Peer reviewed version

Link to published version (if available):
10.1109/ICCW.2016.7503764

Link to publication record in Explore Bristol Research
PDF-document

University of Bristol - Explore Bristol Research
General rights

5G Infrastructures Supporting End-User and Operational Services: The 5G-XHaul Architectural Perspective

Anna Tzanakaki[(1)], Markos Anastasopoulos[(1)], Dimitra Simeonidou[(1)], Ignacio Berberana[(2)], Dimitris Syrivelis[(3)], Thanasis Korakis[(3)], Paris Flegkas[(3)], Daniel Camps Mur [(4)], Ilker Demirkol[(4)], Jesús Gutiérrez[(5)], Eckhard Grass[(5)], Qing Wei [(6)], Emmanouil Pateromichelakis[(6)], Albrecht Fehske [(7)], Michael Grieger [(7)], Michael Eiselt[(8)], Jens Bartelt[(9)], George Lyberopoulos[(10)], Eleni Theodoropoulou[(10)]

(1) HPN Group, University of Bristol, UK, (e-mail: anna.tzanakaki@bristol.ac.uk), (2)Telefonica Investigacion Y Desarrollo, Madrid, ESP, (3)University of Thessaly, Volos, GRC, (4) Mobile and Wireless Internet Group, i2CAT Foundation, ESP (5) IHP GmbH, DE(6) Huawei Technologies Duesseldorf GmbH, (7) Airrays GmbH, Dresden, DE (8) ADVA Optical Networking SE, (9) Vodafone Chair Mobile Communications Systems, Technische Universität Dresden, DE, (10) Network Evolution Dept., Fixed & Mobile, COSMOTE Mobile Communications S.A, GR.

Abstract—We propose an optical-wireless 5G infrastructure offering converged fronthauling/backhauling functions to support both operational and end-user cloud services. A layered architectural structure required to efficiently support these services is shown. The data plane performance of the proposed infrastructure is evaluated in terms of energy consumption and service delay through a novel modelling framework. Our modelling results show that the proposed architecture can offer significant energy savings but there is a clear trade-off between overall energy consumption and service delay.

Keywords—5G, backhauling, fronthauling, small cells, C-RAN

I. INTRODUCTION

To meet the ever increasing growth of mobile traffic demands, the traditional wireless access network architecture based on single layer macro-cells is being currently transformed to an architecture comprising a large number of smaller cells with densely deployed access points (APs), combined with micro and macro-cells. In traditional Radio Access Networks (RANs), baseband units (BBUs) and radio units are co-located suffering several limitations including: i) increased CAPEX to acquire new base stations (BSs) and OPEX due to underutilized resources, ii) limited scalability and flexibility, iii) lack of modularity and limited density as the system is complex to resize after deployment, iv) increased management costs, and v) inefficient power delivery as the BSs processing power cannot be shared.

An alternative solution recently proposed is that of Cloud Radio Access Networks (C-RANs) where distributed APs, referred to as remote radio heads (RRHs), are connected to the BBU pool through high bandwidth transport links known in as fronthaul (FH) (left part of Fig. 1). FH is responsible to carry the RRH wireless signals typically over the optical transport network using either digitized form based on protocols such as the Common Public Radio Interface (CPRI), or in analogue form through radio-over-fiber technology [1]. Recently, several solutions for wireless fronthauling have been proposed, as well
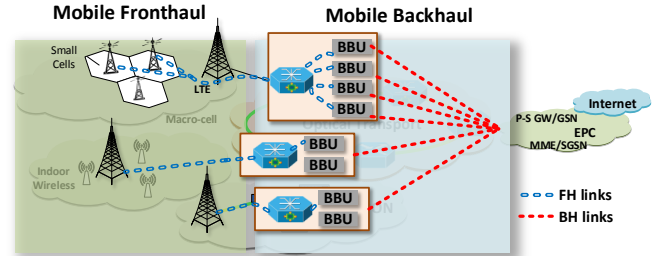

Fig. 1 Mobile fronthaul and backhaul

[2]. The main advantage of digitized transmission is reduced signal degradation allowing data transmission over longer distances, enabling the adoption of longer reach optics offering higher degree of BBU consolidation. C-RAN's main disadvantages include increased transport bandwidth requirements to carry the sampled radio signals, and strict latency and synchronization constraints [3]. For example, in a single LTE 20 MHz 2x2 MIMO sector, the required capacity for the RRH-BBU interconnection is 2.46 Gbps and may increase up to 12.165 Gbps with CPRI line bit rate option 9 [4]. Given that existing optical transport solutions for APs are either based on Passive Optical Networks (PON), Gigabit-capable Passive Optical Networks (GPON) or 10GE technologies offering capacities up to 10 Gbps, it is obvious that the mobile BH network can rapidly become the bottleneck. To relax the stringent FH requirements of C-RAN architectures, while taking advantage of its pooling and coordination gains, alternative architectures proposing flexible splits (Fig. 2) have been proposed [6], [21]. In addition to high bandwidth transport connectivity, this flexible split requires fine bandwidth granularity and elastic resource allocation.

