



# Toward 5G cloud radio access network: An energy and latency perspective

Riccardo Bassoli, Fabrizio Granelli, Sisay Tadesse Arzo, Marco Di Renzo

## ► To cite this version:

Riccardo Bassoli, Fabrizio Granelli, Sisay Tadesse Arzo, Marco Di Renzo. Toward 5G cloud radio access network: An energy and latency perspective. Transactions on emerging telecommunications technologies, 2019, 32, 10.1002/ett.3669 . hal-03377202

**HAL Id: hal-03377202**

**<https://hal.science/hal-03377202>**


Submitted on 14 Oct 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

---

# Toward 5G cloud radio access network: An energy and latency perspective

Riccardo Bassoli<sup>1</sup>  | Fabrizio Granelli<sup>1</sup> | Sisay Tadesse Arzo<sup>1</sup> | Marco Di Renzo<sup>2</sup>

<sup>1</sup>Department of Information Engineering and Computer Science, University of Trento, Trento, Italy

<sup>2</sup>L2S - CNRS, CentraleSupélec, University of Paris-Sud, Paris-Saclay University, Paris, France

## Correspondence

Riccardo Bassoli, Department of Information Engineering and Computer Science, University of Trento, Trento, Italy.  
Email: riccardo.bassoli@unitn.it

## Abstract

Future generation networks will entirely deploy virtualization 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 virtualization 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, which can concurrently satisfy energy efficiency and latency requirements.

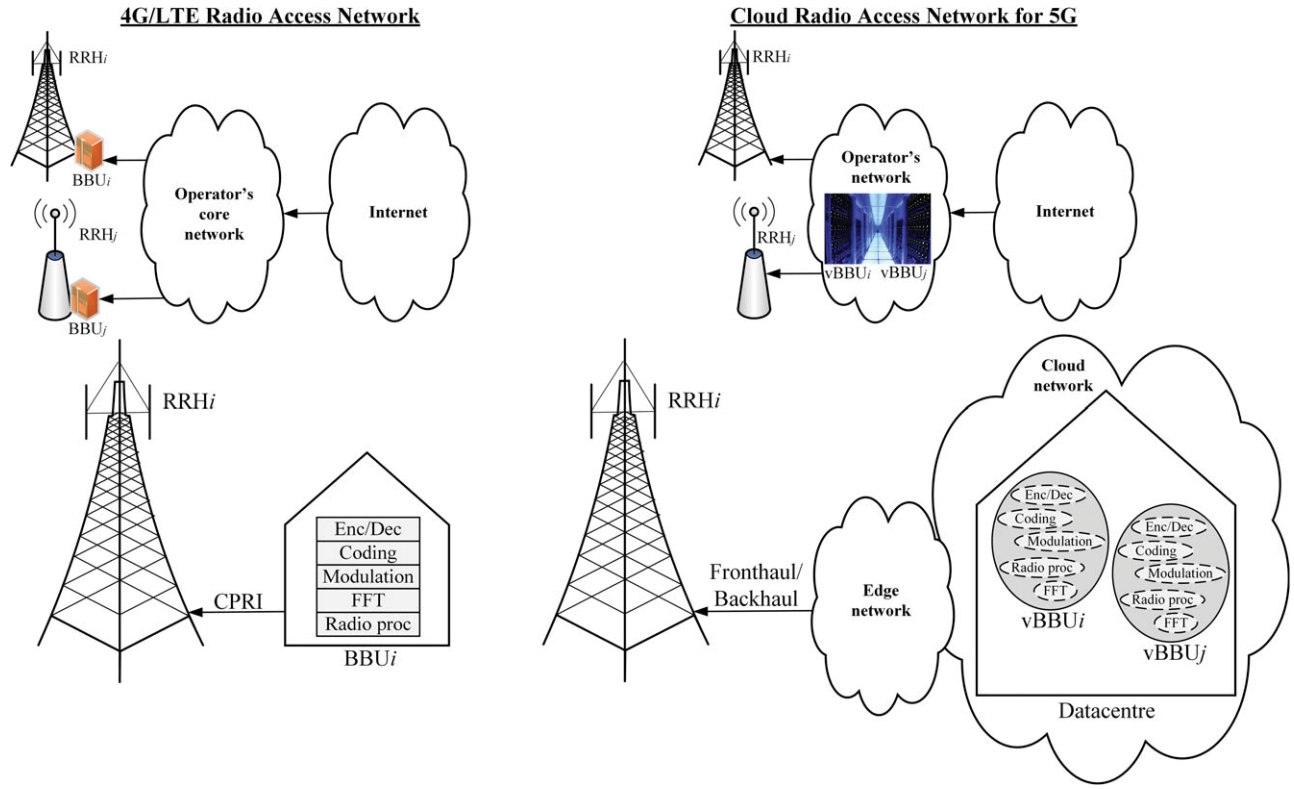
## 1 | INTRODUCTION

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.

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 realize 5G vision while supporting network infrastructure upgrades at an acceptable cost is network virtualization. In particular, network functions virtualization 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)<sup>1,2</sup> is a virtualization paradigm, which aims at moving RAN and baseband functions and procedures to cloud data centers. 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 optimized 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



**FIGURE 1** Downlink communication in heterogeneous 4G/LTE RAN and heterogeneous 5G Cloud RAN. The latter places baseband processing at virtual baseband units (BBUs) in operators' data centers and run them as virtual machines or virtual functions in containers. RRH, radio remote head

BSs; thus, BSs will become pure radio remote heads (RRHs), whereas baseband processing will be moved to dedicated data centers 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<sup>3,4</sup> 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 multiuser detection required to reduce interference, modulation/demodulation, error correction coding (which increases the complexity of the baseband processing at the receiver), radio scheduling, and encryption/decryption of packet data convergence protocol communication (both downlink and uplink). Multicarrier modulation 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. Multiuser detection is used to eliminate the multiple access interference present in CDMA systems.

