Proceedings Paper:
Hassen, F and Mhamdi, L (2017) Congestion-Aware Multistage Packet-Switch Architecture for Data Center Networks. In: 2016 IEEE Global Communications Conference (GLOBECOM). 2016 IEEE Global Communications Conference (GLOBECOM), 04-08 Dec 2016, Washington DC, USA. IEEE . ISBN 978-1-5090-1328-9

https://doi.org/10.1109/GLOCOM.2016.7841681

Congestion-Aware Multistage Packet-Switch Architecture 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*—Data Center Networks (DCNs) have gone through major evolutionary changes over the past decades. Yet, it is still difficult to predict loads fluctuation and congestion spikes in the network switching fabric. Conventional multistage switches/routers used in data center fabrics barely deal with load balancing. Congestion management is often processed at the edge modules. However, neither the architecture of switches/routers, nor their inner routing algorithms tend to consider traffic balancing and congestion management. In this paper, we propose a flexible design of a scalable multistage switch with cross-connected UniDirectional Network-on-Chip based central blocs (UDNs). We also introduce a congestion-aware routing to forward packets adaptively. We compare the current switch architecture to the state-of-the art previous multistage switches under different traffic types. Simulations of various switch settings have shown that the proposed architecture maintains high throughput and low latency performance.**

***Index Terms*—Data Center Networks, Clos, NoC, Congestion, Packet switching, Scheduling**

I. Introduction

DCN switches deal with a huge volume of inter-server traffic. Although there is tremendous interest in designing improved switching architectures for DCNs, few proposals suggest solutions to amend the congestion management at a switch level rather than the DC network level [1] [2]. In the context of data centers, managing the constantly increasing loads is crucial. The move of load balancing functionality in DC networks has been motivated by the apparent necessity of having a global load and congestion administration in the switching fabric. Some of the latest papers struggled to convey load balancing to centralized controllers [3], the network edge [1] [4] or end-hosts [2]. We believe that processing congestion management at a scale of the DCN switching fabric, is not enough. In fact, relying on balancing systems that only use global traffic information, makes response delays too slow as compared with the majority of the short-lived congestion events in the data center.

The expansion of the DC network substrate monitors the design of large-scale switching architectures to fit for the growing demand. Currently, high-performance switches are built using multiple smaller-radix switch chips. Ciscos CRS-3 and Junipers T600 are large capacity routers which all have multistage switching fabrics [5]. Multistage architectures such as Clos arrangements, provide better scheduling management for large sized switches/routers. They are attractive due to their high-modularity and high-expandability. Besides, the low cost of the network components gives a better performance-cost ratio when compared to a single stage crossbar switch. Memory-Space-Memory (MSM) [6] and Memory-Memory-Memory (MMM) [7] are scalable Clos-network based switch designs. Both were proposed to compromise between complexity and performance. Recent proposals advocate the use of Networks-on-Chip (NoC) paradigm to design scalable packet switches. They show that adopting this approach has many advantages over conventional crossbars such as short wires, distributed routers, path diversity and improved scalability [8] [9]. It also obviates the need for costly Virtual Output Queueing (VOQ) [10]. In spite of their performance, none of the aforementioned architectures dealt with congestion since they only fulfill passive routing.

We opt for a micro load-balancing approach [11] as it has the advantage of allowing fine-grained scale decisions (packet level). We argue that it would be even better to mutually adopt a microscopic and a macroscopic approach for fast and accurate routing in DCN fabrics. In particular, we propose a three-stage Clos-network switch that can scale up to large capacity thanks to the NoC-based central stage modules. We add unidirectional cross-interconnections linking the middle-stage's elements and we implement a congestion-aware routing to help spread the traffic load adaptively. We show that with no speedup, the switch performs well under a wide range of traffic patterns.

The remainder of the paper is structured as follows. Section II discusses some relevant existing multistage Clos packet-switch architectures and previous works with their limitations. In Section III, we describe the congestion-aware switch architecture. Section IV gives details of the new adaptive routing algorithm. We outline some implementation issues in section V, and we provide section VI to present and discuss the experimental results. Ultimately, Section VII concludes the paper.

