
Proceedings Paper:
Hassen, FH and Mhamdi, L (2016) Providing Performance Guarantees in Data Center Network Switching Fabrics. In: 2016 IEEE 17th International Conference on. High Performance Switching and Routing (HPSR), 2016 IEEE 17th International Conference on, 14-17 Jun 2016, Yokohama, Japan. IEEE , pp. 155-161. ISBN 978-1-4799-8950-8

https://doi.org/10.1109/HPSR.2016.7525660

Providing Performance Guarantees in Data Center Network Switching Fabrics

Fadoua Hassen Lotfi Mhamdi
School of Electronic and Electrical Engineering
University of Leeds, UK
Email: {elfha, L.Mhamdi}@leeds.ac.uk

***Abstract*—This paper proposes a novel and highly scalable multistage packet-switch design based on Networks-on-Chip (NoC). In particular, we describe a three-stage Clos packet-switch fabric with a Round-Robin packets dispatching scheme where each central stage module is an Output-Queued Unidirectional NoC (OQ-UDN), instead of the conventional single-hop crossbar. We test the switch performance under different traffic profiles. In addition to experimental results, we present an analytical approximation for the theoretical throughput of the switch under *Bernoulli i.i.d* arrivals. We also provide an upper-bound estimation of the end-to-end blocking probability in the proposed switch to help predict performance and to optimize the design.**

Keywords*—*Next-Generation Networking, DCN, Clos-network switch, NoC, Analytical model

I. Introduction

With the help of new designed switching architectures, large-scale networks such as Data Center Networks (DCNs) became able to process petabytes of data. Although the global DCN's reliability directly relies on the global design factors (network architecture, cabling, cooling, etc.), the performance of individual switches/routers can be a bottleneck. DCNs adopt commodity multistage switches/routers that demonstrate little scalability or prohibitive and complex implementation costs [1]. A step up in the design of switching architectures consists on building high-performance NoC fabrics to mitigate a number of limitations inherent to conventional single-hop crossbars, including scalability[1], port speed and path diversity [2]. The purpose of this work is to design a highly scalable packet switch architecture that takes advantage of both a multistage design and NoCs.

After the design step, comes the evaluation of any switching architecture which is usually performed using simulations. The fundamental limitation of event-driven simulations, is that they are extremely slow for large-scale systems. Simulators, often provide little insight on how the different design parameters affect the actual switch performance. The analytical models, however, allow fine-grained analysis and fast evaluation of large systems in early design phase. They also allow rapid trade-off investigations for the switching architecture, accelerate the estimation of the major metrics and ease the design process. In this paper, we suggest a multistage packet-switch suitable for DCNs environments in which we merge a Clos macro-design [3] and an OQ-UDN micro-design. We propose an analytical model to characterize the switch throughput using queuing models and Markov chains analysis. Besides, we estimate an upper bound on the overall blocking probability in the proposed architecture and we compare our model to simulation results.

The reminder of the paper is structured as follows: Section II discusses related work. We describe a three-stage Clos-network switch with OQ-UDN modules and we give details of the analytical modeling of the switch throughput in section III. In section IV, we give an estimate of the blocking probability in the OQ Clos-UDN. Our results are presented in section V and section VI concludes the paper.

II. Related work

NoC fabrics for packet switching have been proposed in some of the latest works [2] [4] [5]. The design approach offers high throughput, low latency and pipelined scheduling. Given the multi-hop fabric nature, the traffic load gets better balanced. Besides, NoC-based packets switches provide speedup and decouple performance/cost dependency to allow a sub-quadratic growth of the fabric's cost as compared to common single-hop crossbars. Recently, a three-stage Clos switch with Input-Queued NoC-based modules (UDN) on the central stage was proposed in [5]. The architecture presents good scalability and parametrization features. However, on-grid routers of the UDNs middle stage modules, require intrinsic speedup for the whole switch to achieve good performance. With output queuing, bandwidth of the UDNs internal interconnects is increased allowing multiple packets to be forwarded at the same time to the same output port where they get queued for transmission on links of the subsequent node. OQ mini-routers (MRs) have several advantages over IQ routers. Mainly, packets in an OQ architecture are delayed by fixed amount unlike IQ routers where contention for links causes delay variations. In this work we propose a three-stage Clos packet switch with OQ UDNs, to provide statistical performance guarantees. We assume that today's technological advances in the field of memory design and synthesis allow the integration of OQ-UDN modules which links run at speedup of 3 for reasonable costs. In what follows of this work, we evaluate the switch performance by simulations, we suggest an analytical model of the switch throughput and we approximate the blocking probability of the switching architecture.

