Towards 5G Cloud Radio Access Network: An Energy and Latency Perspective

Riccardo Bassoli, *Member, IEEE,*, Fabrizio Granelli, *Senior Member, IEEE,*,
Sisay Tadesse Arzo and Marco Di Renzo, *Senior Member, IEEE*

Abstract

Future generation networks will entirely deploy virtualisation paradigms to enhance performance and capabilities of current cellular networks. In order to achieve the vision of fifth generation networks, software-defined networking and network function virtualisation will be applied not only at the core network but also at the radio access network. That will help to achieve significant reduction in power consumption while increasing energy efficiency, flexibility and scalability. This article proposes a general mathematical model that can correctly and accurately describe spatial/topological characteristics, power consumption and latency of Cloud radio access network in future generation networks. Thanks to the development of this novel model based on stochastic geometry, tessellation theory and random multilayer hypergraphs, we can numerically estimate the overall energy efficiency (in bit per Joule) of Cloud radio access network in 5G (considering either edge or cloud computing), and we can compare that to energy efficiency of legacy radio access network of current 4G cellular networks. Moreover, the analysis includes a preliminary discussion about latency: that shows edge computing to be the best paradigm for 5G radio access network, that can concurrently satisfy energy efficiency and latency requirements.

Index Terms

5G, Stochastic geometry, Cloud RAN, Cloud computing, Edge computing, Energy efficiency, Latency.

I. Introduction

Next generation cellular networks represent a new vision, which will guarantee higher performance not only in terms of bandwidth but also of latency and reliability.

R. Bassoli, F. Granelli and S. T. Arzo are with the Department of Information Engineering and Computer Science, at the University of Trento, Trento, Italy (e-mail: {riccardo.bassoli,fabrizio.granelli,sisay.arzo}@unitn.it).

M. Di Renzo is with Paris-Saclay University (L2S - CNRS, CentraleSupélec, Univ Paris Sud), Paris, France. (e-mail: marco.direnzo@l2s.centralesupelec.fr).

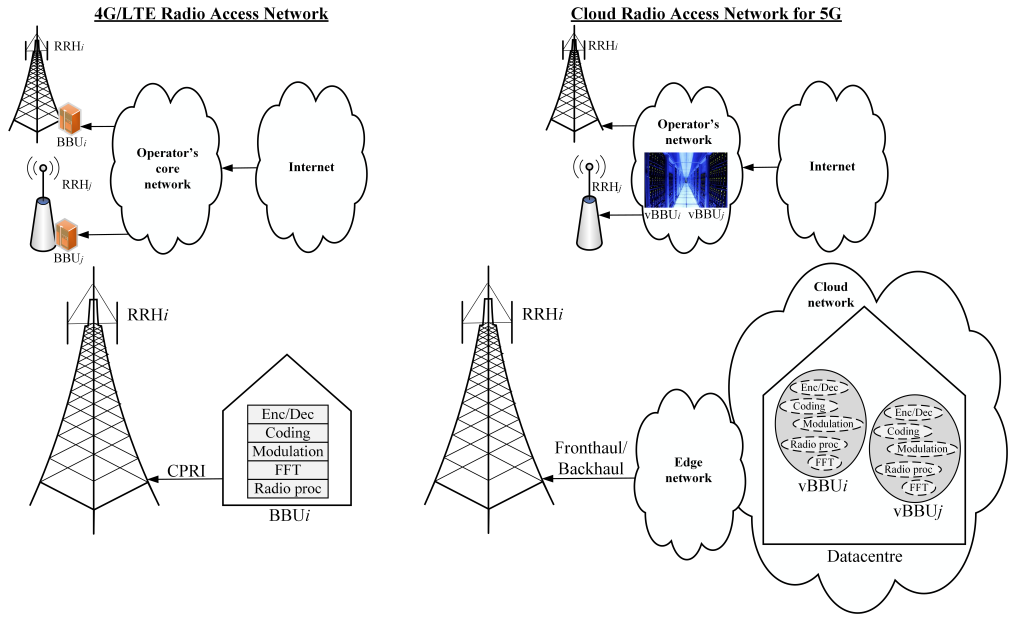

Fig. 1. Downlink communication in heterogeneous 4G/LTE RAN and heterogeneous 5G Cloud RAN. The latter places baseband processing at virtual BBUs in operators' datacentres and run them as virtual machines or virtual functions in containers.

Telecommunications operators aim at achieving those requirements while reducing significantly the expenses due to capital expenditure (CaPEX) and operational expenditure (OPEX). The main means to realise 5G vision while supporting network infrastructure upgrades at an acceptable cost is network virtualisation. In particular, network functions virtualisation (NFV) is the paradigm devoted to mapping specific hardware-based network functions into software-based virtual network functions (VNFs), which are run on general purpose hardware.

Cloud radio access network (C-RAN) [1], [2] is a virtualisation paradigm, which aims at moving RAN and baseband functions and procedures to cloud data centres. That would help to reduce power consumption while increasing energy efficiency of heterogeneous RAN management, deployment and updates.

Figure 1 depicts the idea behind Cloud RAN. Legacy 4G/LTE RAN requires base stations (BSs), which equip a baseband unit (BBU) at each radio site. Nevertheless, this solution is neither scalable nor optimised in large heterogeneous scenarios of future generation networks. On the other hand, by implementing virtual BBUs (v-BBUs), the network achieves higher flexibility in management and configuration of the RAN by detaching baseband processing functionalities from standard BSs: thus, BSs will become pure radio remote heads (RRHs), while baseband processing will be moved to dedicated datacentres with shared processing facilities. This approach is expected to reduce complexity and power consumption of

the RAN. However, the allocation of virtual resources and processing tasks has to be assigned effectively not to increase delays and loads.

In current 4G cellular networks, baseband processing at BBUs [3], [4] includes all the processing due to lower layers of 4G protocol stack. The operations of a BBU involve physical layer processing (4G baseband signal processing components include ASICs, DSPs, microcontrollers, and FPGAs), smart antennas and multi-user detection required to reduce interference, modulation/demodulation, error correction coding (which increases the complexity of the baseband processing at the receiver), radio scheduling, encryption/decryption of packet data convergence protocol (PDCP) communication (both downlink and uplink). Multi-carrier modulation (MCM) is also a baseband process. The subcarriers are created using IFFT in the transmitter, and FFT is used in the receiver to recover the data. A fast DSP is needed for parsing and processing the data. Multi-user detection (MUD) is used to eliminate the multiple access interference (MAI) present in CDMA systems.

Based on preliminary results in [5] and on the initial model published in 2018 at European Wireless conference [6], the main contribution of the article includes a comprehensive and rigorous mathematical model to study C-RAN in the context of 5G cloud and edge computing. The proposed model considers spatial/geographic information to analyse performance such as energy efficiency and latency, which are fundamental targets in the design of future generation cellular networks. Given that, this paper enhances and generalises current models for C-RAN in the literature. To the best of authors' knowledge, such a complete model, based on multilayer random hypergraphs considering power consumption of all areas of the network (included data centres), has never been proposed by now. Next, the work discusses quantitative/analytical comparison between current 4G/LTE RAN and future generation virtual networks with C-RAN, including analysis of total latency of C-RAN in case of 5G edge and cloud computing. It is important to underline that this paper analyses the performance of C-RAN referred to downlink communications. The contribution of uplink communications to BBU processing is not considered.

The article is organised as follows. Section II provides a detailed analysis of C-RAN theoretical models towards evaluation of power consumption and latency. Various works were selected, which represent the spectrum of kinds of models, which are in the current state-of-the-art. In particular, Subsection II-A highlights the details of each model of C-RAN and its power consumption and, eventually, latency. Then, Subsection II-B focuses on motivation and contributions in order to justify the need of content presented in the remainder of the paper. Section III describes in detail the proposed novel model based on random multilayer hypergraphs. In particular, Subsection III-A structures the model of power consumption and energy efficiency of 5G C-RAN system, while Subsection III-B describes the model of latency. Finally, Section IV discusses results referred to power consumption/energy efficiency and latency of C-RAN in

the context of 5G (edge and cloud computing), and compare them with the ones referred to 4G/LTE RAN.

