Computer Systems Laboratory Research Review

Joseph Pasquale
George Polyzos

Computer Systems Laboratory
Department of Computer Science and Engineering
University of California, San Diego
La Jolla, CA 92093–0114

May 1995

UCSD Technical Report CS95-423

Abstract

At the Computer Systems Laboratory at UCSD, we conduct research in operating systems, communication networks, and performance evaluation. This document reviews the current research within the Computer Systems Laboratory. The final section contains descriptions of research projects undertaken by our students.

For more information, contact the authors at the above address, or by FAX at (619) 534-7029. Joseph Pasquale may be contacted by e-mail at *pasquale@ucsd.edu* or by telephone at (619) 534-2673. George Polyzos may be contacted by e-mail at *polyzos@ucsd.edu* or by telephone at (619) 534-3508.

1 Introduction

At the UCSD Computer Systems Laboratory (CSL), we conduct research in operating systems, communication networks, and performance evaluation. In operating systems, we are focusing on high performance I/O software support for a variety of I/O intensive applications, such as multimedia applications and scientific applications. In communication networks, we concentrate on problems associated with supporting continuous media, such as audio and video, over high-speed networks. In performance evaluation, we are interested in characterizing the I/O time-behavior of scientific applications running on supercomputers, such as the Cray Y-MP at the San Diego Supercomputer Center (SDSC).

CSL members include two associate professors (the authors), a visiting scholar, 10 graduate students, and 5 undergraduates. CSL is equipped with a variety of workstations, including two DEC 3000 Model 400 AXP Alpha workstations, four DEC 3000 Model 600 AXP Alpha workstations, six DEC 3000 Model 800 AXP Alpha deskside workstations, seven DECstation 5000/200, four DECstation 5000/240, four DECpc 590 Pentium PCs, six other Intel-based PCs, two Sun SPARCstation 2, two Sun 3/60, one Apple Macintosh IIci and one Apple Macintosh Quadra 840 AV.

Networks within the laboratory include Ethernet, FDDI, and a WaveLan wireless network. CSL has excellent connectivity to the outside world through various UCSD campus networks, the Sequoia 2000 network, and a direct fiber-optic link to the San Diego Supercomputer Center (SDSC). SDSC is a major node on the Internet and other important experimental and production networks such as the CASA Gigabit Testbed, the ATM based vBNS, CERFnet, etc.

CSL is part of the $20 million *Sequoia 2000* project at the University of California (UC). Sequoia 2000 is a collaboration of computer scientists and Earth scientists seeking to address the massive data storage, network, and visualization requirements of Global Change scientific applications. CSL lead the effort on the construction of the Sequoia 2000 network [FPP93], a high-speed computer network that connects the majority of the UC campuses. Much of our work on network-based I/O has been directly applied to the design and implementation of the network.

This document reviews the current research within CSL. In the next three sections, we describe an overview of the work of the three research areas in the laboratory: operating systems, communication networks, and performance evaluation. The following section contains abstracts of past CSL doctoral dissertations and titles of Master theses. The last section contains research summaries of the students.

2 Operating Systems

We are investigating new operating system software structures to improve input/output (I/O) and interprocess communication (IPC) performance for applications requiring communication of very large data objects, e.g. hundreds of megabytes or more. The more important classes of applications supported by this work include multimedia (e.g. images, video, and audio) and scientific (e.g. state-based simulations of atmospheric and oceanic processes).

I/O is one of the lesser understood areas of operating system design, and, with IPC, has become increasingly important due to three important trends: (1) applications are becoming more data-intensive; (2) program structure is based more and more on a set of communicating processes distributable over a set of computers connected by high-speed networks; and (3) the relative speeds and rates of improvement in processor, memory, and I/O (especially disk and network) technologies have shifted dramatically. These trends indicate that the primary bottleneck in computer systems is I/O, and to a lesser degree, memory bandwidth which affects both I/O and IPC.

The UCSD Computer Systems Lab has been at the forefront of this area, having identified why common I/O system software structures are inadequate to support intensive I/O (particularly continuous media) applications, and proposing new I/O software architectures based on the principle of separating control and data movement ([P92], and [P93]). These ideas are supported by concrete experimental research results, including the following contributions on topics which have (or will soon have) each resulted in separate Ph.D. dissertations:

- "Container-shipping", a new interdomain data transfer facility that achieves high throughput and low latency, while maintaining the security provided by separate protection domains [PAM94];
- "In-kernel datapaths", a new interdevice data transfer facility that achieves high throughput and low latency by allowing processes to create data paths directly between I/O devices [FP93, FP94];
- "Multi-structured file system design", integrating three separately organized disk-array storage structures (mirrored for fast response, striped for high throughput, and logging for reliability) to achieve high performance, reliability, and scalability [MP91].

3 Communication Networks

In the area of Communication Networks we explore issues related to (a) the design of network architectures and protocols for high-speed real-time transport of Continuous Media, (b) the performance evaluation of local, metropolitan, and wide area networks through analysis, simulation, and measurement, and (c) network traffic characterization and modeling.

One of our main areas of concentration is the investigation of network architectures and protocols supporting **Continuous Media** (CM), e.g., digital audio and video. CM are often associated with real-time applications such as teleconferencing. The interactive nature of these applications imposes stringent delay requirements on the CM flows. Various guidelines set the end-to-end delay tolerance to tens or at most a few hundreds of milliseconds.

We are investigating network protocols that support various levels of **guaranteed Quality of Service** (QoS) in the form of minimum throughput, maximum delay, and maximum delay jitter, per requested channel. Our work to date has focused on two new network service disciplines: The first, called "Leave-in-Time", provides end-to-end performance guarantees for throughput, delay, and delay jitter [FP95]. Leave-in-Time determines required memory buffer sizes and link bandwidths for the performance bounds it provides. The second, called "Service Curve Based Earliest Deadline First" also provides performance guarantees for throughput and delay and has various other desirable properties, such as, having predictable performance even when a source changes the characteristics of the traffic that has declared to the network. It also provides improved statistical multiplexing gains by requiring and exploiting more detailed characterizations of the traffic [SCP95].

We have also investigated techniques to extend to interconnected FDDI rings the performance guarantees offered by the **FDDI synchronous mode** [B92]. Furthermore, even though the synchronous mode is an integral part of the FDDI standard, as well as a distinguishing characteristic of the FDDI medium access protocol, appropriate system software for the synchronous mode is unfortunately not typically implemented. Therefore, we have made modifications to Ultrix device drivers and have extended network protocols in order to enable applications with real-time requirements to take advantage of the synchronous mode delay guarantees [J92, E93].

We have also developed software structuring techniques which lead to **high performance implementations of network protocols**, particularly TCP/IP. This work includes detailed performance evaluation studies of a common TCP/IP software implementation [KP93, KP93a, KP93b].

In some cases, meeting the stringent delay requirements of interactive CM traffic may require dropping packets when they accumulate between the sender and the receiver because of transient congestion. Fortunately, humans receiving real-time audio and video have some tolerance to transmission errors. The perceived error may vary, however, depending on the particular encoding schemes used and the level of compression. **Hierarchical coding** (also referred to as layered coding) techniques can split CM signals into subcomponents of varying importance. Less important signal components can be dropped without causing an interruption of service; communication can continue in the presence of delays, errors, or network congestion, with only a moderate drop in service quality [P93b]. We have completed one of the first implementations of the hierarchical mode of the JPEG compression standard and are now investigating the tradeoffs of using hierarchical coding for image and continuous media transmission over packet switching networks, and in particular, ATM networks [P95, TP95, HP95].

In addition, we have developed a new communication abstraction for dissemination-based communication of multimedia called the "Multimedia Multicast Channel" (MMC) [PPAK94]. The MMC is based on a Cable-TV model: a source places media streams on a channel, without explicit interaction with receivers; receivers can dynamically "tune in" to channels of interest, without explicit interaction with the source. The use of loose-coupling between a source and the multiple receiver nodes promotes pro-active control mechanisms. In addition, heterogeneous receivers are supported by encouraging hierarchical coding of sourced media streams which receivers can filter arbitrarily. "Filter propagation" is another unique feature of the MMC, whereby a data-stream filtering software module can propagate from a receiver toward a source along a path in a multicast tree, e.g. to reduce bandwidth consumption [PPAK94a].

In connection to **multicasting**, we have developed multicast tree construction algorithms that minimize cost while bounding delay for routing time-sensitive high-bandwidth data (e.g. video) [KPP93, KPP93a].

In the area of **Metropolitan Area Networks** (MANs) we have focused on a key problem in the DQDB protocol. The Distributed Queue Dual Bus (DQDB) protocol is the channel access mechanism of the MAN standard (IEEE 802.6). This is a very interesting system because its frame structure is compatible with ATM/B-ISDN and is believed to be the first form of high-speed connectivity to be offered for the local loop. We have developed the Rotating Slot Generator (RSG) scheme, which overcomes the well known unfairness of DQDB through architectural modifications that allow the system to operate essentially as a dual ring rather than as a dual bus [KPP92, KPP92a, KPP95]. RSG provides higher throughputs and lower delays, and also improves other system characteristics, such as the degree of discrimination between priority classes for multiple priority systems.