Many works propose modeling 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 [6]. In a different approach, authors of [7] introduced a low complexity analytic approach for the mean analysis of some performance metrics

[1]Scalability mainly reports to the port count and the amount of traffic load.

of NoCs. In 2010, Suboh et $al.$ proposed a Network Calculus-based methodology to evaluate the latency, throughput and cost metrics of a NoC-based architecture [8]. In 2012, Fischer and Fettweis presented accurate service estimation model for Round Robin arbiters used in Input-Queued NoC fabrics. The approach is interesting as it takes into account contention of multiple concurrent inputs and characteristics of a RR arbitration [9]. In a similar way in [10], authors considered the flow-control feedback probability between adjacent routers of a NoC to evaluate the performance of the network. Another approach based on G/G/1 queues and priority queues is presented in [11] to estimate the latency in the network. In 2015, Karadeniz et $al.$ presented a low-complexity model for single-stage switch based on Network-on-Chip and OQ routers [4]. This paper suggests analytical models for the throughput and blocking probability of the OQ Clos-UDN switch.

III. Throughput analysis of the OQ Clos-UDN switch

In essence, the design of the switch lifts the performance of the network. We discuss a new multistage switching architecture where we use Output-Queued UDN modules. Our design approach relies on a three-stage Clos macro architecture for its reliability and non-blockingness feature. We alter the classical crossbar/memory modules in the central stage of the Clos-network to plug OQ-UDN modules for additional scalability[2].

A. Terminology of the switch architecture

The following is a description of the terminology that we will use throughout this paper. As represented in Fig. 1, the first

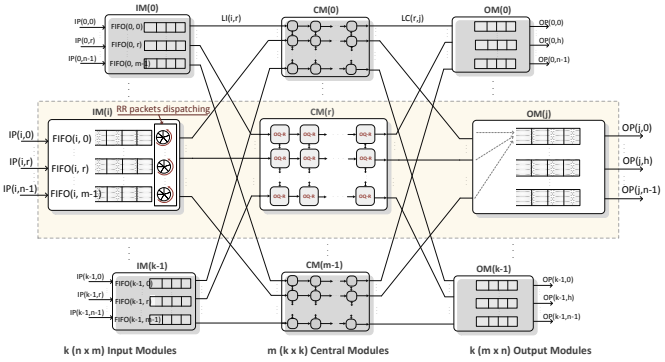


Fig. 1: $(N \times N)$ three-stage OQ Clos-UDN packet-switch architecture

stage of the OQ Clos-UDN comprises k Input Modules (IMs), each of which is of size $(n \times m)$. The second stage is made of m output queued UDN fabric modules, each of dimension $(k \times M)$[3]. The third stage consists of k Output Modules (OMs), each of which has $(m \times n)$ dimension. Although it can be general[4], the proposed OQ Clos-UDN architecture has an expansion factor $\frac{m}{n} = 1$, making it a $Benes$ lowest-cost practical non-blocking fabric. An IM(i) has m FIFOs each of which is associated to one of the m output links denoted as LI(i, r). An LI(i, r) is related to a Central Module (CM(r)). Because $m = n$, each FIFO(i, r) of an input module, IM(i), is associated to one input port. It can receive at most one packet and send at most one packet to one central module at every time slot. A CM(r) has k output links, each of which is denoted as LC(r, j) and is connected to OM(j). OM(j) has n Output Ports, each of which is denoted OP(j, h) and to which is associated an output buffer. An output buffer can receive at most m packets and forward one packet to the output line at every time slot. A CM is defined by the 2-tuple (k, M) where k is the number of I/O ports and M is the depth of the mesh (i.e. the number of pipeline stages). Each mini router has two or three I/Os (referred to as degree of a router) depending on its position on the grid. We use a deadlock-free NoC routing algorithm ($'Modulo\ XY'$) and a credit-based flow control mechanism to avoid elastic buffers.

