ENERGY SAVING SCHEMES FOR SCALABLE MOBILE COMPUTING NETWORKS

by

Ali Alnoman

Master of Science in Electrical Engineering, University of Baghdad, 2012

Bachelor of Science in Electrical Engineering, University of Baghdad, 2009

A dissertation

presented to Ryerson University

in partial fulfilment of the

requirements for the degree of

Doctor of Philosophy

in the program of

Electrical and Computer Engineering

Toronto, Ontario, Canada, 2019

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A DISSERTATION

I hereby declare that I am the sole author of this dissertation. This is a true copy of the dissertation, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this dissertation to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this dissertation by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my dissertation may be made electronically available to the public.

Abstract

Energy Saving Schemes for Scalable Mobile Computing Networks

Ali Alnoman, 2019

Doctor of Philosophy

Electrical and Computer Engineering

Ryerson University, Toronto, Canada

With the growing popularity of smart applications that contain computing-intensive tasks, the provision of radio and computing resources with high quality is becoming more and more challenging. Moreover, supporting network scalability is crucial to accommodate the massive numbers of connected devices. In this thesis, we present effective energy saving strategies that consider the utilization of network elements such as base stations and virtual machines, and implement on/off mechanisms taking into account the quality of service (QoS) required by mobile users. Moreover, we investigate the performance of a NOMA-based resource allocation scheme in the context of Internet of Things aiming to improve network scalability and reduce the energy consumption of mobile users. The system model is mainly built upon the M/M/k queueing system that has been widely used in most relevant works. First, the energy saving mechanism is formulated as a 0-1 knapsack problem where the weight and value of each small base station is determined by the utilization and proportion of computing tasks at that base station, respectively. The problem is then solved using the dynamic programming approach which showed significant energy saving performance while maintaining the cloud response time at desired levels. Afterwards, the energy saving mechanism is applied on edge computing to reduce the amount of under-utilized virtual machines in edge devices. Herein, the square-root staffing rule and the Halfin-Whitt function are used to determine the minimum number of virtual machines required to maintain the queueing probability below a threshold value. On the user level, reducing energy consumption can be achieved by maximizing data rate provision to reduce the task completion time, and hence, the transmission

energy. Herein, a NOMA-based scheme is introduced, particularly, the sparse code multiple access (SCMA) technique that allows subcarriers to be shared by multiple users. Not only does SCMA help provide higher data rates but also increase the number of accommodated users. In this context, a power optimization and codebook allocation problems are formulated and solved using the water-filling and heuristic approaches, respectively. Results show that SCMA can significantly improve data rate provision and accommodate more mobile users with improved user satisfaction.

Acknowledgments

I would like to sincerely thank my supervisor and mentor, Professor Alagan Anpalagan, for his inspiration, patient guidance, and endless support throughout this work. His positive, careful, and insightful feedback has given me the confidence and motivation to successfully achieve this work. It was truly a great privilege working with him.

I would also like to graciously thank my thesis committee members, Prof. Bobby Ma, Prof. Lian Zhao, Prof. Vojislav Misic, Prof. Isaac Woungang, and Prof. Min Dong from UOIT for their valuable time and efforts for reviewing this dissertation. My sincere thanks are extended to all Ryerson University professors and staff who taught, helped, and supported me throughout the course of my PhD study.

I would also like to acknowledge our research group members Dr. Glaucio Carvalho and Dr. Serhat Erkucuk for their helpful discussions, constructive suggestions, and invaluable technical inputs.

Finally, I would like to thank my family for all their support and love they provided me throughout my entire life.

Contents

List of Tables

List of Figures

List of Acronyms

ABS	Almost Blank Subframes
AP	Access Point
BS	Base Station
BBU	Baseband Unit
CA	Carrier Aggregation
CB	Coordinated Beamforming
CC-CRRM	Cloud Computing-based Cooperative Radio Resource Management
CDMA	Code Division Multiple Access
CoMP	Coordinated Multi-Point
CPU	Central Processing Unit
C-RAN	Cloud Radio Access Network
CSI	Channel State Information
CTMC	Continuous-time Markov Chain
DPB	Dynamic Point Blanking
DRX	Discontinuous Reception
DTMDP	Discrete-time Markov Decision Process
DTX	Discontinuous Transmission
eICIC	Enhanced Inter-cell Interference Coordination
H-CRAN	Heterogeneous Cloud Radio Access Network
HetNet	Heterogeneous Network
HPN	High Power Node
HUEs	HPN users
ICT	Information and Communications Technology

IoT	Internet of Things
KKT	Karush-Kuhn-Tucker
LPN	Low Power Node
LTE	Long Term Evolution
MAC	Medium Access Control
MBS	Macro Base Station
MDP	Markov Decision Process
MIMO	Multiple-Input Multiple-Output
MINLP	Mixed-Integer Nonlinear Programming
NOMA	Non-Orthogonal Multiple Access
OFDMA	Orthogonal Frequency Division Multiple Access
OPEX	Operation Expenditures
PA	Power Allocation
PC-RAN	Partially Centralized RAN
PZ	Power Zone
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RAT	Radio Access Technology
RB	Resource Block
RoF	Radio over Fiber
RRH	Remote Radio Head
RUEs	RRH users
SBS	Small Base Station
SCMA	Sparse Code Multiple Access

SDN	Software-defined Networking
S-FFR	Soft Fractional Frequency Reuse
SINR	Signal-to-Interference-plus-Noise Ratio
SLA	Service Level Agreement
SMDP	Semi-Markov Decision Process
SRS	Square-root Staffing
SQNR	Signal-to-quantization-noise Ratio
VM	Virtual Machine
ZFBF	Zero Forcing Beamforming

List of Symbols

B	Bandwidth
b_{fl}	Fiber backhaul link speed
b_m	Total bit rate provided by the MBS
b_s	Total bit rate provided by the SBS
C	Processing speed at each fog node
c_e	Edge device processing speed
d_c	Maximum number of subcarriers per user
D_k	Data size of user k
D_u	Data size of tasks
d_s	Maximum number of users per subcarrier
F_n	Fronthaul capacity of edge device n
k_c	Number of servers in the cloud
k_m	Number of servers in the MBS
k_s	Number of servers in the edge device
l_u	User-edge device association operator
M	Cardinality of the constellation points
N_0	Noise power spectral density
N_e	Number of edge devices
N_s	Number of SBSs
N_{sc}	Number of subcarriers
N_u	Number of users associated with edge device
N_v	Number of VMs per edge device
p_e	Power consumption of an edge device

p_u	User transmit power
P_s	SBS power
p_s	User transmit power to the SBS
p_m	User transmit power to the MBS
p_v	Power consumption of one VM
S_t	Data size of tasks
s_u	Wireless link speed for each user
v_e	CPU clock speed of each VM in the edge device
v_c	CPU clock speed of each VM in the cloud
λ_m	Task arrival rate at MBS
λ_j	Task arrival rate at the jth SBS
α_j	Computing ratio at the jth SBS
μ_m	MBS service rate
μ_s	SBS service rate
μ_c	Cloud service rate
λ_n	Arrival rate at each edge device
μ_n	Departure rate at each edge devices
ζ_k	Processing speed of IoT devices

Chapter 1

Introduction

1.1 Overview

The prosperity of smart devices and their applications in a wide variety of life aspects such as health care, distant learning, road traffic control, public safety, etc., urged the communication and networking society to upgrade traditional networks in order to satisfy future needs. One of the major concerns regarding the mobile communication networks is energy consumption. The information and communication technology (ICT) accounts for 3% of the entire global energy consumption producing about 2% of the total CO_2 emissions [1]. In particular, mobile communications consume about 0.5% of the entire global energy [2]. Moreover, about 70% of the network's energy is consumed by the radio access networks (RANs), this amount of energy has effects on two dimensions: the first dimension is the carbon footprint, where the ICT's carbon footprint is comparable to the entire world's aviation industry. The other dimension is the operational expenditure (OPEX), where energy forms 7 - 20% of the entire network OPEX [3]. Therefore, large amounts of energy are being, and will continue to be wasted if no serious actions are made towards resolving the energy issue which acts as a bottleneck for future networks.

On the computing side, cloud and edge (fog) computing allows mobile devices to benefit the powerful computing capabilities in cloud and edge servers aiming to accomplish those

tasks during shorter time and with less energy consumption. However, due to the variations in user behaviour over time, space, and desires turn the resource and energy management process into a complicated process that necessitates the adoption of dynamic strategies to run and control network elements. In this context, studies such as [4] have shown that less than 30% of the computing resources of some large data centers are consumed while energy is consumed at a rate close to that of a full load. Therefore, the effective cooperation between both radio and computing nodes is crucial to maximize energy saving and system agility.

To this end, the emerging cloud radio access network (C-RAN) that is composed of multi-tier coverage zones such macro cells and small cells, and comprises multiple radio access technologies (RATs) such 5G and WiFi, helps to merge radio resources from all nodes in a unified entity, namely the baseband unit (BBU) pool. The heterogenous architecture of C-RANs helps to improve the total network capacity; however, with the massive number of mobile devices and base stations, large amounts of information needs to be exchanged to establish efficient collaboration and to maintain high quality of service (QoS) standards [5]. Benefiting the advances in software-defined networking (SDN) along with the network function virtualization technologies, the BBU pool facilitates the monitoring and control process for the entire network using programable machines, thus reducing human intervention and operational costs.

One of the efficient energy saving strategies is to monitor the traffic flow associated with base stations and edge devices, and set lightly loaded elements into the off/sleep mode. However, with the strict delay requirements of some applications such as in medical and vehicular systems that can tolerate only few milliseconds of delay [6], energy saving has to be carefully implemented to avoid the undesired delays. Furthermore, due to scarcity in frequency resources, accommodating the continuously growing number of connected devices and improving data rate provision is becoming more challenging. Therefore, allowing users to share frequency resources in non-orthogonal multiple access (NOMA) systems has emerged as an attractive solution to back up spectrum shortage in future cellular systems.

1.2 Thesis Motivation

The motivation of this thesis stems from the growing trend towards integrating both radio and computing systems to accommodate the massive numbers of devices that are constantly joining the network in many areas such as health care, smart homes, autonomous vehicles, etc. Moreover, the Internet of Things (IoT) which has attracted much attention from both industry and academia necessitates the cooperation among network operators and service providers. However, with the large number of devices, base stations, and computing nodes, the management and control of all network elements can be an exhaustive process. To this end, recent advances in network virtualization and SDN can make the process much easier. The C-RAN architecture aggregates all network information on centralized processing units to facilitate efficient network-wide management and global optimization. Furthermore, the cooperation among cloud and edge nodes helps to enhance the computing services and to handle the heterogeneity of computing tasks that have different quality of service (QoS) demands.

Originating from the urgent needs for energy saving policies in future cellular network that are featured with the ultra dense deployment of small radio and computing nodes, this work aims to develop advanced energy policies that consider the joint operation of base stations with cloud and edge computing nodes taking into account the user QoS demands. Energy can be saved not only on the cellular side, but also on edge computing side where the large number of idle edge devices, or virtual machines, can also lead to significant amounts of energy wastage.

On the user-level, saving energy can be achieved by providing larger data rates to reduce the transmission time, thus prolonging the on-device battery lifetime. Moreover, with the continuously increasing number of connected devices, novel radio resource allocation schemes need to be implemented and tested to satisfy user requirements. To this end, NOMA-based techniques, in particular, the sparse code multiple access (SCMA) is proposed and tested in the IoT scenario. In SCMA, the detection complexity of subcarriers intuitively increases

due to the complex nature of codebook design that is intended to overcome the inter-carrier interference; as a result, more delay can be experienced by mobile users. Therefore, the delay experienced by users due to subcarrier detection is also considered when investigating SCMA performance where meeting the task completion deadline indicates user satisfaction.

1.3 Research Contributions

The contribution of this thesis can be summarized as follows:

1. Maximize energy saving in heterogeneous networks using efficient base station sleeping strategy considering the cloud's response time and users deadlines:

    - A small base station (SBS) sleeping mechanism is proposed to save energy in integrated H-CRAN-cloud-edge networks under the constraints of cloud response time and task completion deadline. In other words, two types of constraints are considered namely the long-term statistical cloud response time, and the instantaneous task completion time. In this part of the work, the cloud and edge servers are assumed to have disjoint operation; that is, the workload cannot be shared (disjoint queue model). The problem is formulated as a 0-1 knapsack problem wherein the SBS utilization represents the weight whereas the amount of incoming computing tasks represents the value of that SBS. Here, SBSs serving less amount of computing tasks are given higher values than others. The proposed problem, which is solved using dynamic programming, is a centralized SBS sleeping scheme that aims to select the optimal subset of sleeping SBSs considering cloud and user constraints.

