# A hybrid packet/circuit optical transport architecture for DCN

**Nogol Panahi*, Davide Careglio, Josep Solé Pareta**
*Advanced Broadband Communications Centre, Universitat Politècnica de Catalunya, 08034 Barcelona, Spain*
*Tel: (0034) 93 401 6985, Fax: (0034) 93 401 7055, e-mail: nogol.panahi@upc.edu*

**ABSTRACT**
The aim of this paper is to move away from today's multi-tier, manually operated, and performance limited Data Centre Network (DCN) towards more scalable, flexible, and optimized architecture of tomorrow. We propose a new hybrid optical DCN architecture capable of providing a fully optical interconnection path from server to server within racks and between racks utilizing fast packet switching at the aggregation layer, and high-capacity and low latency circuit switching at the core layer. We have built the DCN control plane upon the Software Defined Networks (SDN) paradigm to leverage the added value of emerging optical technologies for providing flexible and programmable heterogeneous IT and network service. Dynamicity, flexibility, and resiliency are combined in the propose infrastructure to offer advanced services to the data centre providers in order to fulfil the requirements of upcoming applications.
**Keywords**: data centre network architecture, hybrid optical packet and circuit switching, software defined networking, failure recovery, routing algorithm.

## 1. INTRODUCTION

Designing a data centre (DC) architecture has recently been a top research priority for both academic and industry area [1][2]. Today, the specific requirements of DCN are multiple, like high throughput, low latency, vast application diversity, high power density, high reliability, low energy consumption, high resiliency, which are posing significant challenges to their infrastructure and operations [3]. Since today's DCs can contain about millions of servers, scalability is also become a big concern [3].

Optics is still used in DCN mainly (almost exclusively) to provide high bandwidth while electrical switches are used to store and forward operations [4]. This technique has some disadvantages (storing causes delay, fixed routing, doubling equipment and fibres for protection, etc.). On the contrary, nowadays optical technologies can be used to enhance the infrastructure and allow to place more functionalities direct in the optical domains, speeding up the processes, and reducing the energy consumption.

On the other side, advanced control and management with SDN and Network Functions Virtualization (NFV) orchestration opens up new horizons to coordinate and optimise the resources and match the applications and services requirements.

In this direction, this paper contributes in the following aspects:

- Define a clustered packet/circuit optical data plane for the DCN architecture able to meet the aforementioned requirements;
- Design and implement an operational algorithm for the proposed DCN architecture that, thanks to the centralised flavour of SDN, can tune its behaviour according to the network status and traffic patterns;
- Design a strategy to reduce the loss of performance if a failure at both node or link level occurs in any place in the network.

The rest of the paper is organised as follows. Section 2 identifies the main shortcomings of the state of the art and the most promising research solutions. In Section 3, we detail our DCN solution, which is based on the use of both optical packet and circuit technologies to achieve better performance. Section 4 presents a sample of the obtained results. Finally, Section 5 concludes the paper and discusses the future works.

## 2. EXPECTED REQUIREMENTS AND CURRENT TRENDS

Current real cloud architectures like the ones used by Google, Facebook, and Microsoft are using large distributed system with specialized tiers and technologies for different tasks: edge, backbone, and data centres [5]. Such DCNs are applying a fixed statically provision high capacity between all servers using electrical devices. While at the first glance, it seems a way for preventing communication bottleneck assuming arbitrary traffic, it suffers wiring challenges and management complexity [6]. Moreover, full bisectional bandwidth across the entire DCN is not needed for most of the applications and certain parts of the network are rarely used or even sit idle.

Harnessing the power of optics while utilizing the optical flat-fabrics inside the data centres has drawn attention recently to cope with the forecasted traffic growth, overcoming the limitations of legacy hierarchical infrastructure, and having better manageability of increasing data traffic. Optical DCNs are flexible forms for dynamically reconfiguring the infrastructure provisioning high bandwidth resources across the network where it is needed the most. So flexible optical DCNs are becoming more popular compared to those fixed traditional electrical ones [5]. However, prior optical DCNs are even hard to scale, vulnerable in case of failure, and have limited flexible bandwidth for the increasing workload of applications on data centre market. In the other words, they cannot achieve all of the properties simultaneously. Recent research solutions are trying to increase the manageability of

 DOI 10.1109/ICTON.2019.8840229

the traffic inside the DCN and cope with heterogeneous requirements in terms of latency, capacity and availability. These solutions foster the use of hybrid packet/circuit optical technologies [7][8]. Our solution proposed in this paper goes in this direction, where several optical packet switching clusters are used to aggregate traffic from racks in larger flows to be routed through optical circuits of the mesh core nodes.

