Anastasopoulos, M., Tzanakaki, A., & Simeonidou, D. (2017). Optical networking interconnecting disaggregated compute resources: An enabler of the 5G vision. In *2017 21th International Conference on Optical Network Design and Modeling (ONDM 2017): Proceedings of a meeting held 15-17 May 2017, Budapest, Hungary* (pp. 174-179). Institute of Electrical and Electronics Engineers (IEEE).

Optical Networking Interconnecting Disaggregated Compute Resources: An enabler of the 5G Vision

Anna Tzanakaki[(1),(2)], Markos P. Anastasopoulos[(1)], Dimitra Simeonidou[(1)]
(1) University of Bristol, UK, (2) National and Kapodistrian University of Athens, Department of Physics, Greece

***Abstract*— This paper focuses on converged optical-wireless 5G infrastructures and proposes the novel concept of "Dis-Aggregated RAN" (DA-RAN) as an alternative to the RAN and C-RAN solutions. DA-RAN adopts the concept of "disaggregation" of hardware and software components across the wireless, optical and compute/storage domains. The performance of the proposed approach is evaluated through a service provisioning model that takes into consideration both the description of the heterogeneous 5G network and the processor details within the BBUs. Modeling results show significant benefits in terms of power consumption that range between 10-50%**

Keywords—5G network design; optimal functional split; converged optical-wireless infrastructures, resource disaggregation.

I. INTRODUCTION

The explosive growth of mobile internet traffic attributed to the rapidly increasing number of mobile users and smart devices forces network providers to concurrently support a large variety of services that can be either stand alone or interact. This introduces the need to transform traditional closed, static and inelastic network infrastructures into open, scalable and elastic ecosystems that can support a large variety of dynamically varying applications and services. This transformation needs to bring new service capabilities to network operators in terms of: *i)* connectivity for a growing number of very diverse devices, *ii)* high mobility in heterogeneous environments and, *iii)* mission critical services currently handled by specific purpose networks [1], supporting highly variable performance attributes in a cost and energy-efficient manner.

In this context, a future proof infrastructure needs to adopt a flexible architecture offering converged services across heterogeneous technology domains deploying a unified software control [1]. In these environments, where ubiquitous access and user mobility play a key role, network heterogeneity involves integration of advanced wireless with high-capacity wired network domains interconnecting a large variety of end-devices with compute and storage resources in a flexible and scalable manner. Optical network solutions can play a key role in facilitating interconnection of distributed compute and storage resources hosted by data centres (DCs) that can vary in scale (micro- to regional and mega-DCs), as they provide abundant capacity, long reach transmission capabilities, carrier-grade attributes and energy efficiency.

5G wireless access solutions will support a heterogeneous set of integrated air interfaces and will exploit contiguous and wide spectrum bandwidth including Sub-6 GHz and mmWave bands and advanced beam-tracking and MIMO techniques. These will coexist with legacy (2-3G), Long Term Evolution LTE (4G) and Wi-Fi technologies to allow broader coverage and availability, higher network density and increased mobility. To further enhance spectral efficiency and throughput, small cells can be deployed either adopting the traditional Distributed Radio Access Network (D-RAN) paradigm, where Base Band Units (BBUs) and radio units are co-located or the more recently proposed concept of Cloud Radio Access Network (C-RAN). In C-RAN remote units (RUs), are connected to the Central Unit (CU) where the BBU pool is located through high bandwidth transport links known as fronthaul (FH) [2]. Through its pooling and coordination gains, this approach can address the limitations of D-RAN, such as increased capital and operational costs, as well as limited scalability and flexibility. However, C-RAN (depending on the wireless technology adopted) may require tremendous transport bandwidth and impose strict latency and synchronization constraints. In this context, optical network solutions can play a key role offering advanced transport capabilities [2]-[3].

To address the limitations of the D-RAN and C-RAN approaches, this paper proposes a novel architecture exploiting flexible functional splits. The optimal "split" can be flexibly decided, based on a number of factors such as the transport network and the service characteristics with significant resource and energy efficiency benefits [4]. The introduction of these splits allows dividing the processing functions between the CU and the remaining baseband processing functions, through shared compute resources. The required flexibility can be provided by programmable digital hardware, able to support flexible reconfiguration of hardware-accelerated (HWA) and software-realized baseband functions, which can be partitioned at different levels to serve different Key Performance Indicators (KPIs). The shared "pool of resources" required to support this type of activities alleviates the need of owning hardware as it can be hosted either at publicly available micro-Data Centers (DCs) – referred to as Mobile Edge Computing (MEC) – or at remote regional and central large-scale DCs. This alternative RAN approach introduces the need to develop new technology solutions capable of improved performance as well as high levels of power efficiency, flexibility and density.

