Multiplexing Traffic at the Entrance to Wide-Area Networks

Ramón Cáceres

Ph.D. Dissertation

Report No. UCB/CSD 92/717
Computer Science Division
University of California
Berkeley CA 94720

December 1992

ramon@cs.berkeley.edu

Multiplexing Traffic at the Entrance to Wide-Area Networks

by

Ramón Cáceres

A mis padres, Mon y Mirtha.

Abstract

Many application-level traffic streams, or conversations, are multiplexed at the points where local-area networks meet the wide-area portion of an internetwork. Multiplexing policies and mechanisms acting at these points should provide good performance to each conversation, allocate network resources fairly among conversations, and make efficient use of network resources.

In order to characterize wide-area network traffic, we have analyzed traces from four Internet sites. We identify characteristics common to all conversations of each major type of traffic, and find that these characteristics are stable across time and geographic site. Our results contradict many prevalent beliefs. For example, previous simulation models of wide-area traffic have assumed bulk transfers ranging from 80 Kilobytes to 2 Megabytes of data. In contrast, we find that up to 90% of all bulk transfers involve 10 Kilobytes or less. This and other findings may affect results of previous studies and should be taken into account in future models of wide-area traffic.

We derive from our traces a new workload model for driving simulations of wide-area internetworks. It generates traffic for individual conversations of each major type of traffic. The model accurately and efficiently reproduces behavior specific to each traffic type by sampling measured probability distributions through the inverse transform method. Our model is valid for network conditions other than those prevalent during the measurements because it samples only network-independent traffic characteristics. We also describe a new wide-area internetwork simulator that includes both our workload model and realistic models of network components.

We then present a simulation study of policies for multiplexing datagrams over virtual circuits at the entrance to wide-area networks. We compare schemes for mapping conversations to virtual circuits and queueing disciplines for scheduling datagrams onto virtual circuits. We find that networks should establish one virtual circuit per type of traffic flowing between two network points of presence, and provide round-robin service to transmission resources shared by virtual circuits. This multiplexing policy exhibits good performance and consumes moderate amounts of resources at the expense of some fairness among traffic sources of the same type. In particular, it maintains interactive delay nearly constant and close to the possible minimum, and maintains bulk transfer throughput near the possible maximum, even as network load increases beyond saturation. Furthermore, it results in bottleneck buffer consumption that rises slowly with offered load. Other multiplexing policies exhibit interactive delay that increases with offered load, and buffer consumption that rises quickly with offered load.

Again using our traffic characterization, we evaluate mechanisms for multiplexing variable-sized datagrams onto small fixed-size cells. Cells offer performance and implementation advantages to networks that service many types of traffic, but they incur bandwidth inefficiencies due to protocol headers and cell fragmentation. We find that cell-based networks using standard protocols are inefficient in carrying wide-area data traffic. For example, ATM-based networks using SMDS and IEEE 802.6 protocols lose more than 40% of their bandwidth to overhead at the network level and below. Furthermore, we find that viable compression techniques can significantly improve efficiency. For example, a combination of three compression techniques can regain more than 20% of the bandwidth previously lost to overhead.

Acknowledgements

I owe much to my teachers, colleagues, and friends, many of whom fit into more than one of these categories. I would like to specifically thank:

Domenico Ferrari, my research advisor, for giving me the freedom to define my own interests and the guidance necessary to pursue them to a successful conclusion. I am also very grateful for his timely review of this dissertation despite adverse conditions. John Ousterhout and Terry Speed, for serving on my qualifying exam and dissertation committees, and Mike Stonebraker, for chairing my qualifying exam committee. Their comments have substantially improved this work.

Sandy Fraser at Bell Labs, for supervising much of my work there and serving on my dissertation committee. His insight and experience helped guide my research in many ways. Sam Morgan, also at Bell Labs, for his consistent interest in and advice on my work. His thoroughness and enthusiasm have been an inspiration.

Peter Danzig, Deborah Estrin, Sugih Jamin, and Danny Mitzel at USC, for their collaboration on traffic measurement, analysis, and modeling, and their permission to include our joint work in parts of this dissertation. The many people who helped to gather traffic traces: Dave Presotto and Dennis Ritchie at Bell Labs, Dan V. Wilson at Bellcore, Cliff Frost, Balakrishnan Prabhakaran, and Brian Shiratsuki at Berkeley, and Mark Brown, John Conti, Dick Kaplan, and Jim Pepin at USC.