II. Background and Prior Work

Switching architectures can be classified based on their blocking features, inter-stages connection, scheduling scheme, etc. Other criteria such as the nature of the Switching Elements (SEs) might be adopted [12] [7] [13] [5]. Bufferless Clos switches and MSM switches require a global path allocation

which is typically a two-step scheduling process. The matching resolves the contention for paths and output ports using a centralized scheduler. In a bufferless architecture, packets arrive to their destinations in an ordered way which obviate the needs for re-sequencing mechanisms. Still, the need for a central scheduler rises the system complexity and makes the architecture less appealing for large-scale switches. Buffered structures provide higher performance than bufferless switches. They need simple control, but backpressure mechanisms must be implemented to prevent buffers overflowing. Packets of the same flow are likely to experience variable delays depending on their sejourn in the middle stage SEs. This results in an out-of-sequence packets delivery. Some new interesting proposals suggested building scalable high-performance switches/routers using the Networks-on-Chip paradigm [15] [16] [17]. A NoC-based switch brings several advantages over classic crossbars, such as a flexible design, a pipelined scheduling and a sub-quadratic growth of the fabric's cost.

A paramount concern in data center switching fabrics, is to assure continuous load balancing. This is a key point to enhance the network performance and to promote its robustness to floating traffic. Few multistage switch designs have considered load balancing among the SEs using architectural and/or algorithmic solutions. The two-stage load balanced Birkhoff-Von Neumann switch was first introduced in [18] where the first stage balances the traffic load and the second stage performs the switching function. In [19], Smiljanić suggested some load-balancing algorithms for a three-stage Clos network. In [20], Chrysos presented a distributed congestion management scheme for a buffered three-stage Clos switch and evaluated its performance under different traffic patterns. The current work targets micro-load balancing in DCN switching fabrics. We propose modifications to the well-known three-stage Clos arrangement. Unidirectional crossed links between central UDNs and a congestion-aware routing algorithm are employed to flexibly send traffic among different Central Modules (CMs).

Limitation of previous works

Intrinsically bufferless Clos switches cannot balance the load due to the straight point-to-point connections in the crossbar SEs. On the other hand, buffered structures and Clos-UDN switches have internal buffers making them capable of overtaking congestion events. Although the Clos-UDN [8] [9], has good scalability and flexibility features, the switch, as other proposals, deals blindly with congestion. Dimension Order Routing (DOR) methods that a Clos-UDN switch uses, are simple to implement in hardware [21]. Conversely, they poorly disperse the traffic load among links of the NoC. On the contrary, adaptive routings [22] [21] [23] improve the network performance since they tolerate failure and make intelligent arbitration based on the NoC status. CMs of the Clos-UDN switch adopt the $'Modulo\ XY'$ algorithm (which is a variant of DOR) to geometrically route packets through the 2-D mesh. Under strongly unbalanced traffic patterns, Clos-UDN with no speedup ($SP = 1$) suffers bad load distribution which causes local congestion and leads to performance collapse.

In this paper, we modify the Clos-UDN switch architecture and we enhance the $'Modulo\ XY'$ routing for a more robust and reliable switch design. We adopt the Regional Congestion Awareness (RCA) technique [22] to propagate an estimation of the congestion information and to monitor packets routing through the mesh. The approach is interesting for its simplicity and effectiveness. It helps dispersing congestion statistics in a scalable way across the middle stage of the Clos-network with few hardware modifications.

III. The Switch Architecture

Because of the limitation of traditional Clos topology, additional alternative routing resources can provide more network tolerance and further improve the switch performance. Links between the central modules of the three-stage Clos-network are proposed to add connectivity on baseline Clos-UDN architecture. Inter-CMs links reduce traffic congestion in the whole Clos switch under critical traffic patterns and contribute to better load distribution among CMs.

Cross-module interconnection

We mention that above all, we are concerned with designing a scalable and easily configurable switch that meets high performance requirements of today and the next-generation DCN fabrics. For simplicity, we consider $Benes$' lowest-cost practical non-blocking architecture that has an expansion factor $\frac{m}{n} = 1$. The first stage of the switch comprises k Input Modules (IMs), each of which is of size $(n \times m)$. The middle stage is made of m UDNs, each of dimension[1] $(k \times k)$. The third stage consists of k Output Modules (OMs), each of which is of size $(m \times n)$. We maintain m FIFOs per IM, each of which is associated to one of the m output links denoted as LI(i, r). Because $m = n$, each FIFO(i, r) is also associated to one input port, IP(i, r). It can receive at most one packet and send at most one packet to one central module at every time slot. CMs are related to OMs with m links that we call LC(r, j). An OM(j) has n OPs, to which are associated n output buffers. An output buffer can receive at most m packets and forward one packet to the output line card at every time slot.

We consider a static dispatching scheme. Every FIFO constantly delivers packets to the same CM on the connecting LI link. Packets in the Clos-UDN switch are routed minimally using the $'Modulo\ XY'$ algorithm. Traffic flows travel[2] W/E, W/N, W/S, N/S and S/N. Our previous results showed that a static packets dispatching and an oblivious routing scheme, are irrelevant to skewed traffic arrivals. In fact, some UDNs can get more congested than others resulting in longer delays

[1]Unlike conventional Clos networks, the central modules of the switch can be of size $(k \times M)$ crosspoints, where M refers to the NoC depth and $M \leq k$. The switch can 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.