## 3. NEW DCN ARCHITECTURE BASED ON PACKET/CIRCUIT SWITCHING

### 3.1 Architecture

In our proposal, we are considering the following building blocks:

- High-capacity Optical Circuit Switching (OCS) nodes. An OCS is data rate agnostic and extremely energy efficient as it simply reflects the light from one port to another port. OCS can provide lower cost, larger scale, faster switching time, lower insertion loss for overcoming economic, and scale and performance challenges of DCNs [9]. The drawback of OCS is that it cannot change its configuration at the packet timescale but it has to be configured at the beginning and re-configured only if needed.
- Large and scalable optical packet switching (OPS) node. An OPS node is capable of providing data aggregation and fast network connectivity to servers at the packet level [7]. Indeed, OPS supports short-lived traffic and provides well-established and functional switch architectures with lower power requirements, less heat, and less space compared to electronic equipment. To facilitate the aggregation at the OPS, packets from servers are transmitted using a Time Division Multiple Access (TDMA) protocol (i.e., the time is slotted).
- Flexible wavelength multiplexing. Multiple packets can be transmitted at different wavelengths at the same time so then the cabling overhead is reduced and the spectral efficiency of the links is increased [5].
- An SDN-based control plane. This plane is in charge of monitoring the network resources, implementing automated procedures for connectivity setup, and matching the QoS requirements of all diverse services [10]. This implies that a planning phase is needed at the beginning to create the nodes connectivity (the lightpaths) and to assign the correct spectrum to each of them. During the operational phase, the connectivity may need to be reconfigured due to changes in the traffic that are creating long-term congestions or due to failures. In this case, the control plane needs to trigger an operational algorithm to reconfigure part of the DCN without service interruption.

The proposed overall architecture is shown in Figure 1. It consists of several clusters interconnected in a mesh topology by means of OCS nodes which also provide Internet and inter-DC connectivity. To provide reliability, each cluster is connected to at least two different OCS nodes. Therefore, each cluster consists of several racks interconnected by means of a couple of OPS nodes (the redundancy is for protection). Each rack consists of several servers and/or storage devices and a Top of the Rack (ToR) switch, which provides connectivity to other racks and clusters. OPS nodes provide a statistical multiplexing optical communication to other servers in the same cluster using few WDM channels. Besides, it provides an optical aggregation layer using many WDM channels for the communication to servers in other clusters through the core network. This means that the transmission inside the DC is all-optical from ToR to ToR.

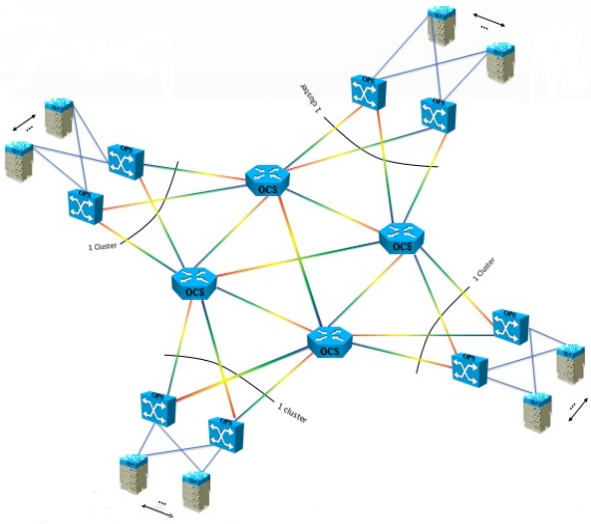



*Figure 1. Proposed hybrid data centre network architecture*

### 3.2 Routing algorithm and failure recovery

The connectivity between racks in the same cluster is provided by the two OPS nodes. On the contrary, lightpaths are needed between racks in different clusters or towards Internet. A lightpath goes from an OPS to another OPS through the OCS nodes. In this proposal, we consider that $k$ shortest paths are pre-calculated from each OPS to each other OPS, each one spanning one to $n$ wavelengths. The number of wavelengths assigned to each lightpath are also pre-calculated balancing the use of all resources. Therefore, when a packet is sent from a ToR to the OPS, the OPS selects the shortest path and the first available wavelength to the destination cluster; if more packets than wavelengths have to be sent to the same cluster, the second shortest path is selected and so on.

To ensure connectivity in case of failure, the shortest paths calculated from each of the two OPS of the same cluster to each other cluster are disjoint. Besides, each cluster is connected to at least two different OCS nodes. If a link, an OPS or an OCS connecting a cluster fails, a connectivity path is always guaranteed to each other cluster.

In addition, we consider that packets can be either unicast or multicast; meaning that a packet can be sent only to another rack or to a group of different racks. OPS nodes are in charge of duplicating and sending the packets to the different destinations using the available shortest paths.

