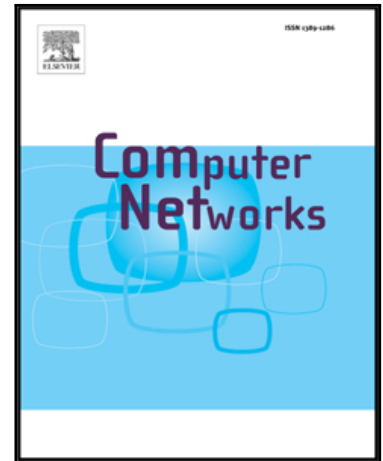


Accepted Manuscript

A Scalable Packet-Switch Architecture Based on OQ NoCs for Data Center Networks

Fadoua Hassen, Lotfi Mhamdi

PII: S1389-1286(17)30310-9
DOI: [10.1016/j.comnet.2017.08.003](https://doi.org/10.1016/j.comnet.2017.08.003)
Reference: COMPNW 6276



To appear in: *Computer Networks*

Received date: 15 November 2016
Revised date: 4 August 2017
Accepted date: 9 August 2017

Please cite this article as: Fadoua Hassen, Lotfi Mhamdi, A Scalable Packet-Switch Architecture Based on OQ NoCs for Data Center Networks, *Computer Networks* (2017), doi: [10.1016/j.comnet.2017.08.003](https://doi.org/10.1016/j.comnet.2017.08.003)

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

A Scalable Packet-Switch Architecture Based on OQ NoCs for Data Center Networks

Fadoua Hassen, *Student Member, IEEE*, and Lotfi Mhamdi, *Member, IEEE*

Abstract—Data Center switches need guarantee high throughput, resiliency and scalability for large-scale networks with constantly floating requirements. Multistage packet switches have been a pervasive solution to implement high-capacity Data Center Networks (DCNs) switches and routers. Yet, classical multistage switching architectures with their Space-Memory variants have shown limited performance. Most proposals prove either too complex to implement or not cost effective. In this paper, we present a highly scalable packet-switch for the DCN environment, in which we exploit the Network-on-Chip (NoC) design paradigm to replace the single-hop crossbars with multi-hop Switching Elements (SEs). In particular, we describe a three-stage switch with Output-Queued Unidirectional NoCs (OQ-UDN) in the central stage of the Clos-network. The design has several advantages over conventional multistage switches. First, it uses a simple Round-Robin (RR) packet dispatching scheme and avoids the need for complex and costly input modules. Besides, it offers better load balancing, a pipelined scheduling and more path-diversity. We assess the performance of the switch in terms of throughput, end-to-end latency and blocking probability using Markov chain analysis, and we propose an analytical model that integrates the various design parameters. Through extensive simulations, we show that the switching architecture achieves high performance under different types of traffic, and that both the analytical and experimental results correlate over wide range of evaluation settings.

Index Terms—Next-generation networking, packet switching, Clos-network, NoC, OQ, analytical model

I. INTRODUCTION

A Data Centre is the nexus from which all the services of the cloud flow and where different types and generations of switches/routers are used to handle the floating workload. Given the growing networking requirements, both today and in-the-future DCN switching fabrics need to rapidly and reliably scale performance, either on a sustained basis or when unexpected load spikes place burden on the bandwidth availability.

The new switching fabrics are expected to improve upon current solutions in many ways to provide better performance. As in many iterative design approaches, each switching architecture improves on the previous ones at better points in a hardware cost/performance curve. The common design trend is founded on building hierarchical switching fabrics as single stage crossbar switches do not fit for the expansion of the network substrate. While they can be implemented for small sized switches, single-stage crossbar switches become complex, unpractical and unscalable for growing port counts (beyond 64 ports) [1], [2]. Multistage switches where many

smaller crossbar fabrics are cascaded have been typical commercial solutions for high-speed routers [3]. They provide good broadcast and multicast features, and they can be incrementally expanded by simply adding more modules to the existing design. The three-stage Clos-network [2] is a popular non-blocking multistage arrangement that is frequently used for telecommunications and networking systems. Despite their scalability potential, almost all existing Clos-network based proposals (from the Space-Space-Space switch – S^3 – to the Memory-Memory-Memory switch – MMM) are too complex to implement, have non satisfactory performance or require costly modules [3], [4]. During the on-going research of packet-switches design, NoC architectures were proposed as a new functional-level design pattern to mitigate the limitations of the classical single-hop crossbars, such as the bottleneck of speedup and scalability in port count. NoCs have interesting characteristics that offer switching fabrics more flexibility, and allow them to operate independently of the switch valency. Moreover, the path diversity that NoC grids provide, help disperse the traffic load and get it better balanced among many intrinsic routes [5].

