Energy Efficiency on Fully Cloudified Mobile Networks: Survey, Challenges, and Open Issues

Ali Alnoman, Glaucio H. S. Carvalho, Alagan Anpalagan, *Senior Member, IEEE*, and Isaac Woungang

***Abstract*—Fully cloudified mobile network infrastructure, which is featured by the joint deployment of heterogeneous cloud radio access networks and edge computing, will successfully cope with the data deluge by densely deploying virtualized wireless base stations and servers while providing the system design with high flexibility, reliability, availability, and scalability. On the other hand, the massive replication of the wireless and computing infrastructure will significantly increase the energy footprint to prohibitive levels. In order to gain actionable insights on energy-efficiency for a fully cloudified mobile network infrastructure, this paper first presents a comprehensive survey of the recent research breakthroughs on each building block of the system, namely: remote radio heads, baseband unit pool, fronthaul, backhaul, HetNet, and edge and cloud computing. Next, we consolidate the discussion with the challenges and open issues of a joint operation.**

***Index Terms*—H-CRANs, HetNets, energy efficiency, edge computing, cloudified mobile network.**

I. Introduction

WIRELESS industry is undergoing a thorough cloudification process of its infrastructure that will deeply impact the way that the networks are designed, deployed, operated, managed, and optimized. Ultimately, this digital transformation, which will serve as a catalyst for industry innovations, will unleash the potential for new services and products, resulting in an increase of competitiveness in the fierce information and communication technology (ICT) landscape with the delivery of unprecedented enterprise services. Giants like Huawei and Ericsson are instances of companies that are promoting this cloudification process.

From a technical perspective, the cloudification process translates into the adoption of heterogeneous cloud radio access networks (H-CRAN) and edge computing. H-CRAN, which is a merge between heterogeneous networks (HetNets) and cloud radio access networks (C-RANs), has recently gained considerable attention in the literature due to its economical (low capital expenditure (CAPEX) and operational expenditure (OPEX)) and technical (high data rates over a large coverage area) benefits. H-CRAN is featured by an evolved base station architecture that performs all the signal processing and resource control virtually and centrally while allowing for an ultra dense deployment of heterogeneous relay stations to cope with signals to and from users equipment. In addition to overhauling the wireless infrastructure, the cloudification process will bring computing closer to the end user. Furthermore, it will support network function virtualization (NFV) and software defined networking (SDN), which ease the dynamic adjustment of network parameters and enable global network control towards maximizing the energy efficiency in base stations, backhaul, and servers. Under the umbrella of interchangeable terms such as mobile edge computing, fog computing, and cloudlets, edge computing will revolutionize the mobile industry by attaching a small cloud one-hop-away from users which will streamline the provision of applications such as face recognition, augmented reality content delivery, big data analytics, and storage, to name a few.

Embedded in this cloudification process, which will re-architecture the network infrastructure, is the mindset to abide by the need to urgently reduce the energy footprint of the ICT sector which alone accounts for 3% of the entire global energy consumption and 2% of the total CO_2 emissions [1]. As for the mobile sector, it bites 0.5% of the entire global energy [2] where radio access networks (RANs) responds for 70% of it. Considering the computing counterpart, data centers have similar energy footprint figures where servers dominate the energy consumption with about 50% to 90% of total energy consumption. Since most the energy is drained at the wireless infrastructure and at the servers, there is a huge concern to address energy efficiency in a joint H-CRAN and edge computing deployment specially for an ultra dense deployment where wireless access points and servers will be massively integrated and replicated.

In order to shed the light on this issue, this article presents a comprehensive review of the recent breakthroughs on energy efficiency on H-CRAN and cloud computing individually and consolidates the discussion with the challenges and open issues of a joint deployment. To the best of our knowledge, this paper stands outs from previous work in the literature by uniquely addressing the full cloudification process where wireless infrastructure and edge computing are individually and jointly taken into account.

Manuscript received April 6, 2017; revised August 24, 2017 and October 16, 2017; accepted November 29, 2017. Date of publication December 6, 2017; date of current version May 22, 2018. The work of A. Anpalagan was supported by the National Science and Engineering Research Council of Canada under Grant RGPIN-2016-253546. The work of I. Woungangn was supported by the National Science and Engineering Research Council of Canada under Grant RGPIN-2017-04423. *(Corresponding author: Alagan Anpalagan.)*

A. Alnoman and A. Anpalagan are with the Department of Electrical and Computer Engineering, Ryerson University, Toronto, ON M5B 2K3, Canada (e-mail: ali.alnoman@ee.ryerson.ca; alagan@ee.ryerson.ca).

G. H. S. Carvalho and I. Woungang are with the Department of Computer Science, Ryerson University, Toronto, ON M5B 2K3, Canada (e-mail: glauciohscarvalho@gmail.com; iwoungan@scs.ryerson.ca).

Digital Object Identifier 10.1109/COMST.2017.2780238

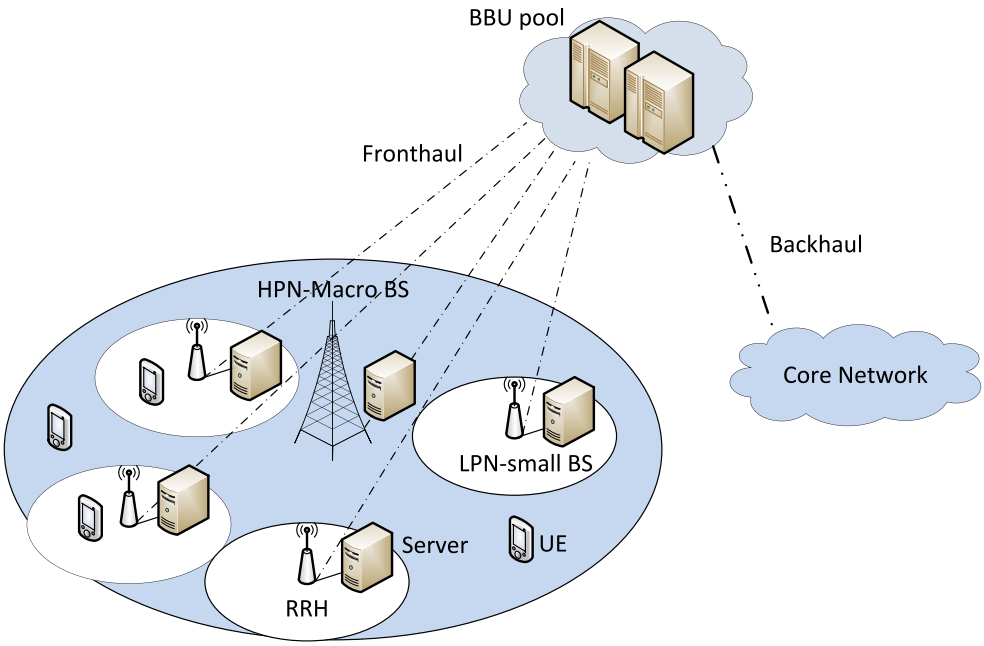

Fig. 1. Fully cloudified mobile network infrastructure.

The rest of this article is organized as follows. In Section II, we overview the architecture for a H-CRAN and edge computing joint deployment. Next, we individually address the sustainable design initiatives in either technology in Sections III and IV, respectively, followed by an analysis of the challenges and open issues in Section V. Finally, Section VI presents the closing remarks.

II. Network Architecture

Fig. 1 illustrates the network architecture in which communication and computing systems are fully virtualized. From the wireless network standpoint, the infrastructure represents the H-CRAN which is a merge of C-RAN and HetNets. C-RANs consist of remote radio heads (RRHs), baseband unit (BBU) pool, and wire/wireless fronthaul links. RRHs operate as soft relays that compress and forward signals from user equipments (UEs) to be processed in the centralized BBUs via wire/wireless fronthaul links, and vice versa. The BBU acts like a virtual base station (VBS) that provides signal processing and resource control. The assignment of BBUs to each RRH can be either in a distributed manner, whereby a RRH directly associates with its exclusive available BBU, or a centralized manner in which all RRHs enter a central processing entity that schedules and allocates processing resources among the RRHs. The former is easier to implement and the latter is more efficient in terms of benefiting the optimal resource management as well as offering the capabilities of implementing various algorithms such as interference cancellation, handover management, and RRH sleeping mechanisms [3]. HetNets, on the other hand, operate as multi-tier networks (e.g., macrocell with femtocells) that comprise multiple radio access technologies (RATs) in a unified entity. The nodes in H-CRANs are referred to as high-power nodes (HPNs) or low-power nodes (LPNs). HPNs provide seamless coverage, control signals (control plane is decoupled from data plane) to avoid unnecessary handovers in small cells, and backward compatibility with existing cellular networks. On the other hand, LPNs provide high data rates in the zones under their coverage, and increase network capacity through spatial frequency reuse [4].

A virtualized physical server, which is attached to the base station, is the pivotal building block of edge computing. Server virtualization optimizes the utilization of physical server's hardware resources by sharing them among multiple virtual machines (VMs). To orchestrate the execution of multiple VMs, a software layer called hypervisor or virtual machine monitor (VMM) is installed on the physical server. This installation process can be performed in two different ways: type-1 and type-2. For a type-1 deployment, the hypervisor runs directly on the physical hardware while for the type-2, it runs on the host operating system. As a result, type-1 deployment has better performance and less overhead when compared to the type-2 one, in addition to providing a more granular control. Each VM represents a specific computing environment with its own operating system, disk space, memory, and CPU cycles. To prevent performance degradation from taking place, VMs are isolated from one another while sharing the same physical host.

The benefits of the server virtualization are beyond the cost-effectiveness of coordinating multiple VMs on a single physical machine rather than purchasing multiple individual physical servers. From a resource management perspective, server virtualization enables the task consolidation, a process at which multiple VMs are clustered into a single physical server for a more efficient use of the computer's resource. Since resource utilization and energy consumption are tightly-coupled, task consolidation algorithms have been designed to save energy in cloud data centers. As it will be discussed latter, an energy-aware coordination between task consolidation

and radio resource management in a fully cloudified mobile network infrastructure stands out as one of the main challenges in a joint H-CRAN and edge computing deployment.

III. Energy-Efficiency on the Wireless Infrastructure

Boosting energy efficiency in cellular networks requires the establishment of large-scale coordination and cooperation among network nodes, resources, and functions. Integrating the cloud in the HetNet architecture can achieve significant progress towards that direction by virtualizing and bringing network functionalities in the central BBU pool. In this section, state-of-the-art techniques regarding energy, infrastructure, and resource optimization in cellular networks are presented.

A. Energy Efficiency in HetNets

The largest portion of energy in HetNets is consumed by the RANs, specifically base stations (BSs), and more than 80 percent of energy in wireless networks is lost as heat [5], [6]. The energy loss is mainly incurred by power amplifiers, which are the most power consuming components in BSs, and hence producing only 5 to 20 percent of useful output power [7]. Moreover, even with light or no traffic load, a BS consumes more than 90 percent of its peak energy [8]. BSs consume power mainly for operational purposes (e.g., cooling and signal processing) and radio transmission. Therefore, the deployment of large number of small cells, despite the fact that they consume small power, will increase the total power consumption in the network. However, small cells require less amount of cooling and most of the consumed power is exploited to broadcast the radio signals [9]. Table I presents recent energy-efficient approaches for HetNet management.

Macro BSs are generally aimed to provide large coverage areas rather than high data rates; therefore, the existence of small BSs is inevitable in future dense networks [20]. However, small cells are more prone to traffic variations than macrocell BSs [10]. A selective activation of femtocell networks in places that are characterized by concentrated traffic load, can significantly reduce the power consumption and outage probabilities [21]. Power consumption can also be reduced if the coverage area of a BS is adaptively reduced according to cell load [22]. Furthermore, achieving optimal handover decisions to reduce the unnecessary handovers can help minimize the power consumption. Song *et al.* [23] formulated the problem of optimal handover triggering as a constrained Markov decision process using stochastic information of handover parameters in areas containing overlapping coverage from multiple BSs.

The ultra dense deployment of small cells in hotspots such as shopping malls and airports leads to the underutilization of these cells at most times, thus to large energy losses [24]. Therefore, traffic offloading provides the opportunity of switching off lightly loaded BSs based on traffic demands [25]. With the adoption of appropriate sleeping strategies, the excessive amounts of consumed energy (e.g., cooling energy) can be avoided. Some of the recent advances that have happened in BS sleeping mechanisms are introduced in Table II.

The intra- and inter-RAT traffic offloading is considered as a promising solution to improving the energy and spectral efficiencies in HetNets. However, the incurred intra- and inter-RAT interference along with the increased burden on the capacity-limited backhauls degrade the total energy efficiency and spectral efficiency gains [11]. In [12], traffic demands have been sorted into real-time services such as video conferencing where fixed and high data rate provisioning are required as well as and non-real-time services such as file transfer with minimum data rate. The authors aimed to maximize the energy efficiency in a 2-tier HetNet considering both power allocation and beamforming design.

Shutting down the BSs with light or no load increases the amount of delay experienced by the users. This delay is generally due to the longer queue of users offloaded to the macro BSs; and SBS activation delay that can reach up to 30 seconds from off to on state [6]. Guo *et al.* [8] and Wu *et al.* [32] studied the energy-delay tradeoff in BS sleeping strategies. Moreover, Niu *et al.* [33] introduced the *N*-policy for optimal energy-delay tradeoff. In this policy, the BS will remain in the sleep mode until N users are accumulated under that BS coverage. The larger the value of N, the lower is the energy consumption and the higher is the delay experienced by the users. Thus, an optimal value of N should be determined to achieve the best energy-delay trade-off in cellular networks. A waiting period (referred to as hysteresis sleep) has been presented in [8] to maintain the system's stability while implementing the sleeping strategies. Hysteresis sleep is a certain amount of time or tasks that must be fulfilled before a sleeping decision is taken by a BS. Moreover, the study in [32] showed that under bursty traffic conditions, the total power consumption is less compared to normal load conditions, given the same average traffic load. This is due to the additional flexibility in determining the threshold of the number of users concentrated within a cell to sleep or wake-up. In other words, the BS will have the chance to sleep more often if the number of users stays below a relatively large threshold value, hence the total power consumption will be reduced.