The terms packet and cell will be used interchangeably in the paper. For simplicity, we consider that packets are of fixed-size with relative routing information stored to their headers. We use the store-and-forward switching mode to transfer traffic across the NoC modules. Several routes exist between blocks of the different stages of the Clos-network. We use a RR selection to depict routes between any element in the input stage and the central stage of the network. We consider the following hypothesis: On each input of the first stage, cells are generated according to independent *Bernoulli* process. We assume that the selection of the LI links is equidistributed and that the equidistribution of paths holds for all stages of the Clos-network [12]. 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 performance of the OQ Clos-UDN switch is very large. However, we focus on the $(k \times M)$ OQ-UDN modules in the middle stage of the Clos switch as the first and last modules' blcks passively forward packets to and from the CMs.

B. Inside the OQ-UDN: Modeling the output-queues

The size of the on-chip queues of MRs impacts not only the small routers performance, but also the silicon area of the design. It is common practice to study the finite capacity queues using Markov chains. In case of the OQ-UDN, all output-queues are of fixed size B. Fig. 2 is a simplified view of one MR. Output buffers work simultaneously. They serve as FIFOs and every output has n' input ports serving it ($n' = 2\ or\ 3$ depending on the MR coordinates in the mesh). Before proceeding, we suppose that packets arrival is a *Bernoulli* process and that P_{arr} is the probability that a packet arrives to an output of the MR. We assume that outputs have deterministic service rate λ_{out} that is the same as the

[2]The NoC design makes the proposed switching architecture easily expandable unlike the conventional designs that rely on crossbars and memory banks. As our simulations shall demonstrate, varying the NoC-based modules' parameters contributes to better performance.

[3]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 $M \leq k$.

[4]The multistage switch can of course be of any size, where $m \geq n$. This would simply require packets insertion policy in the FIFOs should we need to maintain low-bandwidth FIFOs. We consider this to be out of the scope of the current work.

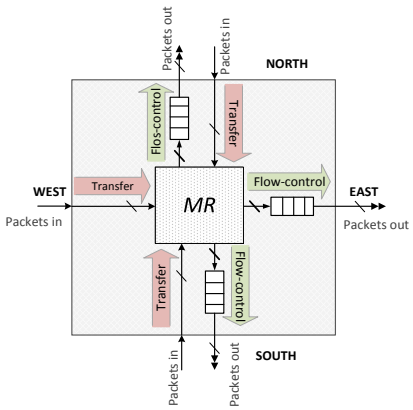


Fig. 2: Block diagram for a mini-router of the Output Queued UDN switching module

probability of packets departure from the output queue P_{dep}. The arrival times and service times are independent and can happen at the same time step. We propose an M/D/1/B queue model to represent an output queue in a MR. Fig. 3 shows the state transition diagram for a single discrete-time M/D/1/B queue where the transition probabilities of a packet moving from one state to another are obtained by considering the ways in which a cell can move between the two states and the probabilities for movements. The state transition in the output queues of the MR mainly consists of two phases: First, check the availability of the buffer space at the subsequent queue and second, move packets forward by one NoC stage. For the M/D/1/B queue, changes in the queue size occur by at most one per time step [6]. We note $b = 1 - p_{arr}$. The probability that a packet remains in the output queue is given by $d = 1-\lambda_{out}$. To describe the state transition diagram for the output queue we define the following intermediate variables:

- α: The probability that a packet arrives to the output buffer but it 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 the output queue is shown in the following Fig. 3. The transition matrix of the output

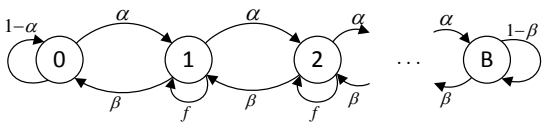


Fig. 3: State transition diagram for a MR output buffer of size B modeled as an M/D/1/B queue

queue is another way to represent information about its state variation. It can be written as:

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

where:

$$\alpha = p_{arr}\ d = p_{arr}\ (1 - \lambda_{out}) \tag{1}$$

$$\beta = (1 - p_{arr})\ \lambda_{out} \tag{2}$$

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

and $\alpha_0 = 1 - \alpha$ and $\beta_0 = 1 - \beta$. We define the state vector S where every component s_i indicates the probability of finding the queuing system in state s_i at that time step [13] and where the first element s_0 reflects the probability that the queue is empty while s_B is the probability that the queue is full.

$$S = [s_0 \quad s_1 \quad s_2 \quad \dots \quad s_B]^t$$

In order to compete the steady state vector elements, we write the equilibrium condition for the output buffer $PS = S$. This 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) by induction and we conclude a generic form of s_i as presented in (5).