[2]North (N), South (S), East (E), West (W).

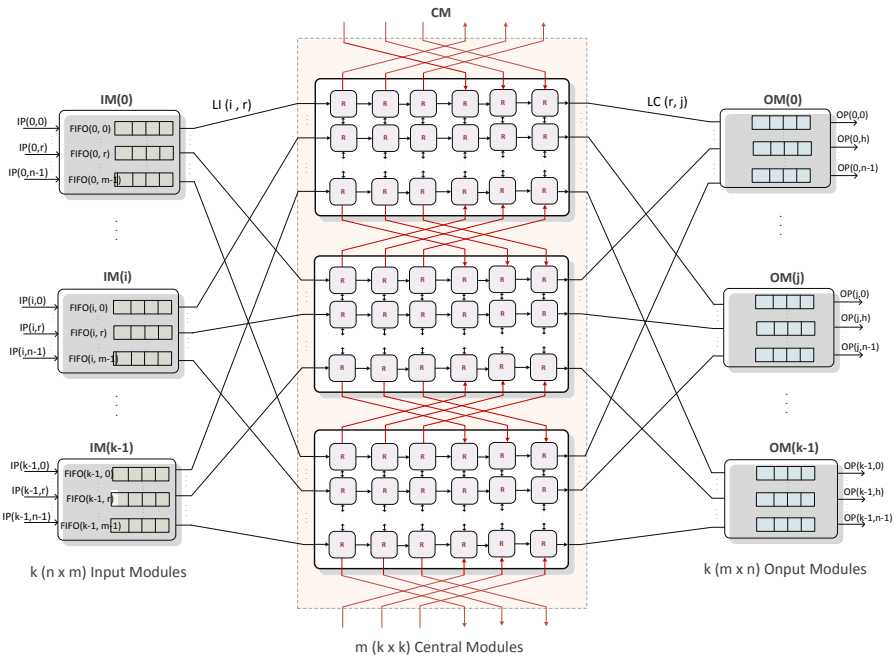

Fig. 1: $N \times N$ Three-stage Clos switch architecture with cross-connected CMs

and poor delivery ratio. We thought of an elegant way to make central modules of the switch share traffic and allow a proper load distribution. We take advantage of the NoC design and we suggest a wrapped-around Clos-network such that $CM(r)$ connects to $CM((r-1) \mod m)$ and $CM((r+1) \mod m)$ by means of M unidirectional links. $\frac{M}{2}$ links serve to send traffic to the upper (or lower) CM neighbour and exactly the same number of links is used to receive traffic from an adjoining module as depicted in Fig.1. We assume that the depth of the UDNs (M) is even and that $node(a, b)$ is a mini-router located at row a and column b of the mesh, where $(0 \leq a \leq k-1)$ and $(0 \leq b \leq M-1)$. For $0 \leq x \leq \frac{M}{2} - 1$, we connect the edge rows of two adjacent CMs such that $node(0, x)$ in $CM(r)$ is related to $node(k-1, x+1)$ in $CM((r-1) \mod m)$. Note that since links are unidirectional, no deadlocks can occur.

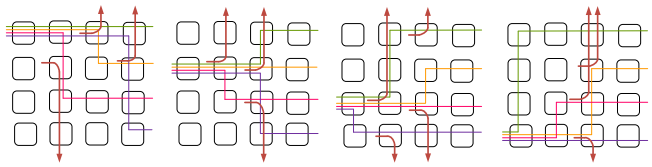

Fig. 2: Example for the Repellent and $'Modulo\ XY'$ algorithms

IV. Congestion-aware routing

Newly arriving packets are stored in FIFOs at the input stage. We implement a static dispatching process where every FIFO is said to persistently send packets to a given CM. Packets scheduling takes place at the heart of the Clos-network: The middle stage, made of UDNs. We add inter-CM links to allow sending and receiving traffic from nearby CMs. The main difference between the proposed switch and previous works is that, the routing unit can select different paths for packets of the same flow depending on the Clos-network condition. A packet might be routed minimally through the current CM or be sent to the nearest less-congested module. It is important to note that NoC-based switches are delay sensitive and that the overall design performance heavily relies on the nature of the routing scheme [22]. Much research has gone into designing routing algorithms with provable behavior. While these approaches typically assume healthy network and fairly distributed load, DCN switches frequently have non-uniform (and sometimes bursty) injection rates and time varying communications. This often leads to temporary congestion known as hot-spots. Schemes that have some flexibility with respect to route choice, provide advantages over oblivious routings that are not able to adapt to the communication pattern and the network status. In this paper, we choose to add functionalities to the minimal-routing $'Modulo\ XY'$ algorithm rather than using a fully adaptive method. A routing decision takes into account the congestion estimate at different points in the Clos-network and forward packets correspondingly.