We have also contributed to the performance analysis of **Conflict Resolution Algorithms** for broadcast communication channels. These protocols are applicable to satellite channels and other wireless packet networks. We have developed a queueing theoretic methodology for the analysis of these protocols which allows us to obtain exact steady state distribution results for the packet delay for many interesting configurations [PMV87, PM94]. In particular, in [PM93] we provide a complete delay analysis for the celebrated FCFS 0.487 Algorithm.

Finally, we have undertaken an extensive NSF-sponsored effort to **characterize and model the traffic** on the NSFNET backbone and other wide area network environments. We focused on traffic and performance from a network rather than an application perspective [CPB93].

We have presented the data collection architecture for the T3 NSFNET, compared it with that for the T1 NSFNET and concentrated on difficulties with using the collected statistics for long-term traffic forecasting [CBP94]. An ARIMA time-series model of long-term NSFNET traffic volume with very good forecasting capabilities is presented in [GP94]. Our study of Internet latencies has shown that the widely used approximation of unidirectional latencies by half the round-trip "ping" delay can be considerably inaccurate [CPB93b]. An observation of discrepancies in traffic measurements obtained through two different mechanisms, prompted interest in sampling for traffic capture, measurement and reporting. We present a methodology for evaluating various sampling approaches and apply it to two specific targets (packet size and interarrival time distribution) in [CPB94].

In [CBP95] we present a methodology for profiling traffic flows on the Internet at a variety of granularities, such as destination network, network pair, destination host, host pair, port, etc. We document that in observed data sets the number of host pair flows is far less than proportional to the square of the number of network number pair flows. This is important for various proposed resource reservation schemes such as RSVP, hierarchical round-robin, etc., that require servicing traffic based on connections or network or host pairs.

Our traffic characterization results allow more accurate capacity planning and network tuning in high-speed realms, such as of routing mechanisms, at various levels of granularity. This research is of particular importance now since new services not only dramatically increase network traffic, but also completely change traffic profiles.

4 Computer System Performance Evaluation

Based on the need to design well-balanced supercomputer systems that deliver the high performance expected of them, we are undertaking a comprehensive study of the relationship between the I/O requirements of scientific applications and the capabilities of the I/O system. We are working closely with the San Diego Supercomputer Center (SDSC) which is one of the four national supercomputer centers sponsored by the National Science Foundation. SDSC currently maintains a Cray C-90 supercomputer (among other machines) which we have been measuring. SDSC offers Cray time to users all over the United States, who generally wish to run large scientific applications. These users comprise 2700 researchers from 170 institutions, spanning 44 states.

The goals of this research are:

1. to characterize the I/O behavior of a wide spectrum of scientific applications
2. to develop models for well-defined classes of scientific application I/O activity
3. to construct an executable I/O benchmark set for the purpose of locating existing or potential I/O bottlenecks
4. to recommend operating system software enhancements to improve supercomputer I/O performance.

The first goal addresses the problem that our understanding of scientific applications needs to be improved. Scientific applications form an interesting set of programs worthy of study because they absolutely need high performance supercomputers and therefore, systems should be designed to meet their requirements. Before we can do the latter, we must understand these requirements. This involves observing an environment supporting a variety of scientific applications and determining what to observe to deduce these requirements in a way that does not interfere with their execution.

The second goal involves summarizing these observations into a form that may be used for descriptive or predictive purposes. The intent of the third goal is to provide a comprehensive set of scientific I/O benchmarks so that users can select the I/O requirements that best represent their own workloads. Finally, the fourth goal is to put results into practice: to improve I/O performance on the supercomputer at our disposal which supports scientific workloads. This will be limited to operating system software enhancements, such as the design and implementation of I/O resource scheduling algorithms, or modification of software cache sizes. This may also involve suggestions to the designers of applications of how to improve their codes.

We are focusing mainly on application characterization rather than overall system behavior because our intent is to investigate not only current but also future potential problems which are not limited to what is observable in current systems. Specifically, we are interested in understanding the issues involved around a potential I/O bottleneck in future supercomputers and to provide tools to prevent it. This scenario is associated with supercomputers that are computationally much more powerful than those of today and also with workloads that are much more I/O intensive. The latter is driven by user expectations about performance of the future machines. Based on current application I/O characterizations one can extrapolate future I/O application requirements, and then through synthesis determine overall I/O workload intensities and patterns.

Results of our analyses support our hypothesis that a significant proportion of I/O intensive, long running, frequently executed scientific applications have predictable I/O requirements [PP93, PP94, PP94a].

In particular, in [PP93] we presented the results of our study of the production workload of SDSC's then Cray Y-MP, identifying I/O intensive scientific applications and examining the regularity in their resource usage. Our workload characterization analysis isolated a set of 50 distinctly named, I/O intensive scientific applications whose aggregate resource usage represented 71% of total system bytes transferred and 9.5% of total system CPU time during a one-month period. Cluster and regression analysis was performed on the most frequently executed applications in this I/O intensive set to determine characteristic patterns and statistical relationships of individual application resource usage.

Cluster analysis revealed three general patterns in application resource usage. Regression analysis revealed that I/O demands, both in terms of the number of characters transferred and the number of logical requests, show considerable correlation with CPU time consumed.

We also performed a workload characterization study of SDSC's Cray X-MP (the predecessor to the Cray Y-MP), running the CTSS operating system [PBK91]. We found that 93% of all jobs were public jobs, yet 90% of all CPU time was consumed by private jobs. Although the public utilities make relatively small demands on the CPU, they consume 22% of the total memory space-time product, and 75% of the total I/O channel time. The latter indicates that a very significant I/O demand is being made by the public component of the workload. Another important observation was that even jobs that make insignificant demands on the CPU (e.g., interactive system utilities) can have a major effect on the I/O system.

Our study presented in [PP94a] focuses on application level I/O behavior with respect to both static and dynamic characteristics. We observed SDSC's Cray C-90 workload and isolated the most I/O intensive applications. The combination of a low-level description of physical resource usage and the high-level functional composition of applications and scientific disciplines for this set reveals the major sources of I/O demand in the workload. We selected two applications from the I/O intensive set and performed a detailed analysis of their dynamic I/O behavior. These applications exhibited a high degree of regularity in their I/O activity over time and their characteristic I/O behaviors can be precisely modeled by one and two, respectively, recurring sequences of data accesses and computation periods.

5 CSL Doctoral Dissertations

This section contains abstracts of past doctoral dissertations by CSL students.

K. Muller: Multi-structured File System Design for High-performance Computing, 1990

File system I/O is increasingly becoming a performance bottleneck in large distributed computer systems. This is due to the slow decline of disk access times (latency and seek) relative to the rapid increase in CPU speeds, memory size, and network bandwidth. Indeed, although disk storage capacity increases while the physical size and cost of disks decrease, file system performance remains limited by long disk access times and disk contention. The problem is further aggravated by new applications with intensive I/O components, such as multi-media or visualization programs.

Technology trends indicate that it is unlikely that disk access delays will significantly decline in the near future. Thus, solutions cannot be based on improvements in disk access time, and must somehow circumvent this problem. Until recently, the high cost of disk storage, limited the design of file systems to techniques which optimize storage efficiency. However, with the rapid decline in disk costs, solutions explicitly designed to optimize performance have become very attractive.

We present a design for a multi-structured file system. Our design is based on the coordinated and cooperative integration of disk caching with three different storage structures each implemented on independent and isolated disk systems. The individual storage structure were designed to optimize a different set of file system access characteristics such as cache writes, directory searches, file attribute requests or large sequential reads/writes. The isolated storage storage structures reduce contention and lower disk access delays, so high bandwidth I/O and decoupled concurrent access to multiple disks can be provided. Both file system performance and storage size can be scaled to meet the requirements of different work loads by adding disks to the appropriate storage structures. As part of our study, we analyze the performance of an existing file system using trace data from UNIX disk I/O-intensive workloads. In trace driven simulations, we show how performance improves using our technique of dividing file system requests based on access type to optimized storage structures, as implemented by our multi-structured file system.

A. Glockner: Coadaptive Decision-making in Distributed Computing Systems, 1991

Decision-making in distributed computing systems is complicated by the introduction of communication delays and the resultant uncertainty both in the global system state and in current actions in the system. This dissertation examines the effects of introducing *coadaptation* into distributed decision-making. By constructing the decision-making process as a set of autonomous agents that can individually learn from feedback, the set of agents adapt together, or *coadapt*, to form a good, though not necessarily optimal, global decision process.