In this paper, we propose the OQ Clos-UDN: A sophisticated switching architecture that defeats several limitations of the classical multistage packet switches. The current design brings about a nested three-stage Clos-network switch with simple FIFO queues at the input modules and a dynamic packet dispatching scheme. Instead of the conventional point-to-point connection crossbars, we use OQ UDN modules in the middle-stage of the Clos-network. The OQ Clos-UDN switch has many advantages over the Memory-Space-Memory (MSM) and MMM architectures since it contributes to: (i) Simplifying the IMs thanks to the NoC central modules¹. (ii) Simplifying the packet dispatching process and avoiding the need for complex and synchronized scheduling algorithms². (iii) Using small and distributed on-chip buffers, and obviating the need for large crosspoint queues that an MMM switch require. (iv) Using a pipelined and distributed routing scheme to move packets across the Central Modules (CMs). (v) Offering speedup, load-balancing and better path diversity as compared with crossbar-based switches.

We use Markov chain analysis to derive an analytical model for the performance metrics of the OQ Clos-UDN switch. In addition to the throughput and the average packet latency, we give an estimation for the upper-bound blocking probability

¹Actually, the Head-of-Line (HoL) problem is hardly noticeable [5] that it becomes possible to use simple FIFO queues instead of the complex and costly VOQs.

²The MSM switches call for complex iterative algorithms to find a conflict-free matching over a high number of input/output modules and port pairs.

The authors are with the Department of Electrical and Computer Engineering, Institute of Integrated Information Systems, University of Leeds, UK (e-mail:elfha@leeds.ac.uk; L.Mhamdi}@leeds.ac.uk).

inside the CMs under uniform *Bernoulli i.i.d* traffic. The analytical model tackles different purposes such as appraising the impact of the design metrics on the switch performance and settling optimal values to achieve given performance for reasonable cost/complexity.

The remainder of the paper is structured as follows. In section II, we review the state-of-the-art switching architectures and the evolution of the design process. In section III, we present the switch terminology. We focus on the OQ-UDN central modules, and we justify the implementation feasibility of the proposed switching architecture. Sections IV and V give details of the analytical modelling for the performance metrics of the switch. In section VI, we evaluate the OQ Clos-UDN's performance, and we compare it to existent multistage switching solutions under various traffic types. We also correlate the analytical models to simulation outputs. Ultimately, section VII summarizes the work and concludes the paper.

II. RELATED WORK

Inspired by Systems-on-Chip (SoC) communications, recent works have proposed the implementation of NoC-based switching fabrics and have discussed their potential and their performance. The NoC design brings numerous advantages. It emerges as a flexible and suitable alternative to single-hop crossbars offering tolerable latencies and good load balancing. As for SoC, the NoC paradigm simplifies the hardware required for the routing functions, and makes the switching fabric reach high throughput. Besides, it offers a pipelined scheduling and allows scaling up the switch size for reasonable costs. Some earlier works [6], [7] evoked Ethernet switches that have been designed using the NoC concept. Later on, a single-stage Input-Queued (IQ) Unidirectional NoC crossbar packet-switch (UDN) was introduced [5], [8]. In 2010, the Multidirectional NoC (MDN) packet-switch was proposed as an extension to UDN [9]. More recent results [10] discussed a possible implementation of a crossbar fabric using NoC-enhanced FPGA, and evaluated its performance for various routing algorithms. In [8], Karadeniz *et al.* suggested one stage packet switch with NoC fabric. They described a wrapped-around grid of OQ mini-routers for which they proposed a low-complexity analytical model.