For a mobile user to access the best candidate BS, a time delay of several hundreds of milliseconds is incurred. This situation can be even worse in ultra dense networks where the coverage radius of SBSs ranges from only several meters to tens of meters. This time delay is mainly due to the large-scale cooperation among different network elements. By decoupling the data and control planes, the MBSs will be responsible for selecting the best SBSs and providing the mobile users with the necessary information to start the access procedure with the SBSs. In this way, the small cell ID, resource block, time and frequency synchronization will be controlled by the MBS. In this paradigm, mobile users will receive data from both the SBSs and the MBSs in the areas supported by both tiers; otherwise, the MBSs will keep the data provisioning wherever the small cell coverage is missing [34]. To deal with the delay-aware radio resource allocation problems, a Markov decision process (MDP), which is a stochastic learning approach, is considered as a successful method since it has been recognized

TABLE I
ENERGY EFFICIENCY IN HETNETS

Reference	Research Direction	Problem Type	Solution Approach
[1]	Energy efficiency maximization considering transmit, backhaul, and circuit power in CoMP OFDMA-based HetNets under the constraint of data rate requirement	Optimization problem with constraints modeled as cubic inequalities	Lagrange multipliers and KKT conditions
[2]	Minimizing energy consumption in OFDM-based HetNets through power and subchannel allocation while satisfying users QoS requirements and inter-cell interference	Mixed-integer nonlinear programming	Iterative algorithm for searching the solution space with small granularity
[5]	Resource allocation for maximizing network utility and stabilizing the queues for media applications in HetNet with multi-homing transmission	Stochastic optimization problem	Lyapunov drift-plus-penalty method, primal-dual decomposition technique, Lagrange multipliers and KKT conditions
[10]	Joint optimization of cell activation, user association, and spectrum allocation	Mixed-integer programming	Reweighted l_1 minimization (majorization-minimization) method
[11]	Joint maximization of energy and spectral efficiencies considering small cell density and offload biasing factor subjected to throughput threshold	Quasi-convex multi-objective function optimization	Fritz-John conditions to determine the Pareto-efficient operational regime
[12]	Power allocation and beamforming design	Mixed combinatorial non-convex optimization	- Energy efficiency maximization is transformed to power minimization and an optimal solution is obtained based on convex programming (semi-definite programming (SDP) with Lagrangian relaxation) - Near-optimal upper-bound solution for more simplicity - Suboptimal zero forcing (ZF)-based solution to further simplify the solution
[13]	BS sleeping using the concept of group sparsity in transmit power vector, BS association, and downlink power allocation in TDD-based HetNets	NP-hard non-convex optimization	Successive convex approximation (SCA)-based algorithm, KKT conditions
[14]	Joint optimization of BS operation, user association, subcarrier assignment, and power allocation	Mixed combinatorial problem	Lyapunov optimization and heuristic algorithm
[15]	Minimizing grid energy consumption through cooperative cell operation with hybrid energy sources	NP-hard problem	The problem is divided into to subproblems using greedy decomposition
[16]	Energy-efficient control of BSs transmit power (cooperative and non-cooperative)	The non-cooperative power control is formulated as a non-cooperative power control game, whereas the cooperative part is a multi-objective optimization	Nash equilibrium and heuristic algorithm
[17]	Energy-efficient trade-off between backhaul energy and throughput subject to QoS and fairness constraints	Multi-objective optimization	The multi-objective optimization is transformed to a single-objective optimization using weighted sum method, then solved using iterative algorithm and Lagrange multipliers
[18]	Maximizing energy efficiency per individual user in multi-RAT HetNets under QoS constraints	Multi-objective optimization problem	Determining the Pareto optimal solution using weighted Tchebycheff method and iterative algorithms
[19]	Joint cell association and on-off policy to minimize energy consumption	General non-convex energy minimization problem	0-1 Knapsack-like optimization

as superior to other optimization techniques such as Lyapunov optimization and the equivalent rate constraint approach [35].

To cope with the network heterogeneity and maintain the hierarchical control, a multi-level decision-making strategy for a macro-femto network was formulated in [36] as a multi-level optimization problem where the decision making in one level (e.g., macro, femto, or user) depends on the decision made in other levels. In this scheme, femto BSs provide access control for their users, while central schedulers allocate radio channels to BSs. Afterwards, macro and femto BSs allocate the available resources to their associated users who in turn select the desirable transmit power levels. Furthermore, Adedoyin and Falowo [37] proposed a joint subcarrier and power allocation scheme to increase the throughput, mitigate the interference, and minimize the power allocated to each user under the constraints of interference (co- and cross-tier interference), QoS, and fairness of subcarriers allocation in a HetNet consisting of macro and femto cells. The optimization problem was solved using Lagrangian duality method and Karush-Kuhn-Tucker (KKT) optimality conditions.

Motivated by the abundance of free and green energy resources, renewable energy harvesting has been studied by many researchers as an alternative resource to empower the SBSs thereby reducing on-grid power consumption. SBS types

TABLE II
BS SLEEPING

Reference	Research Direction	Problem Type	Solution Approach
[6]	Energy efficiency maximization using random and strategic sleeping strategies for small cell BSs under the constraints of coverage and averaged wake-up time	Non-convex optimization	Near optimal solution by maximizing the lower bound of energy efficiency through iterative algorithms
[25]	Joint energy-efficiency and load-balancing in multi-RAT HetNets	Semi-Markov decision process (SMDP)	Optimal policy using Markov decision process
[26]	BS on-off and traffic offloading scheme based on traffic load and renewable energy availability	0-1 knapsack problem	Lagrange multipliers
[27]	Distributed cooperative sleeping strategy for energy saving	Constrained graphical game where BSs act as players under the traffic load constraint	Iterative algorithm is used to find the generalized Nash equilibrium
[28]	BS sleeping strategy and user association in open-access femtocell networks	Binary integer problem	Heuristic algorithm and Lagrange dual method
[29]	Energy efficiency maximization by determining BS density and sleeping strategy	Non-convex optimization	Dynamic gradient iterative algorithm
[30]	Optimal BS activation and cell size determination subject to network coverage	NP-hard	Polynomial-time algorithm
[31]	BS modules (e.g., power amplifiers, cooling, processors, etc.) activation and deactivation based on traffic variation	Discrete time Markov decision process (DTMDP)	Optimal policy is determined based on the probabilistic decisions of the MDP which is solved by linear programming approach

that exploit renewable energy resources include the off-grid SBSs that solely rely on renewable energy resources, or hybrid SBSs that exploit both renewable energy resources along with on-grid power using optimal allocation strategies [26]. Furthermore, mixed renewable energy resources and grid power have been studied in [38] and [39] to minimize the grid power consumption by allocating renewable energy resources efficiently along with BSs on-off strategy.

Instead of shutting down the entire BS site, the schemes of discontinuous transmission (DTX) and reception (DRX) have been considered as successful approaches to save energy in many works in the literature. The DTX, whereby some BS components are switched off when no transmission is required, has been considered in the study of [40] to improve the network's energy efficiency. From the receiver side, in order to save energy and prolong the battery life in mobile devices, a DRX mechanism has been utilized in LTE/LTE-A, where UEs enter a sleep state when no data transmission is required. A lighter sleep state or listen state that lies between the active and sleep states can be incorporated in the sleeping process, to enable UEs to be activated faster when detecting an incoming traffic [41].

Other schemes such as bio-inspired systems have been incorporated in the context of self-organizing networks. For example, the work in [42] defined the analogy between mammalian immune systems and cellular networks, and proposed an artificial immune system for ultra-dense small cell networks where the BS activation is managed autonomously for enhanced energy efficiency and reduced delay depending on traffic variations. Cell zooming is another technique that aims to minimize energy consumption by adaptively shrinking the cell size according to traffic density within the cell [43], [44].

Improving energy efficiency and minimizing power consumption have been alternatively considered in the literature. However, Sun and Wang [2] emphasized on minimizing the energy consumption rather than maximizing the energy efficiency that combines energy consumption with data throughput. This realization is due to the fact that the total network energy efficiency might not satisfy the per-user or per-cell energy and data requirements in HetNets.

B. Energy Efficiency in H-CRANs

Energy efficiency, defined as the ratio of the total data throughput in the network to the total power consumption under the constraints of users' quality-of-experience (QoE) requirements, is presented in this section in the context of H-CRANs. Originating from the necessity of seamless coverage and high data rate provisioning, future cellular networks are categorized as ultra dense and consist of large numbers of BSs and mobile devices. Taking into account the heterogeneity nature of the network along with the fact that the majority of mobile devices are battery-powered, it is becoming more and more important to minimize energy consumption while taking into account the QoS provision such as data rate, end-to-end delay, fairness, and deployment costs [45].

Supporting H-CRANs with a software-defined architecture makes it more convenient to upgrade the performance and services provided by network operators, and facilitates an elastic deployment of technologies and applications for future demands. Moreover, the central controllers help perform network-wide updates in system behavior instead of the individual configuration of network elements [46]. For instance, Cao *et al.* [13] proposed a power model for cloud-assisted HetNets, wherein the large-scale fading imposed on the communication channel is considered to be fixed. Their model jointly optimizes the power consumption of signal processing, circuits, downlink transmission, and backhaul for providing more flexible BS sleeping than the traditional on/off operation. The study that considered 3 macro BSs and 5 pico BSs within each macro BS, showed that the average power consumption of all BSs for a 20 Mbps user data rate was approximately

2.7 kW with the cloud-assisted architecture, which is less than that of the simple on/off strategy that required 3.2 kW. Furthermore, in [47], a scheme was proposed to minimize the energy consumption in dense C-RANs by activating some selected subsets of RRHs according to the dynamic traffic variations. Results showed that the dynamic activation of RRHs was successful in making significant energy saving. For example, within the coverage of 2 micro and 7 pico BSs, the consumed power using the dynamic activation scheme for 50 users was approximately 60 W which is much less compared with the 250 W power consumption incurred by activating all RRHs regardless of the available traffic. When the number of users increased to 250, power consumed using the former scheme increased to 200 W and remained 250 W for the latter. In addition, the deployment of a large number of co-located antennas in massive multiple-input multiple-output (MIMO) can improve the spectral efficiency by up to 10 times, and the energy efficiency in the order of one hundred, as compared to the performance of a single-antenna. Furthermore, implementing efficient cooling systems in the BBU pool helps improve the energy efficiency in H-CRANs [48].

Cloudified networks support the revolutionary SDN architecture which provides a comprehensive control over the entire network nodes in programmable softwares. For instance, incorporating information security schemes such as MD2 Hash functions which require both communication and computing resources with SDN reduces the energy bill to only 3 micro joules per byte compared to 4 micro joules per byte without SDN. This is because the SDN adapts the network parameters such as sleep/wake and link speed according to the real demands. Such features attracted companies like Google to implement SDN in their data centers [49]. The architecture of software defined Energy Internet (SDEI) by which energy is managed in a similar way as data was presented in [50] for smart power systems. This architecture adds another plane; namely the energy plane, to the data and control planes, and confine the functionality of this energy plane to providing intelligent energy control in the network. The SDEI could also be implemented in cellular networks to achieve advanced energy management, where energy routers deliver energy to the network nodes according to the decisions made by the control plane that is aware of the entire network status. Energy is then delivered to the desired IP address in the desired form of energy (e.g., DC, AC, single phase, etc.).

From the computational complexity perspective, the high density of RRHs required to provide high data rates incurs high computational complexity due to the huge amounts of data related to signal processing, resource allocation, and RRHs/BBUs coordination. This complexity is a big challenge facing the establishment of scalable networks. Fan *et al.* [51] introduced several schemes to provide scalability in C-RANs. These schemes were sorted into (a) signal processing techniques such as exploiting the near sparsity in channel matrix to minimize the channel estimation overhead; (b) resource allocation using optimization techniques such as game theory, graph theory, and matching theory to minimize the high computational complexity of solving the combinatorial optimization problems; and (c) RRH/BBU coordination schemes such as the on/off operation of RRHs and BBUs depending on traffic load.

To maintain high data rate provisioning and efficient resource utilization, Peng *et al.* [52] proposed a joint optimization of resource block (RB) assignment and power allocation subject to the constraints of user association and interference between RRHs and HPNs in orthogonal frequency division multiple access (OFDMA)-based H-CRAN systems. The work considered soft fractional frequency reuse (S-FFR) in which the RBs are divided into two sets: the first is dedicated to UEs associated with the RRHs (RUEs) that require high QoS, and the second set is shared between RUEs and UEs that are associated with the HPN (HUEs) and require low QoS. Moreover, the RBs in time/frequency domains have been identified in [53] as power zones (PZs) in which the problem of scheduling each user to specific PZ along with the PZ power level was formulated and solved using graph theory. The problem solution was viewed as a scheduling graph, wherein each vertex represents the individual association of UEs, BSs, and PZs. In [54], a scheme was presented to maximize energy efficiency and maintain the multimedia traffic queue stability in H-CRANs taking into consideration the instantaneous power, average power, fronthaul capacity, and inter-tier interference.

The idea of heterogeneous carrier communication, in which cellular networks are deployed over unlicensed frequency bands, has been extended to the standardization process under the title of licensed-assisted access to break through the obstacle of limited spectrum resources. This technique however, incurs interference with licensed communications thus limiting the power and reliable transmission range of BSs. A proposed solution to the aforementioned concerns, is to allocate control signals to the licensed bands while reserving unlicensed bands for data transmission, thereby providing better long-range control while exploiting the additional bandwidth to improve throughput. Another mechanism referred to as listen-before-talk can help avoiding interference by obligating all transmitters to sense the ambient channels and proceed in transmission only when channels are clear [55].

Moreover, to maximize the spectrum and energy efficiencies, and to achieve ultra-low-latency communications, Lien *et al.* [56] introduced the open-loop communications as a promising technique to fulfill these requirements by avoiding the redundant feedback messages of channel state information (CSI) or reception acknowledgments from the massive number of receivers. Therefore, the transmitter autonomously determines the optimum modulation and coding schemes as well as the required number of repeated transmission in one shot. In regard to uplink transmission in H-CRANs, a scheme was proposed in [57] to jointly optimize power allocation, relay selection, and network selection under the QoS constraints in order to maximize the network's energy efficiency. Table III summarizes recent trends followed to achieve energy-efficiency in H-CRANs.

