MATEC Web of Conferences **41**, 01001 (2016)
DOI: 10.1051/matecconf/20164101001
© Owned by the authors, published by EDP Sciences, 2016

Exploiting Shared-Access Passive Optical Networks for Building Distributed Datacenters

H.-C. Leligou[1,a], T. Orphanoudakis[2], A. Stavdas[3], C. Matrakidis[3]

[1] *Electrical Engineering Department, Technological Educational Institute of Sterea Ellada, Psahna, Evias, 34400, Greece, leligou@teiste.gr*
[2] *Hellenic Open University, School of Science and Technology, Tsamadou 13-15, Patras 26222, Greece, fanis@eap.gr*
[3] *University of Peloponnese, Karaiskaki str., 22100 Tripolis, Greece, astavdas@uop.gr*

Abstract. We propose a novel optical cloud architecture where IT and Telecom resources are used interchangeably as common infrastructure. Key assets are the MAC controlled passive networks for distributed multiplexing and grooming and a node architecture integrating transmission and switching.

1 Introduction

Data-Centers are key-enablers of the Future Internet. The Data-Centers of companies such as Google and Facebook require a vast number of resources to operate which may spread out over a Metropolitan area. This requires a number of federated Data-Centers to act as one.

Today Data-Centers run inefficiently since the deployed IT and Telecom resources are totally decoupled whilst IT policies like virtualisation etc, are of local significance only for they are confined to Data-Center's boundaries with no global implications. Moreover, the actual volume of data that needs to be gathered, stored, processed and routed is driving the size of these Data-Centers to new levels.

The federation of several Data-Centers each one consisting of hundreds of thousands of servers, requires a joint micro and macro-engineering and a global optimization of IT and Telecom resources followed by resources slicing policies for "commoditization". Moreover, this federation is posing stringent performance requirements in terms of QoS performance (mainly latency and packet loss) in the underlying interconnection fabric. In this work, we are proposing a modular, scalable to several Petabit/s capacity, guaranteed QoS and power-efficient federated Data-Centre architecture giving rise to Optical Clouds.

In this work we describe an Optical Cloud which is an IaaS (Infrastructure as a Service) that can: host and efficiently handle a gigantic pool of cloud resources regardless of their physical location; Offer high agility, flexibility and programmability, Guarantee Quality of Service (QoS) performance and in particular, ultra-low latency services; Integrate low CapEx/OpEx systems since cost and power consumption is currently an important bottleneck in current DC and Telco systems.

2 Optical clouds

A key element of an Optical Cloud is the Optical Cloud Interconnection Node (OCIN) which is the node where the IT and Telco systems are integrated. This is an "one-for-all-purposes" ICT infrastructure aiming to serve as both DC and a Core/Metro node at the same time: the IT and Telco services are provided in parallel and they can be dynamically engaged by slicing different portions of IT and/or Telco resources within a given OCIN. By building and integrating the Cloud resources directly onto the optical footprint of a Telco infrastructure, the Cloud functions and services are distributed. This fusion of IT and Telco systems, allowing the processing capabilities and functions to be distributed across the Core/Metro network, manifests a change of paradigm, which is not limited to this network segment only. Actually, we witness the same process spreading across the network: CORD [1] is re-architecting the Central Office in the aggregation network to serve as a DC, whilst HYDRA [2] is making one step further bringing a micro-DC even closer to the end-user which is now placed at the location of a DSLAM or at the Cabinet/Local-Exchange. In fact, the latter is also abolishing the entire aggregation network segment that is replaced by pure transmission facilitating to the Access-Core integration for an enhanced performance from the end-user's perspective.

