Guest Editorial High-Performance Electronic Switches/Routers for High-Speed Internet

I. Introduction

DESPITE the recent slowdown in the telecom equipment market, current estimates and measurements predict that Internet traffic will continue to grow for many years to come. Driving this growth is the fact that the Internet has moved from a convenience to a mission-critical platform for conducting of and succeeding in business. In addition, the provision of broadband services to end users will prolong this growth for many years to come. As a result, there is a great demand for gigabit/terabit electronic routers and switches (IP routers, ATM switches, Ethernet switches) that knit together the constituent networks of the global Internet, creating the illusion of a unified whole. These switches/routers must not only have an aggregate capacity of gigabits/terabits coupled with forwarding rates of billions of packets per second, but they must also deal with nontrivial issues such as scheduling support for differentiated services, a wide variety of interface types, scalability in terms of capacity and port density, and backward compatibility with a wide range of packet formats and routing protocols.

This special issue is a collection of high-quality papers, presenting state-of-the-art design and analysis of high-performance electronic packet switches and routers.

II. Scalability

In the invited paper, "Scalable Electronic Packet Switches," Chiussi and Francini provide an overview of current state of the art of practical large packet switches and routers, and discuss the issues affecting their scalability. The attention falls to three major scalability aspects: implementation, support of quality-of-service (QoS), and multicasting. The impact of these aspects is shown on the most popular switch architectures.

III. IP Lookup

The use of classless interdomain routing (CIDR) allows arbitrary aggregation of network addresses and reduces routing table entries. This complicates the lookup process, requiring a lookup engine to search variable-length address prefixes. Earlier work on fast Internet Protocol Version 4 (IPv4) routing table lookup includes, software mechanisms based on tree traversal or binary search methods, and hardware schemes based on content addressable memory (CAM), memory lookups, and the CPU caching.

Most previous schemes depend on the memory access technology, which limits their performance. Desai *et al.*present a fundamentally different approach in the paper, "Reconfigurable Finite-State Machine Based IP Lookup Engine for High-Speed Router." The IP address lookup problem is presented in the form of a large finite-state machine (FSM), which is then decomposed and implemented into reconfigurable hardware blocks. Performance of the proposed architecture breaks the memory bandwidth limitation and in principle is scalable with very large scale integration (VLSI) technology.

In the paper, "High-Speed IP Routing With Binary Decision Diagrams Based Hardware Address Lookup Engine," Sangireddy and Somani also notice the memory limitation problem. Their solution is a binary decision diagrams (BDDs) based optimized combinational logic, which can be implemented using reconfigurable hardware. The choice of BDD scheme proves to be more beneficial in the scenario that the number of physical ports in a router would increase continuously.

Another practical problem in IP lookup is the high cost of CAMs. In the paper, "Scalable IP Lookup for Internet Routers," Taylor *et al.* present a fast Internet protocol lookup (FIPL) architecture which utilizes tree bitmap algorithm. The architecture only uses a fraction of a reconfigurable logic device and a single commodity SRAM, offering an attractive alternative to expensive CAM-based commercial solutions.

IV. Crossbar Scheduling

The majority of the crossbar scheduling algorithms can be classified as maximum weight/size matching and their iterative pointer-based approximations, driven by either optimal performance or ease of implementation. The next two papers take a shift from legacy scheduling schemes, but are still targeting the provision of efficient crossbar matchings for high performance switches.

In the paper, "DISA: A Robust Scheduling Algorithm for Scalable Crosspoint-Based Switch Fabrics," Elhanany and Sadot present a nonpointer based approach. It performs a synchronized output reservation whereby each input selects a designated output while taking into consideration both local transmission requests and the availability of global resources. The robustness of the scheme under admissible traffic, without the need of speedup, is shown through analysis and computer simulations.

Digital Object Identifier 10.1109/JSAC.2003.813142

In the paper, "Randomized Scheduling Algorithms for High-Aggregate Bandwidth Switches," Giaccone *et al.*exploit hardware parallelism and randomization to yield a set of scheduling algorithms: APSARA, LAURA, and SERENA. Noticing the slow change of queuing lengths in successive time slots, the authors use memory to simplify the implementation and a novel MERGE operation to ensure non-decreasing matching weight. The proposed algorithms are stable under any admissible arrival process. They are simpler than maximum weight matching (MWM) algorithms and achieve comparable delay performance.