These effects are demonstrated with simulation experiments in distributed computer job scheduling; the agents are implemented as learning automata. The resulting behavior is analyzed both in terms of a global performance metric — mean job turnaround time — and in terms of the states of the agents themselves. Several key elements in the behavior of coadaptive systems are illustrated. These include demonstrations of: situtions where apparently good individual decisions result in worse performance, correlations between automata type and decision-making policy of the agents, complementarity in agent behavior, differences in the relative performance of agents both with small changes in agent parameters and and with changes in the environment, and properties of stability in coadaptive systems. Simulations performed with changes to agents and to the distributed computing system model show further principles of coadaptive behavior. The dissertation concludes with a summary of principles of coadaptive behavior and a projection of future methods of coadaptive distributed job scheduling.

E. Billard: Delayed Communication and Dynamic Group Formation in Distributed Systems, 1993

In a centralized computing system, the sole decision-maker, or autocrat, has a perfect view of the global state; it can claim that "it is the state." However, centralized computing is giving way to decentralized computing where the distributed decision-makers each have their own, but uncertain, view of the global state; the uncertainty is caused by the delay in communication of state information. There is a tradeoff in the frequency of communication: too much communication increases overhead and too little communication results in poor decision-making due to an inaccurate view of the global system state. We model this behavior with distributed game automata which are adaptive decision-makers with aged information concerning the strategies of each other. We focus on the performance of these decision-makers with respect to the amount of communication. A particular type of decision considers *with whom* to coordinate and the goal is to find optimal groups within a network of decision-makers. That is, agents may decide to work alone or to work together in groups to derive a mutual benefit. Simulation of a static load balancing application shows that, with sufficient communication, an adaptive strategy in dynamic group formation is globally optimal. Analysis indicates the potential for poor decisions when the agents communicate infrequently. The results illustrate the fundamental nature of the uncertainty present in any distributed system.

V. P. Kompella: Multicast Routing Algorithms for Multimedia Traffic, 1993

This thesis examines the problem of route construction for multimedia traffic directed at multiple destinations. Two significant properties of multimedia data, particularly audio and video, that are taken into account are the high sustained data rates and the low tolerance to delay. In particular, we examine solutions that construct a tree of low cost (which is a measure of the bandwidth utilized by the multicast), and bounded end-to-end delay from source to each destination in the multicast. We call a tree of minimum cost that adheres to the path delay constraints a *constrained Steiner tree*. The major contribution of this work is to present a better approach to multicast routing through this constrained Steiner tree.

We show that the problem of finding the constrained Steiner tree is NP-complete, and present two approaches to construct approximate solutions. The first is a centralized one, wherein all nodes have complete knowledge of the topology and state of the network, in terms of the cost and delay of each link. The source then computes an approximate constrained Steiner tree. There are two variants of the centralized heuristic, each using a different selection function to determine which edges are desirable for the tree. One selection function, f_C, involves the cost of an edge, and the other, f_{CD}, involves both the cost and the delay of the edge.

The second approach is a distributed algorithm, wherein each node knows the costs and delays on adjacent links, and can learn more information through communication. Again, there are two variants of the algorithm, based on the selection functions f_C and f_{CD}.

We show that the algorithms are guaranteed to terminate with an approximate delay-bounded solution. Furthermore, through extensive simulations, we demonstrate the performance of the centralized and distributed algorithms. To this end, we compare the solutions produced by these algorithms to the cost of the optimal solution. The centralized algorithms produce trees that are within 5-10% of the optimal cost. The distributed ones are within 20-30% of the optimal. Current routing strategies, based on minimum delay algorithms, produce trees that cost 70-80% more than optimal.

K. Claffy: Internet Traffic Characterization, 1994

Traffic statistics normally collected during day-to-day operation of wide-area datagram networks are frequently insufficient for researchers to use in studying the workloads and performance of these realistic environments. As wide-area networks become more ubiquitous and service expectations rise, current methods for collecting data will become even less suitable. We examine ways to improve techniques for statistics collection so that the resulting data will enable researchers, and indeed service providers themselves, to develop more accurate Internet traffic models.

We first provide a taxonomy of traffic characterization tasks. We then use operationally collected statistics to characterize traffic of the T1 and T3 NSFNET backbones. Because current infrastructural statistics collection is oriented toward either short term operational requirements or periodic simplistic traffic reports to funding agencies, this data is often not conducive to assessing network workload or performance; we evaluate to what extent they are useful for tasks in the taxonomy, and propose improvements in current statistics collection architectures, with particular application to the NSFNET backbone. We include an investigation of the effects of sampling to characterize traffic and evaluate performance in a high-speed wide-area network environment.

In the second part of the thesis we focus on items in the outlined taxonomy that are not conducive to investigation using operationally collected statistics. These items mostly involve short-term aspects of Internet flows, which operationally collected statistics fail to expose. We develop a general methodology for use in assessing Internet flow profiles and their impact on an aggregate Internet workload. Our methodology for profiling flows differs from many previous studies that have concentrated on end-point definitions of flows defined by TCP connections using the TCP SYN and FIN control mechanism. We focus on the IP layer and define flows based on traffic satisfying various temporal and spatial locality conditions, as observed at internal points of the network. We first define the parameter space and then concentrate on metrics characterizing both individual flows and the aggregate flow. Metrics of individual flows include: volume in packets and bytes per flow, and flow duration. Metrics of the aggregate flow, or workload characteristics from the network perspective, include: counts of the number of active, new, and timed out flows per time interval; flow interarrival and arrival processes; and flow locality metrics. Applying the methodology to our measurements yields significant observations of the Internet infrastructure, which have implications for performance requirements of routers at Internet hotspots, general and specialized flow-based routing algorithms, future usage-based accounting requirements, and traffic prioritization.

Finally, we discuss trends that will affect how Internet service providers collect statistics in the future. Improvements in operational statistics collection, such as support for flow assessment, will help networking activities along various time horizons, from defining service quality patterns to long-term capacity planning. We offer a unique combination of operational and research perspectives, allowing us to reduce the gaps among (1) what network service providers need; (2) what statistics service providers can provide; and (3) what network analysis requires.

K. R. Fall: A Peer-toPeer I/O System in Support of I/O Intensive Workloads, 1994

The I/O subsystem interface present in most modern operating systems stems from the Multics system of the late 1960's. The Multics I/O design places a user program in the center of the flow of I/O data, requiring data movement between I/O devices and user processes. This dissertation argues that an I/O system supporting peer-to-peer communication between I/O objects improves programmability and increases overall system performance in terms of throughput and minimization of transaction delay over the traditional model. Peer-to-peer communication refers to the ability to transfer data between I/O objects (i.e. entities upon which typical read and write I/O operations may be performed) without direct user processes intervention. User processes define data stream endpoints, but need not actively transfer data between objects to facilitate I/O data flow. A system interface and architecture is described which implements a peer-to-peer I/O model. A set of example applications and measured implementations substantiate the claimed flexibility and efficacy of the mechanisms described.

J. S. Kay: Path IDs: A Mechanism for Reducing Network Software Latency, 1995

Network performance is important to an increasing number of com- puter applications. Similarly, network hardware performance is climbing to meet the demands of these new applications, with the advent of 100-megabit Ethernet, ATM, and fast processors to drive them. By contrast, network soft- ware latency is improving relatively slowly. It poses a barrier to the increasing use of networks for a wide variety of applications; as this dissertation shows, minimizing latency is more important to the minimization of processing time of the overall network workload than maximizing throughput.

This dissertation introduces PathIDs, a means of reducing network latency. The latency problem is exacerbated both by the relative lack of research into latency optimization as compared to throughput optimization and by the inherent difficulty of optimizing latency. It is difficult to significantly reduce the average processing time due to non-data-touching overheads because that time is spread across many operations, so that achieving a signif- icant savings would require optimization of many different mechanisms. PathIDs provides a framework for optimizing such a wide variety of overheads.

Optimizing large packet processing is also important; it is the focus of a great deal of study and is believed to be important to applications such as sci- entific data management and digital video transmission. This dissertation describes techniques that greatly reduce processing time for large packets, improving through- put, by eliminating most of all three major CPU-moderated data-touching operations: checksumming, copying data from kernel to user buffers, and copying data from user to kernel buffers. Checksumming is largely eliminated through a technique which avoids checksumming in the frequent case in which it is redundant with the CRC of a LAN adapter. Copying from kernel to user buffer can be largely eliminated by redirect- ing DMA pointers concurrently with packet delivery. Copying from user to kernel buffer can be avoided by arranging for adapters to DMA directly from user buffers.

6 CSL Master's Theses

This section contains the titles of past master's theses by CSL students.

Kirk Taylor, "Prototype Video Multicasting over a Local ATM testbed," M.Sc. Thesis, June 1994.

Nancy Groschwitz, "Traffic and Delay Patterns on the NSFNET Backbone," M.Sc. Thesis, August 1993.