In this paper, a converged optical-wireless 5G network infrastructure interconnecting computational resources with fixed and mobile users is proposed, to support both operational network (C-RAN) and end-user computational services [5], adopting the concept of cloud computing. This infrastructure is being developed in the framework of the EU funded
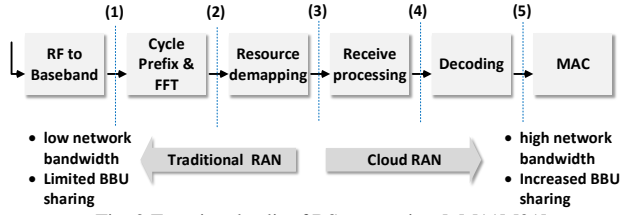
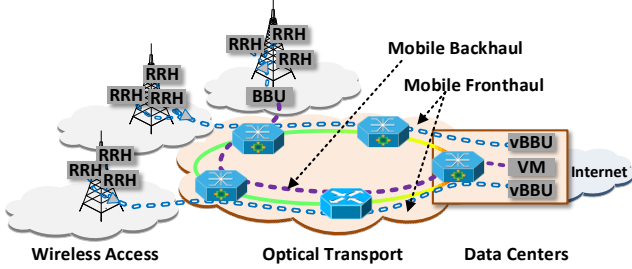Fig. 2 Functional split of BS processing [6],[11],[21]

Fig. 3 Unified mobile FH/BH over converged optical data center networks

HORIZON 2020 5GPPP project 5G-XHAUL. A layered architecture, inspired by the ETSI Network Function Virtualization (NFV) standard [14] and the SDN reference architecture [15], is also presented in detail. This architecture describes the required functions and their interactions, that 5G-XHAUL proposes to effectively and efficiently provision both end-user and operational services over the proposed infrastructure. A novel modelling framework has been developed with the aim to evaluate the performance of this infrastructure. This includes a multi-objective (MOP) service provisioning model used to study a variety of FH and BH options, spanning from the traditional approach where the two functions are supported separately to solutions involving fully or partially converged FH and BH functions (Fig.3). The proposed provisioning model takes a holistic view considering jointly mobile FH and BH functions to ensure appropriate allocation of the required resources across all domains. Its objective is twofold: i) to minimize the operational expenditure of the FH in terms of power consumption under strict delay constraints achieved through the optimal functional split of BS processing as well as through optimal BBU placement [16], and ii) to minimize end-to-end cloud service delay in the BH.

II. OVERVIEW OF THE 5G-XHAUL ARCHITECTURE

A. Data Plane Architecture

The 5G-XHaul data-plane architecture considers an integrated optical and wireless network infrastructure. The wireless domain comprises a dense layer of small cells that are located 50-200 m apart [18]. This small cell layer is complemented by a macro cell layer to ensure ubiquitous coverage. Macro-cell sites are around 500 metres apart. Small cells can be wirelessly backhauled to the macro-cell site using a combination of mm-Wave and Sub-6 wireless technologies. Alternatively, the 5G-XHaul architecture allows small cells to be directly connected to a central office node using optical network technologies and, more specifically, PONs offering enhanced capacity through the deployment of Wavelength Division Multiplexing (WDM). In addition to WDM-PONs, 5G-XHaul adopts the use of a dynamic and flexible/elastic frame based optical network solution that can support more

demanding capacity and flexibility requirements for traffic aggregation and transport. Through this architecture 5G-XHaul aims to efficiently support a large variety of end-user services as they are envisaged for the 5G era (e.g. as defined by the EU project METIS [17]).

