Virtual Resource Consolidation in the Edge for 5G Networks

P. Kokkinos[1,2], A. Kretsis[1] and E. Varvarigos[1,3]

[1] Electrical and Computer Engineering Department, National Technical University of Athens, Athens, Greece
[2] Department of Computer Engineering, Technological Educational Institute of Peloponnese, Sparta, Greece
[3] Department of Electrical and Computer Systems Engineering, Monash University, Melbourne, Australia

***Abstract*—The shift of the radio processing to the cloud, through cloud Radio Access Networks (C-RAN) technologies and of the cloud processing to the edge, through edge computing, form the environment in which 5G systems are being implemented, fostered and transformed from a future technology to a mainstream one. By its nature, global optimization of the edge resource deployment cannot by easily performed considering the number and the diversity of the players that will be involved in the edge computing arena. As a result, building and maintaining more and more edge-located resources for serving radio and application data will eventually lead to increased cost and energy consumption and resource underutilization. One way to overcome this predicament, is through virtual resource consolidation, where separate but efficiently interconnected edge resources appear as a single computing entity, serving radio and application data. In this context, we present the Virtual Elastic Datacenters (VEDC) in the edge notion for 5G networks that can alleviate these issues. We also describe an Integer Linear Programming (ILP) based mechanism for the placement of baseband and application processing loads in a VEDC-based environment, and perform respective experiments. We show that through VEDC resource consolidation better quality services can be provided, while improving resource efficiency.**

Keywords—edge computing, 5G, resource consolidation, processing load placement

I. INTRODUCTION

The central office (CO) was, for decades, the main "local" networking facility of the operator, where data arrived and were switched to/from the end-users and devices. The emergence of 5G networks and the promise for increased performance in terms of capacity, latency and other parameters, leads to the evolution of the COs and their enhancement with new services and responsibilities. In cloud Radio Access Networks (C-RAN), the baseband processing for many cells becomes centralized and is moved from the cell to the CO and to the cloud, leading to performance improvements through cell coordination and efficiency in terms of throughput, energy and cost [1][2][3].

In the cloud arena, cloud providers utilize a number of hyperscale datacenters deployed in several locations around the world, interconnected over private and public networking infrastructures [4][5], with the latency experienced by users varying according to the datacenter's location and networking distance. On the other end, edge computing offers ultra-low latency and high bandwidth cloud-computing services, by deploying computing and storage entities at the edge of the network. Edge computing is also viewed as an enabler for the realization of the 5G use cases for IoT, transportation, robots, smart-grids and other areas [6][7][13], by handling both baseband and application processing loads. In this way, the distinction between the notion of a CO and that of an edge deployed micro-datacenter almost disappears. M-CORD [8] is an open source reference solution for carriers deploying 5G mobile wireless networks, which provides a cloud-based solution located at the edge, for executing virtualized RAN functions (baseband unit - BBU), and implementing virtualized mobile core (Virtual Evolved Packet Core - vEPC) and mobile edge services (for caching, security, billing etc). Today, several companies are offering edge computing products [9].

The deployment of edge computing/storage entities is expected to be driven by network operators, content providers but also by other players, like owners of buildings, stadiums and shopping malls that will operate edge infrastructures in order to serve their own computing needs. As a result, the global optimization of the edge resources deployment cannot by easily performed due to the plethora of edge computing facilities and the diversity of the involved players. This will at some point lead to resource fragmentation and underutilization, as well as increased energy consumption and reduced scalability and resiliency.

To address these issues, we consider a new paradigm of edge computing infrastructure based on the dynamic and virtual consolidation of edge/CO resources, sufficiently interconnected in a programmable way, forming Virtual Elastic Datacenters (VEDC) in the edge, and serving both radio and application loads from multiple cells. This is in contrast to the fixed cell-CO assignment considered in most related works. A VEDC-based environment will overcome the isolation and underutilization of edge cloud resources and provide on demand, low latency, scalable and resilient processing and storage capacity on the edge. In this paper, we explore this virtual resource consolidation vision for edge computing and 5G, and identify the efficient networking interconnection and the volume of baseband and application traffic as critical for its realization. We also present an Integer Linear Programming (ILP)