Regarding systems architecture, today large DCs typically contain 100,000s servers which are interconnected to form a massively parallel super-computing infrastructure making use of Core/Aggregation switches [3] adopting the so-called Spine-and-Leaf topology which is schematically shown in Figure 1. Currently, the connectivity requirements between racks increase at an astonishing rate

[a] Corresponding author: leligou@teiste.gr

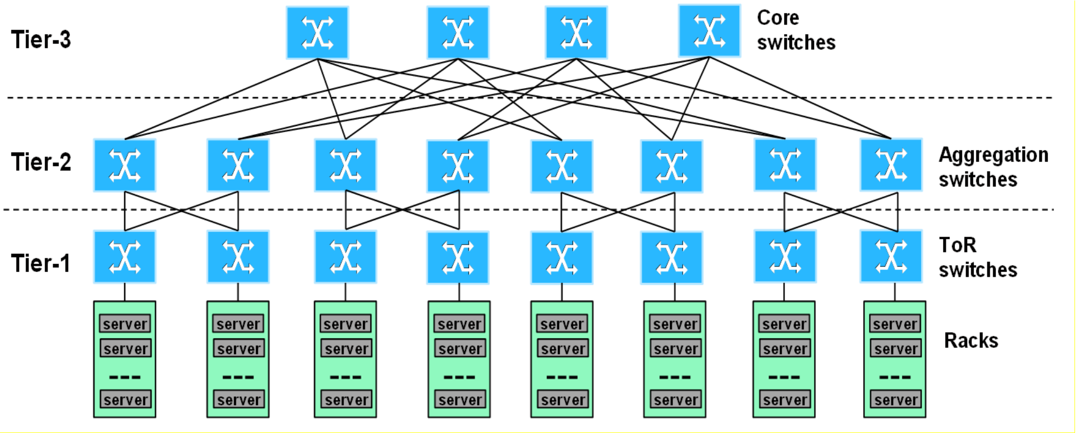

Figure 1. A schematic illustration of DCs Spine-and-Leaf architecture

e.g. 20 times every 4 years, outpacing Moore's Law. In particular, physical layer propagation limitations (loss, parasitic phenomena and capacitance) limit high bit-rate interconnections due to restrictions in the induced power-times-bandwidth product; currently 10 Gb/s Ethernet is adopted as the main intra-DC interface. As it is evident from Figure 1, a considerable number of electronic switches need to be cascaded and crossed over for transporting packets between servers something that is increasing the total DC cost whilst degrading performance, mainly latency. The current trend is the data to be stored, archived or processed exploiting virtualised resources that are distributed over an undefined number of Data-Centers. To accomplish these functions, the Data-Centers are relying on large, power consuming and immensely expensive electronic switches and routers for their data-plane packet forwarding with optical solutions restricted to providing a clear channel for a point-to-point bit-by-bit transmission.

To delegate higher functionality to the optical layer, our solution is adopting an architecture that relies more on transmission and multiplexing and less in switching. Moreover, it is adopting a layered architecture, consisting of three Tiers, to ensure scalability and manageability as shown in Figure 2. The IT and Telecom resources are managed from programmable controller interfaces adopting the SDN approach so as to design a fully programmable and open solution.

The details of the three Tiers are analyzed below:

Tier-One: The purpose of this layer is to provide connectivity between the Racks hosting the IT/NFV or the Telco processing and storage systems, to the node residing at the interface between Tier-1 and Tier 2 in Figure 2 that is called High-Node (HN) hereafter. This node in many respects operates as a backplane for all Tier-1 systems. The connectivity fabric in Tier-1 depends on the type of h/w, the traffic profiles these h/w systems are creating and the services that needs to be provided. The h/w types in Tier-1 can be: NFV is aiming to consolidate many network equipment types onto IT industry's high volume standard server hardware, switches and storage. The IT category also includes storage nodes (IO-nodes), small size Ethernet switches at Top-of-Racks (ToRs), intra-Rack Storage Area Network (SAN) switches, management boxes, etc. An IT/NFV Rack consists of a number of these nodes, typically 48 and up to 128 depending on the supported application. The traffic exchange between IO-nodes for storage synchronization purposes has a relatively low-rate profile with frequent/ constant exchanges. On the other hand, the ToR (Top of Rack) aggregates traffic from V-nodes and for this reason manifest high spatial and temporal traffic asymmetry.