    - A novel shared cloud-edge computing architecture is introduced in coordination with the cellular infrastructure. Here, edge and cloud servers are integrated in a unified queue system i.e., one queue and shared servers. Thereby, edge devices contribute to the improvement of the computing response time by increasing the

total number of functioning servers.

- The optimal subset of sleeping SBSs is then found in the later system using exhaustive search approach. Again, the computing response time and task completion deadline are considered as constraints in this problem.

2. Implement a power saving mechanism on edge device taking into account the QoS requirements of end-users:

   - An SDN-assisted energy saving scheme is proposed for edge computing networks aiming at reducing energy consumption under the queueing probability constraint that directly affects the delay experienced by users. In other words, the on/off mechanism is performed while the queueing probability of users is maintained below a pre-defined value such that the queueing delay remains at satisfactory levels.
   - The proposed scheme is formulated as an optimization problem with the objective to minimize energy consumption under the queueing probability constraint. The problem is then solved using the square-root staffing rule and the Halfin-whitt delay function.
   - A comparison between full- and partial-sleep modes for edge device is conducted to investigate the system performance regarding energy and task migration. In the partial sleep mode, some virtual machines (VMs) are turned off locally within edge devices compared to the entire edge device.
   - A load management strategy is proposed to handle the overloaded edge devices in order to maximize the number of accommodated users under the fronthaul capacity constraint.

3. Improve network connectivity and maximize the data rates provided to users aiming to save the on-device battery energy:

- A SCMA-based scheme is proposed for edge computing to improve IoT system connectivity, throughput, and reduce task completion time in HetNets as compared to orthogonal multiple access schemes.
- An optimization problem is formulated to maximize data rate provisioning under the maximum power constraint of base stations. The problem is subdivided into a power allocation problem which is solved using the water-filling technique, and a codebook allocation algorithm which aims to assign users the codebooks with the highest signal-to-interference-plus-noise-ratio (SINR).
- SCMA parameters are investigated to fulfill the high QoS requirements in IoT systems. Since each IoT application has different processing requirements, CPU cycles are allocated considering the total computing capacity of the fog node. Moreover, each IoT device is assumed to have a particular processor speed to consider the detection time with total experienced delay.

1.4 Thesis Outline

The rest of the thesis is organized as follows. In Chapter 2, a comprehensive literature review on H-CRANs and cloud-fog computing networks regrading architecture, energy efficiency, radio and computing resource management, SDN-assisted cooperative performance among radio and computing nodes, and user-level computing and energy consumption, is presented. A base station sleeping scheme is introduced in Chapter 3, where the objective is to maximize energy saving by allowing small base stations with light load to enter a sleep mode taking into account the effect of task offloading from SBSs to the macro base station (MBS). Herein, the MBS utilization has a key role in deciding the number of potential sleeping SBSs. Moreover, the cloud response time is considered as a quality constraint that has to be met. To this end, a 0-1 knapsack problem is formulated to maximize the number of sleeping base stations with minimum queue delay. Energy saving in edge devices is proposed in Chapter 4, where edge device, assisted by the SDN-controller, forms a cooperative group of computing nodes

that share their available resources to maximize energy saving. Herein, full- versus partial-sleep modes are proposed and compared from the perspectives of energy saving and task migration. To this end, the square-root staffing rule and the Halfin-Whitt function are both employed to achieve the desired performance. In Chapter 5, a NOMA-based technique, namely, the SCMA is implemented to improve the network scalability and to maximize the data rate provision. To this goal, a power optimization is formulated and solved using the water-filling technique, and the detection complexity of SCMA subcarriers is considered to evaluate the SCMA performance in the sense of satisfying users with strict task completion deadlines such as delay-sensitive tasks. Finally, Chapter 6 provides concluding remarks and directions for the future research. Fig. 1.1 depicts the thesis outline.

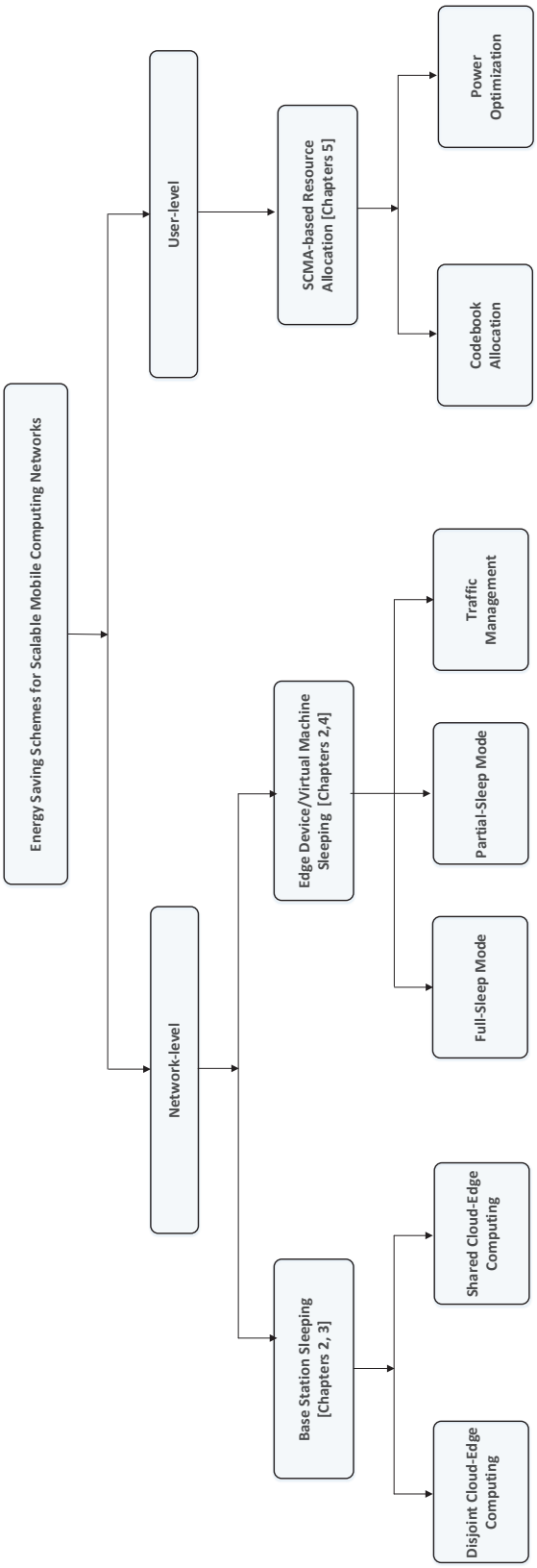


Figure 1.1: Thesis Outline.

Chapter 2

Literature Review on Heterogeneous Cellular-Computing Networks: Energy, Infrastructure, and Resource Management

This chapter introduces insights on future cellular networks, followed by discussing state-of-the art base station sleeping strategies for energy saving.

2.1 Energy Efficiency in HetNets

In future wireless networks, high data rate provisioning is an essence to cope with the ever increasing smart phone applications in a wide variety of life aspects such as health, transportation, remote monitoring, etc. Heterogeneous networks (HetNets) is a promising architecture that enables future networks achieve high data rates everywhere at all times. Small cells form the basic building block in HetNets by facilitating spatial frequency reuse in small geographical areas thus allowing to improve the efficiency of the scarce frequency resources.

The dense deployment of small cells in HetNets, despite the large improvement on spectral efficiency, will incur high amount of energy consumption. The largest proportion of energy in HetNets is consumed by the RANs (specifically BSs), and more than 80 percent of energy in wireless networks is lost as heat [7] [8]. 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 [9]. Moreover, even with no or light traffic load, a BS consumes more than 90 percent of its peak energy [10]. With the adoption of appropriate sleeping strategies, the excessive amounts of power requirements (e.g., for cooling) can be avoided. Power consumption can be further reduced if the coverage area of a BS is adaptively reduced according to cell load [11]. Table 2.1 presents recent energy-efficient approaches for HetNet management.

Table 2.1: Energy Management in HetNets

Ref.	**Research Direction**	**Problem Type**	**Solution Approach**
[1]	EE maximization considering the transmit, backhaul, and circuit power in CoMP OFDMA-based HetNets under the constraint of data rate	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

Continued on next page

Table 2.1 – *Continued*

Ref.	**Research Direction**	**Problem Type**	**Solution Approach**
[7]	Resource allocation for maximizing network utility and stabilizing the queues for media applications in HetNet with multihoming transmission	Stochastic optimization problem	Lyapunov drift-plus-penalty method, primal-dual decomposition technique, Lagrange multipliers and KKT conditions
[12]	Joint optimization of cell activation, user association, and spectrum allocation	mixed-integer program	Reweighted l_1 minimization (majorization-minimization) method
[13]	Joint maximization of EE and SE considering small cell density and offloading biasing factor subjected to throughput threshold	Quasi-convex multi-objective function optimization	Fritz-John conditions to determine the Pareto-efficient operational regime
[14]	Power allocation and beamforming design	Mixed combinatorial non-convex optimization	- EE maximization is transformed to power minimization and an optimal solution is obtained based on convex programming - Near-optimal upper-bound solution - Suboptimal zero forcing (ZF)-based solution to further simplify the solution

Continued on next page

Table 2.1 – *Continued*

Ref.	**Research Direction**	**Problem Type**	**Solution Approach**
[15]	BS sleep mode using the concept of group sparsity in transmit power vector, BS association, downlink power allocation in TDD-based HetNets	NP-hard non-convex optimization	Successive convex approximation (SCA)-based algorithm, KKT conditions
[16]	Joint optimization of BS operation, user association, subcarrier assignment, and power allocation	Mixed combinatorial problem	- Lyapunov optimization, - Heuristic algorithm
[17]	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
[18]	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, - Heuristic algorithm

Continued on next page

Table 2.1 – *Continued*

Ref.	**Research Direction**	**Problem Type**	**Solution Approach**
[19]	Energy-efficient trade-off between backhaul energy and throughput subjected 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
[20]	Maximizing EE 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
[3]	Joint cell association and on-off policy to minimize energy consumption	General non-convex energy minimization problem	0-1 Knapsack-like optimization

Macro BSs are generally aimed at providing large coverage areas rather than high data rates; therefore, the existence of small BSs is inevitable in future dense networks [21]. However, small cells are more prone to traffic variations than macrocell BSs [12]. Selective activation of femtocell networks in places that are characterized by concentrated traffic load, can significantly reduce power consumption and outage probabilities [22].

A multi-level decision-making strategy for a macro-femto network was formulated in [23] as a multi-level optimization problem where decision making in one level (e.g., macro, femto, or user) depends on 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.

The authors in [24] proposed a joint subcarrier and power allocation scheme to increase throughput, mitigate 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.

Renewable energy harvesting has been studied by many researchers as an alternative approach to empower SBSs thereby reducing on-grid power consumption. SBS types 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 [25]. Furthermore, mixed renewable energy resources and grid power have been studied in [26] [27] to minimize grid power consumption by allocating renewable energy resources efficiently along with BSs on-off strategy.

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 [28] 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 [29].

Other schemes such as bio-inspired systems have been incorporated in the context of self-organizing networks. For example, the work in [30] defined the analogy between mam-

malian 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 [31] [32].

The authors in [2] emphasized on minimizing energy consumption rather than maximizing energy efficiency that combines energy consumption with data throughput. This realization is due to the fact that total network energy efficiency may not satisfy the per-user or per-cell energy and data requirements in HetNets.

2.2 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' 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 heterogeneous 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 [33].

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 devices [34]. For instance, the authors in [15] proposed a power model for cloud-assisted HetNets wherein the large-scale fading imposed on the communication channel is considered to be fixed. The model jointly optimized 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 three macro BSs and five 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 compared to the simple on/off strategy that required 3.2 kW. Furthermore, in [35], a scheme was proposed to minimize energy consumption in dense C-RANs by activating selected subsets of RRHs according to the dynamic traffic variations. Results showed that the dynamic activation of RRHs was successful to make significant energy savings. For instance, within the coverage of two micro and seven 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 [36].

From the computational complexity perspective, the high density of RRHs required to provide high data rates incurs high computational complexity due to the huge amount of data related to signal processing, resource allocation, and RRHs/BBUs coordination. This complexity is a big challenge facing the establishment of scalable networks. The authors in [37] 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.

The authors in [38] proposed a joint optimization of resource block (RB) assignment and power allocation subject to the constraints of user association and interference between RRHs and high-power nodes (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-rate QoS requirements, and the second set is shared between RUEs and UEs that are associated with the HPN (HUEs) that require low-rate QoS. The RBs in time/frequency domains have been identified in [39] 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 [40], 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 [41].