978-1-5386-6009-6/18/$31.00 ©2018 IEEE

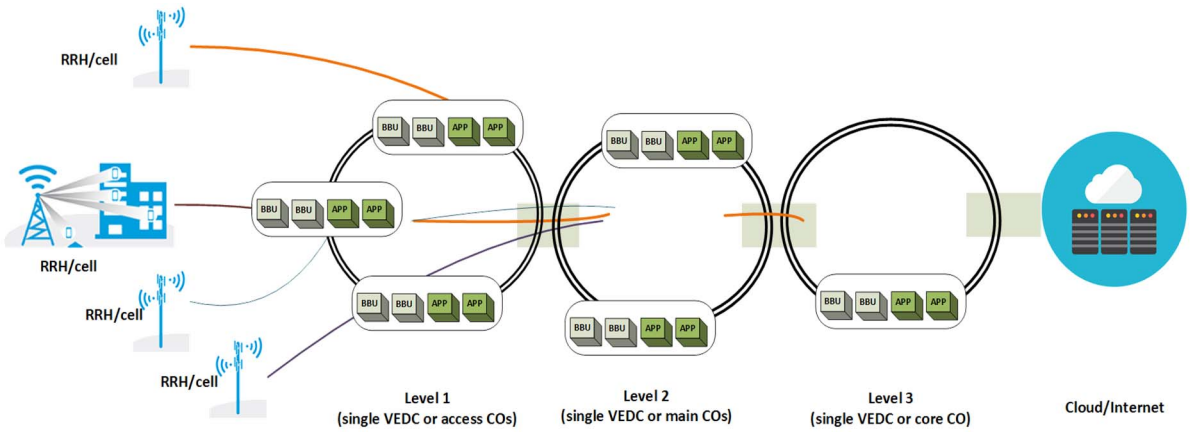

Figure 1 A multi-level VEDC-based architecture for serving baseband and application processing loads.

formulation for the assignment of baseband and application loads to the VEDCs and to different level. The ILP formulation aims at maximizing the size of the application load being executed in the VEDC-based architecture. This is subject to constraints on each level's networking and computation capacities, for different sizes of the respective baseband and application traffic. We show that through this edge resource aggregation, better quality services can be provided, while resource utilization and efficiency are also improved.

Figure 1 illustrates an example multi-level VEDC-based architecture. Baseband loads (different colored-lines) originating from various cells are processed from BBU units, while the processing of the application data these loads carry, is performed by APP units. Usually, each cell has a fixed assignment to a particular CO at each level: Level 1 has three COs, Level 2 has two COs and Level 3 has one, serving four cells. The different levels match the way metro/aggregation networks are typically organized, e.g., with access COs, main COs and core CO [1], with many (edge) resources of smaller capacity operating at lower levels and few resources with larger capacities at higher ones. As shown in Figure, in our proposed VEDC-based approach, the resources (BBU and APP units in the COs) are consolidated/aggregated in a single VEDC or in multiple ones at each level, serving both baseband and application loads from multiple cells. The proposed resource aggregation scheme has more impact on the lower levels of the hierarchy.

The remainder of the paper is organized as follows. In Section II we report on networking and traffic issues related to 5G. In Section III we describe the VEDC vision. The ILP-based load assignment mechanism and the respective simulation results are presented in Sections IV and V, respectively. Finally, Section VI concludes our paper.

II. Networking and Traffic Flows in 5G

In cloud-RANs (C-RANs), the network segment connecting the remote radio head (RRH), where the antenna is located, to the baseband processing units (BBU) pool, in the Central Office (CO), is called fronthaul and its capabilities and efficiency are key to the 5G networks success. The computing resources used for the baseband processing can by either generic resources (e.g. servers or virtual machines) [10] or specialized hardware (e.g., FPGA, ASIC) [2].

C-RAN achieves economies of scale, by having one location serving multiple cells (RRHs) with dynamic associations between RRHs and BBUs, based on traffic [3] or the coordination taking place among adjacent cells to reduce signal interference and improve efficiency [1].