Mark Sullivan, for his support and criticism. Luis Felipe Cabrera, for his advice on technical and career plans. The members of the Tenet, Xunet, Progres, and Dash projects: Riccardo Gusella, Chuck Kalmanek, Srinivasan Keshav, Joann Ordille, Gary Murakami, Colin Parris, Vern Paxson, Stuart Sechrest, Shin-Yuan Tzou, Dinesh Verma, and others. They have been a sounding board for many ideas.

Rebecca Wright, for her companionship and encouragement. The Hillegass Boys: Will Evans, John Hartman, Steve Lucco, and Ken Shirriff, for making our house a home. Nina Amenta, Diane Hernek, Sandy Irani, Dan Jurafsky, Lu Pan, Dana Randall, Ronitt Rubinfeld, and again Will Evans and Mark Sullivan, for Three Winnebagos in the Desert, late-night violin recitals on Golden Gate Bridge, dim sum outings, and many other good things along the way. Marti Hirsch, for helping me through prelims. Randi Weinstein, for making available the notebook computer that made writing this dissertation a lot less painful than it otherwise would have been.

The staff of the CS Division, especially Kathryn Crabtree, for holding my hand and walking me through the bureaucracy. The staff at ICSI, especially Bryan Costales, for providing a productive place to work and a pool of superbly managed workstations on which to run parallel simulations.

Finally, I gratefully acknowledge financial assistance from AT&T Bell Laboratories, BP America, the Corporation for National Research Initiatives, Hitachi, Ltd., the International Computer Science Institute, and the State of California.

Table of Contents

1. Introduction

Lo, soul, seest thou not God’s purpose from the first?
The earth to be spann’d, connected by network,
The races, neighbors, to marry and be given in marriage,
The oceans to be cross’d, the distant brought near,
The lands to be welded together.

Walt Whitman, *Leaves of Grass,* 1892

Communication networks have changed the way people interact. The telegraph first allowed fast, reliable communication over long distances. The telephone network now links a significant portion of the world’s population. Data networks hold the promise of further improving human communication as we increasingly rely on access to widely dispersed information. However, many challenges remain before data networks can serve as many people as today’s telephone network.

This dissertation addresses problems that arise from the aggregation of traffic in large data networks. It characterizes sources of traffic and evaluates ways to mix, or *multiplex*, traffic from many different sources at the entrance to wide-area networks. It presents multiplexing policies and mechanisms that provide good performance and fair treatment to each source of traffic while making efficient use of network resources.

1.1. Context

Throughout this work we are concerned with wide-area internetworks of the type shown in Figure 1.1. *Host* computers communicate through the internetwork with other, geographically distant, hosts. Hosts are directly connected to a local-area network (LAN). Local-area networks are in turn connected to the wide-area portion of the network (WAN) through *router* systems. Routers communicate with other routers through the wide-area network. The wide-area network is itself composed of a collection of switching nodes and communication links (not shown in the figure). In general, data flows from a source host, over a LAN, through an input router, over the WAN, through an output router, over another LAN, and finally to a destination host.

Application programs running on the hosts are the ultimate sources and sinks of network traffic. Examples of *traditional* wide-area network applications are file transfers, electronic mail, network news, and remote terminal sessions. These traditional applications fall into two broad categories: bulk transfer (file transfers, electronic mail, and network news) and interactive (remote terminal sessions). Examples of *non-traditional* applications are voice, high-fidelity audio, and video.

We use the term *conversation* to mean the stream of traffic flowing between an application program on one host and an application program on another host. Conversations are of different *types*, corresponding to the different applications using the network. For example, we speak of file transfer conversations and electronic mail conversations. Each type may offer a different

