Proceedings Paper:
Hassen, F and Mhamdi, L (2019) A New Paradigm to Build Scalable Packet-Switches for Data Center Networks. In: Proceedings of the 2018 IEEE Global Communications Conference (GLOBECOM). GLOBECOM 2018: IEEE Global Communications Conference, 09-13 Dec 2018, Abu Dhabi, United Arab Emirates. IEEE . ISBN 978-1-5386-4727-1

https://doi.org/10.1109/GLOCOM.2018.8647295

A New Paradigm to build Scalable Packet-Switches for Data Center Networks

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

***Abstract*—This paper presents the design, implementation, and evaluation of a class of packet-switching fabric architectures. Based on the well-investigated three-stage Clos-network, we propose a variety of packet-switches that are constructed by adding the most beneficial Network-on-Chip (NoC) paradigm which offers many distinct and practical advantages. Compared to the conventional crossbar switches, the NoC-based architectures provide better path-diversity, simple packet scheduling and speedup. A gradual design method is adopted to enhance the performance of the NoC switch, and several related issues such as the congestion avoidance, micro level load-balancing, and cost-effectiveness are addressed. The NoC switches exhibit a high scalability potential in – both – the port count and traffic volume, making them a good candidate for the next-generation Data Center Networks.**

***Index Terms*—Data Center Networks switching fabric, Clos-network, Network-on-Chip, Packet routing, Packet scheduling**

I. Introduction

The performance of a data center is not only attribute of its computing features. It relies on the capabilities and the performance of the interconnection network infrastructure. The DCN deals with perpetually floating amounts of workloads exchanged among the parallel server machines. It must support the data growth, and amend the bandwidth balancing among the compute nodes. Hence, using the inappropriate interconnection infrastructure can detriment the global DCN performance. Switches and routers are key modules in the DCN, as they are the first elements to process data. They are build in redundancy in the DCN to manage a reliable inter-rack and inter-cluster communications.

Single-stage switches have been long appreciated for their simplicity. However, they show a restricted scalability and a bounded performance spectrum which makes them unsuitable for the ever-growing DCNs. The multistage interconnects have been adopted by both academia and the industry to support large-capacity switching architectures. The approach helps increasing the port count in a cost-effective manner by cascading small-radix Switching Elements (SEs). The three-stage Clos-network [1] has been the one to successfully make the way to an array of commercial products [2], [3]. The Clos-network switches can be classified with reference to the packet buffering strategy. This gives rise to all sorts of switching fabrics, ranging from the Space-Space-Space (S^3) architecture to the Memory-Memory-Memory (MMM) switch [4]–[7]. Other combinations have also attracted attention [8]–[10]. Particularly the Memory-Space-Memory (MSM) [8] switch that has been investigated with different scheduling algorithms. The MSM switch is a popular solution that has been introduced in hopes of achieving a good compromise between the complexity and performance. Although their features, conventional three-stage Clos switches suffer some limitations. For instance, they all implement an excessive number of separate queues in the input modules to resolve the Head-of-Line (HoL) blocking problem [4], [8], [11]. This implies prohibitive costs and complex packet dispatching algorithms. The semi-buffered switches – like the MSM – require a centralized scheduler to resolve the input and output contention, and to perform the global port matching [8]. The MMM switch has been a straightforward way to scale a single-stage buffered switch. It employs distributed buffers [7] to absorb contention, and to dismiss the need for a central arbitration. Yet, the design cost of the fully-buffered architecture is still significant. In an attempt to build large-capacity packet-switches that can easily scale in size and traffic loads, we suggest a class of multistage NoC-based switches. At first place, we alter the middle-stage of the Clos-network, and we plug Uni-Directional NoC (UDN) [12] modules instead of the classic crossbars/memory blocks. Besides, we extend the switching facility by interconnecting the UDN central modules. This improves the path-diversity, and allows a better load distribution across the switching architecture. Ultimately, we describe a three-stage Clos switch with Multi-Directional NoC (MDN) [13] fabric, that scales better and faster than the previous alternatives.

Contributions and paper content

Motivated by the shortcomings of the conventional Clos-network switches, we suggest a class of high-performance NoC-based packet-switching architectures suitable for the DCNs. The primary contributions of this paper span the following:

- We propose a three-stage packet-switch with UDN fabric. The Clos-UDN switch obviates the need for a complex and costly input queueing structure. It also avoids the need for the complex and slow centralised scheduling algorithms, and offers a high degree of modularity, path-diversity and speedup.
- We suggest a simple Round-Robin (RR) dispatching mechanism to send packets from the input First-In-First-Out (FIFO) queues to the central NoC fabric.