C. Resource Management and Network Optimization

To take full advantage of the high computing capabilities provided by the cloud servers, it is paramount to

TABLE III
ENERGY EFFICIENCY IN H-CRANS

Reference	Research Direction	Problem Type	Solution Approach
[37]	Optimizing subcarrier and power allocation for femtocell users based on cognitive radio technology under the constraints of QoS and interference mitigation	Multi-objective optimization	Lagrangian dual decomposition and KKT conditions
[47]	RRH subset activation based on traffic demand and sleeping strategy	Multiple choice multi-dimensional knapsack problem	Lagrange multipliers
[52]	RB assignment and power allocation in OFDMA-based H-CRANs under the constraints of inter-tier interference and RRH/HPN association	Non-convex nonlinear fractional programming optimization	Lagrange dual decomposition and KKT conditions
[54]	Maximizing energy efficiency while maintaining multimedia traffic queue stability under the constraints of instantaneous power, average power, fronthaul capacity, and inter-tier interference	Non-convex stochastic optimization problem formulated by minimizing the drift-plus-penalty function	Lyapunov optimization framework and weighted minimum mean square error
[57]	Joint optimization of power allocation, relay selection, and network selection in uplink H-CRANs under the constraints of QoS requirements	Mixed-integer non-linear non-convex problem	Dinkelbach method and Lagrange dual decomposition
[58]	Optimizing the transmit power of RRHs and HPNs with an interference mitigation strategy	Non-convex optimization	The problem is transformed to a convex optimization using Dinkelbach method and duality gap theorem, then solved by Lagrange dual decomposition method
[59]	Maximizing the average throughput and network stability (queue stability) subject to the constraints of power allocation, RB allocation, and user association	Stochastic optimization containing a mixed-integer subproblem	Lyapunov optimization and Lagrange dual decomposition method (to solve the subproblem)
[60]	Designing coverage areas of macro BSs and RRHs, then allocating resources among RRHs to achieve balanced transmission bandwidth on fronthaul	Mixed-integer program	Lagrange multipliers
[61]	Dynamic subcarrier and power allocation, CoMP, and RRH clustering in TDD-based H-CRANs	Mixed strategy non-cooperative game	Reinforcement learning algorithm is used to achieve Logit equilibrium
[62]	Joint RRH selection and user association to minimize energy consumption considering the fronthaul capacity and spectrum resources	Integer programming problem	- Multiple-choice multidimensional knapsack model is used for user-RRH association - RRH selection is solved using low-complexity heuristic algorithm
[63]	Cross-tier cooperation and cluster formation among LPNs and HPNs towards throughput enhancement	0-1 Multiple knapsack	Heuristic algorithms
[64]	Joint BS selection and beamforming design for power minimization under the constraints of limited fronthaul capacity	l_0 minimization problem	Majorization-Minimization algorithm
[65]	Joint admission control and coordinated beamforming under the constraints of fronthaul capacity, RRH maximum power, and minimum SINR experienced by users	NP-hard optimization	The problem is reformulated to a single-stage semi-definite program using convex relaxation approach

resolve the performance bottlenecks of cellular networks regarding the infrastructure and radio resources. For instance, the performance of task offloading in cell-edge computing could be severely declined under the condition of inter-cell interference especially in ultra-dense networks. Moreover, when a large number of mobile devices tend to offload tasks to the cloud through cellular networks, the transmission delay could increase due to the limitations in radio resources [66]. For this reason, radio and computational resources have to be jointly optimized to achieve the foreseen high QoS requirements regarding energy efficiency, computing performance, end-to-end delay, and throughput for future 5G networks [67] and smart cities [68]. To this end, the issues of frequency allocation, interference coordination, RRH clustering, and fronthaul/backhaul management are thoroughly reviewed in this section.

1) Radio Resource Management: In mobile communication networks, the per-user demand is naturally fluctuating between day and night, weekdays and weekends, residential and commercial areas, in a phenomenon referred to as the tidal effect. To cope with the aforementioned challenge, an elastic resource utilization has been adopted in many research work such as [69], where the RRHs activation and BBUs capacity (e.g., processor speed, memory, etc.) can adapt to the variations in data demands. The work considered two schemes: a) proactive, where resources are provided in advance based on the knowledge of traffic patterns (e.g., weekdays and weekends); and b) reactive prediction that is based

on the time-series analysis of traffic records from real-time or historical data.

As reported in [70], resource sharing in H-CRANs can be divided into three levels: 1) spectrum sharing, which includes RBs sharing in Long Term Evolution (LTE), channel sharing in IEEE 802.11, and the unused spectrum portions named white spaces; 2) infrastructure sharing, where the central workload computations of RRHs and HPNs in the BBU pool facilitate the virtualization of available resources from different physical entities (e.g., base stations, backhaul, and routers) using the techniques of network function virtualization and software-defined networking. Therefore, network functionalities can be decoupled from hardware components. This facilitates the infrastructure sharing among network operators while reducing the CAPEX and OPEX; and 3) network sharing to efficiently manage the available spectrum and infrastructure resources.

Aggregating network information in the central H-CRAN processors helps achieve superior resource and interference management. Dahrouj *et al.* [71] proposed the coordinated scheduling, hybrid backhauling, and multi-cloud association as promising resource allocation schemes for H-CRANs. Unlike the legacy fairness-based allocation schemes, coordinated scheduling is performed in the cloud processors which are responsible for synchronizing the scheduling process in the network. Therefore, scheduling of users to BSs and resource blocks can be performed in the cloud servers. Hybrid backhauling refers to the joint utilization of wireless and wired links, this helps to cope with the fluctuating demands in H-CRANs. The multi-cloud scheme, can benefit the network by reducing the computation burden on central servers, and the complications faced when connecting distant BSs.

Dynamic resource management techniques such as the dynamic load-aware RRH assignment using bin packing algorithm can reduce the number of active BBU servers through many-to-one mapping, thus saving energy and computing resources [72]. Furthermore, a graph-based dynamic frequency reuse has been presented in [73], whereby each RRH within the H-CRAN is viewed as a single vertex in the graph. In addition, graph coloring was used to allocate different bandwidths according to traffic demands, thereby alleviating the inter-tier interference. Adaptive machine learning techniques are also incorporated in the centralized signal processing to achieve an intelligent networking performance that can adapt to data demands (e.g., IoT demands) that fluctuates over time and place [46]. To further increase the system capacity, radio resources could be borrowed from RRHs with low traffic loads, and conveyed to the overloaded neighboring RRHs (homogeneous or heterogeneous) [46]. Moreover, a multi-homing transmission, which is defined as splitting the media traffic simultaneously onto multiple RAN links between UEs and the media content server, can improve the QoS of media applications within the network [5].

Interference mitigation techniques proposed for H-CRANs will be introduced next as one of the main concerns in ultra dense environments. Afterwards, RRHs clustering will be presented as a promising coordination paradigm for enhanced interference cancellation and improved network performance.

2) Interference Coordination: Heterogeneity in cellular networks that comprise base stations with different sizes and RATs, can significantly improve the total system capacity; however, high interference levels will be incurred. Moreover, the dense deployment of RRHs produce severe inter-cell interference due to the relatively short distances between adjacent RRHs leading to higher signal power received by users served by neighboring cells [61]. H-CRANs support the enhanced inter-cell interference coordination (eICIC) through the techniques of advanced carrier aggregation (CA) in the frequency domain, and almost blank subframes (ABS) in the time domain. Moreover, the required signal processing of the related cells is concurrently performed in the same BBU [70]. In addition, the inter-tier interference coordination in HetNets in both time and frequency domains using the ABS technique is achieved by reducing the transmit power of macro BSs to avoid interfering with smaller BSs. Whereas the dynamic point blanking (DPB) technique mutes the interfered signals among the coordinated BSs [74]. Interference coordination techniques for H-CRANs are presented in Table IV.

A careful allocation of the network resources can help minimizing the interference. For instance, dividing the coverage area into sub-regions with different frequency sub-bands in the technique of S-FFR plays an important role in inter-tier and inter-cell interference coordination. Unlike the traditional S-FFR where the allocation of frequency sub-bands are orthogonal, the enhanced S-FFR in H-CRANs enable RUEs to share radio resources with HUEs even at cell-edges [46]. Moreover, a dynamic resource allocation scheme in [75] has been presented to perform global and local resource allocation strategy to optimally share resources among different service providers in C-RANs. The work considered the constraints of limited fronthaul capacity and a threshold-based interference among RRHs in order to achieve optimal resource sharing. Global resource sharing deals with large time-scale traffic variations whereas the local resource sharing performs actions regarding traffic variations in a small time scale. Hierarchical access to frequency resources based on the concept of cognitive radio can also be applied by femtocells which act as secondary users that use frequencies only when no primary users are using that particular frequency. This helps avoiding the overlapping of signals with other users associated with other cells [83].

Incorporating the cloud in the interference cancellation process can significantly facilitate the cooperation among interference sources. Joint cooperative interference mitigation and handover management scheme was proposed in [84] to increase the capacity of H-CRANs by coordinating the functionality of C-RANs and small cells. The work considered the formation of RRH clusters for joint transmission in order to coordinate interference especially for cell-edge users. On the other hand, the handover scheme sorts users based on their speed, and prevents handover from macro to small cells for users characterized as high-speed users. Moreover, the implementation of multiple RATs with different frequency bands can improve radio resource utilization since different RATs use different frequency bands [18]. Moreover, the intra-tier interference among LPNs can be mitigated using

TABLE IV
INTERFERENCE MITIGATION IN H-CRANS

Reference	Research Direction	Problem Type	Solution Approach
[75]	A threshold-based interference control among RRHs that belong to different service providers in order to limit the maximum aggregate interference received by users	Mixed-integer nonlinear programming	First, the problem is linearized, then, a suboptimal solution is obtained using increment-based greedy allocation algorithm
[76]	Interference coordination between MBSs and RRHs	Contract-based optimization	Contract-based game theory, Lagrange multipliers, and KKT conditions
[77]	Suppression of inter-tier interference between RRHs and MBSs using interference collaboration and beamforming	Non-convex optimization	The problem is transformed to a convex optimization, then solved using Lagrange multipliers and KKT conditions
[78]	Inter-tier-interference-aware macrocell paradigm for radio resource allocation such that a macrocell can maximize the interference levels tolerated by its associated users under QoS constraints	Mixed-integer nonlinear programming	Successive convex optimization
[79]	Inter-tier interference reduction using power hierarchy scheme (e.g., macro, femto)	Non-cooperative game (high- and low-power BSs are the players)	Nash equilibrium and KKT conditions
[80]	Joint optimization of RRH clustering, user grouping, and transmit beamforming	Non-convex combinatorial optimization	- Dynamic scheduling algorithm to form user grouping and RRH clustering, - Iterative algorithm for transmit beamforming, - Lagrange multipliers were also used in the solution
[81]	Inter-tier interference mitigation among HPNs and LPNs in H-CRANs through optimized power allocation	Non-convex optimization problem	Perron-Frobenius theory
[82]	Joint user-access point (AP) association and beamforming design for interference coordination in both uplink and downlink transmission	NP-hard	Group-sparse optimization, and relaxed-integer programming

cloud-based large-scale cooperative signal processing. For the inter-tier interference between the high- and low-power nodes, it can be suppressed through cloud-computing-based cooperative radio resource management (CC-CRRM) that incorporates the BBUs and the HPNs via the X2 interface [48]. Furthermore, since downlink and uplink signals are both known by the C-RAN servers, downlink-to-uplink interference can be cancelled by subtracting interference from received signals to recover the original signals [85]. A contract-based interference coordination between RRHs and MBSs in H-CRANs was presented in [76]. In this scheme, the BBU is considered as the principal that offers a contract to coordinate transmission scheduling among RRHs, MBSs, and UEs. The contract is then sent to the agents (MBSs) to be accepted or rejected depending on the acquired benefits regarding spectral efficiency.

In addition to the central cloud processing, a decentralized multiple cloud architecture in C-RAN was proposed in [86] to minimize the total power consumption with reasonable amount of information exchange among clusters. The problem that considered both intra- and inter-cluster interference, achieved energy minimization by determining the sets of active BSs per cluster and the sparse beamforming vectors of users in the network. In [77], a strategy for inter-tier interference suppression between RRHs and MBSs is proposed for H-CRANs using the techniques of interference collaboration, beamforming, and cooperative radio resource allocation. Results showed that the proposed strategy has led to an increased H-CRAN performance depending on the network configurations such as the number of antennas deployed by the MBSs, number of RRHs, and SINR threshold.

Serving large number of users simultaneously in certain zones by macro and small cells incurs strong inter-tier interference. By exploiting the geographic spacing among users, this problem can be alleviated. For instance, the technology of massive-MIMO provides the opportunity of transmitting high directional beams in certain directions and thus providing spatial blanking in other directions. As a result, small cells lying in the blank directions can avoid interfering with the macro cell signals [87]. In a similar manner, device-to-device communication was proposed in [88] to avoid the excessive interference in H-CRANs by establishing D2D links at a certain distance away from the HPNs. Results showed that such strategy can achieve high SINR and low average traffic delivery latency to cope with the limitations of capacity and time-delay in fronthaul links.

3) RRH Clustering: In small networks, modest amounts of CSI acquisition is required, and therefore the interference alignment can be jointly applied on all BSs. In larger networks however, exchanging CSI data among all BSs is large and impractical, thus BS clustering is essential to maintain high QoS [89]. To this end, incorporating large number of cells to form larger clusters will lead to better spectral efficiency and interference cancellation; however, with other factors taken into consideration such as delay and channel estimation (e.g., minimum mean square error) and precoding (e.g., zero-forcing precoders), the performance improvement will not be as high as expected [90]. Moreover, cells from different tiers can cooperate and form a cross-tier cluster that better serves a particular user. This formation is known as user-centric cross-tier clustering wherein the user is geographically located at the center of the cluster [20].

TABLE V
BS CLUSTERING

Reference	Research Direction	Problem Type	Solution Approach
[89]	BS clustering based on long-term user throughput considering CSI overhead and spectral efficiency	Coalition game where BSs are considered as the players	Distributed coalition formation algorithm and a precoding algorithm based on weighted minimum mean square error
[92]	Joint optimization of BS transmit power, BS activation, and backhaul power, using different strategies; namely, the data-sharing and the compression strategies	Discrete non-convex optimization	Reweighted l_1-norm minimization and successive convex approximation
[95]	User-centric BS clustering and sparse beamforming, wherein BSs are equipped with limited storage cache to reduce the burden on backhaul links	Mixed-integer nonlinear programming	Iterative reweighted l_1-norm technique
[96]	Joint RRH clustering and activation optimization under the constraints of coverage and user's QoS	NP-hard problem	Linear-programming relaxation and greedy algorithm
[97]	Cell clustering and activation time for energy minimization subjected to data provisioning and inter-cell interference	NP-hard problem	Column generation, local-enumaration-based bounding scheme, and near-optimal cluster scheduling

It has been shown in [91] that in RRH clustering, the coordinated beamforming (CB) performs better than the zero forcing beamforming (ZFBF). This is because ZFBF aggressively allocates power to RRHs, and thereby incurring higher levels of inter-cluster interference. As a result, no gain was obtained regarding the cluster's sum-rate. On the other hand, the CB improves the sum-rate because it manages the interference more efficiently by controlling the transmit power of coordinated RRHs. It was also shown that global clustering in which all RRHs form one large cluster, achieved better performance than local clustering whereby only few neighboring RRHs form a small cluster. Larger clusters however, require more piloting overhead (training symbols), in addition to the time, frequency, and phase synchronization among clustered RRHs.

A comparison of data-sharing and data compression strategies has been studied in [92]. Data-sharing means that BSs apply beamforming locally after receiving messages from the central server, and then multiple BSs cooperatively transmit to common users. In the compression strategy, the processes of precoding and beamforming are executed in central servers. It is also worth mentioning that in low data rate requirements, data-sharing is found to require less power, whereas in high data rates, the compression is preferred because backhaul will require more power.

Dynamic virtual cluster formation has been proposed in [93] to mitigate inter-cluster interference in OFDMA-based systems. Unlike the traditional omni-subcarrier CoMP, the work considered each cluster as a uni-subcarrier such that each cell could be grouped with different virtual clusters and thus dealing with different subcarriers. Moreover, a branch and bound algorithm has been proposed in [94] to find the global optimum BS clustering considering the inter-cluster interference and CSI overhead. The algorithm was capable of achieving optimality with low complexity compared to exhaustive search algorithms. Table V lists some of the technical approaches applied in cell clustering.