Based on preliminary results in the work of Bassoli et al<sup>5</sup> and on the initial model published in 2018 at European Wireless conference,<sup>6</sup> 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 analyze 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 generalizes current models for C-RAN in the literature. To the best of the authors' knowledge, such a complete model, based on multilayer random hypergraphs considering power consumption of all areas of the network (included data centers), 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 analyzes the performance of C-RAN referred to downlink communications. The contribution of uplink communications to BBU processing is not considered.

This paper is organized as follows. Section 2 provides a detailed analysis of C-RAN theoretical models toward 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, Section 2.1 highlights the details of each model of C-RAN and its power consumption and, eventually, latency. Then, Section 2.2 focuses on motivation and contributions to justify the need of content presented in the remainder of the paper. Section 3 describes in detail the proposed novel model based on random multilayer hypergraphs. In particular, Section 3.1 structures the model of power consumption and energy efficiency of 5G C-RAN system, whereas Section 3.2 describes the model of latency. Finally, Section 4 discusses results referred to power consumption/energy efficiency and latency of C-RAN in the context of 5G (edge and cloud computing) and compares them with the ones referred to 4G/LTE RAN.

## 2 | RELATED WORKS AND MOTIVATION

This section first presents a selection of works in the literature 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, toward an accurate theoretical description of C-RAN in 5G. While Section 2.1 does not strive to be a survey on C-RAN models, its idea is to clarify why stochastic geometry and, subsequently, a more generalized and comprehensive model is needed in C-RAN theory. The following analysis only takes into account the contribution of the works in terms of (i) modeling C-RAN from cellular network (system) point of view, (ii) modeling data centers and virtualization of BBUs, and (iii) modeling power consumption of C-RAN. Their additional contributions are neglected since they are not in the scope of this work.

### 2.1 | Related works

In 2014, Sabella et al<sup>7</sup> considered a scenario where macro cells are replaced with small cells; moreover, BBU processing is virtualized to cloud data centers. They provided an accurate model for power consumption of cellular networks where the overall power is

$$P_{\text{tot}} = P_{\text{vRAN}} + P_{bh} + P_{\text{RAN}}, \quad (1)$$

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

$$P_{\text{srv}} = P_0^{\text{srv}} + \delta_p^{\text{srv}} P_{\text{max}}^{\text{srv}} x_{\text{srv}}, \quad (2)$$

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

Next, the work of Wang et al<sup>8</sup> modeled 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 center management unit) to the RRHs, which are connected to the pool via Ethernet of 10 Gb/s.

In 2015, the work of Zhang et al<sup>9</sup> modeled C-RAN scenario as a heterogeneous network including cloud data centers and heterogeneous BSs, which serve a vector of mobile users. In the same year, Qian et al<sup>10</sup> 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

$$L_{i,j} = \sum_{n=1}^N \beta_{i,j,n} L_{i,n}^{\text{req}}, \quad (3)$$

where  $L_{i,n}^{\text{req}}$  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

$$C_{i,j} = \begin{cases} 0, & \sum_{n=1}^N \beta_{i,j,n} = 0, 1 \\ \delta_{\text{cost}}, & \text{otherwise,} \end{cases} \quad (4)$$

where  $\delta_{\text{cost}}$  is a constant for the communications between BBUs (measured in MOPTS).

In 2016, the work of Zhang et al<sup>11</sup> 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, AC-DC and DC-DC voltage conversion, optical transceivers, and cooling.<sup>12</sup> On the other hand, the model of power consumption of CN 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

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

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 modeled as

$$P_{SDN} = P_{switch} + P_{SDNctl} \quad (6)$$

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

Next, the system model, proposed by Al-Samman et al,<sup>13</sup> modeled the network via a single tier cellular network, with macro-hexagonal regular cells, composed by RRHs containing nine omnidirectional antennas. Next, the BBUs are virtualized and co-located in a single pool. Later, the work of Saxena et al<sup>14</sup> modeled 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

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

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.

Liu et al<sup>15</sup> analyzed C-RAN by modeling session-level dynamics of virtual BSs via Markov model. In particular, heterogeneous virtual BSs are consolidated in a data center and share a number of units, providing computational resources. Next, in the same year, the work of Zhang et al<sup>16</sup> 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  $\lambda_U$  and  $\lambda_R$ , 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 the work of Haenggi and Ganti.<sup>17</sup> Moreover, the proposed latency model for C-RAN is defined by

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

where  $\Delta t_{ce}$  is the channel estimation delay,  $\Delta t_{fb}$  is the average per-channel coefficient feedback delay,  $\Delta t_{pt}$  is the propagation delay,  $\Delta t_{pc}$  is the cloud processing delay,  $\Delta t_{pRRH}$  is the RRH processing delay,  $\Delta t_{BH}$  is the backhaul delay per hop,  $\omega_1$  is the number of channel coefficients to be estimated for a mobile user,  $\omega_2$  is the total number of times channel state information (CSI) is to be fed back for the whole network, and  $\omega_3$  number of backhaul hops. Afterwards, the work of Cai et al<sup>18</sup> 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, whereas C-RAN is composed by a number of RRHs (single tier) with a unique antenna.

In 2017, Mei et al<sup>19</sup> 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

$$U = (F, D), \quad (9)$$



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

$$T = \frac{F}{f} + \frac{D}{r}, \quad (10)$$

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

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

where  $P$  is the transmission power of an RRH,  $\varphi$  is the effective switched capacitance,  $\vartheta$  is a positive constant, and  $\eta$  is a weighted trade-off between energy consumption of a mobile cloud and C-RAN. Finally, the authors provided expression of the signal-to-interference-plus-noise ratio (SINR) to estimate the data rate of a mobile user connected to its serving RRH. In the same year, the work of Xu and Wang<sup>20</sup> 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, Al-Zubaedi and Al-Raweshidy<sup>21</sup> studied the 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<sup>12</sup> of BSs.

