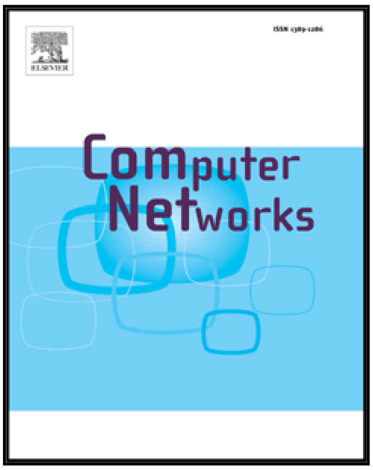

A Scalable Packet-Switch Architecture Based on OQ NoCs for Data Center Networks

Fadoua Hassen, Lotfi Mhamdi

PII: S1389-1286(17)30310-9
DOI: 10.1016/j.comnet.2017.08.003
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

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 modules[1]. (ii) Simplifying the packet dispatching process and avoiding the need for complex and synchronized scheduling algorithms[2]. (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

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).

[1] Actually, 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.

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

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 reminder 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 nodes[3] for the following reasons: (1) Higher throughput is achieved and the overall packets delay is shifted by a fixed amount[4]. (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.

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

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

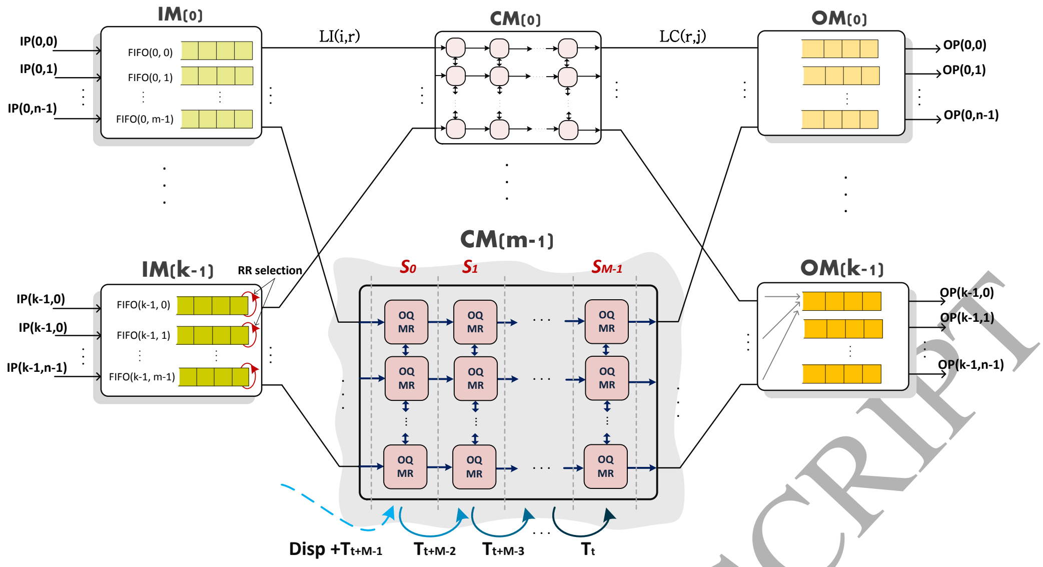

Fig. 1: $(N \times N)$ three-stage OQ Clos-UDN packet-switch architecture

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 FIFOs[5] 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 $(k \times M)$, each[6]. 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 general[7], 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 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 $West \rightarrow East$, $West \rightarrow South$, $West \rightarrow North$, $South \rightarrow East$, $South \rightarrow North$ and $North \rightarrow South$; 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

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

[6] Unlike 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 $M \leq k$.

[7] The multistage switch can be of any size, where $m \geq n$. 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.

[8] A 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.

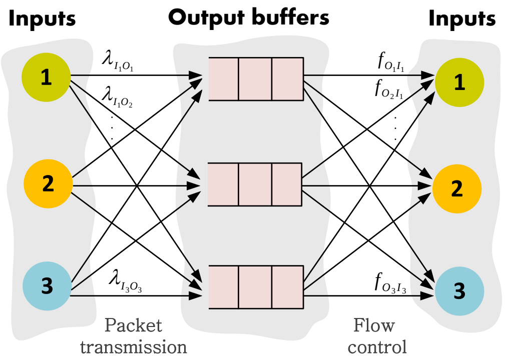

Fig. 2: Routing process in an a mini-router of the OQ-UDN central module

of the mesh, and it takes advantage of the path diversity by introducing an extra turn before the last column in the 2-D mesh topology. The algorithm is simple to implement and inherently deadlock-free. The routing decision is incremental in the sense that path computation is processed at every node in the OQ-UDN which removes the packet overhead that all-at-once routing algorithms create.

The process of packet routing across the NoC fabric is made of two phases: Packets transmission and feedback control; as illustrated in Fig.2[9]. We implement a buffering credit based flow-control that generates a feedback-control signal at each time a packet tries to access an output buffer. This signal throttles down the packets forwarding to a saturated queue, and avoids buffer overflowing. Upon making the routing decision, the adequate switching elements are activated and made ready to move packets to the correct output port. We opt for the store-and-forward switching mode to develop a backlog of frames waiting for the output port facility to become available. Hence, the on-chip routers need to wait for the whole packet before making the forwarding decision which makes the forwarding delay dependent on the size of the packet. However, the current flexibility and performance advancements in the Application-Specific Integrated Circuit (ASIC) design allow successful low-latency store-and-forward implementations (e.g., Cisco Catalyst® 4900M Switch). In the following sub-section, we give a coarse estimation of the OQ Clos-UDN switch implementation complexity, and we justify its feasibility using the currently available technology.

D. Implementation feasibility of the switch

The design challenge includes the implementation of the OQ Clos-UDN switch with competitive performance and feasible complexity, thus satisfying all the identified design goals. Using NoCs for packet-switching fabrics has been a gradual process, with interconnects evolving from a single bus to multiple buses with bridges and crossbars. The design itself is a key offering low communication latency, power consumption and modularity. It enables the fabric to handle various application traffics with different characteristics. NoCs are inherently parallel in nature, with distributed arbitration for resources. Consequently, multiple transactions between the on-chip routers can take place concurrently in different parts of the mesh layout.