– We implement a deadlock-free distributed routing algorithm to forward packets throughout the UDN modules.
– To further improve the path-diversity, we propose the Congestion-Aware (CA) Clos-UDN switch, in which we connect the middle-stage UDNs using interleaved links.
– We implement a congestion-aware routing algorithm that proactively distributes the traffic load among the NoC blocks, and prevents the onset of congestion.
– We introduce the wrapped-around Clos-MDN switch that employs compact NoC modules, and scales faster in size and traffic variations.
– We implement a buffered flow-control, and Virtual Channels (VCs) to send the packet flows in all directions with no deadlocks.
– The Clos-MDN switch works tightly with an appropriate scheduling algorithm to provide high throughput while involving less hardware and less complex scheduling.

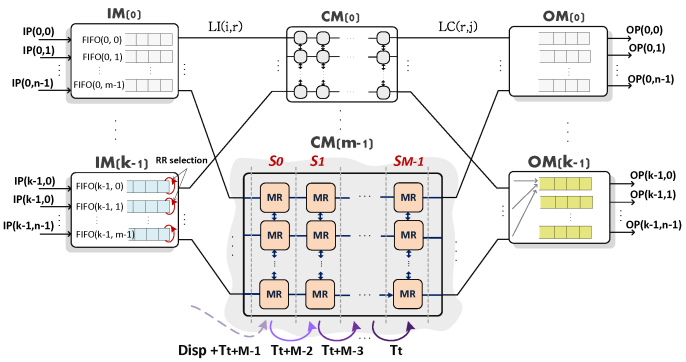


Fig. 1: The generic layout of a three-stage Clos-network switch with NoC fabric.

The rest of the paper is structured as follows. In Section II, we highlight the terminology of the switching architectures. In Section III, we describe the packet scheduling and routing process through the NoC fabrics. In Section IV we discuss the performance of the proposed switching architectures under a range of traffic scenarios, and we evaluate their scalability potential. In section V, we overview the related work, and we conclude the paper in section VI.

II. Clos-packet switches with NoC fabric

In this section we give the terminology of the high-level switching architectures. Besides, we present the Clos-UDN, CA Clos-UDN, and the Clos-MDN switches – respectively. Next, we overview the key features of the UDN and MDN fabrics. The last sub-section describes the packet scheduling mechanism in the different switches.

A. High-level architecture and packet buffers

We consider a three-stage Clos-network macro architecture and a NoC micro architecture to device the different switches as shown in Fig. 1. Although it can be general[1], all proposed architectures have an expansion factor $\frac{m}{n} = 1$, making them *Benes* lowest-cost practical non-blocking fabrics.

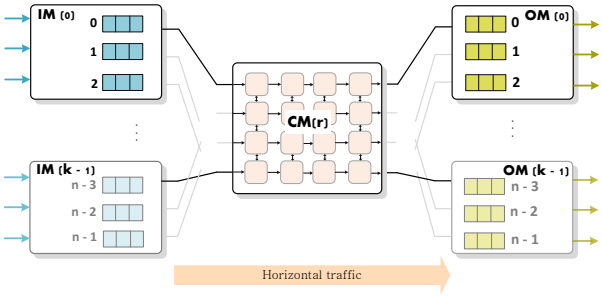


Fig. 2: The Clos-UDN switching architecture.

Both the Clos-UDN, and CA Clos-UDN switches presented in Fig. 2 and Fig. 3 – respectively – use uni-directional NoC modules. They share the same general layout. At the first stage, we use k Input Modules (IMs), each of which is of size $(n \times m)$. An IM(i) has m FIFOs[3], each of which is associated to one of the m output links that we denote as LIs. The link LI(i, r) connects IM(i) to CM(r). It can receive at most one packet, and sends at most one packet to one CM at every time slot. The middle stage of the switch consists of m UDN modules of dimension $(k \times M)$ each[4]. The CM(r) has k output links, that we denote as LCs, and that serve to connect the central module to the Output Modules (OMs). The last stage has k OMs, each of which is of size $(m \times n)$. An OM(j) has n Output Ports (OPs) that we denote OP(j, h), and for which is associated an output buffer. Each output buffer can receive at most m packets, and forwards one packet to the output line card at every time slot.

We add extra links to the baseline Clos-UDN architecture, and we interconnect the middle-stage blocks of the CA Clos-UDN switch, as Fig. 3 and Algorithm 1 show. This extends the advantage of the NoC geometry, boosts the path-diversity, and makes the multistage switch architecture a wrapped-around network. For the Clos-MDN switch, we further exploit the benefits of NoCs to propose a high-radix switching architecture with compact fabric. As depicted in Fig. 4, the generic layout consists of $2k$ Input/Output Modules (IOMs), each of which is of size $(n \times n)$[5], and m central MDNs, each of which

[1]The multistage switch can be of any size, where $m \geq n$. In this case, we would simply require a packet[2] 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.

[3]Because $m = n$, each FIFO(i, r) of an input module, IM(i), is associated to one input port. Each IM needs to maintain only m input queues, and each input port of an IM can send packets to only one FIFO queue per time-slot. Consequently, running the input FIFOs at only twice the line rate is sufficient.

[4]Unlike conventional Clos networks, the central modules of the Clos-UDN and CA Clos-UDN switches can be of size $(k \times M)$ crosspoints, where $M \leq k$ is the number of pipeline stages, generally quoted as the NoC depth.