The types of baseband-related interfaces and respective traffic flows identified in the C-RAN fronthaul network segment include: CPRI (Common Public Radio Interface) for encapsulating radio samples between a RRH and a BBU, S1 for transporting user and control data, and X2 for the interconnection among neighboring cells or respective BBU pools [3]. CPRI is a constant bit rate (CBR) type of traffic that requires higher data rates than the payload it carries. CPRI v7.0 bit rates range from 614 Mbit/s (Rate 1) up to 24330 Mbit/s (Rate 10) [11]. Also, CPRI has tight constraints in terms of maximum latency, jitter and bit error rate between the RRH and the BBU locations. Similarly, for S1 traffic, latencies in the order of tens of milliseconds (ms) are allowed, while X2 traffic requires delays below 1 ms in order to enable the various coordination schemes [14].

The mentioned traffic characteristics, and especially those of CPRI, make necessary the use of high capacity dedicated network resources in the fronthaul. In addition, the use of Multiple-Input and Multiple-Output (MIMO) technologies results in even larger baseband traffic flows, since multiple transmit and receive antennas are simultaneously used.

An option being investigated in order to meet the high capacity and low latency requirements in the fronthaul is the use of various functional splits of the operations performed in the RRH and the BBU [2][12], ranging from the traditional fully distributed RANs, where all processing takes place at the antenna's

location, to the fully centralized C-RAN, where all processing is performed at the BBU. A number of such functional splits have been identified, each one defining different trade-offs, related to processing, bandwidth and latency. In the same context, a modified version of the CPRI protocol, namely eCPRI [11], supporting the concept of functional splits, has been recently presented. eCPRI promises ~10 fold reduction of the required bandwidth, scaling flexibly according to the user plane traffic, while also it can utilize packet-based technologies (e.g., Ethernet). In particular, the packetization of the fronthaul traffic has many benefits in terms of traffic handling but its application is up against a number of difficulties, such as the overhead of the packetization and de-packetization process [12].

The efficient consolidation of edge resources and the realization of the Virtual Elastic Datacenters (VEDC) vision, considered in our work and presented in the next Section, requires a high capacity and flexible substrate network. We expect that such a network will be organized in a number of levels (see Figure 1) to also match the different computing (of the edge devices) and networking capacities, latencies and locations of the equipment. The baseband processing loads will be able to move in any level of this hierarchy, assuming that the respective baseband traffic is comparable in size with the application traffic it carries. New CPRI standards, such as eCPRI, various functional splits, and the packetized fronthaul indicate that a reduction of the baseband traffic relative to the application traffic, will be possible in the future. Also, other directions in support of the VEDC vision is the multiplexing of fronthaul with other kind of traffic (e.g., backhaul) and the integration of the respective network segments into a unified network substrate, driven by software defined networks and network function virtualization [15].

Note that the ring topologies illustrated in Figure 1 are just an example, and any efficient (e.g., PON in the case of optical) interconnection network could be used to connect the edge computing facilities belonging to the same level. The backhaul networks (i.e., the network segment connecting the BBU pools to the internet) are usually based on optical metro and core technologies and architectures [16].

III. Virtul Elastic Datacenters for 5G

Edge devices range in size and capabilities: modular data centers in shipping containers, micro datacenters, specialized computing devices (FPGA, GPU), IoT computing devices (e.g., Arduino, Raspberry Pi), or even computing devices combined with heaters [17]. Their placement, usually very close to the end users, makes them attractive for use in 5G networks. However, the profound variability of the edge resource's characteristics, the fact that these may operate under various administrative domains (from operators, to building owners) and their unpredictable status may result into many of these edge resources being inefficiently used when serving the 5G baseband and processing loads.

To address the problem, we present the notion of Virtual Elastic Datacenters (VEDCs), a new paradigm of edge computing in order to help support highly demanding and dynamic 5G applications, such as those that involve movable data sources (e.g., cars, drones) that are sensitive to latency (Figure 2). VEDCs can be used to reduce the complexities raised by the diversity of the edge resources but also to improve the overall performance and efficiency for 5G and edge computing, executing both baseband and application processing loads. VEDCs are composed of shared edge computing and storage resources of various sizes and characteristics that are consolidated into bigger entities. VEDCs abstract the complexity and the dynamicity of the underlying (cloud and network) infrastructures, while improving the efficiency in the use of the resources. Scalability and resiliency are achieved by transparently adding or updating edge and network resources in a VEDC (Figure 2), as demands in cells change (e.g. more cars, drones and users generate traffic) or in case of failures. We may assume that the resources illustrated in Figure 2 correspond to Level 1 resources, with respect to Figure 1, with more than one however VEDCs.