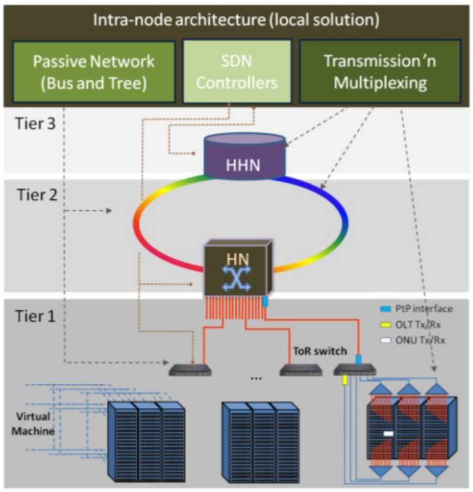
Intra-node architecture (local solution)
Passive Network (Bus and Tree)
SDN Controllers
Transmission'n Multiplexing
Tier 3
HHN
Tier 2
HN
Tier 1
PtP interface
OLT Tx/Rx
ONU Tx/Rx
ToR switch
...
Virtual Machine

Figure 2. The proposed layered architecture

With respect to the need for an open-utility OCIN, the overall traffic profile stemming from a Rack strongly depends on the ratio between the V-nodes and IO-nodes in the Rack and the traffic volume each node type is producing, so the final outcome depends on the application.

Our ambition is to obtain a universal node that dynamically changes role i.e. to serve more as a "Data-Centre" or more as "Telco node" as a result of the high-level decisions of the Global Orchestrator. To do so we should be able to dynamically and interchangeably resort to either the IT/NFV or the Telco h/w placed in OCIN, or to add new h/w, in a straightforward way without complex multi-protocol operations and without scalability problems.

Architecture for connectivity according to the needs: The Tier-1 connectivity should be tailored to the application and the particular implementation policy set by the Global Orchestrator. So, the connectivity fabric is either a tree-topology PON or point-to-point link (ptp) in parallel, as illustrated in figure 2. A number of access PON solutions is possible; a 40/100 Gb/s symmetrical operation at short distances and the low cost/power-consumption target is viable with existing technology and can fit the above requirements.

Thus, the proposed OCIN introduces a number of innovations:

a) *Alleviates bottlenecks:* The Server-to-ToR connectivity is reaching silicon switching limits. The use of PONs is steering away from over-loaded Leaf-and-Spine solutions to a protocol agnostic platform scalable to high throughput.

b) Makes large savings: The use of PON reduces the component inventory; a large number of transceiver pairs and ports in HN are saved compared to the alternative of deploying exclusively ptp links.

c) *Uses resources efficiently:* With an unspecified mixture of V-nodes, IO-nodes and proprietary Telco h/w, a traffic profile with high spatial and temporal asymmetry is expected. A PON allows using the available resources efficiently by dynamically allocating bandwidth to “hot” Racks at the expense of the relatively idle ones. Here, the statistical multiplexing gains of a PON can be substantial.

d) Optimized to Rack connectivity requirements: The point-to-multipoint connectivity requirements of the IO-nodes are ideally served by a PON in the downstream direction. For VM migration, the inbound OCIN traffic exploits the broadcasting capability of a PON to reach all Servers/Racks the PON may span. Likewise, the inter-Rack traffic reaches the HN using a single upstream channel.

e) *A distributed multiplexer:* To provide the aforementioned functionality, the PON relies on a DBA mechanism implemented by the MAC protocol guided by the SDN control plane [4]. Its design takes specific application requirements into account.

f) *Lower cost/power-consumption:* Burst-Mode (BM) PON technology is primarily considered. Given the short distances between the terminals, so options previously excluded from Telco access networks may find their application here. Finally, powering on/off ONUs can be used for power consumption savings.

Why this system is modular: New IT/NFV or Telco Racks are either directly interfaced to existing PON-branches (via free ONUs) or to a new PON interface on the HN. This approach is overcoming the classical architectural server connectivity restriction associated to speed, port and protocol limits of electronic switching systems.

Why the IaaS is possible: The proliferation of the IaaS framework requires a ubiquitous, transparent and technology-independent data-plane allowing the seamless deployment of third-party equipment and/or it is open to the selective employment of the available resources (slicing). The combination of a passive infrastructure with DBA regulated from a MAC and its independence to protocols and the system modularity ensures that the IT/NFV and Telco systems can be directly interfaced supporting interoperability between different vendor's equipment.