A key architectural issue associated with this type of infrastructure is the location of BBUs and radio units. In 5G-XHaul, the concept of C-RAN, where RRHs, are connected to BBU pools through high bandwidth transport links, referred to as fronthaul, is one of the approaches investigated in order to overcome the limitations associated with the traditional RAN approach. Through the need for fronthauling capability, this architectural choice introduces the requirement to support an additional set of services for operational network purposes. More specifically, the densely distributed BSs/RRHs need to be connected to regional data centres that host BBUs with very stringent delay and synchronisation requirements. 5G-XHaul proposes to use a common network infrastructure to support jointly backhauling and fronthauling functions maximising the associated sharing benefits improve efficiency in resource utilisation and provide measurable benefits in terms of overall cost, scalability and sustainability objectives. This can be practically supported through the proposed 5G-XHaul data plane architecture as well as the advanced wireless and optical network technologies that are developed internally within the project. It should be noted that a key enabler supporting the feasibility of the proposed approach is the adoption of a high capacity, flexible optical transport comprising both passive and active solutions. The passive optical network solutions will be based on WDM-PONs, while the active solution adopts the Time-Shared Optical Network (TSON) [18], deployed after being enhanced with novel features for improved granularity and elasticity. These can provide the required connectivity, capacity and flexibility to offer jointly backhauling and fronthauling functions and support a large variety of end-user and operational services. A high level view of the 5G- XHaul data plane architecture is provided in Fig. 4.

B. Overarching Layered Architecture

Through Fig. 4, it is clear that the 5G-XHaul infrastructure exhibits a great degree of heterogeneity in terms of technologies. To address the challenge of managing and operating this type of complex heterogeneous infrastructure in an efficient manner, 5G-XHaul proposes the adoption of Software Defined Networking (SDN) and NFV that will be integrated in a seamless manner. In SDN, the control plane is decoupled from the data plane and is managed by a logically centralized controller that has a holistic view of the network [17]. At the same time, NFV enables the execution of network functions on commodity hardware (general-purpose servers, standard storage and switches) by leveraging software virtualization techniques [17]. Through joint consideration of SDN and NFV significant benefits can be achieved. For example, the separate control plane can be virtualized using NFV, and the SDN controller-related Virtual Network Functions (VNFs) may be deployed dynamically, having the ability to scale up and down on demand based on the associated workloads [19].
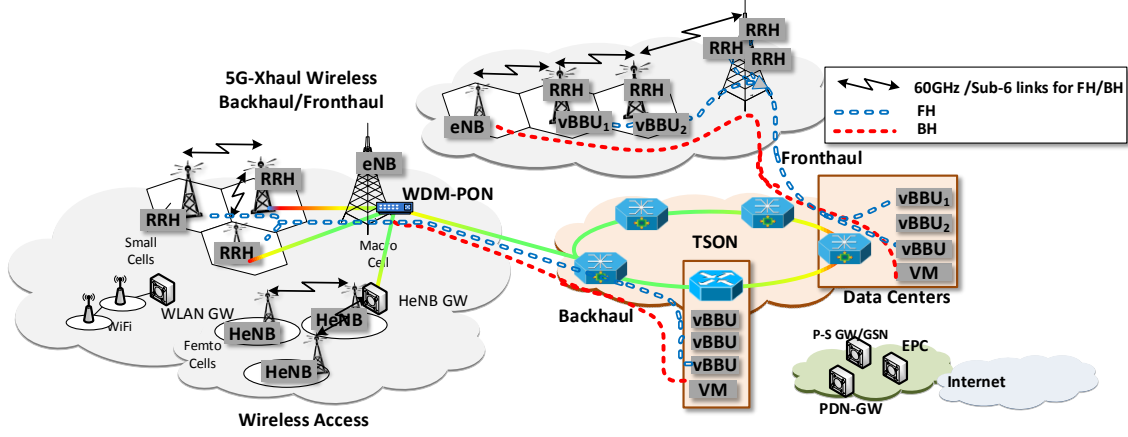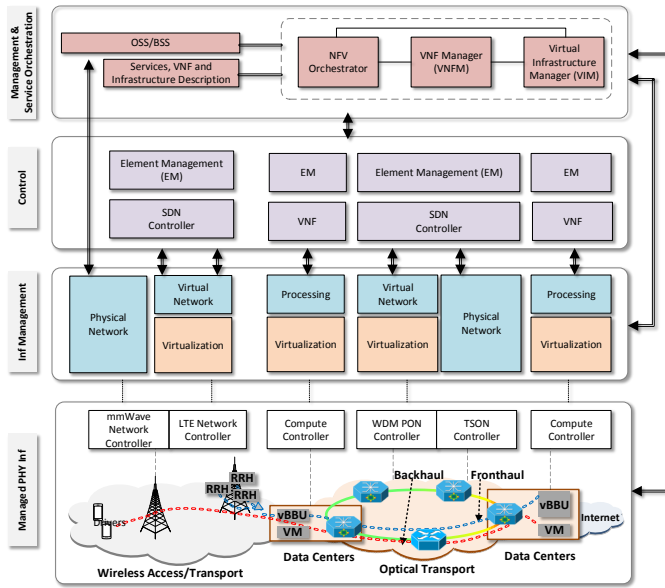
Fig. 4. The 5G-Xhaul Physical Infrastructures

Fig. 5. The overall cross-domain, cross layer architecture

As discussed in detail in [19], SDN network elements can be treated as VNFs since they can be implemented as software running on general-purpose platforms in virtualized environments. Both SDN and non-SDN models can be supported by SDN network elements. On the other hand, network applications can include SDN controller functions, or interact with SDN controllers and can themselves provide VNFs. Service chaining is considered to be one possible network application. Network elements controlled by an SDN controller can be either VNFs or Physical Network Functions (PNFs).