V. QoS Guarantee

As link rate increases, more and more applications require the switching system to provide QoS guarantees. The following four papers involve different aspects of providing QoS, including packet classification, queue management, bandwidth management, and delay bound analysis.

Emerging Internet applications demands advanced packet classifiers. In the paper, "Fast and Scalable Packet Classification," van Lunteren and Engbersen propose a new multifield two-phase classification scheme, parallel packet classification. The scheme uses a novel encoding of the intermediate result vectors, which significantly reduces the storage requirements and minimizes the dependencies within the search structures, thus enabling fast incremental updates. It also involves several encoding styles that can be applied simultaneously and allow the storage efficiency and update dynamics to be tuned at the granularity of individual rules.

Active queue management (AQM) schemes aim to regulate ransmission control protocol (TCP) traffic in an efficient and fair way. Management decision is mainly based on the number of the flows in the buffer and data source rate of a flow. In the paper, "An Active Queue Management Scheme Based on a Capture–Recapture Model," Chan and Hamdi estimate the above parameters by randomly capture/recapture incoming packets. This approach can be implemented with low time/space complexity and experiments show that it closely approximates the "ideal" case where full state information is provided.

Applying fair queuing on output ports maybe ineffective when most of the packets nowadays are waiting at the input buffer. Zhang and Bhuyan realize this and address the problem of fair scheduling of packets in input-queued switch architectures. In the paper, "Deficit Round-Robin Scheduling for Input-Queued Switches," they propose a flow-based fair scheduling algorithm which can allocate the switch bandwidth in proportion to each flow's reservation. Such a scheme is demonstrated to achieve fair scheduling while providing high throughput and low latency. A practical version of the flow-based scheme, based on switch port, is also described.

Recent attention has been paid to the problem of minimizing the worst packet delay. The paper, "Scheduling Reserved Traffic in Input-Queued Switches: New Delay Bounds via Probabilistic Techniques," derives delay bounds for decomposition-based algorithms. Andrews and Vojnović show that by using probabilistic techniques they are able to tighten worst delay bounds in many scenarios.

VI. OQ Emulation

Output queued (OQ) architecture is known to be of optimal performance amongst all queuing schemes. However, memory bandwidth limitation makes it not practical for large switch sizes. Recent research focuses on how to emulate the performance of OQ while using more practical approaches.

In the paper, "Output-Queued Switch Emulation by Fabrics With Limited Memory," Magill *et al.*present a switch architecture with input queuing, fabric queuing, flow-control between the limited fabric buffers and the inputs, and output queuing. This combined input/fabric/output queued (CIFOQ) switch with speedup of two is shown to emulate a broad class of scheduling algorithms operating an OQ switch. The use of limited amount of fabric buffers enables distributed scheduling and significantly reduces the scheduling complexity when compared with the memoryless combined input/output queued (CIOQ) architecture.

Rather than emulating output queuing exactly at the expense of complex algorithms or extra memory, Lee and Seo focus on matching the performance of an output-queued switch statistically using implementable schemes. In their paper, "A Practical Approach for Statistical Matching of Output Queuing," a novel multiple input/output-queued (MIOQ) switch architecture that requires no speedup is proposed. A multitoken-based round-robin arbiter and a virtual FIFO queuing scheme cooporate with the architecture, providing high operation rate and cell order guarantee. Additionally, the proposed switch naturally provides asymmetric bandwidth for inputs and outputs.

VII. Miscellaneous

Small cell-based IP routers normally handle multicast traffic by attaching a bitmap local multicast label (LML) to each cell. Marsan *et al.* point out that this approach would induce intolerable overhead for switches with 128 ports or more. In their paper, "Compression of Multicast Labels in Large IP Routers," static, adaptive, and hybrid lossy compression algorithms to reduce the size of LML are discussed. Analytical and simulation models are used to investigate the performance of the different compression approaches.