The OQ Clos-UDN switch requires two abstraction levels to pinpoint packets routes in the network. We roughly estimate the requirements of the multistage switch in general and the OQ-UDN modules in particular. The process of packet dispatching is non-iterative – in contrast with common semi-buffered Clos switches [19] that require maximum and maximal-weight matching algorithms to define the set of interconnected IMs/CMs and port pairs at a time. The current switch simplifies the scheduling process as following: At every time slot, each of the m input arbiters at an IM selects the CM which priority appears next in the RR selector, and dispatches the HoL packet to it. The operation results in a complexity time of $\mathcal{O}(log\ m)$ and also a hardware complexity of $\mathcal{O}\ (log\ m)$ per IM. Interestingly, the dispatching process and packet routing through CMs work in parallel making the action of dispatching at time slot t $(Disp)$ and the action of packet forwarding through the NoC (T_t) overlapping; as shown in Fig. 1. Assume F_0, the flow of packets sent to a particular central module at time slot $t = 0$. F_0 arrives to the on-chip routers on the first column M_0 of the OQ-UDN. After examining the packets headers, forwarding decisions are made, and the packets are transferred to their next hop. At time slot $t = 1$, a new flow of packets – F_1 – comes to M_0 while F_0 gets routed to the next stage of the CM.

As we mentioned earlier, every central block is made of $(k \cdot M)$ mini-routers all fitted with small output queues that absorb traffic with respect to their capacity. Although in a MR of degree n, all output memories must run n times faster than an input port to handle the worst case scenario, the hardware implementation of the module is still feasible. In [18], authors discuss register-transfer level (RTL) implementation of a single-stage WUDN packet switch that is moderately similar to the OQ-UDN. Considering the current technology, we argue that the HW implementation of a module is perfectly feasible, and that a cost/performance trade-off can be made by varying the NoC-based switch parameters and/or the synthesis technology.

In the next section, we give details of the analytical modelling of the OQ Clos-UDN switch performance metrics: The throughput, the average packet delay and the blocking probability.

IV. Inside the OQ-UDN: Modelling the output-queues

After the design step, comes the evaluation of any architecture which is usually done through simulations. Generally, simulations are extremely slow for large-scale systems and they provide little insight on how the different design parameters affect the actual switch performance. Analytical models, however, allow fast evaluation of large systems in early design phase taking advantage of the rapid trade-off

[9] In the figure we use $\lambda_{I_x,Oy}$ to refer to the rate of traffic flowing from input I_x to output O_y of a mini-router. $f_{Oy,Ix}$ denotes to the probability of the feedback control issued by any output queue to any of the input ports.

design investigations. The variety of parameters in the OQ-UDN modules is the essence for a high flexibility. It spans a large design space making room for both parametrization and optimization. Fig. 3 depicts a high-level diagram of one MR used for the OQ-UDN modules. The output-queues serve simultaneously as FIFOs, and can accommodate up to B packets, each. Every output has n input ports to serve ($n = 2$ *or* 3 depending on the MR coordinates in the mesh).

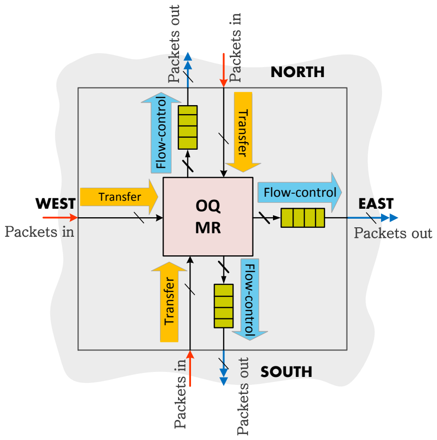

Fig. 3: Block diagram of an output-queued mini-router

It is common practice to study the finite capacity queues using Markov chains analysis. Assume a *Bernoulli i.i.d* packet arrival process with P_{arr} being the probability that a packet arrives to one of the outputs of a mini-router. We denote P_{dep}, the probability of packets departure from a buffer. We also assume that arrival times and service times are independent, and that they can both happen at the same time step. Finally, each output queue of a MR can be modelled as an M/M/1/B queue which state transition diagram is shown in Fig. 4. The transition probabilities of the buffer moving from one state to another are obtained by considering the way in which a packet can move between the two states and the probabilities for movements. Overall, a state transition in an output queue of the MRs consists of two phases: First, verify the availability of the buffer space, and – second move packets forward by one NoC stage. For the M/M/1/B system, changes of the queue size occur by at most one per time step [13]. We denote $b = 1 - p_{arr}$, the probability that no packet arrives to the output buffer and $d = 1 - P_{dep}$, the probability that a packet do not leave the output queue. To describe the transition diagram for the output queue we define the following variables:

- α: The probability that a packet arrives to the output buffer but do not leave it at the current time step. This causes the number of packets in the output queue to increment by a unit.
- β: The probability that a previously arriving cell leaves the output queue. This decrements the number of queued cells in the buffer.
- f: The probability that the queue size remains intact. This can happen in one of two possible scenarios: A currently arrived cell leaves the output queue or no cell arrives or gets removed from the queue at the current time step.

The state transition diagram for an output queue is shown in Fig. 4. The transition matrix is another way to represent information about its state variation. It can be written as follows:

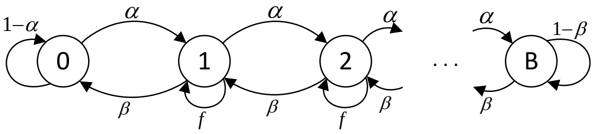

Fig. 4: State transition diagram for a single M/M/1/B queue

$$P = \begin{bmatrix} \alpha_0 & \beta & 0 & \dots & 0 & 0 & 0 \\ \alpha & f & \beta & \dots & 0 & 0 & 0 \\ 0 & \alpha & f & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & f & \beta & 0 \\ 0 & 0 & 0 & \dots & \alpha & f & \beta \\ 0 & 0 & 0 & \dots & 0 & \alpha & \beta_0 \end{bmatrix}$$

where:

$$\alpha = p_{arr}\ d = p_{arr}\ (1 - P_{dep}) \tag{1}$$

$$\beta = (1 - p_{arr})\ P_{dep} \tag{2}$$

$$\begin{aligned} f &= p_{arr}\ P_{dep} + b\ d \\ &= 2\ p_{arr}\ P_{dep} + 1 - (P_{dep} + p_{arr}) \\ &= 1 - (\alpha + \beta) \end{aligned} \tag{3}$$

and $\alpha_0 = 1 - \alpha$ and $\beta_0 = 1 - \beta$. We define the state vector S, for which every element s_i indicates the probability of finding the queuing system in state s_i at that time step [20]. The first element s_0 reflects the probability that the queue is empty, while s_B is the probability that the queue is fully populated.