David C. Etherton, "Approximating FDDI Synchronous Mode Guarantees Through Kernel Modification," M.Sc. Thesis, March 1993.

Michael M. Johnson, "FDDI Performance Guarantees for Real-Time Traffic in the Absence of a Synchronous Mode," M.Sc. Thesis, December 1992.

Riadh Ben Taarit, "Extending FDDI Synchronous Mode Guarantees to Interconnected Networks," M.Sc. Thesis, September 1992.

7 Graduate Student Research Summaries

Research summaries of the graduate students who are members of the Computer Systems Laboratory follow.

Operating System Support for I/O-Intensive Applications

Eric Anderson

Experimental multimedia projects such as the distribution of realtime audio and video by an IP network are often hindered by the inefficient transfer of data between I/O devices on a workstation. Much of the inefficiency results from the design of the I/O interface which Unix presents to programmers. Design assumptions valid several years ago are not always suitable for high performance I/O.

The hardware arrangement of most modern workstations permits the bulk transfer of data from an I/O device into RAM, from RAM to a device, and sometimes directly from device to device. Experimental programs written at the user process level often inspect or modify data. Compression of video, silence detection in audio, and format conversion for either media typically require that the data be accessible by the user's program as the data moves between devices.

For this type of data-manipulating program, the optimal use of hardware occurs when the device-to-RAM transfer places data into the RAM which will be used by the user's program. Likewise, the RAM-to-device transfer should originate from RAM used by the user's program. Protocol processing and the traditional system call interface make this difficult or impossible to accomplish in current Unix systems.

I have developed an alternate interface which will permit this optimal data transfer for all types of I/O. The interface allows a user process to take ownership of one or more hardware pages containing received data, rather than having the data copied byte-for-byte into some additional RAM which the user already owns. A similar process is used when sending data.

This approach yields a variety of security, performance, memory management, and implementation problems. For example, a user process may read an amount of data which is smaller than the hardware page size. However, memory is given to the user process in integral multiples of the hardware page size. In this case, the part of the transferred memory not occupied by the read data must not contain any other kernel data to which the user would not normally have access. There are several possible solutions to this problem which are being studied.

Experiments are being conducted by modifying the OSF/1 operating system, running on 64-bit Alpha AXP workstations connected to T1, Ethernet, T3, and FDDI networks. Raw and JPEG-compressed video provide an excellent source of high-volume data for exercising high-performance networks and operating systems. Additional details of the work to date can be found in the reference below.

Reference

[1] Pasquale, J., Anderson, E., and Muller, P. K., "Container Shipping: Operating System Support for I/O-Intensive Applications," *IEEE Computer*, Vol. 27, No. 3, March 1994.

Transport Protocol Optimization

Barry Brown

The World Wide Web has seen a remarkable explosion is growth since its conception in 1990. The transport protocols devised at its birth are still in use today. Unfortunately, these protocols are inefficient at delivering the large numbers of small documents which make up the vast majority of the Web. I will be investigating previously suggested methods of increasing efficiency and proposing my own protocol for delivering documents as quickly as possible with minimal load on the network and server CPU.

Study of Multimedia Multicast Network Traffic Routing

Fang-Pin Chang

As multimedia gains momentum in research and its appeal gains popularity and acceptance in the marketplace, the proportion of network traffic that is continuous media in nature (e.g., video, audio, etc.) is increasing. High capacity, low latency networks such as ATM are being developed and installed to specifically accommodate this traffic with the required Quality of Service.

My work in this area involves modeling communications networks supporting continuous media traffic using the OPNET network simulator by MIL 3, Inc. Using simulations, we explore network traffic behavior under various traffic load conditions, and investigate ways to improve, maintain, or fine-tune performance of such networks.

Simulation is a convenient performance evaluation technique for large complex systems. It also provides considerable freedom in configuring and modifying the network environment. OPNET was chosen because of its object-oriented, modular approach, which eases the development of the software models. In addition, OPNET's built-in analysis and presentation tools facilitate data manipulation and visualization.

Experiments are now focusing on networks with simple topologies where the simulation results can, at least partially, be compared to known theoretical ones. Future work will concentrate on evaluating the performance and efficiency of network architectures supporting continuous media dissemination using layered coding and multicasting.

Multimedia Multicast Channel (MMC)
and Multimedia Information Dissemination Infrastructure (MIDI)

Frederick C. Chong

Multimedia Multicast Channel (MMC)

The MMC is a communication abstraction analogous to the cable television channel: a source transmits onto a specific channel and receivers which have subscribed to that channel receive media streams without explicit interactions with the source. The channels, ports, streams and segments concepts form the building blocks of the MMC abstraction. The abstraction provides several benefits, among these are: support receivers with heterogenous needs, open-loop control and multicasting. This communication paradigm is highly appropriate for multimedia distribution services required by a large class of multimedia multipoint applications. I have implemented a C library for the MMC programming interfaces. The entire implementation is done at user-level and requires no modification to the UNIX kernel.

Multimedia Information Dissemination Infrastructure (MIDI)

The goal of MIDI is to implement an infrastructure which can make known MMC channel resources to the WAN community, and disseminate multimedia information efficiently on the internet. The MMC facility is used as the base technology for implementing MIDI. The research work entails building an MMC directory service where channels availability can be learned, and the notion of an MMC router, which knows how to disseminate MMC data without wasting excessive network bandwidth.

A New Service Discipline for Control of Real-Time Traffic in a Packet Switching Network

Norival R. Figueira

Some real-time applications generate network traffic that requires stringent performance guarantees in terms of throughput, end-to-end delay, and packet loss rate. The lack of strict performance guarantees in conventional window-based flow control and first-come-first-served (FCFS) service disciplines motivated the creation of rate-based flow control and rate-based service disciplines. Rate-based flow control gives each session a guaranteed minimum data rate, without being affected by the traffic behavior of the other sessions sharing a server. This guaranteed data rate generally requires some admission control mechanism to allocate the finite link capacity of the servers.

Several rate-based service disciplines have been proposed: Delay-EDD [1], Jitter-EDD [7], RCSP [8], VirtualClock [9], PGPS [6], Stop-and-Go [4], and Hierarchical Round Robin [5]. All of these service disciplines provide an upper bound on end-to-end delay (an upper bound on delay for VirtualClock was recently proved in [2]). Jitter-EDD, RCSP, and Stop-and-Go also provide an upper bound on delay jitter.

The approach used with these service disciplines, which we call here the *old* approach, is based on the characterization of sessions according to some pre-defined traffic characterization model, which is generally expressed in terms of upper bounds (or lower bounds) on traffic measures such as interarrival times of packets, peak rate, etc. In the old approach, a network uses the traffic characterization of a session to allocate network resources to the session or to express the service guarantees of the session, i.e. service guarantees (such as upper bounds on delay, delay jitter, etc.) are expressed as functions of the parameters of the traffic characterization of the session. For instance, Delay-EDD and Jitter-EDD use the minimum packet interarrival time of sessions in the admission control tests, and Stop-and-Go uses a traffic parameter (T) in the service guarantees it advertises (In Stop-and-Go, sessions are required to limit their maximum average rate in any interval of time T).

There is no doubt that a session's traffic must be constrained in order to allow a rate-based service discipline to provide service guarantees. The main problem in the old approach is that the defined traffic characterization model is generally one that simplifies the analysis of the service discipline, which may limit the kinds of sessions for which service guarantees are provided. To avoid this problem, we propose a *new* approach in which the network provides a session with service guarantees that are expressed as functions of the *dynamic traffic behavior* of the session, and no specific traffic characterization model is imposed or needed. This implies a major distinction between the new approach and the old one, since the former does not conceal the dynamic traffic behavior of sessions from the calculus of service guarantees (we show that this approach allows a service discipline to provide upper bounds on the probability distribution of some service measures), while the latter can only provide service guarantees that are expressed as upper bounds on service measures.

To show that our new approach is feasible, we are investigating a new service discipline called Leave-in-Time [3]. The Leave-in-Time service discipline is able to provide a session with service guarantees that are completely determined by the dynamic traffic behavior of the session. The Leave-in-Time service discipline provides upper bounds on delay, delay jitter, buffer space requirements, the probability distribution of delays, and the probability distribution of buffer space requirements. The upper bounds on delay, delay jitter, and buffer space requirements are expressed as functions of an upper bound on delay the session would have in a fixed-rate server.

References

[1] D. Ferrari and D. Verma, "A Scheme for Real-Time Channel Establishment in Wide-Area Networks," *In IEEE JSAC*, Vol. 8, No. 4, pp. 368-379, April 1990.

[2] N. Figueira and J. Pasquale, "An Upper Bound on Delay for the VirtualClock Service Discipline," *To Appear in ACM/IEEE Transactions on Networking*, 1995.