II. Related Works and Motivation

This section first presents a selection of works in the literature in order to describe the status of theoretical models on C-RAN research. Second, it highlights the main open issues while justifying the motivation behind our article and the effective contribution it provides, towards an accurate theoretical description of C-RAN in 5G. While Subsection II-A does not strive to be a survey on C-RAN models, its idea is to clarify why stochastic geometry and, subsequently, a more generalised and comprehensive model are needed in C-RAN theory. The following analysis only takes into account the contribution of the works in terms of (i) modelling C-RAN from cellular network (system) point of view (i) modelling data centres and virtualisation of BBUs and (iii) modelling power consumption of C-RAN. Their additional contributions are neglected since they are not in the scope of this work.

A. Related Works

In 2014, authors of [7] considered a scenario where macro cells are replaced with small cells; moreover, BBU processing is virtualised to cloud data centres. They provide an accurate model for power consumption of cellular networks where the overall power is

$$P_{tot} = P_{vRAN} + P_{bh} + P_{RAN} \tag{1}$$

where P_{vRAN} is the power consumed by RAN virtualisation, P_{bh} is the one due to backhaul (or fronthaul) and P_{RAN} is the one consumed by BSs. In order to estimate P_{vRAN}, they approximate data centre (server) power consumption versus its CPU percentage of usage as

$$P_{srv} = P_0^{srv} + \delta_p^{srv} P_{max}^{srv} x_{srv} \tag{2}$$

where P_0^{srv} and P_{max}^{srv} are the power consumption of the server in idle mode and maximum usage respectively, δ_p^{srv} denotes the slope of the equivalent power model of the considered server and x_{srv} is the CPU percentage of usage.

Next, paper [8] modelled C-RAN in heterogeneous cellular scenario where macro RRHs are regularly distributed as hexagonal cells and pico RRHs are circles inside macro cells. A baseband resource pool is connected via a switch (and managed by a centre management unit) to the RRHs, which are connected to the pool via Ethernet of 10 Gb/s.

In 2015, article [9] modelled C-RAN scenario as a heterogeneous network including cloud data centres and heterogeneous BSs, which serve a vector of mobile users. The same year, authors in [10] described

the cellular network as composed by homogeneous RRHs connected to a number of BBUs with equal processing capacity, measured in Mega operations per time slots (MOPTS). Next, the computing resources of a BBU j used by cell i in MOPTS is defined as

$$L_{i,j} = \sum_{n=1}^{N} \beta_{i,j,n} L_{i,n}^{req} \tag{3}$$

where $L_{i,n}^{req}$ is the computing resource needed for task n at cell i, $\beta_{i,j,n} \in \{0,1\}$ is '1' if the task n for cell i is processed by BBU j and '0' otherwise. The model considers tasks can be performed either by a single BBU or multiple BBUs. In the second case, BBUs require additional computing resources for transmission among them. These additional computing resources are defined as

$$C_{i,j} = \begin{cases} 0, \sum_{n=1}^{N} \beta_{i,j,n} = 0,1 \\ \delta_{cost}, otherwise \end{cases} \tag{4}$$

where δ_{cost} is a constant for the communications between BBUs (measured in MOPTS).

In 2016, article [11] defined a model for power consumption of C-RAN by considering the contribution of components of core network (CN) and RRH as $P_{C-RAN} = P_{CN} + P_{RRH}$. Power consumption of RRH is defined as $P_{RRH} = P_{CN} + P_{BS}$, where P_{BS} contains all the components referred to RF, power amplifier (PA), AC-DC and DC-DC voltage conversion, optical transceivers and cooling [12]. On the other hand, the model of power consumption of core network is given by addition of contributions dependent on cooling (P_{cool}), main supply (P_{MS}), DC conversion (P_{DC}), software-defined networking (SDN) (P_{SDN}), SDN controller (P_{cl}), BBUs (P_{BBU}) and the optical devices (P_{opt}). This model somehow can capture the power consumption considering the service diversity and dynamic mapping of RRH-BBUs connections. In particular, the power consumption of a BBU is defined as

$$P_{BBU} = \sum_{i \in I_{BBU}} P_{i,BBU}^{ref} A^{x_i^A} B^{x_i^B} \tag{5}$$

where I_{BBU} is the set of different functions performed by BBUs, measures in Giga operations per second (GOPS), $P_{i,BBU}^{ref}$ is the power consumption of i-th function, A is the total number of antennas/RF transceivers, x_i^A is the scaling exponent of the number of RF chains of the BBU, B is the share of the used bandwidth (measured in Hz) and x_i^B is the scaling exponent of B. Next, the power consumed by SDN equipment is modelled as

$$P_{SDN} = P_{switch} + P_{SDNctl} \tag{6}$$

where P_{switch} is the power consumed by switches (sum of traffic power consumption, P_{flow}, and ports' power consumption P_{port}) and P_{SDNctl} is the one consumed by controller.

Next, the system model, proposed by [13], modelled the network via a single tier cellular network, with macro-hexagonal regular cells, composed by RRHs containing nine omnidirectional antennas. Next, the BBUs are virtualised and co-located in a single pool. Later, article [14] modelled 5G C-RAN as a single-tier cellular network with RRHs transmitting at 31 dBm. The authors use server IBM X3650 to host virtual BSs of their prototype. The model describes the total energy consumption of an RRH serving n users as

$$E_{tot} = E_{const} t_{on} + \sum_{i=1}^{n} E_i t_i + E_{idle} t_{idle} \tag{7}$$

where E_{const} is the constant power consumption of RRHs, t_{on} is the time of power-on of RRHs, E_{idle} is the power consumption in idle mode, t_{idle} is the idle time of RRHs and E_i is transmission power of a mobile user.

The authors in [15] analysed C-RAN by modelling session-level dynamics of virtual BSs via Markov model. In particular, heterogeneous virtual BSs are consolidated in a data centre and share a number of units, providing computational resources. Next, the same year, paper [16] studied C-RAN in a single tier cellular network. It considers BS and mobile users randomly distributed according to two Poisson point processes, of density λ_U and λ_R, into d-dimensional space. Each RRH is equipped with M antennas and each mobile user with a single one. Their stochastic-geometric model is based on [17]. Moreover, the proposed latency model for C-RAN is defined by

$$\Delta t = \omega_1 \Delta t_{ce} + \omega_2 (\Delta t_{fb} + \Delta t_{pt}) + \Delta t_{pc} + \Delta t_{pRRH} + \omega_3 \Delta t_{BH} \tag{8}$$

where Δt_{ce} is the channel estimation delay, Δt_{fb} is the average per-channel coefficient feedback delay, Δt_{pt} is the propagation delay, Δt_{pc} is the cloud processing delay, Δt_{pRRH} is the RRH processing delay, Δt_{BH} is the backhaul delay per hop, ω_1 is the number of channel coefficients to be estimated for a mobile user, ω_2 is the total number of times channel state information (CSI) is to be fed back for the whole network and ω_3 number of backhaul hops. Afterwards, article [18] considered a system composed by two subsystems C-RAN and cloud computing, which are connected via either optical or wireless backhaul. Cloud computing is represented by a virtual BS pool while C-RAN is composed by a number of RRHs (single tier) with a unique antenna.

In 2017, authors of [19] proposed a description of C-RAN consisting of small-cell RRHs serving the user equipments (UEs) in their cells. Each mobile user has a task, defined as

$$U = (F, D) \tag{9}$$

where F is the total number of CPU cycles needed to complete the task U and D is the whole size of the task for the transmitting data. Then, the delay to complete a task becomes

$$T = \frac{F}{f} + \frac{D}{r} \tag{10}$$

where f is the computational capacity allocated to the mobile user for task U and r is the data rate of the UE. Next, the energy cost of a task of a mobile user is defined as

$$E = \varphi(f)^{\vartheta - 1} F + \eta P \left(\frac{D}{r} \right) \tag{11}$$

where P is the transmission power of an RRH, φ is the effective switched capacitance, ϑ is a positive constant and η is a weighted trade-off between energy consumption of a mobile cloud and C-RAN. Finally, the authors provide expression of the signal-to-interference-plus-noise ratio (SINR) in order to estimate the data rate of a mobile user connected to its serving RRH. The same year, article [20] described C-RAN via a set of RRHs, with a demand for traffic processing, connected to a set of candidate sites, which host the BBU pool. The latency due to communication between RRH and BBU is a fixed constrain. Next, authors in [21] studied architecture of C-RAN as a set of RRHs connected to a BBU pool via optical fibres. In particular, the virtual pool contains a set of physical servers, each of which hosts a number of CPU cores. Given these premises, the article enhances power consumption model [12] of BSs.