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

The OQ state of occupancy can be one among the s_i states at a given time step which means that $\sum_{i=0}^{B} s_i = 1$. We deduce 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:

$$\tau = \frac{\alpha}{\beta} = \frac{P_{arr}(1 - \lambda_{out})}{\lambda_{out}(1 - P_{arr})} \tag{7}$$

Finally, we express the throughput of a single M/D/1/B queue [14].

$$th_{M/D/1/B} = P_{arr}\ \lambda_{out}\ s_0 + \sum_{i=1}^{B} \lambda_{out}\ s_i \tag{8}$$

C. Characterization of the throughput of Clos-UDN switch

The throughput of the switch is the rate of packets delivered to their ultimate destinations. At low traffic loads, the delivery rate is equal to the packets arrival rate while it saturates with the increasing load [13]. 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). The current switching architecture can be described as a nested network in which 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 simulated switch performance, we consider that buffers of the OMs have infinite capacity, which means that analyzing the throughput of the OQ Clos-UDN switch can be reduced to evaluating the packets delivery rate in the OQ-UDN modules.

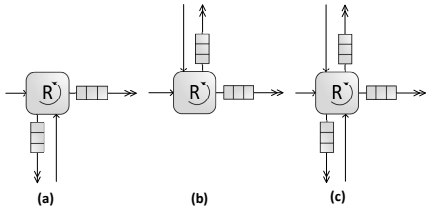


Fig. 4: Types of MRs in an OQ-UDN switch based on the degree of routers

The OQ-UDN switching fabric consists of three different types of MRs based on the degree of the routers as Fig. 4 shows. Overall there are $2M$ MRs of degree 2 and $M(k-2)$ routers of degree 3. As we mentioned earlier, the central modules are supposed to work independently. Hence, to characterize the total throughput of the multistage switch, we analyze the average number of packets that exit one central block. At this level, we assume that packet arrivals to a given node on the NoC grid (MR) and the departure process are independent of each other which means that the average throughput of a single OQ-UDN module can be seen as the summed contributions of the last column's MRs. It can be described using the following equation:

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

Where th_{deg_2} and th_{deg_3} are respectively, the throughput of MRs of degree 2 and degree 3. Since all I/O(s) of a MR work independently and simultaneously to contribute to the average throughput of the node, we can express th_{deg_2} and th_{deg_3} using (11) as a summation of as many M/D/1/B queues as the degree of the MR.

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

Output buffers of the MRs are limited in 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. Generally, in complex networks, deriving the end-to-end blocking probability of a path 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. However, there are constraints that introduce dependencies. 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 queues at the output ports of the on-grid routers. It is mainly the availability of the buffering resource in outputs of the downstream MRs that result in dependencies and adds complexity to what could be a simple estimation of the end-to-end blocking probability. In this section we show that it is possible to estimate an upper-bound on the probability that any path in the OQ-UDN switch is blocked.

A. Feedback flow-control probability in a single output queue

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

$$P_{ctr} = s_B = \tau^B\ \frac{1-\tau}{1-\tau^{B+1}} \tag{10}$$

B. Upper bound on the end-to-end blocking probability in the central modules

We consider the following model. We call node, any MR on the grid of a central module. A path (also called route) is a set of successive links and output buffers that a packet has to cross from a source node to a destination node. 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 intermediate links. 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.

$$B(r) \leq \sum_{j=1}^{\chi} P_{ctr}(j) \tag{11}$$

Where χ is the number of buffers that belong to route r.

$Proof$: We prove (11) in two stages: First, we approach the availability probability in the central modules of the OQ Clos-UDN switch, then we infer an upper bound for the blocking probability. We define $\mathscr{R}_{r_j} \subseteq \mathscr{R}$, $r_j = 1, \ldots, m$, 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 of the switch, an output buffer is used by at least one path in $\mathscr{R}$. For our analysis, we denote ν_r, the offered end-to-end traffic load for a given route $r \in \mathscr{R}$.

We assume that traffic getting out of the IMs of the OQ Clos-UDN switch is still stationary. Although we consider uniform traffic, the proof stands even for an arbitrary traffic type. At this point, we have no idea about the traffic intensity ρ_j entering an output buffer at a time since it is not an input parameter like ν_r. However, this proportion of traffic is the

superposition of the load carried over the previous stretch of a given path. Clearly, ρ_j depends on the availability of the upstream MRs' output buffers that may in turn depend on other factors.