[3] N. Figueira and J. Pasquale, "Leave-in-Time: A Service Discipline for Real-Time Communications in a Packet Switching Data Network," *To Appear in Proceedings of ACM SIGCOMM 1995*.

[4] S. J. Golestani "A Stop-and-Go Queueing Framework for Congestion Management," *In Proceedings of ACM SIGCOMM 1990*, pp. 8-18.

[5] C. Kalmanek, H. Kanakia, and S. Keshav, "Rate Controlled Servers for Very High-Speed Networks," *In Proceedings of IEEE GlobeCom 1990*, pp. 300.3.1-300.3.9.

[6] A. K. Parekh, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks," *Ph.D thesis, Massachusetts Institute of Technology, LIDS-TH-2089*, February 1992.

[7] D. Verma, H. Zhang, and D. Ferrari, "Delay Jitter Control for Real-Time Communication in a Packet Switching Network," *In Proceedings of IEEE TriCom 1991*, pp. 35-43.

[8] H. Zhang and D. Ferrari, "Rate-Controlled Static-Priority Queueing," *In Proceedings of IEEE INFOCOM 1993*, pp. 227-236.

[9] L. Zhang, "VirtualClock: A New Traffic Control Algorithm for Packet Switching Networks," *In ACM Transactions on Computer Systems*, Vol. 9, No. 2, pp. 101-124, May 1991. Also in Proceedings of ACM SIGCOMM 1990, pp. 19-29.

Applications of Hierarchical Coding in High-speed Packet Switching Networks

James K. Han

Hierarchical coding (also referred to as layered coding) of images and continuous media can be used to effectively control congestion in high-speed networks supporting interactive multimedia communications [1]. However, the tradeoffs of the use of hierarchical coding have not yet been adequately investigated. We have undertaken an empirical study to investigate the effectiveness of hierarchical coding from a network perspective using the hierarchical mode of the JPEG standard. A software implementation of hierarchical JPEG (HJPEG) codec has been completed for this study and a static analysis of HJPEG images in terms of Quality of Service (QoS) management and comparison with baseline JPEG (BJPEG) has been performed.

To investigate the performance of the two different modes of JPEG over packet- switching networks, we simulate a switching node as a discrete-time single- server queue where different traffic sources are merged together: Poisson arrivals are used for traditional data traffic and deterministic distributions of interarrival times of frames for Variable Bite Rate (VBR) video sources using motion JPEG. The size of the video frames is obtained from real video sources and from statistical models of video sources. We measured the size of images after compression and the time to compress and decompress using HJPEG and compared them to similar metrics for BJPEG. Our measurements indicate that the overhead resulting from the multiple layers of HJPEG may be insignificant [2].

We are currently focusing on the analysis of hierarchical coding from the perspective of detailed delay distributions, cell loss statistics and measures of parameter sensitivities for the two different coding methods of JPEG. An other important issue for interactive real-time applications over ATM and other packet-switching networks is their behavior when cell or packet loss occurs. We are conducting experiments to investigate this issue.

The implications of these results for network transmission are significant considering the potential gains from statistical multiplexing in ATM-like networks.

References

[1] Pasquale, J. C., Polyzos G. C., Anderson E. W., and Kompella V. P., “The Multimedia Multicast Channel,” *Network and Operating System Support for Digital Audio and Video*, Lecture Notes in Computer Science #712, Springer-Verlag, pp. 197-208, 1992.

[2] Han, J. K. and Polyzos, G. C. “ Networking Applications of the Hierarchical Mode of the JPEG Standard” (submitted for publication).

Wireless Mobile Multimedia Communications

Kenji Hayashi

I am now working in the area of Wireless Mobile Multimedia Communications.

(1)Performance analysis One of the research goals is to know how to use relatively narrow bandwidth effectively in Wireless Mobile communication. To achieve this goal, I am analyzing the wireless communication characteristics by changing communication parameters and environment. Such parameters and environment include packet length, retransmission timers, signal-to-noise ratio, network configuration and so forth. The analysis is based on the data collected by both network simulator and real system. In the next stage, I plan to analyze the characteristics of Multimedia communication over wireless network. In the wireless environment, some of the applications, such as video on demand, may permit a certain amount of data loss as today's TV does. Assuming that an appropriate ACK-less protocol is used for multimedia communication over wireless network, it may improve the throughput of the communication within the limited bandwidth. I plan to analyze the characteristics of such a protocol on a network simulator and real systems.

(2)Mobile IP routing The additional research I am working is routing scheme in Mobile IP. Mobile IP is an enhanced Internet Protocol which allows a node to roam on the Internet without changing its IP address. It is now under standardization in IETF. I am thinking about a modified routing scheme of Mobile IP which reduces redundant routing.

A Study of Queueing Delays in the Service Curve-based Earliest Deadline First Scheduling Discipline

Jesse S. Hose

With the proliferation of multimedia applications, high speed packet switched networks are being called upon to provide a wide variety of services. A crucial problem is the ability to provide applications with quality of service (QOS) guarantees. Many applications, such as packetized voice and video, have stringent delay requirements. Thus, an interesting measure of QOS is the delay experienced by packets in the network.

In this study, through simulation, we investigate the behavior of a scheduling discipline for performance guarantees via service curves[1]. A comparison of delay distributions is made between the Service Curve-based Earliest Deadline First (SC-EDF) discipline and the first-come first-served (FCFS) discipline. A single node case is considered with traditional voice (ON-OFF) source models along with variant models of ON-OFF based sources.

Further study is made into the service curve discipline's ability to isolate a stream's service from being influenced by the traffic characteristics of other streams concurrently being serviced by the node. This property, known as a firewall, is especially important in the presence of streams transmitting at a rate greater than agreed upon at admission time, thus potentially denying conforming sources of their guaranteed service rate. In this study we look at queueing delay only, propagation and packet processing delays are not considered.

Reference

[1] Sariowan, H.; Cruz, R.L.; Polyzos, G.C. "Scheduling for Quality of Service Guarantees via Service Curves" *SIGCOMM paper submission, 1995*,

mmctool: A Testing Program for the Multimedia Multicast Channel

Diego Kaplan

Mmctool is a multipurpose communications program built on top of the MMC library developed by Fred Chong at the CSL lab. It creates a user interface to the MMC API, and allows creation of channels and streams. It allows the user to open ports as SOURCE or RECEIVER into those channels, and manages production and consumption of standard multimedia data types. In its present form, only video and text are supported, but eventually sound and a more flexible stream specification will be added, providing the potential to build interesting connections (i.e. video conferencing) easily.

Nanokernel & Wireless Radio

David T. Moore

Operating system research currently involves modification of an existing large complex or the design of a new one. Both of these tasks can consume large amounts of time. Use of an existing design may prevent you from readily testing your design.

We are currently exploring designing a small, easily extensible kernel. The functionality provided would be as minimal as possible, yet provide sufficient functionality for easily testing OS research ideas.

This operating system should provide primarily mechanism, leaving policy for non-kernel code. Hopefully, the mechanism will work for real-time systems, hundreds of small independent threads, as well as supporting larger systems similar to current OSs.

I am additionally interested in wireless network technologies, and how operating systems and user processes can be designed to work efficiently on this medium.

Descartes

Max Yoshio Okumoto

For our Descartes project, we are interested in creating a small portable kernel over which many interesting operating systems can be implemented. The issues we are interested are realtime, resource protection, resource allocation/deallocation.

I am also interested in communication in a mobile computing enviorment. The cost of 'wireless technologies' has drop significantly in the last few years. I am currently experimenting with the NCR WaveLan products.

Characterizing the I/O Behavior of Scientific Applications

Barbara K. Pasquale

The growing dependence on I/O in high performance systems (e.g., large-scale mathematical modeling and simulation, real-time visualization, and high speed network communication), has focused attention on research and development for high performance I/O subsystems. Finding effective solutions to potential I/O bottlenecks in high performance systems will require an integrated approach. I/O issues must be addressed at all system levels, including applications, operating systems, file systems and disk architectures.

To better understand the sources and characteristics of I/O demand with respect to the processed workload, we focus on application level I/O behavior. Focusing on the behavior of individual I/O intensive applications rather than overall system I/O allows us to identify a specific and important source of I/O demand and describe it in terms of individual applications, computations, and scientific discipline. This high level view of the I/O demand is relatively independent of the underlying architecture and can be used not only to describe current demands, but to study changes in I/O demand relating to specific applications or for classes of applications. Based on current application I/O characteristics and changes in these characteristics over time, one can extrapolate future application I/O requirements and thus determine the overall effects on system I/O workload intensities and patterns.