A. Congestion evaluation

We adopt a metric that is suitable for the routing scheme to correlate well with the global Clos-network congestion and be inexpensive to compute. We consider the RCA [22] approach to evaluate and propagate congestion information across a UDN of index r and its direct neighbour CMs (blocs of indices $((r-1) \mod m)$ and $((r+1) \mod m)$). Thanks

to RCA, we compare locally competed congestion metrics with those propagated from a neighbour CM before taking the routing decision. We define a routing quadrant to be the sub-network limited by the packet's current position in the 2-D mesh and the egress port through which it exits the current CM to the third stage of the Clos. We also define the local CM information to be the information readily available at a given CM module and representing the status of all nodes (also called mini-routers) that figure in the routing quadrant. Given its current position, a packet can travel in one of the two quadrants N/E and S/E with each quadrant having exactly two possible output directions excluding the local port. Buffers occupancy is a classic congestion metric that reflects the load distribution in points of the network. To keep on routing traffic adaptively through minimal paths, we combine two metrics: the buffers occupancy and the hops count.

B. Repellent routing

For NoC-based switches, adaptive routings are better than oblivious schemes whenever the traffic is non-uniform. However adaptive methods can disrupt load balance due to local decisions that lack knowledge of the network state beyond the nearest neighbours of a node. In case of 2-D mesh, they congest the middle of the NoC and steer the traffic towards the center leaving the edge nodes/links underutilized. In this proposal, we modify the routing policy to make it possible for packets to exit a currently congested CM towards a less crowded module. We call this scheme: Repellent routing as it tends to push a portion of the traffic to borders of the mesh as shown in Fig.2. Colored paths are used to illustrate routing decisions taken by the $'Modulo\ XY$ algorithm. Whereas, red thick lines show the effect of Repellent routing in changing a packet's path towards a neighbor CM.

At every time slot and any position in a UDN module, a packet is subject to two levels of decision making: First, select the closest CM neighbour. Next, elect the less-congested routing quadrant. CM$((r+1) \mod m)$ is said to be closer than CM$((r-1) \mod m)$, if the vertical distance from the current node to the first row of the mesh is less than that to the last row. As mentioned earlier, coupling the distance information with information about the load distribution in the routing quadrant, minimizes the impact of pushing packets back away from their destinations to be routed through another CM module.

If the cell is going to be routed locally, then $'Modulo\ XY'$ algorithms is used. Otherwise the packet is sent vertically $North$ (or $South$) until the first (or last) row of the NoC where it leaves the CM to another block. Algorithm.1 gives details of the routing logic that our NoC-based modules adopt. We mention that the crossed inter-CMs connections reduce the number of NoC stages that a cell must go through until its corresponding LC link to avoid cumulating latencies and declining performance of the switch.

Algorithm 1 : Repellent routing

```
1: if (pck_repulsed = TRUE) then
2:     port ← routing_direction
3: else
4:     fct : choose closest CM
5:     if (local routing quadrant is less congested) then
6:         'Modulo XY',
7:         pck_repulsed ← FALSE
8:     else
9:         if (chosen CM is 'UP') then
10:            routing_direction ← North,
11:            port ← North,
12:            pck_repulsed ← TRUE,  //Override bit
13:        else
14:            routing_direction ← South,
15:            port ← South,
16:            pck_repulsed ← TRUE
17:        end if
18:    end if
19: end if
```