Afterwards, work [22] analysed C-RAN consisting of heterogeneous RRHs (e.g. macro and pico) regularly distributed to form hexagonal grid (macro cells), which contain various small cells. Each RRH is equipped with a number of transmitting antennas while the UE has one receiving antenna. Information comes from the backbone network towards the mobile user (downlink). The RRH are connected to the BBU pool via switch using Ethernet of 10 Gb/s.

Finally, paper [23] modelled C-RAN as a system with a unique macro cell, containing various small cells. Each RRH is connected to the BBU pool in the core network via backhaul/fronthaul links. Each virtual BBU is associated with one UE and has specific computational capacity, expressed in terms of user's data rate.

B. Motivation and Contribution

Subsection II-A described some examples of theoretical system models in detail, which have been used to study properties and performances of C-RAN itself or C-RAN in the context of 5G. As it is possible to see, while they are correct and suitable to analyse very specific aspects of power consumption and latency, they cannot capture the complexity of system and requirements of 5G C-RAN. First, future 5G networks [24] will be heterogeneous networks, where the distribution of different kinds of BSs will not be

regular. Furthermore, in this scenario, C-RAN involves not only the wireless access network but also wired networks and sub-networks (data centre internal architecture): thus, a correct system-level analysis of C-RAN should provide spatial-topological information of the networks, while capturing the heterogeneity of nodes and links. Next, C-RAN in 5G networks, considering 5G key performance indicators (KPIs), cannot be correctly investigated and studied without a system-level analysis because of end-to-end nature of performance in 5G: in fact, characteristics of areas in the network can affect performances of other parts, in terms of specific KPIs.

According to these premises, we can identify four main open issues in theoretical research about 5G C-RAN, which arise from the detailed description of Subsection II-A:

- *5G C-RAN is not modelled as heterogeneous system with spatial information*. The study of C-RAN, in the context of 5G, cannot neglect the characterisation of signal-to-noise-plus-interference-ratio (SINR), which requires knowledge of network geometry [25], [26]. In order to circumvent the difficulty to characterise SINR, stochastic geometry and random graphs were proposed. Regular models of radio coverage (e.g. hexagonal and square lattices) were used in the past but they are highly inaccurate for heterogeneous networks in urban and suburban scenarios, where cells' radii considerably change because of transmission power and density. The previous subsection has showed that the main methods used in research to model C-RAN were based on regular mathematical structures, thus resulting inaccurate.
- *Virtualisation of RAN is not contextualised in a framework, that models the actual architecture of a data centre*. To the best of authors' knowledge no existing work, that deals with 5G C-RAN, has flexibly analysed how data centre's architecture, interacting with rest of the network, affects performance of C-RAN.
- *The evaluation of power and latency does not consider all the parts of the network.* The works, which were previously listed, do not consider the contribution of all the areas of the network in evaluation of characteristics of C-RAN. Especially, the different impact of edge and cloud computing or the specific architecture of the data centre are frequently neglected.
- *In 5G C-RAN investigation, it is not analysed the trade-off between power consumption/energy efficiency and latency*. The works about C-RAN listed in previous subsection analyse either power or latency in C-RAN while not considering the combination of them towards a placement of BBU VNFs in the wired operators' network. Our analysis permits to identify and to justify where to perform baseband processing (either edge or cloud) and why.

In respect to the above open issues, the contribution of this article includes:

- Section III. A general, flexible, coherent and comprehensive mathematical model, capable to capture the intrinsic and complex characteristics of 5G C-RAN. This model, based on random multilayer hypergraphs, can include and merge different specific theoretic tool (e.g. stochastic geometry) to investigate effectively the complexity and heterogeneity of 5G C-RAN as a system.
- Subsection III-A and Section IV. A power consumption model, included in and supported by the random multilayer hypergraph, which permits a reasonably detailed study of power consumption and energy efficiency of 5G C-RAN as a system. This model considers the contributions referred to RAN, backhaul/fronthaul, edge, core and data centres during downlink communications. Moreover, it includes a detailed characterisation of baseband processing requirements because of UEs, provided by [27].
- Subsection III-B and Section IV. Since 5G KPIs are not to be satisfied singularly but concurrently, the analysis in terms of power consumption and energy efficiency is drawn up to an evaluation of latency. That helps to complete and to detail the final considerations about 5G C-RAN.

These contributions will help towards a more significant and accurate modelling and characterisation of C-RAN properties and behaviour in 5G.

III. 5G System Model

Graph theory is the area of mathematics that has allowed effective modelling of communication networks as a whole. Wired networks have always been modelled as planar graphs, composed by a set of nodes (e.g. switches, routers, etc.) and a set of edges (i.e. wired links). Side by side, a planar hypergraph is a graph's generalisation where edges can connect group of nodes to each others (i.e. not connecting only two nodes as in normal graphs). By the advent of stochastic geometry and random graphs to model wireless cellular networks, hypergraphs have lost their central role in modelling wireless networks. However, while random graphs are useful to model the nature of legacy access cellular networks, the complexity of virtual networks in 5G requires a more complex and flexible architecture: in fact, the theoretical description should be able to consider random wireless links and fixed wired links in the same multilevel scenario. That's why this article proposes a new generalised model to study effectively C-RAN in future 5G networks based on very general multilayer random hypergraphs.

The 5G reference scenario of this paper is a multi-tier heterogeneous cellular network, which comprises different kinds of BSs. Next, there are data centres, which can be located either in core network (called large data centre, cloud computing) or in edge network (small data centre, edge/fog computing), hosting v-BBUs, which can run as VNFs in virtual machines or containers. According to preliminary research in [5], [6] and to previous discussion in Subsection II-B, we propose to model virtualised RAN via a

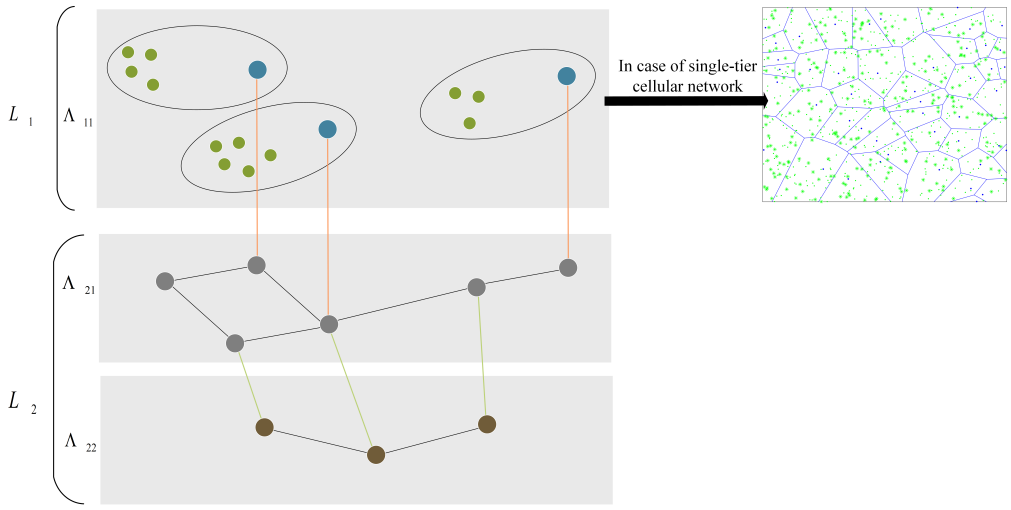

Fig. 2. Example of structure of a multilayer hypergraph (on the left). The subsets of nodes are represented with different colours.

random multilayer hypergraph, a mathematical object that can flexibly describe the properties and the characteristics of 5G C-RAN.

Let $H = (X, E)$ be a planar hypergraph representing the physical network, where X is the set of nodes and E is the set of non-empty subsets of X, called hyperedges. Next, the set X can be partitioned into subsets $X = \{X_1, X_2, \ldots\}$ respectively referred to mobile end users, BSs, network nodes and internal nodes of data centres' network (hosting the v-BBUs).