We have developed a general methodology to isolate the predominant I/O intensive scientific applications from a high performance sequential workload and to characterize both the static and dynamic attributes of their I/O resource usage. First we observe a typical, high performance, production workload to isolate the I/O intensive population. From this population, we select applications which are representative of the major I/O resource consumers with respect to scientific discipline, I/O volume and rate, CPU execution time, and frequency of execution. For each selected application, run-time I/O activity is monitored and profiles are constructed, depicting I/O behavior over time. The graphical representation of these time series profiles enables us to observe the application's characteristic I/O behavior and the presence or absence of regular and recurring cycles of data access and computation. Finally, through statistical analysis of the I/O trace data, the application's characteristics I/O behavior is translated into a model. The resulting models of application I/O behavior can be used to construct artificial workloads for empirical I/O performance studies, and to provide input parameters for analytic and simulation performance models.

Using our methodology, we conducted an extensive I/O workload characterization study of the San Diego Supercomputer Center's current Cray C90 [1] and former Cray YMP [2]. The conclusions we can draw from these studies are as follows.

The set of I/O intensive applications in a supercomputer workload is small, but consumes the vast majority of system resources. Both the Cray YMP February 1992 and Cray C90 January 1994 workloads, contained a small set of I/O intensive applications, representing less than 51992 I/O intensive set was responsible for 71transferred and 10January 1994 I/O intensive set was responsible for 88I/O and 33explained in part by the increase in C90 CPU performance which is estimated to be 1.5 - 2 times faster than the YMP. However, the overlap of seventeen distinctly named applications and their increasing average I/O rates for January 1994 shows that applications are placing more demand on the I/O subsystem.

The applications in the I/O intensive set exhibit a high degree of regularity in resource usage, including I/O. Regression analysis reveals that individual applications I/O demand, in terms of logical I/O requests and number of bytes transferred, shows a high degree of correlation with CPU time consumed. Cluster analysis revealed three general patterns in individual application resource usage: 1) fairly stable logical I/O rates, 2) short executions on the order of seconds together with long executions on the order of minutes, and 3) consistent memory usage.

The applications from atmospheric sciences consume the most resources with respect to the I/O intensive set and to the entire workload. The applications in the I/O intensive set span 12 difference scientific disciplines. Although it ranks number six in terms of the number of applications executed, atmospheric sciences ranks number one in terms of I/O and CPU time usage. These applications are responsible for 27of I/O intensive bytes transferred and 24

The dynamic I/O behaviors of applications investigated could be modeled by regular cycles of data accesses and computation periods. Three atmospheric sciences applications were studied, one analysis code and two simulation codes. Although the I/O behaviors of these codes are different, each exhibited a basic cycle of I/O activity. The individual I/O events within the cycle are regular and predictable as well as the duration of computations periods. On every iteration, the same files are accessed, in the same sequence, transferring the same amount of data. In addition, all three applications used large data files (e.g., tens and hundreds of megabytes), accessed files sequentially, read input data incrementally (i.e., data staging) and periodically checkpointed the current state of the computation.

References

[1] Pasquale, B. K. and Polyzos, G. C. "Dynamic I/O Characterization of I/O Intensive Scientific applications," *Proceedings of Supercomputing '94*, Washington, DC, November 1994, pp. 660–669.

[2] Pasquale, B. K. and Polyzos, G. C. "A Static Analysis of I/O Characteristics of Scientific Applications in a Production Workload," *Proceedings of Supercomputing '93*, Portland, OR, November 1993, pp. 388–397.

Scheduling for Quality of Service Guarantees via Service Curves

Hanrijanto Sariowan

A new scheduling policy which provides guarantees to virtual circuits in packet-switched networks is proposed. This scheduling policy, called the SC-EDF policy, is developed for a general framework for service provisioning based on Service Curves, proposed by Cruz [1]. Instead of explicitly guaranteeing specific quality of service measures, such as maximum delay and throughput, SC-EDF guarantees the service curve for a connection. The service curve guarantee, combined with constraints on the traffic burstiness of a connection, can then be translated into quality of service guarantees. The service curve specification gives flexibility to a network element in allocating its resources to meet various delay and throughput requirements of the connections sharing the resources. In contrast to many scheduling disciplines which only provides a lower bound on throughput, SC-EDF decouple delay and throughput requirements of a connection, leading to a larger schedulability region. It can be shown that SC-EDF provides a larger schedulability region than scheduling policies such as VirtualClock [2] and PGPS [3] with rate proportional assignment. A simple and convenient condition under which SC-EDF can simultaneously guarantee a set of service curves is proved. This condition is needed for the connection admission policy and thus complements the scheduling policy, making SC-EDF a complete scheme for provisioning of performance guarantees. The service curve-based scheduling in a single network element can be extended to the end-to-end scheduling based on the notion of *network* service curves. The main issue which is currently under investigation is the sufficient condition under which a set of arbitrary network service curves can be simultaneously guaranteed.

References

[1] R. L. Cruz, "Quality of service guarantees in virtual circuit switched networks," % Invited paper to appear in *IEEE Journal on Selected Areas in Communications*, 'Advances in the Fundamentals of Networking'.

[2] L. Zhang, "VirtualClock: A new traffic control algorithm for packet switching networks," *ACM Transactions on Computer Systems*, vol. 9, no. 2, May 1991, pp. 101-124.

[3] A. K. Parekh and R.G. Gallager, " A generalized processor sharing approach to flow control in integrated services networks: the single-node case," *IEEE/ACM Transactions on Networking*, vol.1, no.3, June 1993, pp. 372-85.

Architectures and Service Models for New Generation Networks

George Xylomenos

My research interests revolve around the largely complementary topics of building distributed operating systems and systems supporting distributed applications. This entails research not only in operating systems architecture and implementation, but it also requires close investigation of the enabling technology for distributed systems, namely, the new generation of high-bandwidth and low-latency communications networks. These emerging networks are increasingly making the original ARPAnet architecture and its associated protocols obsolete. To harness all this new hardware power we need to proceed in two directions: first, we must avoid wasting resources by optimizing the system software to deal with very high-speed networks; second, we must offer to the users an advanced service model, to enable them to expand the abilities of their applications in ways not supported by the current network architectures.

The first issue is one that also arises in the context of I/O systems, and a large body of research has already been devoted to operating system optimizations for solving the problems that it encompasses. The second issue on the other hand is not so well-studied, largely because it requires that we reconsider our ideas on what a network can do for its users. But, if we are to replace the current Internet architecture to use more efficient and technically advanced hardware and software solutions, we get for free the chance to provide to all network users an advanced service model that will ease the writing and use of distributed applications.

New generation applications that have been proposed to take advantage of high-speed networks, like Digital Continuous Media communications and Multiparty Conferencing, require hard or soft real-time guarantees on delays and multicast capabilities, among other things. Generally, distributed applications call for advanced functionality, such as multipoint communications, and Quality of Service guarantees, such as delay and error loss rate bounds. A network capable of providing these services would need to support mechanisms for multicast routing, resource reservations, admission control and packet scheduling policies, coupled with protocols for specification and negotiation of the services requested by applications [1].

Although it is hard to separate these software components from one another, as the services provided by each one will have an effect on the others, by decoupling we enable them to collaborate with multiple versions of the other modules. This decoupling makes possible the use of any new component even on the current Internet, given some kind of support from other experimental modules that are being constantly developed; this is probably the best way to test the abilities of the components in the real world and make them accepted by users and researchers.

My research currently focuses on two of the above components: multicast routing and resource reservations. There are many things to be done in these areas, including development and analysis of protocols, integration with existing hardware and software, investigation of the interactions between the components and assuring compatibility and interoperability among pieces of a heterogeneous network, including WANs, MANs and LANs, from rusty Ethernets to high-tech ATMs. Naturally, these interests touch also on the other components and even on the form of the new network architectures and service models proposed. I hold the view that all proposals in this area should be eventually validated on real networks using actual applications and workloads; the way to make preliminary theoretical evaluations predict with more accuracy the behavior of the real world, may be an even more significant and rewarding research topic.

References

[1] L. Zhang, S. Deering, D. Estrin, S. Shenker and D. Zappala "RSVP: A New Resource ReSerVation Protocol," *IEEE Network*, Vol. 7, No. 5, 1993, pp. 8–18.

8 CSL Publications

[B92] Ben Taarit, R. "Extending FDDI Synchronous Mode Guarantees to Interconnected Rings," M.Sc. Thesis, Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, September 1992.

[BP92] Billard, E. and Pasquale, J., "Effects of Periodic Communication on Distributed Decision-making," *Proc. IEEE International Conference on Systems, Man, and Cybernetics*, Chicago, October 1992, pp. 49-54.

[BP93] Billard, E. and Pasquale, J., "Effects of Delayed Communication in Dynamic Group Formation," *IEEE Trans. Systems, Man, and Cybernetics*, accepted for publication, January 1993.

[BP93a] Billard, E. and Pasquale, J., "Dynamic Scope of Control in Decentralized Job Scheduling," *Proc. 1st International Symposium on Autonomous Decentralized Systems*, Kawasaki, Japan, March 1993.