In scheduling a single IQ/CIOQ switch, maximum weight matching (MWM) is identified as optimal due to 100% throughput under admissible traffic and satisfying delay performance. However, a usual MWM policy turns to be unstable in networks of interconnected IQ/CIOQ switches. The stability among switches is addressed in the paper "On the Stability of Local Scheduling Policies in Networks of Packet Switches With Input Queues." Marsan *et al.* analyze two scheduling policies, Birkhoff–von Neumann based and modified weight MWM based, using fluid models methodology. They identify these policies require no coordination among switches and guarantee 100% throughput in a network of IQ/CIOQ switches.

In the paper, "The Sliding-Window Packet Switch: A Shared-Memory Switch Architecture With Plural Memory Modules and Decentralized Control," Kumar proposes a new class of shared-memory packet switches. The architecture has separated multiple memory modules that are logically

shared among all the ports of the switch and the control is decentralized. Decentralized switching functions enable the sliding-window switch to operate in a pipeline fashion and enhance scalability and switching capacity.

ACKNOWLEDGMENT

The Guest Editors would first like to acknowledge all authors for having chosen this special issue to disseminate their research results. We also appreciate the reviewers' time and effort in controlling the quality by providing detailed and constructive feedbacks. Last but not least, their sincere gratitude for the remarkable efforts of all the editorial staff of JSAC.

MOUNIR HAMDI, *Guest Editor-in-Chief*
Hong Kong University of Science and Technology
Kowloon, Hong Kong

DANIEL J. BLUMENTHAL, *Guest Editor*
University of California at Santa Barbara
Santa Barbara, CA 93106 USA

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

EMILIO LEONARDI, *Guest Editor*
Politecnico di Torino
Torino 10129, Italy

CHUNMING QIAO, *Guest Editor*
University at Buffalo
Buffalo, NY 14222 USA

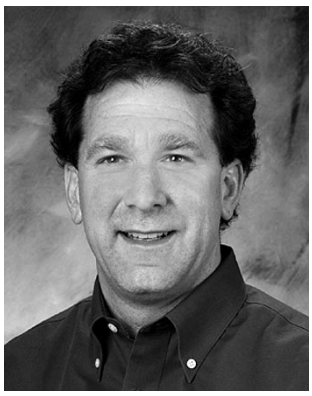
KENNETH Y. YUN, *Guest Editor*
University of California at San Diego
San Diego, CA 92037 USA

R. RAMASWAMI, *J-SAC Board Representative*

Mounir Hamdi (S'89–M'90) received the B.S. degree in computer engineering (with distinction) from the University of Louisiana, Lafayette, LA, in 1985, and the M.S. and the Ph.D. degrees in electrical engineering from the University of Pittsburgh, Pittsburgh, PA, in 1987 and 1991, respectively.

He has been a Faculty Member in the Department of Computer Science, Hong Kong University of Science and Technology, since 1991, where he is now Associate Professor of Computer Science and the Director of the Computer Engineering Program that has some 350 undergraduate students. From 1999 to 2000, he held Visiting Professor positions at Stanford University, Stanford, CA and the Swiss Federal Institute of Technology, Zürich, Switzerland. His general areas of research are in high-speed packet switches/routers and all-optical networks, in which he has published more than 180 research publications, received numerous research grants, supervised some 20 postgraduate students, and for which he has served as Consultant to various international companies. Currently, he is working on high-speed networks including the design, analysis, scheduling, and management of high-speed switches/routers, wavelength division multiplexing (WDM) networks/switches, and wireless networks. He is currently leading a team that is designing one the highest capacity chip sets for terabit switches/routers. This chip set is targeted toward a 256×256 OC-192 switches, and includes a crossbar fabric chip, a scheduler/arbiter chip, and traffic management chip.

Dr. Hamdi is/was on the Editorial Board of IEEE TRANSACTIONS ON COMMUNICATIONS, *IEEE Communications Magazine*, *Computer Networks*, *Wireless Communications and Mobile Computing*, and *Parallel Computing*, and has been on the program committees of more than 50 international conferences and workshops. He was a guest editor of *IEEE Communications Magazine*, IEEE JOURNAL ON SELECTED AREAS OF COMMUNICATIONS, and *Optical Networks Magazine*, and has Chaired more than five international conferences and workshops including the IEEE GLOBECOM/ICC Optical networking workshop, the IEEE ICC High-Speed Access Workshop, and the IEEE IPPS HiNets Workshop. He is the Chair of IEEE Communications Society Technical Committee on transmissions, access, and optical systems, and Vice-Chair of the optical networking Technical Committee, as well as ComSoc Technical Activities Council. He received the Best Paper Award at the International Conference on Information and Networking in 1998 out of 152 papers. In addition to his committment to research and professional service, he is also a dedicated teacher. He received the Best 10 Lecturers Award (through university-wide students voting for all university faculty held once a year) and the Distinguished Teaching Award from the Hong Kong University of Science and Technology. He is a member of ACM.