$$S = \begin{bmatrix} s_0 & s_1 & s_2 & \dots & s_B \end{bmatrix}^t$$

The equilibrium condition of the output buffer can be written as: $PS = S$, which yields the following set of difference equations.

$$\begin{cases} \alpha s_0 - \beta s_1 & = 0 \\ \alpha s_{i-1} - g s_i + \beta s_{i+1} & = 0, \quad 0 < i < B \end{cases} \tag{4}$$

where $g = \alpha + \beta$. We resolve the system of equations in (4), and we conclude the generic form of s_i that is the following:

$$s_i = (\frac{\alpha}{\beta})^i\ s_0, \quad 0 \leq i \leq B \tag{5}$$

The OQ state of occupancy changes over time. It can be one among the s_i states at a given time step which means that at a time, the summation $\sum_{i=0}^{B} s_i = 1$. Thus, we infer the probability s_0 that the queue is empty.

$$s_0 = \frac{1 - \tau}{1 - \tau^{B+1}} \tag{6}$$

Where τ is the magnitude of the distribution vector S given by:

$$\tau = \frac{\alpha}{\beta} = \frac{P_{arr}(1 - P_{dep})}{P_{dep}(1 - P_{arr})} \quad (7)$$

Using the previous equations, we readily compute the throughput of a single M/M/1/B queue [21].

$$Th_0 = P_{arr}\ P_{dep}\ s_0 + \sum_{i=1}^{B} P_{dep}\ s_i \quad (8)$$

In the following section, we elaborate analytical models for the OQ Clos-UDN switch throughput, end-to-end delay, as well as the blocking attitude of the architecture.

V. Analytical Modelling of Performance Metrics of the Switch

The OQ Clos-UDN switch has a NoC fabric instead of the classical crossbar/memory modules. In addition to intermediate links relating two successive Clos-network stages, multiple routes are available within a single central module providing better path diversity and better load distribution. Starting their journey in the switch fabric, packets need be sent from the input module buffers to the OQ-UDN blocks. For simplicity and fairness, we use a RR selection of the routes between any IM and CM.

The analytical approach of modelling needs complete knowledge of the parameters and inputs of the multistage switch to rigorously describe the architecture. In the following, we assume that on each input of the first stage, cells are generated according to an independent *Bernoulli* process. We also consider that the choice of a CM at the packet dispatching phase is independent and equidistributed [22] among the central SEs. Hence, we can break the analysis to separately model the switching stages of the OQ Clos-UDN architecture. We build an approximated analytical model to get the switch performance mainly by making use of the queuing theory and Markov chains. The set of eventual parameter values that impact the performance of the OQ Clos-UDN switch is very large. However, we focus on the OQ-UDN modules in the middle stage of the Clos switch given their interesting proprieties. In the next sub-sections, we give an estimation of the throughput of the proposed multistage switch by making use of the previous single output-queue model.

A. Characterization of the the throughput of Clos-UDN switch

In general, the throughput of the network is the rate of packets delivered to their ultimate destinations. At low traffic loads,the delivery rate is equal to the packet arrival rate while it saturates with the increasing load [20]. The factors contributing to the throughput saturation are substantially the topology of the network, the routing algorithm and the feedback control mechanisms (if any is used). As for our proposal, exits of the MRs in the last column of the OQ-UDN modules are related to output buffers in the OMs. For the sake of comparison with the simulated switch performance, we consider that buffers of the OMs have infinite capacity, which means that analysing the throughput of the OQ Clos-UDN switch can be reduced to evaluating the packet delivery rate in the OQ-UDN central modules.

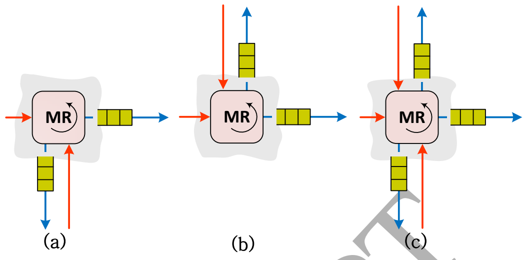

Fig. 5: Types of MRs in an OQ-UDN switch based on their degrees

The NoC-based SEs regroup three different types of MRs based on the degree of the routers as Fig. 5 shows. Overall there are $2M$ nodes of degree 2 and $M(k-2)$ mini-routers of degree 3. As we mentioned earlier, the central modules are supposed to work independently one from the other. Subsequently, to characterize the total throughput of the multistage switch, we examine the average number of packets that exit one central block. At this level, we assume that the processes of packet arrival and departure to and from on-chip nodes, are independent of each other. Consequently, the average throughput of a single OQ-UDN module can be seen as the summed contributions of the last column's MRs and it can be described using the following equation:

$$Th_{CM} = \sum^{2}(Th_{deg_2}) + \sum^{k-2}(Th_{deg_3}) \quad (9)$$

Where Th_{deg_2} and Tth_{deg_3} are the average throughput of MRs of degree 2 and degree 3 – respectively. Since all I/O(s) of a MR work independently and simultaneously to contribute to the average throughput of the node, we can derive expressions of Th_{deg_2} and Th_{deg_3} using (16) as the average contribution of as many M/M/1/B queues as the degree of the MR.

B. Average end-to-end delay

The average packets latency is an important metric that helps evaluate the performance of a switching network especially in multi-hop networks that are more delay-sensitive than single-stage point-to-point connected crossbars. The average end-to-end delay in the OQ Clos-UDN switch might be viewed as the contribution of – mainly – the input delay at the first stage of the Clos-network and the delay across the NoC fabric[10].

Considering *Bernoulli i.i.d* packet arrivals to happen at the input ports of the OQ Clos-UDN switch, we define λ to be the average arrival rate and μ to be the service rate. The input queues at the first stage of the Clos-network switch can be approximated with an M/M/1 system which mean waiting time is given by:

[10] To map with the simulation environment, we consider that output buffers associated to the switch output ports in the third stage of the Clos-network are of infinite size and that once packets exit the OQ-UDN central modules, they all leave the output buffers to their corresponding output ports after a fixed time.

$$\bar{W}_{M/M/1} = \frac{\rho'}{\mu(1-\rho')} \tag{10}$$

ρ' is the utilization factor equal to λ/μ' and μ' is the modified input FIFOs' service time subject to the MR's buffers availability, P_{fwd}.