Moreover, to maximize the spectrum and energy efficiencies, and to achieve ultra-low-latency communications, the authors in [42] introduced the open-loop communications as a promising technique to fulfill these requirements by avoiding the redundant feedback mes-

sages 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, and the required number of repeated transmission in one shot. In regard to uplink transmission in H-CRANs, a scheme was proposed in [43] to jointly optimize power allocation, relay selection, and network selection under the QoS constraints in order to maximize the network's energy efficiency. Table 2.2 summarizes recent trends followed to achieve energy-efficiency in H-CRANs.

Table 2.2: Energy Management in H-CRANs

Ref.	**Research Direction**	**Problem Type**	**Solution Approach**
[24]	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
[38]	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
[35]	Active RRH subset determination based on traffic demand and sleeping strategy	Multiple choice multi-dimensional knapsack problem	Lagrange multipliers

Continued on next page

Table 2.2 – *Continued*

Ref.	**Research Direction**	**Problem Type**	**Solution Approach**
[40]	Maximizing EE 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
[43]	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
[44]	Optimizing the transmit power of RRHs and HPNs along with 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
[45]	Maximizing the average throughput and network stability (queue stability) subject to the constraints of power allocation, resource block allocation, and user association	Stochastic optimization containing a mixed-integer subproblem	- Lyapunov optimization, - Lagrange dual decomposition method (to solve the subproblem)

Continued on next page

Table 2.2 – *Continued*

Ref.	**Research Direction**	**Problem Type**	**Solution Approach**
[46]	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
[47]	Dynamic resource allocation (subcarrier and power allocation, RRH clustering, and CoMP) in TDD-based H-CRANs	Mixed strategy noncooperative game	Reinforcement learning algorithm is used to achieve Logit equilibrium
[48]	Joint RRH selection and user association to minimize energy consumption	Integer programming problem	- Multiple-choice multidimensional knapsack model is used for user association with each RRH, with the consideration of fronthaul capacity and radio resources - RRH selection is achieved using low-complexity heuristic algorithm
[49]	Cross-tier cooperation and cluster formation among LPNs and HPNs towards throughput enhancement	0-1 Multiple knapsack	Heuristic algorithms

Continued on next page

Table 2.2 – *Continued*

Ref.	**Research Direction**	**Problem Type**	**Solution Approach**
[50]	Joint BS selection and beamforming design fro power minimization under the constraints of limited fronthaul capacity	l_0 minimization problem	Majorization-Minimization algorithm
[51]	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 a convex relaxation approach

2.3 Resource Management and Network Resource Optimization

To take full advantages of the high computing capabilities provided by the cloud servers, it is paramount to 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 [52]. 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 [53] and smart cities [54]. To this end, the issues of frequency allocation, inter-

ference coordination, RRH clustering, and fronthaul/backhaul management are reviewed in this section.

2.3.1 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 [55], where the RRHs activation and BBUs capacity (e.g., processor speed, memory capacity, etc.) can adapt to the variations in data demands. This 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.

Based on [56], resource sharing in H-CRANs can be divided into three levels: 1) spectrum sharing, this 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, the central workload computations of RRHs and HPNs in the BBU pool facilitate the virtualization of available resources of 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 infrastructure sharing among network operators and reduces the CAPEX and OPEX; 3) network sharing to efficiently manage the available spectrum and infrastructure resources.

The authors in [57] 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, that 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 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 [58]. Furthermore, a graph-based dynamic frequency reuse has been presented in [59], 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 intelligent networking performance that can adapt to data demands (e.g., IoT demands) that fluctuates over time and place [34].

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) [34]. 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 [7].

For a mobile user to access the best candidate BS, a time delay of several hundreds of milliseconds is required. 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, MBSs will be responsible for selecting the best SBSs and providing 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 SBSs and MBSs in the areas supported by both tiers; otherwise, MBSs will keep data pro-

visioning wherever small cell coverage is missing [60]. To deal with the delay-aware radio resource allocation problems, Markov decision process (MDP) which is a stochastic learning approach, is considered as a successful method, and has been recognized to outperform other optimization techniques such as Lyapunov optimization and the equivalent rate constraint approach [61].

Interference mitigation techniques proposed for H-CRANs will be introduced next sas 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.

2.3.2 Interference Coordination

Heterogeneity in cellular networks that consist of base stations with different sizes and RATs, can significantly improve the total system capacity; however, high interference levels will be encountered. Moreover, the dense deployment of RRHs produces 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 [47]. 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 [56]. 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 [62]. Interference coordination techniques for H-CRANs are presented in Table 2.3.

Table 2.3: Interference Mitigation in H-CRANs

Ref.	**Research Direction**	**Problem Type**	**Solution Approach**
[63]	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
[64]	Interference coordination between MBSs and RRHs	Contract-based optimization	- Contract-based game theory, - Lagrange multipliers, - KKT conditions
[65]	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
[66]	Inter-tier interference-aware macrocells paradigm for radio resource allocation, such that macrocells can maximize the interference levels tolerated by associated users under the QoS constraints	Mixed-integer nonlinear programming	Successive convex optimization

Continued on next page

Table 2.3 – *Continued*

Ref.	**Research Direction**	**Problem Type**	**Solution Approach**
[67]	Inter-tier interference reduction based on power hierarchy (e.g., macro, femto)	Noncooperative game (high- and low-power BSs are the players)	Nash equilibrium and KKT conditions
[68]	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
[69]	Inter-tier interference mitigation among HPNs and LPNs in H-CRANs through optimized power allocation	Non-convex optimization problem	Perron-Frobenius theory
[70]	Joint user-access point (AP) association and beamforming design for interference coordination in both uplink and downlink transmission	NP hard	- Group-sparse optimization, - Relaxed-intger programming

Dividing the coverage area into sub-regions with different frequency sub-bands in the technique of soft fractional frequency reuse (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 [34].

A dynamic resource allocation scheme in [63] 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.

A joint cooperative interference mitigation and handover management scheme was proposed in [71] 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 [20].

Serving a large number of users simultaneously concentrated in certain zones by macro and small cells incurs a strong inter-tier interference. The technology of massive-MIMO provides the opportunity for 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 macro cell signals [72].

The intra-tier interference among low power nodes (LPNs) can be mitigated using 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 [36]. 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 [73].

A contract-based interference coordination between RRHs and MBSs in H-CRANs was

presented in [64]. 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.

A device-to-device communication scheme was proposed in [74] 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.

A hierarchical access to frequency resources based on the concept of cognitive radio can also be applied by femtocells that can act as secondary users who use frequencies only when no primary users are using that particular frequency. This helps to avoid the overlapping of signals with other users associated with other cells [75].

A decentralized multiple cloud architecture in C-RAN was proposed in [76] 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 [65], 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 led the H-CRAN to perform better depending on the network configurations such as the number of antennas deployed by the MBSs, number of RRHs, and SINR threshold.

2.3.3 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 sometimes impractical, thus BS clustering is essential to maintain

high QoS [77]. 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 [78]. 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 [21].

It has been shown in [79] 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 [80]. 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 [81] to mitigate inter-cluster interference in OFDMA-based systems. Unlike the traditional omni-subcarrier CoMP, this 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 [82] 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 2.4 lists some of the technical approaches applied in cell clustering.

Table 2.4: BS Clustering

Ref.	**Research Direction**	**Problem Type**	**Solution Approach**
[77]	BS clustering based on long-term user throughput considering CSI overhead and spectral efficiency	Coalition game where BSs are considered the players	Distributed coalition formation algorithm and a precoding algorithm based on weighted minimum mean squared error
[80]	Joint optimization of BS transmit power, BS activation, and backhaul power are compared under data-sharing and compression strategies considering the data rate requirements	Discrete non-convex optimization	Re-weighted l_1-norm minimization and successive convex approximation
[83]	User-centric BS clustering and sparse beamforming, wherein BSs are equipped with limited storage cache to reduce burden on backhaul links	Mixed-integer nonlinear programming	Iterative re-weighted l_1-norm technique

Continued on next page

Table 2.4 – *Continued*

Ref.	**Research Direction**	**Problem Type**	**Solution Approach**
[84]	Joint RRH clustering and activation optimization under the constraints of coverage and user's QoS	NP-hard problem	- Linear-programming relaxation and, - Greedy algorithm
[85]	Cell clustering and activation time for energy minimization subjected to data provision within a specific time and inter-cell interference	NP-hard problem	- Column generation, - Local-enumaration-based bounding scheme, - Near-optimal cluster scheduling

2.3.4 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 geo-spatial fluctuations that characterize the traffic trend. 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 1/3 of the total bandwidth [86].

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 [57]. Based on the observations of [87] [88], 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 [89] as a sharing technique that can maximize energy efficiency in small cell HetNets.

The authors in [90] 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 [91].

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 [36]. 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 [92].

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 [93], 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 [61].

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 [94].

In [95], 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 [96].

To overcome the limited capacity in fronthaul links, time-reversal (TR)-based communications for air interfacing have been proposed in [97] 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 multi-path effects of the communication environment by acting as a matched filter that adjusts the temporal and spatial effects.

2.4 Base Station Sleeping

BSs consume power mainly for operational purposes (e.g., cooling and signal processing) and radio transmission. 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 consumed power is exploited to broadcast radio signals [98].

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 [21]. However, small cells are more prone to traffic variations than macro cell BSs [12]. Selective activation of femto cell networks in places that are characterized by concentrated traffic load, can significantly reduce power consumption and outage probabilities [22].

The ultra dense deployment of small cells in hotspots such as shopping malls and airports leads to under-utilization of these cells at most times and thus large energy losses [99]. Therefore, traffic offloading has been considered as a promising solution to give the opportunity of switching off lightly loaded BSs based on traffic demands [100]. Thus, the intra- and inter-RAT traffic offloading, can improve both energy and spectral efficiencies in HetNets. However, the incurred intra- and inter-RAT interference along with the increased burdens on the capacity-limited backhauls degrades the total energy efficiency and spectral efficiency gains [13]. In [14] traffic demands have been sorted into real-time services such as video conferencing where fixed and high data rate provisioning is required, 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.

The study in [101] 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 extra flexibility in deciding 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, and hence reducing the total power consumption. Some of the recent advances that happened in BS sleeping mechanisms are introduced in Table 2.5.

Table 2.5: Related Work on BS Sleeping

Ref#	Research Direction	Problem Type	Solution Approach
[8]	EE 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 EE lower bound through iterative algorithms
[100]	Joint energy-efficiency and load-balancing in multi-RAT HetNets	Semi-Markov decision process (SMDP)	Optimal policy using Markov decision process
[25]	BS on-off and traffic offloading scheme based on traffic load and renewable energy availability	0-1 knapsack problem	Lagrange multipliers
[102]	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
[103]	BS sleeping strategy and user association in open-access femtocell networks	Binary integer problem	- Heuristic algorithm, - Lagrange dual method
[104]	EE maximization by determining BS density and sleeping strategy	Non-convex optimization	Dynamic gradient iterative algorithm

Continued on next page

Table 2.5 – *Continued*

Ref#	**Research Direction**	**Problem Type**	**Solution Approach**
[105]	Energy-efficient optimal BS activation and cell size subjected to network coverage	NP-hard	Polynomial-time algorithm
[106]	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 solved by linear programming approach

Energy saving can be obtained either by adjusting the transmit power of BSs according to traffic load or by letting the entire BS go to sleep when light or no traffic exists. The former method is considered to have more energy saving due to the fact that most of BSs energy is consumed by circuits rather than transmit power [107]. Furthermore, controlling the operation of of BS components (e.g., electric circuits, power amplifiers, etc.) yields different sleeping modes with different activation periods. For instance, activating a BS from light sleep (standby) mode incurs a time delay of 0.5 seconds, whereas activating BS from deep sleep mode incurs 10 seconds, and activating a BS that is turned off requires 30 seconds [8].

Furthermore, shutting down BSs with light or no load increases the amount of delay experienced by users. This delay is generally due to the longer queue of users offloaded to Macro BSs, and SBS activation delay that can reach up to 30 seconds from off to on state [8]. Researchers in [101] [10] studied the energy-delay tradeoff in BS sleeping strategies. The authors in [108] 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 users. Thus an optimal value of N has to 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 [10] to maintain system stability while implementing 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.

2.5 Chapter Summary

In this chapter, an extensive literature review on energy efficiency in both C-RAN and cloud-edge networks was presented. First, the architecture of C-RANs has been introduced in details with a highlight on the SDN and NFV technologies implemented in the BBU pool. Several energy-efficient techniques have been introduced and tabulated to showcase the importance of energy efficiency in mobile networks. The cooperative performance of the heterogeneous network elements have shown to be crucial for improved network-wide performance. The system model in this chapter forms the starting point for the next chapters where the C-RANs along with the layered cloud-edge architecture is considered in the proposed energy saving schemes in this thesis.