A path is said to be blocked with a probability $B(r)$, if and only if at least one of the buffers that a packet must go to on its route is blocked. 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 may not be simply independent of other buffers somewhere located in the mesh network which makes the situation delicate to handle mathematically [15]. Therefore, we choose to go for an estimation of an upper bound of the end-to-end blocking probability of a route rather then determining an exact value for $B(r)$.

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 IV-A). We introduce a set of useful terms that will be used to prove (11).

Let $\alpha_r(\rho_j) = \alpha_{r_j}$, be the probability that an output queue is available in route r. We call route $r \in \mathscr{R}$, available with a probability $A(r)$, if the whole set of output queues on the path are simultaneously available. Since output buffers of MRs are not necessarily independent, then the availability of the set of buffers all along a path is not always the 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 $r = r_j$, then $A(r) = (1 - P_{ctr}(j))$. We denote $r^{[j]}$, the j first output 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 arbitrary dependency pattern, the probability that a route $r \in \mathscr{R}$ is available always satisfies the following equation:

$$A(r) = \prod_{j=1}^{\chi} \alpha_{r_j} \left(\sum_{s \in \mathscr{R}_{r_j}} \nu_s \; A(s - r^{[j]}) \right); r \in \mathscr{R} \setminus \{\emptyset\} \quad (12)$$

To prove (12), 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

$$A(r_1) = \alpha_{r_1} \underbrace{\left(\sum_{s \in \mathscr{R}_{r_1}} \nu_s \; A(s - r_1) \right)}_{\psi} \quad (13)$$

The term ψ is the sum of the traffic intensities offered on all routes that contain 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 as the $route{-}carried$ traffic load that ends up at queue r_1 and that we previously denoted as ρ_1. From (13) we can write $A(r_1) = \alpha_{r_1}(\rho_1) = \alpha_{r_1}$. Hence (12) holds for $\chi = 1$.

In the following, we shall prove that the system of equations in (12) is still valid for routes of length $\chi > 1$. We introduce the set of events e_j, $j = 1, \ldots, \chi$ to mark 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 depend on the all previous buffers.

$$\begin{aligned} A(r) &= Pr(e_1) \frac{Pr(e_1 e_2)}{Pr(e_1)} \cdots \frac{Pr(e_1 e_2 \ldots e_\chi)}{Pr(e_1 e_2 \ldots e_{\chi-1})} \\ &= Pr(e_1) \prod_{j=2}^{\chi} Pr(e_k | e_{j-1} \ldots e_1) \end{aligned} \quad (14)$$

Using the initial input traffic load of a 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 (14) shows. In other terms, it depends on $r - \{r_1\} - \{r_{j-1}, \ldots r_1\} = r - r^{[j]}$. Finally, we can write (14) in a different way:

$$Pr(e_j | e_{j-1} \ldots e_1) = \alpha_{r_j} \left(\sum_{s \in \mathscr{R}_{r_j}} \nu_s \; A(s - r^{[j]}) \right) \quad (15)$$

Taking into account that $Pr(e_1) = A(\{r_1\}) = \alpha_{r_1}$ and using the system of equations in (15), we infer (12).

Being a probability, the factor $A(r - r^{[j]}) \leq 1$. Thus removing $A(r - r^{[j]})$ from the right-hand side of (12), should result in the following inequality:

$$A(r) \geq \prod_{j=1}^{\chi} \alpha_{r_j} \left(\sum_{s \in \mathscr{R}_{r_j}} \nu_s \right) \geq \prod_{j=1}^{\chi} \alpha_{r_j}(\breve{\rho_j}) \quad (16)$$