Taking advantage of the SDN concept and the benefits of cross-technology virtualization 5G-XHaul proposes a heterogeneous network infrastructure and an overarching layered architecture able to efficiently and effectively support 5G services. Fig. 5 presents the overall multi-domain multi-layer architecture. The lower layer of the layered architecture i.e. the Managed Physical Infrastructure comprises a converged optical/wireless transport network able to interconnect RRHs, end-users and computing resources. The wireless access part comprises a dynamically programmable, high capacity, low latency, point-to-multipoint (p2mp) mm-Wave transceivers,

cooperating with sub-6-GHz systems. The optical transport relies on a hybrid passive (WDM-PON) and active optical network solution. The optical metro network solution supporting frame-based sub-wavelength switching granularity, cooperating with advanced passive optical networks is based on TSON.

The second layer (Infrastructure Management) is responsible for the management of the different technology domains and the creation of virtual and physical infrastructure slices comprising heterogeneous resources. The Infrastructure Management Layer (IML) communicates with the various network and compute controllers that are responsible for retrieving information and communicating with the individual domains. Once the information has been collected, the resources are abstracted and virtualized. From the architectural and functional perspective, IML addresses all virtualization associated functions as well as the virtual resource management functions. Management of traditional non-virtualized physical infrastructures can be also supported.

Cross-domain orchestration of the virtual and physical infrastructures, created and exposed by the IML to the higher layers, is carried out by the control layer. This layer, has a holistic view of all network segments and technology domains and implements converged control and management procedures for dynamic and automated provisioning of end-to-end connectivity services (i.e., service chaining) according to specific QoS considerations. Configuration of virtualized (or non-virtualized) wireless and optical network resources is carried out by a set of distributed SDN controllers. Control of legacy devices directly from the Operational Support System (OSS) is also supported. Besides network configuration capabilities offered by the SDN controllers, further enhanced VNFs that run on top of the virtualized infrastructures can be developed in order to operate the entire heterogeneous infrastructure in a seamless manner.

Finally, the Management and Service Orchestration Layer is responsible for the converged orchestration of cloud and network services. It is also used for the composition and delivery of multi-tenant chains of virtualized network functions. In addition, it performs Resource Orchestration through NFV resources across multiple Virtual Infrastructure Managers (VIMs) and includes lifecycle management of

Network Services, supporting Network Service Orchestration functions.

It should be noted that the proposed architecture also allows direct interaction of the OSS with physical network devices that do not deploy SDN control. This provides a framework supporting smooth interoperability with legacy software and hardware technologies and architectures.

III. USE CASE: JOINT OPTIMIZATION OF FH/BH IN SUPPORT OF C-RAN AND CONTENT DELIVERY SERVICES

In this section we concentrate on the evaluation of the performance of the proposed data plane architecture. More specifically we consider a physical infrastructure (PI) that interconnects RRHs and end-users with a set $\mathcal{S}$ of S geographically distributed general-purpose servers ([7]) through a heterogeneous frame-based WDM optical metro network [8]. The PI is represented as a weighted graph $\mathcal{G} = (\mathcal{N}, \mathcal{E}, \mathcal{D})$ where $\mathcal{N}$ represents the set of PI nodes, $\mathcal{E}$ the set of PI links and $\mathcal{D}$ describes the set of demands. $\mathcal{D}$ is partitioned into $\mathcal{D}_F$ and $\mathcal{D}_B$ i.e., $\mathcal{D} = \mathcal{D}_F \cup \mathcal{D}_B$, where $\mathcal{D}_F, \mathcal{D}_B$ are the set of demands originating from the FH and BH, respectively. At this point, it should be noted that FH demands are generated at the BSs, therefore, in the remaining part of the paper it is assumed that $\mathcal{D}_F$ is identical to the set of BSs. In order to abide to the strict latency constraints of the C-RAN flows, the FH is modelled using network calculus theory, where each C-RAN flow $d \in \mathcal{D}_F$ is constrained by an arrival curve a_{r_d,b_d} and a service curve β_{c_d,T_d}. Arrival curves of the form a_{r_d,b_d} allow sources d to transmit bursts with size b_d bits at once, but no more than r_d bits/s in the long run [9]. Service curves β_{c_d,T_d} can serve traffic with rate c_d after T_d time delay.