Despite the high potential of the NoC based crossbar fabrics, their application has been restricted to single-stage switch design. In [11], authors first introduced a three-stage Clos switch with IQ NoC-based modules (UDN) in the central stage. The switch has good scalability and parametrization features. However, on-grid routers of the UDNs modules require speedup for the whole Clos switch to achieve good performance. As for single-stage crossbar switches, an output queuing design scheme contributes to increasing the bandwidth, and allows many cells to be simultaneously forwarded to the same output port resulting in higher throughput [12]. All in all, adopting OQ Mini-Routers (MRs) in the UDN blocks is much effective than using IQ nodes³ for the following reasons: (1) Higher throughput is achieved and the overall packets delay

³The terms mini-routers and nodes are used interchangeably throughout the paper to refer to on-chip routers.

is shifted by a fixed amount⁴. (2) The internal links of an OQ-UDN module run at the same rate as the external line and no speedup is required. Assuming the technological advances in the field of memory design and synthesis, it became possible to implement the OQ-UDN modules for reasonable costs. Overall, the proposed switching architecture offers great scalability degree and high performance making it a good candidate for the next-generation DCN switches/routers.

As part of the performance evaluation process, there is a great deal of interest in developing analytical models for the switching architectures as purely simulation analysis is not only inflexible, but also time consuming. In this paper, we analyse the performance of the proposed switch but above all, we focus on the OQ-UDN modules which geometrical features differ from a simple crossbar. In this context, we review some works that conferred modelling of NoCs and NoC-based switching fabrics. In 2009, Elmiligi *et al.* proposed an empirical model to address the queue size problem in OQ routers for NoCs using Markov chains analysis [13]. In a different method, authors in [14] introduced a low complexity analytic approach for the mean analysis of some performance metrics of NoCs. In 2010, Suboh *et al.* used a Network Calculus-based methodology to evaluate the latency, throughput and cost metrics of a NoC architecture [15]. In 2012, Fischer and Fettweis presented an accurate service estimation model for the IQ NoC fabrics with RR packet arbitration. Their approach is interesting as it takes into account the contention of multiple concurrent inputs and the characteristics of the RR arbitration [16] to elaborate a delay model. Authors of [17] studied the flow-control feedback probability between adjacent routers of a NoC as key step to evaluate the total performance of the network. In 2015, Karadeniz *et al.* presented a low-complexity model for a single-stage switch based on Network-on-Chip and OQ routers [18]. In a similar way, we propose a detailed model for the primal performance metrics of the switch; throughput and packet delay. We also give an estimation for the upper-bound of the blocking probability inside the central-stage OQ-UDN modules of the switch. The analytical models give feedback about the switch behaviour which is useful in the design optimization loop (specifying the central modules' size, the expansion of the NoC modules, as well as the output buffers capacity).

Next, we provide a full description of the switch design, the OQ-UDN modules and the packet routing process before deriving into the analytical modelling and the performance evaluation.

III. CLOS-UDN SWITCH WITH OUTPUT-QUEUED MINI-ROUTERS

In this section, we provide a description of the multistage switch, the central-stage OQ-UDNs and the packet routing process. We also give a rough estimation of the switch complexity and we justify its implementation feasibility.

⁴Unlike with IQ routers where contention for links causes random delay variations.

