component of these measurement servers to collect data on dynamically requested calls from the RSVP-TE Management Information Base (MIBs) located at routers/switches [7]. The collection of CL traffic information is well developed and understood; current utilization and link capacity can be derived from router/switch interface measurements stored in interface SNMP MIB variables. However, the availability and validity of information provided for CO traffic is less clear and we are currently exploring the extent to which the MIBs have been defined and whether network switch vendors have implemented these MIBs. If these are not available, we will implement snooping solutions to capture RSVP-TE messages in/out of switch control cards to gather this data. Since dynamic call arrivals are not expected to occur at very high speeds (Ethernet interfaces to switch control cards over which RSVP-TE messages are transported are typically 100Mbps) we think this approach is feasible.

These servers are necessary to support our laboratory proof-of-concept testbed. As we test application software programs that generate RSVP-TE messages, these servers will capture those messages. Plans are to run some of the monitoring hosts on this testbed to capture data on applications that we execute on the testbed.

5.2.4 Forecasting

An important feature of specifying whether a particular hop is capable of the dynamic CO traffic is whether at any time during the transfer will the interface experience congestion. For example, if we were to set up a very long file transfer and a particular interface is currently un-congested, then it may be determined that the link does not require a CO path for this hop. However, say predictable competing transfer occurs (e.g. site A replicates a large catalogue to site B every day at midnight), then this competing traffic will steal bandwidth from our CL flow (on this hop). Therefore it is important to be able to forecast that throughout the duration of the transfer, congestion is or is not expected.

As such, in parallel with the development of the modeling and testbed, performance data from IEPM-BW, AMI and perfSONAR measurement servers located in real production networks will be used to evaluate and develop short and long-term (hours to days) forecasting techniques for predicting bottleneck magnitude and location. The forecasts will take into account seasonal patterns and long term trends in the data and will build on the existing work by SLAC in this area [20]. These forecasts, including confidence levels, and will eventually be used by the SOCRATES servers.

The forecasting will initially be based on the Holt-Winters [12] triple Exponential Weighted Moving Averages (EWMA) technique for time series that exhibit short term variations, long term trends and seasonal changes. This technique will be applied to the various time-series of active and passive measurements. We will also evaluate other techniques, such as Principal Component Analysis, wavelets, neural networks, and hidden-Markov models (described in the next section).

5.3 Modeling

The approach of the SOCRATES project is to allow end host applications to request and reserve bandwidth only on the congested links, rather than on every link in the end-to-end route. However, congestion

points may change over time so bandwidth allocation will be driven by measurements.

Even as network administrators (humans) consider providing MPLS tunnel services to end users on a provisioned (requested through a web site) basis, they will likely impose an arbitrary maximum on the amount of connection-oriented (CO) traffic permitted on their router interfaces. (i.e., a maximum on the total amount of bandwidth available for allocation to MPLS tunnels). The SOCRATES project participants will develop sound algorithms to determine this value based on the mathematical principal of maximizing utility. Our proposed research builds on our extensive previous work on related resource management [5, 8, 9, 16, 17, 21, 28, 29, 33, 45, 50, 51, 52, 53, 55, 56, 57, 58].

5.3.1 The PPC-Bandwidth Threshold

The PPC-bandwidth threshold parameter is set on every link on which both CO and CL services are supported. It is the maximum value up to which bandwidth can be assigned to connections as bandwidth-reservation requests are received by signaling engines at routers/switches. Note that while this threshold limits bandwidth available to CO services, if there are no such CO requests, the entire link is available to CL traffic.

We will develop algorithms to calculate this PPC-bandwidth threshold parameter, so that the aggregate utility over time is maximized. These algorithms use as input data measurements of the CL and CO traffic over time. These algorithms will be implemented in the UOS described in Section 5.1. The PPC-bandwidth threshold value may vary from one link to another and from one time epoch to another.

As described in Section 5.1, the PPC-bandwidth threshold values determined for each link with both CO/CL service will be communicated to the RSVP-TE control-plane engines running at the routers/switches. As these control-plane engines receive requests for bandwidth via the RSVP-TE protocol, requests will be fulfilled as long as the PPC-bandwidth threshold is not exceeded. If the threshold value is reached, a higher priority reservation can replace a lower priority one. If a reservation cannot be met at all points along the path (if congested links are consecutive, the partial connection setup triggered by the PPC-Triggering Process (see Section 5.1) could lead to a multi-hop partial connection), the request is denied and the requestor receives a *call failed* error message.[7]

5.3.2 Utility

The control of resources in the SOCRATES project is based on the mathematical concept of utility (an old concept in economics, but currently used extensively in network resource control problems). A host or application completing service on the network receives a certain degree of "satisfaction," usually called *utility*. Typically, the utility depends on the amount of resources consumed (usually a convex function). The utility is gained only if the call completes as desired and is not preempted. The algorithms are designed to ensure maximum aggregate utility over time. Because network services compete for fixed resources, the utility values experienced by different services have to be traded off to maximize the overall (aggregate) utility. This kind of tradeoff is a common approach to addressing **fairness**. The utility formulation also yields naturally to pricing models through an application of duality theory [16].