Arrival curves a_{r_d,b_d} depend primarily on the functional split options of the BS processing and the characteristics of the LTE system. For example, assuming an LTE system with transmission bandwidth B_w=20 MHz, sampling frequency of 30.72 MHz, bit resolution per I/Q 2, oversampling factor 2 and 2 antennas, $r_{d(1)}$ under split option (1) in Fig. 2 will be 2.46 Gbps. However, when employing split option (2) this is reduced to $r_{d(2)}$=720 Mbps, assuming 1200 subcarriers and Fast Fourier Transformation (FFT) period of 66.67μsec [6], [21]. Let Σ_d be the set of split options for demand d (see Fig. 2 for a graphical representation of the split options set) and σ_{di} a binary variable taking value equal to 1 if split option $i \in \Sigma_d$ is adopted, 0 otherwise. The following demand constraints should be satisfied:

$$\sum_{s \in \mathcal{S}} \sum_{p \in \mathcal{P}_d} \wp_{ds} x_{dp_s} = \sum_{i \in \Sigma_d} \sigma_{di} a_{r_{d(i)}, b_{d(i)}}, \qquad d \in \mathcal{D}_F \quad (1)$$

$$\sum_{i \in \Sigma_d} \sigma_{di} = 1, \quad d \in \mathcal{D}_F \quad (2)$$

where $\mathcal{P}_d = 1, 2, \ldots, P_{ds}$, is the set of paths transferring FH demands $d \in \mathcal{D}_F$ to server $s \in \mathcal{S}$, $\wp_{ds}$ is a binary coefficient taking values equal to 1 if $d \in \mathcal{D}_F$ is processed at server $s, s \in \mathcal{S}$, 0 otherwise and x_{dp_s} is the non- negative capacity allocated to

path p supporting demand d. Summing up the paths through each link e ($e \in \mathcal{E}$), the capacity constraints should be satisfied:

$$\sum_{d \in \mathcal{D}_F} \sum_{s \in \mathcal{S}} \sum_{p \in \mathcal{P}_d} \delta_{edp} x_{dp_s} \leq u_{FH,e}, \qquad e \in \mathcal{E} \ (3.1)$$

$$u_{FH,e} \leq \mathcal{U}_e \ e \in \mathcal{E} \qquad (3.2)$$

In (3.1), δ_{edp} is a binary coefficient with value 1 if link e belongs to path p for traffic flow d and 0 otherwise, $u_{FH,e}$ is the link e capacity allocated for FH functions and $\mathcal{U}_e$ is the total capacity of e. x_{dp_s} is viewed as the arrival curve for flow d using path p to reach server s. Based on (3), the aggregated arrival curve for all flows $j \in \mathcal{D}_F, j \neq d$ at e, denoted as $a_{-d,e}$, is given through:

$$a_{-d,e} = \sum_{j \in \mathcal{D}_F, j \neq d} \sum_{s \in \mathcal{S}} \sum_{p \in \mathcal{P}_j} \delta_{ejp} x_{jp_s} \quad (4)$$

Substituting in β_{c_d,T_d} the parameters $c_d = u_e$ and $T_d = T_e$, then, β_{u_e,T_e} can now be seen as the service curve of e, where T_e is the propagation delay. According to network calculus, for a flow traversing a system with arrival curve a and service curve β, the upper delay bound is $h(a, \beta) = \inf\{t \geq 0 : (a \oslash \beta)(-t) \leq 0\}$, where $\oslash$ is the in min-plus de-convolution operator. The upper bound $\hat{t}_{de}$ of the delay introduced by link e for flow d can be evaluated through:

$$\hat{t}_{de} = h\left(x_{dp_s}, \left[\beta_{u_e,T_e} - a_{-d,e}\right]^+\right) \quad (5)$$

Once $\hat{t}_{de}$ has been determined, the total delay introduced across all links forming the path is evaluated. To ensure seamless operation of C-RAN, this delay should be limited below a certain threshold, usually, between 100-200 μs.