Towards this direction the recently proposed concept of "*disaggregation of resources*" is expected to play a key role. Disaggregation relies on physically decoupling components and mounting them on remote locations, instead of tightly coupling all components on one integrated system. Disaggregation facilitates independence across technologies and technology subsystems, offering increased granularity in the control of resources and the way they are allocated and provisioned [5]. To exploit the concept of disaggregation in RAN environments novel 5G architectures and technology solutions are needed to increase the density and power efficiency of the "pool of resources", while supporting at the same time high bandwidth connectivity between them.

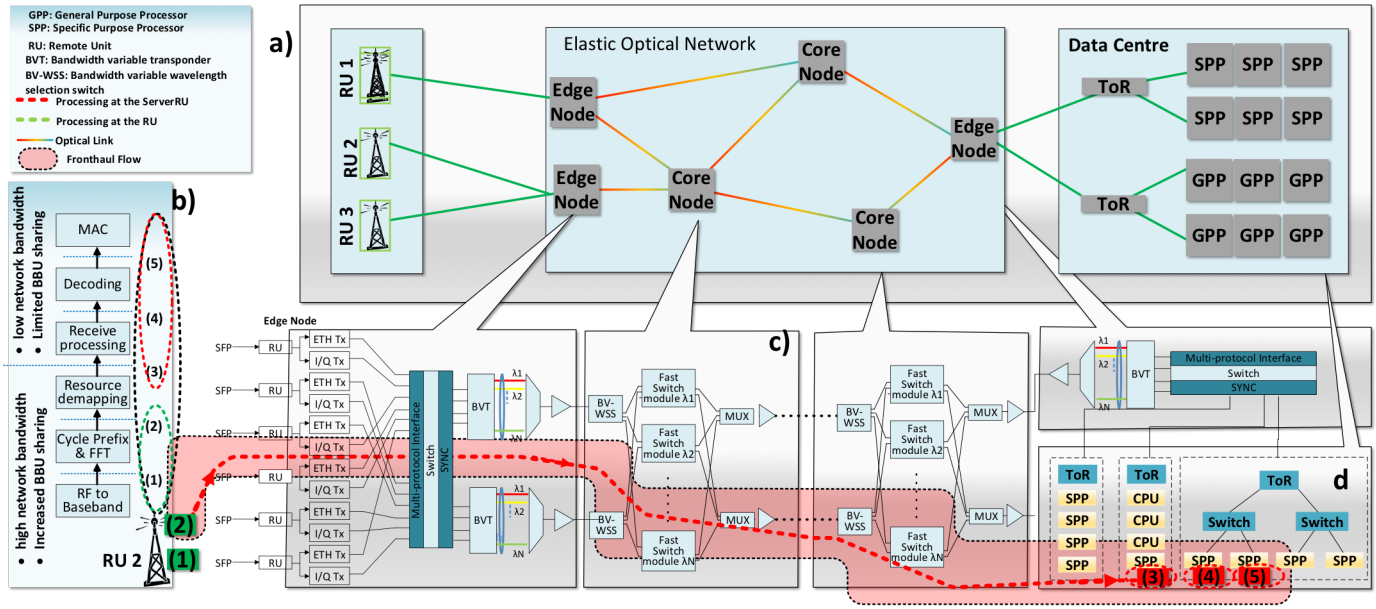

Fig. 1. a) Modelling of network components and resources: a) Multitechnology network infrastructure, b) BBU processing chain and functional split [3]-[4], c) Data path interfaces, AxC data stream generation (upper part) and multiplexing (middle) over an elastic optical metro network solution, d) Dissagregated DC network with GPP/SPP.

This paper proposes a paradigm shift, from the traditional RAN and recent C-RAN to the "Dis-Aggregated RAN" (DA-RAN) approach. DA-RAN is a novel concept adopting the notion of "disaggregation" of hardware (HW) and software (SW) components across the wireless, optical and compute/storage domains. "Resource disaggregation" allows decoupling of HW and SW components creating a common "pool of resources" that can be independently selected and allocated on demand. These HW and SW components form the basic set of building blocks that, in principle, can be independently combined to compose any infrastructure service. Apart from increased flexibility, disaggregation, due to its modular approach, offers enhanced scalability, upgradability and sustainability potential that are particularly relevant to 5G environments supporting enormous and continuously growing number of end devices and services.