Let $\mathcal{H} = (X, E, X_{\mathcal{H}}, E_{\mathcal{H}}, L)$ be a multilayer random hypergraph where

- X is the set of random nodes, which can be distributed according to either random point processes (e.g. BSs) or deterministic spatial distributions (e.g. wired operator's network);
- E is the set of random hyperedges, whose cardinality $|E_i|$ can be defined by either Voronoi tessellation in $\mathbb{R}^2$ (e.g. wireless cellular networks) or deterministic values (e.g. links in wired networks);
- $L = \{L_1, \ldots, L_a\}$ is the set of layers, where a is the number of aspects; each layer can be a set of sub-layers Λ_{ij}, where i is the number of layer it belongs to and j is the number of sub-layer ($j = 1, \ldots, |L_i|$);
- $X_{\mathcal{H}}$ is the set of node-layer elements;
- $E_{\mathcal{H}}$ is the set of hyperedge-layer elements.

Figure 2 depicts an example of random multilayer hypergraph. This example of hypergraph has two layers L_1 and L_2 ($a = 2$), where L_1 is composed by a single sub-layer Λ_{11} and L_2 is composed by two sub-layers Λ_{21} and Λ_{22}. If L_1 represents a single-tier cellular network, its hyperedges can be identified via Voronoi tessellation. In the rest of the article, since we will work on multi-tier networks, the specific

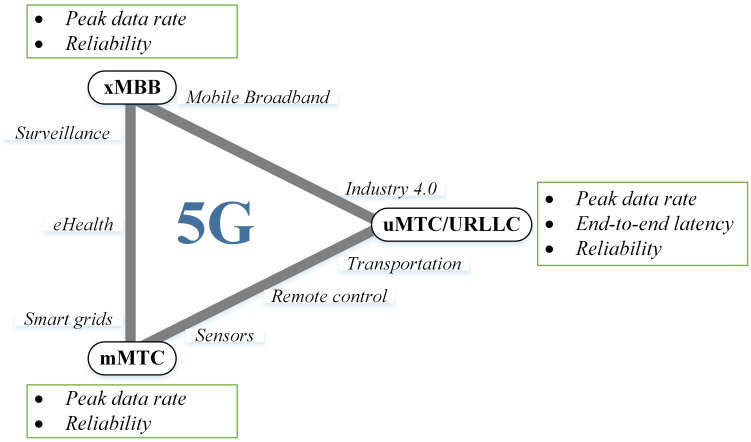

Fig. 3. 5G legacy classification of end-users according to service requirements.

tessellation will be a multiplicatively-weighted (MW) Voronoi tessellation. Side by side, L_2 sub-layers are planar graphs, modelling different areas of wired network. The red links, connecting blue and green nodes, may model backhaul links. If this links had been wireless backhaul links (random hyperedges), they would have been represented via another Voronoi tessellation (that is the case considered in the analytical evaluation below).

Next, Figure 3 depicts the legacy classification of end users in future 5G networks, which are divided into three main categories. Extreme Mobile Broadband (xMBB) will enhance significantly current support for mobile broadband and mobile video streaming mainly in terms of bandwidth. Next, ultra-reliable Machine-Type Communications (uMTCs) or ultra-reliable low-latency communications (URLLCs) represent the major framework for verticals referred to transportations and industry 4.0. Their requirements are mainly focused on bandwidth, latency and reliability. Finally, massive Machine-Type Communications (mMTCs) will support all the universe of Internet-of-Things (IoT), eHealth, smart grids and surveillance. These verticals' requirements are significantly focused on bandwidth supply for massive number of devices and reliability of the communications.

Given these premises, each node representing a mobile end user is identified by the ith commodity flow, thus a quadruple (s_i, σ_i, D_i), where $s_i \in S$ is the source (S is the set of sources) and $\sigma_i \in \Sigma$ is the sink (Σ is the set of sinks). Then, let D_i be the *demand set*, which defines the attributes of mobile end user i. To the best of authors' knowledge, there are no reliable traffic models for uMTC and mMTC services. Then, we consider only xMBB users in the evaluation of the next sections. In fact, a reasonable traffic model for xMBB end users can be established by using statistics provided in [28]. Figure 4 shows the fraction of xMBB end users, which are active in average during the hours of the day in Europe.

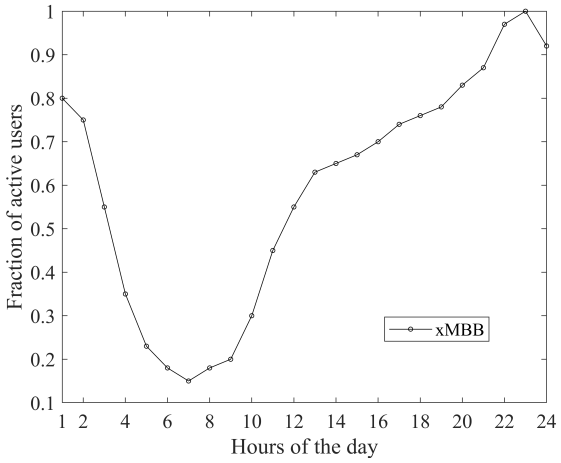

Fig. 4. Average variation of fraction of active end users (xMBB) according to the hours of the day.

A. Model of Power Consumption

The general system model of 5G C-RAN described above is now specified to estimate the power consumption. Let's consider a three-tier heterogeneous cellular network, with nodes belonging to X_1, X_2 and X_3 (subsets of X) following three homogeneous PPP Φ_{BS_m} (micro BSs), Φ_{BS_p} (pico BSs) and Φ_{BS_f} (femto BSs) of intensity λ_{BS_m}, λ_{BS_p} and λ_{BS_f} respectively. Next, let Φ_{mw-sw} be the homogeneous PPP, with intensity λ_{mw-sw}, representing the spatial distribution of microwave aggregate switches for wireless backhaul: in particular, BSs connects to the nearest aggregate switch. Finally, let Φ_{xMBB} be the homogeneous PPP describing the distribution of mobile broadband users, of intensity λ_{xMBB}. All the PPPs Φ_{BS_m}, Φ_{BS_p}, Φ_{BS_f}, Φ_{mw-sw} and Φ_{xMBB} are assumed to be independent. In the network, each end user is associated to the nearest BS.

Given the heterogeneous transmit power (and as a consequence difference transmission range) of BSs belonging to different tiers, the coverage is modelled using a multiplicatively-weighted Voronoi tessellation, since the points of Φ_{BS_m}, Φ_{BS_p} and Φ_{BS_f} have different weights [29]. It is important to notice that this article focuses on downlink baseband communications. Given that, the hyperedges of L_1 follow a MW Voronoi tessellation. On the other hand, the hyperedges at L_2 follow a Voronoi tessellation. The *average fraction of nodes (xMBB users) served by j-th tier* [29] can be expressed as

$$N_j = \frac{\lambda_j {P_{trx,j}}^{2/\alpha} {\theta_j}^{2/\alpha}}{\sum_{i=1}^{3} \lambda_i {P_{trx,i}}^{2/\alpha} {\theta_i}^{2/\alpha}} \tag{12}$$

where $P_{trx,i}$ is the transmission power of the ith tier, θ_i is the signal-to-interference-plus-noise ratio (SINR) threshold and α is the path loss exponent: these are attributes referred to nodes, which identify

BSs. As a consequence, each BS of the jth tier has an average load of N_j/λ_j. In order to minimise the propagation delay, we assume BSs are connected to the nearest aggregation switch of backhaul via wireless link. This implies that aggregation switches at backhaul, belonging to subset X_4, serve the BSs that are placed in their respective Voronoi cell (i.e. connected via random hyperedge). The probability mass function (pmf) of the number of nodes (BSs) that are connected to an aggregate switch [30] is N_{BS}, expressed as

$$P\left[N_{BS}=n\right]=\frac{3.5^{3.5}\Gamma\left(n+3.5\right)\left(\lambda_{BS}/\lambda_{mw-sw}\right)^{n}}{\Gamma\left(3.5\right)n!\left(\lambda_{BS}/\lambda_{mw-sw}+3.5\right)^{n+3.5}} \tag{13}$$

where λ_{BS} is the sum of all the intensities of BSs and $\Gamma\left(x\right)$ represents the gamma function.