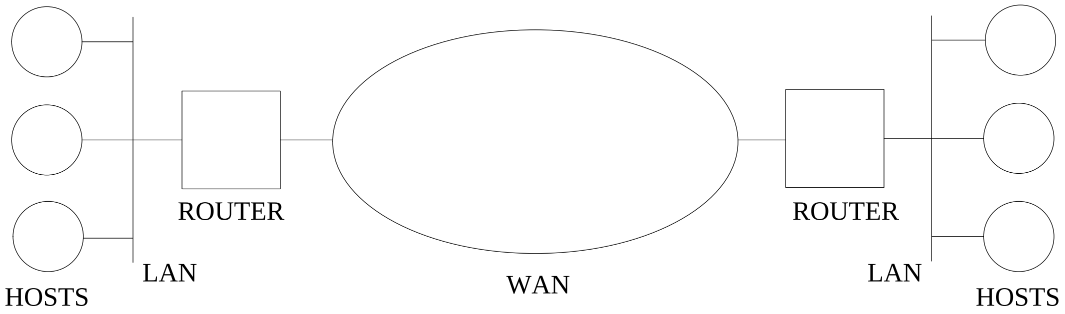

Figure 1.1. A Wide-Area Internetwork

workload to the network and demand a different level of performance from the network.

1.2. Traffic Multiplexing Objectives

Many conversations of different types come together at the entrance to wide-area networks. Our work aims to define policies and mechanisms for multiplexing traffic at a router that satisfy the following criteria: *performance*, *fairness*, and *efficiency*. We motivate these three objectives below.

First, each conversation should obtain the performance desired by the application that drives it. For example, bulk transfer applications transmit batch data such as files. They are concerned mostly with throughput, or how fast data is carried by the network. In contrast, interactive applications transmit interactive data such as keystrokes. They are concerned with delay, or the response time of the network. Within the limits of available resources, network control policies should provide high throughput to bulk transfer applications and low delay to interactive applications.

Second, the network should allocate its resources fairly among all conversations. In particular, it should provide low delay to interactive applications in spite of high offered loads from bulk transfer applications. It should not let high volumes of bulk data monopolize the network at the expense of low volumes of interactive data. In addition, all conversations of the same type should obtain equal performance.

Third, the network should make efficient use of its resources. In particular, it should not waste bandwidth or buffer space. High-speed wide-area networks include costly hardware components such as routers and switches that are made more expensive by their memory requirements. These networks also use expensive long-haul communication lines. Network control schemes with low buffer requirements are preferable over schemes with high buffer requirements. Similarly, protocols and multiplexing techniques that conserve bandwidth are preferable over those that waste bandwidth.

Our discussion is in the context of networks whose users are willing to pay for the level of performance desired by their applications, or networks that charge all users equally (including networks that do not charge). In networks where these conditions do not hold, the above three multiplexing objectives take on a slightly different form. First, a conversation should obtain the grade of service that has been mutually agreed upon by the network user and provider. A conversation may obtain a lower grade of service than that desired by the associated application. Second, the network should allocate its resources according to grade of service contracts. It may allocate more resources to conversations that have contracted for a higher grade of service than to those who have contracted for a lower grade, even if the underlying application is the same. Third, although efficiency remains an important consideration, an overriding goal is to minimize the investment required to provide the services contracted for. Inefficiencies may occur if they are unavoidable to insure a grade of service that users are willing to pay for.

1.3. Motivation

Our work is motivated by the failure of past networks to meet the stated multiplexing objectives and by trends in future networking.

We use two examples to illustrate the shortcomings of past networks. First, consider NSFnet [82]. NSFnet is a 1.5 Megabit/second, U.S. cross-country network that forms the backbone for the current Internet. Although its communication links have been gradually upgraded to 45 Megabit/second over the last two years, it continues to suffer from congestion due to the growth of the Internet. More to the point, in the face of this congestion, NSFnet gives poor performance to interactive applications – cross-country response times longer than 100-200 milliseconds are common, and much longer delays are not rare.† Studies have found that humans perceive interactive response to be ‘‘bad’’ when it takes longer than 100-200 milliseconds [90].

This behavior is a result of the networks’s traffic multiplexing policies. NSFnet does not separate different conversations or types of traffic. It treats all traffic the same way in a first-in first-out basis. Short interactive packets compete for resources with much larger bulk transfer packets and thus incur high delays. It is often difficult for interactive users to work productively across NSFnet.