Afterwards, the work of Wang et al<sup>22</sup> analyzed C-RAN consisting of heterogeneous RRHs (eg, 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, whereas the UE has one receiving antenna. Information comes from the backbone network toward the mobile user (downlink). The RRH are connected to the BBU pool via switch using Ethernet of 10 Gb/s.

Finally, the work of Lee et al<sup>23</sup> modeled C-RAN as a system with a unique macro cell, containing various small cells. Each RRH is connected to the BBU pool in the CN 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.

## 2.2 | Motivation and contribution

Section 2.1 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 analyze 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<sup>24</sup> 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 subnetworks (data center 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 Section 2.1.

- *5G C-RAN is not modeled as heterogeneous system with spatial information.* The study of C-RAN, in the context of 5G, cannot neglect the characterization of SINR, which requires knowledge of network geometry.<sup>25,26</sup> In order to circumvent the difficulty to characterize SINR, stochastic geometry and random graphs were proposed. Regular models of radio coverage (eg, 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 section has showed that the main methods used in research to model C-RAN were based on regular mathematical structures, thus resulting inaccurate.
- *Virtualization of RAN is not contextualized in a framework, which models the actual architecture of a data center.* To the best of authors' knowledge, no existing work that deals with 5G C-RAN has flexibly analyzed how data center's architecture, interacting with the 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 center is frequently neglected.
- *In 5G C-RAN investigation, it is not analyzed the trade-off between power consumption/energy efficiency and latency.* The works about C-RAN listed in the previous section analyze either power or latency in C-RAN while not considering the combination of them toward 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.

With respect to the aforementioned open issues, the contribution of this article includes the following.

- Section 3. 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 (eg, stochastic geometry) to investigate effectively the complexity and heterogeneity of 5G C-RAN as a system.
- Sections 3.1 and 4. 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 centers during downlink communications. Moreover, it includes a detailed characterization of baseband processing requirements because of UEs, provided by Werthmann et al.<sup>27</sup>
- Sections 3.2 and 4. 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 toward a more significant and accurate modeling and characterization of C-RAN properties and behavior in 5G.

### 3 | 5G SYSTEM MODEL

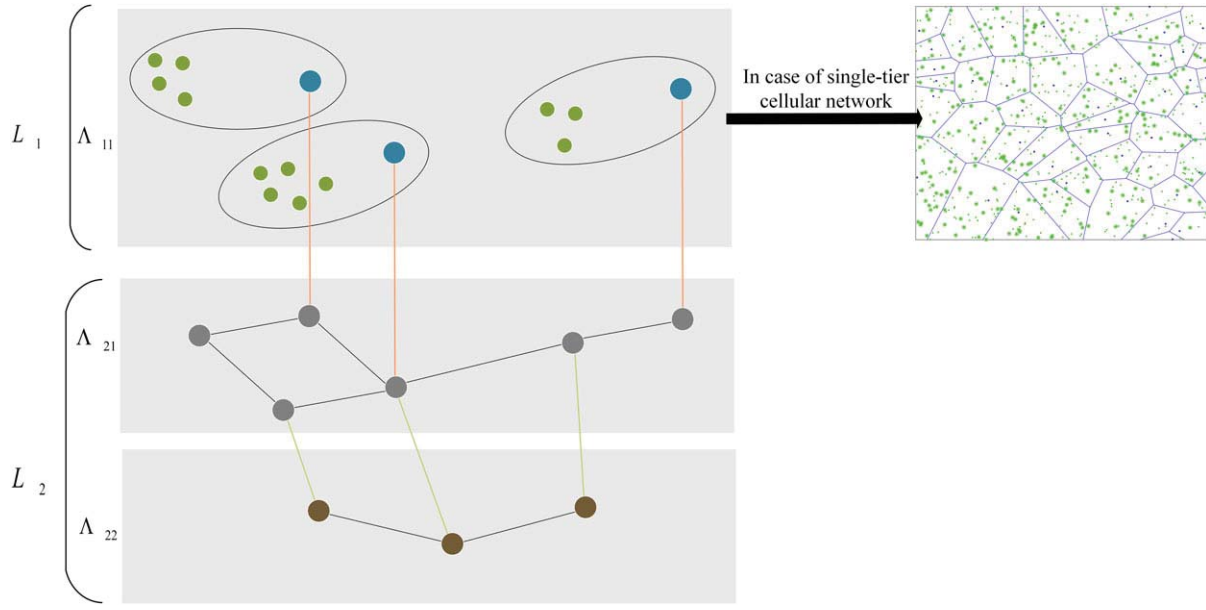
Graph theory is the area of mathematics that has allowed effective modeling of communication networks as a whole. Wired networks have always been modeled as planar graphs, composed by a set of nodes (eg, switches, routers, etc) and a set of edges (ie, wired links). Side by side, a planar hypergraph is a graph's generalization where edges can connect group of nodes to each others (ie, 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 modeling 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 is why this article proposes a new generalized 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 multitier heterogeneous cellular network, which comprises different kinds of BSs. Next, there are data centers, which can be located either in CN (called large data center, cloud computing) or in edge network (small data center, edge/fog computing), hosting v-BBUs, which can run as VNFs in virtual machines or containers. According to preliminary research in the works of Bassoli et al<sup>5</sup> and Granelli et al<sup>6</sup> and to previous discussion in Section 2.2, we propose to model virtualized RAN via a random multilayer hypergraph, a mathematical object that can flexibly describe the properties and the characteristics of 5G C-RAN.