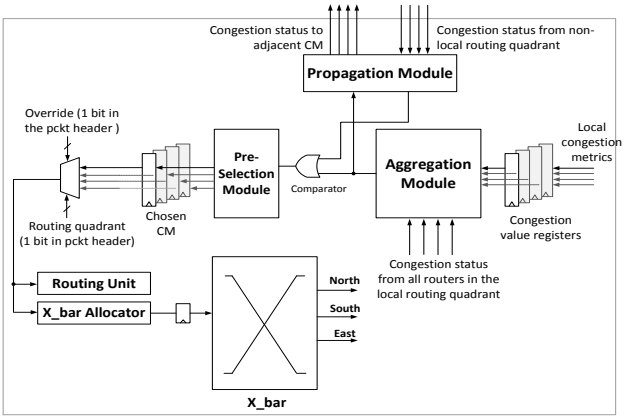

Fig. 3: Design of mini router

V. Implementation Issues

A. Mini-Routers Micro-architecture

On-chip routers use distance and buffers occupancy of downstream nodes within a routing area to evaluate congestion. In conventional adaptive routers, only intrinsic congestion information is used to select a preferred port (occupancy of output buffers as a common example). RCA approach helps aggregating local and non-local information to better estimate the congestion status [22]. Fig.3 shows a high-level design of a mini-router. The aggregation module uses the specified congestion metrics to combine values and feeds the result to both, the comparator and the propagation module. The major difference between a typical RCA mini-router and the present design, is that congestion information from different CMs is used for routing decision making. The pre-selection module keeps reference to CM(r), CM($(r-1) \mod m$) and CM($(r+1 \mod m$). Based on the output of the comparator, packets might be routed locally or sent to an adjacent bloc. The propagation unit transfers congestion information from and to other nodes

in the routing quadrant. Unlike the common RCA router that sends information in a single direction, our propagation module requires additional logic to convey congestion estimate to nodes of a remote CM. In the default case scenario, a packet is routed locally using the $'Modulo\ XY'$ algorithm until the routing unit indicates that it should go through another less-congested CM. Next, the cell will have to go the way UP or $DOWN$ to exit the current UDN. Consequently, it must be able to override the value indicating the routing CM. This action is accomplished via an override bit in the packet's header.

B. Re-ordering packets

Out-of sequence packets delivery is a common problem to all multistage packet switch architectures with buffered middle stages. A re-sequencing mechanism at the output stage of the switch [24] is a popular solution to this phenomenon. In an extension of our previous work [8], we suggested a static cells dispatching to prevent out-of-order packets delivery. However, the current proposal breaks this asset. Sending packets across the middle stage of the Clos-network in a flexible way to enhance load balancing and to mitigate congestion, mis-sequences packets order. We may consider one of the several ways to resolve this issue. In [24], authors discussed two re-ordering schemes based on time-stamp monitoring. Although both alternatives do not require synchronization among the different SEs, many buffers and arbiters have been introduced making the solutions unscalable. In [14], H. J. Chao et al. proposed other re-sequencing mechanisms such as: Static and dynamic hashing, and window-based resequencing. Our switch requires a re-sequencing stage to re-establish the correct packets order, but we reserve this part to future work and consider it to be out of the scope of this paper.

VI. Experimental Results

Throughput and delay are the two most important performance metrics used to evaluate packet switches and routers. In this section, we test the delay performance of the congestion-aware Clos-UDN switch under different scenarios and we compare it to the Clos-UDN with static packets dispatching, MSM [6] and Memory-Memory-Memory (MMM) [24] switches. We vary the switch size and the traffic profile. In all our simulations, the depth of the UDN mesh (M) in all the central modules of the Clos-network is such as $M = k$, if not explicitly mentioned. The delay is the averaged value over all packet queuing delay measured in a simulation. We start evaluating the switch's performance under uniform traffic. We consider Bernoulli arrivals. Results are shown in Fig.4. Both the congestion-aware switch and the Clos-UDN switch perform poorly under light loads. All the same, our focus will be mainly on heavy loads as they are more relevant to the context of data centers. With $SP = 1$, a congestion-aware architecture improves upon the Clos-UDN with static packets dispatching. Thanks to the inter-CM connections and the Repellent routing scheme, packets continue to be routed minimally across the Clos-network taking into consideration the congestion levels in the CM modules. We note that the average packets delay is slightly reduced and that the throughput of the switch is boosted. Under heavy loads, rising the speedup of the UDN units to 2 makes our proposal outperform MSM (even if the Concurrent Round-Robin Dispatching (CRRD) algorithm is iterated 4 times) and MMM with crossbar buffers worth of one packet each. In Fig.5, we vary the switch valency. Simulations show that a congestion-aware switching architecture ameliorates the overall packets delay and the throughput even if no speedup is used ($SP = 1$).