4) Backhaul and Fronthaul Management: Radio resources cannot be fully exploited without having sufficient capacity in the fronthaul and backhaul links. Fronthaul links are generally defined as the connecting media (wired/wireless) between the RRHs the BBU pool, whereas backhaul links maintain the connection between the BBU pool and the core network. Thus, providing high bandwidth transmission in the fronthaul and backhaul links is considered as one of the major challenges facing the implementation of H-CRANs especially with the implementation of intra- and inter-cell CoMP techniques. Besides, the under-utilization of the full backhaul capacity, which is designed for peak bandwidth provisioning, is another challenge due to the geospatial fluctuations that characterize the traffic nature. Fortunately, the decoupling of data and control planes along with the support of HPNs, has made significant improvements in alleviating the load burdens on backhaul and fronthaul links.

In [1], two types of backhaul links have been proposed, namely inter-backhaul between MBSs and the central server, and intra-backhaul between the RRHs and the local server that is located within the boundaries of one large cell. The inter-backhaul links consist of optical fiber cables whereas the intra-backhaul contains both fiber cables and wireless links. In order to minimize the transmission bandwidth in backhaul links, data compression techniques are envisioned as a promising solution. Such techniques could be applied in the time domain, such as reducing the sampling rate or using non-linear quantization, or in the frequency domain such as subcarrier compression with FFT/IFFT. Moreover, workload balancing algorithms can assist in reducing the peak data transmission and decrease the requirements to less than one third of the total bandwidth [98].

With the dense utilization of small cells, wireless backhaul links are considered as a scalable and cost-efficient approach compared to fiber optical cables that are more suitable for cells characterized as large or medium cells. However, wireless backhaul relies on the wireless medium which is delay prone [71]. Based on the observations of [99] and [100], two-tier networks with wireless backhaul are more energy efficient than single-tier networks, provided that an optimal bandwidth division is conducted between the wireless backhaul and radio access links. Furthermore, bandwidth partitioning between wireless backhaul and wireless access links for both uplink and downlink transmission has been presented in [101] as a

sharing technique that can maximize energy efficiency in small cell HetNets.

Zhang *et al.* [102] found that the co-located call patterns at the same BS are highly correlated due to their social interplay. In other words, a mobile user pair tends to make a face-to-face communication after their call. By extracting and analysing a large-scale mobility traces, user locations can be predicted several hours ahead. This location prediction process can be implemented in the cloud to improve resource management and QoS provisioning; moreover, it fosters the addition of location-based social services. To make use of these social patterns, traffic caching is envisioned as a promising solution for reducing traffic loads on the backhaul. Caching strategies aim to store redundant and frequently accessed contents in the BBU pool, thereby enabling direct access by UEs and avoiding the need to access the core network through the backhaul [103]. For instance, if the data of a particular user is predicted in advance and cached during off-peak hours, then it can be retrieved and transmitted during peak load hours without adding burdens on backhaul links [104]. This feature transforms the network behaviour from being reactive to proactive.

In [105], a compress-and-forward scheme for transferring data from BSs to central cloud processors in uplink C-RANs has been introduced. It has been shown that by maintaining the quantization noise levels proportional to the background noise gives a near optimal performance for backhaul capacity allocation especially when the signal-to-quantization-noise-ratio (SQNR) level is high. BSs perform the compress-and-forward process to achieve more efficient transmission through fronthaul links. In this process, received signals are quantized within BSs using various techniques such as single-user compression and Wyner-Ziv coding. Unlike the single-user compression, Wyner-Ziv coding utilizes the correlation between signals received in other BSs and hence improves the total compression performance [106].

From the fronthaul perspectives, establishing multiple connections between a mobile user and multiple RRHs within the same tier or with other tiers can improve the spectral efficiency through coordination techniques (e.g., CoMP); however, the costs on fronthaul resources (e.g., energy and bandwidth) will be high. Therefore, optimizing the size of the associated RRH/HPN clusters is essential to maintain the tradeoff between benefiting the spectral efficiency or wasting the fronthaul resources [48]. Each cluster is controlled by a single server via the fronthaul links. In addition, the connection between the RRHs and the BBUs may have a single-hop or multi-hop topology by relaying through other RRHs until reaching the desired server [107].

In order to carry the massive amounts of data from the RRHs to the BBU pool, two forms of data transportation have been introduced in [108], namely radio over fiber (RoF) whereby data are transferred in an optical form, or digitized IQ samples which can be carried on wired or wireless links. The authors also presented the concept of the partially centralized C-RAN (PC-RAN) in which baseband signal processing is divided between the BBU and RRHs. Thus, precoding and data modulation are processed in the BBUs, whereas radio transmission is performed by the RRHs. Furthermore, integrating the functionalities of the physical, medium access control (MAC), and network layers incurs significant signaling overhead on fronthaul. Thus, partial centralization which incorporates only physical layer functionalities in the RRHs, can significantly reduce the burden on fronthaul links since the physical layer computation requirements are the highest compared to other layer requirements. However, the performance of RRH coordination techniques such as CoMP could be degraded. A promising solution is the clustering of RRHs based on their geographical locations. This can reduce the scale of cooperative processing in the BBU pool; and as a result, reduce the load on fronthaul links. RRH clustering can take the form of disjoint clustering or user-centric. In disjoint clustering, the coverage area is pre-divided into specific zones to provide common service. This technique, however, subjects mobile users to face inter-cluster interference especially at cluster borders. On the other hand, user-centric clustering combines neighboring RRHs to form local clusters wherein users are located in the cluster center [35].

A one-to-multiple mapping between a BBU and RRHs can be applied to reduce the load on fronthaul links and to efficiently utilize the BBU computing resources. This configuration enables addressing the spatial and temporal traffic load variations; and moreover, supports saving energy in the BBU pool by switching off the BBUs identified with light loads [109].

To overcome the limited capacity in fronthaul links, time-reversal (TR)-based communications for air interfacing have been proposed in [110] to exploit the characteristic of location-specific signature in order to combine multiple signals and send them concurrently through fronthaul links without additional bandwidth requirements. In TR-based communications, a pilot signal is received by the transceiver prior to transmission. The normalized time-reversed conjugate of that signal is then being saved as the waveform used for transmission. With this strategy, TR-based communication overcomes the multipath effects of the communication environment by acting as a matched filter that adjusts the temporal and spatial effects.

IV. Energy-Efficient Cloud Computing

The plethora of information and mobile applications that are expected to dominate the future life style is mainly fueled by the powerful computing capabilities provided by the cloud. However, many challenges are still facing the cloud computing world regarding energy, task scheduling, etc. This section sheds light on recent computing advances in both servers and mobile devices.

A. Cloud Computing

In future networks, the massive number of machines are expected to maintain always-online connections with high upload demands. Cloud computing is envisioned as a promising technology for alleviating the challenges that arise from the bursty nature of the traffic load, in which machines generate large amounts of data in short periods of time, and remain silent for a longer time thereafter [111].

Servers dominate the energy consumption in data centers with about 50 to 90 percent of total energy consumption; as a result, it is essential to consider energy-efficient servers in future networks. From the processing perspective, the utilization of reduced instruction set computers (RISCs) is more energy-efficient due to the lesser number of integrated transistors compared to the complex instruction set computers (CISCs). Moreover, integrating the processor cores, memory, and input/output modules on a single chip namely system-on-chip (SoC) design reduces electric interconnections and hence power requirements. Graphic processing units (GPUs), on the other hand, provide more energy-efficient performance than central processing units (CPUs) especially in floating point calculations [112].

The dynamic workload in the cloud comes mainly from the fluctuation in computing demand and computing priorities. In addition to being dynamic, cloud servers are also described as heterogeneous. The heterogeneity in cloud servers is due to the utilization of various hardware components that belong to different models and generations. For example, processors could have different speeds and architectures, memory modules could posses various capacities, and energy consumption depends on how much workload is performed [113].

About 18% of the operational expenditures of data centers goes to the energy sector [113], and as time goes by, the traditional concept of central cloud servers, where the entire traffic load is concentrated, will lead to excessive end-to-end delay, energy consumption, and capacity limitations [114]. It has been shown in the study of [115] that computing resources have non-linear relation with user density. That is to say, computing resources will increase sharply when high density hotspots exist in the network. Moreover, applications are divided into small processes running in containers (e.g., VMs). The resource utilization in these containers is low and only a few of these containers are fully utilized which means that huge amounts of data processing capabilities and energy are lost [113].

The power requirement for medium-size data centers could be as high as 4 MW; therefore, supporting data centers with efficient power distribution mechanisms and deploying devices with lesser energy losses is critical to reduce that significant amount of power. Besides, the hardware reliability falls about 50 percent for each 10 °C increase in temperature over 21 °C. As a result, the implementation of cooling mechanisms which account for 38 percent the data centers power is inevitable to maintain a high quality performance [112]. It is also worth mentioning that idle servers consume as much as 50 percent of the full load power; therefore, switching the servers to sleep or off states is envisioned as a promising solution to save the excessive amounts of energy [116]. However, due to the unpredicted temporal variations in traffic loads along with the diversity of resource demands (e.g., CPU, memory, etc.) urges for sophisticated prediction techniques to reduce the time delay and energy losses. Moreover, the elasticity in cloud server geographical distribution could be exploited in assigning tasks to servers that exploit green energy resources. Hybrid cooling mechanism that uses both air and liquid to reduce the temperature of data centers is recognized as more energy and cost effective than utilizing air or liquid individually. With this cooling system server components with high power density such as CPUs and memory units are cooled down using water, whereas air is used to cool other auxiliary components [117]. Moreover, in [117] a joint optimization of server consolidation and inlet cooling water temperature was proposed to attain cost benefits. On the other hand, computing failures due to software or hardware crashes lead to significant energy waste to recover and repeat the computing tasks [118]. Therefore, failure analysis and diagnosis is essential to maintain energy-efficient clouds.

Network virtualization plays an essential role in VM consolidation, whereby multiple client requests can be executed by fewer servers and thus minimizing the number of active servers. The VM consolidation problem can be solved using the bin packing optimization technique which aims to pack VMs in fewer possible physical machines (bins). Another way to save energy in servers is the VMs migration, in which VMs from one physical machine are transferred to another machine in order to concentrate the load in fewer machines giving the opportunity to switch off machines that have light loads. Although obtaining accurate traffic prediction is complicated and impractical, approximation techniques can be utilized to achieve an adaptive and energy-efficient VM migration [119]. In addition, the actual resource requirement to execute an application is much lower than what is requested, for instance, only 55 percent of memory and 35 percent of CPU resources are actually exploited. To overcome the latter problem, resource over-commitment technique that assigns more VMs on a physical machine than the actual capacity of that machine to avoid resource under-utilization [120].

Information-centric networking (ICN) has been presented in [121] as a promising architecture to cope with the sharp increase in traffic load. In the ICN architecture, users care about how to obtain information regardless of where to get that information from. Moreover, the authors introduced multiple techniques to save energy such as switching off idle devices, adjusting the processor speed based on the quantity of received packets, and energy-efficient caching techniques. Moreover, nano data centers that distribute and host contents in a peer-to-peer fashion with end-users were introduced in [122] to reduce energy consumption and latency through the cooperative performance with central data centers. It was also shown that energy saving is guaranteed when nano data centers are used to run applications characterized by high data rates and low access requests such as video surveillance. To provide fair admittance to mobile devices, fairness-aware resource allocation mechanism in heterogeneous cloud computing systems was proposed in [123]. The work incorporated all heterogeneous servers and emphasized that a mobile device should not prefer other device resources for its own benefit; moreover, all users should report their resource demands truthfully.

In regard with mobile devices, the amount of energy consumption is dependant on the application workload, voltage and frequency configurations of the processor cores, and core scheduling mechanism. The computation energy can be minimized by adjusting the processor's clock frequency via the technology of dynamic voltage and frequency scaling whereby

the value of frequency is approximately linearly proportional to that of the voltage [124]. Furthermore, Hu and Wang [125] proposed a scheme for exploiting the under-utilized computational capabilities of mobile devices in the form of user-centric local mobile cloud without accessing the network servers, thus minimizing energy consumption and signaling overhead. The role of the network in this scheme is confined to the initialization and control provision for the local clouds, afterwards, mobile devices will take the responsibility of dividing the application into subtasks, and process each subtask in a local cloud that consists of multiple mobile devices sharing their computing capabilities. In the same context, an open source router platform with advanced power management strategies in [126] was successful in achieving 37 percent energy saving by optimally adjusting the forwarding operations according to the estimated traffic load. To this purpose, the clock frequency of active processor cores were tuned whereas the idle ones were put into sleep mode. Accordingly, the total system power consumption dropped from over 500 W at peak load hours (e.g., 3-8 pm) to 20 W at light load hours (e.g., 6 am).

From the cost and benefit perspective, the concept of carrier cloud emerged to enable mobile network operators accommodate the changing traffic load by requesting only adequate amount of resources (e.g., memory and CPU) to satisfy the requested services. Therefore, mobile operators need to pay only the per requested resources to the cloud providers, thus reducing the costs and increasing the profits. Furthermore, this scheme which is supported by the NFV and SDN technologies, allows scalability, elasticity, and fast network adaptability to the unpredicted traffic changes compared to the slow hardware modification in traditional mobile networks [127]. In a similar context, joint cloud service billing and device energy consumption was studied in [128] in regard with deriving the optimal idle-active operation in both device and cloud. The work was focused on queries initiated by smart phones for audio/visual recognition and classification in IoT services.

B. Task Offloading and VM Migration

The temporal and spatial variations of computing demands incur significant load unbalance in cloud servers. From this point, task offloading (or cyber foraging) and VM migration emerged as promising approaches to improve load balancing and hence computing and energy efficiencies.

1) Task Offloading: Originating from the limited computing capabilities of mobile devices, intensive computing tasks can be offloaded to close cloudlet servers in order to benefit from the powerful computing resources. However, the power limitation of the on-device batteries can degrade the overall offloading process. By comparing the evolution of integrated circuits in which the number of transistors doubles every two years according to Moore's law, with that of battery capacity which increases by only 5 percent every year, it can be realized that the energy gap between supply and demand increases by 4 percent every year. This gap will grow faster with the increasing popularity of smart phone applications [129].

To cope with this issue, several device-level offloading approaches have been developed. For instance, the MAUI approach in [130] aimed to maximize energy saving during runtime by continuously estimating the offloading cost regarding device energy, application requirements, network connectivity, bandwidth, and latency. Depending on the application type, MAUI was able to achieve at least 27 percent increase in energy saving by offloading to nearby servers with short round-trip time (RTT) such as 10 ms. Whereas, the energy consumed by offloading the same application to remote servers with RTT of 220 ms can be as high as twice that of offloading to nearby servers, and sometimes exceed the on-device processing energy due to the increase in communication overhead. However, MAUI needs the developer annotations to decide which partitions in the code to be offloaded. To this end, the CloneCloud in [131] focused on improving the flexibility in program partitioning by automatically choosing the code execution points; that is to say, enabling the device to select the program migration and re-integration points without requiring the developer annotations or even the source code. The work considered offloading via WiFi to nearby servers and via the cellular network to remote servers. It can be noticed that when the experiment was conducted on virus scanning application with 100 kB filesystem, the amounts of consumed energy using on-device, local servers (via WiFi), and remote servers (via the cellular network) were almost the same. However, when the filesystem size changed to 10 MB, offloading to local and remote servers reduced energy consumption by 12 and 5.6 times, respectively, compared with the on-device consumption. The reason behind this energy gain is that larger workloads benefit more from computing speed-up which allows faster task completion, thus reducing the computing and communication overheads.