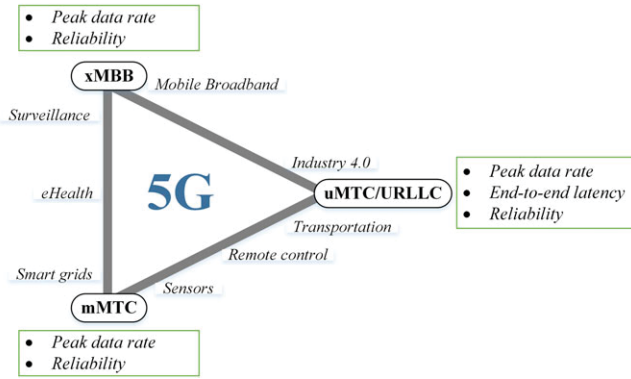
Let  $H = (X, E)$  be a planar hypergraph representing the physical network, where  $X$  is the set of nodes and  $E$  is the set of nonempty subsets of  $X$ , called hyperedges. Next, set  $X$  can be partitioned into subsets  $X = \{X_1, X_2, \dots\}$  respectively referred to mobile end users, BSs,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

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

- $X$  is the set of random nodes, which can be distributed according to either random point processes (eg, BSs) or deterministic spatial distributions (eg, wired operator's network);
- $E$  is the set of random hyperedges, whose cardinality  $|E_i|$  can be defined by either Voronoi tessellation in  $\mathbb{R}^2$  (eg, wireless cellular networks) or deterministic values (eg, links in wired networks);



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



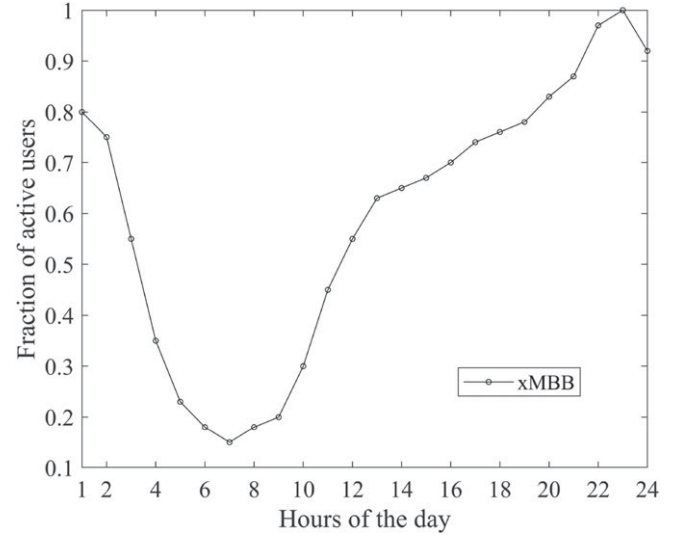
**FIGURE 3** 5G legacy classification of end-users according to service requirements. mMTC, massive Machine-Type Communication; uMTC, ultra-reliable Machine-Type Communication; URLLC, ultra-reliable low-latency communication; xMBB, Extreme Mobile Broadband

- $L = \{L_1, \dots, L_a\}$  is the set of layers, where  $a$  is the number of aspects; each layer can be a set of sublayers  $\Lambda_{ij}$ , where  $i$  is the number of layer it belongs to and  $j$  is the number of sublayer ( $j = 1, \dots, |L_i|$ );
- $X_H$  is the set of node-layer elements;
- $E_H$  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 sublayer  $\Lambda_{11}$  and  $L_2$  is composed by two sublayers  $\Lambda_{21}$  and  $\Lambda_{22}$ . If  $L_1$  represents a single-tier cellular network, its hyperedges can be identified via Voronoi tessellation. In the rest of this paper, since we will work on multitier networks, the specific tessellation will be a multiplicatively-weighted (MW) Voronoi tessellation. Side by side,  $L_2$  sublayers are planar graphs, modeling 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 later).

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.





**FIGURE 4** Average variation of fraction of active end users (xMBB) according to the hours of the day. xMBB, Extreme Mobile Broadband

Given these premises, each node representing a mobile end user is identified by the  $i$ th commodity flow, thus a quadruple  $(s_i, \sigma_i, D_i)$ , where  $s_i \in S$  is the source ( $S$  is the set of sources) and  $\sigma_i \in \Sigma$  is the sink ( $\Sigma$  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 the 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 the work of Auer et al.<sup>28</sup> Figure 4 shows the fraction of xMBB end users, which are active in average during the hours of the day in Europe.

### 3.1 | Model of power consumption

The general system model of 5G C-RAN described earlier is now specified to estimate the power consumption. Let us consider a three-tier heterogeneous cellular network, with nodes belonging to  $X_1$ ,  $X_2$ , and  $X_3$  (subsets of  $X$ ) following three homogeneous Poisson point process (PPP)  $\Phi_{BS_m}$  (micro BSs),  $\Phi_{BS_p}$  (pico BSs), and  $\Phi_{BS_f}$  (femto BSs) of intensity  $\lambda_{BS_m}$ ,  $\lambda_{BS_p}$ , and  $\lambda_{BS_f}$  respectively. Next, let  $\Phi_{mw-sw}$  be the homogeneous PPP, with intensity  $\lambda_{mw-sw}$ , representing the spatial distribution of microwave aggregate switches for wireless backhaul; in particular, BSs connects to the nearest aggregate switch. Finally, let  $\Phi_{xMBB}$  be the homogeneous PPP describing the distribution of mobile broadband users, of intensity  $\lambda_{xMBB}$ . All the PPPs  $\Phi_{BS_m}$ ,  $\Phi_{BS_p}$ ,  $\Phi_{BS_f}$ ,  $\Phi_{mw-sw}$ , and  $\Phi_{xMBB}$  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 modeled using a multiplicatively-weighted Voronoi tessellation, since the points of  $\Phi_{BS_m}$ ,  $\Phi_{BS_p}$ , and  $\Phi_{BS_f}$  have different weights.<sup>29</sup> 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*<sup>29</sup> can be expressed as

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

where  $P_{trx,i}$  is the transmission power of the  $i$ th tier,  $\theta_i$  is the SINR threshold, and  $\alpha$  is the path loss exponent; these are attributes referred to nodes, which identify BSs. As a consequence, each BS of the  $j$ th tier has an average load of  $N_j / \lambda_j$ . In order to minimize the propagation delay, we assume that 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 (ie, connected via random hyperedge). The probability mass function (pmf) of the number of nodes (BSs) that are connected to an aggregate switch<sup>30</sup> is  $N_{BS}$ , expressed as