To exploit the concept of disaggregation in RAN environments, novel 5G technology solutions are needed to increase the density and power efficiency of the "pool of resources", supporting at the same time high bandwidth connectivity between them. Such novel networking approaches can facilitate increased functionality and flexibility infrastructures, offering simplified management and advanced capabilities including slicing and virtualization that allow the disaggregated resource pool to be shared and accessed remotely. On-demand selection and allocation of these resources (flexible mix-and-match) will enable provisioning of any service without the prerequisite of owning and installing any specific HW or SW, adopting novel approaches such as the notion of service chaining (SC). SC can support orchestrated service provisioning over heterogeneous environments combining together a set of functions performed by different resources as appropriate [2]

To address these issues, an optimization framework is proposed that jointly minimizes the overall power consumption of the network and compute infrastructure and as such the associated operational expenditure, subject to a set of constraints including the tight FH delay requirements [2]. This is achieved by identifying the optimal functional splits as well as optimal BBU placement [7] offering minimum power consumption. In this study we assume that the BBU processing resource pool comprises a mix of general purpose processors (GPPs) (i.e. x86 CPUs, GPUs) and specific purpose processors (SPPs) (i.e., ASICS. FPGAs). These resources are hosted at regional or mobile edge DCs and are adopted to process in a parallel manner the various FH functions. In general, these functions can be mapped to GPPs or SPPs adopting either the *pipelining* or *parallel* (or *sequential*) processing mode. In the former, each processing unit handles a specific function adopting 1:1 mapping, whereas in the latter the same function is distributed across multiple processing units (1:N). To keep the analysis tractable, this study focuses on the pipelining model to support the vBBU SC, however, it can be easily extended support the *parallel* processing mode, as well.

So far several studies have focused either on addressing 5G topological design problems through the modeling energy performance of C-RAN with optical transport considerations [6], optimal BBU placement [7], impact optimal placement of microwave links for small cell backhauling [8] and optimal placement of optical network units (ONUs), Remote Nodes (RNs) and fibres or identifying optimal split options over integrated wireless/optical infrastructures [2]. To the best of the authors knowledge this is the first time that the 5G FH service provisioning problem is extended beyond the consideration of the heterogeneous network and some high level description of the associated computation requirements and addresses the details of optimal processor allocation within the BBUs comprising a combination of GPPs and SPPs. The performance of the proposed approach is examined using realistic traffic statistics and validated over the 5G Bristol is Open network topology.

II. Network Description and Problem definition

A. Network scenario

The present study focuses on a multi-technology network infrastructure deploying a set of optical and wireless network technologies to interconnect RUs with compute resources. The RUs are uniformly distributed across the served area in a typical hexagonal cellular fashion. Backhauling of the RUs is provided through an active optical metro solution aggregating traffic demands generated at the RUs, providing also the necessary capacity for the interconnection of compute resources. A typical example of such a network configuration is shown in Figure 1 (a).

A key architectural issue associated with this type of infrastructure is the placement of BBUs with respect to the RUs. In addition to this, recognizing the stringent delay and synchronization requirements of the existing FH protocol implementations, the concept of functional split processing is also considered. As illustrated in Figure 1 (b) the range of "split options", spans between the "traditional distributed RAN" case where "all processing is performed locally at the Access Point (AP)" to the "fully-centralized C-RAN" case where "all processing is allocated to a CU". All other options allow allocating some processing functions at the RU, while the remaining processing functions are performed remotely at the CU. The optimal allocation of processing functions to be executed locally or remotely i.e. the optimal "split", can be decided based on a number of factors such as transport network characteristics, network topology and scale as well as type and volume of services that need to be supported. In addition to the optimal split selection, mapping of the FH functions to the suitable GPP or SPP within the DC is also part of the optimization process. In the following subsection a high level description of the key optical transport characteristics and the disaggregated DC network together with the relevant modeling assumptions is provided.

B. Optical Transport Netowork

For the metro network, we consider a frame-based optical network solution [9] where the ingress edge nodes aggregate the incoming traffic into optical frames, which are then assigned to suitable time-slots and wavelengths for further transmission. At the egress point the reverse function takes place. The optical edge nodes are also equipped with elastic bandwidth allocation capabilities supported through the deployment of Bandwidth Variable Transponders (BVTs). The objective of the optical transport network is to provide connectivity for a number of RUs and end-users with a set of general GPPs. The use of GPPs enables the concept of virtual BBUs (vBBUs), facilitating efficient sharing of compute resources. This joint functionality is enabled by the edge nodes that comprise a subsystem able to handle both continues (CPRI data streams) and packetized flows (Ethernet flows). The design of such a subsystem is out of the scope of the present study, however, an indicative architecture supporting both type of services is provided in [10]. A practical implementation of this subsystem could be facilitated through a hybrid CPRI-Ethernet switch. The CPRI switch handles transport classes with strict synchronization and bandwidth constraints - i.e. split options (1) and (2) - while the Ethernet data switch handles BH traffic and relaxed FH transport classes, i.e. split options (3)-(5). An analysis of these splits is provided in [12]. FH data streams are supported by a synchronization block that manages the synchronization signals between the end points.