In the next chapter, an energy saving mechanism is implemented in C-RANs by allowing SBSs to enter a sleep mode taking into account the cloud response time and the task completion deadline of mobile users. The C-RAN structure will consider an MBS, multiple SBSs, BBU pool, cloud, and edge devices. Both computing and non-computing tasks are considered in this energy saving mechanism such that both types of tasks are accommodated by the MBS when SBSs enter the sleep mode.

Chapter 3

Computing-Aware Base Station Sleeping Mechanism in H-CRAN-Cloud-Edge Networks

3.1 Introduction

Future cellular networks are characterized by their capability to satisfy the stringent needs of mobile users in regard with latency and data rate [109]. Providing network coverage using small base stations along with macro base stations has emerged as an attractive solution to improve network scalability and to cope with the growing number of mobile devices. However, with the large number of base stations and RATs employed in heterogeneous networks, the management of mobile networks is becoming more and more complicated. To this end, performing data aggregation from all network nodes in the centralized BBU pool for processing, in the well-known architecture of heterogeneous cloud radio access networks (H-CRANs), can achieve huge success in this direction [110]. The RRHs and SBSs in H-CRANs are basically deployed to provide high data rates by exploiting the spatial reuse of frequencies. Meanwhile, MBSs are in charge of providing cross-tier management such as user association, handover management, traffic flow, and network-wide coverage. In other words,

SBSs belong to the data plane whereas MBSs belong to the control plane.

From the computing perspective, having the complex computing tasks such as computer vision and data analytics processed in the central cloud is a big step towards improving the computing performance for users and machines [111]. Nevertheless, the increasing number of connected devices in the context of Internet of Things, smart homes, autonomous driving, etc., will eventually overload or even crash cloud servers. Thus, it is essential to filtrate data to reduce the burden on the cloud and network resources, and to improve the QoE especially in regard with end-to-end delay [112] [113].

Bringing computing services at the vicinity of mobile users in the paradigm of edge (fog) computing can significantly reduce the end-to-end delay experienced by users. This reduced delay helps support the emerging delay-sensitive applications such as E-health, real-time control, and vehicular communications [114] that can tolerate a delay of only few milliseconds [6]. Edge devices are equipped with the necessary hardware to enable small-scale cloud-like functions such as computing and storage. Moreover, edge computing benefits the close proximity with mobile users to offer geo- and context-aware services such as content caching. It is thus obvious why edge computing which complements the cloud is described as "fog" because fog physically resides closer to the ground (users) compared to the "cloud" seen in the sky [115]. To take full advantage of edge computing, it is necessary to coordinate edge devices with the central cloud on one hand, and with the H-CRAN on the other hand [116]. With the help of SDN technology, efficient coordination of computing and communication nodes can be achieved with less complexity.

One of the main constraints that stands in the way of future networks is the high energy consumption. Not only because energy raises the operational expenditures, but also because it causes detrimental impacts on our planet. Adopting smart SBS operation mechanisms can significantly reduce energy consumption since base stations account for 80% of the overall energy consumption in cellular networks [107]. Controlling the operation of SBSs can be achieved in a distributed or centralized manner. In the former, an SBS operates as a stand alone entity using intelligent self-organized features. Whereas in the centralized

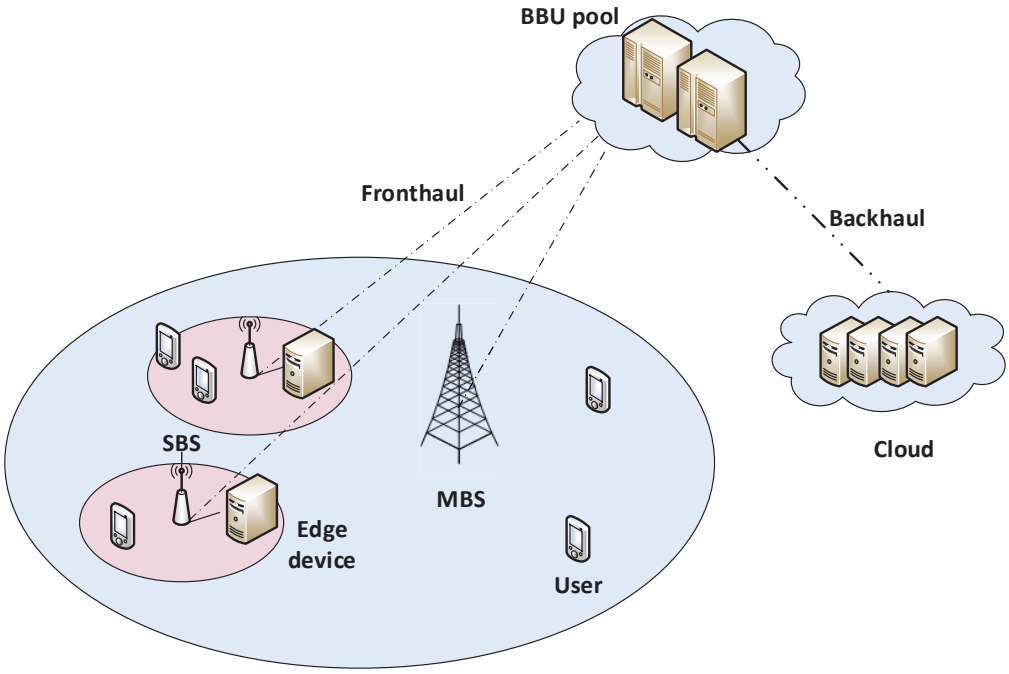


Figure 3.1: H-CRAN-CE system layout.

control, data from SBSs, MBSs, and other supporting nodes enter the central BBU pool for an optimal network-aware processing. The overhead of the centralized control is naturally higher compared to the distributed one; however, the informed and ceratin decisions of the centralized control boosts the overall system performance. Therefore, prior to initiating the On/Off and traffic offloading processes, network nodes should be well coordinated to maintain high QoS [117].

Similar to traffic offloading in cellular networks, computing tasks can also be offloaded from edge devices to the cloud and vice versa depending on the desired QoS requirements such as energy and delay [118]. In other words, computing tasks can be processed either locally by the edge device or remotely by the cloud via the MBS through backhaul links [119]. However, offloading tasks to the central cloud will inherently increase the burden on cloud servers, communication resources, and backhaul links. Moreover, adopting coordinated task offloading in the layered cloud-fog architecture can increase the communication overhead and

thus extra delay [120]. Therefore, it is essential to take into account the consequences of task offloading on both the communication and computing nodes. From the aforementioned, we propose a coordinated cellular-computing architecture that considers both communication and computing resources towards optimal SBS sleeping operation. Fig. 3.1 depicts the state-of-the-art H-CRAN-cloud-edge system.

Over the last few years, SBS sleeping gained considerable attention in the context of HetNets. Nevertheless, limited amount of research considered SBS sleeping from both communication and computing perspectives. In [100], a sleeping strategy was proposed by which all RATs are activated when resource utilization in the MBS reaches a threshold value. The N-policy scheme in [107] is concerned with the energy-delay tradeoff in SBS sleeping without considering traffic offloaded from sleeping SBSs to the MBS. Furthermore, the SBS activation delay was the goal of [8], wherein authors used iterative approaches to maximize energy efficiency considering wake-up times and coverage probability regardless of the MBS traffic load. All aforementioned works were considering performance in a communication environment; that is to say, no computing aspects were involved.

However, the proliferation of computing hungry applications have brought the attention of both academic and industrial communities recently. For instance, a hierarchical edge-cloud architecture was proposed in [121] to achieve workload balancing among different computing tiers. By dynamically distributing the workload on different servers, over 25% improvement in program execution time was obtained. In a similar context, authors in [115] considered workload scheduling to find the optimal power-delay tradeoff in cloud-fog computing systems. Furthermore, a scheduling algorithm was proposed in [122] to minimize the queue delay in cloud servers in order to guarantee the ultra-low latency in Internet services.

Since communication nodes play a major role in linking computing tasks with computing infrastructure, it is essential to consider both communication and computing nodes in contemporary research work. Here, a joint energy harvesting and SBS sleeping was studied in [123] aiming at minimizing energy consumption and improving the caching performance in cache-enabled SBS networks. The work considered the effect of SBS sleeping while maxi-

mizing the hit ratio of cached contents.

Table 3.1: Base Station Sleeping Strategies

Reference	Network Model	Computing Model	Performance Indicator(s)	Sleeping Initiator
[107]	HetNet	None	Energy-delay tradeoff	Number of tasks
[100]	HetNet	None	Energy and blocking probability	Traffic load
[8]	HetNet	None	Energy efficiency and coverage probability	Traffic load
[123]	HetNet	Cache-enabled SBSs	Power consumption & cache hit ratio	Harvested energy and traffic load
[124]	HetNet	None	Power consumption and throughput	Traffic load and user location
[103]	HetNet	None	Power Consumption	Traffic load
[105]	HetNet	None	Power consumption and coverage probability	Traffic load and network coverage
This work	H-CRAN	Cloud-edge	Power consumption, cloud response time, and user energy	Traffic, cloud response time, and task completion time

Unlike most related work in the literature, we aim to maximize power saving considering the SBS load, MBS load, cloud response time, delay experienced by users, and traffic offloaded from sleeping SBSs to the MBS. The joint operation of both communication and computing nodes can improve the network-wide performance and provide sophisticated sleeping mechanism for future networks. Table 3.1 compares this work with related ones in the literature.

The organization of this Chapter is as follows. Section 3.1 provides an overview, related work, and main contributions. Section 3.2 describes the power, network, and computing models. The computing-aware SBS sleeping scheme is introduced in Section 3.3, followed by SBS sleeping in the proposed shared computing model in Section 3.4. In Section 3.5, simulation setup and results are demonstrated, and finally, Section 3.6 provides the Chapter summary.

3.2 System Model

In this section, power, network, and computing models are presented. The general view of the integrated H-CRAN-cloud-edge system can be well perceived in Fig. 3.1.

3.2.1 Power Model

The MBS is assumed to remain active all the time in order to provide coverage, cross-tier control, and to accommodate users offloaded from sleeping SBSs. Accordingly, the MBS has approximately a constant power and thus does not affect the SBS sleeping performance. For this reason, the MBS is not taken into account when calculating the total network power. The SBSs, on the other hand, coexist with the MBS and carry out a flexible On/Off operation. The power consumption of the jth SBS is given by

$$P_j = \begin{cases} P_s, & \text{if SBS is On} \\ 0, & \text{if SBS is Off,} \end{cases} \tag{3.1}$$

where P_s denotes the power consumption during the active mode. It is worth to mention that in the proposed sleeping mechanism the power associated with the sleep/Off mode is considered always 0. Hence, the power consumed by all active SBSs can be expressed as

$$P_t = \sum_{j=1}^{N_s} x_j P_j, \tag{3.2}$$

where x_j is the On/Off indicator of the jth SBS, such that $x_j = 1$ and $x_j = 0$ indicate the On, Off mode, respectively. Now, let $x'_j = 1 - x_j$ denotes the complement of x_j such that $x'_j = 1$ indicates the Off mode, then the total power saving P'_t can be written as

$$P'_t = \sum_{j=1}^{N_s} x'_j P_j. \tag{3.3}$$

Thus, two modes of operation are considered, namely "On" (SBS in full operation) with 100 % power consumption, and "Off" with 0 % power consumption [8]. It should also be noted that the term "sleeping SBS" indicates an SBS that is operating in the "Off" mode and has 0 % power consumption.

3.2.2 Network Model

We consider a heterogeneous network consisting of one MBS and a set of N_s SBSs denoted by $\mathcal{S}$, where users can be associated with either the MBS or a nearby SBS. The management

of all network elements is performed in the central BBU pool which is capable of taking network-wide decisions. In the context of H-CRANs, RRHs generally have lighter processing capabilities compared to SBSs; nevertheless, both SBSs and RRHs are denoted as SBSs in this work assuming that they have similar functionality. The MBS is modeled as an $M/M/k_m$ queueing system in which k_m servers (radio channels) can serve k_m users simultaneously without waiting in the queue. Similarly, each SBS is modeled as an $M/M/k_s$ system with equal service rate but different arrival rates. Now, let λ_m, k_m, and μ_m denote the arrival rate of tasks (users) within only the MBS coverage (no SBS coverage), number of MBS servers, and MBS service rate, respectively, then the MBS utilization can be expressed as

$$\rho_m = \frac{\lambda_m}{k_m \mu_m}, \tag{3.4}$$

where ρ_m must be less than or equal to 1 in order to maintain system stability. To showcase the effect of SBS sleeping on cloud computing, it is assumed that tasks arriving at the MBS have no computing demands. In other words, ρ_m has no direct effect on the the cloud response time; nevertheless, it affects the number of sleeping SBSs (edge devices), and as a consequence, the amount of computing tasks offloaded on to the cloud.

