Guest Editorial: Switching and Routing for Scalable and Energy-Efficient Networking

A. Smiljanić, J. Chao, C. Minkenberg, E. Oki, and M. Hamdi

DATACENTERS are the Internet brains, where scalability and power consumption take their toll. Cost-effective computing and storage are provided in datacenters due to economy of scale. Their key challenge is scalability, as a datacenter may house thousands of servers interconnected through many switches and routers, and the physical infrastructure is shared by multiple tenants, each deploying many applications. Tremendous throughputs and processing performed in datacenters require enormous amounts of energy. Datacenters consume more than 1% of the total electricity, according to recent data. It is important to reduce power consumption of datacenters not only to reduce their operational costs, but also to allow their sustainable growth, and protect our environment.

In this special issue, three different aspects of the energy savings are addressed: savings at the link layer by appropriate power management, savings within a datacenter by optimization of data migration and routing mechanisms, and carbon emission reduction by routing traffic between datacenters in order to efficiently utilize renewable sources of energy.

Many applications (e.g., cluster based storage and MapReduce) in modern datacenters require a high fan-in, incast (many-to-one) type of data communication, which could cause severe congestion in switches and result in TCP goodput collapse, substantially degrading the application performance. As a result, traditional congestion control needs to be either revised, or replaced by a new paradigm which may require an active role of switches and routers. Three novel transport protocols based on different concepts were proposed and analyzed in the papers of the issue.

It has been recognized that the datacenter throughputs can be improved by load balancing. However, known load balancing techniques are not readily applicable to the multicast traffic, and need to be revisited since one-to-many applications are common in datacenters, e.g. distribution of executable binaries, or replication in distributed file systems. Also, multicast addressing has been problematic as it does not allow aggregation, and each multicast flow needs to have its unique address. Both of the aformentioned multicast issues have been addressed in this special issue.

Packet classification is another critical functionality in a datacenter network that has been addressed, since it should be fast and scalable to handle many customers and their applications. Also, datacenter virtualization is causing a dramatic rise of the number of entries in forwarding tables.

Ling Wang et al. consider in their paper a general framework for saving the energy in datacenters. They utilize the mobility of virtual machines in datacenters, and consolidate virtual machines to a subset of servers to minimize the power consumption. Then, authors reduce the number of active switches and their power consumption by optimizing the routing of the traffic exchanged among the virtual machines. It was shown that the proposed two-step optimization procedure can save up to 50% of the energy consumed by datacenters.

Rafaele Bolla et al. propose a closed-form model to accurately estimate both power consumption and network performance of Energy Efficient Ethernet links (IEEE 802.3az) which can adjust energy consumption to the actual traffic load. In contrast with the existing analytical frameworks, the proposed approach allows obtaining the average energy consumption of the link and the mean latency time in closed form, without any upper or lower bound approximations. The model considers relevant traffic parameters such as the packet length distribution, burst size distribution, and average burst inter-arrival rates.

Mirko Gattulli et al. address a low-carbon cloud-computing solution both to achieve significant reduction in CO_2 emissions and to cope with the growing power requirements of datacenters. The paper proposes two routing strategies, designed to route optical connections supporting cloud-computing service requests, minimizing the CO_2 emissions of datacenters by means of renewable energy which are coming from sun and wind. The paper addresses the tradeoff between the energy consumption of data transport and the energy consumption required for processing the cloud-computing requests inside datacenters. The performance evaluations observe that relevant reductions, up to about 30%, in CO_2 emissions can be achieved.