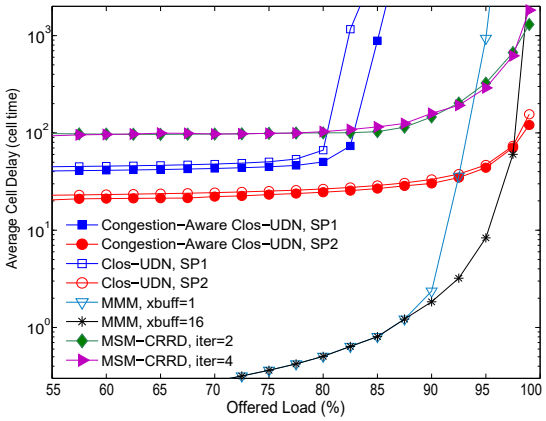

Fig. 4: Delay performance of (256×256) switches under Bernoulli uniform traffic

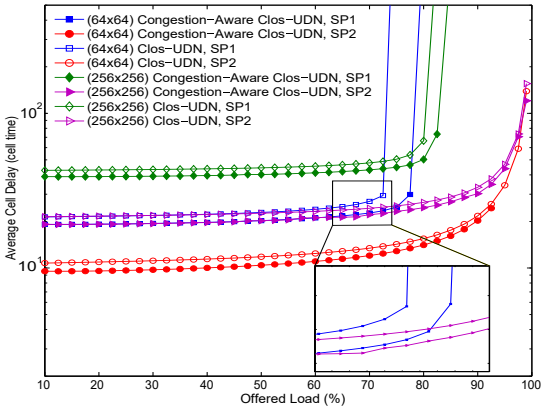

Fig. 5: Delay performance the Congestion-Aware Clos-UDN switch under Bernoulli uniform traffic for different switch sizes

Workloads in the DCN are perpetually changing. Many high-bandwidth demanding applications make a bursty traffic relevant to DCNs with high-levels of peak utilization. In our simulations, we set the default burst length to 10 packets. A bunch of packets that arrive at the same On-period are destined to the same output port. As presented in Fig.6, the currently proposed switch decreases the end-to-end latency and slightly improves the throughput. Experimental results show that it is possible to improve the switch response to burstiness by speeding-up the UDN modules. With $SP = 2$, our switch beats MSM and MMM architectures under heavy traffic loads.

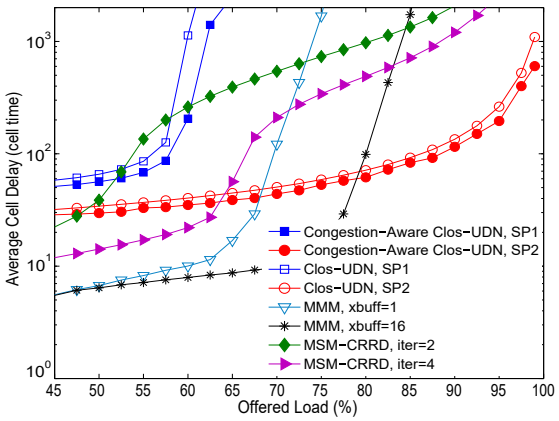

Fig. 6: Delay performance of (256×256) switches under Bursty uniform traffic

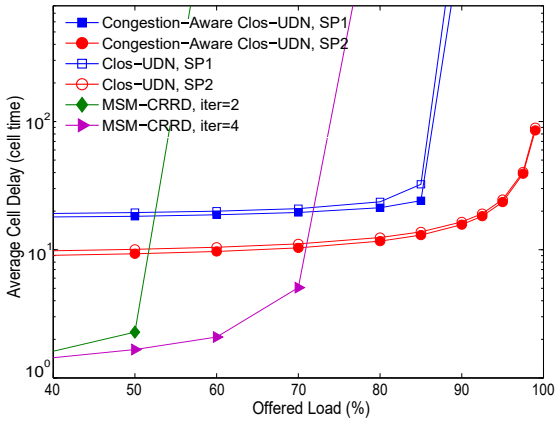

Fig. 7: Delay performance of (64×64) switches under hot-spot traffic, $\omega = 0.5$

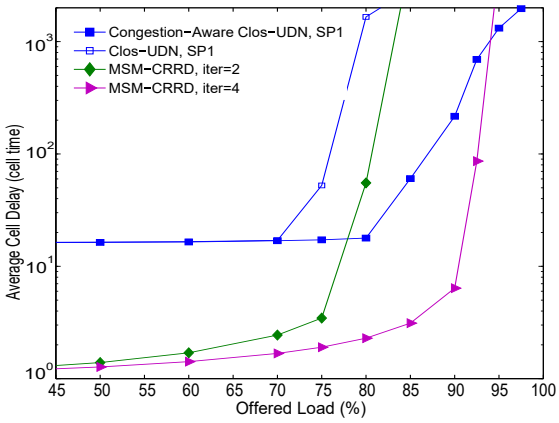

Fig. 8: Delay performance of (64×64) switches under Double-Diagonal traffic

The uniform traffic is not realistic. Thus, we run the next set of simulations under unbalanced traffic patterns to test our design's robustness to non-uniformity. We consider the following scenarios: Bernoulli unbalanced, diagonal and hot-spot