Disregarding the intermittent nature of network connectivity might limit the benefits of previous approaches. Thus, implementing dynamic offloading paradigms is essential to maintain robust computing services. Such paradigms can be found in [132] and [133], where the former carried out a dynamic code repartitioning based on the runtime network and device status, whereas the latter considered the temporal variations in network capacity (i.e., bandwidth). Both approaches were experimented on real data traces and made significant enhancement to the computation offloading process regarding both energy saving and task completion time. Moreover, in [134], two types of clones (software) were proposed to support the task offloading in the cloud; the first was dedicated to support the computation offloading, and the second was intended to support the data backup to restore applications and users data. The study showed that the synchronization of the first clone incurred higher costs regarding both energy and network traffic because it occurs more frequently (e.g., system files) compared to the second clone which deals with big and long lasting-data (e.g., files saved by users). Furthermore, in [135] an algorithm was proposed to dynamically optimize the offloading decision, CPU frequency, and the offloading transmit power in energy harvesting (EH) mobile devices using Lyapunov optimization.

From the cellular network perspectives, inter-cell interference incurred at densely deployed areas can significantly degrade the offloading performance. To this end, Sardellitti *et al.* [67] formulated an optimization problem

that aims to minimize the total energy consumed by mobile devices, under the cloud-access delay constraint which involved both radio and computational resources. Furthermore, radio resource allocation schemes were proposed in [66] and [136] to optimize the task offloading performance in cloud networks. It is now obvious that cloudified cellular networks require smart offloading schemes that take into account the costs and rewards regarding task offloading, communication overhead, network conditions, and energy efficiency [137].

Processing tasks can also be carried out by other neighboring devices that have available resources to be shared. A network-assisted D2D task offloading was presented in [138], where devices collaborate in processing tasks while considering the incentives and fairness issues to avoid the under- or over-loaded device conditions. The objective of the problem, which was formulated as a Lyapunov optimization problem, was to minimize the averaged energy consumption of task execution for all users. Social relationships among users can also be exploited to establish a guaranteed and more efficient offloading process. Basically, this process is based on statistical measurements of call patterns, where users who have close social relationships tend to contact each other more frequently and spend shorter contact-time [139].

It is also important to notice that the energy consumed by a device to upload files is higher than to download. In addition, using MIMO technology by mobile phones incurs higher energy costs compared to single-input single-output (SISO) transmission. In this context, a case study was undertaken by [129] for task offloading in different types of smart phones for uploading a 23.97 MB file and downloading an 8.21 MB file with task offloading showed that using two 4G antennas required about 58 Joules which is almost double the energy consumed by SISO transmission that consumed 30 Joules. In addition, energy can be saved in wearable devices by offloading tasks to nearby smart phones or to the cloud. However, for small data size (e.g., 1 MB), processing tasks locally in the device could bring more energy saving than offloading to the cloud [140]. Furthermore, an adaptive algorithm presented in [141] was capable of minimizing the energy and cost of data uploading in mobile devices by approximately 50 and 60 percent, respectively. The algorithm divides users into data-plane and non-data plane users. The goal is to minimize energy consumption for the former group of users by using the most energy-efficient network available or by piggybacking data over phone calls. Whereas the idea for cost minimization in non-data plane users is offloading data from the cellular network to a local network such as WiFi or Bluetooth. Furthermore, the offloading decision strategy shown in [142] was capable of achieving 56 percent energy saving in the device battery by choosing executing tasks either in the device CPU, device GPU, or the cloud.

2) VM Migration: The motivation behind VM migration is to improve the provided QoS by running applications at close proximity with mobile devices [143]. The formation of VMs is initiated when a user requests offloading certain tasks to the cloud for processing. These VMs are placed in a physical server and managed by the VM manager. Whenever it is required, VMs can be transferred from one physical server to another, for single or multiple times during the task life-time to improve the computing capabilities, reduce the execution time of a task, or to achieve traffic balancing. Nevertheless, the VM migration process results in time delay due to the following reasons: 1) stopping the operating server, 2) transferring task-related data, and 3) initializing a new server. Therefore, efficient VM migration mechanisms must consider the incurred time delay for optimal performance. Several policies can be adopted to perform VM migration at different system levels. Cloud-wide migration, despite its complexity, aims at minimizing the execution time for the entire system. On the other hand, migration decisions can be initiated by the servers or even tasks themselves [144]. It is worth mentioning that software-defined networking and network function virtualization are key enabling technologies that allow a hardware platform to share multiple running tasks simultaneously, thus facilitating VM migration and improving resource utilization [145].

Several techniques regarding VM migration have been studied recently. In [146], dynamic VM migration was proposed using Bucket code learning and binary graph matching to divide VMs into groups and place these groups in particular servers. The work considered energy efficiency, migration cost, and VM to VM communication. In [147] VMs were named based on the service they provide instead of the traditional IP addressing. In such a way, routing will use the service name directly thus avoiding service interruption and enables decoupling services from the VM locations.

The migration process involves transferring all or part of the memory to the destination server. The pre-copy VM migration scheme captures and transfers the entire memory to the new server, whereas the post-copy scheme copies only minimum amount of memory and system state for the new server [143]. The pre-copy approach is more widely used but requires larger and variable bandwidth size; therefore, different algorithms have been adopted in [148] and [149] to reduce the bandwidth requirements and migration time for such a VM migration scheme.

An online VM scheduling scheme was proposed in [150] to optimally relocate VMs across data centers according to users mobility. The study on that scheme showed its ability to double the achievable throughput compared to data centers without such VM mobility feature. Moreover, a comparison between VM live migration in which only recent VM memory is transferred from source to destination to maintain synchronization, and VM bulk migration which transfers the entire VM memory for a mobility-aware network showed that the former approach is suitable for delay-sensitive services whereas the latter consumes much resources and considered as feasible only when the VM file size is small.

C. Cloudlet/Edge/Fog Computing

To achieve high computing agility, high bandwidth, and low latency, tasks can be offloaded from mobile devices to cloud servers located at close physical proximity. Bringing servers closer to mobile devices can also save energy due to the short-range transmission. This paradigm is known as mobile

TABLE VI
ENERGY-EFFICIENT CLOUD COMPUTING

Reference	Research Direction	Problem Type	Solution Approach
[119]	Power minimization in data centers using prediction-based VM migration. The work utilized Auto Regressive Integrated Moving Average (ARIMA)-based Kalman filter for estimation	Integer linear/quadrtic programming optimization	- Column generation, - The cut-and-solve-based algorithm and call back method were used to reduce complexity
[158]	Resource allocation and power consumption optimization for enhanced C-RANs (cloudlets integrated with C-RANs)	Non-cooperative matrix game	Nash equilibrium and KKT conditions
[159]	Energy-efficient multimedia data dissemination in vehicular clouds	Stochastic-reward-nets-based coalition game	Coalition game based on stochastic reward nets (SRNs)
[160]	Optimal offloading policy for intermittent connectivity in cloudlets	Markov decision process (MDP)-based optimization	The problem is expressed by Bellman's equation and solved using iterative algorithms

edge computing (MEC), and after comparing the energy cost of local computing and task offloading, a mobile device can decide whether to offload tasks to the MEC server or process them locally within the same device. For example, it was shown in [151] that the number of devices who decided to offload tasks decreased when the energy cost of processing tasks at the MEC server increased.

The study in [152] revealed that replicating popular contents to distributed clouds can save up to 43 percent of energy compared to centralized unaware clouds. Moreover, 48 percent of the energy can be saved when contents are migrated according to the access frequency of that particular content. Furthermore, slicing VMs and placing them in close physical proximity to mobile users can save 25 percent of the total energy. In addition, performing computations locally in cloudlets save the travel time through the backhaul into the central servers, and thus support low-latency applications [153]. Furthermore, the deployment of these mobile edge entities facilitates social information sharing while reducing transmission delay and the burdens on central servers [154]. Table VI summarizes some of the research work related to cloudlet computing.

Traffic routing in cellular networks is performed by the Serving Gateways (S-GWs) and Packet data Gateways (P-GWs). The S-GW devices are responsible for mobility anchoring whereas the P-GWs perform central control for traffic optimization, filtering, and firewalls. These P-GWs are located at the edge between Internet and the cellular network. However, having massive amounts of functionalities centered at the P-GWs and the network core will add undesired delay and traffic congestion. To improve network scalability and avoid the aforementioned challenges, the softcell concept was proposed in [155] such that access switches are distributed closer to the base stations to perform packet processing tasks, while keeping simple switches at the core network. Moreover, a SDN-based controller can be utilized at the P-GWs to perform central scheduling of tasks for cooperative offloading to save energy for mobile devices [156]. It is also worth mentioning that cloudlets could be placed either at the access points, aggregation nodes, or core nodes where the interconnection among these network elements follows different topologies such as mesh, ring, tree, or hybrid topologies [157].

A similar hierarchical management architecture is called fog-to-cloud (F2C). The fog computing concept falls inline with that of the cloudlet in being located at close proximity from mobile device to provide low-latency and high-bandwidth to support applications such as medical services and vehicle-to-vehicle (V2V) communications [161]. In vehicular clouds, for instance, a group of vehicles form a cooperative computing and communication entity that share information and resources. Vehicular clouds have been investigated in [162] for task migration, considering the cloud and cloudlets as cooperative service providers. Furthermore, Kumar *et al.* [159] formulated an energy-efficient multimedia data dissemination in a vehicular cloud environment by stochastic reward nets (SRNs)-based coalition game. Vehicles were considered as the players that organize their strategies based on the penalties and profits regarding frame delivery and latency that depend on the amount of computing demands and the available resources in the nearest cloudlet located on vehicles or at the road side.

The synergy between C-RANs and edge computing servers has been analytically presented in [163] as a promising approach to provide caching and processing capabilities which are controlled by the C-RAN. This network architecture is known as fog RAN (F-RAN). Furthermore, an enhanced C-RAN (EC-RAN) paradigm was presented in [158] to improve scalability and computation capabilities in 5G vehicular networks by integrating geographically distributed cloudlets with the central cloud to provide local services more efficiently. An algorithm for centralized task scheduling was proposed in [164] to manage a collaborative operation between mobile cloudlets and infrastructure-based cloudlets. The work considered the energy-efficient dynamic task offloading such that mobile cloudlets assist in performing task execution when the infrastructure-based cloudlet has insufficient resources.

Task scheduling involves multiple steps in order to achieve a better energy saving in MCC. These steps start by determining which tasks to be computed within mobile devices and which are offloaded to the cloud. Locally computed tasks are then mapped onto the processor cores and allocated suitable frequencies for optimal energy efficiency [165]. Markov decision process (MDP) was proposed in [160] to find an optimal offloading policy that avoids offloading failures and minimizes the computing and communication costs in cloudlets characterized with intermittent connectivity. Based on this policy, mobile users can decide between processing tasks locally or offloading them to the cloudlet depending on threshold parameters decided by the MDP policy. Furthermore, a policy for wireless energy transfer along with computation offloading

towards maximum energy saving has been proposed in [166]. The policy aims at deciding whether to offload computations to the cloud or to jointly compute within the same device and perform wireless power transfer along with controlling CPU cycles for optimal energy efficiency.

Rather than offloading the entire task, partial task offloading facilitates parallel computing and reduces the demand on wireless bandwidth [167] thus reducing energy consumption and time delay. To this end, a joint optimization problem has been presented regarding the smart device computation speed, transmit power, and offloading ratio under the constraints of energy consumption and application delay. To obtain further energy savings, an energy management policy for wireless energy charging has been presented in [168], where MDP was utilized to find the optimum policy for the cloudlet to decide whether or not to buy wireless energy when the battery level is below a certain threshold, and whether it is willing to accept offloaded tasks from mobile users.

V. Challenges and Open Issues

A. Energy-Efficient Joint H-CRAN and Edge Computing Deployment

The loosely-coupled characteristic of a fully cloudified mobile network infrastructure such as the one on Fig. 1 provides high flexibility, reliability, availability, and scalability. At the same time, it supplies network engineers with more network parameters and settings that might turn the system optimization into a daunting task.

Considering the massive deployment of RRHs and servers, the joint management of wireless and computing resources arises as one of the most challenging and promising ways to green the network infrastructure. In this sense, a joint activation/deactivation of base stations and task consolidation algorithm have the potential to significantly reduce the energy consumption on the whole cloudified mobile network infrastructure at the same way that they individually have done in their own realm. However, to unleash the energy saving potential behind this joint operation, the mobile network operator should create the appropriate technical and physical conditions to promote it. To expedite the communications between sites and therefore the VM migration between different cell sites, a high-capacity inter-site communication network should be put in place. Although this condition can be relaxed under the presence of a high-capacity fronthaul, its sharing between the ever-growing mobile traffic and VM migration traffic may result in performance degradation for both services. Particularly, for the edge computing, it may potentially compromise the live migration, which is the process of seamlessly moving a running VM between different physical machines. Thus, an inter-site communication network makes the traffic segmentation and isolation between both services viable.

In a fully interconnected air interface, two energy-aware server consolidation schemes might stand out: the local server consolidation and the network-wide consolidation. As the name implies, the local server consolidation is featured by a VM migration among servers within the same site while the network-wide server consolidation is realized between servers in different sites. From a sustainable operation perspective, the local server consolidation can exploit the short-term variations on the workload to consolidate servers in the same cell site and economize energy while the network-wide server consolidation can take advantage of long-term variations on the workload to accomplish a profound energy savings since the server consolidation might involve servers residing on multiple cell sites. To achieve further energy savings, local and network-wide server consolidation can be combined with base station activation and deactivation. For instance, under a light traffic situation, the base station may be switched off. In order to keep the connectivity between the UE and the mobile edge, a network-wide server consolidation can be performed moving the serving VM to an active server in a different cell site while handing the UE off to the corresponding base station. Similarly, under heavy traffic load where all base stations are active, local variations on the traffic load can be intelligently exploited to save energy by consolidating servers locally.

B. Energy-Aware Revenue Maximization

While the success of mobile cloud computing and IoT pave the way for the digital transformation, their massive needs for communication and computing resources might push mobile network operators towards the formation of coalitions with public cloud service providers in order to keep up the service provisioning without making major upfront investment.

The work in [169] presents a framework for revenue maximization in a coalition between a mobile network operator and public service providers considering a HetNet deployment and a multi-cloud system. The core aspect of the framework is the integer linear programming model that provides the maximum revenue for each coalition formed among the players while giving the optimal user association. By user association, it is meant an end-to-end mapping between the UE and a cloud data center through a base station. Next, the concept of Shapley value is applied to individualize the contribution of each player based on the maximum revenue for the optimal user association. Since the negative economical footprint of energy may considerably diminish the revenue, an energy-aware user association that binds UE to clouds through RRHs taking into account their joint on/off operation might potentially increase it while lowering the respective CO_2 emissions.