C. Intra-Data Center Network

For the intra-DC case, we consider a standard indirect or switch-based topology where connectivity between any two nodes is supported through switches. In such systems, multiple layers of switches are interconnected forming a hierarchical networking model. Switches may be organized either using simple tree topologies [13] (usually two-tier or three-tier [14]) or more sophisticated structures e.g. fat trees [15], [16]. Based on the type of interconnected devices used to form the pool of resources various switching solutions can be adopted. For example, PCIe switches can be used to interconnect multiple GPUs hosted in the same rack adopting the GPUDirect protocol or 40G/100G Ethernet switches for the interconnection of remotely located processing units. In the latter case, the GPUDirect RDMA over 40Gbps Ethernet protocol can be used.

A simple hierarchal network interconnecting CPUs and GPUs is shown in Fig. 1 d. In this figure, the SPP unit supporting FH function (3) (Fig. 1b), communicates through a set of high speed Ethernet switches with the SPP hosting function (4). The output of this SPP unit will be then sent to the SPP (5) through a PCIe switch. Following this approach, the entire SC implementing the FH service of RU1 shown in Fig. 1 b) can be realized.

III. End-to-end Network Modeling and Optimization

This section focuses on a two-stage optimization of the converged 5G infrastructures, comprising both heterogenoues network and compute/storage resources, in terms of energy consumption for both the inter and the intra-DC network. To achieve this, initially, the *optical transport network planning* problem is formulated aiming at identifying the necessary optical network resources for the interconnection of the RUs with the DCs. Once the *optical transport network* problem has been formulated a second sub-problem linked to the allocation of the FH functions to the disaggregated pool of resources is provided. To keep the analysis tractable, it is assumed that optical metro network topology (location of the optical nodes fibers) as well as the location of the RUs is kept fixed, whereas the FH functions supporting the operation of a specific RU are instantiated within the same DC.

Sub-problem 1: Transport Network Optimization. The first sub-problem tries to identify the optical network resources and the location of the DCs where the vBBUs are processed so that the total power consumption of the resulting network infrastructure is minimized. This formulation extends the work in [2] to include the elastic features of the optical transport network technology. For this sub-problem, the following parameters/indices/variables and constraints are introduced:

Indices:

$\mathbf{R} = 1, \dots, R$	Set of RUs
$\mathbf{D} = 1, \dots, D$	Set of data centers

$\boldsymbol{p_{rd}} = 1, \dots, P_{rd}$ Set of paths interconnecting RU r to DC $d \in \mathbf{D}$

$\mathbf{E} = 1, \dots, E$ Set of optical network links

$\mathbf{P}_r = \cup \boldsymbol{p}_{rd}$ Set of all paths interconnecting edge node $o \in \mathbf{O}$ to the DCs

$\mathbf{\Sigma} = 1, \dots, 5$ Set of split options. A summary of these options and the associated processing chain provided in Figure 1 b)

Constants

H_{ri} Transport network requirement of RU r under split option $i \in \mathbf{\Sigma}$

p_{ri} Total processing requirement of the flow generated at RU $r \in \mathbf{R}$ under split option $i \in \mathbf{\Sigma}$

p_{ri}^{RU} Local processing requirements of the flow generated at RU $r \in \mathbf{R}$ under split option $i \in \Sigma$

p_{ri}^{d} Processing load at the remote DC $d \in \mathbf{D}$ for the flow generated at RU $r \in \mathbf{R}$ under split option $i \in \Sigma$.

$\mathcal{E}_i$ Power consumption of element i.