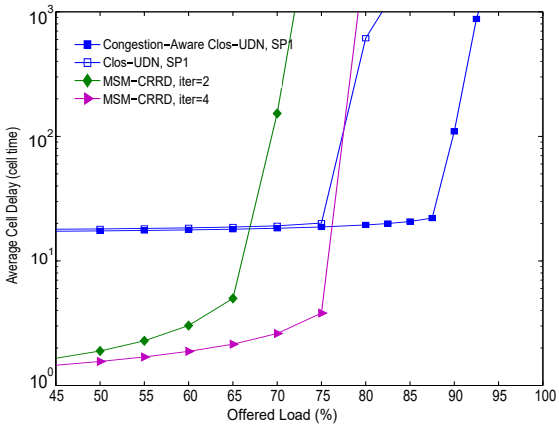

Fig. 9: Delay performance of (64×64) switches under Log-Diagonal traffic

arrivals. An unbalanced traffic pattern uses a probability, ω as the fraction of input load sent to a predetermined output, while the rest of the input load is uniformly directed to other output ports. As compared to the Clos-UDN with static dispatching scheme, the congestion-aware switch provides lower delays thanks to the adaptive routing we use. More importantly, it outperforms the MSM switch with CRRD scheduling under medium and high loads even if $SP = 1$. We conduct more simulations while considering the minimum SP value and a switch size (64×64) under a diagonal traffic. A diagonal traffic can be represented as $d\rho(i,j) = d\rho_i$ for $i = j$ and $(1-d)\rho_i$ for $((i+1) \mod N)$, where N is the generic switch size and ρ_i is the load at input i. Fig.8 and Fig.9 compare the delay performance of the congestion-aware switch to the baseline switch and to MSM with CRRD scheduling. We observe that a cross-connected Clos-UDN architecture used along with an appropriate routing is more effective under skewed traffic pattern. With no speedup, our switch distributes better the load across the Clos-network and achieves high throughput.

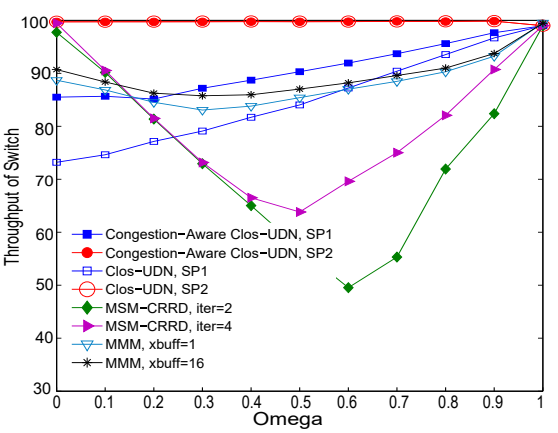

Fig. 10: Throughput stability of different switches under Bernoulli traffic, var ω

In Fig.10, we compare the throughput of the different switches under Bernoulli traffic. Changing the coefficient ω from 0 to 0.5 corresponds to shifting from a uniform traffic

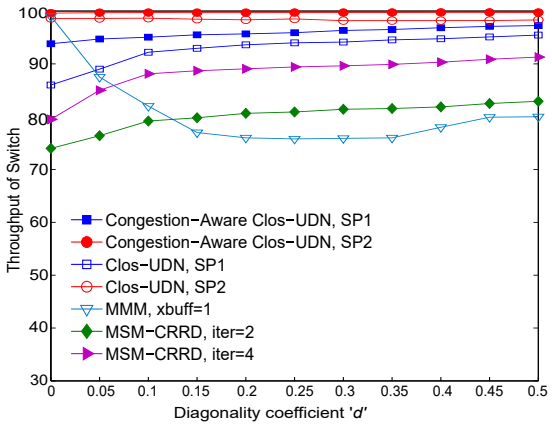

Fig. 11: Throughput stability of different switches under Diagonal traffic, var d

to a hot-spot traffic. We point that for $SP = 1$, a congestion-aware design increases the throughput of the baseline Clos-UDN. Additionally, our design provides better performance than both MSM and MMM under medium and heavy traffic loads. Increasing SP to 2, makes our switch insensitive to traffic variation as it unconditionally delivers full throughput. In Fig.11, we alter the value of the diagonality coefficient d to see how the throughput of different switches evolve. Simulations show that the response of MSM and MMM to diagonal traffic is poor under almost the whole range of d. On the contrary, our proposal delivers up to 96% throughput assuming no speedup is used and full throughput for $SP \geq 2$.