C. Cost- and Energy-Aware Cell Site Selection in Hybrid Power Supplied Deployment

While differences in energy prices due to the geographical diversity have been used as a criterion to reduce the total cost of distributed cloud data centers, the same criterion might become meaningless for the minimization of the total cost on edge computing given the need to process the application close to the UE. In order to minimize the energy bills and the dependence on fossil fuel-based energy, mobile network operators can invest on renewable energy. However, the intermittent nature of solar and wind generation exposes an inconvenient liability of a full green cell site deployment which makes a pure clean deployment questionable.

On the other hand, a hybrid power supplied cell site, where grid and renewable energy are symbiotically integrated, arises as a potential solution for the problem of cost and energy minimization on a fully cloudified mobile network. To get most out of this solution, mathematical tools such as Markov Decision Process and stochastic optimization stand out as prospective candidates to fine tune the use of grid and renewable source considering the randomness of supply and demand characteristics.

D. Energy-Efficient Data-Oriented Design

Irrefutably, the majority of works in literature assume a Poisson model as a representation for the data traffic despite the fact that data traffic is bursty and correlated in nature. It has been shown in [32] that the traffic autocorrelation has negligible effect on the total power consumption while the burstiness's contribution cannot be put aside. In fact, it was shown that the more bursty the traffic is, the less total power is consumed. When it comes to burstiness and autocorrelation, Markov-Modulated Poisson Process (MMPP) arises as an important stochastic process on the realm of traffic engineering by capturing these features while still addressing simplicity, tractability, versatility, and accuracy.

Considering the data-driven nature of IoT and mobile cloud computing, the orchestration of a joint on/off operation of H-CRAN and edge computing empowered by local and network-wide server consolidation under a MMPP data source might unfold the potential for energy savings in an fully cloudified mobile network infrastructure. This study is unprecedented in literature.

VI. Conclusion

In this article, we have provided an in-depth analysis of the research progress on energy-efficiency on H-CRAN and edge computing in order to gain actionable insights on the design of sustainable fully cloudified mobile network infrastructure. Energy efficiency concerns have been discussed based on each building block of the network infrastructure, which paved the way for a holistic analysis of the challenges and the open issues of a synergistic H-CRAN and edge computing operation.

With the unprecedented data deluge, motivated by the IoT and mobile cloud computing, mobile network operators will inevitably shift to an ultra dense deployment of virtualized wireless access points and servers which unless eco-friendly approached will considerably grow their energy footprint to unaffordable levels. In this sense, we believe that the design of an energy-aware cloudified mobile infrastructure arises as an important enabler for a more sustainable development.

References

[1] K. M. S. Huq *et al.*, "Green HetNet CoMP: Energy efficiency analysis and optimization," *IEEE Trans. Veh. Technol.*, vol. 64, no. 10, pp. 4670–4683, Oct. 2015.

[2] X. Sun and S. Wang, "Resource allocation scheme for energy saving in heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 8, pp. 4407–4416, Aug. 2015.

[3] R. Wang, H. Hu, and X. Yang, "Potentials and challenges of C-RAN supporting multi-RATs toward 5G mobile networks," *IEEE Access*, vol. 2, pp. 1187–1195, 2014.

[4] T. O. Olwal, K. Djouani, and A. M. Kurien, "A survey of resource management toward 5G radio access networks," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1656–1686, 3rd Quart., 2016.

[5] W. Wu, Q. Yang, B. Li, and K. S. Kwak, "Adaptive cross-layer resource optimization in heterogeneous wireless networks with multi-homing user equipments," *J. Commun. Netw.*, vol. 18, no. 5, pp. 784–795, Oct. 2016.

[6] C. Liu, B. Natarajan, and H. Xia, "Small cell base station sleep strategies for energy efficiency," *IEEE Trans. Veh. Technol.*, vol. 65, no. 3, pp. 1652–1661, Mar. 2016.

[7] J. Wu, Y. Zhang, M. Zukerman, and E. K.-N. Yung, "Energy-efficient base-stations sleep-mode techniques in green cellular networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 803–826, 2nd Quart., 2015.

[8] X. Guo, Z. Niu, S. Zhou, and P. R. Kumar, "Delay-constrained energy-optimal base station sleeping control," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1073–1085, May 2016.

[9] M. Kamel, W. Hamouda, and A. Youssef, "Ultra-dense networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 2522–2545, 4th Quart., 2016.

[10] B. Zhuang, D. Guo, and M. L. Honig, "Energy-efficient cell activation, user association, and spectrum allocation in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 823–831, Apr. 2016.

[11] J. B. Rao and A. O. Fapojuwo, "Analysis of spectrum efficiency and energy efficiency of heterogeneous wireless networks with intra-/inter-RAT offloading," *IEEE Trans. Veh. Technol.*, vol. 64, no. 7, pp. 3120–3139, Jul. 2015.

[12] J. Tang, D. K. C. So, E. Alsusa, K. A. Hamdi, and A. Shojaeifard, "Resource allocation for energy efficiency optimization in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, pp. 2104–2117, Oct. 2015.

[13] P. Cao, W. Liu, J. S. Thompson, C. Yang, and E. A. Jorswieck, "Semidynamic green resource management in downlink heterogeneous networks by group sparse power control," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1250–1266, May 2016.

[14] Y. Li, M. Sheng, Y. Sun, and Y. Shi, "Joint optimization of BS operation, user association, subcarrier assignment, and power allocation for energy-efficient HetNets," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3339–3353, Dec. 2016.

[15] Y.-H. Chiang and W. Liao, "Green multicell cooperation in heterogeneous networks with hybrid energy sources," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 7911–7925, Dec. 2016.

[16] Y. Kwon, T. Hwang, and X. Wang, "Energy-efficient transmit power control for multi-tier MIMO HetNets," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, pp. 2070–2086, Oct. 2015.

[17] H. Pervaiz, Z. Song, L. Musavian, Q. Ni, and X. Ge, "Throughput and backhaul energy efficiency analysis in two-tier HetNets: A multiobjective approach," in *Proc. IEEE Int. Workshop Comput. Aided Model. Design Commun. Links Netw. (CAMAD)*, Guildford, U.K., Sep. 2015, pp. 69–74.

[18] G. Yu, Y. Jiang, L. Xu, and G. Y. Li, "Multi-objective energy-efficient resource allocation for multi-RAT heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, pp. 2118–2127, Oct. 2015.

[19] E. Chavarria-Reyes, I. F. Akyildiz, and E. Fadel, "Energy consumption analysis and minimization in multi-layer heterogeneous wireless systems," *IEEE Trans. Mobile Comput.*, vol. 14, no. 12, pp. 2474–2487, Dec. 2015.

[20] W. Nie, F.-C. Zheng, X. Wang, W. Zhang, and S. Jin, "User-centric cross-tier base station clustering and cooperation in heterogeneous networks: Rate improvement and energy saving," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1192–1206, May 2016.

[21] M. Lin, S. Silvestri, N. Bartolini, and T. L. Porta, "On selective activation in dense femtocell networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 7018–7029, Oct. 2016.

[22] C. Jia and T. J. Lim, "Resource partitioning and user association with sleep-mode base stations in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 7, pp. 3780–3793, Jul. 2015.

[23] Y. Song, P.-Y. Kong, and Y. Han, "Potential of network energy saving through handover in HetNets," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 10198–10204, Dec. 2016.

[24] J. Kim, W. S. Jeon, and D. G. Jeong, "Effect of base station-sleeping ratio on energy efficiency in densely deployed femtocell networks," *IEEE Commun. Lett.*, vol. 19, no. 4, pp. 641–644, Apr. 2015.

[25] G. H. S. Carvalho, I. Woungang, A. Anpalagan, and E. Hossain, "QoS-aware energy-efficient joint radio resource management in multi-RAT heterogeneous networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6343–6365, Aug. 2016.

[26] S. Zhang *et al.*, "Energy-aware traffic offloading for green heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1116–1129, May 2016.

[27] J. Zheng, Y. Cai, X. Chen, R. Li, and H. Zhang, "Optimal base station sleeping in green cellular networks: A distributed cooperative framework based on game theory," *IEEE Trans. Wireless Commun.*, vol. 14, no. 8, pp. 4391–4406, Aug. 2015.

[28] J. Kim, W. S. Jeon, and D. G. Jeong, "Base-station sleep management in open-access femtocell networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 3786–3791, May 2016.

[29] L. Li, M. Peng, C. Yang, and Y. Wu, "Optimization of base-station density for high energy-efficient cellular networks with sleeping strategies," *IEEE Trans. Veh. Technol.*, vol. 65, no. 9, pp. 7501–7514, Sep. 2016.

[30] C.-Y. Chang, W. Liao, H.-Y. Hsieh, and D.-S. Shiu, "On optimal cell activation for coverage preservation in green cellular networks," *IEEE Trans. Mobile Comput.*, vol. 13, no. 11, pp. 2580–2591, Nov. 2014.

[31] P.-Y. Kong, "Optimal probabilistic policy for dynamic resource activation using Markov decision process in green wireless networks," *IEEE Trans. Mobile Comput.*, vol. 13, no. 10, pp. 2357–2368, Oct. 2014.

[32] J. Wu, Y. Bao, G. Miao, S. Zhou, and Z. Niu, "Base-station sleeping control and power matching for energy–delay tradeoffs with bursty traffic," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 3657–3675, May 2016.

[33] Z. Niu, X. Guo, S. Zhou, and P. R. Kumar, "Characterizing energy–delay tradeoff in hyper-cellular networks with base station sleeping control," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 4, pp. 641–650, Apr. 2015.

[34] X. Zhang *et al.*, "Macro-assisted data-only carrier for 5G green cellular systems," *IEEE Commun. Mag.*, vol. 53, no. 5, pp. 223–231, May 2015.

[35] M. Peng, C. Wang, V. Lau, and H. V. Poor, "Fronthaul-constrained cloud radio access networks: Insights and challenges," *IEEE Wireless Commun.*, vol. 22, no. 2, pp. 152–160, Apr. 2015.

[36] B. Niu and V. W. S. Wong, "Network configuration for two-tier macro–femto systems with hybrid access," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2528–2543, Apr. 2016.

[37] M. Adedoyin and O. Falowo, "Self-organizing radio resource management for next generation heterogeneous wireless networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[38] J. Gong, J. S. Thompson, S. Zhou, and Z. Niu, "Base station sleeping and resource allocation in renewable energy powered cellular networks," *IEEE Trans. Commun.*, vol. 62, no. 11, pp. 3801–3813, Nov. 2014.

[39] D. Liu, Y. Chen, K. K. Chai, T. Zhang, and M. Elkashlan, "Two-dimensional optimization on user association and green energy allocation for HetNets with hybrid energy sources," *IEEE Trans. Commun.*, vol. 63, no. 11, pp. 4111–4124, Nov. 2015.

[40] W. Nie, Y. Zhong, F.-C. Zheng, W. Zhang, and T. O'Farrell, "HetNets with random DTX scheme: Local delay and energy efficiency," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6601–6613, Aug. 2016.

[41] J. Liu, H. Guo, Z. M. Fadlullah, and N. Kato, "Energy consumption minimization for FiWi enhanced LTE-A HetNets with UE connection constraint," *IEEE Commun. Mag.*, vol. 54, no. 11, pp. 56–62, Nov. 2016.

[42] H. Klessig *et al.*, "From immune cells to self-organizing ultra-dense small cell networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 800–811, Apr. 2016.

[43] Z. Niu, Y. Wu, J. Gong, and Z. Yang, "Cell zooming for cost-efficient green cellular networks," *IEEE Commun. Mag.*, vol. 48, no. 11, pp. 74–79, Nov. 2010.

[44] H. Y. Lateef, M. Z. Shakir, M. Ismail, A. Mohamed, and K. Qaraqe, "Towards energy efficient and quality of service aware cell zooming in 5G wireless networks," in *Proc. IEEE 82nd Veh. Technol. Conf. (VTC)*, Boston, MA, USA, Sep. 2015, pp. 1–5.

[45] G. Wu, C. Yang, S. Li, and G. Y. Li, "Recent advances in energy-efficient networks and their application in 5G systems," *IEEE Wireless Commun.*, vol. 22, no. 2, pp. 145–151, Apr. 2015.

[46] M. Peng, Y. Li, Z. Zhao, and C. Wang, "System architecture and key technologies for 5G heterogeneous cloud radio access networks," *IEEE Netw.*, vol. 29, no. 2, pp. 6–14, Mar./Apr. 2015.

[47] A. Li, Y. Sun, X. Xu, and C. Yuan, "An energy-effective network deployment scheme for 5G cloud radio access networks," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, San Francisco, CA, USA, Apr. 2016, pp. 684–689.

[48] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, "Heterogeneous cloud radio access networks: A new perspective for enhancing spectral and energy efficiencies," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 126–135, Dec. 2014.

[49] D. B. Rawat and S. R. Reddy, "Software defined networking architecture, security and energy efficiency: A survey," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 1, pp. 325–346, 1st Quart., 2017.

[50] W. Zhong, R. Yu, S. Xie, Y. Zhang, and D. H. K. Tsang, "Software defined networking for flexible and green energy Internet," *IEEE Commun. Mag.*, vol. 54, no. 12, pp. 68–75, Dec. 2016.

[51] C. Fan, Y. J. Zhang, and X. Yuan, "Advances and challenges toward a scalable cloud radio access network," *IEEE Commun. Mag.*, vol. 54, no. 6, pp. 29–35, Jun. 2016.

[52] M. Peng, K. Zhang, J. Jiang, J. Wang, and W. Wang, "Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 11, pp. 5275–5287, Nov. 2015.

[53] A. Douik, H. Dahrouj, T. Al-Naffouri, and M.-S. Alouini, "Coordinated scheduling and power control in cloud-radio access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2523–2536, Apr. 2016.

[54] M. Peng, Y. Yu, H. Xiang, and H. V. Poor, "Energy-efficient resource allocation optimization for multimedia heterogeneous cloud radio access networks," *IEEE Trans. Multimedia*, vol. 18, no. 5, pp. 879–892, May 2016.

[55] S.-Y. Lien, S.-M. Cheng, K.-C. Chen, and D. Kim, "Resource-optimal licensed-assisted access in heterogeneous cloud radio access networks with heterogeneous carrier communications," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 9915–9930, Dec. 2016.

[56] S.-Y. Lien, S.-C. Hung, K.-C. Chen, and Y.-C. Liang, "Ultra-low-latency ubiquitous connections in heterogeneous cloud radio access networks," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 22–31, Jun. 2015.

[57] Y. Zhang, Y. Wang, and W. Zhang, "Energy efficient resource allocation for heterogeneous cloud radio access networks with user cooperation and QoS guarantees," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Doha, Qatar, Apr. 2016, pp. 1–6.

[58] Y. Zhang and Y. Wang, "A framework for energy efficient control in heterogeneous cloud radio access networks," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC Workshops)*, Chengdu, China, Jul. 2016, pp. 1–5.

[59] J. Li, M. Peng, Y. Yu, and Z. Ding, "Energy-efficient joint congestion control and resource optimization in heterogeneous cloud radio access networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 9873–9887, Dec. 2016.

[60] C. Ran and S. Wang, "Resource allocation in heterogeneous cloud radio access networks: A workload balancing perspective," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, Dec. 2015, pp. 1–6.