P_d Total processing capacity of data center $d \in \mathbf{D}$

P_r Total processing capacity of RU $r \in \mathbf{R}$

δ_{erp} Binary coefficient taking value 1 if link $e \in \mathbf{E}$ belongs to path p realizing the traffic flow generated at the RU r; 0 otherwise

h_r Transport network requirements of RU r

C_e Capacity of link $e \in \mathbf{E}$

ξ_e Cost of link $e \in \mathbf{E}$

Variables

σ_{ri} Binary variable taking value equal to 1 if split option $i \in \Sigma$ is adopted, 0 otherwise.

a_{rd} Binary variable taking value equal to 1 if data center $d \in \mathbf{D}$ hosts the BBU service chain (or some of its parts) of the RU $r \in \mathbf{R}$

u_{rp} Binary variable forcing a single flow to be transferred from RU r over a single path $p \in \mathbf{P}_r$

C_e Capacity of link $e \in \mathbf{E}$

Constraints

1.1) Single-path flow from RU r to DC

$$\sum\nolimits_{p \in \mathbf{P}_r} u_{rp} = 1, \quad r \in \mathbf{R}$$

1.2) Transport network capacity requirements

$$\sum\nolimits_{r \in \mathbf{R}} h_r \sum\nolimits_{p \in \mathbf{P}_r} \delta_{erp} u_{rp} \leq C_e \quad e \in \mathbf{E}$$

1.3) Single-split processing enforcing constraints:

$$\sum\nolimits_{i \in \Sigma} \sigma_{ri} = 1, \ r \in \mathbf{R}$$

1.4) RU Demand constraint:

$$h_r = \sum\nolimits_{i \in \Sigma} H_{ri} \sigma_{ri}, \ r \in \mathbf{R}$$

1.5) BBU processing constraint: To support the BBU processing chain shown in Figure 1 (b), specific compute resources need to be allocated. Based on the functional split option adopted, a subset of the BBU chain functions can be executed at the RUs whereas the remaining ones are placed at remote servers through the optical metro network. For example, for the heavy FH flows shown in Figure 1 (b), processing for the "RF to baseband" and "Cycle Prefix and FFT" functions will be carried out at the RUs while the remaining ones ("Receive processing", "Decoding" , "MAC") at the servers placed at the right hand side of Figure 1 (a). The associated processing constraints per split option for the RU are:

$$\sum\nolimits_{i \in \Sigma} p_{ri}^{RU} \sigma_{ri} \leq P_r, \quad r \in \mathbf{R}$$

whereas of the case of DCs, the processing flows of all RUs shall not exceed data center's $d \in \mathbf{D}$ capacity constraints:

$$\sum\nolimits_{r \in \mathbf{R}} \sum\nolimits_{i \in \Sigma} p_{ri}^{d} \sigma_{ri} \leq P_d, \ d \in \mathbf{D}$$

1.6) Single-server BBU processing: The processing requirements per RU flow can be supported by a single DC at most.

$$\sum\nolimits_{d \in \mathbf{D}} a_{rd} \leq 1, \ r \in \mathbf{R}$$

1.7) BBU chain processing conservation constraints:

$$p_{ri}^{RU} + \sum\nolimits_{d \in \mathbf{D}} a_{rd} p_{ri}^{d} = p_{ri}, \quad r \in \mathbf{R}, i \in \mathbf{\Sigma}$$

Constraint (1.7) indicates that the sum of the processing load perfomed either at the RU or at the DC $d \in \mathbf{D}$ equals the total processing requirement of the FH flow generated at the RU $r \in \mathbf{R}$ under split option $i \in \mathbf{\Sigma}$.

1.8) Objective:

$$\min F_1 = \sum\nolimits_{r \in \mathbf{R}} \mathcal{E}_r \left(\sum\nolimits_{i \in \Sigma} p_{ri}^{RU} \sigma_{ri} \right) + \sum\nolimits_{e \in \mathbf{E}} \mathcal{E}_e C_e +$$

In (1.8) the first term accounts for the total power consumption at the RUs for partially processing the associated BBU chain whereas the second term indicates the total power consumption at the transport network interconnecting the RUs with the DCs.

Sub-problem 2: Disaggregated DC Network Optimization. The second sub-problem tries to identify the optimal processing modules where the remaining parts of the FH SC have to be allocated. To achieve this, once the FH data reach a DC hosting the