[5]For an arbitrary non-blocking Clos-network, the number of outlets in any of the first-stage modules – m – can differ from the number of its inlets – n. However, in this paper we consider the simple design case of *Bene's* network for which $n = m$

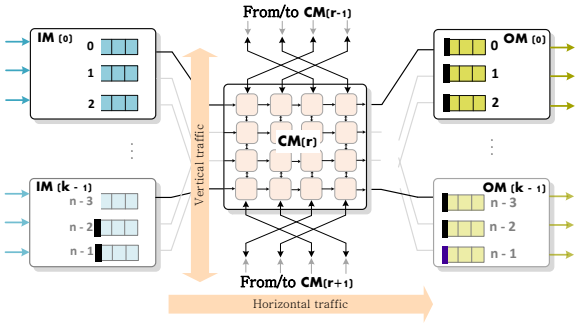


Fig. 3: The Congestion-Aware Clos-MDN switching architecture.

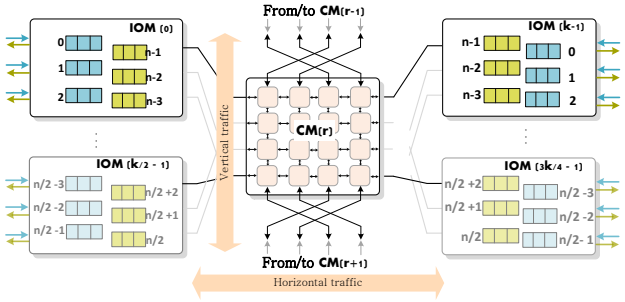


Fig. 4: The Clos-MDN switching architecture.

Algorithm 1 : Inter-CM interleaved connections

1: **for** $r \in \{1, \ldots, m-1\}$ **do**
2: $r' \leftarrow ((r+1)\ mod\ m)$ and $r'' \leftarrow ((r-1)\ mod\ m)$
3: **for** $i \in \{1, \ldots, k-1\}$ **do**
4: $j \leftarrow ((\frac{k}{2}+i)\ mod\ k)$
5: $\text{MR}^r(k-1, i)$ connects to $\text{MR}^{r'}(0, j)$
6: $\text{MR}^r(0, i)$ connects to $\text{MR}^{r''}(k-1, j)$
7: **end for**
8: **end for**

is of size $(k \times k)$. The LI and LC links still work at the same rate as the external lines. The essence of the Clos-MDN is in its capacity to scale faster in size while compact MDN modules are put in use. Typical flattened[6] three-stage Clos switches tend to move packets in one direction – from the IMs to the corresponding OMs. Yet, the Clos-MDN switch allows the traffic circulation in opposite directions across the network.

In the next sub-section, we overview the UDN and MDN fabrics, and we specify the key differences between both of them.

B. The UDN and MDN fabric

A UDN [12] module is a 2-D mesh network with a total of $(k \cdot M)$ small on-chip packet-switched routers. The Mini-Routers (MRs) are fitted with input FIFO queues of limited capacity, that we call *buffer depth* – buff. To preserve their integrity, packets are fully received and stored in one of the mini-router's buffers before going to the next hop. It is worth mentioning that a typical UDN module[7] has its input ports at the *Western* side of the mesh layout, while the output ports are placed at the *Eastern* side. In the contrary, an MDN switching element has the set of its I/O ports speared over the perimeter of the mesh. Thus, it allows traffic to flow in all directions. Originally, the MDN [13] fabric was introduced as an extension of the UDN proposal. It explores the assets of the NoC paradigm, and efficiently builds a compact switching fabric. It also alleviates deadlocks by means of an adequate routing algorithm and two virtual channels. The VCs merely isolate the traffic flows moving in opposite directions. The distribution of the buffering space amongst the VCs can be symmetrical or asymmetrical depending on the design choice[8]. Fig. 5 is a view of the high-level diagram of the on-chip routers used in the UDN and MDN fabrics – respectively. Both UDN and MDN fabrics are used along with a credit-based flow-control mechanism to manage the downstream buffers reservation. In the following parts, we will use the term speedup to refer to the speed ratio at which the on-chip links of the NoC fabric can run with respect to the LI/LC links speed. Saying that the NoC switching elements run at a speedup SP, is equivalent to the on-chip routers removing up to SP packets from one input, and sending up to SP packets to one output per time-slot.

In the following section, we present the packet scheduling and routing algorithms used for the NoC-based switches.

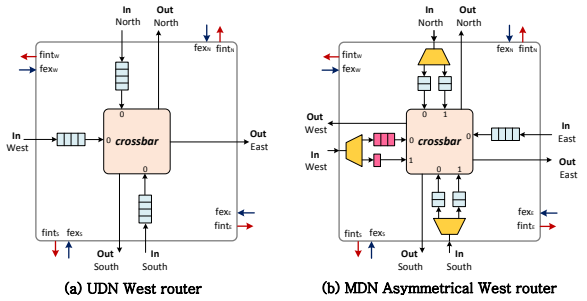


Fig. 5: High-level diagram of the UDN and MDN *Western* mini-routers.