Second, consider Xunet 1 [96]. Xunet 1 was an experimental 1.5 Megabit/second, U.S. cross-country network that was dismantled in early 1991. It was based on Datakit [33] switches, which separated the traffic from different conversations and gave priority to short packets. This policy avoided interactive delay problems – cross-country response times remained well below 100 milliseconds even under congestion. However, Xunet 1 did not make efficient use of its bandwidth resources – the aggregate throughput of bulk transfer applications never reached the full bandwidth of the network.‡

This behavior was due to a combination of the network’s multiplexing policies and its lack of buffer memory. The Datakit design dictates that, at each switch, each conversation be allocated an amount of buffer space equal to the network’s *round-trip window*. The round-trip window is the network’s inherent storage capacity, that is, the product of its bandwidth and round-trip propagation delay. Since congestion at a link can cause an entire window of data to temporarily accumulate in the buffer in front of the congested link, this much buffer space insures that the network does not drop data and senders can successfully utilize the full bandwidth of the network. However, even allowing for rapidly declining memory prices, the costs of implementing this design in high-speed wide-area networks can be prohibitive.

Table 1.1 shows the size of one round-trip window for a hypothetical continental network with different speeds. It assumes a 45-millisecond round-trip delay that approximates the speed-of-light propagation delay across 6,000 miles of glass fiber. This much memory is

---

† Measured between the University of California in Berkeley and AT&T Bell Laboratories in Murray Hill, New Jersey, on a typical weekday afternoon.

‡ Measured between the University of California at Berkeley and AT&T Bell Laboratories in Murray Hill, New Jersey, under various load conditions.

Link Speed (Megabit/second)	Window Size (Kilobytes)
1.5	8
45	253
150	844
600	3,375
1,700	9,562

Table 1.1. Round-trip windows for a network with a 45-millisecond round-trip time

necessary for each conversation flowing through each Datakit switch. Xunet 1 switches did not have enough buffer memory to support high throughput across long distances.

Five trends involving wide-area internetworks suggest that the problems just described will intensify. First, long-haul communication speeds are increasing. Second, local-area network speeds are increasing. Third, host speeds are increasing. Fourth, the number of traffic types is increasing. Fifth, the number of local-area networks and hosts connected to wide-area internetworks is increasing. For wide-area networks, these trends portend the following three conditions. First, there will be increasing demands for resources, both bandwidth and memory. Second, there will be a need for more grades of service, in particular delay and throughput. Third, there will continue to be congestion due to the aggregation of traffic.

We see a need for better traffic multiplexing policies and mechanisms. We aim to find techniques that avoid the shortcomings of past networks and accommodate the demands of future networks. In particular, we would like to prevent a stream of traffic from interfering with the performance of another stream, and to make efficient use of memory and bandwidth.

1.4. Scope of the Investigation

All wide-area networks multiplex data at their entrance. There are many types of networks carrying many different workloads. It would be impractical to explore every aspect of the multiplexing problem as described so far. Instead, we concentrate on an important class of network and workload, namely, *cell-based virtual circuit networks that carry traditional datagram traffic*. We define the scope of our investigation below.

1.4.1. Datagrams over Virtual Circuits

Datagrams and virtual circuits are widely used data communication styles. The two terms have been defined many ways. A useful analogy compares datagrams to the common postal service, and virtual circuits to ordinary telephone service [98]. Like letters flowing through the postal system, datagrams are individually addressed and delivered items of information. A datagram service delivers information with no guarantees regarding ordering or reliability. In contrast, like voice conversations in the telephone system, virtual circuits set up and tear down a path through the network before and after their use. A virtual circuit service delivers information in the same order it was sent, although not necessarily free of errors.

Throughout this work, we use the following restricted definition of virtual circuits. Virtual circuits imply only three things: a connection establishment and teardown procedure, a fixed path through the network for the lifetime of a virtual circuit, and a separate queue for each virtual circuit at each queueing point. We do not assume many other features that are sometimes associated with virtual circuits. For example, we do not assume any reservation of resources, either memory or bandwidth; all data is statistically multiplexed. We also do not assume any error recovery beyond checks for the integrity of protocol headers.

There is a long history of wide-area networks that transport datagrams over virtual circuits, and such networks continue to exist. In these networks, hosts outside the wide-area portion of the network exchange datagrams, while nodes inside and at the edge of the wide-area portion of the network communicate through virtual circuits. The Advanced Research Projects Agency Network (ARPANET) [82], built in the late 1960's and used until the late 1980's, used virtual circuits for communication between its internal nodes, or Interface Message Processors (IMPs). However, ARPANET hosts exchanged datagrams from the Internet Protocol (IP) [15]. More recent networks built in the 1980's carried IP datagrams over X.25 virtual circuits [65], for example the Defense Data Network (DDN) [24] and parts of the Computer Science Network (CSNET) [14].