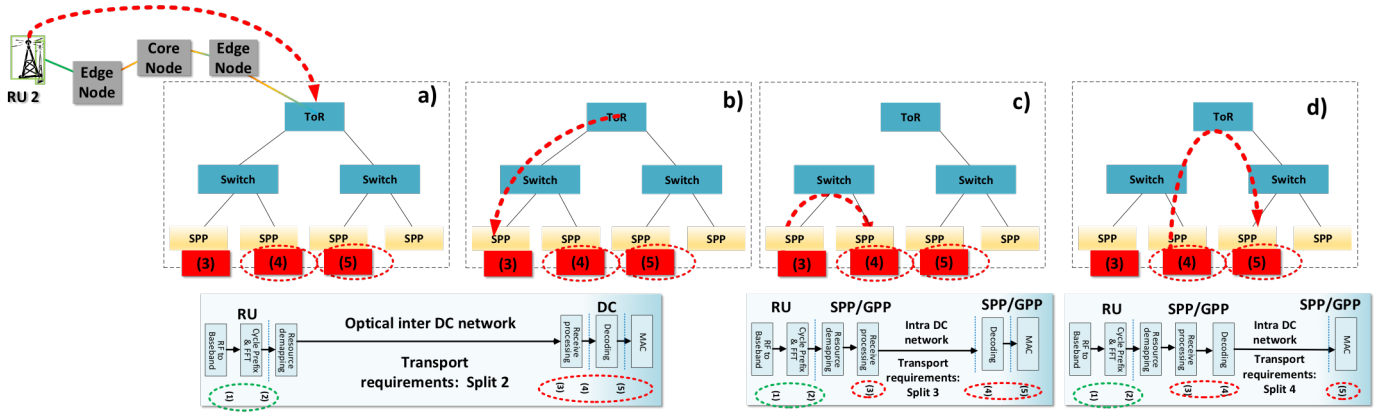

Fig. 2 Time evolution of FH service chain for split option (2) over dissagregated DC resources for the pipelining processing mode.

candidate pool of resources, a path interconnecting the edge DC node with the GPP/SPP modules that will process the remaining FH functions is established. The order of FH functions processing is defined by the corresponding SC shown in Figure 1b). The following functions/parameters/variables are introduced:

Indices:

$\mathbf{M}^{d} = 1, \dots, M^{d}$ Set of processing modules belonging to DC $d \in \mathbf{D}$

$\mathbf{FH}_{ri}^{d} = 1, \dots, FH_{ri}^{d}$ Ordered set of remaining FH functions for the flow generated at the RU $r \in \mathbf{R}$ under split option $i \in \Sigma$ processed at DC $d \in \mathbf{D}$.

$\mathbf{E}^{d} = 1, \dots, E^{d}$ Set of inter-DC network links

$\boldsymbol{p}_{\varphi km}^{d} = 1, \dots, P_{\varphi km}^{d}$ Set of paths interconnecting module $k \in \mathbf{M}$ hosting function $\varphi \in \{1, \dots, FH_{ri}^{d} - 1\}$ to module $m \in \mathbf{M}$ hosting the next function $\varphi + 1$ of the FH SC at DC $d \in \mathbf{D}$.

$\mathbf{P}_{\varphi k}^{d} = \cup\, \boldsymbol{p}_{\varphi km}^{d}$ Set of all paths interconnecting module $k \in \mathbf{M}$ supporting function φ to any other module of the DC $d \in \mathbf{D}$

$\boldsymbol{H_{k\varphi}}$ Transport network requirement of module k hosting function $\varphi \in \{1, \dots, FH_{ri}^{d} - 1\}$

Constants

P_k Processing capacity of module $k \in \mathbf{M}^{d}$

p_{φ} Processing requirements of function $\varphi \in \mathbf{FH}_{ri}^{d}$

$\zeta_{e\varphi p}$ Binary coefficient taking value 1 if link $e \in \mathbf{E}^{d}$ belongs to path $p \in \boldsymbol{p}_{\varphi km}^{d}$ interconnecting modules k and m; 0 otherwise

Variables

$a_{\varphi k}$ Binary variable taking value equal to 1 if module $k \in \mathbf{M}$ hosts FH function $\varphi \in \mathbf{FH}_{ri}^{d}$

u_{kp} Binary variable forcing a single egress flow from module $k \in \mathbf{M}$ over a single path $p \in \mathbf{P}_{k}^{d}$

Constraints

2.1) Pipelining processing constraints: a) Every function $\varphi \in \mathbf{FH}_{ri}^{d}$ has to be processed by a single module $k \in \mathbf{M}$:

$$\sum\nolimits_{k \in \mathbf{M}^{d}} a_{\varphi\kappa} = 1, \quad \varphi \in \mathbf{FH}_{ri}^{d}, \qquad r \in \mathbf{R}, i \in \mathbf{\Sigma}, d \in \mathbf{D}$$

b) Processing capacity constraints of each module must not violated:

$$\sum\nolimits_{r \in \mathbf{R}} \sum\nolimits_{i \in \Sigma} \sum\nolimits_{\varphi \in \mathbf{FH}_{ri}^{d}} p_{\varphi}\, a_{\varphi\kappa} \le P_{k}, \quad k \in \mathbf{M}^{d}, d \in \mathbf{D}$$