3.2.3 Computing Model

Tasks offloaded from sleeping SBSs are accommodated by the central cloud which is modeled as an $M/M/k_c$ queueing system with k_c servers or virtual machines (VMs). We generally classify tasks into two categories, computing tasks that require realtime processing and feedback from edge or cloud servers, and non-computing tasks that require telephony services without powerful computing capabilities thus can be handled by the cellular nodes. Let λ_j and α_j denote respectively the arrival rate of tasks and the ratio of computing tasks to all incoming tasks (computing plus non-computing) at SBS j, then the total arrival rate of computing tasks at the cloud is

$$\lambda_c = \sum_{j=1}^{N_s} x'_j \lambda_j \alpha_j, \tag{3.5}$$

Accordingly, the cloud utilization ρ_c is expressed as

$$\rho_c = \frac{\lambda_c}{k_c \mu_c}, \tag{3.6}$$

where μ_c denotes the service rate of each server. The performance metric of the system under consideration is the response time offered by the cloud. To this end, we consider the steady state analysis based on the continuous-time Markov chain (CTMC) of the $M/M/k_c$ cloud system as shown in Fig. 3.2.

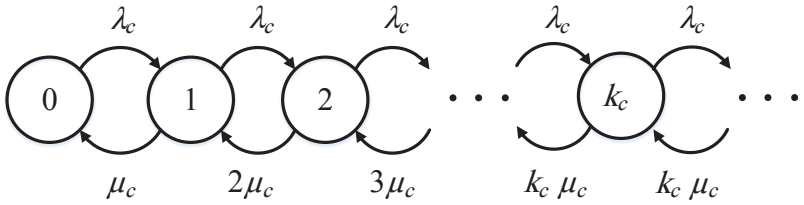


Figure 3.2: Cloud queue model.

The probability that a user will have to queue (all servers are occupied) can be calculated as

$$\begin{aligned} P_Q &= \sum_{i=k_c}^{\infty} \pi_i \\ &= \pi_0 \frac{k_c^{k_c}}{k_c!} \frac{\rho_c^{k_c}}{1-\rho_c}, \end{aligned} \tag{3.7}$$

where π_i represents the steady state probability that i servers are occupied. π_0, which is the steady state probability that zero tasks exist in the cloud, can be written as

$$\pi_0 = \left[\sum_{i=0}^{k_c-1} \frac{(k_c\rho_c)^i}{i!} + \frac{k_c^{k_c}}{k_c!} \frac{\rho_c^{k_c}}{1-\rho_c} \right]^{-1}. \tag{3.8}$$

Therefore, the cloud response time can be obtained by

$$\begin{aligned} E[T_c] &= \frac{1}{\lambda_c} \cdot \frac{\rho_c}{1-\rho_c} \cdot P_Q + \frac{1}{\mu_c}, \\ &= \frac{k_c^{k_c}}{\lambda_c k_c!} \cdot \frac{\rho_c^{k_c+1}}{(1-\rho_c)^2} \cdot \left[\sum_{i=0}^{k_c-1} \frac{(k_c\rho_c)^i}{i!} + \frac{k_c^{k_c}}{k_c!} \frac{\rho_c^{k_c}}{1-\rho_c} \right]^{-1} + \frac{1}{\mu_c}. \end{aligned} \tag{3.9}$$

3.2.4 Cost of Task Migration from Edge to Cloud

In the proposed system, a VM is allocated to each computing task arriving to the edge device or the central cloud with fixed CPU speed. When the serving SBS enters the sleep mode, all associated tasks will be migrated to the central cloud. Thus, the cost of task migration is considered as a delay constraint in the sleeping mechanism. Each task has a particular data size S_t that includes both application data and VM state [125], and a completion deadline θ_t by which the task must be executed and delivered to end-user. However, the migration delay does not consider the setup delay of VMs at the new host. Therefore, the total experienced delay for accomplishing task t consists of three components (a) task execution time (b) data transmission time and (c) response time. Mathematically we can formulate the total delay as follows

$$d_t = \begin{cases} \frac{S_t}{v_e} + 2\frac{S_t}{b_s} + E[T_e], & \text{if SBS is On} \\ \\ \frac{S_t}{v_c} + 2\frac{S_t}{b_m} + 2\frac{S_t}{b_{fl}} + E[T_c], & \text{if SBS is Off,} \end{cases} \tag{3.10}$$

where v_e and v_c denote the VM's CPU clock speed (cycles/sec) in the edge device and the cloud respectively. b_s, b_m, and b_{fl} are respectively the bit rate provided by the SBS, MBS, and the fiber backhaul link. $E[T_e]$ and $E[T_c]$ denote the mean response time at the edge device and the cloud, respectively. The multiplier "2" is used to calculate the time required for both uplink and downlink transmission. Taking into account the cost of task migration, the energy consumed by a mobile device thus depends on whether the task t is processed by the near edge device or the distant cloud:

$$e_t = \begin{cases} p_s d_t, & \text{if SBS is On} \\ \\ p_m d_t, & \text{if SBS is Off,} \end{cases} \tag{3.11}$$

where p_s and p_m denote the user transmit power to the SBS and MBS, respectively. It should be noticed that we consider fixed transmit powers p_s and p_m without considering

the power variations due to the different data size or application requirements of tasks. Since this work incorporates different communication and computing aspects, we assume an appropriate service level agreement (SLA) is committed by cloud providers to ensure all clients have access to cloud facilities [126]. Moreover, the SLA between cloud operators and mobile devices ensures that all computing tasks are completed before the deadline and are compensated for the extra energy consumption due to sleeping SBSs.

3.3 Computing-Aware SBS Sleeping

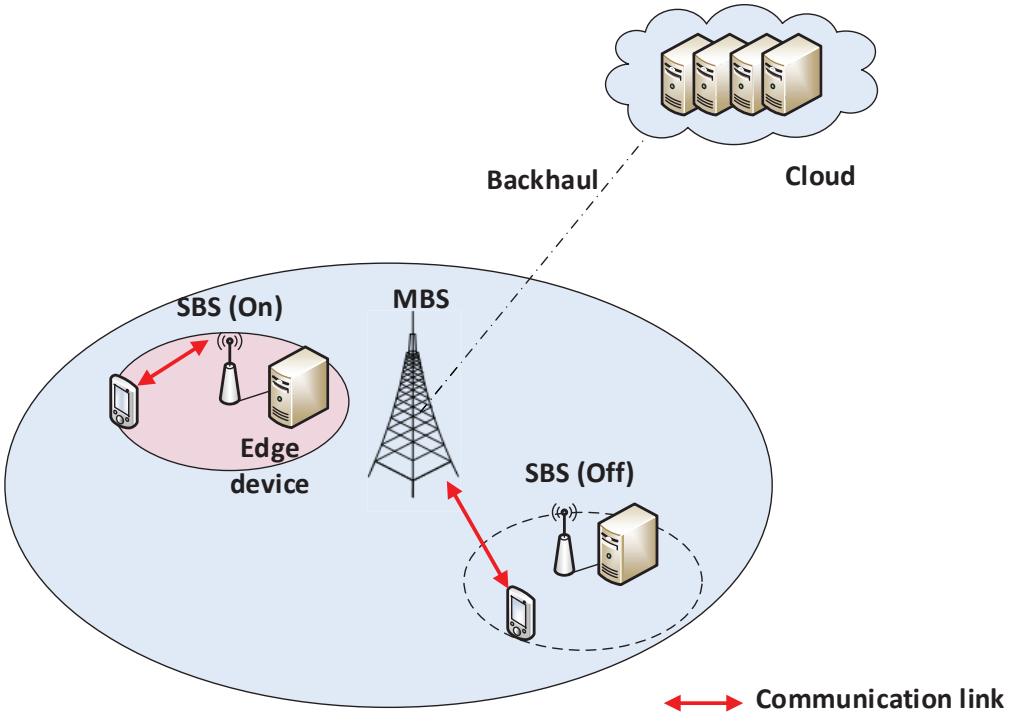


Figure 3.3: Proposed SBS sleeping in the H-CRAN-CE system layout.

As mentioned earlier, edge devices are assumed to coexist with SBSs and follow the same On/Off operation. Also, computing tasks associated with a sleeping SBS are offloaded to the cloud. However, offloading tasks from the edge to the central cloud can increase the

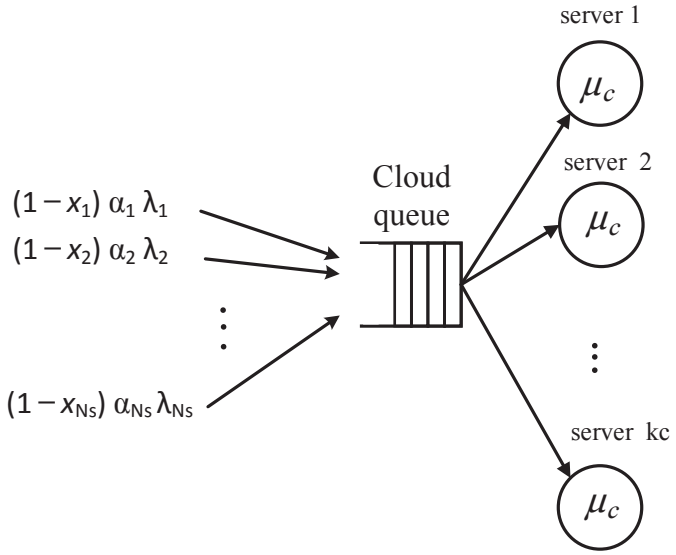


Figure 3.4: Cloud queue model for the proposed SBS sleeping mechanism.

time delay experienced by users; thus, it is essential to take into account the loading effect of computing tasks before deciding on whether to put an SBS in "On" or "Off" mode. Figs. 3.3 and 3.4 illustrate the system and queue models for the proposed sleeping mechanism. The computing-aware SBS sleeping mechanism is formulated as a 0-1 knapsack problem which in general aims to optimize the total value under the total weight constraint. Here, the arrival rate of tasks at SBS j (λ_j), which directly affects the utilization of the SBS $\rho_s = \frac{\lambda_j}{k_s \mu_j}$, is considered as the weight of the SBS. Let $\alpha'_j = 1 - \alpha_j$ denote the ratio of non-computing tasks to all incoming tasks at SBS j, then $(\alpha'_j \lambda_j)$ is considered the value of that SBS. The values of α follows a continuous uniform distribution between 0 and 1. In other words, the objective of the problem is to maximize the number of sleeping SBSs that have less computing duties as follows:

$$
\begin{aligned}
\textbf{P1:}\quad & \underset{x'_j}{\textbf{Maximize:}} \quad \sum_{j=1}^{N_s} \alpha'_j \lambda_j x'_j \\
& \textbf{Subject to:}\ C1: \ \sum_{j}^{N_s} \frac{\lambda_j x'_j + \lambda_m}{k_m \mu_m} \leq 1, \\
& \qquad\qquad C2: \ E[T_c] < \theta_c, \\
& \qquad\qquad C3: \ d_t < \theta_t, \forall t, \\
& \qquad\qquad C4: \ x'_j \in \{0,1\}, \quad \forall j \in \mathcal{S}, \\
& \qquad\qquad C5: \ \alpha'_j \in [0,1], \quad \forall j \in \mathcal{S}.
\end{aligned}
\tag{3.12}
$$

It is worth mentioning that $x'_j = 1$ indicates that the jth SBS is Off and all computing tasks associated with that SBS are offloaded to the cloud through the MBS. The rationale behind this optimization problem which is solved using dynamic programming, is that SBSs with less computing duties (i.e., more non-computing tasks or higher α'_j) are considered to have higher values compared to other SBSs and thus put into the Off mode by setting $x'_j = 1$. In other words, the optimization problem aims to maximize the amount of tasks offloaded from the SBSs to the MBS with minimum burden on cloud servers. In the general dynamic programming problem, we have a number of items (SBSs) each with an associated weight and value (benefit or profit). The objective is to fill the knapsack (MBS) with items such that we have a maximum profit (less computing tasks to reduce the burdens on the cloud) without crossing the weight limit of the knapsack. The constraint C1 ensures that the total incoming tasks at the MBS (offloaded tasks plus MBS tasks) will not exceed the MBS utilization limit (i.e., $\rho_m = 1$) which is considered the total weight limit in the 0-1 knapsack problem. C2 and C3 set the upper time limit for the cloud response and task completion, respectively. C4 indicates that x'_j is a binary variable. C5 shows that α'_j can have any real value from 0 to 1.

It should be noted that the average arrival rates of tasks at all SBSs underlying an MBS are used to drive the base station sleeping mechanism rather than the instantaneous number of tasks. This is because the sleeping mechanism in this work is centralized compared to other

distributed sleeping schemes such as the N-policy in [107] that allows SBSs to individually decide on the sleep decision based on the instantaneous number of tasks.