Also, VEDCs fit in the multi-layer architecture described in Section II and presented in Figure 1, where smaller, more dynamic (due to the dynamicity of their constituent edge resources) but also offering lowering latencies, VEDCs, are in the lower layers, with their size and stability increasing at higher layers along with the provided latency. Each layer can have multiple VEDCs, while the number of layers and the capabilities of each one may differ based on various performance characteristics or the capabilities of the edge devices and the networking infrastructure interconnecting them. Also, under this VEDC based operation there is not a computing facility dedicated to particular cells as the BBU hotel, but instead any VEDC with any characteristics (static and dynamic) can be utilized for this purpose, e.g. based on the latency constraints or the current loads.

Figure 2 The VEDC paradigm of edge computing in 5G networks.

Of course, the realization of the VEDC vision requires a number of components to be in place, including the actual edge resources, which will additionally have to be highly and flexibly (through programmable networks) interconnected so as to perform seamlessly as a single virtual datacenter. A VEDC orchestrator is also necessary for the creation and the updates of VEDCs, for their multi-layer operation and the overall joint management of the resources and of the workloads. These are in accordance with the 5G roadmap for the integration of networking, computing and storage resources into one programmable and unified infrastructure [18].

IV. Load Assingment in VEDC-based Edge

The baseband and application processing is constrained by the available computation and networking capacity. In a VEDC-based environment, resources in each level are consolidated and shared between different cells. So, the proposed VEDCs would relax the computational constraints compared to the fixed cell/RRH-CO assignment (Figure 1).

Also, baseband and application loads can be flexibly assigned to the various resources, in the same and at different levels. The network constraints affect the traffic flows that can move upwards the hierarchy (Figure 1). In particular, if the baseband and application traffic loads are comparable in size, then it will be possible that some of the baseband processing (originating from a cell/RRH) is moved up in the hierarchy in favor of serving (in lower levels) high priority applications (originating from other cells/RRHs) that require low latency. Also, serving application traffic reduces the traffic load that is carried to higher levels or even to the internet and the cloud.

Under these considerations, in what follows we present an ILP formulation for assigning baseband and application processing loads in a multi-level VEDC-based architecture, so as to maximize the size of the application loads, originating from various cells, being executed.

Table I – Processing load assignment in VEDCs, ILP formulation

ILP formulation

Input

C: Number of cells/RRH

L: Number of VEDC levels

P_c: Aggregated processing load from cell c

N_c: Baseband processing load from cell c

G: Array of percentage divisions of application processing load

Q: Percentage of network traffic per application processing unit

R_l: Aggregated VEDC processing capacity at level l

W_l: Network capacity at level l

vc_{max}, vc_{min}: maximum and minimum percentage of application load from cell c executed in the VEDC layers

F: maximum number of levels where application processing load can be executed

Ic: CPRI traffic network load originated from each cell c

Variables

S_{cgl}: Boolean variable if cell's c application processing load is executed in percentage $g \in G$ in VEDC-level l

D_{cl}: Boolean variable if a cell's c BU processing load is executed in VEDC-level l

Objective

$$\max \left(\sum_{c=1}^{C} \sum_{g=1}^{|G|} \sum_{f=l}^{L} S_{cgf} \cdot P_c \cdot G[g] \right)$$

Constraints

$$\sum_{g=1}^{|G|} \sum_{l=1}^{L} S_{cgl} \leq F, \text{ for all } c \in [1, \dots, C] \quad (1)$$

$$\sum_{l=1}^{L} D_{cl} == 1, \text{ for all } c \in [1, \dots, C] \quad (2)$$

$$\frac{P_c - \sum_{g=1}^{|G|} \sum_{f=1}^{l+1} S_{cgf} \cdot P_c \cdot G[g]}{Pc} \geq 1 - \sum_{f=1}^{l+1} D_{cf} \text{ , for all } c \in [1, \dots, C] \text{ and for all } l \in [1, \dots, L-1] \quad (3)$$

$$vc_{min} \leq \sum_{g=1}^{|G|} \sum_{l=1}^{L} S_{cgl} \cdot P_c \cdot G[g] \leq vc_{max} \text{ , for all } c \in [1, \dots, C] \quad (4)$$

$$\sum_{c=1}^{C} \sum_{g=1}^{|G|} S_{cgl} \cdot P_c \cdot G[g] + \sum_{c=1}^{C} D_{cl} \cdot N_c \leq R_l, \text{for all } l \in [1, \dots, L] \quad (5)$$

$$\sum_{c=1}^{C} \sum_{g=1}^{|G|} \sum_{f=l}^{L} S_{cgf} \cdot P_c \cdot G[g] \cdot Q + \sum_{c=1}^{C} \sum_{f=l}^{L} D_{cf} \cdot I_c \leq W_{l-1}, \text{ for all } l \in [2, \dots, L] \quad (6)$$