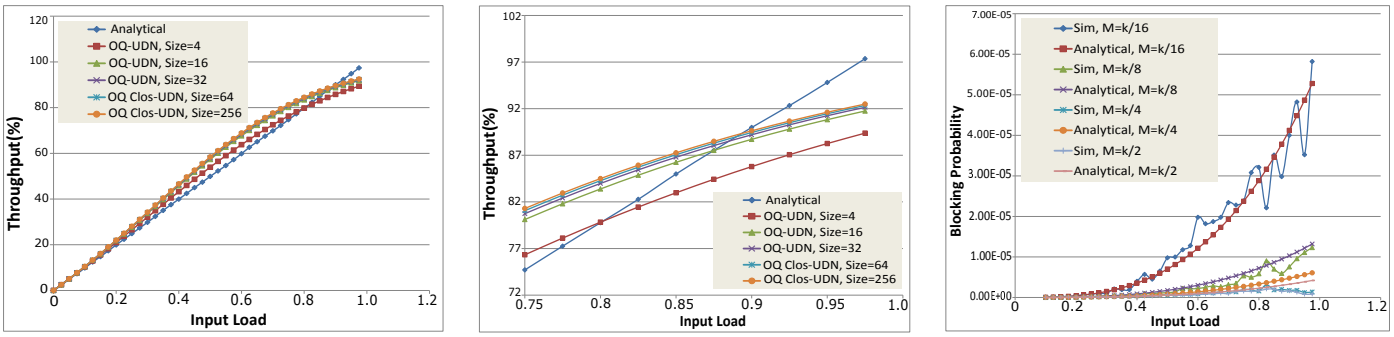
Where $\breve{\rho_j} = \sum_{s \in \mathscr{R}_{r_j}} \nu_s$, is the traffic intensity that results in queue r_j from all routes $r \in \mathscr{R}_{r_j}$. Since in general $1 - \prod_{s=1}^{S} a(j) \leq \sum_{s=1}^{S} (1 - a(i))$, we derive an upper bound on $B(r)$.

$$B(r) \leq 1 - \prod_{j=1}^{\chi} \alpha_{r_j}(\breve{\rho_j}) \leq \sum_{j=1}^{\chi} \left(\underbrace{1 - \alpha_{r_j}(\breve{\rho_j})}_{P_{ctr(j)}} \right) \quad (17)$$

We conclude (11).

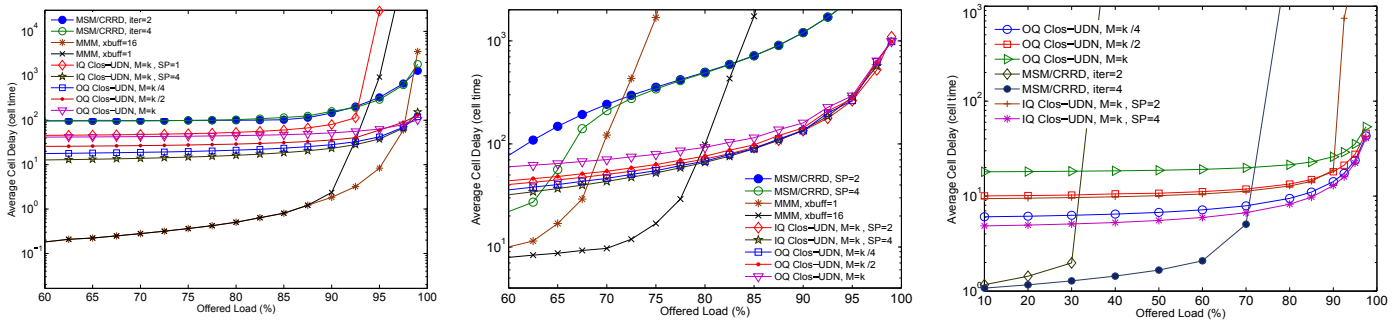
V. Performance evaluation

In the first part of this section, we compare the analytical results to simulation outputs. Next, we present further performance analysis of the switch's performance for different settings. The parameter k denotes the number of I/O ports of the Clos switch's central modules and M is the depth of the 2-D mesh. Buffer Depth BD of on-chip output queues is set to the default value of 3.



(a) Throughput of switches (b) Throughput under heavy loads (c) $B(r)$ in a (64×64) OQ Clos-UDN switch

Fig. 5: Analytical model Versus simulations under $Bernoulli\ i.i.d$ traffic



(a) $Bernoulli\ i.i.d$ traffic, (256×256) switch (b) $Bursty$ uniform traffic, (256×256) switch (c) Hot-spot traffic, (64×64) switch

Fig. 6: Delay performance of MSM, MMM, IQ Clos-UDN and OQ Clos-UDN, $BD = 3$

A. Analysis of the theoretical model

1) Throughput performance of the switch: To demonstrate the accuracy of our approximation, we compare the analytical results to the simulations for variable switch sizes and a minimum MRs' buffers depth, $BD = 3$. Fig. 5(a) shows the variation of percentage of throughput under *Bernoulli i.i.d* arrivals. The proportion of throughput increases linearly when the input load increases because of the number of packets generated. We can see that the values obtained using the analytical model approach those of simulation. Under light loads, simulations perfectly match the analytical model. We note that for a single-stage OQ-UDN, the deviation is about 4.5% under relatively medium input loads. This margin increases when the number of ports gets higher and the traffic load becomes heavier. According to Fig. 5(b), 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, deviation still do no not go beyond 7.98% for the smallest single-stage switch of size (4×4) and 5.2% for a (64×64) multistage OQ Clos-UDN.