Daniel J. Blumenthal (F'03) received the B.S.E.E. degree from the University of Rochester, NY, in 1981, the M.S.E.E. degree from Columbia University, NY, in 1988, and the Ph.D. degree from the University of Colorado at Boulder, in 1993.

In 1981, he worked at StorageTek of Louisville, CO, in the area of optical data storage. From 1986 to 1988, he worked at Columbia University in the areas of photonic switching and ultrafast all-optical networks. From 1993 to 1997, he was an Assistant Professor in the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta. He is the Associate Director for the Center on Multidisciplinary Optical Switching Technology (MOST) and Professor in the Department of Electrical and Computer Engineering, University of California, Santa Barbara. His current research areas are in optical communications, photonic packet switched and all-optical networks, all-optical wavelength conversion, optical subcarrier multiplexing, integrated-optic chip scale WDM, and nanophotonic technologies. He has authored or coauthored over 100 papers in these and related areas.

Dr. Blumenthal is recipient of a 1999 Presidential Early Career Award for Scientists and Engineers (PECASE) from the White House and the DoD, a 1994 NSF Young Investigator (NYI) Award, and a 1997 Office of Naval Research Young Investigator Program (YIP) Award. He has served as an Associate Editor for the IEEE PHOTONICS TECHNOLOGY LETTERS and the IEEE TRANSACTIONS ON COMMUNICATIONS. He was a Guest Editor for the JOURNAL OF LIGHTWAVE TECHNOLOGY, Special Issue in Photonic Packet Switching Systems (December 1998) and is currently a Guest Editor for the IEEE JOURNAL OF SELECTED AREAS IN COMMUNICATIONS, Special Issue on High-Performance Optical/Electronic Switches/Routers for High-Speed Internet. He has served as the General Program Chair for the 2001 OSA Topical Meeting on Photonics in Switching and as Program Chair for the 1999 Meeting on Photonics in Switching. He has also served on numerous other technical program committees including the Conference on Optical Fiber Communications OFC (1997, 1998, 1999, and 2000) and the Conference on Lasers and Electrooptics CLEO (1999 and 2000). He is also a Member of the Lasers and Electrooptic Society and the Optical Society of America.

H. Jonathan Chao (F'01) received the B.S. and M.S. degrees in electrical engineering from National Chiao Tung University, Taiwan, and the Ph.D. degree in electrical engineering from Ohio State University, Columbus.

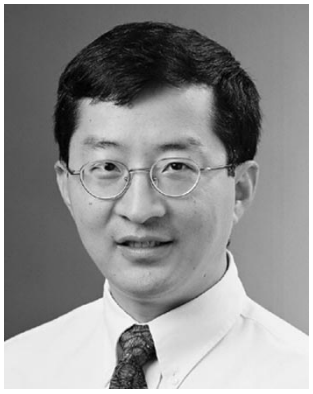
He is a Professor of electrical and computer engineering at Polytechnic University, NY, where he joined in January 1992. He has been doing research in the areas of terabit switches/routers, QoS control, optical networking, and network security. He holds more than 20 patents and has published over 100 journal and conference papers in the above areas. He has also served as a consultant for various companies, such as Lucent Technologies, NEC, and Telcordia. He has been giving short courses to industry people in the subjects of SONET/ATM/IP/MPLS networks for over a decade. From 2000 to 2001, he was Cofounder and CTO of Coree Networks, NJ, where he led a team to implement a multiterabit MPLS switch router with carrier-class reliability. From 1985 to 1992, he was a Member of Technical Staff at Telcordia, NJ, where he was involved in transport and switching system architecture designs and ASIC implementations, such as the first SONET-like framer chip, ATM layer chip, Sequencer chip (the first chip handling packet scheduling), and ATM switch chip. From 1977 to 1981, he was a Senior Engineer at Telecommunication Labs of Taiwan performing circuit designs for a digital telephone switching system. He coauthored two networking books *Broadband Packet Switching Technologies* (New York: Wiley, 2001) and *Quality of Service Control in High-Speed Networks* (New York: Wiley, 2001).