In the above ILP formulation, G is an array of discrete possible percentages or fractions (e.g., 0.15, 0.25, 0.35, 0.45) of application processing load executed in any level. Each entry of G may represent part of an application's total workload or the processing load of a particular Virtual Network Function (VNF) for that application. A number of VNFs for various 5G applications (IoT, video services, traditional broadband) and their static placement in the edge and in the core cloud is illustrated in [19]. The objective of the presented ILP formulation is to maximize application load being executed in the VEDC-based architecture; other objectives can also be considered, such as minimizing the latency experienced by applications. The values taken by the variables in the ILP solution indicate the level in which baseband and application loads are being processed. Also, constraint (1) ensures that a cell's aggregated application processing load cannot be executed in more that F VEDC-levels. Generally, a large value for F increases application related transfers, but also in practice increases and complicates application management operations. Constraint (2) dictates that a cell's baseband processing load must be executed at a single level and single VEDC, assuming each level is a VEDC. In constraint (3) it is specified that a cell's baseband processing load will be executed in the same or a lower level than the lowest execution level of the respective cell's application processing loads. Constraint (4) ensures that the percentage of application processing load being

executed for each cell *c* is between the min and max values received as input. Constraint (5) specifies that the overall processing application load executed in each level cannot by larger than the available capacity. Finally, Constraint (6) ensures that the network capacity constraints at each level are also satisfied.

V. EXPERIMENTS

We performed a number of simulation experiments, implementing the above ILP formulation and respective scenario, using the python Pulp library and utilizing the CPLEX LP/MIP solver.

In our experiments, we consider a number of cells/RRHs whose baseband and processing load is served by a three level infrastructure (as in Figure 1). The computational and networking capacity increases in each subsequent level, representing an environment where multiple low capacity edge resources exist near the end users and the cells, while higher capacity ones operate in more central locations of the network. This of course is reflected by the consolidated capacity of the respective VEDCs, assuming one VEDC at each level. The main parameter considered in the experiments was the network load of the baseband traffic (CPRI traffic) versus the size of the application-related traffic.

Figure 3 shows the number of baseband processing loads (BU) executed in different levels of the hierarchy, parametrizing the size of the baseband traffic versus the respective application traffic. The BU loads are indivisible and there is one-to-one relation of such loads with their respective cells. The application computing loads (AP) are actually the percentages of application processing loads originating from various cells and correspond to different users or application processing operations (e.g. in the form of Virtual Network Functions). When baseband and application network traffic flows are of similar size (1x) then there is a lot of flexibility, meaning that applications loads can be executed in lower levels so as to reduce the respective application latency, while baseband loads (belonging to different cells) can also be executed in higher levels, assuming that the latency constraint of the application traffic these carry is not so strict. However, as the baseband network traffic becomes larger in relation to the application traffic (10x, 100x) then the referenced flexibility is reduced, and baseband loads are executed in lower layers, while few applications loads of larger size (Figure 4) are executed in higher levels. This is because by executing baseband loads in lower levels, the resulting application loads (that originate from the same cell) have lower traffic sizes that can be served/transferred by the network. The latter cases (>10x, 100x) in particular correspond to the basic CPRI protocol, while the former cases (1x, <10x) form the target of eCPRI and of other related methods for transferring baseband traffic (e.g., using packetization).