Jiao Zhang et al. propose in their paper a novel transport protocol which is based on fair sharing of the bottleneck switch buffer in a datacenter. In SAB (Sharing by Allocating the switch Buffer), the controller associated with a buffer updates the number of flows passing this buffer based on SYN and FIN packets, and updates the window size in the headers of the passing packets as a function of the number of active flows. While simple for implementation, SAB outperforms

A. Smiljanić is with Belgrade University, Serbia.

J. Chao is with New York University Polytechnic School of Engineering, USA.

C. Minkenberg is with IBM Research - Zurich, Switzerland.

E. Oki is with The University of Electro-Communications, Tokyo, Japan.

M. Hamdi is with Hong Kong University of Science and Technology, China.

M. Zukerman is the J-SAC board representative for this issue of IEEE Journal on Selected Areas in Communications.

Digital Object Identifier 10.1109/JSAC.2014.140101

0733-8716/14/$31.00 © 2014 IEEE

transport protocols previously proposed for datacenters. SAB follows quickly the changes of traffic loads, and rarely loses packets. Consequently, completion times of both the short queries and the background flows are reduced by SAB.

The paper of Changlin Jiang, Dan li and Mingwei Xu addresses the TCP incast problem, by a new transport protocol LTTP (Luby Transform based Transport Protocol) for many-to-one communications. LTTP includes a data channel and a control channel. In the data channel from servers to the client, instead of using the timeout-and-retransmit mechanism as in TCP, LTTP improves the UDP-based LT code for reliable delivery, which is based on FEC (Forward Error Correction) with data redundancy. LTTP also utilizes TFRC (TCP Friendly Rate Control) to adjust the sending data rates at servers for the congestion control.

The root cause of the incast collapse in datacenters is the long idle period of the Retransmission Timeout (RTO) that is triggered at one or more senders by packet losses in congested switches. The paper of Shikhar Shukla et al. develops a packet labeling scheme, named PLATO, to improve the loss detection capabilities of TCP NewReno. Packets carrying the special label are preferentially enqueued at the switch. This allows TCP to detect packet loss using three duplicate acknowledgements, instead of the time expensive RTO; thus avoiding the goodput collapse. PLATO makes minor modifications to NewReno and does not alter its congestion control mechanism. Simulation results using ns-3 tool show that PLATO performs orders of magnitude better than NewReno as well as the state-of-art incast solution ICTCP (Incast Control TCP).

While many overlay prototypes are proposed for datacenter networks, they focus mainly on functionality and security, with little being known yet about their impact on the system level performance. The paper of Daniel Crisan et al. evaluates the impact of the overlay network on two representative datacenter workloads, Partition/Aggregate and 3-Tier, using query completion time as the primary performance metric. Partition/Aggregate is a protocol-sensitive application which is the core of MapReduce and Hadoop mechanisms, notoriously latency-sensitive and exposed to the TCP incast congestion. The performance study showed that latency-sensitive workloads do not necessarily lose performance in virtualized networks.

Eitan Zahavi, Isaac Keslassy and Avinoam Kolodny study in their paper the convergence of dynamic routing in Clos-based datacenter networks. They demonstrate that distributed adaptive routing may be used to provide scalable and non-blocking routing for long-lived flows in a rearrangeably-non-blocking Clos network assuming that none of the flows exceeds more than half the link capacity. Their scheme was shown to converge within a few iterations, while causing minimal out-of-order packet delivery.

Zhiyang Guo and Yuanyuan Wang propose a novel routing algorithm for path assignments to multicast flows, BCMS (Bounded Congestion Multcast Scheduling). BCMS is based on a centralized controller according to the OpenFlow trends, and it includes balancing through the core switches. BCMS has a moderate complexity, while it achieves bounded congestion of the network links under any arbitrary sequence of multicast flow requests that satisfy the hose model.

Wen-Kang Jia and Li-Chun Wang propose an efficient stateless source-routing scheme for handling both unicast and multicast packets in SDN-based datacenter networks. The proposed scheme is shown to reduce the processing time and the protocol overhead associated with conventional methods. This is achieved by using a novel decoding mechanism of bitmaps corresponding to the packet output ports which avoids using addressing of multicast groups.

Pi-Chung Wang proposes an algorithm, called encoded rule expansion, to transform packet classification rules into an equivalent set of rules with fewer distinct prefix-length combinations. The numerical results show that more than 90% of hash tables can be eliminated. The paper also shows that the software implementation of the proposed scheme without using any hardware parallelism can support up to one thousand customer VLANs and one million rules.