Prof. Chao is a Fellow of the IEEE for his contributions to the architecture and application of VLSI circuits in high-speed packet networks. He received the Telcordia Excellence Award in 1987. He is a corecipient of the 2001 Best Paper Award from the IEEE Transaction on Circuits and Systems for Video Technology. He has served as a Guest Editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS (JSAC) on the special topics of Advances in ATM Switching Systems for B-ISDN (June 1997), Next-Generation IP Switches and Routers (June 1999), and the recent issue on "High-Performance Optical/Electronic Switches/Routers for High-Speed Internet." He also served as an Editor for IEEE/ACM TRANSACTIONS ON NETWORKING from 1997 to 2000.

Emilio Leonardi (M'99) received the Dr.Ing. degree in electronics engineering and the Ph.D. degree in telecommunications engineering, in 1991 and 1995, respectively, both from Politecnico di Torino, Torino, Italy.

He is an Assistant Professor in the Electronics Department, Politecnico di Torino. In 1995, he was Visiting Scholar in the Computer Science Department, University of California, Los Angeles (UCLA). He was Visiting Researcher in the High-Speed Networks Research Department, Bell Laboratories, Lucent Technologies, Holmdel, NJ (summer 1999) and in the Electrical Department, Stanford University, Stanford, CA (summer 2001), hosted by Prof. B. Prabakhar. He participated in several National and European projects, IST-SONATA and IST DAVID. He is also involved in several consulting and research projects with private industries, including Lucent Technologies, British Telecom, and TILAB. He has coauthored over 100 papers published in international journals and presented in leading international conferences. His areas of interest are in all-optical networks, queueing theory, and scheduling policies for high-speed switches.

Dr. Leonardi received the "IEEE TCGN Best Paper Award" for a paper presented at IEEE GLOBECOM 2002, High-Speed Networks Symposium. He has participated in the technical program committees of several conferences, including IEEE INFOCOM and IEEE GLOBECOM.



Chunming Qiao is currently an Associate Professor at the University at Buffalo (SUNY), where he directs the Lab for Advanced Network Design, Evaluation, and Research (LANDER). He has published more than 100 papers in leading technical journals and conference proceedings, and is recognized for his pioneering research on Optical Internet and in particular, the optical burst switching (OBS) paradigm. His work on integrated cellular and *ad hoc* networking systems (iCAR) is also internationally acclaimed.

He is an Editor of several journals and magazines including IEEE/ACM TRANSACTIONS ON NETWORKING (ToN) and IEEE Optical Communications, as well as a Guest Editor for several issues of IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS and other publications. He has Chaired and Co-Chaired many conferences and workshops including the Symposium on Optical Networks at ICC 2003 and OPTICOMM 2002. He is also the Founder and Chair of the Technical Group on Optical Networks (TGON) sponsored by SPIE, and a Vice Chair of the IEEE Technical Committee on Gigabit Networking (TCGN).

Kenneth Y. Yun received the S.M. degree in electrical engineering and computer science from Massachusetts Institute of Technology, Cambridge (MIT) and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA.

He is currently an Associate Professor in the Department of Electrical and Computer Engineering, University of California, San Diego. His current research interests include the design and implementation of network systems and protocols and high-speed VLSI circuits. He was a Founder and the Chief Technical Officer of a networking IC startup, YuniNetworks, and a Chief Technical Officer of a telecom IC company, AMCC, while on leave of absence from University of California. He hads worked as a Consultant for Intel Corporation, CA, on the Asynchronous Instruction Decoder Project from 1996 to 1999. He has organized ASYNC'98 as a Program Co-Chair.

Dr. Yun is the recipient of a National Science Foundation CAREER Award and a Hellman Faculty Fellowship. He has received the Charles E. Molnar Award for a paper that best bridges theory and practice of asynchronous circuits and systems at ASYNC'97 and a Best Paper Award at ICCD'98.