We assume that utility is a function of bandwidth. For example, the utility function for a session with bandwidth x may be linear in $x : U_S(x) = ax$. Alternatively, a concave log utility function is also common: $U_S(x) = a\log(x+b)$. Since we are trading off CL bandwidth for CO bandwidth, we will also use utility functions for the CL traffic. We expect CL flows to be "elastic" (in the sense that they can tolerate a variety of bandwidths), yielding to analysis based on utility functions. Indeed, dominant traffic types on the Internet—web traffic, email, and file transfers—are all elastic in this sense. We assume for simplicity that all the CL flows have the same utility function (denoted u), that the total bandwidth x for CL flows is divided equally among them, and that there are n CL flows. Then, the aggregate utility (per unit time) for the CL traffic

[7] An important point about "call-failure" in the context of this work is that it is only the reserved capacity allocation that fails. The application could still proceed using standard IP datagram (CL) service.

is $U_{CL}(x) = nu(x/n)$. If u is linear, then the dependence on n disappears: $U_{CL}(x) = u(x)$. In this case, the entirety of CL flows can be treated as a single "virtual" flow with utility function u. It is reasonable to assume that u is linear because of the elastic nature of CL service—we will make this assumption, at least initially, in our approach. In the general, U_{CL} may depend on n, making it necessary to measure (or estimate) the number of ongoing CL flows.

The simple model of aggregate utility above will be refined as part of our proposed research to more accurately reflect the actual amount of bandwidth that a new CL flow would receive on an existing link that already serves existing CL traffic. This refinement involves analytical and measurement-based models. Indeed, as pointed out in Section 5.2, part of our measurement efforts will address the issue of estimating the bandwidth of a new CL flow on a loaded link.

5.3.3 Markov decision theoretic formulation

It is often the case that our goal involves maximizing aggregate utility over a large time horizon, taking into account the random variations in system behavior over time. This is the case in dynamic situations, where requests arrive continually over time. Such resource allocation problems can be posed in great generality under the framework of *Markov decision theory* [15, 31, 41]. The problem here actually yields to a formulation based on a special case called a *partially observable Markov reward process (POMRP)*.

Recall that our problem is to set and update the PPC-bandwidth threshold parameter. The updates can be done, for example, once every hour. The factors on which we base the threshold parameter values are time of day and prior history of requests (e.g., from request logs and network measurement). The expected aggregate utility over one hour is the criterion that drives our optimization of this parameter value. The expectation depends on how the probabilistic features of the system evolve over time, which is captured by a *hidden Markov model*. Specifically, the evolution of the system over time is due to the random nature of the calls in the system (request arrivals, duration of ongoing calls, etc.).

In our problem, a hidden Markov model is used to model the following probabilistic components of the system: the arrivals of CO requests over one hour, including all call-specification parameters (start time, bandwidth, etc.), and the starts/completions of CL flows over the hour. The underlying Markov state captures random changes in the behavior of the requests and flows over time. We specifically include the time-of-day as a component of the state, because we expect that call requests and flows are modulated naturally by the time-of-day (e.g., fewer call requests at 1am than at 9am). The observation model captures the factors that are available to us for decision-making.

The hidden Markov model for our system can be obtained from a variety of methods to "train" or "learn" from empirical data, including the well-known EM algorithm [22]. Data, such as request arrivals and CL traffic over time, is collected and used as input for the training algorithms. The model can be updated from time to time to adapt to changing conditions over time.

To simplify the training process, we could impose some approximating assumptions to the model. First, we could limit the size of the state space. Second, we could assume some structure on the form of the observation model. For example, the request arrival distribution given an underlying state could be modeled as Poisson, so that the training involves only the fitting of a single parameter, the Poisson rate. Similarly, we could model the call duration as a truncated Pareto, again with one parameter to train. Finally, the simplifying assumption that the arrival process is conditionally independent given the state allows us to factor the arrival and duration distributions. The training and updating of the hidden Markov model from empirical data is a nontrivial task, and constitutes one component of our efforts.

Given the hidden Markov model, the expected aggregate utility resulting from any PPC-bandwidth threshold setting applied over the next hour can be calculated. This is the basis for searching over the space of threshold values to find the one that maximizes the aggregate utility. To do this we apply an optimization algorithm (e.g., [18]).