3.4 SBS Sleeping in Shared Cloud-Edge Computing System

In the shared cloud-edge computing model, both cloud and edge servers cooperate in a sense that allows the workload to be shared among all available cloud and edge servers provided that at SBSs operate in the On mode. In other words, the queueing model of the shared computing systems has one queue and joint cloud-edge servers as elaborated in Figs. 3.5 and 3.6. In the proposed shared computing system, the arrival rate and server utilization are respectively written as:

$$\lambda_{sh} = \sum_{j=1}^{N_s} \lambda_j \alpha_j, \tag{3.13}$$

$$\rho_{sh} = \frac{\lambda_{sh}}{k_{sh}\mu_{sh}}, \tag{3.14}$$

where k_{sh} is the total number of cloud and active edge servers ($k_{sh} = k_c + \sum_j^{N_s} x_j k_s$).

Lemma 1. *The cloud response time in the proposed shared computing model is faster than central cloud system by a factor of* $\left(1 + \frac{\sum_j^{N_s} x_j k_s}{k_c}\right)$.

Proof.

$$\frac{E[T_c]}{E[T_{sh}]} = \frac{\frac{1}{\lambda_c}\frac{\rho_c}{(1-\rho_c)}P_Q^c + \frac{1}{\mu_c}}{\frac{1}{\lambda_{sh}}\frac{\rho_{sh}}{(1-\rho_{sh})}P_Q^{sh} + \frac{1}{\mu_{sh}}}, \tag{3.15}$$

where $E[T_{sh}]$ and P_Q^{sh} denote the response time and queuing probability for the shared computing system, respectively. By comparing the two systems at the same load level and queuing probability (i.e., $\rho_c = \rho_{sh} = \rho$, and $P_Q^c = P_Q^{sh} = P_Q$), we get

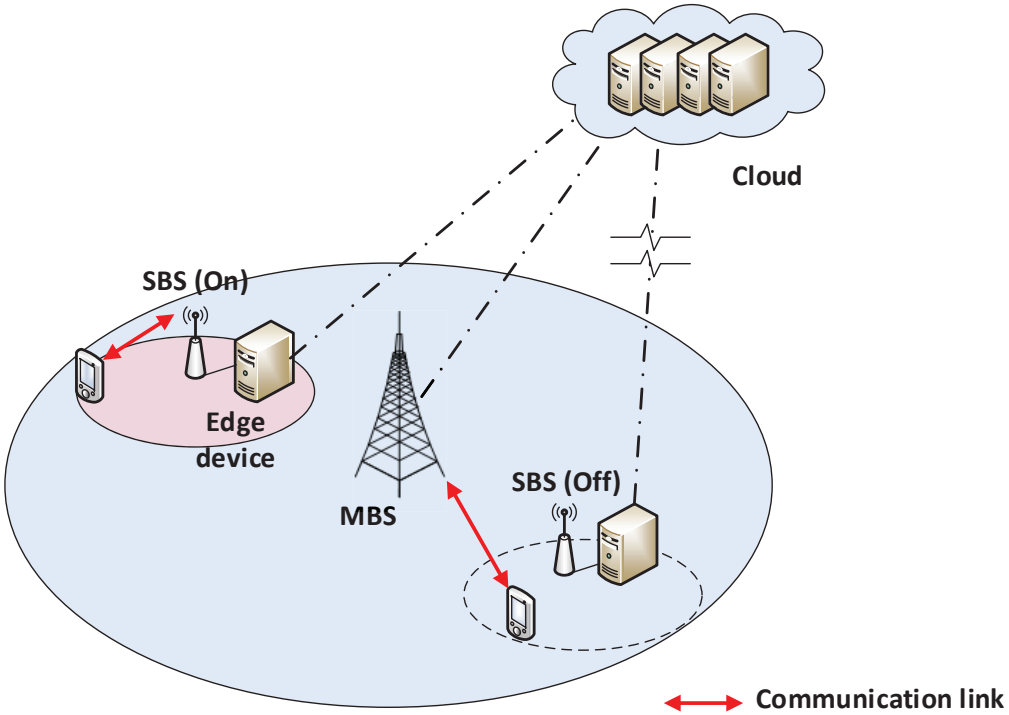


Figure 3.5: Shared computing system layout.

$$\begin{aligned}\frac{E[T_c]}{E[T_{sh}]} &= \frac{\frac{1}{\lambda_c}\frac{\rho}{(1-\rho)}P_Q + \frac{1}{\mu_c}}{\frac{1}{\lambda_{sh}}\frac{\rho}{(1-\rho)}P_Q + \frac{1}{\mu_{sh}}} \\ &= \frac{\frac{1}{\lambda_c}\rho P_Q + \frac{(1-\rho)}{\mu_c}}{\frac{1}{\lambda_{sh}}\rho P_Q + \frac{(1-\rho)}{\mu_{sh}}}\end{aligned}$$

where the last step is obtained by multiplying both numerator and denominator with $(1-\rho)$. When the system is heavily loaded (i.e., $\rho \approx 1$ and $P_Q \approx 1$), then the arrival rate equals the service rate (i.e., $\lambda_{sh} = \mu_{sh}$ and $\lambda_c = \mu_c$), thus

$$\frac{E[T_c]}{E[T_{sh}]} = \frac{\lambda_{sh}}{\lambda_c} = \frac{k_{sh}\mu_{sh}}{k_c\mu_c} = \frac{(k_c\mu_{sh} + \sum_j^{N_s} k_s\mu_{sh}x_j)}{k_c\mu_c}.$$

Assuming that the service rate of both systems are equal (i.e., $\mu_{sh} = \mu_c$), then

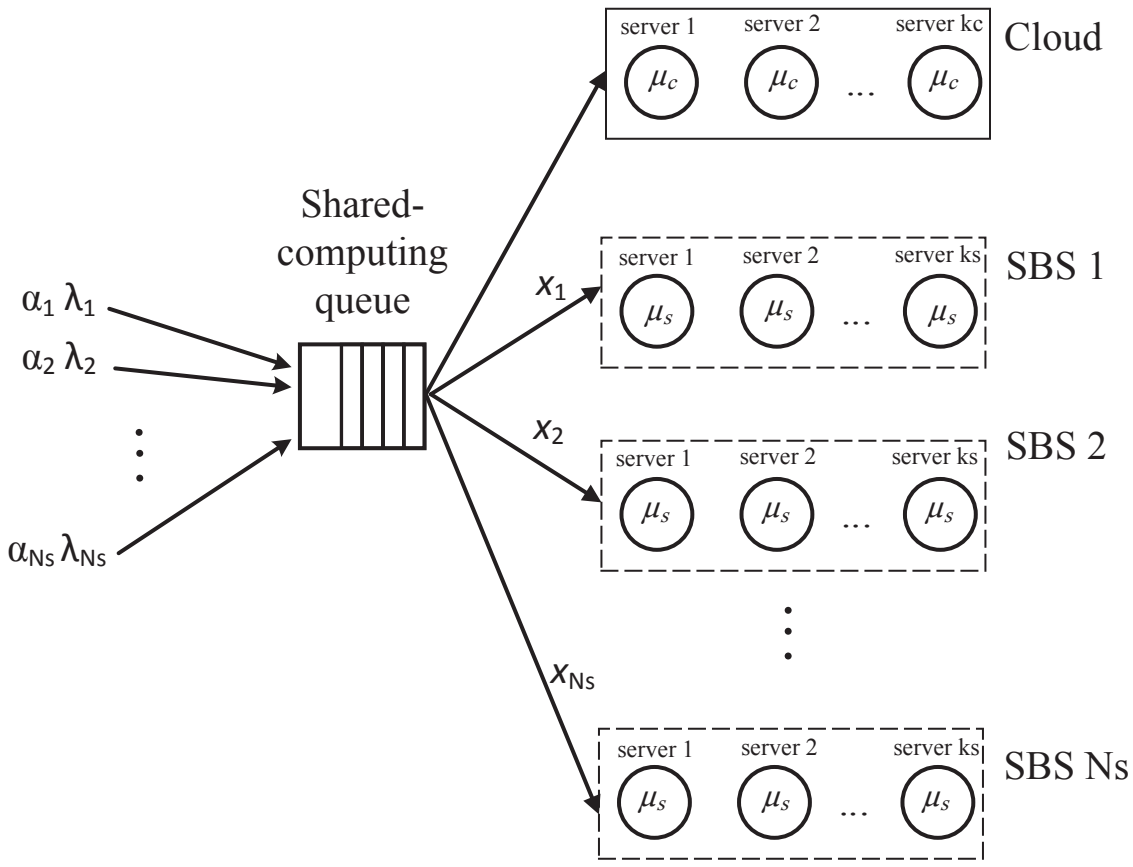


Figure 3.6: Shared computing queue model.

$$E[T_c] = \left(1 + \frac{\sum_j^{N_s} x_j k_s}{k_c}\right) E[T_{sh}].$$

□

To find the optimal set of sleeping SBSs taking into account power consumption and cloud response time, the cost function is formulated as:

$$C(\mathbf{x}) = \beta\left(\frac{P_t(\mathbf{x})}{\max\{P_t\}}\right) + (1-\beta)\left(\frac{E[T_{sh}(\mathbf{x})]}{\max\{E[T_{sh}]\}}\right) \tag{3.16}$$

where $\mathbf{x} = \{x_1, x_2, ..., x_{N_s}\}$ represents the operation status of all SBS in the system. Moreover, $\mathbf{x}$ has 2^{N_s} different combinations of binary numbers in a truth table style. Whereas

$C(\mathbf{x})$, P_t, and $E[T_{sh}]$ are respectively the cost, total power consumption, and response time associated with $\mathbf{x}$. β is a weighting factor that determines whether to prioritize the minimization of power consumption or cloud response time. Thus, the problem of SBS sleeping in the shared computing model can be formulated as:

$$\begin{aligned} \textbf{P2:} \quad & \underset{\mathbf{x}}{\textbf{Minimize:}} \quad \mathbf{C}(\mathbf{x}) \\ & \textbf{Subject to:} \quad C1: \ \sum_{j}^{N_s} \frac{\lambda_j x'_j + \lambda_m}{k_m \mu_m} \leq 1, \\ & \qquad\qquad\quad C2: \ E[T_c] < \theta_c, \\ & \qquad\qquad\quad C3: \ d_t < \theta_t, \forall t, \\ & \qquad\qquad\quad C4: \ x'_j \in \{0,1\}, \quad \forall j \in \mathcal{S}. \end{aligned} \tag{3.17}$$

The constraint C1 ensures that the total incoming tasks at the MBS do not exceed the MBS utilization limit. C2 and C3 set the time threshold for the cloud response and task completion, respectively. C4 indicates the On/Off operation of SBS j. To solve this mixed-integer optimization problem, exhaustive search which has been successfully used to find the optimal solution in similar problems [103] will be used. The optimal solution for this problem is obtained by testing 2^{N_s} different combinations of $\mathbf{x}$. For instance, if the system contains two SBSs, then four iterations will be conducted to test all possible SBS configurations 00, 01, 10, and 11, where these two digits represent the operation mode for each SBS. Therefore, when more SBSs exist in the system, the number of iterations to find the optimal solution will increase. Algorithm 1 illustrates the solution search strategy, where $\mathbf{x}^*$ represents the optimal operation for the SBSs under consideration.

3.5 Simulation Setup and Results

To evaluate the performance of the proposed computing-aware sleeping mechanism, simulation setup and results are provided and elaborated in this section. Table 3.2 lists the description, notation, and value for each parameter used in the simulation. The data size

Algorithm 1: Searching optimal solution for **P2**

Initialize $\mathbf{x_n}$, $n = 1, 2, ..., 2^{N_s}$;

while $n < 2^{N_s}$ **do**

  Calculate ρ_m according to (3.4);

  if $\sum_j^{N_s} \frac{\lambda_j x'_j + \lambda_m}{k_m \mu_m} \leq 1 - \rho_m$ **then**

    Calculate $\mathbf{C}(\mathbf{x_n})$ according to (3.16);

  end

  $n \leftarrow n + 1$;

end

$\mathbf{x}^* = \underset{\mathbf{x_n}}{\text{argmin}}\{\mathbf{C}\}$.

and task completion deadline are uniformly distributed between [0.5-2] MB and [2-4] sec, respectively. Following the work in [124], the total provided cellular throughput is 27 Mbps by the MBS and 15 Mbps by the SBS. It should be noted that the data rates here are shared by all users such that when the number of users increases, the per-user data rate will decrease. Other parameter settings are inspired by [127] and [128].