2.2) Pipelining communication constraints: a) Connectivity between module $k \in \mathbf{M}^{d}$ hosting function $\varphi \in \mathbf{FH}_{ri}^{d}$ and module $m \in \mathbf{M}^{d}$ hosting the subsequent function in the SC should be provided over a single path:

$$\sum\nolimits_{p \in \mathbf{P}_{\varphi k}^{d}} u_{kp} = 1, \quad \varphi \in 1, \dots, FH_{ri}^{d} - 1, k \in \mathbf{M}^{d}, d \in \mathbf{D}$$

b) Module-to-module demand requirements: The egress traffic from module $k \in \mathbf{M}^{d}$ hosting function $\varphi \in 1, \dots, FH_{ri}^{d} - 1$ should be redirected to module m hosting function $\varphi + 1$. The associated network requirement in this case is given by:

$$H_{k\varphi} = H_{ri+1}, \quad k \in \mathbf{M}^{d}, \varphi \in 1, \dots, FH_{ri}^{d} - 1, r \in \mathbf{R}, i \in 1, .., \Sigma - 1$$

Module-to-module communication capacity requirements:

$$\sum\nolimits_{r \in \mathbf{R}, i \in \Sigma, k \in \mathbf{M}^{d}} \sum\nolimits_{\varphi \in \mathbf{FH}_{ri}^{d}} H_{k\varphi} \sum\nolimits_{p \in \mathbf{P}_{\varphi k}^{d}} \zeta_{e\varphi p} u_{kp} \le C_{e} \qquad e \in \mathbf{E}^{d}$$

2.3) Objective:

$$\min F_{1} = \sum\nolimits_{k \in \mathbf{M}^{d}} \mathcal{E}_{k} \left(\sum\nolimits_{r \in \mathbf{R}} \sum\nolimits_{i \in \Sigma} \sum\nolimits_{\varphi \in \mathbf{FH}_{ri}^{d}} p_{\varphi}\, a_{\varphi\kappa} \right) + \sum\nolimits_{e \in \mathbf{E}^{d}} \mathcal{E}_{e}\, C_{e} \; +$$

In (2.3) the first term accounts for the total power consumption of the pool of computing resources for processing FH functions for all RUs under various possible split options whereas the second

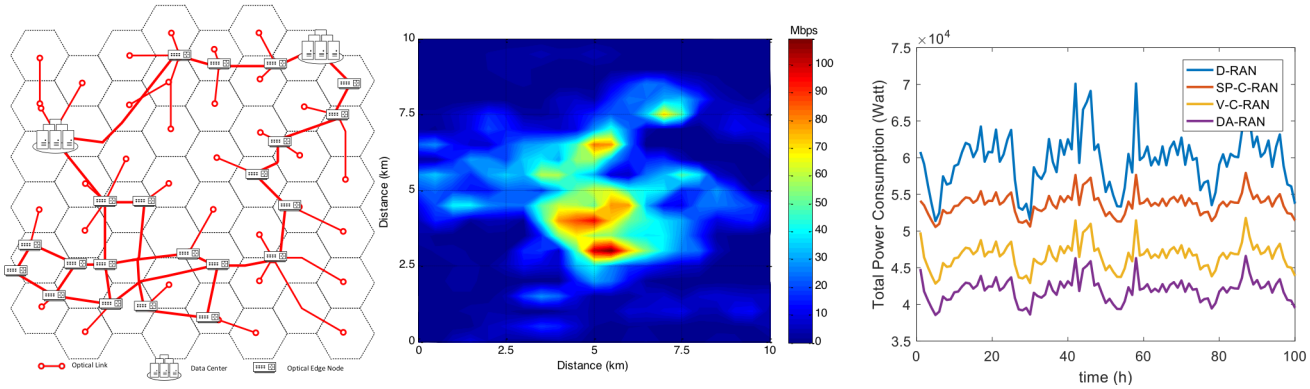

Fig. 3a) The 5G City of Bristol network topology, b) Spatial distribution of traffic for 50 BS over an 10x10km2 area, c) Evolution of the total power consumption for the various RAN schemes using the traffic statistics in [11]

term indicates the total power consumption for the DC network infrastructures.