The main reason our formulation here yields to the special case of a Markov *reward* process (rather than

the more general Markov *decision* process) is that we have implicitly assumed that the service durations are short relative to the time between threshold updates. In the more general situation where many calls will span multiple update epochs, we will need to consider a more full-blown partially observable Markov decision process (POMDP). It turns out that the threshold optimization procedure in this case is similar to what we have described so far for the simpler POMRP formulation, with some modification to the objective function being optimized at each decision epoch. To elaborate briefly, we would need to augment our expected aggregate utility over one hour to include an expectation of utility over some horizon into the *future* (the extent of this horizon depends on the duration of calls). This augmentation forms what is called a Q-function. Other than this modification, the procedure described in this section applies. We have had great success recently in developing and applying approximation methods to Markov decision problems under reasonable constraints on computational burden [14, 15, 17, 23, 29, 30, 39, 40, 55, 56, 57, 58, 59].

We will implement the algorithms described above, not only for use in our laboratory test-bed but more importantly in our parallel network simulator. Since the monitoring and measurement infrastructure described in Section 5.2 will feed real live data directly into the simulator, the algorithms can execute in this environment on real data, extract the hidden Markov models and compute the PPC-bandwidth threshold values. These will be communicated to our simulated routers/switches.

5.3.4 Congestion Identification

As pointed out before, the determination of which links are congested is critical to the SOCRATES solution. Further, as previously noted, a sensible definition of "congestion" in our context (with respect to a particular call request) is a link with the following property: the probability that the bandwidth available on the link—*if it were not reserved for the call*—falls below the required bandwidth at some time during the call, exceeds some probability-threshold value (say 0.001). In other words, a link is congested if it is likely to violate the call bandwidth requirement. Based on this definition, it makes sense that we have to set up a PPC reservation over congested links, and *only* over such links.

We will develop algorithms to determine congested links based on forecasts of available bandwidth over the duration of the call request on each potential congestion point. This forecasting (prediction) is based on a measurement-driven model, and will be provided by the Measurement servers as described in Section 5.2. We will develop models to determine if the probability of "bandwidth-requirement violation" exceeds the preset probability-threshold value.

5.4 Experimental Plan

An important aspect of the SOCRATES research plan is to extend the project beyond the laboratory to a wide-area test environment. The HOPI testbed was built to encourage experimentation into the use of both packet- and circuit-switched networks. The main networking nodes of the HOPI testbed are GMPLS-enabled Ethernet switches with IEEE 802.1q virtual LAN (VLAN) and 802.1p QoS capabilities. These switches connect to IP routers on the Abilene network, which makes it an ideal testbed for our SOCRATES architecture. We are currently planning to connect the CHEETAH testbed into HOPI, which will allow for interesting heterogeneous CO testing (given CHEETAH is SONET based).

We plan to use a combination of Abilene and the HOPI testbed to test the SOCRATES architecture. Given that the loading on Abilene backbone links is currently light, we plan to provision 100Mbps MPLS virtual circuits between strategically selected Abilene IP routers, and limit traffic routed to these VCs to datagrams sourced from and destined to the SOCRATES-experimental subnets (which will be located at the four project institutions). For example, traffic between SOCRATES hosts at UVA and G.Tech will be routed on the provisioned VC between Washington, DC, and the Atlanta routers, while traffic between SOCRATES hosts at CSU and SLAC will be routed on a VC between the Sunnyvale and Denver routers. We plan to run multiple flows between pairs of SOCRATES hosts so that the virtual-circuit links allocated to our project on Abilene backbone links become highly loaded. This would allow the measurements and PPC-determination

algorithm to require bypassing the CL path with a VC (PPC) setup on-demand through the HOPI path. The team members from Internet 2 (Summerhill/Riddle) will be heading up this aspect of the project.

5.4.1 Demonstration Applications

We intend to demonstrate the SOCRATES architecture through the development of several applications that will be integrated with the measurement and PPC setup services through a well-defined API. Demonstrations will include a file-distribution application based on the need for large file updates from the software distribution servers currently operated at Georgia Tech. Voice and video telephony applications will be applied through our ongoing work in high quality remote conferencing services with DVTS. Gaming applications will be explored as well with a special emphasis on the Undergraduate Research Competition (UROC) [11] at Georgia Tech.

Research work in the applications area will consist of (1) determining the exact networking needs of these applications, and (2) designing and implementing software modules that can be integrated into the application tools. These software modules will include control-plane functions, such as the ability to request partial or end-to-end connections for immediate usage, as well as data-plane functions, such as maintaining fixed sending rates to match reserved bandwidth on the connections.