Cloud RAN paradigm will be a subsystem of future 5G networks, involving four main areas: RAN (P_{RAN}), backhaul/fronthaul (P_{bh}), edge network and core network (cloud). Thus, when power consumption of C-RAN is evaluated, it is important that the contribution of each area is included such that

$$P_{tot5G}=P_{RAN}+P_{bh}+P_{net}+P_{dc} \tag{14}$$

where P_{dc} is the average power consumed by data centre either in the core (cloud computing) or in the edge (edge computing).

An accurate and detailed model to study power consumption of legacy multi-tier 4G cellular networks is published in [12], [28]. In particular, the linear approximation of the *power consumption of a BS* [28] (this is an attribute of BSs nodes) can be expressed as

$$P_{BS}=N_{trx}\left((1-\rho)P_{BSidle}+\rho\Delta_{p}P_{BSmax}\right) \tag{15}$$

where N_{trx} is the number of transmission chains (i.e. ratio between transmit and receive antennas per site), P_{BSidle} is the power consumption calculated at the minimum possible power, Δ_p is the slope of load dependent power consumption P_{BSmax} is the the maximum RF output power at maximum load and ρ is the fraction of load variation. This parameter is referred to the variable fraction of active users, which follows the pattern in Figure 4, according to different hours of the day. In 5G, the power consumption of RAN only considers the contribution of RRHs since BBU is virtualised, while each legacy 4G/LTE BS has to consider BBU power consumption. Next, the *total power consumption of the RAN* is the sum of P_{BS} of all the BSs.

Since 5G will be a virtualised network, the *power consumption of backhaul/fronthaul* will be mainly affected by the number of switches aggregating traffic and connecting the BS with data centre. Moreover, it should be added the contribution due to microwave antennas connecting RRHs and backhaul network. Then, power consumption P_{bh} (attribute of aggregate switch nodes) [7] can be estimated as

$$P_{bh}=\sum_{n=1}^{N_{cell}}P_{sw}^{n}+N_{mw}^{n}P_{link}^{n} \tag{16}$$

where N_{cell} is the number of aggregation switches, P^n_{sw} is the power consumed by aggregation switches, N^n_{mw} is the number of antennas to transmit/receive aggregate backhaul traffic and P^n_{link} is the power consumption of backhaul links. Variable P_{net} represents the power consumption due to the network between the backhaul and the data centre, thus contribution of either edge or edge and core networks. The nodes belonging to edge and core networks belongs respectively to X_5 and X_6 (their subsequent idle and maximum powers are respective attributes assigned to these nodes). Next, P_{net} can be estimated as

$$P_{net} = (1-\rho)\left(h_e P_{e-idle} + h_c P_{c-idle}\right) + \rho\left(h_e P_{e-max} + h_c P_{c-max}\right) \tag{17}$$

where h_e is the number of hops in the edge network P_{e-idle} and P_{e-max} are the power consumptions of an edge router in idle and maximum load status respectively, h_c is the number of hops in the core network and P_{c-idle} and P_{c-max} is the power consumption of a core router in idle and maximum load status respectively. Next, the power consumption P_{dc} depends on the number of switches and servers, which compose the data centre: in particular, the number of processing servers depends on the processing load, required by each mobile user at a specific time. This load can be estimated as [27]

$$p_{UE} = \left(3A + A^2 + \frac{1}{3}MCL\right)\frac{R}{10} \tag{18}$$

where A is the number of antennas, M the modulation bits, C the code rate, L the number of spatial MIMO-layers and R the number of physical resource blocks (PRBs). The processing load p_{UE} is measured in GOPS. Variable p_{UE} is attribute belonging to demand vector D_i. By considering a data centre with a three-tier structure and the linear approximation in [31], the *power consumption of a data centre* P_{dc} can be evaluated as

$$P_{dc} = P_{dc-sw} + P_{dc-s} \tag{19}$$

where the linear approximation of the average power consumption of switches is

$$P_{dc-sw} = (1-\rho)P_{sw-idle} + \rho P_{sw-max} \tag{20}$$

and the linear approximation of the average power consumption of servers is

$$P_s = (1-\rho)P_{s-idle} + \rho P_{s-max} \tag{21}$$

where $P_{sw-idle}$ and P_{s-idle} are the power consumption in idle mode and P_{sw-max} and P_{s-max} are the power consumption at maximum load for switches and servers of data centre's network respectively. Switches and servers in data centre's network belongs to subsets X_7 and X_8, and their idle and maximum power consumption are nodes' attributes respectively assigned to them.

Finally, the *total energy efficiency* [32] of the 5G network is calculated as

$$EE = \frac{C}{P_{tot5G}} \tag{22}$$

Fig. 5. Snapshot of Manchester's map, obtained from Google maps. The line shows the distance between north and south of the city centre in order to have an idea of the order of magnitude of involved distances.

where C is the transmission capacity (measured in b/s).

B. Model of Latency

While legacy 4G/LTE networks places BBUs at each BS just connected with CPRI wire, future 5G networks will employ v-BBUs located in data centres either in the edge or in the cloud. That implies additional delays for transmission via the edge and/or core network towards the data centre. Thus, the total latency of future 5G networks can be expressed as

$$\tau_{5G} = \tau_{RAN} + \tau_{bh} + \tau_{edge} + \tau_{core} + \tau_{dc} \tag{23}$$

where τ_{RAN} is the time for transmission between RRH and UE, τ_{bh} is the delay due to wireless backhaul link, τ_{edge} is the time due to transmission on edge networks, τ_{core} is the time due to transmission via the core network and τ_{dc} is the latency at the data centre. In particular, τ_{edge} and τ_{core} are the combined contribution of propagation delay and load delay while τ_{dc} considers processing delay and propagation delay inside data centre's network. These delays are attributes assigned to respective hyperedges, belonging to set E.

IV. Results and Discussions

The urban scenario considered in this article is based on the available data from the city of Manchester (Figure 5). Given the statistics for the city of Manchester provided by [33], we consider a density of 37 BS/$\mathcal{A}$, where $\mathcal{A}$ = 1.8 km^2. If we consider the city centre as a square of side 15 km (see Figure 5),

TABLE I

PARAMETERS FOR EVALUATION DEPENDING ON THE TIER [28].

	Micro	Pico	Femto
N_{trx}	2	2	2
Δ_p	3.1	4	7.5
P_{BSidle} (W)	6.3	0.13	0.05
P_{BSmax} (W)	53	6.8	4.8
P_{BBU} (W)	27.3	3	2.5
P_{trx} (W)	3.4	0.4	0.2
α	3	3	3
θ	4	2	1

TABLE II

PARAMETERS FOR EVALUATION BASEBAND PROCESSING LOAD PER UE [27].

A	M	C	L	R
2	4 [16QAM] 6 [64QAM]	3/4	2	2 [5 MHz – 25 PRBs] 4 [10 MHz – 50 PRBs] 6 [15 MHz – 75 PRBs] 9 [20 MHz – 100 PRBs]

we have about 125 areas $\mathcal{A}$ in the city centre, containing 4625 BSs in total. Let's consider a three-tier cellular network, consisting of micro, pico and femto BSs. According to their technical specifications, it is reasonable to split the 37 BS/$\mathcal{A}$ as λ_{BS_m} = 2 BS/$\mathcal{A}$, λ_{BS_p} = 7 BS/$\mathcal{A}$ and λ_{BS_f} = 27 BS/$\mathcal{A}$. It is important to notice that we do not consider the presence of mmWave base stations in this article. Moreover, the results presented in [33] allow to model correctly the distribution of BSs as independent two-dimensional homogeneous Poisson point processes (PPP) on an Euclidean plane $\mathbb{R}^2$, called Φ_{BS_m}, Φ_{BS_p} and Φ_{BS_f}, where λ_{BS_m}, λ_{BS_p} and λ_{BS_f} are the respective densities of the point processes.

Next, Figure 4 depicts the average variation of density of active xMBB UEs according to the hours of the day (i.e. the hourly variation of λ_{xMBB}). Regarding the density of end users, λ_{xMBB} can be assumed to be ten times the number of BSs [33]. In order to make the comparison between 4G/LTE and 5G consistent, we only assume the contribution of xMBB users, neglecting uMTC and mMTC since 4G networks do not support low-latency ultra-reliable and massive communications.