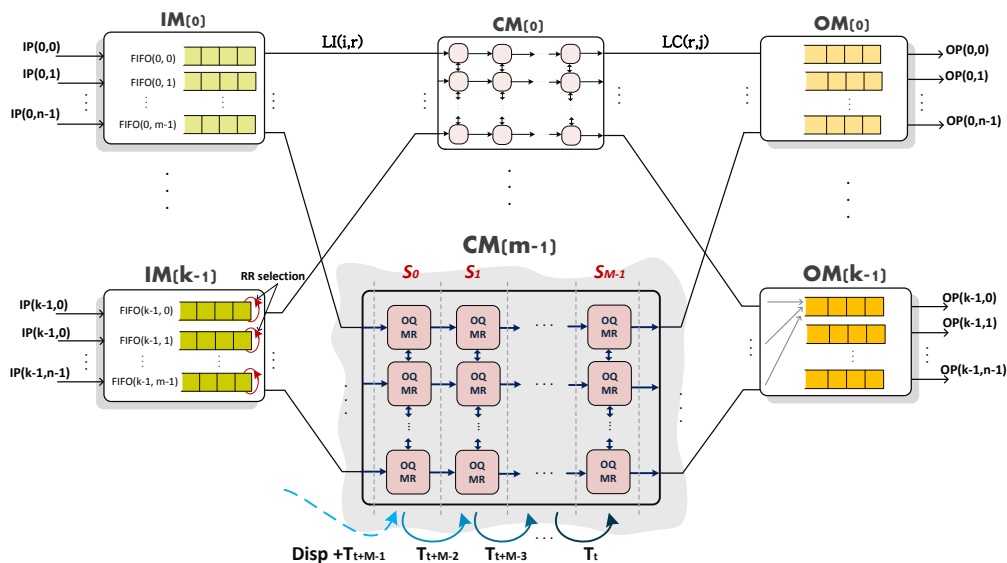


Fig. 1: $(N \times N)$ three-stage OQ Clos-UDN packet-switch architecture

A. Nomenclature of the switching architecture

We describe a three-stage Clos-network switch with output-queued NoC fabric. The design is a nested network as Fig.1 depicts. The first stage of the switch is made of k Input Modules (IMs), each of which is of size $(n \times m)$. An $IM(i)$ has m FIFOs⁵ each of which is associated to one of the m output links that we denote as $LI(i, r)$. An $LI(i, r)$ link connects the $IM(i)$ to the $CM(r)$. It can receive at most one packet from an input FIFO, and sends at most one packet to one CM at every time slot. The middle stage of the switch consists of m OQ UDN modules of dimension $(k \times M)$, each⁶. The $CM(r)$ has k output links that we denote as $LC(r, j)$. The LC links serve to connect a given CM to the different OMs at the third stage. The last stage has k Output Modules (OMs), each of which is of size $(m \times n)$. An $OM(j)$ has n Output Ports (OPs) that we denote $OP(j, h)$ for which is associated an output buffer. Each output buffer can receive at most m packets and forwards one packet to the output line at every time slot. Although it can be general⁷, the proposed OQ Clos-UDN architecture has an expansion factor $\frac{m}{n} = 1$, making it a *Benes* lowest-cost practical non-blocking fabric.

B. The OQ-UDN modules

In what follows, we describe the middle-stage modules of the OQ Clos-UDN packet-switch. A central-stage OQ-UDN module is a 2-D mesh fabric that can be fully defined by the

⁵Because $m = n$, each $FIFO(i, r)$ of an input module, $IM(i)$, is associated to one input port.

⁶Unlike conventional Clos networks, the central modules of the OQ Clos-UDN can be of size $(k \times M)$ crosspoints, where M refers to the NoC depth and $M = k$.

⁷The multistage switch can be of any size, where $m = n$. In this case, we would simply require a packets insertion policy in the input queues in order to maintain low-bandwidth FIFOs and to avoid the design purpose disruption (simple input modules). We consider this to be out of the scope of the current work.

2-tuples (k, M) where k^8 is the number of LI/LC links, and M is the depth of the mesh layout (*i.e.*, the number of pipeline stages, or also expansion factor of the mesh network). The NoC assimilates $k \cdot M$ mini-routers with two or three I/Os (referred to as degree of a MR) depending on its position on the grid. We implement a deadlock-free routing algorithm “*Modulo XY*” [5] and a credit-based flow-control mechanism to transfer packets across the mesh. This allows the upstream MRs to keep track of the free room in each output buffer downstream, and avoids elastic buffers. In the rest of the paper, we assume that packets are of fixed-size, and that all relative routing information are stored to their headers. Next, we describe the routing process inside the central-stage of the Clos switch.

C. Routing in the OQ NoC fabric

Packets are dispatched from the IMs to the central stage modules in a RR manner. At every time step, each input arbiter among the m arbiters associated to the FIFO queues in an IM, selects an OQ-UDN module to which it sends the HoL packet. At their arrival to the selected CM, packets start being routed locally until exiting the NoC module to the output stage of the Clos-network architecture. They travel *West East, West South, West North, South East, South North* and *North South*; from the left-most MRs to the right-most nodes of the OQ-UDN modules until exiting the central stage to the corresponding OMs. We propose a dimension order algorithm to forward packets across the NoC fabric. The routing approach is called “*Modulo XY*”, and it is an advanced version of the classic “*XY*” scheme. It routes packets along one dimension, then along the second dimension

⁸A UDN module has k input/output ports and M NoC stages. When the UDN is part of the multistage switch, the term k is reserved for the LI and LC links that relate the middle stage modules to the first and last stage blocks – respectively.