[61] Z. Yu, K. Wang, H. Ji, X. Li, and H. Zhang, "Dynamic resource allocation in TDD-based heterogeneous cloud radio access networks," *China Commun.*, vol. 13, no. 6, pp. 1–11, Jun. 2016.

[62] A. Li, Y. Sun, X. Xu, and C. Yuan, "Joint remote radio head selection and user association in cloud radio access networks," in *Proc. IEEE 27th Annu. Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, Valencia, Spain, Sep. 2016, pp. 1–6.

[63] P.-H. Huang, H. Kao, and W. Liao, "Hierarchical cooperation in heterogeneous cloud radio access networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[64] S. Kuang and N. Liu, "Energy minimization via BS selection and beamforming for cloud-RAN under finite fronthaul capacity constraints," in *Proc. IEEE 83rd Veh. Technol. Conf.*, Nanjing, China, May 2016, pp. 1–6.

[65] V. N. Ha and L. B. Le, "Joint coordinated beamforming and admission control for fronthaul constrained cloud-RANs," in *Proc. IEEE Glob. Commun. Conf.*, Austin, TX, USA, Dec. 2014, pp. 4054–4059.

[66] Y. Cao, T. Jiang, and C. Wang, "Optimal radio resource allocation for mobile task offloading in cellular networks," *IEEE Netw.*, vol. 28, no. 5, pp. 68–73, Sep./Oct. 2014.

[67] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile-edge computing," *IEEE Trans. Signal Inf. Process. Over Netw.*, vol. 1, no. 2, pp. 89–103, Jun. 2015.

[68] D. Mazza, D. Tarchi, and G. E. Corazza, "A unified urban mobile cloud computing offloading mechanism for smart cities," *IEEE Commun. Mag.*, vol. 55, no. 3, pp. 30–37, Mar. 2017.

[69] D. Pompili, A. Hajisami, and T. X. Tran, "Elastic resource utilization framework for high capacity and energy efficiency in cloud RAN," *IEEE Commun. Mag.*, vol. 54, no. 1, pp. 26–32, Jan. 2016.

[70] M. A. Marotta *et al.*, "Resource sharing in heterogeneous cloud radio access networks," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 74–82, Jun. 2015.

[71] H. Dahrouj, A. Douik, O. Dhifallah, T. Y. Al-Naffouri, and M.-S. Alouini, "Resource allocation in heterogeneous cloud radio access networks: Advances and challenges," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 66–73, Jun. 2015.

[72] D. Mishra, P. C. Amogh, A. Ramamurthy, A. A. Franklin, and B. R. Tamma, "Load-aware dynamic RRH assignment in cloud radio access networks," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Doha, Qatar, Apr. 2016, pp. 1–6.

[73] K. Wang, M. Zhao, and W. Zhou, "Traffic-aware graph-based dynamic frequency reuse for heterogeneous cloud-RAN," in *Proc. IEEE Glob. Commun. Conf.*, Austin, TX, USA, Dec. 2014, pp. 2308–2313.

[74] M. Wang, H. Xia, and C. Feng, "Joint dynamic point blanking and ABS for ICIC in cloud cooperated heterogeneous network," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Shenzhen, China, Nov. 2015, pp. 1–5.

[75] B. Niu, Y. Zhou, H. Shah-Mansouri, and V. W. S. Wong, "A dynamic resource sharing mechanism for cloud radio access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8325–8338, Dec. 2016.

[76] M. Peng, X. Xie, Q. Hu, J. Zhang, and H. V. Poor, "Contract-based interference coordination in heterogeneous cloud radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1140–1153, Jun. 2015.

[77] M. Peng, H. Xiang, Y. Cheng, S. Yan, and H. V. Poor, "Inter-tier interference suppression in heterogeneous cloud radio access networks," *IEEE Access*, vol. 3, pp. 2441–2455, 2015.

[78] A. Abdelnasser and E. Hossain, "Resource allocation for an OFDMA cloud-RAN of Small cells underlaying a macrocell," *IEEE Trans. Mobile Comput.*, vol. 15, no. 11, pp. 2837–2850, Nov. 2016.

[79] N. Abuzainab and W. Saad, "Cloud radio access meets heterogeneous small cell networks: A cognitive hierarchy perspective," in *Proc. IEEE 17th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Edinburgh, U.K., Jul. 2016, pp. 1–5.

[80] X. Huang, G. Xue, R. Yu, and S. Leng, "Joint scheduling and beamforming coordination in cloud radio access networks with QoS guarantees," *IEEE Trans. Veh. Technol.*, vol. 65, no. 7, pp. 5449–5460, Jul. 2016.

[81] K. Zhang, M. Peng, C. Wang, and S. Yan, "Perron–Frobenius theory based power allocation in heterogeneous cloud radio access networks," in *Proc. IEEE 82nd Veh. Technol. Conf. (VTC)*, Boston, MA, USA, Sep. 2015, pp. 1–5.

[82] S. Luo, R. Zhang, and T. J. Lim, "Downlink and uplink energy minimization through user association and beamforming in C-RAN," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 494–508, Jan. 2015.

[83] K. A. Meerja, A. Shami, and A. Refaey, "Hailing cloud empowered radio access networks," *IEEE Wireless Commun.*, vol. 22, no. 1, pp. 122–129, Feb. 2015.

[84] H. Zhang, C. Jiang, J. Cheng, and V. C. M. Leung, "Cooperative interference mitigation and handover management for heterogeneous cloud small cell networks," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 92–99, Jun. 2015.

[85] W.-T. Lin, C.-H. Lee, and H.-J. Su, "Downlink-to-uplink interference cancellation in cloud radio access networks," in *Proc. IEEE 79th Veh. Technol. Conf.*, Seoul, South Korea, May 2014, pp. 1–5.

[86] O. Dhifallah, H. Dahrouj, T. Y. Al-Naffouri, and M.-S. Alouini, "Decentralized group sparse beamforming for multi-cloud radio access networks," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, San Diego, CA, USA, Dec. 2015, pp. 1–6.

[87] A. Adhikary, H. S. Dhillon, and G. Caire, "Massive-MIMO meets HetNet: Interference coordination through spatial blanking," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 6, pp. 1171–1186, Jun. 2015.

[88] M. A. Abana, M. Peng, Z. Zhao, and L. A. Olawoyin, "Coverage and rate analysis in heterogeneous cloud radio access networks with device-to-device communication," *IEEE Access*, vol. 4, pp. 2357–2370, 2016.

[89] R. Brandt, R. Mochaourab, and M. Bengtsson, "Distributed long-term base station clustering in cellular networks using coalition formation," *IEEE Trans. Signal Inf. Process. Over Netw.*, vol. 2, no. 3, pp. 362–375, Sep. 2016.

[90] L. Zhang *et al.*, "Performance analysis and optimal cooperative cluster size for randomly distributed small cells under cloud RAN," *IEEE Access*, vol. 4, pp. 1925–1939, 2016.

[91] M. M. U. Rahman, H. Ghauch, S. Imtiaz, and J. Gross, "RRH clustering and transmit precoding for interference-limited 5G CRAN downlink," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, San Diego, CA, USA, Dec. 2015, pp. 1–7.

[92] B. Dai and W. Yu, "Energy efficiency of downlink transmission strategies for cloud radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 1037–1050, Apr. 2016.

[93] A. Hajisami and D. Pompili, "DJP: Dynamic joint processing for interference cancellation in cloud radio access networks," in *Proc. IEEE 82nd Veh. Technol. Conf.*, Boston, MA, USA, Sep. 2015, pp. 1–5.

[94] R. Brandt, R. Mochaourab, and M. Bengtsson, "Globally optimal base station clustering in interference alignment-based multicell networks," *IEEE Signal Process. Lett.*, vol. 23, no. 4, pp. 512–516, Apr. 2016.

[95] E. Chen and M. Tao, "User-centric base station clustering and sparse beamforming for cache-enabled cloud RAN," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Shenzhen, China, Nov. 2015, pp. 1–6.

[96] H. M. Soliman and A. Leon-Garcia, "QoS-aware joint RRH activation and clustering in cloud-RANs," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Doha, Qatar, Apr. 2016, pp. 1–6.

[97] L. Lei, D. Yuan, C. K. Ho, and S. Sun, "Optimal cell clustering and activation for energy saving in load-coupled wireless networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6150–6163, Nov. 2015.

[98] C. Ran, S. Wang, and C. Wang, "Balancing backhaul load in heterogeneous cloud radio access networks," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 42–48, Jun. 2015.

[99] H. H. Yang, G. Geraci, and T. Q. S. Quek, "Energy-efficient design of MIMO heterogeneous networks with wireless backhaul," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4914–4927, Jul. 2016.

[100] H. H. Yang, G. Geraci, and T. Q. S. Quek, "MIMO HetNets with wireless backhaul: An energy-efficient design," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[101] T. M. Nguyen, A. Yadav, W. Ajib, and C. Assi, "Achieving energy-efficiency in two-tiers wireless backhaul HetNets," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[102] D. Zhang, M. Chen, M. Guizani, H. Xiong, and D. Zhang, "Mobility prediction in telecom cloud using mobile calls," *IEEE Wireless Commun.*, vol. 21, no. 1, pp. 26–32, Feb. 2014.

[103] C. Yang, Z. Chen, B. Xia, and J. Wang, "When ICN meets C-RAN for HetNets: An SDN approach," *IEEE Commun. Mag.*, vol. 53, no. 11, pp. 118–125, Nov. 2015.

[104] M. Jaber, M. A. Imran, R. Tafazolli, and A. Tukmanov, "5G backhaul challenges and emerging research directions: A survey," *IEEE Access*, vol. 4, pp. 1743–1766, 2016.

[105] Y. Zhou and W. Yu, "Optimized backhaul compression for uplink cloud radio access network," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1295–1307, Jun. 2014.

[106] Y. Zhou and W. Yu, "Fronthaul compression and transmit beamforming optimization for multi-antenna uplink C-RAN," *IEEE Trans. Signal Process.*, vol. 64, no. 16, pp. 4138–4151, Aug. 2016.

[107] O. Simeone, A. Maeder, M. Peng, O. Sahin, and W. Yu, "Cloud radio access network: Virtualizing wireless access for dense heterogeneous systems," *J. Commun. Netw.*, vol. 18, no. 2, pp. 135–149, Apr. 2016.

[108] S. Park, C.-B. Chae, and S. Bahk, "Large-scale antenna operation in heterogeneous cloud radio access networks: A partial centralization approach," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 32–40, Jun. 2015.

[109] K. Sundaresan, M. Y. Arslan, S. Singh, S. Rangarajan, and S. V. Krishnamurthy, "FluidNet: A flexible cloud-based radio access network for small cells," *IEEE/ACM Trans. Netw.*, vol. 24, no. 2, pp. 915–928, Apr. 2016.

[110] H. Ma, B. Wang, Y. Chen, and K. J. R. Liu, "Time-reversal tunneling effects for cloud radio access network," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 3030–3043, Apr. 2016.

[111] X. Zhou *et al.*, "Toward 5G: When explosive bursts meet soft cloud," *IEEE Netw.*, vol. 28, no. 6, pp. 12–17, Nov./Dec. 2014.

[112] J. Shuja *et al.*, "Survey of techniques and architectures for designing energy-efficient data centers," *IEEE Syst. J.*, vol. 10, no. 2, pp. 507–519, Jun. 2016.

[113] Q. Zhang and R. Boutaba, "Dynamic workload management in heterogeneous cloud computing environments," in *Proc. IEEE Netw. Oper. Manag. Symp. (NOMS)*, Kraków, Poland, May 2014, pp. 1–7.

[114] M. Barcelo *et al.*, "IoT-cloud service optimization in next generation smart environments," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 4077–4090, Dec. 2016.

[115] Y. Liao, L. Song, Y. Li, and Y. A. Zhang, "How much computing capability is enough to run a cloud radio access network?" *IEEE Commun. Lett.*, vol. 21, no. 1, pp. 104–107, Jan. 2017.

[116] T. Mastelic and I. Brandic, "Recent trends in energy-efficient cloud computing," *IEEE Cloud Comput.*, vol. 2, no. 1, pp. 40–47, Feb. 2015.

[117] S. Chen, S. Irving, and L. Peng, "Operational cost optimization for cloud computing data centers using renewable energy," *IEEE Syst. J.*, vol. 10, no. 4, pp. 1447–1458, Dec. 2016.

[118] P. Garraghan, I. S. Moreno, P. Townend, and J. Xu, "An analysis of failure-related energy waste in a large-scale cloud environment," *IEEE Trans. Emerg. Topics Comput.*, vol. 2, no. 2, pp. 166–180, Jun. 2014.

[119] S. Vakilinia, B. Heidarpour, and M. Cheriet, "Energy efficient resource allocation in cloud computing environments," *IEEE Access*, vol. 4, pp. 8544–8557, 2016.

[120] M. Dabbagh, B. Hamdaoui, M. Guizani, and A. Rayes, "Toward energy-efficient cloud computing: Prediction, consolidation, and overcommitment," *IEEE Netw.*, vol. 29, no. 2, pp. 56–61, Mar./Apr. 2015.

[121] C. Fang, F. R. Yu, T. Huang, J. Liu, and Y. Liu, "A survey of green information-centric networking: Research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 3, pp. 1455–1472, 3rd Quart., 2015.

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

[123] W. Wang, B. Liang, and B. Li, "Multi-resource fair allocation in heterogeneous cloud computing systems," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 10, pp. 2822–2835, Oct. 2015.

[124] C. Luo, L. T. Yang, P. Li, X. Xie, and H.-C. Chao, "A holistic energy optimization framework for cloud-assisted mobile computing," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 118–123, Jun. 2015.

[125] H. Hu and R. Wang, "User-centric local mobile cloud-assisted D2D communications in heterogeneous cloud-RANS," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 59–65, Jun. 2015.

[126] R. Bolla, C. Lombardo, R. Bruschi, and S. Mangialardi, "DROPv2: Energy efficiency through network function virtualization," *IEEE Netw.*, vol. 28, no. 2, pp. 26–32, Mar./Apr. 2014.

[127] T. Taleb, "Toward carrier cloud: Potential, challenges, and solutions," *IEEE Wireless Commun.*, vol. 21, no. 3, pp. 80–91, Jun. 2014.

[128] F. Renna, J. Doyle, V. Giotsas, and Y. Andreopoulos, "Media query processing for the Internet-of-Things: Coupling of device energy consumption and cloud infrastructure billing," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2537–2552, Dec. 2016.

[129] M. Altamimi, A. Abdrabou, K. Naik, and A. Nayak, "Energy cost models of smartphones for task offloading to the cloud," *IEEE Trans. Emerg. Topics Comput.*, vol. 3, no. 3, pp. 384–398, Sep. 2015.

[130] E. Cuervo *et al.*, "MAUI: Making smartphones last longer with code offload," in *Proc. ACM MobiSys*, San Francisco, CA, USA, Jun. 2010, pp. 49–62.