Besides network capacity, FH requires specific computing resources allocated for BBU processing. The processing power per demand depends on the sub-components of the BBU (Fig. 1) including FFT, error correction, processing-resource mapping/demapping etc. calculated in Giga Operations per Second (GOPS) via an equation of the form [10]-[11]:

$$P_{\mathcal{I}_{BBU},d} = \sum_{c \in \mathcal{I}_{BBU}} P_{c,d}\left(\mathcal{X}_{act}, \mathcal{X}_{ref}, s_{c,x}\right) \qquad (6)$$

In (7), $\mathcal{I}_{BBU}$ is the set of BBU sub-components, $P_{c,d}$ is the processing power required to execute tasks related to component c for demand d and $\mathcal{X}_{act}, \mathcal{X}_{ref}, s_{c,x}$ are reference parameters [10]. These parameters depend on the configuration of the LTE system (i.e. number of antennas, bandwidth, modulation, coding, number of resource blocks). Based on the functional split adopted, part of the processing can be performed either at a local BS with cost w_d per GOPS or at a remote server s with cost w_s per GOPS ($w_s > w_d$). The total information to be processed by server s for FH is:

$$\pi_{FH,s} = \sum_{d \in \mathcal{D}_F} \sum_{i \in \Sigma_d} \wp_{ds} P_{\mathcal{I}_{BBU(i)},d}, \quad (7)$$

while the portion of demand d that is processed locally:

$$\pi_{FH,d} = \sum_{i \in \Sigma_d} P_{\mathcal{I}_{BBU(-i)},d} \quad (8)$$
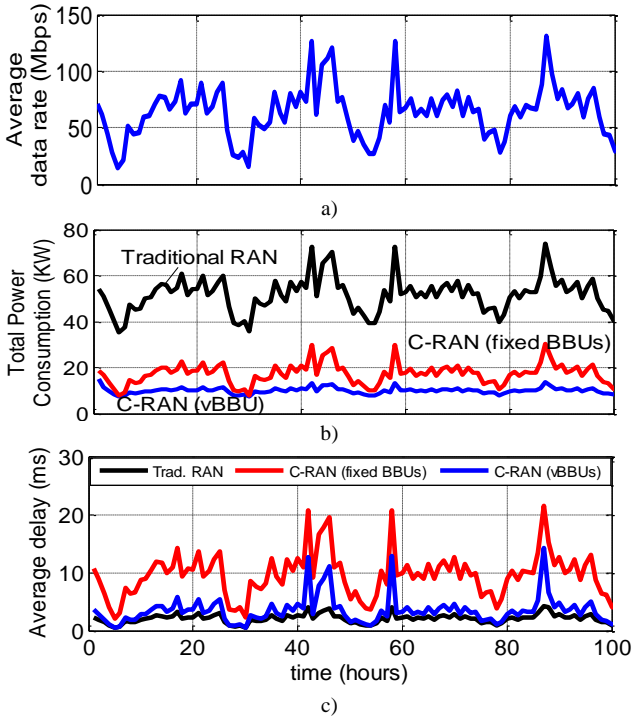
Fig. 6 a) Average traffic/BS based on the dataset [10] during 8/2012, b)-c) Total power consumption and delay over time for the traditional RAN and the C-RAN with fixed and vBBUs

$\mathcal{J}_{BBU(i),d}$ is the subset of components processed at the remote BBUs in case of split $i \in \Sigma_d$ and $\mathcal{J}_{BBU(-i),d}$ is the subset of the remaining components processed locally. During this process, the servers' capacity constraints should not be violated: $\pi_{FH,s} \leq \mathcal{C}_s$ ($s \in \mathcal{S}$) and $\pi_{FH,d} \leq \mathcal{C}_d$ ($d \in \mathcal{D}_F$) where $\mathcal{C}_s$ is the processing capacity of server s and $\mathcal{C}_d$ the processing capacity of a small-scale server that is co-located with the BS where demand d is generated. Assuming that the cost per link e is w_e, the optimal FH network is identified through the minimization of the cost:

$$\min \mathcal{FH}(\boldsymbol{u}, \boldsymbol{\pi}) = \sum_{e \in \mathcal{E}} w_e\, u_{FH,e} + \sum_{s \in \mathcal{S}} w_s \pi_{FH,s} + \sum_{d \in \mathcal{D}_F} w_d\, \pi_{FH,d} \quad (9)$$

subject to capacity, functional split and demand constraints.

Due to the inherent energy efficient operation of the optical network, improved performance is achieved, in terms of power consumption for higher degree of centralization i.e. C-RAN compared to traditional RAN approach. However, this comes at the expense of overloading the optical transport to support the FH requirements, leaving limited resources for the BH functions. To address this issue, the secondary optimization objective is to minimize the end-to-end delay in the BH :