III. Packet Scheduling

In contrast with the single-stage switches, multistage switching architectures require a dispatching process to manage

[6]Other 3-D architectures such as the hypercube and layered switches, would obviously allow traffic circulation in a cubic way or in-between the layers.

[7]As will be described and used in this paper.

[8]Consider two virtual channels, VC1 and VC2. A symmetrical port buffer distribution would allocate the same buffer space to both channels. An asymmetrical distribution makes the *Western* mini-routers reserve 2/3 of the buffer depth for VC1 and 1/3 for VC2. Correspondingly, the *Eastern* mini-routers use 1/3 of the port buffering space for VC1 and 2/3 for VC2. The central mini-routers allocate equal buffering space to VC1 and VC2).

the packet forwarding from the IMs to the CMs. The complexity of the dispatching process correlates with the nature of the middle-stage fabric. It can be highly complex in a bufferless switch (*e.g.*, MSM switch), and relatively simple in a fully buffered architecture (*e.g.*, MMM switch). The different switches described in this paper fall into the category of buffered architectures. They greatly simplify the process of packet dispatching and scheduling. In what follows, we describe a dynamic and a static packet dispatching processes.

A. The process of packet dispatching

The Clos-UDN switch maintains m RR input schedulers in every IM – one per FIFO queue. The input schedulers are initialized to different positions, and they keep updating their selection pointers to one position at the end of every time-slot. This guarantees that all pointers are always desynchronized and that no conflict in the LI links selection happens. At the start of every time-slot, a scheduler selects an LI link among m links. Then, the HoL packet is transferred from the input FIFO to the corresponding CM. A packet is accepted to the CM module if the left-most NoC mini-router still has room in its left buffer. Once at the UDN fabric, the "$Modulo\ XY$" routing algorithm takes over, and routes the packet to its outgoing LC link. The dynamic packet dispatching distributes packets of the same flow to different UDNs, where they are likely to experience variable queueing delays and get to their OPs out-of-order. We suggest the static dispatching scheme whereby the LI links are statically configured. This means that an input FIFO at the first stage persistently sends packets to the same CM. The static dispatching used with the "$Modulo\ XY$" on-chip routing guarantees in-order packet delivery in the baseline Clos-UDN switch [14].

B. On-chip packet routing

In the rest of the paper, we consider a static packet dispatching scheme for all of the switches, in which case the performance disparity is mainly attribute of the NoC fabric.

The packet routing process can be described as follows. Once in the UDN fabric, packets are routed until the corresponding LC links using the "$Modulo\ XY$" algorithm [12]. Unlike the classic "XY" routing [15] that moves packets in one direction, and then in the other direction[9], the "$Modulo\ XY$" algorithm introduces an extra turn in an intermediate pipeline stage before the last one to better balance the load in the mesh. Packets are routed in a very similar way across the MDN fabric. Upon arriving to the MDN modules, packets are assigned the correct VC depending on their destination IOMs. Next, the "$Modulo\ XY$" routing takes over to move forward the traffic flows.

[9]We assume a packet is at the mini-router of position (x, y), and with a destination node located at position (x', y'). In the general case, the classic "XY" algorithm moves the packet *East* if $(x' > x)$, and *West* if $(x' < x)$. When reaching the line where the corresponding destination node is located, a packet moves *North* if $(y' > y)$, and *South* if $(y' < y)$. The "XY" in its basic form poorly distributes the traffic across the network as it tucks the biggest load in the middle of the NoC.

The traffic load moves horizontally in one direction in the Clos-UDN switch. However, the CA Clos-UDN and Clos-MDN switches have – both – a wrapped-around architectures, where packets are allowed to travel vertically between the CMs. Both proposals are motivated by the need for a micro load-balancing [16] to be processed at the switch-level, rather than the macro techniques performed at the DC network level [17], [18]. In the following sub-section, we present the *repellent routing* algorithm that we use to manage the inter-CM traffic.