[B+93] Braun, H. W., Chinoy, B., Claffy, K., and Polyzos, G. C., "Analysis and Modeling of Wide-Area Networks: Annual Status Report," Technical Report GA-A21224, SDSC and UCSD, La Jolla, CA, February 1993.

[B+92] Braun, H. W., Claffy, K., Chinoy, B., and Polyzos, G. C. "Analysis and Modeling Tools for High-Speed Networks: Project Status Report," Technical Report CS92-237, Dept. of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, April 1992.

[BCP93] Braun, H. W., Claffy, K., and Polyzos, G. C., "A Framework for Internet Accounting," Proceedings of *SICON'93,* Singapore, pp. 847–851, September 1993.

[CBP93] Claffy, K.C., Braun, H.-W., and Polyzos, G.C., "Long-term Traffic Aspects of the NSFNET," Proceedings of *Inet'93,* San Francisco, CA, August 1993.

[CBP94] Claffy, K.C., Braun, H.-W., and Polyzos, G.C., "Tracking Long-term Growth of the NSFNET," *Communications of the ACM,* vol. 37, no. 8, pp. 34–44, August 1994. (Special Issue on Internet Technology).

[CBP95] Claffy, K.C., Braun, H.-W., and Polyzos, G.C., "A Parameterizable Methodology for Internet Traffic Flow Profiling," *IEEE Journal on Selected Areas in Communications*, 1995 (in press).

[CP92] Claffy K., and Polyzos, G. C., "Location Transparent Connection Management: A Survey of Protocol Issues," *Proc. 11th Annual IEEE International Phoenix Conference on Computers and Communications*, Scottsdale, AZ, April 1-3, 1992.

[CPB93a] Claffy, K., Polyzos, G. C., and Braun, H. W., "Traffic Characteristics of the T1 NSFNET Backbone," *Proc. IEEE INFOCOM '93*, San Francisco, CA, March 28 - April 1, 1993. (Also as Technical Report CS 92-252, Dept. of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, July 1992)

[CPB93b] Claffy, K., Polyzos, G. C., and Braun, H. W., "Measurement Considerations for Assessing Unidirectional Latencies," *Internetworking: Research and Experience*, vol. 4, no. 3, pp. 121-132, 1993.

[CPB93c] Claffy, K., Polyzos, G. C., and Braun, H. W., "Application of Sampling Methodologies to Network Traffic Characterization," Proceedings of *ACM SIGCOMM'93,* San Francisco, CA, pp. 194–203, September 1993.

[E93] Etherton, D. C., "Approximating FDDI Synchronous Mode Guarantees Through Kernel Modification," M.Sc. Thesis, Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, March 1993.

[FP93] Fall, K., and Pasquale, J., "Exploiting In-Kernel Data Paths to Improve I/O Throughput and CPU Availability," *Proc. USENIX Winter Conference*, San Diego, January 1993.

[FP94] Fall, K., and Pasquale, J., "Improving continuous-media playback performance with in-kernel data paths ," *Proc. IEEE Intl. Conf. on Multimedia Computing and Systems (ICMCS)*, Boston, MA, June 94, pp. 100-109.

[FPP92] Ferrari, D., Pasquale, J., and Polyzos, G. C., “Network Issues for Sequoia 2000,” *Proc. IEEE Compcon Spring ’92*, San Francisco, CA, February 24-28, 1992.

[FP95] Figueira, N., and Pasquale, J., "An Upper Bound on Delay for the VirtualClock Service Discipline," *ACM/IEEE Transactions on Networking*, 1995 (in press).

[FP95a] Figueira, N., and Pasquale, J., "Leave-in-Time: A Service Discipline for Real-Time Communications in a Packet Switching Data Network," *Proceedings of ACM Communications Architectures and Protocols Conf. (SIGCOMM)*, Cambridge, MA, September 95, to appear.

[G90] Glockner, A. “Coadaptive Decision-making in Distributed Computer Systems,” Ph.D. Thesis, Department of Computer Science and Engineering, University of California, San Diego, December 1990.

[GP92] Glockner A., and Pasquale, J., “Co-adaptive Decision-making in Distributed Systems,” *IEEE Transactions on Systems, Man, and Cybernetics*, accepted for publication.

[GP93] Glockner, A., and Pasquale, J., “Coadaptive Behavior in a Simple Distributed Job Scheduling System,” *IEEE Transactions on Systems, Man, and Cybernetics*, March 1993.

[GP94] Groschwitz, N., and Polyzos, G.C., “A Time-Series Model of Long-term Traffic on the NSFNET Backbone,” Proceedings of the *IEEE International Conference on Communications (ICC’94),* New Orleans, LA, May 1994.

[J92] Johnson, M. M., “FDDI Performance Guarantees for Real-Time Traffic in the Absence of a Synchronous Mode,” M.Sc. Thesis, Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, December 1992.

[KPP92] Karvelas, D., Papamichail, M., and Polyzos, G. C. “The Rotating Slot Generator Dual Bus,” *Proc. of the 23rd Pittsburgh Conference on Modeling and Simulation*, Pittsburgh, PA, April 1992.

[KPP92c] Karvelas, D., Papamichail, M., and Polyzos, G. C., “Performance Analysis of the Rotating Slot Generator Scheme,” *Proc. IEEE INFOCOM ’92*, Florence, Italy, May 4-8, 1992.

[KPP95] Karvelas, D., Papamichail, M., and Polyzos, G. C., “Fair DQDB: The Rotating Slot Generator Scheme,” *Journal of High Speed Networks*, 1995 (in press).

[KP92] Kay, J., and Pasquale, J., “Network Sloth: Where the Time Goes in Network Software,” Technical Report CS92-238, Dept. of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, April 1992.

[KP93] Kay, J., and Pasquale, J., “Measurement, Analysis, and Improvement of UDP/IP Throughput for the DECstation 5000,” *Proc. USENIX Winter Conference*, San Diego, January 1993, pp. 249-258.

[KP93a] Kay, J., and Pasquale, J., “A Summary of TCP/IP Networking Software Performance for the DECstation 5000,” *ACM Conference on Measurement and Modeling (SIGMETRICS)*, (extended abstract), Santa Clara, CA, May 1993.

[KP93b] Kay, J., and Pasquale, J., “The importance of non-data touching processing overheads in TCP/IP,” *Proc. ACM Communications Architectures and Protocols Conf. (SIGCOMM)*, San Francisco, CA, September 93, pp. 259-269.

[KPP91] Kompella, V. P., Pasquale, J. C., and Polyzos, G. C., “Multicast Routing for High-Bandwidth Delay-Sensitive Applications,” Technical Report CS91-185, Dept. of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, March 1991.

[KPP92a] Kompella, V., Pasquale, J., and Polyzos, G., “Multicasting for Multimedia Applications,” *Proc. IEEE INFOCOM Conference on Computer Communications*, Florence, Italy, May 1992, pp. 2078-2085.

[KPP93] Kompella, V. P., Pasquale, J. C., and Polyzos, G. C., “Multicast Routing for Multimedia Communication,” *IEEE/ACM Transactions on Networking,* vol. 1, no. 3, pp. 286–292, June 1993.

[KPP93a] Kompella, V. P., Pasquale, J. C., and Polyzos, G. C., "Two Distributed Algorithms for Multicasting Multimedia Information," Proceedings of the *Second International Conference on Computer Communications and Networks,* San Diego, CA, pp. 343–349, June 1993.

[M90] Muller, K. "Multi-structured File System Design for High Performance Computing," Ph.D. Thesis, Department of Computer Science and Engineering, University of California, San Diego, December 1990.

[MP91] Muller K., and Pasquale, J., "A High-performance Multi-structured File System Design," *Proc. 13th ACM Symposium on Operating System Principles (SOSP)* , Asilomar, CA, October 1991, pp. 56-67.

[PP93a] Pasquale, B. K. and Polyzos, G. C., "A Static Analysis of I/O Characteristics of Scientific Applications in a Production Workload," *Proceedings of Supercomputing '93*, pp. 388-397, Portland, Oregon, November 15-19, 1993. (Also appears as "A Static Analysis of I/O Characteristics of a Broad Class of Scientific Applications," Technical Report No. CS92-270, University of California, San Diego, Nov. 1992, and Project Sequoia 2000 Technical Report No. 92/19, University of California, Berkeley, Nov. 1992.)

[PP93b] Pasquale, B. K. and Polyzos, G. C., "I/O Profiles of a Scientific Application on a Workstation and a Supercomputer," Technical Report No. CS93-299, Dept. of Computer Science and Engineering, University of California, La Jolla, CA, July 1993. (pre-MASCOTS version)

[PP93c] Pasquale, B. K. and Polyzos, G. C., "Real-time Dissemination of Continuous Media in Packet-Switched Networks," Proceedings of *IEEE Compcon Spring '93,* San Francisco, CA, February 1993.