## 4. RESULTS

### 4.1 Scenario

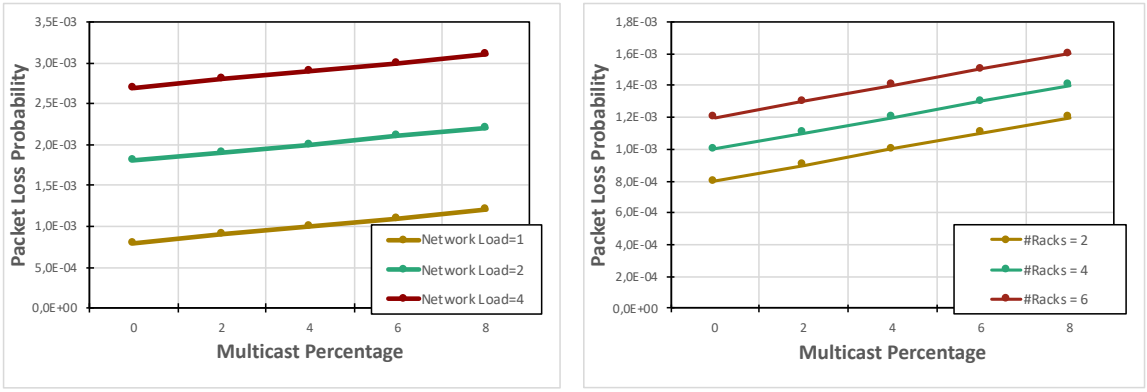
A system simulation was used to examine the evaluation of the proposed DCN consisting of 8 clusters, 16 OPS, and 4 OCS nodes as shown in Fig. 1. The OCS nodes are in the fully connected mesh network. The clusters can have different number of wavelengths, racks, and servers. In this evaluation, we have considered the parameters listed in Table 1.

*Table 1. Simulation system parameters*

| **Parameters** | **Value** |
|---|---|
| Servers per rack | {40-80} |
| Racks per cluster | {2-8} |
| Number of OCS nodes | 4 |
| Number of clusters | 8 |
| OPS nodes per cluster | 2 |
| Connectivity of each rack to the OPS nodes in its cluster | 2 links |
| Wavelengths of each link from each ToR to its connected OPS | 1 |
| Connectivity from each OPS node to an OCS node | 1 link |
| Connectivity degree of a cluster | 4 links |
| Number of connected OCS nodes to each OPS node | 2 |
| Number of shortest paths between each pair of OCS nodes | 5 |
| Wavelengths belong to each shortest path | {2-7} |
| Maximum number of packets received at each destination at a time slot | 2 packets |
| Number of wavelengths of each link in the OCS network and between OPS/OCS | {10-60} |
| Percentage of multicast traffic | {0-8}% |
| Failure probability of each component | {0,1,5,10,15}% |

### 4.2 Regular scenario

In this part, we analyse the network the packet loss probability by varying the multicast percentage, the network load, and the number of racks. Figure 2(a) shows that incrementing the multicast percentage causes an increase of the loss. As expected, incrementing the network load also causes more losses. This is because of the increase of average number of transmitted packets which creates congestion in the links between the racks and the OPS. It is worth mentioning that in this scenario we are considering a single wavelength is available in this link.




*(a)* *(b)*

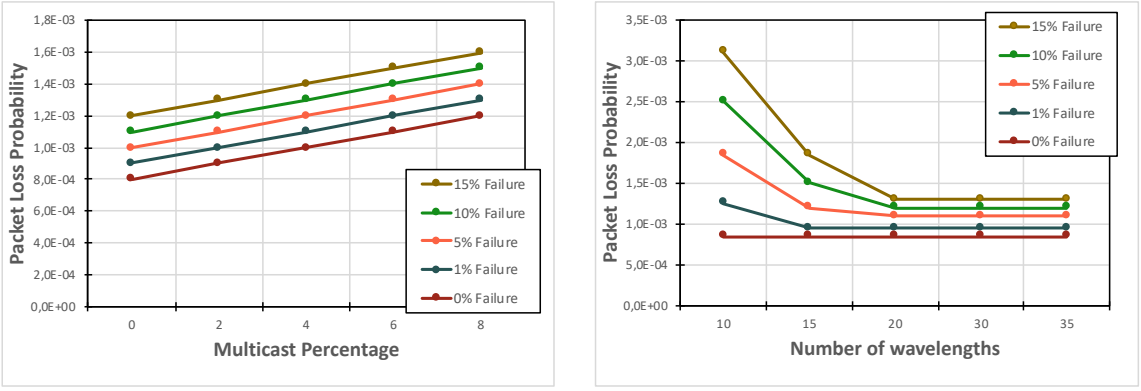
*Figure 2. Packet loss probability as a function of multicast percentage considering (a) different network loads (b) different racks per cluster*

As it can be seen in Fig. 2(b), increasing the number of racks per cluster causes more traffic load in the network which leads to more packet loss as in the previous case.