C. Proactive congestion management

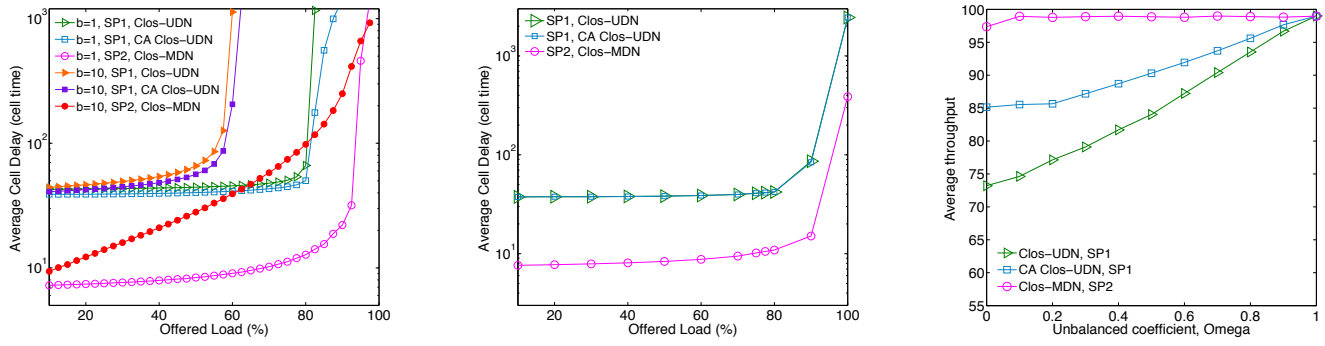
Load-balancing in DCNs has been long relying on centralized controllers or end-hosts [17], [18] to collect the global network information, and to distribute the traffic load. However, a good proportion of the congestion spikes encountered in the DCN are short-lived events [19]. Motivated by this observation, recent works dived into designing new switches with micro load-balancing capabilities [16]. In this paper, we piggyback on these design trends. To improve the load distribution in the switch, we make the central-stage modules of the CA Clos-UDN and the Clos-MDN switches capable of sharing traffic via intermediate links. We opt for an interleaved configuration to ensure that sending packets from their original congested CMs to a neighbouring module, does not affect the overall latency[10]. Our ultimate goal is to maximize the switch throughput under coarse traffic without affecting the delay performance. Therefore, we adopt a metric that is suitable for the routing scheme to correlate well with the global Clos-network congestion status while being inexpensive to compute. We consider the Regional Congestion Awareness (RCA) [20] to evaluate and to propagate the congestion information proactively, across the central module of index r and its direct neighbours (blocs of indexes $((r-1)\ mod\ m)$ and $((r+1)\ mod\ m)$). The congestion metric weights both distance (hops count until the exit port) and buffers occupancy to make sure that the traffic is adaptively transferred through minimal paths, and that the average packet delay is little affected by the inter-module routing decision. Algorithm 2 gives details of the repellent routing.

In the next section, we assess the performance of the different switching architectures under a wide range of synthetic traffic.

IV. Performance Analysis

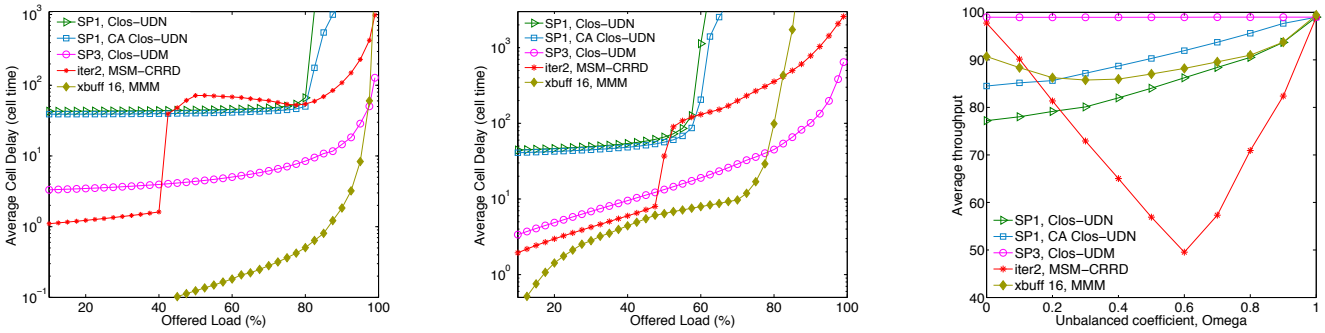
Our models are built on top of an event-driven simulator written in C language. In what follows, we assume that the capacity of the on-chip buffers (buff) is 4 packets each – unless it is otherwise stated. In all plots, we consider that the Clos-UDN and the CA Clos-UDN switches use NoC modules with full mesh depth (*i.e.*, $k = M$). Since the Clos-MDN switch

[10]In the worst case scenario, a packet will do the same number of hops in the neighbour CM as it would have in its non-congested CM for two reasons: First, the inter-module routing algorithm considers the distance metric, and second packets are minimally routed within a single CM.



(a) Uniform traffic (b) *Hot-spot* traffic (c) Throughput stability

Fig. 6: The performance of 256-ports Clos switches with NoC fabrics.



(a) Uniform *Bernoulli i.i.d* traffic (b) Uniform *Bursty* traffic (c) Throughput stability

Fig. 7: The performance of 256-ports MSM, MMM, and UDN/MDN Clos switches.