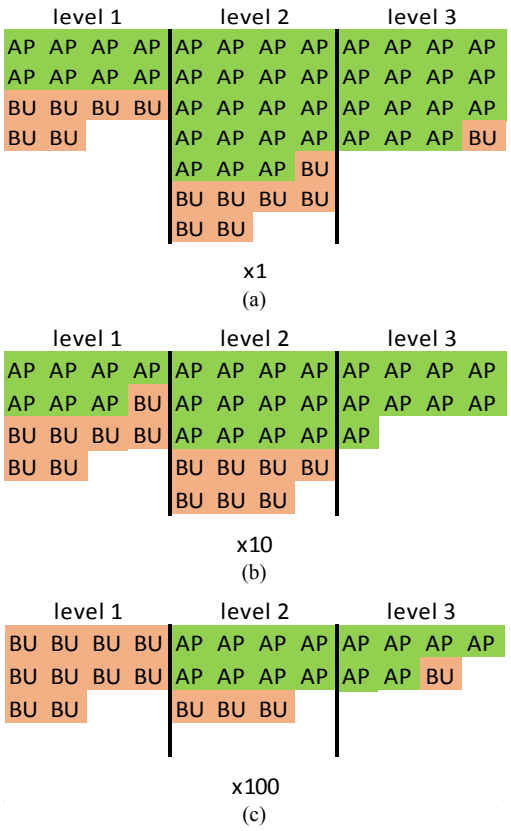

Figure 3 Baseband (BU) and application (AP) processing loads executed in the levels of the VEDC-based architecture under various baseband versus application traffic size ratios: 1x, 10x, 100x.

Figure 4 shows the total size of baseband and application processing loads that are executed in all levels, parametrizing again the size of the baseband traffic versus the respective application traffic. Figure 4 actually shares similar information with Figure 3. In all cases the baseband processing loads executed remain the same since these must be executed in the edge computing infrastructure, while fewer application loads in size (Figure 4) and in number (illustrated as the #app loads line in Figure 4 and as AP boxes in Figure 3) are executed. This means that the high volume baseband network traffic leads in prioritizing baseband loads.

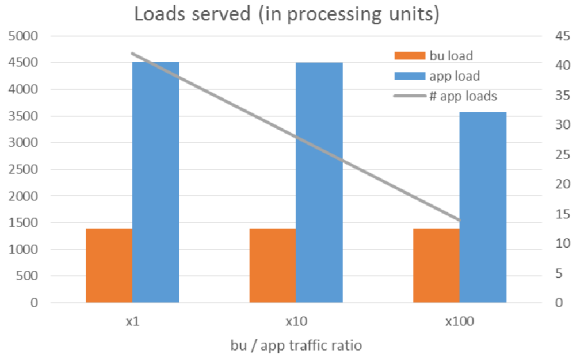

Figure 4 The total size of baseband and application processing loads and the number of applications loads that are executed in the levels of the VEDC-based architecture under various baseband versus application traffic size ratios: x1, x10, x100.

The benefits of the VEDC-based operation are exhibited in Figure 5, where we compare the total load executed in the edge computing infrastructure when resources in each level are consolidated/aggregated, forming a VEDC or not. In the former case resource sharing between different cells adds more flexibility in serving the loads and increases the utilization efficiency of the resources in lower levels. In the particular experiment illustrated, in the VEDC/aggregated scenario, all loads are served by the first two levels of the hierarchy.

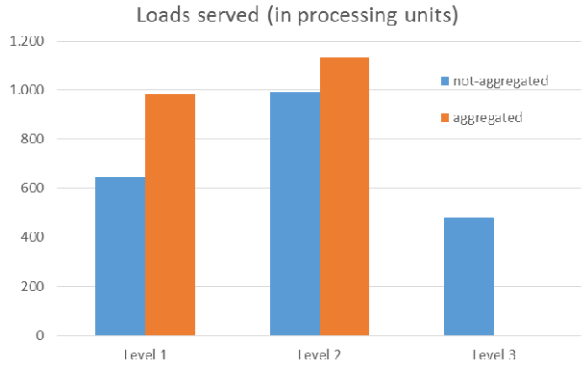

Figure 5 The total size of baseband and application processing loads that are executed in a multi-level edge computing architecture where consolidation/aggregation is applied in the form of VEDC or not.

VI. CONCLUSIONS

In this work, we presented a new paradigm of edge computing based on the dynamic and virtual consolidation of edge resources, sufficiently interconnected in a programmable way, forming Virtual Elastic Datacenters (VEDC) in the edge and serving both radio and application loads. The efficient assignment of baseband processing and application loads in a multi-level VEDC-based architecture is of major importance. We also described an ILP based mechanism for deciding such assignments and performed respective simulation experiments. We illustrated the benefits of the VEDC aggregation in terms of resource utilization efficiency and the importance of the relation between baseband and application traffic sizes, in serving the respective loads in a multi-level hierarchy. The reduction of baseband traffic's volume is the target of eCPRI, which as illustrated by our results, can increase the flexibility in using the computing resources deployed in the edge.