$$\min \mathcal{BH}(\boldsymbol{u}, \boldsymbol{\pi}) = \sum_{e \in \mathcal{E}} \left[\mathcal{U}_e - u_{FH,e} - u_{BH,e} \right]^{-1} + \sum_{s \in \mathcal{S}} \left[\mathcal{C}_s - \pi_{FH,s} - \pi_{BH,s} \right]^{-1} \quad (10)$$
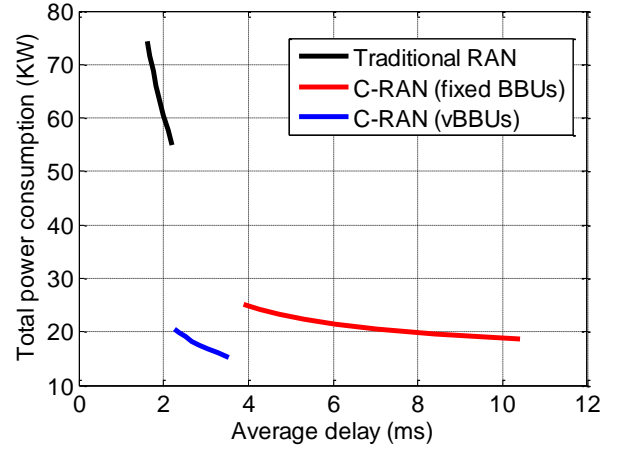

Fig.7. Pareto front of the MOP

subject to demand processing and capacity constraints in the backhaul, where $u_{BH,e}$, $\pi_{BH,s}$ represent the network and server capacity allocated to the BH, respectively.

The MOP problem described through equations (1)-(10) can now be formulated as follows:

$$\min \mathcal{F}(\boldsymbol{u}, \boldsymbol{\pi}) = [\mathcal{FH}(\boldsymbol{u}, \boldsymbol{\pi}), \quad \mathcal{BH}(\boldsymbol{u}, \boldsymbol{\pi})] \quad (11)$$

```
Subject to constraints (1)-(8).
```

This problem, can be transformed from a MOP problem into a single objective optimization using traditional scalarization techniques. For example, if the Pascoletti-Serafini scalarization method is adopted [24], the MOP problem (12), can be written to the following equivalent form:

$$\min \quad t$$

Subject to

$$a + tr - \mathcal{F}(\boldsymbol{u}, \boldsymbol{\pi}) \in \mathbf{K}$$

```
constraints (1)-(8)
```

where $a = [a_1\ a_2]^T \in \mathbb{R}^2$, $r = [r_1\ r_2]^T \in \mathbb{R}^2$ and $\mathbf{K} = \mathbb{R}_+^2$ is the closed pointed convex cone. Then, the problem can be solved using relaxation schemes i.e. Lagrangian Relaxation.

IV. NUMERICAL RESULTS

The proposed MOP scheme is evaluated using the metro optical network topology presented in [13] covering a 10x10 km² area over which 50 BSs are uniformly distributed. End-users served by the BSs generate demands according to real datasets reported in [12]. Fig. 6a presents the evolution of the average traffic per BS for the wireless access domain, respectively. This traffic needs to be processed by specific computing resources. The proposed optimization scheme is focusing on three different scenarios:

a) *"Traditional RAN"* giving emphasis on the optimization of the cloud services supported by the BH. Power consumption per BS ranges between 600 and 1200 Watt under idle and full load conditions, respectively. Small scale commodity servers are deployed for user cloud services.

b) *"C-RAN with fixed BBUs"* where remotely located specialized hardware is used for BBU processing with 200 GOPS capacity/BBU and 1.2W/GOPS power consumption. In this scenario, cloud computing demands originating from the end-users are processed at small scale servers as before.

c) *"C-RAN with virtual BBUs (vBBUs)"* where large-scale commodity servers are used to support both BBU processing (through the creation of vBBUs [4], [7]) and user cloud services.

When adopting the C-RAN approach over the proposed integrated wireless-optical infrastructure and comparing it with the traditional RAN approach, significant energy savings (ranging between 60-75%) can be achieved (Fig. 6b). However, due to overloading of network resources to support FH requirements, C-RAN leads to an increase of the end-to-end service delay in the BH (Fig. 6c), which however remains below 20 ms for a 100 Mbps flow request. It is interesting to note that the BH service delay calculated for the C-RAN vBBU case is lower compared to the delay calculated for the C-RAN fixed BBU case. This is due to the fact that in the C-RAN vBBU case lower processing times are required by the large commodity servers to execute the user cloud services.

Fig. 7 shows the Pareto front indicating optimal operating points of the proposed MOP framework, in terms of energy consumption and end-to-end service delay, for all three scenarios considered. The C-RAN scheme with vBBUs achieves the optimal balance between energy consumption and end-to-end service delay. Traditional RAN provides minimum end-to-end service delays as its functions do not consume any backhaul bandwidth, but suffers high energy consumption due to the lack BBU sharing. The C-RAN with fixed BBUs scheme, offers relatively low energy consumption, but higher delays as execution of end users services is not exploiting the benefit of fast processing times available through the large scale servers.