Table II lists the parameters to evaluate baseband processing of each xMBB user. The different frequencies of transmission implies different number of available PRBs per slot: since all the mobile

TABLE III

PARAMETERS FOR NUMERICAL EVALUATION OF BACKHAUL, EDGE, CORE AND DATA CENTRE'S POWER CONSUMPTION.

P_s [7]	53 W
$f_{cell-bh}$ [7]	128%
Y_{max} [7]	84.4 Mb/s
C_{sw} [7]	36 Gb/s
P^n_{link} [7]	22.2 W (idle) 37 W (low traffic) 92.5 W (high traffic)
N_{mw} [7]	2
Edge router [34]	P_{e-idle} = 4095 W P_{e-max} = 4550 W h_e = 3
Core router [34]	P_{c-idle} = 11070 W P_{c-max} = 12300 W h_c = 6
P_{dc-sw} [31]	$P_{sw-idle}$ = 200 W P_{sw-max} = 300 W
P_{dc-s} [31]	P_{s-idle} = 544 W P_{s-max} = 750 W

users are assumed transmitting at same rate and with equal importance, we schedule the same number of PRBs per each user. According to parameters in Table II, it also possible to identify the transmission rate of each user[1].

Table III summarises the values of parameters for numerical evaluation of backhaul, edge, core and data centre's power consumption in Matlab. Regarding the specifications of processing capacity of a data centre's server and its conversion in GOPS, we can assume that each server of the three-tier data centre [35, Section 4] has a capacity of 54 GOPS. Given these premises, the number of servers daily changes as depicted in Figure 6. Next, given the processing load required by xMBB users, the architecture of the data centre and the number of active servers, it is possible to estimate the average variation of power consumption at the data centre during the day according to equation (21). In particular, Figure 7 shows the average power consumed by the three-tier data centre during the day in order to satisfy baseband processing requirements of xMBB users.

[1]The transmission rate is $\frac{PRB \cdot M \cdot R \cdot sub \cdot cp}{\tau_{slot}}$, where PRB is the number of PRBs, R is the coding rate, sub is the number of subcarriers, cp is the number of CP symbols and τ_{slot} is the duration of the slot (0.5 ms).

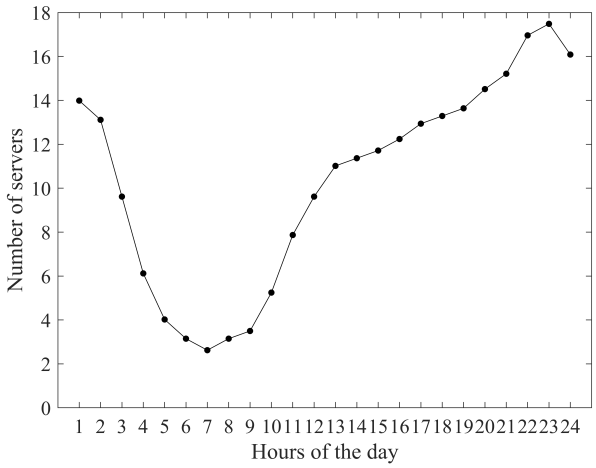

Fig. 6. Number of active servers at the three-tier data centre, which are required to support the traffic load due to xMBB end users.

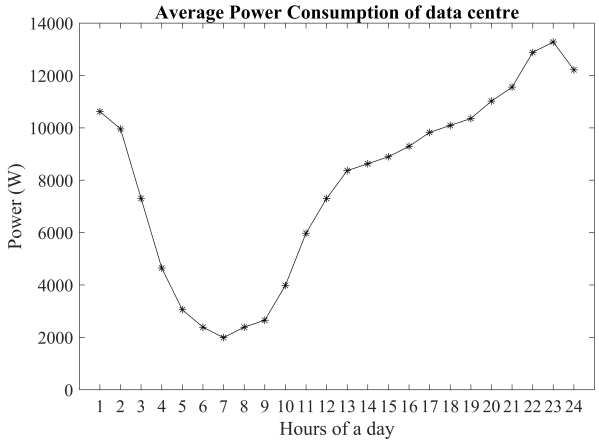

Fig. 7. Variation of average power consumption of three-tier data centre according to hours of the day.

Next, given the model in Subsection III-A and the values in Table I, II and III, we can evaluate the total average power consumption of 4G/LTE and 5G C-RAN (edge and cloud computing). Figure 8(d) describes the comparison between the three architectures. Virtualisation of RAN can significantly reduce the total power consumption of future 5G networks but only in case of edge computing: in fact, 5G C-RAN based on cloud computing slightly increases the average total power consumption of the system. Why does this happen? In order to understand that, we need to look at Figure 8 as a whole. 4G/LTE network has higher power consumption at RAN than 5G because it has BBU for each BS. However, it has lower power consumption at backhaul, edge and core network since their devices do

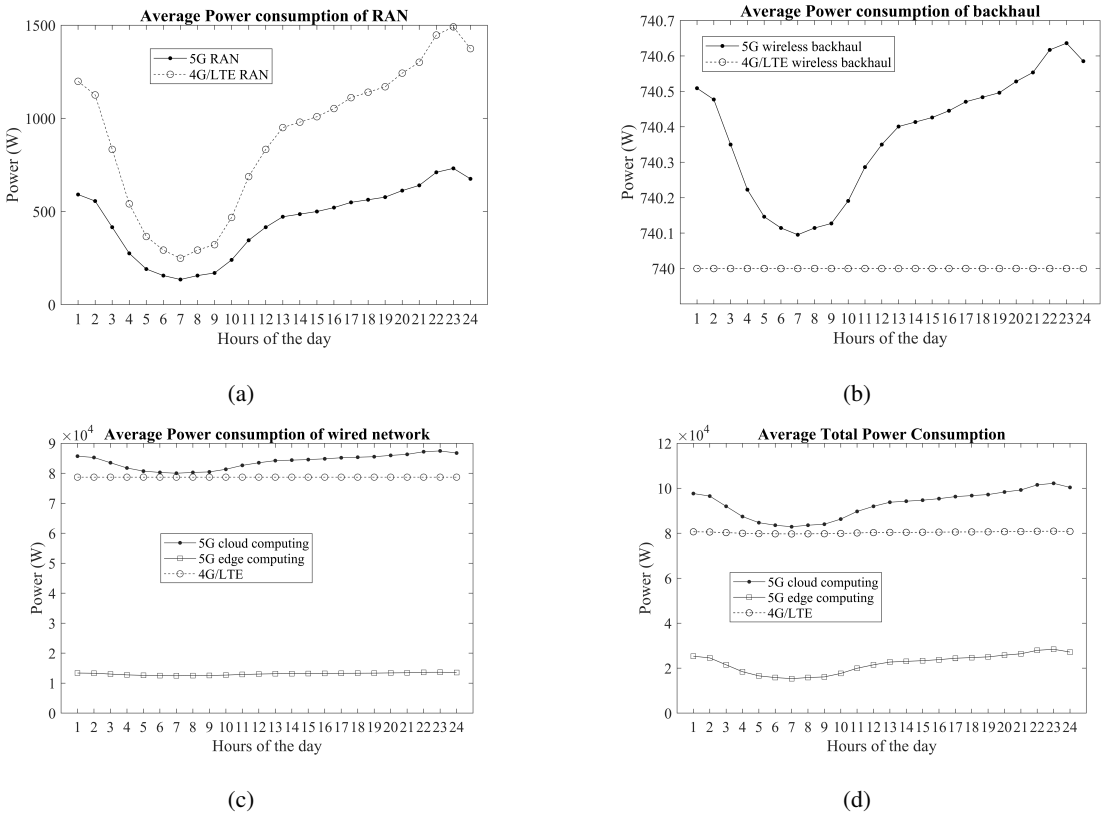

Fig. 8. Comparison of average variation of power consumption between 4G/LTE C-RAN and 5G C-RAN (cloud and edge computing) according to hours of the day (a) Average power consumption at RAN (b) Average power consumption at wireless backhaul (c) Average power consumption of the wired network (d) Total average power consumption considering the entire network.

not transmit RAN traffic: because of that, we assume them in idle mode. Finally, it does not have the power consumption due to data centre. On the other hand, 5G C-RAN with cloud computing achieves the highest power consumption since it uses all parts of the network. Then, the best choice seems to be 5G C-RAN with edge computing, which places data centre in the edge: that allows significant reduction of power consumption (devices in the core network have the highest power consumption when increasing load).