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

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

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

$$P_{\text{tot5G}} = P_{\text{RAN}} + P_{bh} + P_{\text{net}} + P_{dc}, \quad (14)$$

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

An accurate and detailed model to study power consumption of legacy multitier 4G cellular networks is published in the works of Auer et al.<sup>12,28</sup> In particular, the linear approximation of the *power consumption of a BS*<sup>28</sup> (this is an attribute of BSs nodes) can be expressed as

$$P_{BS} = N_{\text{trx}} \left( (1 - \rho) P_{\text{BSidle}} + \rho \Delta_p P_{\text{BSmax}} \right), \quad (15)$$

where  $N_{\text{trx}}$  is the number of transmission chains (ie, ratio between transmit and receive antennas per site),  $P_{\text{BSidle}}$  is the power consumption calculated at the minimum possible power,  $\Delta_p$  is the slope of load dependent power consumption,  $P_{\text{BSmax}}$  is the maximum RF output power at maximum load, and  $\rho$  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 virtualized, whereas 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 virtualized network, the *power consumption of backhaul/fronthaul* will be mainly affected by the number of switches aggregating traffic and connecting the BS with data center. 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)<sup>7</sup> can be estimated as

$$P_{bh} = \sum_{n=1}^{N_{\text{cell}}} P_{sw}^n + N_{mw}^n P_{\text{link}}^n, \quad (16)$$

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

$$P_{\text{net}} = (1 - \rho) (h_e P_{e-\text{idle}} + h_c P_{c-\text{idle}}) + \rho (h_e P_{e-\text{max}} + h_c P_{c-\text{max}}), \quad (17)$$

where  $h_e$  is the number of hops in the edge network  $P_{e-\text{idle}}$  and  $P_{e-\text{max}}$  are the power consumptions of an edge router in idle and maximum load status respectively,  $h_c$  is the number of hops in the CN and  $P_{c-\text{idle}}$  and  $P_{c-\text{max}}$  are 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 center; 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<sup>27</sup>

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

where  $A$  is the number of antennas,  $M$  is the modulation bits,  $C$  is the code rate,  $L$  is the number of spatial MIMO-layers, and  $R$  is 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 center with a three-tier structure and the linear approximation in the work of Rui et al.,<sup>31</sup> the *power consumption of a data center*  $P_{dc}$  can be evaluated as

$$P_{dc} = P_{dc-sw} + P_{dc-s}, \quad (19)$$

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

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

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

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

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 center's network respectively. Switches and servers in data center's network belong 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*<sup>32</sup> of the 5G network is calculated as

$$EE = \frac{C}{P_{tot5G}}, \quad (22)$$

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

### 3.2 | 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 centers either in the edge or in the cloud. That implies additional delays for transmission via the edge and/or CN toward the data center. Thus, the total latency of future 5G networks can be expressed as

$$\tau_{5G} = \tau_{\text{RAN}} + \tau_{bh} + \tau_{\text{edge}} + \tau_{\text{core}} + \tau_{dc} \quad (23)$$

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

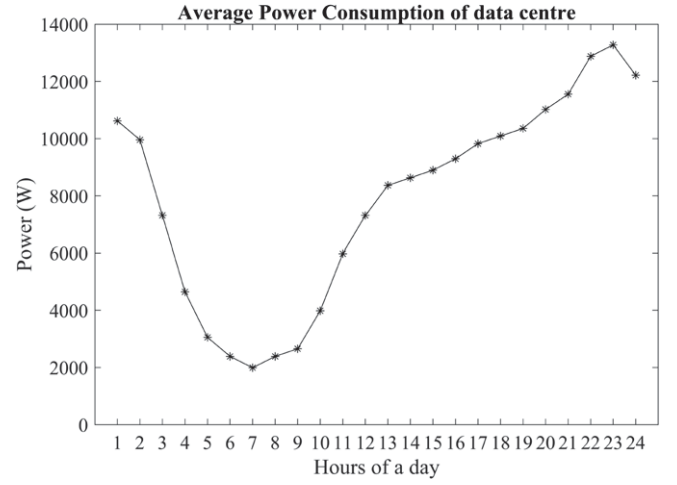
## 4 | RESULTS AND DISCUSSIONS

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 Lu and Di Renzo,<sup>33</sup> we consider a density of 37 BS/ $\mathcal{A}$ , where  $\mathcal{A} = 1.8 \text{ km}^2$ . If we consider the city center as a square of side 15 km (see Figure 5), we have about 125 areas  $\mathcal{A}$  in the city center, containing 4625 BSs in total. Let us 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  $\lambda_{BS_m} = 2 \text{ BS}/\mathcal{A}$ ,  $\lambda_{BS_p} = 7 \text{ BS}/\mathcal{A}$ , and  $\lambda_{BS_f} = 27 \text{ BS}/\mathcal{A}$ . It is important to notice that we do not consider the presence of mmWave BSs in this article. Moreover, the results presented in the work of Lu and Di Renzo<sup>33</sup> allow to model correctly the distribution of BSs as independent two-dimensional homogeneous PPPs on a Euclidean plane  $\mathbb{R}^2$ , called  $\Phi_{BS_m}$ ,  $\Phi_{BS_p}$ , and  $\Phi_{BS_f}$ , where  $\lambda_{BS_m}$ ,  $\lambda_{BS_p}$ , and  $\lambda_{BS_f}$  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 (ie, the hourly variation of  $\lambda_{\text{xMBB}}$ ). Regarding the density of end users,  $\lambda_{\text{xMBB}}$  can be assumed to be 10 times the number of BSs.<sup>33</sup> 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 2 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 users are assumed transmitting at same rate





**FIGURE 7** Variation of average power consumption of three-tier data center according to hours of the day

**TABLE 1** Parameters for evaluation depending on the tier<sup>28</sup>

	Micro	Pico	Femto
$N_{trx}$	2	2	2
$\Delta_p$	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
$\alpha$	3	3	3
$\theta$	4	2	1