6 Management Plan

6.1 Coordination

We have assembled a strong team of investigators that is uniquely qualified to carry out the planned work. Russell Clark will manage the project and work on the measurement testing and applications interface. Warren Matthews will work on measurement data services and software. Les Cottrell will work extensively on the measurement aspects of the project and enhance the measurement tools. Edwin Chong will lead the analysis and algorithm design efforts. Malathi Veeraraghavan will provide network control plane expertise, focus on the design and simulation of new control protocols and lead the IETF standardization activities. Bob Riddle will provide support by facilitating research using the testbed and other components from the Internet2 Abilene network. Matt Zekauskas will work on measurement and Rick Summerhill will coordinate HOPI integration efforts.

The group will coordinate this project with regular email exchanges and twice-monthly conference calls. A Wiki web site has been created at GT for sharing documents and collaborative space for use by the team. The project is tightly integrated and the interaction will involve working group coordination in addition to individual updates, sharing research results and planning publications. As tools are developed, a web site will be available with detailed descriptions of all aspects of the project and information on using the tools. The group will meet face-to-face at conferences, I2 meetings and at least one dedicated meeting per year is budgeted. In addition, the group will coordinate with outreach to user communities not directly supported by the project but who will benefit from the results.

6.2 WorkPlan

Our workplan consists of four tracks:

- **Track I:** Large-scale parallel network modeling/simulation with real-live measurement feeds from production networks
- **Track II:** Implementing a laboratory testbed
- **Track III:** Wide-area testbed deployment using Internet 2 and HOPI networks
- **Track IV:** Input to standardization efforts

The following table outlines the tracks, individual tasks and key researcher. The years identify the primary year in which each task will be performed.

Track	Milestones and Deliverables	Date	Responsible PI
I	Algorithms for Congestion Detection	Year 1	Chong
I	Addition of Link Utilization Data	Year 1	Cottrell/Zekauskas
I	Forecasting	Year 1	Cottrell
II	Overall Socrates Software Architecture	Year 1	Veeraraghavan/Clark/Matthews
II	Web Services (PerfSONAR) Front End to Measurement Data	Year 1	Cottrell
III	Installation of Socrates Servers	Year 1	Matthews
III	Configuration of MPLS tunnels	Year 1	Riddle
I	Algorithms for Utility Optimization	Year 2	Chong
I	Simulation Studies	Year 2	Veeraraghavan
I	Characterization of Long Lived Flows	Year 2	Cottrell
II	Implementation of Measurement and Congestion Detection Processes	Year 2	Matthews
II	Implementation of Utilization Optimization	Year 2	Matthews
II	Implementation of Socrates RSVP process	Year 2	Veeraraghavan
II	Implementation of Socrates PPC-triggering Process	Year 2	Summerhill
III	Socrates Neighbor Discovery	Year 2	Matthews
III	Testbed Server	Year 2	Clark
I	Refinement and Evaluation of Algorithms and Models	Year 3	Chong
I	Federated Data Access	Year 3	Cottrell
II	Applications	Year 3	Clark
II	Develop a subset of IEPM tools for Hosts	Year 3	Cottrell
III	Phase I Wide-area Experiments	Year 3	Clark/Riddle
III	Deploy Socrates processes	Year 3	Matthews
III	Cross Domain Cheetah-HOPI test	Year 3	Veeraraghavan/Summerhill
I	Further Refinement and Evaluation of Algorithms and Models	Year 4	Chong
I	Study of Netflow Records	Year 4	Cottrell/Zekauskas
II	RSVP client for Applications	Year 4	Matthews
III	Phase II Wide-area Experiments	Year 4	Clark/Riddle
III	Applications and Billing	Year 4	Clark
IV	IETF Draft for partial path connections	Year 4	Veeraraghavan

7 Broader Impacts

The broader impact of the proposed work is in the significant enhancements to the infrastructure for end-users and network providers. The proposed networking solutions are intended to make moderate-speed, low-delay network services available to more users at a fraction of the cost required today. We plan to leverage the geographic and intellectual diversity of our team, covering a breadth of topics from modeling to mathematics to operational networking, to develop inter-disciplinary courses, expose our students to national laboratory research (through the participation of SLAC), and bring greater opportunities to our minority and women students.

8 Results from Prior NSF Support

Edwin K. P. Chong: 0099137-ANI, 783, 104(*Chong´sportion*195,776), July 1, 2001 to June 30, 2004, "Design and Control of Next Generation Networks: A Measurement Analytical Approach," 31 publications to date resulting from this support.

Malathi Veeraraghavan: ANI-0335190, "End-To-End Provisioned Optical Network Testbed for Large-Scale eScience Applications," Jan. 2004 - Dec. 2006, published 9 papers (2 journal, 1 magazine, rest conferences/workshops), two technical reports, have posted specifications and software on web site: http://cheetah.cs.virginia.edu. Main accomplishment: We have deployed a wide-area circuit-switched high-speed testbed (NC-GA-TN) and provided scientists the software to use this testbed for large file transfers and remote visualization.