ACKNOWLEDGMENT

This work was partially supported by the EC through the Horizon 2020 5G-PHOS project (project id: 761989).

REFERENCES

[1] F. Musumeci, C. Bellanzon, M. Tornatore, A. Pattavina and J. T. Gijon, "Enhancing RAN throughput by optimizec controller placement in optical metro networks", IEEE International Conference on Communications, 2017.

[2] A. Tzanakaki, M. P. Anastasopoulos and D. Simeonidou, "Optical networking interconnecting disaggregated compute resources: An enabler of the 5G vision", International Conference on Optical Network Design and Modeling, 2017.

[3] L. Velasco, A. Castro, A. Asensio, M. Ruiz, G. Liu, C. Qin, R. Proietti, S. J. B. Yoo, "Meeting the requirements to deploy cloud RAN over optical networks", IEEE/OSA Journal of Optical Communications and Networking, Vol. 9, No. 3, pp. B22-B32, 2017.

[4] Hyperscale Data Center Market - Global Opportunity Analysis and Industry Forecast, 2014 - 2022", Market Research Reports, Inc, 2016.

[5] P. Kokkinos, I. Gravalos, A. Kretsis and E. Varvarigos,, "Inter-datacenter virtual capacity services: Reality and mechanisms", IEEE International Conference on Cloud Networking, 2017.

[6] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta and D. Sabella, "On Multi-Access Edge Computing: A Survey of the Emerging 5G Network Edge Cloud Architecture and Orchestration", IEEE Communications Surveys & Tutorials, Vol. 19, No. 3, pp. 1657-1681, 2017.

[7] NGMN Alliance, "NGMN 5G White Paper," 2015.

[8] M-CORD, www.opennetworking.org/solutions/m-cord/, last accessed on January 2017.

[9] 10 MEC Products That Enable Edge Computing, www.sdxcentral.com/mec/definitions/mec-products/, last accessed on January 2017.

[10] J. Tang, W. P. Tay, T. Q. S. Quek and B. Liang, "System Cost Minimization in Cloud RAN with Limited Fronthaul Capacity", IEEE Transactions on Wireless Communications, Vol. 16(5), pp. 3371-3384, 2017.

[11] CPRI, www.cpri.info, last accessed on January 2017.

[12] C. Y. Chang, R. Schiavi, N. Nikaein, T. Spyropoulos and C. Bonnet, "Impact of packetization and functional split on C-RAN fronthaul performance", IEEE International Conference on Communications (ICC), 2016.

[13] B. Rimal, D. Van and M. Maier, "Mobile Edge Computing Empowered Fiber-Wireless Access Networks in the 5G Era", IEEE Communications Magazine, Vol. 55(2), pp. 192-200, 2017.

[14] A. Asensio, M. Ruiz, L. Contreras, and L. Velasco, "Dynamic Virtual Network Connectivity Services to Support C-RAN Backhauling", IEEE/OSA Journal of Optical Communications and Networking, Vol. 8, No. 12, pp. B93-B103, 2016.

[15] F. Cavaliere et al, "Towards a unified fronthaul-backhaul data plane for 5G The 5G-Crosshaul project approach", Computer Standards & Interfaces, Vol. 51, pp. 56-62, 2017.

[16] J. M. Fabrega, M. S. Moreolo, L. Nadal, F. J. Vílchez, J. P. Fernández-Palacios and L. M. Contreras, "Mobile front-/back-haul delivery in elastic metro/access networks with sliceable transceivers based on OFDM transmission and direct detection", International Conference on Transparent Optical Networks, 2017.

[17] Nerdalize, www.nerdalize.com, last accessed on January 2017.

[18] 5GPP, Vision on Software Networks and 5G, 5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP_SoftNets_WG_whitepaper_v20.pdf, last accessed on January 2017.

[19] S. Vassilaras et al, "The algorithmic aspects of network slicing", IEEE Communications Magazine, Vol. 55(8), pp. 112-119, 2017.