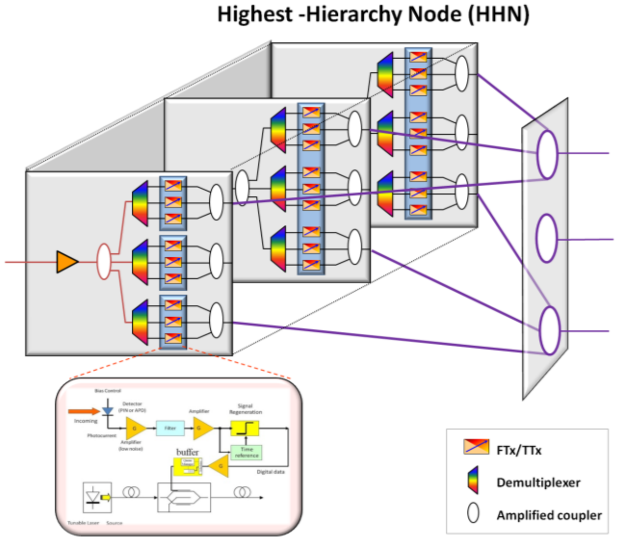

Figure 3. A schematic layout of the HN

Tier-Two: This is the aggregation layer of the solution where a number of PONs systems are interfaced to the HN. The HN in figure 3 consists of a) the optical interfaces and b) a L2 switch. Apart from the L2 switch, the HN may also include: L3 addressing /lookup and/or a Layer 3 gateway; a groups of cards built upon ASIC/commodity hardware implementations to incorporate data-base functions, or video encoders or security encryptors and/or that are cards providing BRAS functionality or Mobile Gateway Functionality. To forward packet flows within Tier-2 under guaranteed QoS performance, the HNs are interfaced to a passive optical bus, hereafter identified as ring. The upstream/downstream traffic is carried out using different fibres to minimize the number passive splitters in the path. The HNs in a ring are interconnected via the Highest-Hierarchy Nodes (HHN) using a common pool of WDM channels that are TDMA-shared between the HNs following the approach of which provides for optimized QoS performance.

Tier-Three: The rings with their HNs and the HHN are forming up a cluster which is either part of a Data-Centre, or a self-contained one, and is interconnected to the Internet and/or other clusters through the HHN that serves as the gateway. The case of federated Data-Centers is schematically depicted in figure 4.

We propose a *switchless* and *bufferless* HHN architecture as illustrated in figure 5. The slot forwarding in the HHN is done as follows: The WDM comb in the upstream fiber is entering the amplified coupler. With N rings in the Data-Centre, the coupler should have N ports, i.e., a *1:N* coupler. At the exit of the coupler the multiple copies of the original WDM comb are forwarded to a “WDM card” consisting of a demultiplexer and an array of the transceivers with a fixed-receiver (burst mode), a full 3R regeneration stage and followed by a tunable transmitter and a passive multiplexer materialized via a *M:1* amplified coupler; *M* being the number of WDM channels.

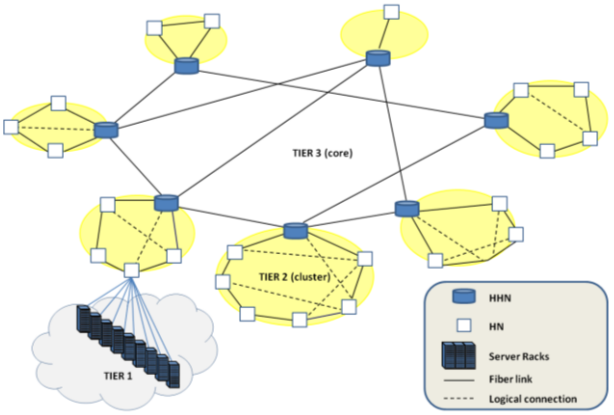

Figure 4: A number of federated Data-Centres

Apparently, the HHN includes no switching elements and no packet processing hardware. The architecture allows for broadcasting and it is wavelength and link modular. In fact the modularity at subsystem and system level is an important asset it allows WDM cards with different technology to co-exist.