**TABLE 2** Parameters for evaluation baseband processing load per UE<sup>27</sup>

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

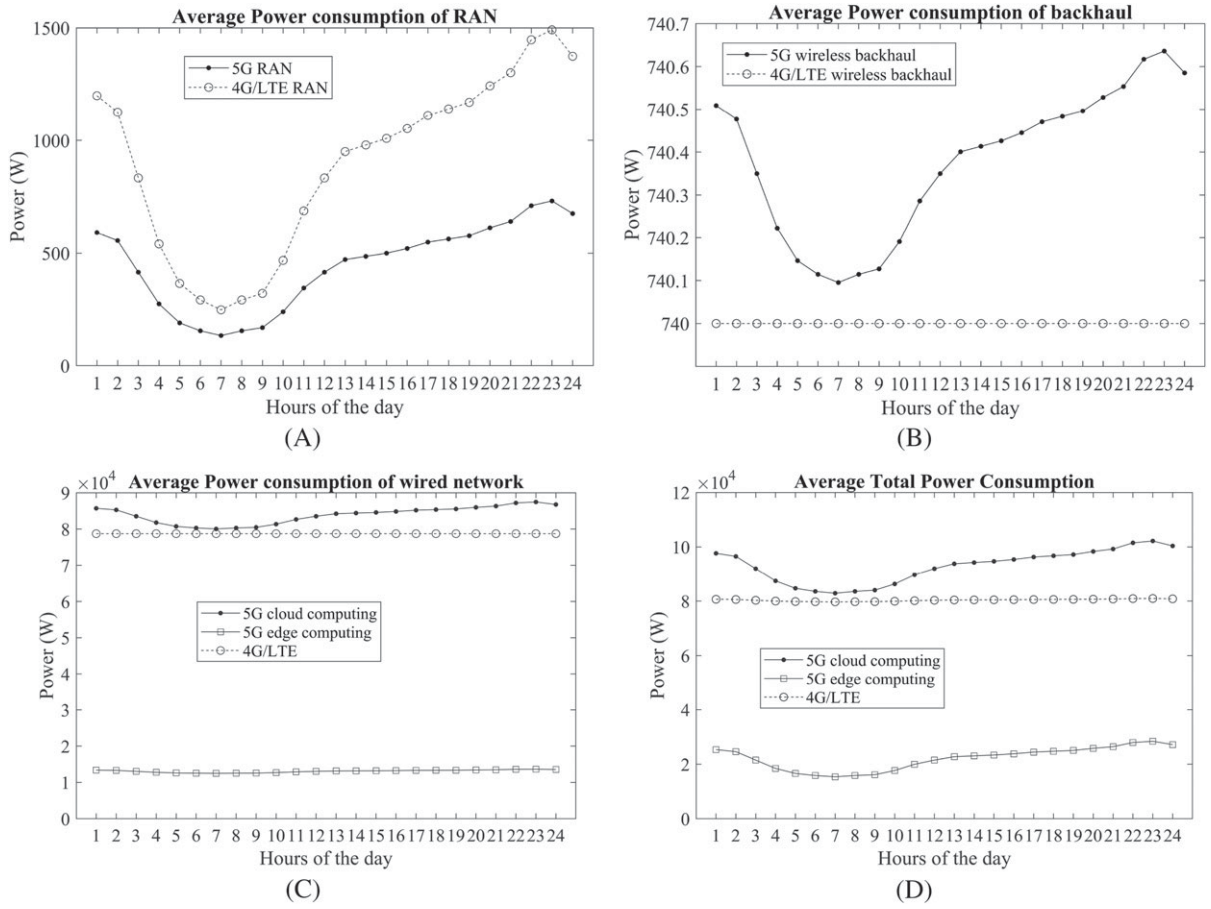
Abbreviations: PRB, physical resource block.

**TABLE 3** Parameters for numerical evaluation of backhaul, edge, core and data center's power consumption

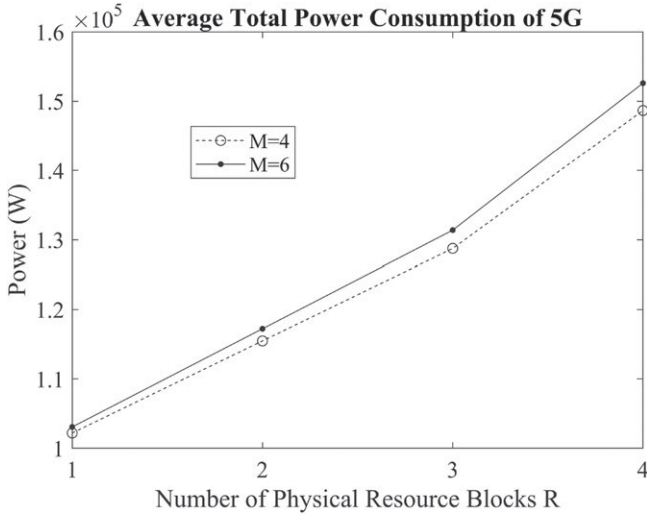
$P_s^7$	53 W
$f_{cell-bh}^7$	128%
$Y_{max}^7$	84.4 Mb/s
$C_{sw}^7$	36 Gb/s
$P_{link}^n$	22.2 W (idle) 37 W (low traffic) 92.5 W (high traffic)
$N_{mw}^7$	2
Edge router <sup>34</sup>	$P_{e-idle} = 4095$ W $P_{e-max} = 4550$ W $h_e = 3$
Core router <sup>34</sup>	$P_{c-idle} = 11070$ W $P_{c-max} = 12300$ 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

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 center in the edge; that allows significant reduction of power consumption (devices in the CN have the highest power consumption when increasing load).