Algorithm 2 : The Repellent routing

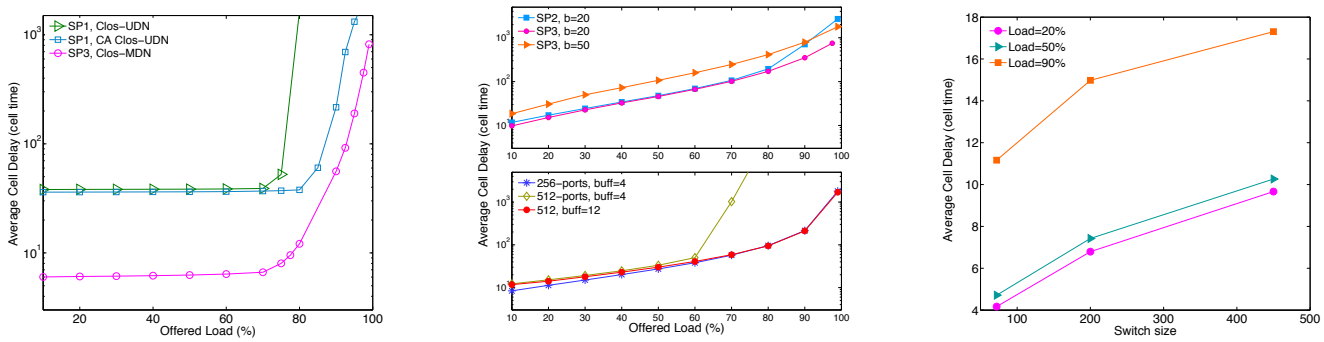
if ($pck\_repulsed = TRUE$) **then**
  $port \leftarrow routing\_direction$
else
  fct : choose closest CM
  if (local routing quadrant is less congested) **then**
    "$Modulo\ XY$",
    $pck\_repulsed \leftarrow FALSE$
  else
    if (chosen CM is *UP*) **then**
      $routing\_direction \leftarrow North$,
      $port \leftarrow North$,
      $pck\_repulsed \leftarrow TRUE$, //Override bit
    else
      $routing\_direction \leftarrow South$,
      $port \leftarrow South$,
      $pck\_repulsed \leftarrow TRUE$
    end if
  end if
end if

employs NoC blocks with less[11] pipeline stages, we consider running the fabric at a $SP \geq 2$ to compensate for the design mismatch. We argue that it is possible to trade area by speedup in the Clos-MDN switch, since it is not expensive to run short on-chip links a bit faster using the current technology.

[11]In comparison with the UDN-based proposals.

We start by assessing the performance of the Clos-network switches with NoC fabrics under both uniform and non-uniform traffic arrivals. Fig.6 (a), depicts the simulation results under uniform traffic (Bernoulli *i.i.d*, for a burst size of 1 packet, and bursty uniform, for a burst size of 10 packets), The Clos-UDN and CA Clos-UDN switches achieve comparable performance under uniform packet arrivals. They both yield a higher latency than the Clos-MDN switch under light loads as they use a higher number of NoC stages. The initial delay correlates with the number of NoC stages at the middle stage of the Clos-network. Filling in the pipeline takes few cycles before the latency variation becomes quasi constant. We note that with few on-chip mini-routers fitted in the MDN modules, and a small speedup factor ($SP = 2$), the Clos-MDN switch outperforms the two Clos-UDN switch variations. Overall, trading the area by speedup improves the Clos-MDN throughput by approximately 20% under Bernoulli *i.i.d* traffic, and 30% under bursty traffic. Under *hot-spot* traffic, the Clos-MDN architecture still defeats the switches with uni-directional NoC fabrics in terms of packet latency and throughput as shown in Fig.6 (b) and Fig.6 (c) – respectively.

Next, we evaluate the delay and throughput performance of the proposed switches to the well-investigated MSM switch with a Concurrent Round-Robin Dispatching (CRRD) [8]



(a) *Diagonal* traffic (b) *Bursty* non-uniform traffic – Clos-MDN (c) *Hot-spot* traffic

Fig. 8: The impact of the design parameters on a 256-ports Clos switch with NoC fabrics.

algorithm, and the MMM switch[12] as has been described in [6]. Note that the NoC-based switches fit into the category of buffered multistage packet-switches. However, we strongly think that comparing their performance to the MSM, helps analysing the switches response with respect to their features (mesh expansion, on-chip buffering space, scheduling complexity, etc.). Fig.7 (a) and Fig.7 (b), show the simulation results for a Bernoulli *i.i.d* and uniform bursty traffic – respectively. The MSM switch performs well under light loads. However it becomes unstable under moderate loads where the average packet delay rises sharply (at around 40% of the load for a Bernoulli *i.i.d* traffic, and and 55% of the load, for a uniform arrival of bursts), and never pulls down. The MMM switch yields the lowest latency under Bernoulli traffic. Yet, its throughput saturates at around 84% under bursty traffic even if the crosspoint buffers are worth of 16 packet, each. Clearly, the switches with NoC fabric experience higher delays when they are lightly loaded. Still the delay varies smoothly and no sudden throughput collapse happens. This is a key asset showing a good scalability of the proposed switches in load fluctuation. Fig.7 (c) depicts the variation of the average throughput when the traffic gets more and more skewed. It shows that the MSM has the less stable throughput, while buffered architectures deal better with critical traffic patterns. With small on-chip buffering and small speedup factors, the NoC-based switches – namely the Clos-MDN switch – provide high throughput across the whole range of the unbalanced coefficient[13], ω.