Figure 9 shows the variation of total power consumption of 5G according to the number of PRBs assigned to mobile users. This comparison helps to see that increasing the number of PRBs per user (and so the transmission rate) can significantly affect the C-RAN power consumption. Furthermore, the number of symbols in the modulation scheme has some influence as well.

Next, Figure 10 compares the average total energy efficiency of 4G C-RAN with the one of 5G C-RAN with edge and cloud computing: in particular, it shows results for different bandwidth and

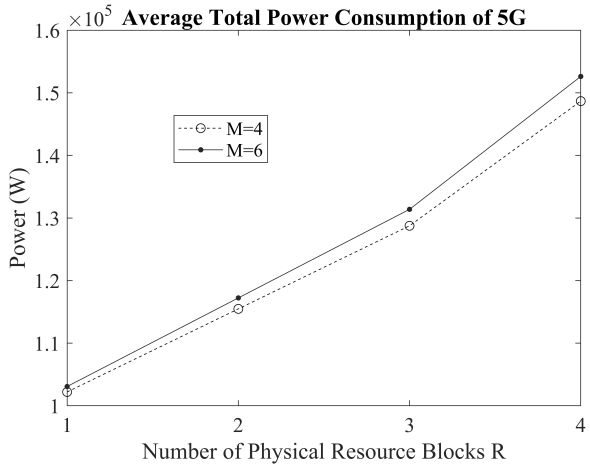

Fig. 9. Average total power consumption of 5G versus the number of PRBs per user (R).

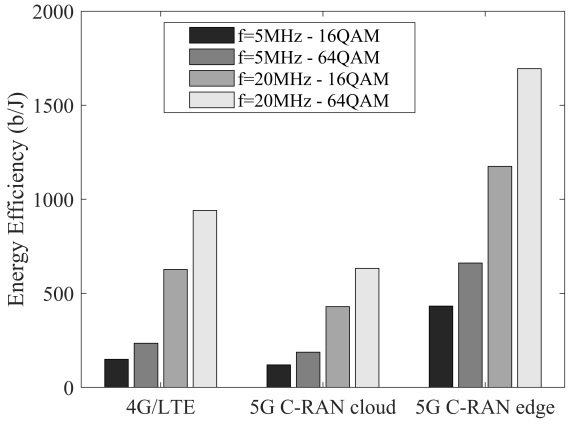

Fig. 10. Average total energy efficiency of 4G C-RAN and 5G C-RAN with edge and cloud computing.

different modulation schemes. The results in terms of energy efficiency confirm the benefits of 5G C-RAN with edge computing because it avoids BBUs at each BS while guaranteeing a more efficient use of operator's network to run v-BBUs. By increasing bandwidth, the gain of energy efficiency moves from $\approx$ 64% to $\approx$ 65% for 16QAM modulation, while it changes between $\approx$ 44% and $\approx$ 46% for 64QAM modulation. Higher frequencies and modulations decrease the energy efficiency gain of 5G C-RAN with edge computing in respect of 4G/LTE. That demonstrates the potential increase in energy efficiency achievable with deployment of edge computing in C-RAN of future 5G networks.

Regarding latency, we can now refer to equation (23). If we split the component of delay due to RAN τ_{RAN} into τ_{RRH} and τ_{BBU}, we can see that τ_{RRH} is similar to 4G/LTE RAN and 5G C-RAN. The

TABLE IV

PARAMETERS FOR COMPARISON OF LATENCY BETWEEN 4G/LTE RAN AND 5G C-RAN [35], [36].

τ_{UDCL}	15.7 μs
τ_{ISCL}	28.34 μs
τ_{DAL}	18.11 μs
τ_{RRH}	1 ms
τ_{bh}	200 μs
τ_{core}	866.6 μs
τ_{edge}	50 μs

value of τ_{BBU} is now to compare to τ_{dc} in order to analyse if there is a latency gain in virtualisation. Moreover, it is important to estimate the impact of propagation time in case of 5G C-RAN edge and cloud computing since baseband processing is moved from the BS to the data centre in the wired network. In this context, the contribution of delay, due to traffic load on the link, becomes negligible in comparison with the magnitude of delay of propagation. Especially, in cloud computing, we consider a large data centre located in London, thus causing a $\tau_{core} \approx 866$ μs, while, in edge computing (small data centre around Manchester), we estimate a delay $\tau_{edge} \approx 50$ μs (propagation delay is calculated using distances obtained from Google Maps and using speed of light).

The time at three-tier data centre can be calculated as

$$\tau_{dc} = \tau_{UDCL} + \tau_{ISCL} + \tau_{DAL} + \tau_{proc} \tag{24}$$

where τ_{UDCL} is the uplink/downlink communication latency in the data centre, τ_{ISCL} is the inter-server communication latency, τ_{DAL} is the delay to access data base in the server and τ_{proc} is the time due to processing (calculations at the server) [35]. Table IV lists the values, which are used for latency evaluation. The values of τ_{tot} for the three technologies are

$$\begin{aligned} \tau_{5G-cloud} &= 1000 + 1129 + \tau_{proc} \\ \tau_{5G-edge} &= 1000 + 312 + \tau_{proc} \\ \tau_{5G-edge} &= 1000 + \tau_{proc} \end{aligned} \tag{25}$$

If we consider 4G/LTE latency as baseline and we neglect the time for processing baseband tasks, we can notice that cloud-based 5G C-RAN adds $\approx 53\%$ higher latency while edge-based C-RAN only $\approx 23\%$. Moreover, if data centres can guarantee higher processing speed (less processing time) than 4G/LTE BBUs, edge computing can perform better than legacy 4G/LTE RAN.

At this point, we can express some final considerations. Cloud RAN paradigm has the potentials to reduce significantly power consumption of current 4G/LTE networks: especially, that would only happen

in the case of edge computing by achieving maximum power gains of $\approx 84\%$. That in line with results previously obtained about possible advantages of edge (fog) computing on cloud computing in terms of energy [34]. Since 5G networks will require simultaneous satisfaction of various performance indicators, with particular attention to latency, we can claim that C-RAN based on edge computing will be the only paradigm to be ahead of legacy 4G/LTE C-RAN. An optimisation of baseband processing at data centres and efficient parallelisation will be a key aspect to permit a significant latency gain. Moreover, it is important to underline that we have assumed the same channel characteristics of both 4G RAN and 5G C-RAN: however, an expected reduction of τ_{RRH} in 5G [36], due to new radio channel structures, will increase the latency gain of C-RAN edge computing and will make comparable the ones of 4G/LTE RAN and 5G C-RAN with cloud computing. Thus, our previous analysis can enforce cloud computing could be reserved (via network slicing techniques) to xMBB and some mMTC users while edge computing to uMTC and some mMTC users (with stringent delay requirements). By considering energy efficiency point of view, future 5G networks expect a "reduction in energy usage by almost 90%" [24]. We have see above that the implementation of efficient 5G C-RAN, based on edge computing, will help to achieve that percentage till $1/3$ of the desired value. Thus, our results highlight the importance of RAN virtualisation.

V. Conclusion

The article has designed a mathematical model based on random multilayer hypergraphs, which takes advantage of results of multilayer graphs, stochastic geometry and tessellation theory to describe characteristics and behaviour of Cloud RAN in future generation networks. Such a general, accurate and flexible model can also be further extended with additional attributes and characteristics of nodes and hyperedges to target more detailed analyses. First, the results has focused on numerical evaluation of virtual resources requirements (in terms of number of servers) and power consumption of data centre. Second, the discussion has analysed the power consumption of each sector of the network and the one of the network as a whole, for 4G/LTE C-RAN, 5G C-RAN with edge and cloud computing. Finally, we estimated the total average energy efficiency and latency of these three network paradigms. Edge computing for 5G C-RAN resulted to be the most efficient way to use network resources for RAN, considering concurrently power consumption/energy efficiency and latency. Finally, we can claim edge computing in C-RAN will be the promising technique to achieve the targeted trade-off between energy efficiency and latency in future 5G networks.