### 4.3 Failure scenario

Figure 3 analyses the case of failure. We consider that any link, OPS or OCS can fail with a given probability at the start of the simulation (but the connectivity within and between the clusters is always guaranteed). For each given failure probability, we run 5 different executions and calculate the average packet loss probability. Figure 3 (a) shows the packet loss probability as a function of the multicast percentage and the failure probability. As we can observe, the network experienced more losses if the failure probability increases. Nonetheless, the losses

between the working scenario (0% failure) and the worst considered case (15% failure) only experience a little increase; for example, from 0.8 $10^{-3}$ to 1.2 $10^{-3}$ when no multicast traffic is considered. In Fig. 3 (b), we analyse the effect of adding more WDM to the links. We can observe that there is a point where adding more wavelengths do not decrease the packet losses. This analysis provides a guideline for the dimensioning of the network. For example, in case of no failure, more than 10 wavelengths are useless; more than 20 wavelengths with a 15% failure probability are also not needed. We can also see that (as observed already in Fig. 1), once the wavelengths are not the bottleneck, the losses are due to congestion at the links between OPS and racks.




*(a)* *(b)*

*Failure 3. Packet loss probability as a function of failure probability: (a) multicast percentage and (b) WDM*

## 5. CONCLUSIONS

We have analysed and evaluated the performance of the proposed network architecture under different conditions. The metric of interest in this performance evaluation is the packet loss probability which has an especially critical role in interactive communications and gives the means to estimate the provided quality of service. The analysis of the results shows that low packet loss probability can be achieved, included if failures occur in the network. For example, in case of failures up to 15% of the entire network, the packet loss probability is maintained to 0.16% thanks to the designed topology of the network and the routing algorithm.

For future, we plan to deploy dynamic wavelength and bandwidth allocation algorithm for having energy efficiency while upgrading our network structure techniques in order to guarantee the network performance and decreasing the packet loss probability even to lower level (near zero).

## ACKNOWLEDGEMENTS

This work has been partially funded by the Spanish Ministry of Economy and Competitiveness under contract FEDER TEC2017-90034-C2-1-R (ALLIANCE project) and supported but not funded by the Generalitat de Catalunya under contract 2017SGR-1037.

## REFERENCES

[1] N. Panahi, D. Careglio, J. Solé-Pareta, "A flexible optical network architecture providing enhanced performance to data centres", in *Proc. 18th ICTON 2016*, Trento, Italy, July 2016.

[2] B. Wang, Z. Qi, R. Ma, H. Guan, A. V. Vasilakos, "A survey on data center networking for cloud computing", *Comp. Netw.*, vol. 91, pp. 528-547, Nov. 2015.

[3] W. Xia, P. Zhao, Y. Wen, H. Xie, "A survey on Data Center Networking (DCN): infrastructure and operations", *IEEE Commun. Surv. & Tut.*, vol. 19, no. 1, pp. 640-656, Firstquarter 2017.

[4] H. Hajabdolali Bazzaz, et al., "Switching the optical divide: fundamental challenges for hybrid electrical/optical datacenter networks", in *Proc. 2nd ACM SOCC 2011*, Cascais, Portugal, October 2011.

[5] C.F. Lam, et al., "Fiber optic communication technologies: What's needed for datacenter network operations", *IEEE Commun. Mag.*, vol. 48, no. 7, pp. 32-39, July 2010.

[6] S. Leon Gaixas, et al., "Scalable topological forwarding and routing policies in RINA-enabled programmable data centers", *Trans. Emerg. Telecommun. Technol.*, vol. 28, no. 12, Nov. 2017.

[7] J. Perelló, et al., "All-Optical Packet/Circuit Switching-based Data Center Network for Enhanced Scalability, Latency and Throughput", *IEEE Netw. Mag.*, vol. 27, no. 6, pp. 14-22, Nov. 2013.

[8] K. Kitayama, et al., "Torus-Topology Data Center Network Based on Optical Packet/Agile Circuit Switching with Intelligent Flow Management", *OSA/IEEE J. Lightw. Technol.*, vol. 33, no. 5, Mar. 2015.

[9] A. Vahdat, H. Liu, X. Zhao, C. Johnson, "The emerging optical Data Center", in *Proc. OSA OFC 2011*, Mar. 2011.

[10] B. Jennings, R. Stadler, "Resource management in clouds: survey and research challenges", *J. Netw. Syst. Manag.*, vol. 23, no 3, pp. 567-619, July 2015.