IV. Numerical Results and Conclusions

The proposed optimization scheme is evaluated using the optical network topology shown in Fig. 2a) covering a 10x10 km^2 area over which 50 BSs are uniformly distributed. RUs demands are generated according to real datasets reported in [11]. A snapshot of the spatial distribution of this traffic is shown in Figure 2 b). As already discussed, some of the FH service functions need to be processed by specific compute resources. Based on the type and location of the compute resources the following cases are examined:

a) "Distributed-RAN (D-RAN)": In this scheme, RUs and BBUs are co-located and FH service processing is carried out exclusively by specific purpose hardware. Sharing of BBUs between multiple RUs is not supported and sizing of BBUs is performed based on worst case traffic statistics. The power consumption per AP ranges between 600 and 1200 Watts under idle and full load conditions, respectively.

b) "C-RAN with specific purpose (SP) BBU hardware (SP-C-RAN)": In this scheme, specific purpose BBUs are placed at a centralized location. Compared to D-RAN, this scheme offers the advantage of BBU sharing. As before, sizing of BBUs is performed under worst case traffic conditions.

c) "C-RAN with virtual BBU (V-C-RAN)-": Compared to SP-C-RAN where BBUs run on SP hardware, this scheme allows BBUs to be instantiated as virtual functions and run on general purpose processors. This scheme allows sharing of resources and on-demand resizing of compute resources to match the FH service requirements. The main disadvantages of this approach include higher processing cost per bit associated with GPPs when compared to specific purpose hardware.

d) "Disaggregated-RAN (DA-RAN)": This novel scheme combines the benefits of SP-C-RAN and V-C-RAN allowing FH functions to be processed either at SPP or GPP based on their specific characteristics. Through this approach, intensive FH functions are performed at SPP (ASICS) hosted at the DCs whereas the remaining functions are instantiated on shared GPPs.

Fig.3c provides the total infrastructure power consumption as a function of time for the four schemes under consideration. As expected, the DA-RAN approach outperforms all alternative approaches. The benefits of the DA-RAN is attributed to the sharing gains it offers in both the space and time domains due to its flexible and on demand resource allocation capabilities. DA-RAN minimizes overprovisioning requirements present in alternative approaches leading to 10--50% power consumption savings.

Acknowledgment

This work has been supported by the EU Horizon 2020 5GPPP project 5G-XHaul.

References

[1] 5G Vision. The 5G Infrastructure Public Private Partnership: the next generation of communication networks and services. 2015. [Online]

[2] A. Tzanakaki *et al*., "5G infrastructures supporting end-user and operational services: The 5G-XHaul architectural perspective," *IEEE ICC*, 2016.

[3] M. Ruffini, Multi-Dimensional Convergence in Future 5G Networks. IEEE/OSA Journal of Lightwave technology, Vol. 35, No. 3, March 2017.

[4] U. Dötsch *et al*., Quantitative Analysis of Split Base Station Processing and Determination of Advantageous Architectures for LTE, *Bell*

[5] S. Han et al., "Network support for resource disaggregation in next-generation datacenters", In Proceedings of HotNets-XII. ACM, New York, NY, USA, Article 10 , 2013.

[6] M. Fiorani, S.Tombaz, J.Martensson, B.Skubic, L.Wosinska, P. Monti, "Modeling energy performance of C-RAN with optical transport in 5G network scenarios," IEEE/OSA JOCN vol.8, no.11,pp.B21-B34, 2016.

[7] F. Musumeci *et al.,* "Optimal BBU Placement for 5G C-RAN Deployment Over WDM Aggregation Networks," J. Lightwave Technol. **34**, 2016.

[8] U Siddique et al., Wireless backhauling of 5G small cells: Challenges & solution approaches, *IEEE Wireless Comm.*. vol.22, no.5, pp.22-31, 2015

[9] Y. Yan *et al*., "High performance and flexible FPGA-based time shared optical network (TSON) metro node," Opt. Exp **21**, 5499-5504,2013

[10] Xilinx, CPRI Switch, https://www.xilinx.com/products/intellectual-property/1-6kiivn.html

[11] X. Chen et al., "Analyzing and modeling spatio-temporal dependence of cellular traffic at city scale," *IEEE ICC*, pp.3585-3591, 2015

[12] 5G-XHaul Project, Deliverable D2.1 "Requirements Specification and KPIs Document", March 1st, 2016.

[13] Cisco Systems, "Data Center Design – IP Network Infrastructure," http://www.cisco.com/en/ US/docs/solutions/Enterprise/Data Center/DC 3 0/DC-3 0 IPInfra.pdf, Oct. 2009

[14] http://www.cisco.com/c/en/us/td/docs/solutions/Enterprise/Data_Center/DC_Infra2_5/DCInfra_1.html

[15] M. Al-Fares et al., "A scalable, commodity data center network architecture," in proc. of SIGCOMM, pp. 63–74, 2008

[16] A. Greenberg et al., VL2: a Scalable and Flexible Data Center Network," in proc. of SIGCOMM, pp. 51–62, 2009