$$\mu' = P_{fwd}\ \mu \tag{11}$$

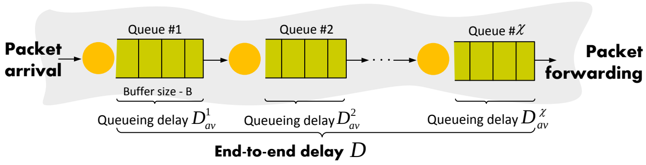

Fig. 6: Packet delay in a tandem queue

After being dispatched to the middle stage of the Clos-network, packets cross the central NoC fabric hop-by-hop until reaching the LC links. We call path (or route), the set of successive links and output buffers that a packet has to cross from a source node to a destination node. We suggest modelling a route as a tandem queue with χ output buffers, as shown in Fig.6. Packets that are successfully received at the receiving side of each link are buffered in an output queue for either to be transmitted to the next hop or to be delivered to the third stage of the Clos switch, otherwise. We consider that the delay of transmission over an intermediate link between two successive queues is negligible in comparison to the buffering delay itself that depends on the following factors:

- The queue size, B
- The probability that a queue has i packets, s_i
- The probability of packet arrival to the queue, p_{arr}
- The probability of service at the queue, P_{dep}

The average number of time steps that a packet spends inside the queue is given by:

$$D_{av} = \frac{Q_{av}}{Th_0} \tag{12}$$

where Q_{av} is the average queue size given by:

$$Q_{av} = \sum_{i=0}^{B} i.s_i = \frac{\tau[1-(B+1)\tau^B + B\tau^{B+1}]}{(1-\tau)(1-\tau^{B+1})} \tag{13}$$

and Th_0 is the throughput of the queue which expression is given in (8).

Using Little's formula for a tandem of queues [23], the end-to-end delay can be written as:

$$D = \sum_{j=0}^{\chi} D_j \tag{14}$$

where D_j is the average queueing delay at queue j and given in (12).

C. End-to-end blocking probability in the switch

The MRs output buffers of the OQ Clos-UDN switch have limited capacity which means that it is necessary to control packets transfer to them. Under certain traffic patterns, packet flows invading output queues may rise the network's blocking attitude. In general, deriving the end-to-end blocking probability of a path in complex networks would be straightforward from the individual blocking probability of a single link (unitary portion of the path) if we assume that they are statistically independent. In case of the OQ-UDN switch, a path is made of passive input links (*i.e.,* that eventually impose no real constraints on packets transfer) and output queues of the on-grid routers. The availability of the buffering resource in outputs of the downstream MRs results in dependencies, and adds complexity to what could be a simple estimation of the end-to-end blocking probability. Next, we show that it is possible to estimate an upper-bound on the probability that any path in the OQ-UDN switch is blocked.

We call P_{ctr}, the probability that an output queue issues a feedback control signal at a time step. Referring to our previous analysis, we note that s_B is the state where a single output port's buffer is fully occupied which corresponds to the probability of the flow-control feedback generation. We have:

$$P_{ctr} = s_B = \tau^B \frac{1-\tau}{1-\tau^{B+1}} \tag{15}$$

Similar to the previous modelling approach, we unfold the OQ-UDN structure, and we analyse the packets sojourn across a route. The end-to-end blocking probability of a route r in the OQ-UDN fabric is bounded by the sum of the blocking probabilities of its output queues.

$$B(r) \leq \sum_{j=1}^{\chi} P_{ctr}(j) \tag{16}$$

The proof is provided in appendix A.

The next section assimilates the performance evaluation of the proposed switch under for different settings and traffic types.

VI. Performance Evaluation

We test the performance of the proposed switching architecture using an event-driven simulator where we consider different settings and a variety of traffic patterns.

A. Uniform packets arrivals

We first investigate the average end-to-end packet delay in the switch for different switch sizes, mesh depths M, and traffic types. Unless it is mentioned, the output buffers' capacity B, is set to the default value 3. We evaluate the delay performance of the OQ UDN switch working as a single-stage switch, and when being part of the three-stage Clos switch under smooth traffic arrivals. In all figures, we use the notation OQ-UDN to refer to a single-stage switch, while OQ Clos-UDN is reserved for the multistage architecture.

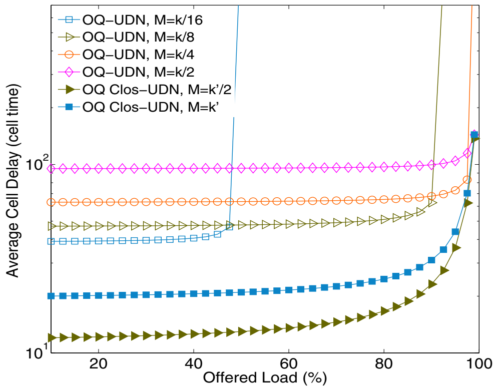

Fig. 7: Delay performance of 64-ports single-stage and multistage switches, under *Bernoulli i.i.d* traffic.

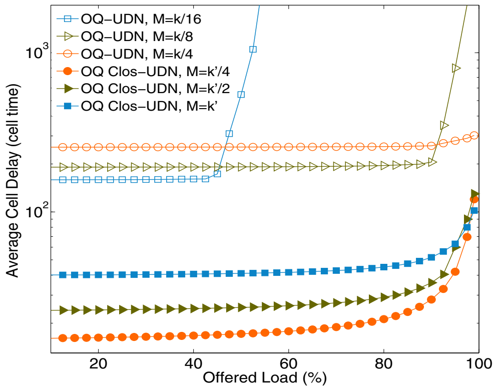

Fig. 8: Delay performance of 256-ports single-stage and multistage switches, under *Bernoulli i.i.d* traffic.

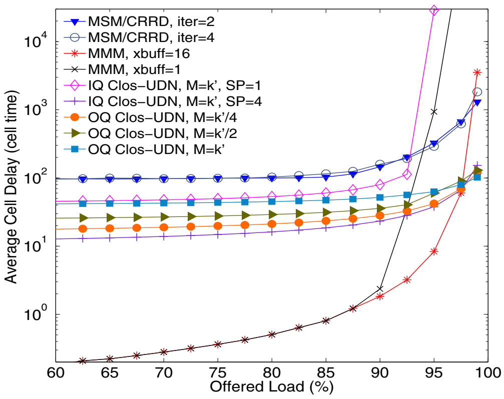

Fig. 9: Delay performance of a 256-ports MSM, MMM, IQ Clos-UDN and OQ Clos-UDN switch, under *Bernoulli i.i.d* traffic.