Using the initial input traffic load of the route r and the number of buffers that a packet runs through, we compute $A(r)$. Clearly, the probability of availability of the route r concerns with the remaining subset of queues after we exclude the first j buffers as (A.3) shows. This means that it depends on $r - \{r_1\} - \{r_j, \dots, r_1\} = r - r^{(j)}$. Hence, we can write (A.3) in a different way:

$$Pr(e_j | e_{j-1} \dots e_1) = \prod_{s=2}^{r_j} \frac{A(s - r^{(j)})}{s} A(r - r^{(j)}) \quad (A.4)$$

Taking into account that $Pr(e_1) = A(\{r_1\}) = r_1$ and using the system of equations in (A.4), we infer (A.1).

Being a probability, the factor $A(r - r^{(j)})$ should result in the following inequality:

$$A(r) \leq \prod_{j=1}^r \frac{A(r - r^{(j)})}{r_j} \quad (A.5)$$

Where $\tilde{j} = \sum_{s=1}^{r_j} s$, is the traffic intensity that comes from all routes $r \in R_{r_j}$ and falls into to queue r_j . Finally, we use the general inequality $1 - \prod_{s=1}^Q a(j) \leq \prod_{s=1}^Q (1 - a(i))$, and we derive an upper bound on the blocking probability $B(r)$.

$$B(r) \leq \prod_{j=1}^r \frac{A(r - r^{(j)})}{r_j} \leq \prod_{j=1}^r \frac{A(r - r^{(j)})}{P_{ctr(j)}} \quad (A.6)$$

We conclude (16).

ACKNOWLEDGMENT

This work was supported by the EU Marie Curie Grant (SCALE: PCIG-GA-2012-322250).