Figs. 3.7 and 3.8 show the system performance using both minimum load and computing-aware mechanisms. The minimum load approach is greedy-based and controls SBS sleeping according to only the SBS load (i.e., arrival rate) without considering the computing demand. On the other hand, the computing-aware mechanism determines the sets of active and sleeping SBSs considering both the arrival rate and the amount of computing tasks. It can be seen in Fig. 3.7 how θ_c affects the SBS power saving since it acts as a constraint on the cloud response time and thus the number of sleeping SBSs. Moreover, the computing-aware mechanism is found to achieve better power saving since it considers the computing load when selecting the sleeping SBSs and that also leads to reduced cloud response time.

It is also clear how the MBS utilization (ρ_m) significantly impacts the overall performance. When the MBS is lightly loaded (e.g., $\rho_m = 0.1$), both power saving and response time were

Table 3.2: Simulation Parameters

Description	**Notation**	**Value**
SBS power	P_s	50 W
Number of SBSs	N_s	10
Number of servers in the MBS	k_m	100
Number of servers in the edge device	k_s	10
Number of servers in the cloud	k_c	50
Task arrival rate at MBS	λ_m	1-100 task/sec
Task arrival rate at the jth SBS	λ_j	1-10 task/sec
Computing ratio at the jth SBS	α_j	[0,1]
MBS service rate	μ_m	1 task/sec
SBS service rate	μ_s	1 task/sec
Cloud service rate	μ_c	1 task/sec
CPU clock speed of each VM in the edge device	v_e	3.2 GHz
CPU clock speed of each VM in the cloud	v_c	3.2 GHz
Total bit rate provided by the SBS	b_s	15 Mbps
Total bit rate provided by the MBS	b_m	27 Mbps
Fiber backhaul link speed	b_{fl}	10 Gbps
Data size of tasks	S_t	0.5-2 MB
User transmit power to the SBS	p_s	0.05 W
User transmit power to the MBS	p_m	0.5 W

found to achieve higher values since more MBS servers are free and willing to accept more offloaded tasks from more sleeping SBSs. Nevertheless, having more sleeping SBSs increased the cloud response time because more computing tasks are directed to the central cloud instead of being processed locally by edge devices. In contrast, when the MBS is heavily

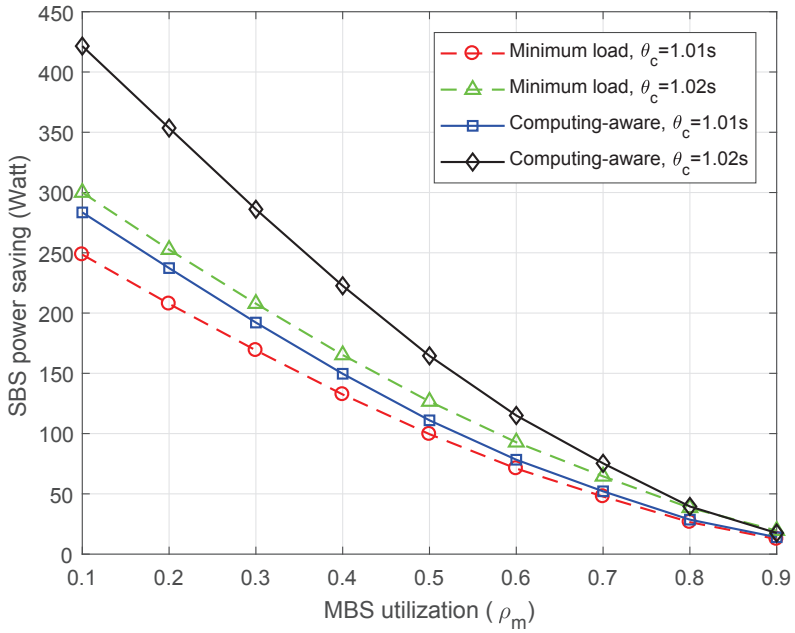


Figure 3.7: SBS power saving under different values of θ_c in disjoint cloud-edge system, $\bar{\theta}_t = 3s$.

loaded (e.g., $\rho_m = 0.9$), both power saving and cloud response time are decreased since SBSs have smaller opportunities to enter the sleep mode; as a result, less computing tasks are offloaded to the central cloud. The response time in this system does not fall below $1s$ since the proposed service rate in the cloud (μ_c) is set to $1s$ and thus the service time is $1/\mu_c = 1s$. This service time in addition to the cloud queue delay constitutes the cloud response time as shown in (3.9). Also, it should be noted that the task completion deadline θ_t follows a uniform distribution between 2 and 4 seconds (i.e. mean $\bar{\theta}_t$=3s).

Fig. 3.8 shows that having more power saving due to SBS sleeping will oblige users to spend more energy since tasks will be forwarded to the longer cloud path via the MBS rather than being processed by the edge device. Moreover, the non-linearity in the user energy consumption comes as a result of the more abundant frequency resources offered by the MBS when it is lightly loaded which is reflected by the higher SBS power saving. As a result, less time will be required to complete tasks since higher data rates will be offered by the MBS, and thus less energy consumption.

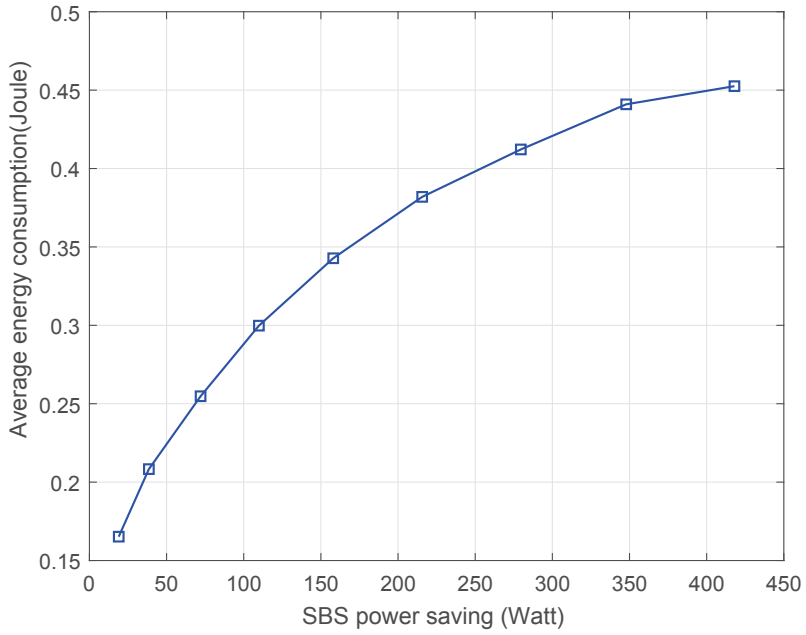


Figure 3.8: User energy consumption under different values of θ_c in disjoint cloud-edge system, $\bar{\theta}_t = 3s$.

Table 3.3: Solution Search Time

Number of SBSs (N_s)	**Time to find x^***
5	0.015 s
10	1.25 s
15	75 s
16	165 s
17	360 s
18	800 s

In Figs. 3.9 and 3.10, comparisons between disjoint and shared computing cloud-edge systems are conducted. As proved in Lemma 1, having more active SBSs in the system

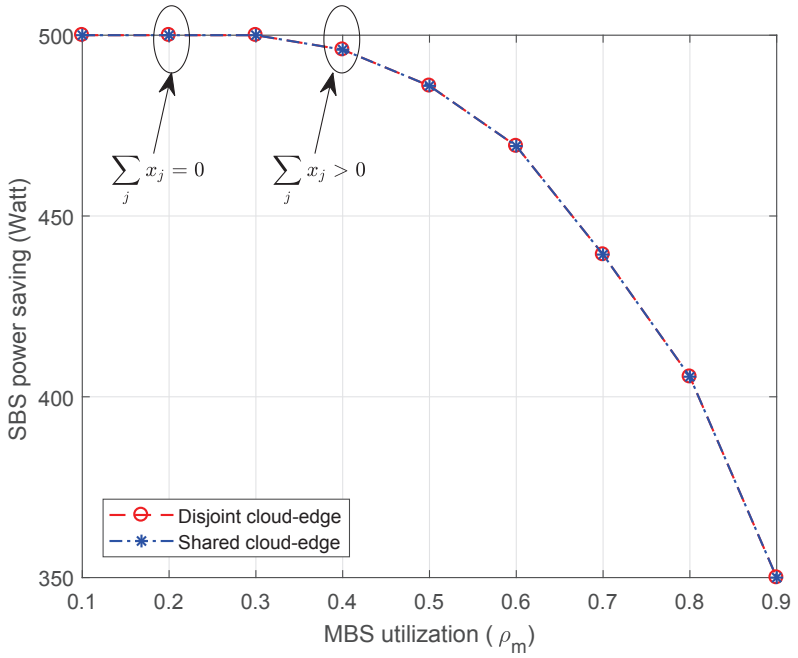


Figure 3.9: Comparing the SBS power saving between disjoint and shared computing systems.

reduces the cloud response time. This can be observed especially when ρ_m is high which forces more SBSs to remain active thus reducing the overall response time. To maintain fair evaluation, comparisons in Figs. 3.9 and 3.10 were obtained using only the minimum load approach without imposing a delay constraint on the cloud response time nor having a task completion deadline, and that is why they seem to have different shapes compared to other results in this section.

The optimal sleeping solution in the shared computing system is shown in Fig. 3.11. By adjusting the value of the weighting factor β, preference can be given to either saving energy or reducing the response time. Here, when $\beta = 0.8$ more emphasis is put on power saving than reducing cloud response time. Furthermore, adding more stringent requirements such as θ_c on the cloud response time and θ_t on the task completion time will affect the power saving significantly as seen in Figs. 3.12 and 3.13.

Finding the optimal solution requires searching all possible operation modes for all SBS

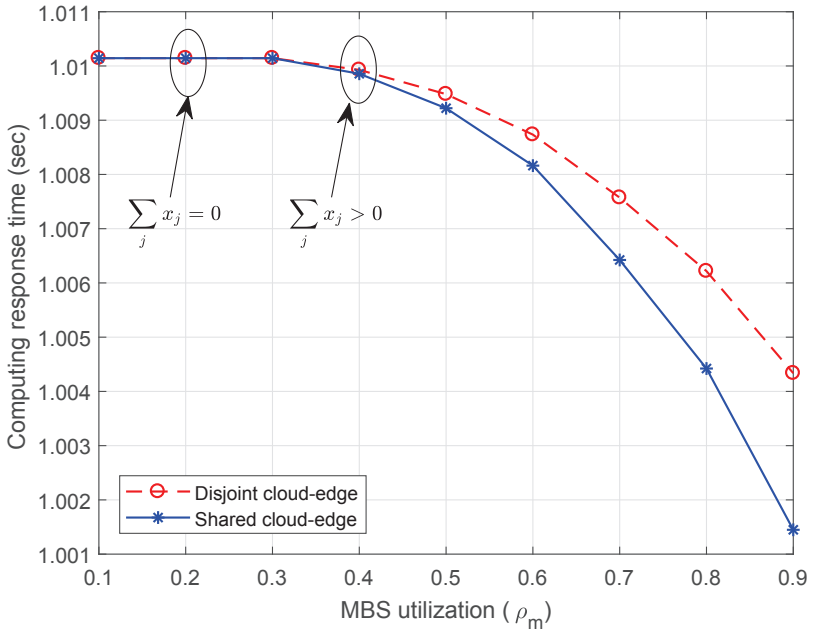


Figure 3.10: Comparing the cloud response time between disjoint and shared computing systems.

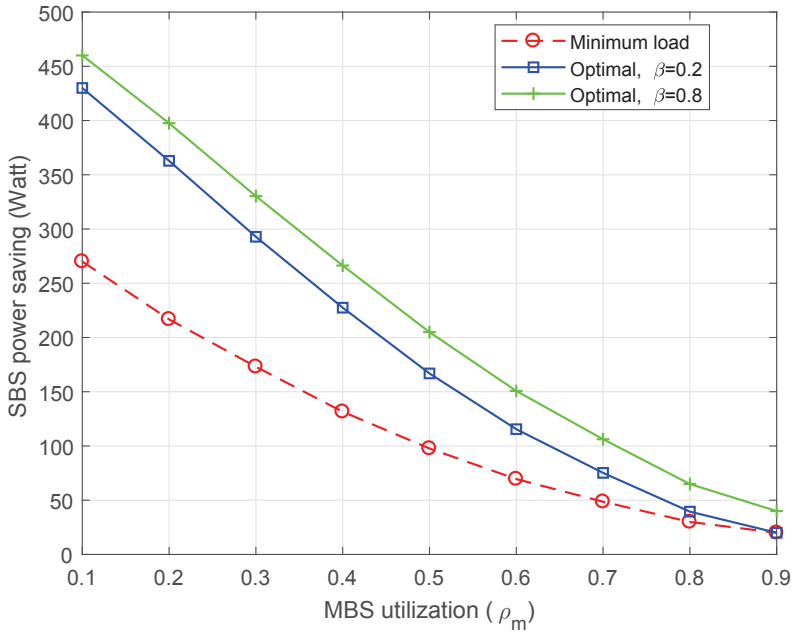


Figure 3.11: SBS Power saving in the shared cloud-edge system using different values of β.