References

[1] B. Haberland, F. Derakhshan, H. Grob-Lipski, R. Klotsche, W. Rehm, P. Schefczik, and M. Soellner, “Radio base stations in the cloud,” *Bell Labs Technical Journal*, vol. 18, no. 1, pp. 129–152, Jun. 2013.

[2] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for mobile networks – a technology overview," *IEEE Communications Surveys Tutorials*, vol. 17, no. 1, pp. 405–426, Firstquarter 2015.

[3] R. Arunadevi and S. Selvakumari, "Mobile communication in 4G technology," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 2, Nov. 2014.

[4] V. Q. Rodriguez and F. Guillemin, "Cloud-RAN modeling based on parallel processing," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 3, pp. 457–468, Mar. 2018.

[5] R. Bassoli, M. D. Renzo, and F. Granelli, "Analytical energy-efficient planning of 5G cloud radio access network," in *2017 IEEE International Conference on Communications (ICC)*, May 2017, pp. 1–4.

[6] F. Granelli, R. Bassoli, and M. D. Renzo, "Energy-efficiency analysis of cloud radio access network in heterogeneous 5G networks," in *European Wireless 2018; 24th European Wireless Conference*, May 2018.

[7] D. Sabella, A. de Domenico, E. Katranaras, M. A. Imran, M. di Girolamo, U. Salim, M. Lalam, K. Samdanis, and A. Maeder, "Energy efficiency benefits of RAN-as-a-service concept for a cloud-based 5G mobile network infrastructure," *IEEE Access*, vol. 2, pp. 1586–1597, 2014.

[8] K. Wang, M. Zhao, and W. Zhou, "Traffic-aware graph-based dynamic frequency reuse for heterogeneous cloud-ran," in *2014 IEEE Global Communications Conference*, Dec. 2014, pp. 2308–2313.

[9] H. Zhang, H. Ji, X. Li, K. Wang, and W. Wang, "Energy efficient resource allocation over cloud-RAN based heterogeneous network," in *2015 IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom)*, Nov. 2015, pp. 483–486.

[10] M. Qian, W. Hardjawana, J. Shi, and B. Vucetic, "Baseband processing units virtualization for cloud radio access networks," *IEEE Wireless Communications Letters*, vol. 4, no. 2, pp. 189–192, Apr. 2015.

[11] R. S. Alhumaima and H. S. Al-Raweshidy, "Evaluating the energy efficiency of software defined-based cloud radio access networks," *IET Communications*, vol. 10, no. 8, pp. 987–994, 2016.

[12] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. A. Imran, D. Sabella, M. J. Gonzalez, O. Blume, and A. Fehske, "How much energy is needed to run a wireless network?" *IEEE Wireless Communications*, vol. 18, no. 5, pp. 40–49, Oct. 2011.

[13] I. Al-Samman, M. Artuso, H. Christiansen, A. Doufexi, and M. Beach, "A framework for resources allocation in virtualised C-RAN," in *2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Sep. 2016, pp. 1–7.

[14] N. Saxena, A. Roy, and H. Kim, "Traffic-aware cloud RAN: A key for green 5G networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, pp. 1010–1021, Apr. 2016.

[15] J. Liu, S. Zhou, J. Gong, Z. Niu, and S. Xu, "Statistical multiplexing gain analysis of heterogeneous virtual base station pools in cloud radio access networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 8, pp. 5681–5694, Aug. 2016.

[16] L. Zhang, A. U. Quddus, E. Katranaras, D. Wübben, Y. Qi, and R. Tafazolli, "Performance analysis and optimal cooperative cluster size for randomly distributed small cells under cloud RAN," *IEEE Access*, vol. 4, pp. 1925–1939, 2016.

[17] M. Haenggi and R. K. Ganti, *Interference in Large Wireless Networks*. Hanover, MA, USA: Now Publishers Inc., Feb. 2009, vol. 3, no. 2.

[18] Y. Cai, F. R. Yu, and S. Bu, "Dynamic operations of cloud radio access networks (C-RAN) for mobile cloud computing systems," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 3, pp. 1536–1548, Mar. 2016.

[19] H. Mei, K. Wang, and K. Yang, "Multi-layer cloud-RAN with cooperative resource allocations for low-latency computing and communication services," *IEEE Access*, vol. 5, pp. 19 023–19 032, 2017.

[20] S. Xu and S. Wang, "Baseband unit pool planning for cloud radio access networks: An approximation algorithm," *IEEE Communications Letters*, vol. 21, no. 2, pp. 358–361, Feb. 2017.

[21] W. Al-Zubaedi and H. S. Al-Raweshidy, "A parameterized and optimized BBU pool virtualization power model for C-RAN architecture," in *IEEE EUROCON 2017 -17th International Conference on Smart Technologies*, Jul. 2017, pp. 38–43.

[22] K. Wang, W. Zhou, and S. Mao, "On joint BBU/RRH resource allocation in heterogeneous cloud-RANs," *IEEE Internet of Things Journal*, vol. 4, no. 3, pp. 749–759, Jun. 2017.

[23] Y. L. Lee, J. Loo, T. C. Chuah, and L. Wang, "Dynamic network slicing for multitenant heterogeneous cloud radio access networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 4, pp. 2146–2161, Apr. 2018.

[24] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 1617–1655, thirdquarter 2016.

[25] M. Haenggi, J. G. Andrews, F. Baccelli, O. Dousse, and M. Franceschetti, "Stochastic geometry and random graphs for the analysis and design of wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 7, pp. 1029–1046, Sep. 2009.

[26] J. G. Andrews, F. Baccelli, and R. K. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Transactions on Communications*, vol. 59, no. 11, pp. 3122–3134, Nov. 2011.

[27] T. Werthmann, H. Grob-Lipski, S. Scholz, and B. Haberland, "Task assignment strategies for pools of baseband computation units in 4G cellular networks," in *2015 IEEE International Conference on Communication Workshop (ICCW)*, Jun. 2015, pp. 2714–2720.

[28] G. Auer, V. Giannini, I. Godor, P. Skillermark, M. Olsson, M. A. Imran, D. Sabella, M. J. Gonzalez, C. Desset, and O. Blume, "Cellular energy efficiency evaluation framework," in *IEEE 73rd Vehicular Technology Conference (VTC Spring)*, May 2011, pp. 1–6.

[29] H. S. Dhillon, R. K. Ganti, F. Baccelli, and J. G. Andrews, "Modeling and analysis of k-tier downlink heterogeneous cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 550–560, Apr. 2012.

[30] M. D. Renzo, W. Lu, and P. Guan, "The intensity matching approach: A tractable stochastic geometry approximation to system-level analysis of cellular networks," vol. abs/1604.02683. [Online]. Available: http://arxiv.org/abs/1604.02683

[31] P. Ruiu, A. Bianco, C. Fiandrino, P. Giaccone, and D. Kliazovich, "Power comparison of cloud data center architectures," in *2016 IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–6.

[32] M. Ismail, W. Zhuang, E. Serpedin, and K. Qaraqe, "A survey on green mobile networking: From the perspectives of network operators and mobile users," *IEEE Communications Surveys Tutorials*, vol. 17, no. 3, pp. 1535–1556, thirdquarter 2015.

[33] W. Lu and M. D. Renzo, "Stochastic geometry modeling of cellular networks: Analysis, simulation and experimental validation," in *ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, Cancun, Mexico, Nov. 2015.

[34] F. Jalali, K. Hinton, R. Ayre, T. Alpcan, and R. S. Tucker, "Fog computing may help to save energy in cloud computing," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1728–1739, May 2016.

[35] C. Fiandrino, D. Kliazovich, P. Bouvry, and A. Y. Zomaya, "Performance and energy efficiency metrics for communication systems of cloud computing data centers," *IEEE Transactions on Cloud Computing*, vol. 5, no. 4, pp. 738–750, Oct. 2017.

[36] S. Nagata, L. H. Wang, and K. Takeda, "Industry perspectives," *IEEE Wireless Communications*, vol. 24, no. 3, pp. 2–4, Jun. 2017.