Guest Editorial
Next-Generation IP Switches and Routers

I. Introduction

UNTIL a few years ago, the performance of the Internet was limited by the speed of the long-haul links, not by the capacity of switches and routers. Things have changed: with the rapid provisioning of higher performance links (such as OC-48c) and the introduction of wavelength division multiplex (WDM), the bottleneck has moved from the links to the switches and routers. The response from network equipment vendors and the research community has been twofold. On one hand, there are new products and techniques that provide differentiated qualities-of-service (QoS's) over existing congested links. And on the other hand, there are new protocols for fast switching and routing (such as IP switching and label swapping) and novel architectures (such as switched backplanes, and multistage fabrics). The papers in this special issue reflect both of these developments.

There are a broad range of architectures and techniques being proposed. For example, recent proposals show ways to construct switching devices with aggregate bandwidths in excess of 1 Terabit/s, orders of magnitude beyond the routers that constitute the Internet core of today. At the processing level, we see arrays of the latest, most powerful processors competing with semicustom circuits to implement packet-forwarding, classification, and scheduling algorithms. At the protocol level, many believe that some form of label-swapping is required for traffic engineering or for high throughput. Other researchers believe that conventional per-hop routing is both viable and preferable when combined with the latest in switching technology, allowing simpler integration into current networks.

This special issue received paper submissions from all aspects with the common goal to improve today's Internet infrastructure. The selected papers cover seven broad areas of technologies that strive for this goal.

II. Router Architectures

New router architectures are introduced to meet the demand of the ever-growing bandwidth requirement of the Internet. Transforming from the centralized architecture in the traditional router design, two major approaches have been taken to realize the high-performance routers: distributed- and parallel-router architectures. The distributed architecture utilizes hardware-based forwarding engines and switching fabric to achieve its high throughput, where the forwarding engine is part of the line-card architecture. The parallel architecture maintains separate banks of forwarding engines and line cards, which are interconnected by switching fabrics. This separation allows the forwarding engines to be shared among multiple line cards, thus increasing the port density. On the other hand, a high-speed line card can use multiple forwarding engines to support a higher port throughput.

Chan *et al*. develop an analytical framework for modeling and analyzing the impact of technical factors on the cost-performance tradeoffs in distributed- and parallel-router architectures in their paper, "A Framework for Optimizing the Cost and Performance of Next-Generation IP Routers."

III. Switching Architectures

In their paper, "Matching Output Queueing with a Combined Input/Output-Queued Switch," Chuang *et al*. demonstrate that a combined input/output-queueing (CIOQ) switch running twice as fast as an input-queued switch can provide precise emulation of a broad class of packet-scheduling algorithms, including weighted-fair queueing (WFQ) and strict priority queueing. The authors introduce a variety of algorithms that configure the crossbar so that emulation is achieved with a speedup of two and consider their running time and implementation complexity. They also show that the exact emulation holds for all input traffic patterns.

In their paper, "Linear-Complexity Algorithms for QoS Support in Input-Queued Switches with No Speedup," Kam and Siu present several fast and practical linear-complexity scheduling algorithms that enable provision of various QoS guarantees in an input-queued switch without any speedup. The authors derive their innovation from judicious choices of edge weights in a bipartite matching problem, where the edge weights are certain functions of the amount and waiting times of queued cells and credits received by a virtual circuit. By using a linear-complexity variation of the well-known stable-marriage matching algorithm, the authors show the edge weights are bounded by proofs and simulations.

In their paper, "On the Speedup Required for Work-Conserving Crossbar Switches," Krishna *et al*. present a novel crossbar arbitration algorithm, named the lowest occupancy output first algorithm (LOOFA), which is work conserving for all traffic patterns and switch sizes for a speedup of only two. To provide a wide range of delay and fairness properties, the authors present several schemes that can be used in combination with LOOFA. They also describe a mechanism to provide delay bounds for real-time traffic, and these delay bounds can be achieved without requiring output-buffered switch emulation.

IV. Address-Lookup Algorithms

A self-contained problem of considerable practical and theoretical interest that is related to the forwarding speed of IP traffic is the longest-prefix lookup operation, which is perceived as one of the decisive bottlenecks in IP forwarding.

Publisher Item Identifier S 0733-8716(99)06869-9.

0733–8716/99$10.00 © 1999 IEEE

Tzeng and Przygienda survey the novel lookup algorithms and classify them based on the applied techniques, accompanied by a set of practical requirements which are critical in design of high-speed routing devices. In their paper, "On Fast Address-Lookup Algorithms," the authors also present two novel approaches to the problem, giving two new possible tradeoffs in the solution space. The first new approach utilizes a prefix-trie compression technique to reduce the memory requirement and the number of memory accesses to support fast address lookup. The second approach is based on multiresolution tries or multibit tries, where update and lookup of the tries can be performed in parallel, without locking any part of the trie structure.