[PP94a] Pasquale, B. K. and Polyzos, G. C., "A Case Study of Scientific Application I/O Behavior," *Proceedings of MASCOTS '94, International Workshop on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*, pp. 101-105, Durham, North Carolina, January 31-February 2, 1994. (Also appears as Project Sequoia 2000 Technical Report No. 93/36, University of California, Berkeley, December 1993.)

[PP94b] Pasquale, B. K. and Polyzos, G. C., "Dynamic I/O Characterization of I/O Intensive Scientific Applications." *Proceedings of Supercomputing '94*, pp. 660-669, Washington, DC, November 14-18, 1994. (Also appears as Technical Report No. CS94-364, University of California, San Diego, April 1994.)

[P88] Pasquale, J., "Using Expert Systems for Managing Distributed Systems," *IEEE Network*, Vol. 2, No. 5, Sept. 1988.

[P89] Pasquale, J., "The Effects of Using Aging State Information in Decentralized Load Balancing," Technical Report CS89-161, Dept. of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, October 1989.

[P89b] Pasquale, J., "Decentralized Resource Management Research in Large Distributed Computer Systems," Technical Report CS89-162, Dept. of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, December 1989.

[P90] Pasquale, J., "Beyond Client Server Architectures," *NCR Perspective*, Vol. 4, No. 2, Spring 1990.

[P91] Pasquale, J., "Decentralized Control in Large Distributed Systems," *Coordination Theory and Collaboration Technology Workshop Proceedings*, National Science Foundation, Washington D.C., June 1991, pp. 9-20.

[P91a] Pasquale, J., "Decentralized Control in Large Distributed Systems," *Proc. NSF Coordination Theory and Collaboration Technology Workshop*, National Science Foundation, Washington D.C., June 1991.

[P92b] Pasquale, J., "I/O System Design for Intensive Multimedia I/O," Proceedings *IEEE Workshop on Workstation Operation Systems (WWOS)*, Key Biscayne, FL, April 1992, pp. 29-33.

[P93] Pasquale, J., "Randomized Coordination in an Autonomous Decentralized System," *Proc. 1st International Symposium on Autonomous Decentralized Systems*, Kawasaki, Japan, March 1993.

[P93a] Pasquale, J., "System Software and Hardware Support Considerations for Digital Video and Audio Computing," *Proc. Hawaii International Conference on System Sciences (HICSS)*, Maui, Hawaii, January 1993.

[PAM94] Pasquale, J., Anderson, E., and Muller, P. K., "Container Shipping: Operating System Support for I/O-Intensive Applications," *IEEE Computer*, Vol. 27, No. 3, March 1994, pp. 84-93

[P+90] Pasquale, J., Anderson, E., Bubien, M., Fall, K, Kompella, V., McMullan, S., Ranganathan, D., and Terek, R. "Operating System and Window System Research for Multi-media Applications: A Status Report," Technical Report CS90-176, Dept. of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, July 1990.

[PBK91] Pasquale, J., Bittel, B., and Kraiman, D. "A Workload Characterization Study of a Cray XMP Running CTSS," Technical Report CS91-222, Dept. of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, April 1991.

[PBK91b] Pasquale, J., Bittel, B., and Kraiman, D., "A Static and Dynamic Workload Characterization of the SDSC Cray XMP," *1991 ACM SIGMETRICS Proceedings* (extended abstract), San Diego, May 1991.

[PFF92] Pasquale, J., Fall, K., and Forrest, J., "The Sequoia 2000 Network (S2Knet) Handbook," Sequoia 2000 Technical Report 92/5, U. C. Berkeley, November 1992.

[PP90] Pasquale, J. and Polyzos, G. C. "System Support for Multimedia Applications," *Proc. First International Workshop on Network and Operating System Support for Digital Audio and Video*, International Computer Science Institute, Berkeley, CA, November 8-9, 1990.

[P+91] Pasquale, J., Polyzos, G. C., Anderson, E. W., Fall, K. R., Kay, J. S. , Kompella, V. P., McMullan, S. R., and Ranganathan, D. "Network and Operating System Support for Multimedia Applications," Technical Report CS91-186, Dept. of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, March 1991.

[PPAK92] Pasquale, J., Polyzos, G. C., Anderson, E. W., and Kompella, V. P., "The Multimedia Multicast Channel," *Proc. Third International Workshop on Network and Operating System Support for Digital Audio and Video*, La Jolla, CA, November 12-13, 1992. (Reprinted in Network and Operating System Support for Digital Audio and Video, Lecture Notes in Computer Science #712, Springer-Verlag, pp. 197–208.)

[PPAK94] Pasquale, J., Polyzos, G. C., Anderson, E. W., and Kompella, V. P., "The Multimedia Multicast Channel," *Journal of Internetworking: Research and Experience*, Vol. 5, No. 4, pp. 151-162, December 1994.

[PPAK94a] Pasquale, J., Polyzos, G. C., Anderson, E. W., and Kompella, V. P., "Filter Propagation in Dissemination Trees: Trading Off Bandwidth and Processing in Continuous Media Networks," Proceedings of the *Fourth International Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV),* D. Shepherd, G. Blair, G. Coulson, N. Davies and F. Garcia (eds), Lecture Notes in Computer Science, Vol. 846, Springer-Verlag, 1994.

[P+91b] Pasquale, J., Polyzos, G. C., Fall, K., Kompella, V. "Internet Throughput and Delay Measurements Between Sequoia 2000 Sites," Technical Report CS91-219, Dept. of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, December 1991.

[PPFK91] Pasquale, J., Polyzos, G., Fall, K., and Kompella, V., "Internet Throughput and Delay Measurements Between Sequoia 2000 Sites," Sequoia 2000 Technical Report 91/7, U. C. Berkeley, December 1991 (also available as U. C. San Diego Technical Report CS91-219, December 1991).

[PPK93] Pasquale, J., Polyzos, G. C., and Kompella, V. P., "Real-time Dissemination of Continuous Media in Packet-Switched Networks," *Proc. IEEE Compcon Spring '93*, San Francisco, CA, February, 1993.

[P92] Polyzos, G. C., "Networking for Sequoia 2000," *Proc. 4th IEEE COMSOC International Workshop on Multimedia Communications*, MULTIMEDIA '92, Monterey, CA, April 1-4, 1992.

[P93b] Polyzos, G. C., "Hierarchical Coding of Images and Continuous Media for Transmission over Packet-Switching Networks," Technical Report CS 92-273, Dept. of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, January 1993.

[P94] Polyzos, G. C., "Image Browsing and Animation over a Local ATM Testbed," Proceedings of the *Fifth IEEE COMSOC International Workshop on Multimedia Communications, MULTIMEDIA'94,* Kyoto, Japan, May 1994.

[P95] Polyzos, G. C., "Continuous Media Dissemination over ATM Networks," Proceedings of the *Seventh IEEE Workshop on Local and Metropolitan Area Networks,* Marathon, FL, March 1995.

[PKP93] Polyzos, G. C., Karvelas, D., and Papamichail, M., "R-DQDB: Fair DQDB through Rotation of the Responsibility for Slot Generation," Proceedings of the *Sixth IEEE Workshop on Local and Metropolitan Area Networks,* La Jolla, CA, October 1993.

[PMV87] Polyzos, G.C., Molle, M.L., and Venetsanopoulos, A.N., "Performance Analysis of Finite Nonhomogeneous Population Tree Conflict Resolution Algorithms Using Constant Size Window Access," *IEEE Transactions on Communications*, Vol. COM-35, No. 11, pp. 1124-1138, November 1987.

[PM90] Polyzos, G. C. and Molle, M. L. "Conflict Resolution Algorithms for High Error-rate Multi-access Channels," *IEEE International Symposium on Information Theory*, San Diego, CA, January 1990.

[PM90a] Polyzos, G. C. and Molle, M. L. "Delay Analysis for the FCFS 0.487 Conflict Resolution Algorithm," Technical Report CS90-179, Dept. of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, October 1990.

[PM93] Polyzos, G. C. and Molle, M. L., "A Queueing Theoretic Approach to the Delay Analysis for the FCFS 0.487 Conflict Resolution Algorithm," *IEEE Transactions on Information Theory*, Vol. 39, no. 11, pp. 1887-1906, November 1993.

[PM94] Polyzos, G. C., and Molle, M. L., "A Queueing Theoretic Methodology for Performance Analysis of Separable Window Access Conflict Resolution Algorithms with Variable Length Elementary Events," *Queueing Systems: Theory and Applications,* vol. 16, pp. 339–361, 1994.

[PT95] Polyzos, G.C., and Taylor, K., "A Prototype Video Dissemination Application over ATM," Proceedings of the *IEEE International Conference on Communications (ICC'95),* Seattle, WA, June 1995 (in press).

[TP91] Terek, R., and Pasquale, J., "Experiences with Audio Conferencing Using X, UNIX, and TCP/IP over an Ethernet," *1991 Summer Usenix Proceedings*, Nashville, June 1991.