Ori Rottenstreich et al. investigates a novel technique for the efficient representation of forwarding tables in datacenter switches. In particular, the paper introduces a novel forwarding table architecture with separate encoding in each column using a dedicated variable-length binary prefix encoding, i.e., an encoding in which any codeword is not a prefix of any other codeword.

Aleksandra Smiljanić received M.A. and Ph.D. degrees in electrical engineering from Princeton University in 1996 and 1999, respectively. She completed B.Sc. in electrical engineering at Belgrade University in 1993. Currently, Aleksandra works as an associate professor at Belgrade University in Serbia. She worked for two summers at NEC USA on a design of the packet switch with terabit capacity. She had worked for AT&T Labs from 1999 until 2004. At AT&T Labs Aleksandra worked on the research in the area of high-capacity packet switches and routers, and on the evaluation of Internet core routers and multiservice edge routers offered by various vendors. She also worked as an adjunct professor at Stony Brook University in New York, and as a research professor at Polytechnic Institute of NYU.

Aleksandra Smiljanić is the author of numerous conference and journal papers in the area of high performance switching and routing. She is the inventor of ten US patents. Some of these patents have been patented in Europe, Japan and China as well. In 2009, Aleksandra got the Ilija Stojanović Award for the best journal paper in the area of communications sponsored by the Telenor Fondation. Aleksandra Smiljanić is the author of the Best Papers at IEEE Conference on High Performance Switching and Routing 2000, and IEICE/IEEE Workshop on High Performance Switching and Routing 2002. She got the Research Excellence Award at AT&T Labs in 2000. She is a recipient of the Aleksandar Damjanović Award as the best student in her class at Belgrade University, 1993. Before university, she won numerous prizes in Yugoslav and international competitions in mathematics and physics.

Aleksandra Smiljanić was the editor of OSA Journal on Optical Networking in a period 2003-2009. She served as as the editor of IEEE Communication Letters from 2005 until 2011. Aleksandra was the General Chair of IEEE Conference on High-Performance Switching and Routing 2012.

H. Jonathan Chao (IEEE Fellow) is Department Head and Professor of Electrical and Computer Engineering at New York University Polytechnic School of Engineering, where he joined in January 1992. He has been doing research in the areas of data center network designs, terabit switches/routers, network security, network on chip, and medical devices. He holds 46 patents with 11 pending and has published more than 200 journal and conference papers. During 2000–2001, he was Co-Founder and CTO of Coree Networks, NJ, where he led a team to implement a multi-terabit MPLS (Multi-Protocol Label Switching) switch router with carrier-class reliability. From 1985 to 1992, he was a Member of Technical Staff at Telcordia. 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 co-recipient of the 2001 Best Paper Award from the IEEE Transaction on Circuits and Systems for Video Technology. He coauthored three networking and switching books.

Cyriel Minkenberg is a Research Staff Member and manager of the Systems Fabrics group in the Systems department at IBM Research - Zurich. His group pursues architectural and protocol innovation, performance evaluation, and practical implementation of interconnection networks for high-performance computing and data center networks, with a focus on workload-level performance impact of design aspects of the entire communication subsystem. He obtained MSc and PhD degrees in electrical engineering from the Eindhoven University of Technology, the Netherlands, in 1996, and 2001, respectively, and has been a Research Staff Member at IBM Research - Zurich since 2001.

His research interests include interconnection networks, switch architectures, networking protocols, performance modeling, and simulation. Minkenberg has co-authored more than 50 papers in peer-reviewed journals and conferences and holds 14 US patents. He received the 2001 IEEE Fred W. Ellersick Award for the best paper published in an IEEE Communications Society magazine in 2000, as well as conference best paper awards at IEEE Hot Interconnects 2005, IEEE IPDPS 2007 Architectures Track, IEEE HPCC 2012, and ICPP 2012. He served as Technical Program Chair for IEEE Hot Interconnects 2010 and 2013, IEEE HPSR 2012, HiPEAC INA-OCMC 2012, as Tutorial Chair for IEEE Hot Interconnects 2011 and 2012, and as General Chair for HiPEAC INA-OCMC 2013.