We further test the performance of the proposed switches under diagonal traffic. Fig. 8 (a) shows that the saturation throughput is limited under high loads. Yet, the same figure shows that the Clos-MDN architecture greatly improves the delay performance over the baseline Clos-UDN switch using a small speedup factor. In Fig. 8 (b), we investigate the effect of speedup, buffer depth (buff), and switch size on the performance of the Clos-MDN switch under bursty non-uniform arrivals. A few conclusions can be drawn: Using a higher speedup factor contributes towards larger on-chip bandwidth, and improves the response of the switch to heavy bursty arrivals. Extending the on-chip buffering capacity, remarkably improves the performance of the Clos-MDN switch, and resolves the backlogs that form inside the fabric under heavy loads. In Fig. 8 (c), we plot the impact of the switch valency on the average delay performance under *hot-spot* traffic. We vary the Clos-MDN switch size, and we inspect the packet latency under the light (20%), medium (50%) and heavy (90%) traffic loads. We can see that for light and medium loads, the delay is little sensitive to the switch size showing a good scalability potential. When the Clos-MDN switch is heavily loaded, the delay increases slowly with the port count. For all simulations, the delay is always smaller than 18 time slots when the load is as high as 90%.

V. Related work

The appearance of bandwidth-demanding applications, have imposed stringent requirements of the interconnect infrastructure of data centers. Today, there is more interest in deploying scalable switches and routers in DCNs, to ensure the flawless growth of the whole system in traffic volume and number of connected end-points. Building scalable packet-switches has attracted much attention, and a plethora of proposals exits. So far, the multistage switches – namely the Clos-network packet-switches – have been a good alternative to build cost-effective high-radix switches. A combination of bufferless, semi-buffered [8], [9], [21] and fully-buffered [4], [6] switching architectures have been investigated. However, the previous solutions are too complex [8], too costly [6], or perform sub-optimally under some traffic patterns [22], [23].

The NoC paradigm has emerged as an attractive candidate to design packet-switches [24], [25]. However, their application was restricted to single-stage models [12], [13], [25]–[28]. In our previous work, we have suggested multistage switches with input-queued [14] and output-queued [29] NoC fabrics.

[12]We test MSM using the CRRD with 2 iterations since even with more iterations the performance converges to nearly the same values [8]. We also set the MMM crosspoint buffers (xbuff) to 16 packets as with only one-packet crosspoint buffering, the throughput do not exceed 65% under bursty traffic [6].

[13]When $\omega = 0$, the traffic is uniform. When $\omega = 0.5$, then the traffic is *hot-spot*. When $\omega = 1$, the traffic is directional.

The architectures described in this paper offer a variety of features that make each architectures easily customizable and highly scalable. It is possible to vary the on-chip buffers capacity, the number of pipeline stages, or the internal speedup of the NoC fabric to trade-off the cost to the performance. The uni-directional NoC-based switches offer good throughput levels and smooth latency variation. However, they need large NoC modules to perform well under critical traffic. The Clos-MDN switch employs compact CM modules. It implements VCs, a deadlock-free routing and a proactive congestion-control mechanism to forward packets to their corresponding output line cards. Running the on-chip links a little bit faster, makes the Clos-MDN achieve high-performance, scale fast in size and load, and promote the micro load-balancing [16]. On the hardware front, the NoC-based multistage switches improve the path-diversity, and simplify the queuing structure at the IMs. Costly Virtual Output Queues (VOQs) are replaced with simple FIFOsSimple and distributed packet-scheduling come to replace the complex dispatching processes that bufferless switches need, and the large buffers of the conventional MMM switches are also replaced with small and distributed on-chip queues.

VI. Conclusion

In this paper, we propose a novel way to make high-radix packet-switching architectures for the next-generation DC interconnect infrastructure. A class of multistage switches with NoC fabrics is implemented and assessed under a range of traffic patterns. The switches overcome the shortcomings of the conventional buffered multistage design. They simplify the queuing and packet scheduling mechanisms. The CA Clos-UDN and the Clos-MDN switches are two wrapped-around architectures with improved path-diversity, congestion-aware routing, and a high scalability degree which makes them good candidates for the demanding DCNs. In-depth performance analysis of the proposed switches is reserved for future work.

VII. Acknowledgement

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

References

[1] C. Clos, "A study of non-blocking switching networks," *Bell Labs Technical Journal*, vol. 32, no. 2, pp. 406–424, 1953.

[2] "Cisco," 2016. [Online]. Available: http://www.cisco.com/c/en/us/products/switches/nexus-5000-series-switches/datasheet-listing.html

[3] "Juniper Networks," June 2015. [Online]. Available: http://www.juniper.net/assets/us/en/local/pdf/datasheets/1000414-en.pdf

[4] Z. Dong, R. Rojas-Cessa, and E. Oki, "Memory-memory-memory Clos-network packet switches with in-sequence service," in *HPSR, International Conference on*. IEEE, 2011, pp. 121–125.