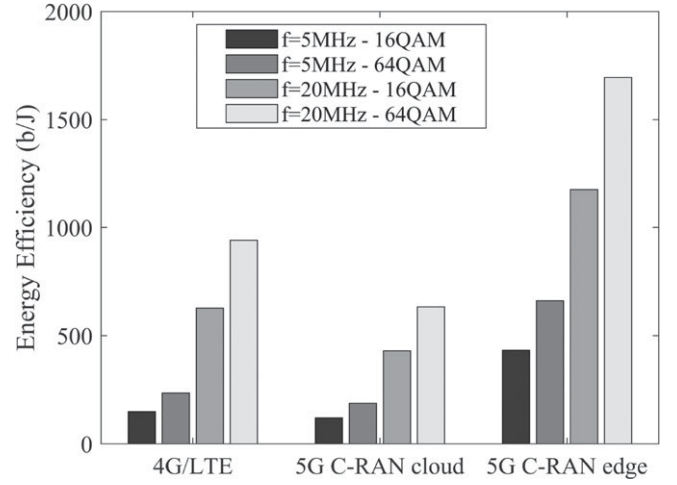


**FIGURE 8** Comparison of average variation of power consumption between 4G/LTE C-RAN and 5G C-RAN (cloud and edge computing) according to hours of the day. A, Average power consumption at RA; B, Average power consumption at wireless backhaul; C, Average power consumption of the wired network; D, Total average power consumption considering the entire network



**FIGURE 9** Average total power consumption of 5G versus the number of physical resource blocks per user ( $R$ )

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.



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

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 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, whereas 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  $\tau_{\text{RAN}}$  into  $\tau_{\text{RRH}}$  and  $\tau_{\text{BBU}}$ , we can see that  $\tau_{\text{RRH}}$  is similar to 4G/LTE RAN and 5G C-RAN. The value of  $\tau_{\text{BBU}}$  is now to compare to  $\tau_{\text{dc}}$  to analyze if there is a latency gain in virtualization. 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 center 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 center located in London, thus causing a  $\tau_{\text{core}} \approx 866 \mu\text{s}$ , whereas, in edge computing (small data center around Manchester), we estimate a delay  $\tau_{\text{edge}} \approx 50 \mu\text{s}$  (propagation delay is calculated using distances obtained from Google Maps and using speed of light).

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

$$\tau_{\text{dc}} = \tau_{\text{UDCL}} + \tau_{\text{ISCL}} + \tau_{\text{DAL}} + \tau_{\text{proc}}, \quad (24)$$

where  $\tau_{\text{UDCL}}$  is the uplink/downlink communication latency in the data center,  $\tau_{\text{ISCL}}$  is the inter-server communication latency,  $\tau_{\text{DAL}}$  is the delay to access data base in the server, and  $\tau_{\text{proc}}$  is the time due to processing (calculations at the server).<sup>35</sup> Table 4 lists the values, which are used for latency evaluation. The values of  $\tau_{\text{tot}}$  for the three technologies are

$$\begin{aligned} \tau_{5G\text{-cloud}} &= 1000 + 1129 + \tau_{\text{proc}} \\ \tau_{5G\text{-edge}} &= 1000 + 312 + \tau_{\text{proc}} \\ \tau_{5G\text{-edge}} &= 1000 + \tau_{\text{proc}}. \end{aligned} \quad (25)$$

**TABLE 4** Parameters for comparison of latency between 4G/LTE RAN and 5G C-RAN<sup>35,36</sup>

$\tau_{\text{UDCL}}$	15.7 $\mu\text{s}$
$\tau_{\text{ISCL}}$	28.34 $\mu\text{s}$
$\tau_{\text{DAL}}$	18.11 $\mu\text{s}$
$\tau_{\text{RRH}}$	1 ms
$\tau_{\text{bh}}$	200 $\mu\text{s}$
$\tau_{\text{core}}$	866.6 $\mu\text{s}$
$\tau_{\text{edge}}$	50 $\mu\text{s}$

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, whereas edge-based C-RAN only  $\approx 23\%$ . Moreover, if data centers 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 is in line with the results previously obtained about possible advantages of edge (fog) computing on cloud computing in terms of energy.<sup>34</sup> 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 optimization of baseband processing at data centers and efficient parallelization 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  $\tau_{\text{RRH}}$  in 5G,<sup>36</sup> 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 that cloud computing could be reserved (via network slicing techniques) to xMBB and some mMTC users, whereas 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%”.<sup>24</sup> We have seen earlier 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 virtualization.

## 5 | 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 behavior 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 center. Second, the discussion has analyzed 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.

## ORCID

Riccardo Bassoli  <https://orcid.org/0000-0002-6132-7985>

## REFERENCES

1. Haberland B, Derakhshan F, Grob-Lipski H, et al. Radio base stations in the cloud. *Bell Labs Tech J*. 2013;18(1):129-152.
2. Checko A, Christiansen HL, Yan Y, et al. Cloud RAN for mobile networks — a technology overview. *IEEE Commun Surv Tutor*. 2015;17(1):405-426.
3. Arunadevi R, Selvakumari S. Mobile communication in 4G technology. *Int J Innov Res Comput Commun Eng*. 2014;2(11).
4. Rodriguez VQ, Guillemin F. Cloud-RAN modeling based on parallel processing. *IEEE J Sel Areas Commun*. 2018;36(3):457-468.
5. Bassoli R, Di Renzo M, Granelli F. Analytical energy-efficient planning of 5G cloud radio access network. Paper presented at: 2017 IEEE International Conference on Communications (ICC); 2017; Paris, France.
6. Granelli F, Bassoli R, Di Renzo M. Energy-efficiency analysis of cloud radio access network in heterogeneous 5G networks. Paper presented at: European Wireless 2018; 24th European Wireless Conference; 2018; Catania, Italy.
7. Sabella D, de Domenico A, Katranaras E, et al. Energy efficiency benefits of RAN-as-a-service concept for a cloud-based 5G mobile network infrastructure. *IEEE Access*. 2014;2:1586-1597.
8. Wang K, Zhao M, Zhou W. Traffic-aware graph-based dynamic frequency reuse for heterogeneous cloud-ran. Paper presented at: 2014 IEEE Global Communications Conference; 2014; Austin, TX.
9. Zhang H, Ji H, Li X, Wang K, Wang W. Energy efficient resource allocation over cloud-RAN based heterogeneous network. Paper presented at: 2015 IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom); 2015; Vancouver, Canada.