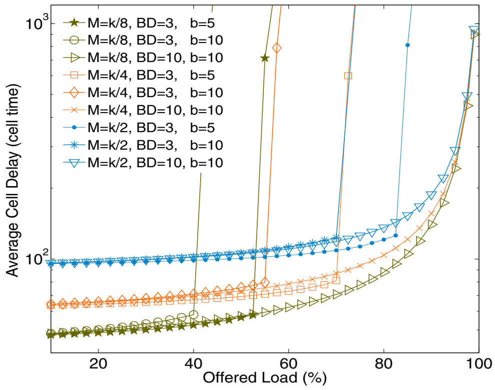

Fig. 10: Delay performance of a 64-ports single-stage OQ-UDN switch under *Bursty* uniform traffic.

1) Uniform Bernoulli traffic: Fig.7 depicts the variation of the average packet delay for a single and three-stage OQ-UDN switch under *Bernoulli i.i.d* arrivals. The parameters k and k' indicate the number of I/O ports of the standalone OQ-UDN and the Clos switch's central modules – respectively. Whether used in a single or multistage architecture, OQ-UDN design offers smooth delay variability for all proportions of the input load. Reducing the number of pipeline stages M, deteriorates the performance of a (64×64) single-stage switch. Unlikely, the OQ Clos-UDN seems less affected as it keeps on delivering up to 99% throughput even with small M values. This mainly reports to what a multistage architecture brings over a single-stage design. Actually, breaking the unique large NoC into small units mounted in a Clos fashion reduces the size of the central modules. It becomes possible to distribute packet flows to various CMs where they are routed through smaller UDNs with much reduced congestion level. We note that at some point, reducing M leads to the saturation of the single-stage OQ-UDN, and that the multistage architecture offers better control on the absolute delay in large-scale switches as Fig.8 depicts. Simulation results in Fig.8 clearly show that a single-stage design is unscalable in both the port count and traffic load. A (256×256) single-stage OQ-UDN switch can achieve full throughput only by expanding the NoC layout and setting $M = k/4 = 64$. However, this alternative is still unpractical and cost-prohibitive.

In Fig.9, we compare the delay performance of the proposed Clos switch to an MSM with the Concurrent RR Dispatching scheme (CRRD) [19], the MMM switch [4] and the IQ Clos-UDN switching architecture as being described in [11]. Our proposal outperforms MSM under heavy workloads where it categorically provides full throughput. The throughput of the IQ Clos-UDN switch saturates at around 90% provided that on-chip links run as fast as the external LI/LC links (*i.e.,* $SP = 1$). An MMM architecture affords lower delays. Yet, we still need large crosspoint buffers to achieve full throughput (*e.g.,* a total of 16 packets per crosspoint buffer). In the contrary, the OQ Clos-UDN switch running with small on-chip buffers $(B = 3)$ and $M = k'/4$ (that is only equal to 4 for (256×256) switch ports) ensures almost constant delay variations and provides high throughput.

2) Uniform Bursty traffic: In reality, workloads in the DCN are constantly changing. Distributed file systems in Big Data

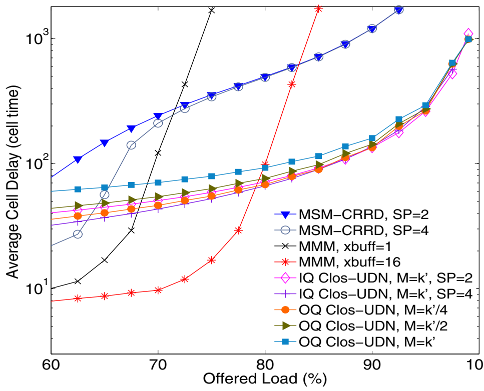

Fig. 11: Delay performance of a 256-ports MSM, MMM,IQ Clos-UDN and OQ Clos-UDN switch, under *Bursty* uniform traffic.

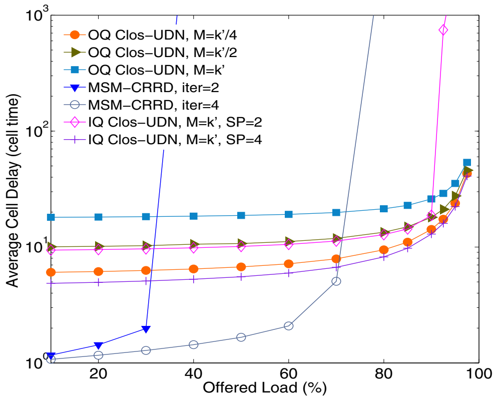

Fig. 12: Delay performance of a 64-ports MSM, MMM, IQ Clos-UDN and OQ Clos-UDN switch, under *Hot-spot* traffic.

analytics, streaming media services and many other high-bandwidth demanding applications make the bursty traffic pattern prevalent in a data centre network with high-levels of peak utilization. We presume that it is useful to examine how a bursty traffic impacts the proposed switch performance.

Fig.10 shows the latency of a (64×64) single-stage OQ-UDN under bursty traffic, for which we vary M, the on-chip queues' capacity and the size of the burst. Visibly, increasing M improves the switch throughput. Still, for $B = 3$ (minimum queues depth) and a burst size (b) of 10 packets, the NoC structure saturates, and the blocking ratio rises exponentially. Simulation results show that it is possible to ameliorate the switch response to burstiness by reducing the burst size. However, the throughput expansion is limited to about 14%. Providing larger queues for the mini-routers proves much effective to resolve the saturation problem at the expense of additional cost. On the whole, the standalone OQ-UDN as it is, do not scale in the switch size under bursty traffic unlike the multistage architecture that shows robustness and flexibility.

Fig. 11 illustrates the average end-to-end latency in the MSM, MMM and IQ/OQ Clos-UDN switches. Under heavy loads, both the conventional semi-buffered and fully-buffered Clos architectures yield worse delay performance than the NoC-based switches. The MMM switch cannot achieve full throughput even if the middle stage buffers are worth of 16 packets, each. The flexibility of NoCs and the multistage interconnect used along with a dynamic RR packet dispatching work towards better distributing the traffic load, and to conserve high throughput.

B. *Unbalanced traffic*

We evaluate a (64×64) Clos with OQ-UDN modules under non-uniform traffic whereby one fraction of the total input load is uniformly distributed among the switch outputs, and the other fraction goes to the output port with the same index as the issuing input port. If the unbalanced coefficient $\omega = 0$, then the traffic is perfectly uniform. On the other hand, if $\omega = 1$, then the switch deals with a directional traffic. Fig. 12 shows the average packet delay for the different switching architectures with variable settings, input loads and values of $\omega = 0.5$. As for uniform traffic, OQ Clos-UDN switch outperforms the MSM proposal with CRRD scheduling. Both IQ and OQ Clos-UDN switches can achieve comparable latencies if the parameters are adjusted (mainly the speedup SP and the mesh expansion factor M for the input-queued type, and M and the output buffers' depth B for an OQ-UDN module). Although both designs are highly customizable, an input-queued structure with no speedup and full mesh depth ($M = k' = 8$) do not achieve full throughput.