VII. Conclusion

We proposed a three-stage Clos switch with UDN central modules and inter-CM connections. Used with an appropriate routing algorithm, the wrapped-around architecture allows better load balancing among the middle stage blocs. We adopt the Regional Congestion Awareness (RCA) and we modify the micro-architecture of the on-chip routers to make them capable of evaluating and comparing congestion status at different points of the Clos-network. For more effective routing, we combine the minimal path and buffers occupancy metrics to estimate the local and remote congestion in corresponding routing quadrants. Our focus is mainly on the switch performance under high loads as they are more relevant to the context of DC networks. Experimental results show that our design delivers good throughput and bearable delays under a variety of traffic types. Yet, packets are likely to cross different CMs which results in an out-of-order delivery urging the need for a re-sequencing stage to re-establish the correct cells' order.

References

[1] 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 *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 4. ACM, 2014, pp. 503–514.

[2] K. He, E. Rozner, K. Agarwal, W. Felter, J. Carter, and A. Akella, "Presto: Edge-based Load Balancing for Fast Datacenter Networks," in *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*. ACM, 2015, pp. 465–478.

[3] J. Perry, A. Ousterhout, H. Balakrishnan, D. Shah, and H. Fugal, "Fastpass: A centralized zero-queue datacenter network," *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 4, pp. 307–318, 2015.

[4] P. Wang and H. Xu, "EXPEDITUS: Distributed Load Balancing with Global Congestion Information in Data Center Networks," in *Proceedings of the 2014 CoNEXT on Student Workshop*. ACM, 2014, pp. 1–3.

[5] Y. Xia, M. Hamdi, and J. Chao, "A Practical Large-capacity Three-Stage Buffered Clos-network Switch architecture."

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

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

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

[9] ——, "A Scalable Packet-Switch Based on Output-Queued NoCs for Data Centre Networks," in *Communications (ICC), 2016 IEEE International Conference on*. IEEE, 2016, p. in press.

[10] K. Goossens, L. Mhamdi, and I. V. Senin, "Internet-Router Buffered Crossbars based on Networks on Chip," in *Digital System Design, Architectures, Methods and Tools, 2009. DSD'09. 12th Euromicro Conference on*. IEEE, 2009, pp. 365–374.

[11] 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.

[12] 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.

[13] X. Li, Z. Zhou, and M. Hamdi, "Space-Memory-Memory architecture for Clos-Network Packet Switches," in *ICC 2005*. IEEE, 2005, pp. 1031–1035.

[14] 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, 2005*. IEEE, 2005, pp. 246–253.

[15] L. Mhamdi, K. Goossens, and I. V. Senin, "Buffered Crossbar Fabrics Based on Networks on Chip." in *CNSR*, 2010, pp. 74–79.

[16] T. Karadeniz, A. Dabirmoghaddam, Y. Goren, and J. Garcia-Luna-Aceves, "A New Aproach to Switch Fabrics based on mini-router grids and Output Queueing," in *Computing, Networking and Communications (ICNC), 2015 International Conference on*. IEEE, 2015, pp. 308–314.

[17] A. Bitar, J. Cassidy, N. Enright Jerger, and V. Betz, "Efficient and Programmable Ethernet Switching with a NoC-enhanced FPGA," in *Proceedings of the tenth ACM/IEEE symposium on Architectures for networking and communications systems*. ACM, 2014, pp. 89–100.

[18] C.-S. Chang, D.-S. Lee, and Y.-S. Jou, "Load Balanced Birkhoff-Von Neumann Switches," in *High Performance Switching and Routing, 2001 IEEE Workshop on*. IEEE, 2001, pp. 276–280.

[19] A. Smiljanić, "Load Balancing Mechanisms in Clos Packet Switches," in *Communications, 2004 IEEE International Conference on*, vol. 4. IEEE, 2004, pp. 2251–2255.

[20] N. I. Chrysos, "Congestion Management for Non-Blocking Clos Networks," in *Proceedings of the 3rd ACM/IEEE Symposium on Architecture for networking and communications systems*. ACM, 2007, pp. 117–126.

[21] C. Wang, W.-H. Hu, and N. Bagherzadeh, "Congestion-Aaware Network-on-Chip Router Architecture," in *Computer Architecture and Digital Systems (CADS), 2010 15th CSI International Symposium on*. IEEE, 2010, pp. 137–144.

[22] P. Gratz, B. Grot, and S. W. Keckler, "Regional Congestion Awareness for Load Balance in Networks-on-Chip," in *High Performance Computer Architecture, 2008. HPCA 2008. IEEE 14th International Symposium on*. IEEE, 2008, pp. 203–214.

[23] X. Chang, M. Ebrahimi, M. Daneshtalab, T. Westerlund, and J. Plosila, "Pars An efficient Congestion-Aware Routing method for Networks-on-Chip," in *Computer Architecture and Digital Systems (CADS), 2012 16th CSI International Symposium on*. IEEE, 2012, pp. 166–171.

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