REFERENCES

- [1] N. I. Chrysos, "Request-grant scheduling for congestion elimination in multi-stage networks," Crete University, 2006, Tech. Rep.
- [2] C. Clos, "A study of non-blocking switching networks," *Bell System Technical Journal*, vol. 32, no. 2, pp. 406–424, 1953.
- [3] F. M. Chiussi, J. G. Kneuer, and V. P. Kumar, "Low-cost scalable switching solutions for broadband networking: The ATLANTA architecture and chipset," *Communications Magazine on*, vol. 35, no. 12, pp. 44–53, 1997.
- [4] Z. Dong and R. Rojas-Cessa, "Non-blocking memory-memory-memory Clos-network packet switch," in *34th Sarnoff Symposium on*. IEEE, 2011, pp. 1–5.
- [5] K. Goossens, L. Mhamdi, and I. V. Senin, "Internet-router buffered crossbars based on networks-on-chip," in *DSD'09, 12th Euromicro Conference on*. IEEE, 2009, pp. 365–374.
- [6] E. Bastos, E. Carara, D. Pigatto, N. Calazans, and F. Moraes, "MOTIM—a scalable architecture for Ethernet switches," in *ISVLSI'07, Symposium on*. IEEE, 2007, pp. 451–452.
- [7] F. Moraes, N. Calazans, A. Mello, L. Möller, and L. Ost, "HERMES: An infrastructure for low area overhead packet-switching networks on chip," *INTEGRATION, the VLSI journal*, vol. 38, no. 1, pp. 69–93, 2004.
- [8] T. Karadeniz, L. Mhamdi, K. Goossens, and J. Garcia-Luna-Aceves, "Hardware design and implementation of a network-on-chip based load balancing switch fabric," in *ReConFig, Conference on*. IEEE, 2012, pp. 1–7.
- [9] L. Mhamdi, K. Goossens, and I. V. Senin, "Buffered crossbar fabrics based on networks on chip," in *CNSR, Conference on*. IEEE, 2010, pp. 74–79.
- [10] A. Bitar, J. Cassidy, N. Enright Jerger, and V. Betz, "Efficient and programmable Ethernet switching with a NoC-enhanced FPGA," in *ANCS, 10th Symposium on*. ACM/IEEE, 2014, pp. 89–100.
- [11] F. Hassen and L. Mhamdi, "A multi-stage packet-switch based on noc fabrics for data center networks," in *Globecom Workshops (GC Wkshps)*. IEEE, 2015, pp. 1–6.
- [12] N. McKeown, A. Mekkittikul, V. Anantharam, and J. Walrand, "Achieving 100% throughput in an input-queued switch," *Transactions on Communications*, vol. 47, no. 8, pp. 1260–1267, 1999.
- [13] H. Elmiligi, M. El-Kharashi, and F. Gebali, "Modeling and implementation of an output-queueing router for networks-on-chips," *Embedded Software and Systems*, pp. 241–248, 2007.
- [14] U. Y. Ogras and R. Marculescu, "Analytical router modeling for network-on-chip performance analysis," in *DATE'07, Conference on*. IEEE, 2007, pp. 1–6.
- [15] S. Suboh, M. Bakhouya, J. Gaber, and T. El-Ghazawi, "Analytical modeling and evaluation of network-on-chip architectures," in *HPCS, International Conference on*. IEEE, 2010, pp. 615–622.
- [16] E. Fischer and G. P. Fettweis, "An accurate and scalable analytic model for round-robin arbitration in network-on-chip," in *NoCS, 7th International Symposium on*. IEEE/ACM, 2013, pp. 1–8.
- [17] Y. Zhang, X. Dong, S. Gan, and W. Zheng, "A performance model for network-on-chip wormhole routers," *Journal of Computers*, vol. 7, no. 1, pp. 76–84, 2012.
- [18] T. Karadeniz, A. Dabirmoghaddam, Y. Goren, and J. Garcia-Luna-Aceves, "A new approach to switch fabrics based on mini-router grids and output queueing," in *ICNC, International Conference on*. IEEE, 2015, pp. 308–314.
- [19] E. Oki, Z. Jing, R. Rojas-Cessa, and H. J. Chao, "Concurrent round-robin-based dispatching schemes for Clos-network switches," *Networking, Transactions on*, vol. 10, no. 6, pp. 830–844, 2002.
- [20] F. Gebali, *Computer communication networks: Analysis and design*. Northstar Digital Design, Incorporated, 2005.
- [21] —, *Analysis of computer networks*. Springer, 2015.
- [22] A.-L. Beylot and M. Becker, "Dimensioning an ATM switch based on a three-stage Clos interconnection network," in *Annales des télécommunications*, vol. 50, no. 7-8. Springer, 1995, pp. 652–666.
- [23] L. Le and E. Hossain, "Tandem queue models with applications to qos routing in multihop wireless networks," *Mobile Computing, Transactions on*, vol. 7, no. 8, pp. 1025–1040, 2008.
- [24] M. E. Ekpenyong and J. Isabona, "Performance modeling of blocking probability in multihop wireless networks," *Journal of Applied Science & Engineering Technology*, vol. 4, 2011.
- [25] A. E. Kiasari, Z. Lu, and A. Jantsch, "An analytical latency model for networks-on-chip," *VLSI Systems, Transactions on*, vol. 21, no. 1, pp. 113–123, 2013.
- [26] H. Yoon, K. Y. Lee, and M. T. Liu, "Performance analysis of multi-buffered packet-switching networks in multiprocessor systems," *Computers, Transactions on*, vol. 39, no. 3, pp. 319–327, 1990.
- [27] X. Li, Z. Zhou, and M. Hamdi, "Space-memory-memory architecture for Clos-network packet switches," in *ICC, International Conference on*. IEEE, 2005, pp. 1031–1035.
- [28] A. Faragó, "Efficient blocking probability computation of complex traffic flows for network dimensioning," *Computers & Operations Research*, vol. 35, no. 12, pp. 3834–3847, 2008.

Fadoua HASSEN received her M.S. degree in Telecommunications Communication engineering (with distinction) from the Higher School of Communication of Tunis, SUPCOM University of Carthage, in 2011. She is currently working towards the Ph.D degree in Electrical Engineering at the University of Leeds. Her research interests include high performance packet-switch design, scalable switching architectures and switching/routing in Data Center Networks. She is a student member of the IEEE.

Lotfi MHAMDI received the Master of Philosophy (MPhil.) degree in computer science from the Hong Kong University of Science and Technology (HKUST) in 2002 and the Ph.D. degree in computer engineering from Delft University of Technology (TU Delft), The Netherlands, in 2007. He continued his work at TU Delft as post-doctoral researcher, working on high-performance networking topics within various European Union funded research projects. Since July 2011, he has been a Lecturer with the school of Electronic and Electrical Engineering at the University of Leeds, UK. Dr. Mhamdi is/was a technical program committee member in various conferences, including the IEEE International Conference on Communications (ICC), the IEEE GLOBECOM, the IEEE Workshop on High Performance Switching and Routing (HPSR), and the ACM/IEEE International Symposium on Networks-on-Chip (NoCS). His research work spans the area of high-performance networks including the architecture, design, analysis, scheduling, and management of high performance switches and Internet routers. He is a member of the IEEE.