In Fig.15 we vary ω, and we observe the behaviour of a (64×64) OQ Clos-UDN switch as packet arrivals become more and more critical. The simulation scenario comprises three traffic types: Uniform, hot-spot and diagonal. We note that the blocking ratio in the central modules of the Clos switch evolves in the same way as the transferred packets ratio when we shift from a perfectly uniform to a diagonal traffic. When $\omega = 1$ (*i.e.,* diagonal traffic) and using the "$Modulo\ XY$" routing algorithm, packet flows travel horizontally in the NoC fabric towards the LC$(r,\ j)$ links. Averaging the proportion of packets over all nodes, results in a fraction being equally distributed among the rows of the OQ-UDN module. For a hot-spot traffic, packets cross northern and southern links to reach to outputs of the UDNs. This explains the disproportion of the load sent in the X and Y directions. We note that the amount of traffic crossing the NoC-based modules doubles under uniform traffic, and that the blocking probability gets less equalized across the mini-routers.

C. *Scalability of the switch*

Throughput stability is primordial. It reflects how resilient is, the switching architecture to traffic fluctuations and harshness. With the help of small on-chip queues, the incoming traffic is absorbed and transferred from one stage of the NoC to the subsequent stage. Fig.16 depicts that the IQ Clos-UDN with full depth ($M = k' = 8$) and $SP = 1$ achieves only up to 90% throughput. A buffered MMM architecture provides better throughput than the MSM with CRRD scheduling (60% throughput if $iter = 4$ and $\omega = 0.5$). In the contrary, the

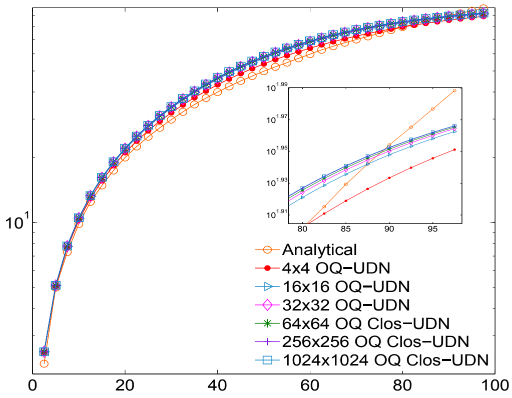

Fig. 13: The average switch throughput under *Bernoulli i.i.d* traffic arrivals.

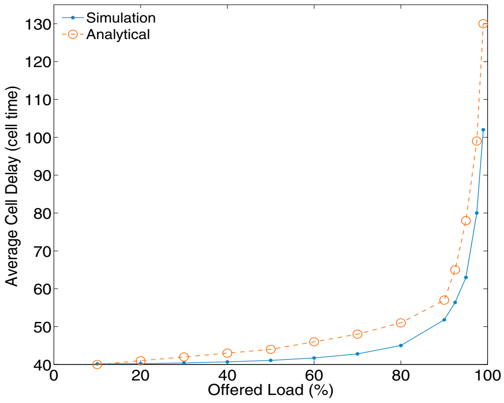

Fig. 14: Delay performance of a 256-ports switch under *Bernoulli i.i.d* traffic.

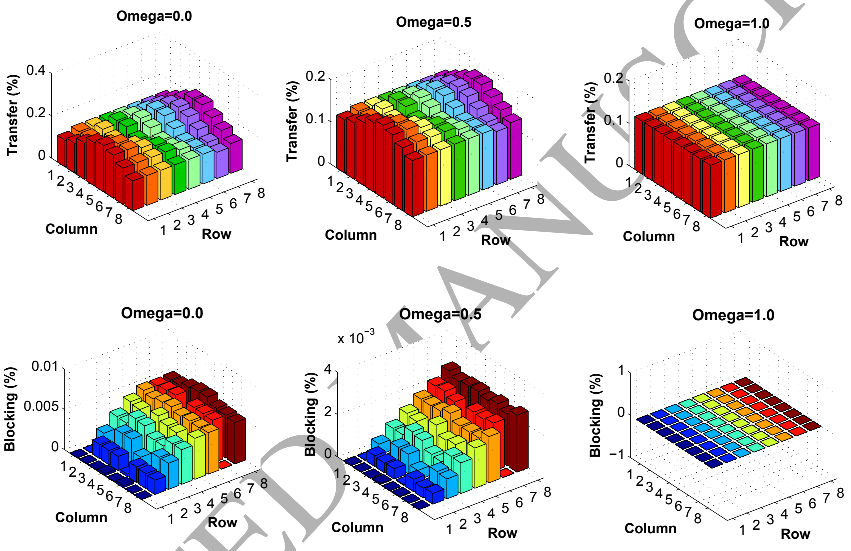

Fig. 15: Variation of the transfer and blocking ratios in the central modules of a 64-ports switch.

OQ Clos-UDN offers full throughput under the whole range of ω, even when the minimum-value settings ($B = 3$ and $M = k'/4 = 2$ for a 64-ports switch) are used.

DCN switches/routers must fulfil the large-scale in addition to the data-intensive communication prerequisites. In this context, we evaluate the impact of the switch size on the end-to-end packet latency. We vary the switch valency from 4 to 256, and we measure the overall delay in the OQ Clos-UDN switch under light loads (20%), medium loads (50%) and heavy loads (90%). We can see from Fig.17 that the delay variation under light and medium traffic loads is approximately the same no matter the ports count is. When the switch is heavily loaded, the latency increases slowly with the switch size. Yet, it does not exceed 50 time slots when the load is as high as 90%. In the following part, we compare the analytical results to outputs of the simulation.