In the paper "IP-Address Lookup Using LC-Tries," Nilsson *et al.* demonstrates that IP routing tables can be succinctly represented and efficiently searched by structuring them as level-compressed tries. For a large class of distributions the average depth of an LC-trie is $O(\log \log n)$, where n is the number of entries in the table. The authors' experiments show that in some cases the average depth is smaller for a larger table. The authors also discuss modifications required to support more general classifications of packets, needed for link sharing, quality of service provisioning, and for multicast and multi-path routing.

The paper "A Novel IP-Routing Lookup Scheme and Hardware Architecture for Multigigabit Switching Routers" by Huang *et al.* presents a novel address-lookup mechanism that only needs very small SRAM and can be implemented in a pipelined skill in hardware. In their paper, the authors show that most address lookups can be accomplished in one memory access and only require three memory accesses in the worst case.

V. Packet Scheduling

Chao *et al*. propose a novel RAM-based searching engine (RSE) for packet-fair queuing implementation in their paper, "Design of Packet-Fair Queuing Schedulers Using a RAM-Based Search Engine." The RSE employs commercial memory chips and uses the concept of hierarchical searching with a tree data structure such that the total time complexity is independent of the number of sessions in the system. With the extension of the RSE, the authors propose a two-dimensional RSE architecture to implement a general shaper-scheduler. The authors consider the problems of time-stamp overflow and aging in implementing packet-fair queuing algorithms, and they present an implementation architecture for the RSE to demonstrate the feasibility of the proposed approach.

Liebeherr *et al*. propose a novel approach to approximate the sorted queue of an earliest-deadline-first scheduler. Their paper "Priority Queue Schedulers with Approximate Sorting in Output-Buffered Switches," suggests using a set of prioritized first-in/first-out (FIFO) queues that are periodically relabeled. Since the scheme can be efficiently operated with a constant number of pointer manipulations, the hardware implementation becomes feasible. They also investigate the necessary and sufficient conditions for the worst-case delays of the scheduler with the approximate sorting, in addition to some numerical results based on some MPEG video traces.

In the paper by Stephens *et al*., "Implementing Scheduling Algorithms in High-Speed Networks," the authors present an architecture to implement a low-complexity high-performance WFQ scheduler. They do this by restricting the number of guaranteed rates that are supported. For ATM networks, this helps simplify the job of sorting time stamps, policing, and choosing the next cell to leave.

VI. Buffer Management

Limited-buffering capability and low-latency requirements demand better buffer management schemes for high-level protocols such as TCP, HTTP, FTP, Telnet, etc. The paper, "Buffer Management Schemes for Supporting TCP in Gigabit Routers with Per-Flow Queueing," investigates the extent to which fair queueing (and its variants) can be employed in conjunction with appropriately tailored buffer management schemes. In the paper, Suter *et al*. suggest two schemes, FQ-RND and FQ-LQD, and provide the best performance in terms of throughput and fairness in various scenarios. They also conclude that FCFS-RED, FQ-RED, or any other global dropping policy performs very poorly when TCP competes with more aggressive sources or in asymmetric networks with perpetually congested reverse paths.

VII. Label-Switching Routers

Label switching enables high-speed packet forwarding based on the fixed length label attached to each packet. A label-switching router (LSR) can forward traffic based on its layer-3 addresses (maybe in conjunction with higher layer information) or the label attached to it. Nagami *et al*. consider both the traffic-driven and topology-driven mapping policies based on real backbone traffic traces in the paper "Flow Aggregated, Traffic Driven Label Mapping in Label-Switching Networks." Their "flow aggregated traffic driven" policy effectively reduces the number of required labels and increases the cut-through ratio. Triggered by the actual arrival traffic, the policy maps the same label for an aggregated packet stream toward a specific destination.

A "VC-merging capable" LSR can map many IP routes into the same VC label. The paper "Performance Issue in VC-Merge Capable Switches for Multiprotocol Label Switching" investigates the impact of VC merging on the buffering requirement for the reassembly buffers. Widjaja *et al.* propose a realistic architecture to support VC merging and also study the performance of the proposed architecture based on both analytical and simulation.

VIII. Firewall

To keep out unwanted IP traffic, an IP firewall filters the traffic based on packet characteristics. Firewalls usually process entire IP packets, so in an ATM environment cells would then need to be reassembled into packets before firewall processing occurs. Although there are next-generation routers that can filter IP packets at wire speed, it is not always desirable to terminate the layer-2 traffic

at the firewall. ATM traffic can instead be filtered at the signaling phase of connection setup. Xu *et al*. extend the basic ATM functionality to perform packet-level (IP) filtering at the maximum throughput of 2.88 Gbit/s in the paper "Design and Evaluation of A High-Performance ATM Firewall Switch and Its Applications." Since the ATM firewall switch can filter IP traffic without terminating the ATM connection, data can flow through the ATM cloud at high speed and low latency.

ACKNOWLEDGMENT

The Guest Editors received a large number of submissions, and many high quality papers could not be accepted. They would like to thank the authors for their submissions and the reviewers for their high quality reviews.