There will also be future virtual-circuit-based networks that carry datagrams. Although there are competing opinions as to how networks will evolve, there is widespread agreement that future networks will be integrated. An integrated network is one that carries many different types of traffic – traditional data traffic as well as non-traditional traffic. A prominent design calls for Broadband Integrated Services Digital Networks (B-ISDN) based on Asynchronous Transfer Mode (ATM) virtual circuits [41] [72]. New traffic types are appearing because increasing communication speeds now allow networks to support applications with stringent bandwidth and delay requirements, for example audio and video. It is not yet clear what protocols these new applications will use. However, traditional traffic will continue to use well-established datagram protocols such as IP over the new ATM virtual circuits.

The reasons for this mix of styles are many. Datagrams are attractive for end-to-end internetwork applications, while virtual circuits offer advantages to providers of integrated-services wide-area networks. Some network applications like Remote Procedure Call (RPC) [8] are natural datagram applications. They exchange data in largely independent pieces that do not fit well in the virtual circuit mold. On the other hand, a large network based on virtual circuits is easier to manage than one based on datagrams. A virtual circuit is a convenient unit on which to carry out flow control and congestion control policies, allocate bandwidth and delay, and perform accounting. For example, it is easier for a virtual circuit network than for a datagram network to apply back-pressure along a path back to a misbehaving source, because each virtual circuit uniquely identifies a complete path from source to destination, while datagrams between the same source and destination may follow different paths.

1.4.2. Cell Networks

There will also be future cell-based networks, again due to the push for service integration. *Cells* are small, fixed-length units of data. For example, ATM networks use cells that are 53 bytes long. Before transporting application data units such as variable-sized datagrams, the network fragments them into cells. Cells are the basic multiplexing unit throughout the network. They allow fine control over bandwidth and delay allocation. Small cells provide lower delay than large datagrams because a short multiplexing unit quickly frees up resources for units from other conversations. No multiplexing unit is forced to wait for a long unit to complete transmission. The fixed-size nature of cells also allows efficient hardware implementations since the network always deals with homogeneous units of data. Thus, cells offer performance and implementation advantages to integrated networks.

1.4.3. Traditional Data Traffic

Our choice of workload is traditional data traffic, by which we mean traffic from existing wide-area network applications such as file transfers, electronic mail, network news, and remote terminal sessions. We chose this workload because we can accurately characterize it through direct measurements of current network traffic. Future networks will begin to carry audio and video traffic with characteristics different from current traffic. However, representative quantities of this traffic are not available for measurement. Furthermore, traditional forms of traffic will prevail for several years and will continue to be used for much longer. As a result, future networks, including B-ISDN networks based on ATM, must efficiently support traditional traffic.

1.5. Applicability of Results

Although we have narrowed the scope of our investigation, our results will be applicable to a wide range of networks. For example, our results could contribute to the design of future datagram-based networks as well as future reservation-based networks. We will return to the applicability issue at the end of the dissertation, after we have presented our work in detail.

1.6. Underlying Approach

We take an empirical approach to our work that is seldom used in wide-area network research. There is considerable research activity addressing the needs of future types of traffic, in particular real-time traffic like audio and video [44] [45]. These studies make educated guesses as to the characteristics of future traffic and use these best-guess models to make network design decisions. No other approach is possible with traffic that does not yet exist.

However, we can do much better with mature traffic types. File transfers and remote terminal sessions have been important applications of wide-area networks for close to 20 years [61]. We expect that many characteristics of these traffic types remain stable or can be parametrized to track changes in technology. Nevertheless, no realistic per-conversation models of wide-area traffic exist. We should be able to accurately characterize this traffic from observations of real networks, without relying on our intuition. We can use what we learn to tune existing networks and design new ones.

This dissertation exemplifies this empirical approach to network research. We measure real networks and identify the characteristics of their traffic. We find that many important characteristics are indeed stable. We then derive traffic models directly from the measurements. Finally, we use these models to study traffic multiplexing at the entrance to wide-area networks. Through simulation, we evaluate different multiplexing policies in terms of their performance, fairness, and efficient use of buffer memory. We also evaluate different network protocols and multiplexing techniques in terms of their efficient use of bandwidth.

1.7. Outline