D. Accuracy of the analytical model

We compare the analytical and simulation results for different switch sizes while se set the output buffers capacity B to 3 packets, each. Fig. 13 shows the variation in the throughput percentage under *Bernoulli i.i.d* arrivals. The proportion of throughput increases linearly when the input load increases because of the number of packets generated in the system. We can see that overall values obtained using the analytical model approach those of simulation. Under light loads, simulations perfectly match the analytical model. For a single-stage OQ-UDN switch, the deviation is about 4.5% when moderately medium traffic loads come to the switch inputs. This difference margin increases when the number of ports becomes higher and the traffic load becomes heavier. According to Fig.13, the more we increase the switch size, the bounded becomes our approximation. The fact that we simplified the model by dropping some architectural considerations, partially accounts for this lack of accuracy. However, the disparity between the

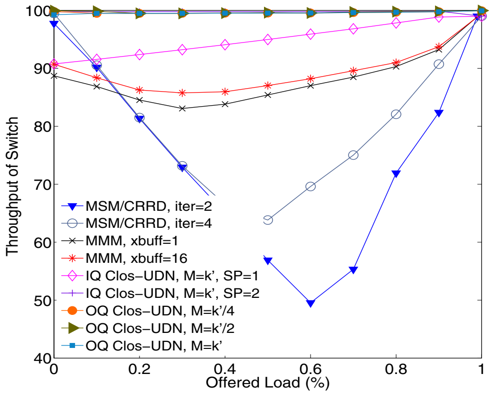

Fig. 16: Throughput stability of 64-ports switches, under *Unbalanced* traffic.

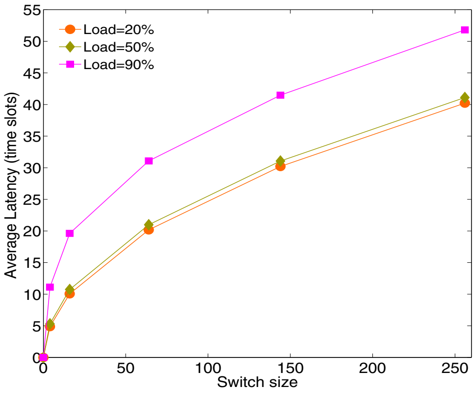

Fig. 17: The impact of the switch size variation on the delay performance under *Hot-spot* traffic.

analytical and the experimental results still do no not go beyond 7.98% for the smallest single-stage switch of size 4-ports, and 5.2% for a 64-ports multistage OQ Clos-UDN.

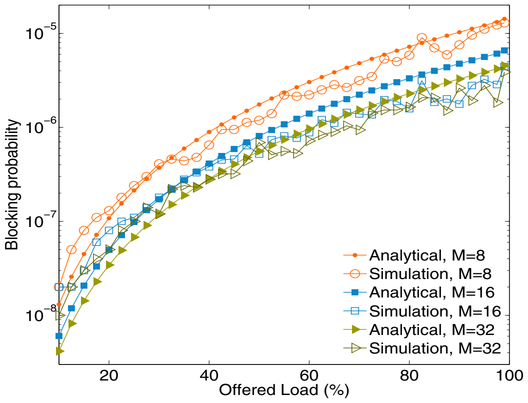

Fig. 18: The variation of the blocking probability in a 64-ports switch under *Bernoulli i.i.d* traffic.

Using the value of the offered input load, we can calculate the average arrival rate to the OQ Clos-UDN switch inputs λ. In the steady state where all inputs perpetually have pending packets, the service rate at the input FIFOs is restricted to the availability of the first output buffer that they request in the selected CM. This probability is given by $P_{fwd} = 1 - P_{ctr}$, and can be dynamically calculated whenever we have information about the arrival and departure probabilities in a single on-chip node. At the steady state operation mode, it is easy to calculate P_{dep}. Every output queue serves one of the n associated inputs with equal probability making $P_{dep} = 1/n$. Fig.14 shows the delay variation for a 256-ports switch for full mesh depth, $B = 3$ and uniform packet arrivals. It demonstrates that the simulation results decently support the analytical model.

The next set of simulations is performed to trace the end-to-end blocking ratio in the OQ Clos-UDN switch. Given the multistage switching topology and the assumptions that we made for the analytical analysis, we claim that the central modules are the one and only bottleneck of the design. Hence tracing $B(r)$ in the central stage modules, reflects the same blocking attitude of the OQ Clos-UDN switch[11]. In the following experiment, we choose a switch of size (64×64) that can work as single stage or be plugged into the middle stage of the Clos architecture for larger switch valency. We set $B = 3$ and we consider a *Bernoulli* uniform packet arrival process. We note that the blocking probability rises exponentially with the input load and that the proposed mathematical approximation of $B(r)$ approaches the simulation curves as depicted in Fig.18. For light workload intensities that are less or equal to 30%, the simulation results are located above the analytical curves. This is mainly due to the fact that the OQ Clos-UDN switch is not over-loaded, and that the packets travel flawlessly across the NoC fabric without a noticeable blocking. Therefore, our model do not correlate with the experimental results. As the traffic load becomes higher, the blocking likelihood of the NoC modules rises, and the analytical model starts reflecting the real behaviour of the switch. Regardless of the size of the NoC layout M; the blocking ratio $B(r)$ remains less than 10^{-5}.

VII. Conclusion

It is difficult to predict the growth of data center networks requirements, but we know that the switching fabrics need to make a significant progress towards bridging the performance and scalability gap in large scale clusters using high-performance switches and routers that handle challenging tasks. Many approaches have been considered to address latency, throughput and scalability issues in the design of multistage packet switches. Some of them achieve good performance at the expense of prohibitive cost and complexity. The OQ Clos-UDN is a highly-scalable multistage switch suitable for the DCN environment. It comprises NoC-based modules at

[11] We map to the simulation environment where input queues at the first stage and output queues at the last stage are of infinite capacity. This means that only the central modules contribute to blocking packets during their transfer.

the middle stage of its Clos switching fabric where every single on-chip node is an output-queued router. The design exploits the NoC topology as well as the output queueing approach to provide high-performance, scalability and parametrization that are all important in the DCN switching fabric context.

In this paper, we study the performance of the switch by simulations. We show that the proposed architecture outperforms the MSM, MMM and IQ Clos-UDN switches under a variety of traffic types. We jointly propose an analytical approximation for the theoretical throughput using the queueing theory and Markov chains. Although we drop some architectural considerations and simplify the analysis, the experimental results show that the average deviation of the model is about 7.9% for a single stage OQ-UDN (size < 64) and is around 5.2% for a Clos-UDN switch (size $\geq$ 64). Besides, we propose an approximation for the overall packet delay, and we estimate an upper bound on the end-to-end blocking probability in the central modules of the OQ Clos-UDN switch taking into consideration the inter-dependencies that the architectural design imposes. Given the buffered nature of the OQ Clos-UDN switch, packet are likely to arrive to their corresponding output ports out-of-order. Resolving this issue, and providing a hardware implementation of the switch are reserved for future work.

Appendix A
Proof of the Equation (16)

We prove (16) in two steps: First, we approach the availability probability in the central modules of the switch, then we infer an upper bound for the blocking probability.