2) Blocking in the switch: Given the OQ Clos-UDN switch topology and the assumptions that we made for the analytical analysis, we argue that the central modules are the one and only bottleneck of the design. Hence tracing $B(r)$ of the intermediate stages, reflects the same behavior in the multistage switch operating under uniform traffic. We consider a switch of size (64×64) that can work as a standalone single stage switch, or be plugged into the middle stage of the Clos-network architecture (for larger switch valencies). We set the output queues size to the default value $BD = 3$ and we consider a *Bernoulli* uniform traffic. Note that the proposed mathematical approximation of $B(r)$ approaches the curves' envelopes. Overall, the blocking probability rises exponentially with an increasing input load as Fig. 5(c) shows. For light workload intensities, the network is hardly blocking. Regardless of the value of M, $B(r)$ is less than 10^{-5}. In the worst-case scenario where we set $M = k/16 = 4$, we note an average error that is around 3.9×10^{-6}.

B. Further experiments

We further study the switch performance using an event-driven simulator for various settings. We show that our design is scalable to switch size and robust to traffic variability [16]. In Fig. 6(a), we compare the delay performance of the current switch to Memory-Space-Memory (MSM) with the iterative Concurrent Round Robin Dispatching algorithm (CRRD) [17], Memory-Memory-Memory (MMM) [18] and the IQ Clos-UDN switches [5]. Our proposal outperforms MSM under heavy workloads. It always provide full throughput unlike the IQ Clos-UDN switch that saturates at around 90% if no speedup is used ($SP = 1$). A fully buffered architecture offers lower delays. Yet, large crosspoint buffers are required to achieve full throughput. On the contrary, our switch running with small on-chip buffers ($BD = 3$) and $M = k/4$ (that is only equal to 4 for (256×256) switch ports) ensures almost

constant delay variations and high throughput.

Fig. 6(b) depicts the average delay in the different switching architectures under bursty uniform traffic with bursts worth of 10 packets, each. Our simulations show that heavy bursts arrivals make both MSM and MMM perform poorly. Moreover, MMM do not achieve full throughput even if the middle stage buffers' size is increased to 16 packets, each. The flexibility of Networks-on-Chip, Clos interconnect pattern and the dynamic packets dispatching collaborate to better distribute the load and to conserve high throughput.

We consider a more realistic traffic where the load distribution among the switch inputs vary based on an unbalancing coefficient $\omega \in [0, 1]$ indicates the nature of the traffic. If $\omega = 0.5$, then, the switch deals with hot-spot traffic. Fig. 6(c) shows the delay performance of different switches with variable settings, a variable input load and $\omega = 0.5$. As for uniform arrivals (Fig. 6(a) and Fig. 6(b)), the OQ Clos-UDN outperforms MSM switch with CRRD scheduling. IQ and OQ Clos-UDN switches have comparable delay performance if the architectural parameters are properly set (mainly speedup and M for the input-queued type and M and BD for an OQ-UDN module). Although both designs are highly customizable, an input-queued Clos-UDN with no speedup and $M = k = 8$ do not provide full throughput.

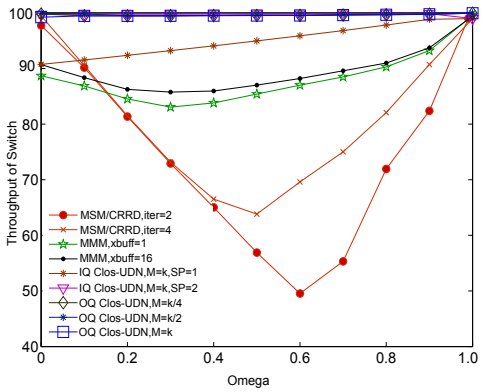


Fig. 7: Throughput stability of different switching architectures for (64×64) switch and variable ω.