Eiji Oki is a Professor at The University of Electro-Communications, Tokyo, Japan. He received the B.E. and M.E. degrees in instrumentation engineering and a Ph.D. degree in electrical engineering from Keio University, Yokohama, Japan, in 1991, 1993, and 1999, respectively. In 1993, he joined Nippon Telegraph and Telephone Corporation (NTT) Communication Switching Laboratories, Tokyo, Japan. He has been researching network design and control, traffic-control methods, and high-speed switching systems. From 2000 to 2001, he was a Visiting Scholar at the Polytechnic Institute of New York University, Brooklyn, New York, where he was involved in designing terabit switch/router systems. He was engaged in researching and developing high-speed optical IP backbone networks with NTT Laboratories. He joined The University of Electro-Communications, Tokyo, Japan, in July 2008. He has been active in standardization of path computation element (PCE) and GMPLS in the IETF. He wrote more than ten IETF RFCs.

Prof. Oki was the recipient of the 1998 Switching System Research Award and the 1999 Excellent Paper Award presented by IEICE, the 2001 Asia-Pacific Outstanding Young Researcher Award presented by IEEE Communications Society for his contribution to broadband network, ATM, and optical IP technologies, and the 2010 Telecom System Technology Prize by the Telecommunications Advanced Foundation. He has authored/co-authored four books, Broadband Packet Switching Technologies, published by John Wiley, New York, in 2001, GMPLS Technologies, published by CRC Press, Boca Raton, FL, in 2005, Advanced Internet Protocols, Services, and Applications, published by Wiley, New York, in 2012, and Linear Programming and Algorithms for Communication Networks, CRC Press, Boca Raton, FL, in 2012. He is an IEEE Fellow.

Mounir Hamdi is a Chair Professor at the Hong Kong University of Science and Technology, and the Head of the Department of Computer Science and Engineering. He is an IEEE Fellow for contributions to design and analysis of high-speed packet switching. He received the B.S. degree in Electrical Engineering - Computer Engineering minor (with distinction) from the University of Louisiana in 1985, and the MS and the PhD degrees in Electrical Engineering from the University of Pittsburgh in 1987 and 1991, respectively. He has been a faculty member in the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology since 1991, where he was a founding member of the University and the Department. He is now the Head and Chair Professor of the Department. He held visiting professor positions at Stanford University, USA, and the Swiss Federal Institute of Technology, Lausanne, Switzerland. His general area of research is in high-speed wired/wireless networking in which he has published more than 300 research publications, received numerous research grants, and graduated more than 30 graduate students. In addition, he has frequently consulted for companies and governmental organizations in the USA, Europe and Asia. He is a frequent keynote speaker at international conferences and forums.

Prof. Hamdi is/was on the Editorial Board of various prestigious journals and magazines including IEEE Transactions on Communications, IEEE Communication Magazine, Computer Networks, Wireless Communications and Mobile Computing, and Parallel Computing as well as a guest editor of IEEE Communications Magazine, IEEE Journal on Selected Areas of Communications, and Optical Networks Magazine. He has chaired more than 20 international conferences and workshops, and has been on the program committees of more than 200 international conferences and workshops. He was 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 member of the ComSoc technical activities council. He received the best paper award at the IEEE Globecom 2012, IEEE International Conference on Communications in 2009 and the IEEE International Conference on Information and Networking in 1998. In addition to his commitment to research and professional service, he is also a dedicated teacher and renowned quality-assurance educator. He received the best 10 lecturers award (through university-wide student voting for all university faculty held once a year), the distinguished engineering teaching appreciation award from the Hong Kong University of Science and Technology, and various grants targeted towards the improvement of teaching methodologies, delivery and technology.