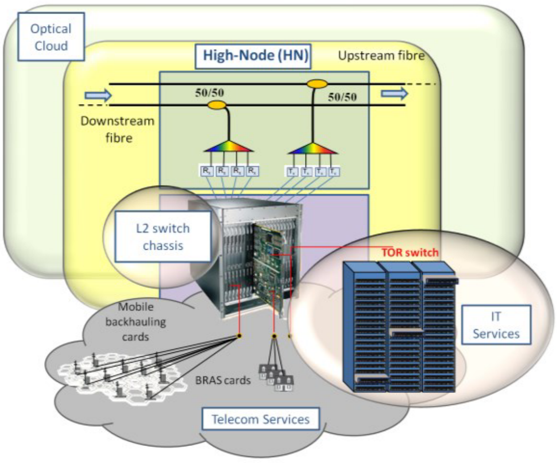

Figure 5. The proposed HHN architecture

3 Concluding evaluation remarks

A summary of the proposed innovations are the following: a) two different passive network topologies are introduced to implement distributed multiplexing and grooming so large electronic switches are obviated; b) the HHN is switchless and bufferless that integrates transmission and switching since the operations are used interchangeably and c) an end-to-end arbitration mechanism that allow the Data-centre fabric to operate efficiently under guaranteed QoS performance.

Thanks to the modular architecture, the fabric can smoothly scale from few hundred Gb/s to several Pb/s throughput while the pay-as-you-grow approach minimizes the up-front capital outlay without performance compromise. The overall operations limits buffering to HN's only with no buffering between HNs belonging either to the same or distant cluster and, as such, any delay in slot forwarding is completely minimized.

The use of a protocol-agnostic transmission as the main means of a packet forwarding fabric also paves the way to integrate IT (Data-centre) and Telecom (Core node) infrastructures where are use of an SDN-based control plane to coordinate both will allow a global resources handling policy and third party applications.

Finally, to demonstrate the scalability of the architecture, figure 6 shows the BER performance of a HHN with a variable number of Tier-2 rings, each consisting of 10 HNs. The scenarios displayed are with 40 WDM channels at 100 GHz spacing carrying 40 Gb/s traffic and 80 WDM channels at 50 GHz spacing carrying 10 Gb/s traffic. More than 40 rings can be interconnected, even without FEC in all scenarios, and with FEC the number is more than 100, even for 40 Gb/s transmission.

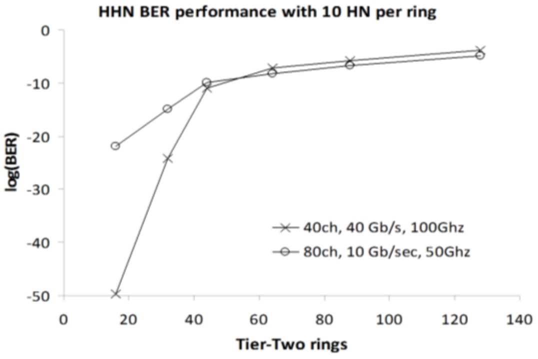

Figure 6. BER performance of HN/HHN combination

ACKNOWLEDGEMENT
This work was partially supported by the FP-7 IDEALIST project under grant agreement number 317999 and the European Union (ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the (NSRF) Research Funding Program: ARCHIMEDES III.

References

1. SDN/NFV whitepaper, "CORD Fabric: Proof of concept demonstration", Open Networking Summit, June 2015
2. C. Matrakidis, T.G. Orphanoudakis, A. Stavdas, "HYDRA: A Scalable Ultra Long Reach/High Capacity Access Network Architecture Featuring Lower Cost and Power Consumption", Journal of Lightwave Technology Vol. 33 (2), 339-348, Jan. 2015
3. C. Kachris, I. Tomkos. A Survey on Optical Interconnects for Data Centers. IEEE Communications Surveys & Tutorials, Volume: PP , Issue: 99, 2012 , Pp: 1 - 16
4. S. Peng, et al., "A novel SDN enabled hybrid optical packet/circuit switched data centre network: The LIGHTNESS approach," IEEE European Conference on Networks and Communications (EuCNC), Bologna, Italy, June 2014