[5] H. J. Chao, J. Park, S. Artan, S. Jiang, and G. Zhang, "Trueway: A highly scalable multi-plane multi-stage buffered packet switch," in *HPSR, International Conference on*. IEEE, 2005, pp. 246–253.

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

[7] Y. Xia, M. Hamdi, and H. J. Chao, "A practical large-capacity three-stage buffered Clos-network switch architecture," *Trans. Parallel Distrib. Syst.*, vol. 27, no. 2, pp. 317–328, 2016.

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

[9] 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.

[10] K. Pun and M. Hamdi, "Distro: A distributed static round-robin scheduling algorithm for bufferless Clos-network switches," in *GLOBECOM, International Conference on*, vol. 3. IEEE, 2002, pp. 2298–2302.

[11] Y. Xia and H. J. Chao, "Module-level matching algorithms for MSM Clos-network switches," in *HPSR, International Conference on*. IEEE, 2012, pp. 36–43.

[12] 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.

[13] L. Mhamdi, K. Goossens, and I. V. Senin, "Buffered crossbar fabrics based on networks on chip," in *CNSR, Eighth Annual*. IEEE, 2010, pp. 74–79.

[14] F. Hassen and L. Mhamdi, "A scalable multi-stage packet-switch for data center networks," *Journal of Communications and Networks*, vol. 19, no. 1, pp. 65–79, 2017.

[15] W. Zhang, L. Hou, J. Wang, S. Geng, and W. Wu, "Comparison research between XY and odd-even routing algorithm of a 2-dimension 3x3 mesh topology network-on-chip," in *IEEE WRI Global Congress on Intelligent Systems.*, vol. 3, 2009, pp. 329–333.

[16] S. Ghorbani, B. Godfrey, Y. Ganjali, and A. Firoozshahian, "Micro load balancing in data centers with DRILL," in *Proceedings of the 14th ACM Workshop on Hot Topics in Networks*. ACM, 2015, p. 17.

[17] M. Alizadeh, T. Edsall, S. Dharmapurikar, R. Vaidyanathan, K. Chu, A. Fingerhut, F. Matus, R. Pan, N. Yadav, G. Varghese *et al.*, "CONGA: Distributed congestion-aware load balancing for datacenters," in *SIGCOMM Computer Communication Review*, vol. 44, no. 4. ACM, 2014, pp. 503–514.

[18] K. He, E. Rozner, K. Agarwal, W. Felter, J. Carter, and A. Akella, "Presto: Edge-based load balancing for fast datacenter networks," *SIGCOMM Comp. Comm. Rev.*, vol. 45, no. 4, pp. 465–478, 2015.

[19] S. Kandula, S. Sengupta, A. Greenberg, P. Patel, and R. Chaiken, "The nature of data center traffic: Measurements & analysis," in *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, 2009, pp. 202–208.

[20] P. Gratz, B. Grot, and S. W. Keckler, "Regional congestion awareness for load balance in networks-on-chip," in *HPCA, International Symposium on*. IEEE, 2008, pp. 203–214.

[21] J. Kleban and A. Wieczorek, "CRRD-OG: A packet dispatching algorithm with open grants for three-stage buffered Clos-network switches," in *Workshop on HPSR*. IEEE, 2006, pp. 6–pp.

[22] N. Chrysos, C. Minkenberg, M. Rudquist, C. Basso, and B. Vanderpool, "Scoc: High-radix switches made of bufferless clos networks," in *21st HPCA, International Symposium*. IEEE, 2015, pp. 402–414.

[23] H. J. Chao, Z. Jing, and S. Y. Liew, "Matching algorithms for three-stage bufferless Clos network switches," *IEEE Commun. Mag.*, vol. 41, no. 10, pp. 46–54, 2003.

[24] A. Bitar, J. Cassidy, N. Enright Jerger, and V. Betz, "Efficient and programmable ethernet switching with a noc-enhanced FPGA," in *ANCS'10, Proceedings of*. ACM, 2014, pp. 89–100.

[25] 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.

[26] E. Bastos, E. Carara, D. Pigatto, N. Calazans, and F. Moraes, "MOTIM-A scalable architecture for Ethernet switches," in *ISVLSI'07*. IEEE, 2007, pp. 451–452.

[27] 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, International Conference on*. IEEE, 2012, pp. 1–7.

[28] 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*. IEEE, 2015, pp. 308–314.

[29] F. Hassen and L. Mhamdi, "A scalable packet-switch based on output-queued nocs for data centre networks," in *ICC, International Conference on*. IEEE, 2016, pp. 1–6.

[30] K. Goossens, J. Dielissen, and A. Radulescu, "Æthereal network on chip: Concepts, architectures, and implementations," *IEEE Design & Test of Computers*, vol. 22, no. 5, pp. 414–421, 2005.