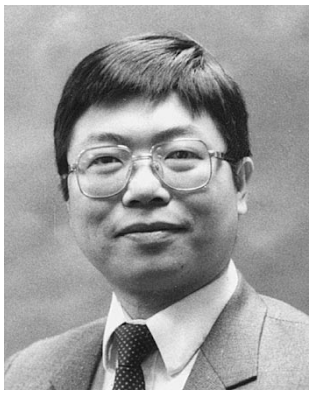
H. JONATHAN CHAO, *Guest Editor*
Polytechnic University
Brooklyn, NY 11201 USA

MIKAEL DEGERMARK, *Guest Editor*
Lulea University of Technology
Sweden

NICK MCKEOWN, *Guest Editor*
Stanford University
Stanford, CA 94305 USA

HENRY HONG-YI TZENG, *Guest Editor*
Bell Labs
Holmdel, NJ 07733-3030 USA

R. RAMASWAMI, *J-SAC Board Representative*



H. Jonathan Chao (S'82–M'85–SM'95) received the B.S.E.E. and M.S.E.E. degrees from National Chiao Tung University, Taiwan, in 1977 and 1980, respectively, and the Ph.D. degree in electrical engineering from The Ohio State University, Columbus, in 1985.

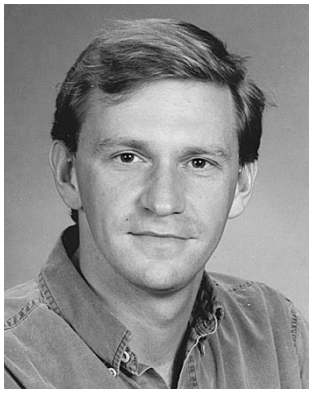
He is a Professor in the Department of Electrical Engineering at Polytechnic University, Brooklyn, NY, which he joined in January 1992. His research interests include large-capacity packet switches and routers, packet scheduling and buffer management, and congestion-flow control in IP/ATM networks. From 1985 to 1991, he was a Member of Technical Staff at Bellcore, Red Bank, NJ, where he conducted research in the area of SONET/ATM broadband networks. He was involved in architecture designs and ASIC implementations, such as the first SONET-like framer chip, ATM layer chip, and sequencer chip (the first chip handling packet scheduling). He received the Bellcore Excellence Award in 1987.

Dr. Chao served as Guest Editor for IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS for the special topic, "Advances in ATM Switching Systems for B-ISDN" (June 1997). He is currently serving as Editor for IEEE/ACM TRANSACTIONS ON NETWORKING.

Mikael Degermark received the M.Sc and Ph.D. degrees in computer science from Lulea University of Technology, Sweden, in 1987 and 1997, respectively.

He currently holds positions at Lulea University of Technology, the Swedish Institute of Computer Science, and Effnet AB. His research interests include protocols, algorithms, and efficient implementations for high-speed networks and mobile (radio) networks. He has worked on address-lookup algorithms, packet classification, resource reservation problems, and header-compression algorithms.

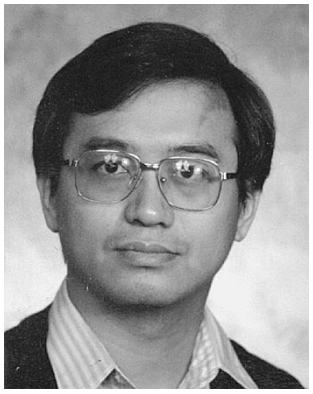
Dr. Degermark received two student paper awards while pursuing the Ph.D degree. (Sigcomm'97, MobiCom'96), and he has won research awards in Sweden.



Nick McKeown (S'91–M'95–SM'97) received the Ph.D. from the University of California, Berkeley, in 1995.

He is an Assistant Professor of Electrical Engineering and Computer Science at Stanford University, Stanford, CA. From 1986 to 1989, he worked for Hewlett-Packard Labs in the Network and Communications Research group in Bristol, England. During 1995, he worked briefly for Cisco Systems, where he helped architect the GSR12000 router. He researches techniques for high-speed networks, including high-speed Internet routing and architectures for high-speed switches.

Dr. McKeown serves as an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS. He is a Robert Noyce Faculty Fellow at Stanford University and a Fellow of the Alfred P. Sloan Foundation.



Henry Hong-Yi Tzeng (S'92–M'95) received the B.S. degree from the Tatung Institute of Technologies, Taiwan, R.O.C., in 1988 and the M.S. and Ph.D degrees in electrical engineering from the University of California, Irvine, in 1993 and 1995, respectively.

Since 1995, he has been a Member of Technical Staff, Bell Labs, Lucent Technologies, Holmdel, NJ. His research interests are in algorithm design for high-speed networks and fault-tolerant distributed systems. He is currently working on high-performance routing technologies, such as fast-search algorithms, link-state routing, and inter-domain routing protocols, in addition to his publications on reliable multicast protocols, congestion-control protocols for unicast/multicast ATM-ABR services, and TCP performance over ATM networks.

Dr. Tzeng was a corecipient of the 1997 IEEE Browder J. Thompson Memorial Prize Award and also the UC Regent's Dissertation Fellowship in 1995.