V. Conclusions

This paper presents the converged optical-wireless 5G network infrastructure interconnecting fixed and mobile users and computational resources to support both operational network (C-RAN) and end-user computational services proposed by 5G-XHAUL. An overarching layered architecture, inspired by the ETSI NFV standard and the SDN reference architecture is also presented. A novel modelling framework has been developed to evaluate the performance of the 5G-XHaul infrastructure. Our study has considered a variety of FH and BH options, spanning from the traditional approach where the two functions are supported separately to solutions involving fully or partially converged FH and BH functions. . Our modelling results show that the proposed architecture can offer significant energy savings but there is a clear trade-off between overall energy consumption and service delay.

References

[1] M. Sauer et al., "Radio over fibre for picocellular network architectures," *IEEE J. on Lightwave Tech.*, vol. 25, no. 11, 2007

[2] J. Chen et al "Demonstration of Analog Millimeter-wave Fronthaul Link for 64-QAM LTE Signal Transmission" *in Proc. IEEE VTC,* Fall 2015

[3] D. Bladsjo, et.al., "Synchronization aspects in LTE Small Cells," *IEEE Comm. Mag.*, vol.51, no.9, pp.70,77, Sep. 2013

[4] C-RAN. The Road Towards Green RAN. White Paper 3.0 (Dec. 2013), http://goo.gl/Sw6bfw

[5] Mobile Fronthaul: Transmode's unique active and passive solution enables mobile operators to migrate to Cloud-RAN architecture

[6] U. Dötsch et al., Quantitative Analysis of Split Base Station Processing and Determination of Advantageous Architectures for LTE, *Bell Labs Tech. J*, vol. 18, no.1, pp. 105-128, May 2013

[7] http://tinyurl.com/nlfup2l

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

[9] J.-Y. Le Boudec, P.Thiran, "Network calculus: a theory of deterministic queuing systems for the internet", Springer, 2001

[10] C. Desset et al., "Flexible power modeling of LTE base stations," *in proc. of IEEE WCNC* 2012, pp.2858-2862, 1-4 April 2012

[11] T. Werthmann, T., "Approaches to adaptively reduce processing effort for LTE Cloud-RAN systems," *in proc. of ICCW*, pp.2701-2707, 8-12, 2015

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

[13] A. Tzanakaki et al., "Virtualization of heterogeneous wireless-optical network and IT infrastructures in support of cloud and mobile cloud services," *IEEE Commun. Mag.* , vol.51, no.8, pp.155-161, Aug. 2013

[14] http://www.etsi.org/technologies-clusters/technologies/nfv

[15] https://www.opennetworking.org/

[16] N. Carapellese, M. Tornatore, A. Pattavina, "Energy-Efficient Baseband Unit Placement in a Fixed/Mobile Converged WDM Aggregation Network," *IEEE JSAC* , vol.32, no.8, pp.1542-1551, Aug. 2014

[17] METIS Project D6.1 [Online] http://tinyurl.com/joohwkt

[18] Bo Han; Gopalakrishnan, V.; Lusheng Ji; Seungjoon Lee, "Network function virtualization: Challenges and opportunities for innovations," *IEEE Commun. Mag.*, vol.53, no.2, pp.90-97, Feb. 2015

[19] ETSI GS NFV-SWA 001 V1.1.1 (2014-12)

[20] 4G Americas. Bringing Network Function Virtualization to LTE. November 2014

[21] D. Wubben et al., "Benefits and Impact of Cloud Computing on 5G Signal Processing: Flexible centralization through cloud-RAN,*" IEEE Signal Proc. Mag.*, vol.31, no.6, pp.35-44, Nov. 2014

[22] A. Tzanakaki et al, "SDN for Converged Access/Aggregation Networks", chapter in book entitled "Fiber-Wireless Convergence in Next-Generation Communication Networks: Systems, Architectures, and Management", M. Tornatore, G. Ellinas, and G-K. Chang (Eds), Springer.

[23] G. Zervas, J. Triay, N. Amaya, Y. Qin, C. Cervelló-Pastor, and D. Simeonidou, "Time Shared Optical Network (TSON): a novel metro architecture for flexible multi-granular services," *Opt. Express* 19, B509-B514 (2011).

[24] A. Pascoletti, P. Serafini, "*Scalarizing vector optimization problems*", *J Optimization Theory Appl*, vol. 42, pp. 499-524, 1984.

[25] M. P. Anastasopoulos, A. Tzanakaki, and D. Simeonidou, "Stochastic Planning of Dependable Virtual Infrastructures Over Optical Datacenter Networks," *J. Opt. Commun. Netw*. 5, 968-979 (2013)