10. Qian M, Hardjawana W, Shi J, Vucetic B. Baseband processing units virtualization for cloud radio access networks. *IEEE Wirel Commun Lett.* 2015;4(2):189-192.
11. Alhumaima RS, Al-Raweshidy HS. Evaluating the energy efficiency of software defined-based cloud radio access networks. *IET Communications.* 2016;10(8):987-994.
12. Auer G, Giannini V, Dessel C, et al. How much energy is needed to run a wireless network? *IEEE Wirel Commun.* 2011;5(18):40-49.
13. Al-Samman I, Artuso M, Christiansen H, Doufexi A, Beach M. A framework for resources allocation in virtualised C-RAN. Paper presented at: 2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC); 2016; Valencia, Spain.
14. Saxena N, Roy A, Kim H. Traffic-aware cloud RAN: a key for green 5G networks. *IEEE J Sel Areas Commun.* 2016;34(4):1010-1021.
15. Liu J, Zhou S, Gong J, Niu Z, Xu S. Statistical multiplexing gain analysis of heterogeneous virtual base station pools in cloud radio access networks. *IEEE Trans Wirel Commun.* 2016;15(8):5681-5694.
16. Zhang L, Quddus AU, Katranaras E, Wübben D, Qi Y, Tafazolli R. Performance analysis and optimal cooperative cluster size for randomly distributed small cells under cloud RAN. *IEEE Access.* 2016;4:1925-1939.
17. Haenggi M, Ganti RK. *Interference in Large Wireless Networks*. Vol 3. Hanover, MA: Now Publishers Inc; 2009.
18. Cai Y, Yu FR, Bu S. Dynamic operations of cloud radio access networks (C-RAN) for mobile cloud computing systems. *IEEE Trans Veh Technol.* 2016;65(3):1536-1548.
19. Mei H, Wang K, Yang K. Multi-layer cloud-RAN with cooperative resource allocations for low-latency computing and communication services. *IEEE Access.* 2017;5:19023-19032.
20. Xu S, Wang S. Baseband unit pool planning for cloud radio access networks: an approximation algorithm. *IEEE Commun Lett.* 2017;21(2):358-361.
21. Al-Zubaedi W, Al-Raweshidy HS. A parameterized and optimized BBU pool virtualization power model for C-RAN architecture. Paper presented at: IEEE EUROCON 2017-17th International Conference on Smart Technologies; 2017; Ohrid, Macedonia.
22. Wang K, Zhou W, Mao S. On joint BBU/RRH resource allocation in heterogeneous cloud-RANs. *IEEE Internet Things J.* 2017;4(3):749-759.
23. Lee YL, Loo J, Chuah TC, Wang L. Dynamic network slicing for multitenant heterogeneous cloud radio access networks. *IEEE Trans Wirel Commun.* 2018;17(4):2146-2161.
24. Agiwal M, Roy A, Saxena N. Next generation 5G wireless networks: a comprehensive survey. *IEEE Commun Surv Tutor.* 2016;18(3):1617-1655.
25. Haenggi M, Andrews JG, Baccelli F, Dousse O, Franceschetti M. Stochastic geometry and random graphs for the analysis and design of wireless networks. *IEEE J Sel Areas Commun.* 2009;27(7):1029-1046.
26. Andrews JG, Baccelli F, Ganti RK. A tractable approach to coverage and rate in cellular networks. *IEEE Trans Commun.* 2011;59(11):3122-3134.
27. Werthmann T, Grob-Lipski H, Scholz S, Haberland B. Task assignment strategies for pools of baseband computation units in 4G cellular networks. Paper presented at: 2015 IEEE International Conference on Communication Workshop (ICCW); 2015; London, UK.
28. Auer G, Giannini V, Gódor I, et al. Cellular energy efficiency evaluation framework. In: 2011 IEEE 73rd Vehicular Technology Conference (VTC Spring); 2011; Yokohama, Japan.
29. Dhillon HS, Ganti RK, Baccelli F, Andrews JG. Modeling and analysis of K-tier downlink heterogeneous cellular networks. *IEEE J Sel Areas Commun.* 2012;30(3):550-560.
30. Di Renzo MD, Lu W, Guan P. The intensity matching approach: a tractable stochastic geometry approximation to system-level analysis of cellular networks. 2016. abs/1604.02683. <http://arxiv.org/abs/1604.02683>
31. Rui P, Bianco A, Fiandrino C, Giaccone P, Kliazovich D. Power comparison of cloud data center architectures. Paper presented at: 2016 IEEE International Conference on Communications (ICC); 2016; Kuala Lumpur, Malaysia.
32. Ismail M, Zhuang W, Serpedin E, Qaraqe K. A survey on green mobile networking: from the perspectives of network operators and mobile users. *IEEE Commun Surv Tutor.* 2015;17(3):1535-1556.
33. Lu W, Di Renzo MD. Stochastic geometry modeling of cellular networks: Analysis, simulation and experimental validation. In: Proceedings of the 18th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems; 2015; Cancun, Mexico.
34. Jalali F, Hinton K, Ayre R, Alpcan T, Tucker RS. Fog computing may help to save energy in cloud computing. *IEEE J Sel Areas Commun.* 2016;34(5):1728-1739.
35. Fiandrino C, Kliazovich D, Bouvry P, Zomaya AY. Performance and energy efficiency metrics for communication systems of cloud computing data centers. *IEEE Trans Cloud Comput.* 2017;5(4):738-750.
36. Nagata S, Wang LH, Takeda K. Industry perspectives. *IEEE Wirel Commun.* 2017;24(3):2-4.