[131] B. Chun, S. Ihm, P. Maniatis, M. Naik, and A. Patti, "CloneCloud: Elastic execution between mobile device and cloud," in *Proc. 6th Conf. Comput. Syst.*, Salzburg, Austria, Apr. 2011, pp. 301–314.

[132] L. Yang, J. Cao, S. Tang, D. Han, and N. Suri, "Run time application repartitioning in dynamic mobile cloud environments," *IEEE Trans. Cloud Comput.*, vol. 4, no. 3, pp. 336–348, Jul./Sep. 2016.

[133] J. L. D. Neto *et al.*, "ULOOF: A user level online offloading framework for mobile edge computing," Working Paper, Jun. 2017. [Online]. Available: http://hal.upmc.fr/hal-01547036

[134] M. V. Barbera, S. Kosta, A. Mei, and J. Stefa, "To offload or not to offload? The bandwidth and energy costs of mobile cloud computing," in *Proc. IEEE INFOCOM*, Turin, Italy, Apr. 2013, pp. 1285–1293.

[135] Y. Mao, J. Zhang, and K. B. Letaief, "Dynamic computation offloading for mobile-edge computing with energy harvesting devices," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3590–3605, Dec. 2016.

[136] Y. Zhao, S. Zhou, T. Zhao, and Z. Niu, "Energy-efficient task offloading for multiuser mobile cloud computing," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Shenzhen, China, Nov. 2015, pp. 1–5.

[137] L. Zhang, D. Fu, J. Liu, E. C.-H. Ngai, and W. Zhu, "On energy-efficient offloading in mobile cloud for real-time video applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 1, pp. 170–181, Jan. 2017.

[138] L. Pu, X. Chen, J. Xu, and X. Fu, "D2D fogging: An energy-efficient and incentive-aware task offloading framework via network-assisted D2D collaboration," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3887–3901, Dec. 2016.

[139] P. Yang *et al.*,"'Friend is treasure': Exploring and exploiting mobile social contacts for efficient task offloading," *IEEE Trans. Veh. Technol.*, vol. 65, no. 7, pp. 5485–5496, Jul. 2016.

[140] C. Ragona, F. Granelli, C. Fiandrino, D. Kliazovich, and P. Bouvry, "Energy-efficient computation offloading for wearable devices and smartphones in mobile cloud computing," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, San Diego, CA, USA, Dec. 2015, pp. 1–6.

[141] L. Wang, D. Zhang, Z. Yan, H. Xiong, and B. Xie, "effSense: A novel mobile crowd-sensing framework for energy-efficient and cost-effective data uploading," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 45, no. 12, pp. 1549–1563, Dec. 2015.

[142] Y.-D. Lin, E. T.-H. Chu, Y.-C. Lai, and T.-J. Huang, "Time-and-energy-aware computation offloading in handheld devices to coprocessors and clouds," *IEEE Syst. J.*, vol. 9, no. 2, pp. 393–405, Jun. 2015.

[143] B. Hu, S. Chen, J. Chen, and Z. Hu, "A mobility-oriented scheme for virtual machine migration in cloud data center network," *IEEE Access*, vol. 4, pp. 8327–8337, 2016.

[144] L. Gkatzikis and I. Koutsopoulos, "Migrate or not? Exploiting dynamic task migration in mobile cloud computing systems," *IEEE Wireless Commun.*, vol. 20, no. 3, pp. 24–32, Jun. 2013.

[145] K. Xu, C. Lin, Z. Chen, K. Meng, and M. Hakmaoui, "An effective policy relocation scheme for VM migration in software-defined networks," in *Proc. Int. Conf. Comput. Commun. Netw. (ICCCN)*, Las Vegas, NV, USA, Aug. 2015, pp. 1–8.

[146] F. Tao, C. Li, T. W. Liao, and Y. Laili, "BGM-BLA: A new algorithm for dynamic migration of virtual machines in cloud computing," *IEEE Trans. Services Comput.*, vol. 9, no. 6, pp. 910–925, Nov./Dec. 2016.

[147] R. Xie, Y. Wen, X. Jia, and H. Xie, "Supporting seamless virtual machine migration via named data networking in cloud data center," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 12, pp. 3485–3497, Dec. 2015.

[148] Y. Ruan, Z. Cao, and Z. Cui, "Pre-filter-copy: Efficient and self-adaptive live migration of virtual machines," *IEEE Syst. J.*, vol. 10, no. 4, pp. 1459–1469, Dec. 2016.

[149] J. Zhang, F. Ren, R. Shu, T. Huang, and Y. Liu, "Guaranteeing delay of live virtual machine migration by determining and provisioning appropriate bandwidth," *IEEE Trans. Comput.*, vol. 65, no. 9, pp. 2910–2917, Sep. 2016.

[150] S. Secci, P. Raad, and P. Gallard, "Linking virtual machine mobility to user mobility," *IEEE Trans. Netw. Service Manag.*, vol. 13, no. 4, pp. 927–940, Dec. 2016.

[151] K. Zhang *et al.*, "Energy-efficient offloading for mobile edge computing in 5G heterogeneous networks," *IEEE Access*, vol. 4, pp. 5896–5907, 2016.

[152] A. Q. Lawey, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Distributed energy efficient clouds over core networks," *J. Lightw. Technol.*, vol. 32, no. 7, pp. 1261–1281, Apr. 1, 2014.

[153] S. Barbarossa, S. Sardellitti, and P. D. Lorenzo, "Communicating while computing: Distributed mobile cloud computing over 5G heterogeneous networks," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 45–55, Nov. 2014.

[154] S.-Y. Lien, S.-C. Hung, H. Hsu, and K.-C. Chen, "Collaborative radio access of heterogeneous cloud radio access networks and edge computing networks," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 193–199.

[155] X. Jiny, L. E. Li, L. Vanbevery, and J. Rexfordy, "SoftCell: Scalable and flexible cellular core network architecture," in *Proc. 9th ACM Conf. Emerg. Netw. Exp. Technol. (CoNEXT)*, Santa Barbara, CA, USA, Dec. 2013, pp. 163–174.

[156] Y. Cui *et al.*, "Software defined cooperative offloading for mobile cloudlets," *IEEE/ACM Trans. Netw.*, vol. 25, no. 3, pp. 1746–1760, Jun. 2017.

[157] A. Ceselli, M. Premoli, and S. Secci, "Mobile edge cloud network design optimization," *IEEE/ACM Trans. Netw.*, vol. 25, no. 3, pp. 1818–1831, Jun. 2017.

[158] R. Yu *et al.*, "Optimal resource sharing in 5G-enabled vehicular networks: A matrix game approach," *IEEE Trans. Veh. Technol.*, vol. 65, no. 10, pp. 7844–7856, Oct. 2016.

[159] N. Kumar, J. Lee, N. Chilamkurti, and A. Vinel, "Energy-efficient multimedia data dissemination in vehicular clouds: Stochastic-reward-nets-based coalition game approach," *IEEE Syst. J.*, vol. 10, no. 2, pp. 847–858, Jun. 2016.

[160] Y. Zhang, D. Niyato, and P. Wang, "Offloading in mobile cloudlet systems with intermittent connectivity," *IEEE Trans. Mobile Comput.*, vol. 14, no. 12, pp. 2516–2529, Dec. 2015.

[161] X. Masip-Bruin, E. Marín-Tordera, G. Tashakor, A. Jukan, and G.-J. Ren, "Foggy clouds and cloudy fogs: A real need for coordinated management of fog-to-cloud computing systems," *IEEE Wireless Commun.*, vol. 23, no. 5, pp. 120–128, Oct. 2016.

[162] H. Zhang, Q. Zhang, and X. Du, "Toward vehicle-assisted cloud computing for smartphones," *IEEE Trans. Veh. Technol.*, vol. 64, no. 12, pp. 5610–5618, Dec. 2015.
[163] R. Tandon and O. Simeone, "Harnessing cloud and edge synergies: Toward an information theory of fog radio access networks," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 44–50, Aug. 2016.
[164] Z. Wu, L. Gui, J. Chen, H. Zhou, and F. Hou, "Mobile cloudlet assisted computation offloading in heterogeneous mobile cloud," in *Proc. Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Yangzhou, China, Oct. 2016, pp. 1–6.
[165] X. Lin, Y. Wang, Q. Xie, and M. Pedram, "Task scheduling with dynamic voltage and frequency scaling for energy minimization in the mobile cloud computing environment," *IEEE Trans. Services Comput.*, vol. 8, no. 2, pp. 175–186, Mar./Apr. 2015.
[166] C. You, K. Huang, and H. Chae, "Energy efficient mobile cloud computing powered by wireless energy transfer," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1757–1771, May 2016.
[167] Y. Wang, M. Sheng, X. Wang, L. Wang, and J. Li, "Mobile-edge computing: Partial computation offloading using dynamic voltage scaling," *IEEE Trans. Commun.*, vol. 64, no. 10, pp. 4268–4282, Oct. 2016.
[168] D. Niyato, P. Wang, P. C. H. Joo, Z. Han, and D. I. Kim, "Optimal energy management policy of a mobile cloudlet with wireless energy charging," in *Proc. IEEE Int. Conf. Smart Grid Commun. (SmartGridComm)*, Venice, Italy, Nov. 2014, pp. 728–733.
[169] G. H. S. Carvalho, I. Woungang, A. Anpalagan, M. Jaseemuddin and E. Hossain, "Intercloud and HetNet for mobile cloud computing in 5G systems: Design issues, challenges, and optimization," *IEEE Netw.*, vol. 31, no. 3, pp. 80–89, May/Jun. 2017.

Ali Alnoman received the B.Sc. and M.Sc. degrees in electrical engineering from the University of Baghdad, Iraq, in 2009 and 2012, respectively. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Ryerson University, Canada. From 2012 to 2014, he was a Faculty Member with Ishik University, Erbil, Iraq. His research interests include energy efficiency and resource allocation in HetNets and cloud computing. He also served as a Technical Program Committee Member in the IEEE Vehicular Technology Conference VTC2017-Fall, in Toronto.

Glaucio H. S. Carvalho received the Ph.D. degree in electrical engineering from the Federal University of Para (UFPA), Brazil, in 2005. He is currently pursuing the second Ph.D. degree with the Department of Computer Science, Ryerson University with emphasis on Cybersecurity. He was a Professor with the Department of Computer Science, UFPA from 2005 to 2015. He worked twice as a Post-Doctoral Fellow with the Department of Computer Science, Ryerson University doing research on the field of cloud systems, distributed systems, and networks. He has served as the Chair for the IEEE Toronto Section Signals & Computational Intelligence Joint Society. His research interests include security and performance analysis of cloud systems, distributed systems, and networks.

Alagan Anpalagan (S'98–M'01–SM'04) received the B.A.Sc., M.A.Sc., and Ph.D. degrees in electrical engineering from the University of Toronto, Canada. He joined the ELCE Department, Ryerson University, Canada, in 2001, where he was promoted to a Full Professor in 2010 and served as a Graduate Program Director from 2004 to 2009 and the Interim Electrical Engineering Program Director from 2009 to 2010. During his sabbatical (2010–2011), he was a Visiting Professor with the Asian Institute of Technology, Thailand, and a Visiting Researcher with Kyoto University, Japan. His industrial experience includes working for three years with Bell Mobility, Nortel Networks, and IBM. He directs a research group working on radio resource management (RRM) and radio access and networking areas within the WINCORE Laboratory. He also completed a course on Project Management for Scientist and Engineers with the University of Oxford CPD Center, Oxford, U.K. He has co-authored three edited books entitled *Design and Deployment of Small Cell Networks* (Cambridge University Press, 2016), *Routing in Opportunistic Networks* (Springer, 2013), and *Handbook on Green Information and Communication Systems* (Academic Press, 2012) and a book entitled *Game-Theoretic Interference Coordination Approaches for Dynamic Spectrum Access* (Springer, 2016). His research interests include 5G wireless systems, energy harvesting and green communications technologies, cognitive RRM, wireless cross layer design and optimization, cooperative communication, M2M and sensor communication, small cell, and heterogeneous networks. He was a recipient of the Deans Teaching Award in 2011, the Faculty Scholastic, Research and Creativity Award in 2010 and 2014, the Faculty Service Award from Ryerson University in 2011 and 2013, the IEEE M.B. Broughton Central Canada Service Award in 2016, the Exemplary Editor Award from IEEE ComSoc in 2013, the Editor-in-Chief Top10 Choice Award in Transactions on Emerging Telecommunications Technology in 2012, and the IEEE SPS Young Author Best Paper Award in 2015 for his co-authored paper. He served as an Editor for the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS from 2012 to 2014, the IEEE COMMUNICATIONS LETTERS from 2010 to 2013, and the *EURASIP Journal of Wireless Communications and Networking* from 2004 to 2009. He also served as a Guest Editor for six special issues the IEEE WIRELESS COMMUNICATIONS on Sustainable Green Networking and Computing in 5G Systems, the IEEE ACCESS on Internet of Things in 5G Systems, *IET Communications* on Evolution and Development of 5G Wireless Communication Systems, *EURASIP* on Radio Resource Management in 3G+ Systems and on Fairness in Radio Resource Management for Wireless Networks, and *MONET* on Green Cognitive and Cooperative Communication and Networking. He served as the TPC Co-Chair for IEEE VTC Fall 2017, TPC Co-Chair for IEEE INFOCOM'16: First International Workshop on Green and Sustainable Networking and Computing, IEEE Globecom15: SAC Green Communication and Computing, IEEE PIMRC11: Cognitive Radio and Spectrum Management. He served as the Vice Chair for IEEE SIG on Green and Sustainable Networking and Computing With Cognition and Cooperation from 2015 to 2016, the IEEE Canada Central Area Chair from 2012 to 2014, the IEEE Toronto Section Chair from 2006 to 2007, the ComSoc Toronto Chapter Chair from 2004 to 2005, and the IEEE Canada Professional Activities Committee Chair from 2009 to 2011. He is a Registered Professional Engineer in the province of Ontario, Canada and a fellow of the Institution of Engineering and Technology.

Isaac Woungang received the M.Sc. degree in mathematics from the University of Aix Marseille II, France, in 1990, the Ph.D. degree in mathematics from the University of South, Toulon, France, in 1994, and the M.Sc. degree from INRS-Materials and Telecommunications, University of Quebec in Montreal, QC, Canada, in 1999. From 1999 to 2002, he was a Software Engineer with Nortel Networks, Ottawa, Canada. Since 2002, he has been with Ryerson University, Toronto, Canada, where he is currently a Professor of computer science and the Director of the Distributed Applications and Broadband Research Laboratory. His current research interests include radio resource management in next generation wireless networks and network security. He has published eight books and over 90 refereed technical articles in scholarly international journals and proceedings of international conferences. He served as an Associate Editor for the *Computers and Electrical Engineering* (Elsevier) and the *International Journal of Communication Systems* (Wiley). He has guest edited several special issues with various reputed journals, such as *IET Information Security*, *Computer Communications* (Elsevier), *Computers and Electrical Engineering* (Elsevier), and *Telecommunication Systems* (Springer). He served as the Chair of Computer Chapter, IEEE Toronto Section from 2012 to 2016.