As shown in Fig. 7, IQ Clos-UDN with full depth ($M = k = 8$) and $SP = 1$ delivers 90% throughput. MMM provides better throughput than MSM which throughput saturates at 60% if CRRD is iterated four times ($iter = 4$) and $\omega = 0.5$. OQ Clos-UDN gives full throughput under the whole range of ω even when minimum settings are used ($BD = 3$ and $M = k/4 = 2$, for a (64×64) switch).

VI. Conclusion

We propose a highly-scalable multistage switching architecture for DCN environments based on NoCs and Output Queuing. We study the performance of the switch by simulations and we propose an analytical approximation for the theoretical throughput that the switch achieves. Although we dropped some architectural considerations and we imposed independence assumptions to simplify the model, the experimental results still show that the average deviation of the model is about 7.9% for single stage OQ-UDN (size < 64) and around 5.2% for a Clos-UDN switch (size ≥ 64). Furthermore, we estimated an upper bound on the end-to-end blocking probability in the central modules of the OQ Clos-UDN switch taking into consideration inter-dependencies that the architectural design imposes. Results show that for *Bernoulli i.i.d* packet arrivals, the model approaches simulation results.

Acknowledgment

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

References

[1] F. M. Chiussi, J. G. Kneuer, and V. P. Kumar, "Low-cost scalable Switching Solutions for Broadband Networking: the ATLANTA architecture and chipset," *IEEE*, vol. 35, no. 12, pp. 44–53, 1997.

[2] K. Goossens, L. Mhamdi, and I. V. Senin, "Internet-router buffered crossbars based on Networks-on-Chip," in *DSD, 12th Euromicro Conf.* IEEE, 2009, pp. 365–374.

[3] C. Clos, "A Study of Non-Blocking Switching Networks," *Bell System Technical Journal*, vol. 32, no. 2, pp. 406–424, 1953.

[4] 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, conf.* IEEE, 2015, pp. 308–314.

[5] F. Hassen and L. Mhamdi, "A Multi-Stage Packet-Switch Based on NoC Fabrics for data center networks," in *IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2015, pp. 1–6.

[6] 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.

[7] U. Y. Ogras and R. Marculescu, "Analytical router Modeling for Network-on-Chip Performance Analysis," in *DATE conf.* IEEE, 2007, pp. 1–6.

[8] S. Suboh, M. Bakhouya, J. Gaber, and T. El-Ghazawi, "Analytical modeling and evaluation of Network-on-Cchip architectures," in *HPCS, conf.* IEEE, 2010, pp. 615–622.

[9] E. Fischer and G. P. Fettweis, "An Accurate and Scalable Analytic Model for Round-Robin arbitration in Network-on-Chip," in *NoCS, Seventh IEEE/ACM*. IEEE, 2013, pp. 1–8.

[10] 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.

[11] A. E. Kiasari, Z. Lu, and A. Jantsch, "An Analytical Latency Model for Networks-on-Chip," *VLSI Systems, IEEE Transactions on*, vol. 21, no. 1, pp. 113–123, 2013.

[12] 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.

[13] F. Gebali, *Computer communication networks: Analysis and design*. Northstar Digital Design, Incorporated, 2005.

[14] ——, *Analysis of Computer Networks*. Springer, 2015.

[15] M. E. Ekpenyong and J. Isabona, "Performance modeling of blocking probability in multihop wireless networks," *Journal of Applied Science & Engineering Technology*, vol. 4, 2011.

[16] F. Hassen and L. Mhamdi, "A Scalable Packet-Switch Based on Output-Queued NoCs for Data Centre Networks," in *ICC, conf.* IEEE, 2016, p. in press.

[17] E. Oki, Z. Jing, R. Rojas-Cessa, and H. J. Chao, "Concurrent round-robin-based dispatching schemes for Clos-network switches," *IEEE/ACM*, vol. 10, no. 6, pp. 830–844, 2002.

[18] Z. Dong and R. Rojas-Cessa, "Non-blocking Memory-Memory-Memory Clos-network packet switch," in *Sarnoff Symposium, 34th*. IEEE, 2011, pp. 1–5.

[19] H. Yoon, K. Y. Lee, and M. T. Liu, "Performance Analysis of Multibuffered Packet-Switching Networks in Multiprocessor Systems," *Computers, IEEE Transactions on*, vol. 39, no. 3, pp. 319–327, 1990.