Let $1, \ldots, m$, be the set of output buffers of the OQ-UDN module, and $\mathscr{R}$ the set of paths in the mesh network where each route $r \in \mathscr{R}$ is a non-empty set of output buffers connected by means of physical links. We define $\mathscr{R}_{r_j} \subseteq \mathscr{R}$, $r_j = 1, \ldots, m$, as the subset of paths that intersect in output buffer j. We assume that $\mathscr{R}_{r_j} \neq \emptyset$, and that at the steady state an output buffer is used by at least one path in $\mathscr{R}$. We also denote ν_r, the end-to-end traffic load of route $r \in \mathscr{R}$.

Assuming that the traffic getting out of the IMs of the OQ Clos-UDN switch after the dispatching phase is still stationary. It is worth-mentioning that although a uniform traffic is considered for this analysis, the current proof stands even for an arbitrary traffic type. At this point, we have no idea about the traffic intensity ρ_j coming to an output buffer at a time since it is not an input parameter like ν_r. However, this proportion of traffic, ρ_j, is the superposition of the load carried over the previous stretch of a given path. Obviously, ρ_j depends on the availability of the upstream MRs' output buffers that may in turn depend on other factors. To simplify, we set an upper bound on ρ_j that we denote $\breve{\rho}_j$ with respect to $\{\nu_r, \breve{\rho}_j \leq 1\}$.

A path is said to be blocked with a probability $B(r)$, if and only if at least one of the buffers a packet must go to on its route is not available. The blocking probability of any output buffer is an increasing function of its input traffic intensity. Diversely, the availability probability is a decreasing function of the same parameter. The blocking events might not be simply independent of other buffers located somewhere in the mesh network which makes the situation delicate to handle mathematically [24]. We suppose that a route has χ buffers and that for any output queue, with an input load ρ_j, the blocking probability is $P_{ctr}(j)$ (evaluated in sub-section V-C). We introduce a set of useful terms to prove (16). Consider $\alpha_r(\rho_j) = \alpha_{r_j}$, the probability that an output queue is available in a route r. We say that a route $r \in \mathscr{R}$ is available with a probability $A(r)$, if the whole set of buffers on the path are simultaneously available. Since the output buffers of MRs are not necessarily independent, the availability of the set of buffers all along a path is not always a simple product of the individual buffers availability probabilities. In other words, $A(r) \neq \nu_r \prod_{j \in r} (1 - P_{ctr}(j))$, $r \in \mathscr{R}$.

Note that if $r = r_j$ (a single output on the route which is possible if the number of pipeline stages of the OQ-UDN mesh, $M = 1$), then $A(r) = (1 - P_{ctr}(j))$. We denote $r^{[j]}$, the j first queues on the initial segment of the route r where $j \leq |\chi|$. If the initial segment is such that $j = 0$, then the route is an empty set of buffers for which we define $A(\emptyset) = 1$. We show that with an arbitrary dependency pattern, the probability that a route $r \in \mathscr{R}$ is available always satisfies the following equation:

$$A(r) = \prod_{j=1}^{\chi} \alpha_{r_j} \left(\sum_{s \in \mathscr{R}_{r_j}} \nu_s \; A(s - r^{[j]}) \right) ; r \in \mathscr{R} \setminus \{\emptyset\} \tag{A.1}$$

To prove (A.1), we consider $\tilde{\mathscr{R}}$, such that $\tilde{\mathscr{R}} = \mathscr{R} \setminus \{\emptyset\}$. If $\chi = 1$ then $r = r_1$, and we simply have

$$A(r_1) = \alpha_{r_1} \underbrace{\left(\sum_{s \in \mathscr{R}_{r_1}} \nu_s \; A(s - r_1) \right)}_{\psi} \tag{A.2}$$

The term ψ in (A.2), is the sum of the traffic intensities offered on all routes that enclose output queue r_1 multiplied by the probability that the remaining stretch of the route $s \in \mathscr{R}_{r_1}$ is available ($A(s - r_1)$). We can describe ψ differently as the $route-carried$ traffic load that ends up at the input of queue r_1 and that we previously denoted as ρ_1. Using (A.2), we conclude the following equivalence: $A(r_1) = \alpha_{r_1}(\rho_1) = \alpha_{r_1}$. We conclude that (A.1) holds for $\chi = 1$.

Now, we are ready to prove that the system of equations in (A.1) is still valid for routes of length $\chi > 1$ (*i.e.,* an arbitrary number of pipeline stages such that $M > 1$). We introduce the set of events $\{e_j, j = 1, \ldots, \chi\}$ to spot whenever an output queue is available with the probability $Pr(e_j)$. The probability that the whole path is not blocked can be expressed as a conditional probability in such a way that the availability of a set of outputs in the route depends on the previous buffers.

$$\begin{aligned} A(r) &= Pr(e_1) \frac{Pr(e_1 e_2)}{Pr(e_1)} \cdots \frac{Pr(e_1 e_2 \ldots e_\chi)}{Pr(e_1 e_2 \ldots e_{\chi-1})} \\ &= Pr(e_1) \prod_{j=2}^{\chi} Pr(e_k | e_{j-1} \ldots e_1) \end{aligned} \tag{A.3}$$

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 $r - \{r_1\} - \{r_{j-1}, \dots r_1\} = r - r^{[j]}$. Hence, we can write (A.3) in a different way:

$$Pr(e_j|e_{j-1}\dots e_1) = \alpha_{r_j}\left(\sum_{s\in\mathscr{R}_{r_j}} \nu_s \; A(s - r^{[j]})\right) \tag{A.4}$$

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

Being a probability, the factor $A(r - r^{[j]}) \leq 1$. Thus removing $A(r-r^{[j]})$ from the right-hand side of (A.1), should result in the following inequality:

$$A(r) \geq \prod_{j=1}^{\chi} \alpha_{r_j}\left(\sum_{s\in\mathscr{R}_{r_j}} \nu_s\right) \quad \geq \prod_{j=1}^{\chi} \alpha_{r_j}(\breve{\rho}_j) \tag{A.5}$$

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

$$B(r) \leq 1 - \prod_{j=1}^{\chi} \alpha_{r_j}(\breve{\rho}_j) \quad \leq \sum_{j=1}^{\chi} \left(\underbrace{1 - \alpha_{r_j}(\breve{\rho}_j)}_{P_{ctr(j)}}\right) \tag{A.6}$$

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-queuing 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-cchip 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 PhD. 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 tropics 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.