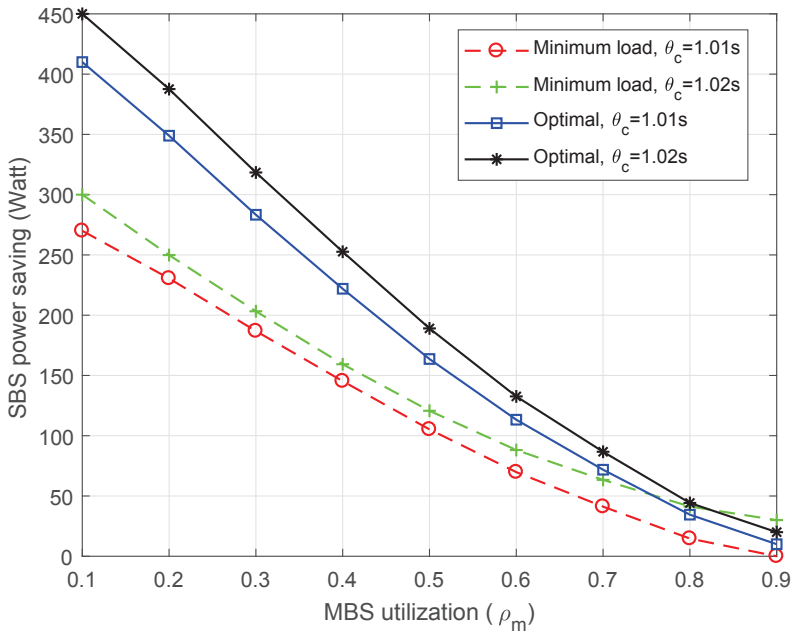


Figure 3.12: SBS Power saving in the shared cloud-edge system under different cloud response constraints, ($\bar{\theta}_t = 3s$, $\beta = 0.8$).

within the MBS coverage. Since there are only two operation modes, the total number of required iterations is 2^{N_s} as illustrated in Algorithm 1. Here, it is helpful to measure the time required to find the optimal solution for different numbers of SBSs although 10 SBSs were considered in this work. Table 3.3 lists the time required to reach the optimal solution.

It can be observed that finding the optimal solution requires longer time as the number of SBSs underlying an MBS increases; in which case, the search space can be reduced by clustering SBSs into smaller sets or assigning particular SBSs a fixed mode of operation using network and cloud characteristics to get a sub-optimal solution with reduced complexity [126]. Furthermore, reinforced learning techniques can be implemented to extract the features of user behaviour to help decide on each particular SBS operation and hence reduce the search space.

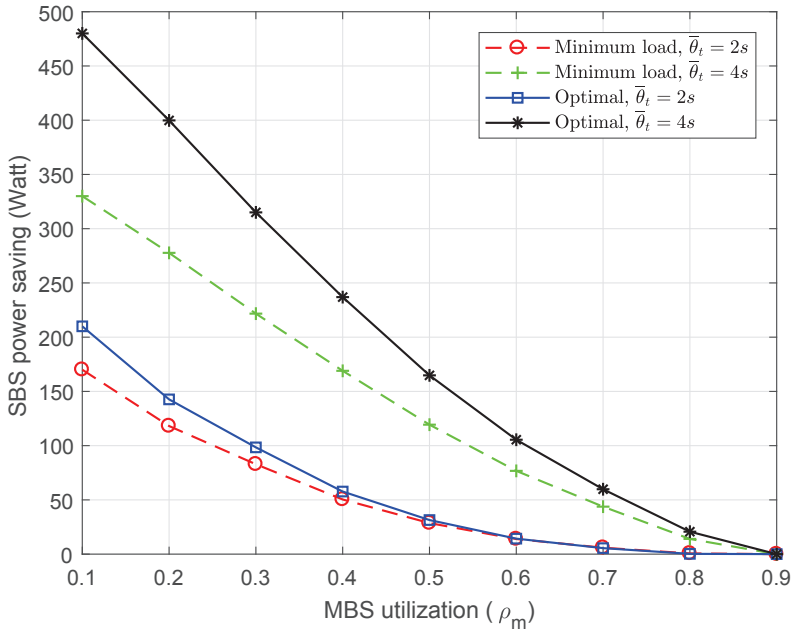


Figure 3.13: SBS power saving in the shared cloud-edge system under different task completion deadlines, ($\theta_c = 1.02s$, $\beta = 0.8$).

3.6 Chapter Summary

The problem of SBS sleeping in integrated H-CRAN-cloud-edge networks has been studied in this chapter. First, a SBS sleeping mechanism was proposed to save energy taking into account the constraints of task completion deadline and cloud response time. The problem was formulated as a 0-1 knapsack problem and solved using dynamic programming. Secondly, a joint cloud-edge computing model was introduced such that edge devices contribute to the total network computing resources beside the cloud to improve the system computing capability. Finally, finding the optimal power saving in the later system was found using an exhaustive search strategy. Abiding by the fact that traffic associated with sleeping SBSs will be eventually served by the MBS, the MBS utilization was considered as a major practical constraint that defines the observations and results obtained in this Chapter.

In Chapter 4, we will focus on achieving energy saving on edge devices taking into account the QoE requirements of mobile users. Two approaches will be investigated and compared,

namely, full- and partial-sleep modes depending on whether all or some virtual machines at an edge device are allowed to enter the sleep mode. The system model considers the existence of an SDN-based controller similar to the BBU pool used in this chapter; however, in the next chapter the SDN controller is in charge of only a group of cooperative edge devices rather than the entire network.

Chapter 4

QoS-aware Energy Saving Scheme for SDN-assisted Edge Computing Networks

4.1 Introduction

With the rapid growth of smart applications that help to promote autonomy, safety, and precision in a variety of life aspects, more and more mobile devices are continuously joining the existing networks. Large number of these devices came as a result of the increasing popularity of IoT that has attracted both industry and academia recently. Moreover, machine-type communication that forms a large proportion of IoT applications is expected to occupy 45% of the entire Internet traffic by 2022 [129], and that necessitates the efficient exploitation of the limited frequency and computing resources to provide satisfactory QoS for end users. However, many challenges are still facing the establishment of smart IoT systems especially in regard to energy, connectivity, and latency. The computing-intensive tasks of many smart applications require larger amounts of processing (CPU), memory, and battery capabilities compared to the on-device resources [130]. However, with the emerging edge computing networks, users will have the opportunity to offload their tasks for processing at the edge de-

vices benefitting the powerful computing resources. As a result, more devices can be served with better QoS, and that helps to relieve cloud servers and to enhance network scalability by running computing tasks on the small-size agile edge devices near mobile users.

Table 4.1: Related Works

Reference	Computing Paradigm	Objective	Approach
[131]	Mobile edge computing	Energy saving	Data compression, and resource allocation
[132]	Multilayered fog computing	Energy saving and processing delay	Partial task offloading among fog nodes
[133]	Cooperative computing in wireless sensor networks	Energy saving	Task partitioning and offloading among nodes
[134]	Content caching on edge nodes	Energy saving	Cooperative content caching
[115]	Fog-cloud computing	Reduce delay and energy consumption	Fog-cloud workload allocation
[135]	Virtualized edge computing for wireless sensor networks	Energy saving	Turning off camera nodes during inactivity
[136]	Partial SDN deployment	Energy saving	Shutting down unnecessary network elements
[137]	SND/OP backbone networks	Energy saving	SDN-assisted IP routing
[4]	SDN-based data centers	Energy saving	Overbooking computing resources
Our work	SDN-assisted cooperative edge computing	Energy saving	QoS-aware on/off operation of edge devices/VMs

It is not hard to realize that energy acts as a major challenge in modern communication and computing networks due to the massive amounts of computing-intensive applications that require dense deployment of base stations and cloud services. Herein, the under-utilization of available resources incurs large amounts of energy wastage in both computing and communication facilities. For instance, the average utilization of servers in large data centers ranges between 10 to 30 % [4], and the average link utilization in the backbone network of one large service provider has a utilization that does not exceed 40% whereas the energy consumption of that link is 95% of the fully loaded link [136]. This necessitates the adoption of effective energy saving strategies to reduce energy wastage and greenhouse gas emissions [110]. One of the efficient energy saving strategies is to monitor the traffic flow associated with base stations and edge devices, and set lightly loaded base stations or edge devices into the off/sleep mode. However, without careful implementation of such strategies, more service delay can be experienced by end-users and that can have serious impact on delay-sensitive applications such as e-health and autonomous vehicles which can tolerate few milliseconds of delay [6].

The under-utilization of computing resources leads to large amounts of energy wastage. Implementing SDN-assisted routing schemes to shut down under-utilized links and hardware components was proposed in several studies such as [137], [136], [138], and [139]. Authors in [4] proposed overbooking network resources by allocating more requests on the same resources to avoid resource and energy wastage. A hierarchical edge-cloud architecture was proposed in [121] to achieve workload balancing between the cloud and edge devices. In a similar context, authors in [115] considered workload scheduling between cloud and fog nodes to find the optimal energy-delay tradeoff. A scheduling algorithm was proposed in [122] to minimize the queue delay in cloud computing in order to reduce latency for Internet services.

From the communication perspectives, turning off base stations with light load to save energy has been considered in several studies such as [107] and [8] to save energy in heterogeneous networks since power amplifiers consume the largest portion of energy in cellular networks. The joint communication-computing aspects of SBS sleeping was studied in [140] and [141] to save energy in H-CRANs along with cloud-edge networks considering the computing delay experienced by users. In this work, we aim to save energy by turning off unused VMs in an SDN-assisted cooperative edge computing model where VMs of all edge devices constitute a shared pool of computing resources. Therefore, turning off VMs at any edge device must be governed by a strict queueing probability threshold since the on/off operation of VMs at any edge device affects the queueing delay experienced by mobile users. Table 5.1 summarizes the most relevant and recent works.

The rest of the Chapter is organized as follows. In Section 4.2, the system model is introduced. The problem formulation and solution approach in addition to a comparison between the full and partial sleeping modes are presented in Section 4.3. A load management strategy for overloaded edge devices under the fronthaul capacity constraint is proposed in Section 4.4. Simulation results and discussions are elaborated in Section 4.5. Finally, Section 4.6 provides a short summary for the Chapter.

4.2 System Model

In this section, the system, computing, and communication models are presented. Fig. 4.1 depicts the proposed joint communication-computing system model.

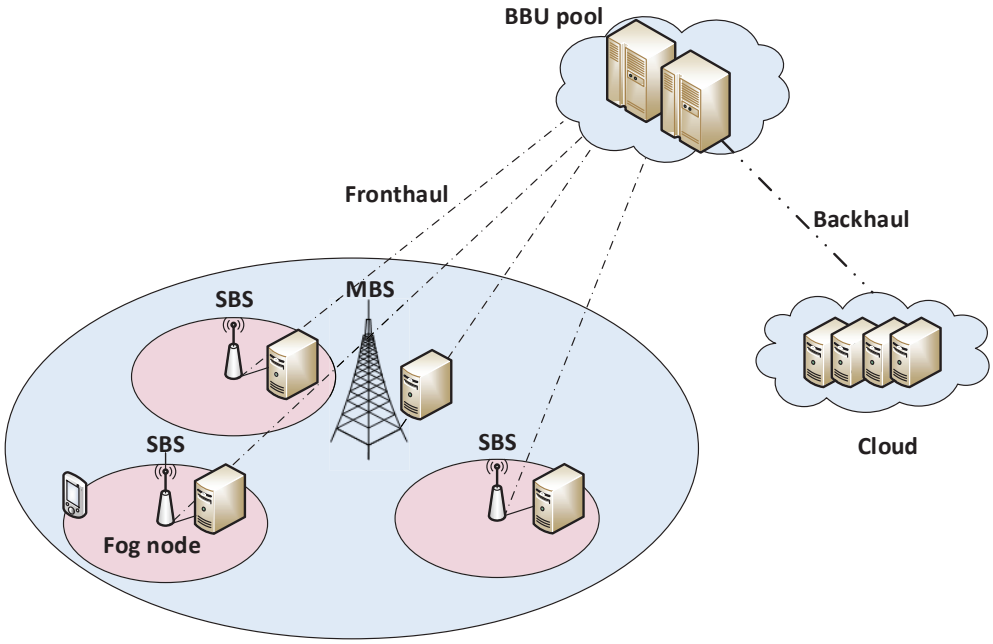


Figure 4.1: System model.

We consider a distributed computing network in which wireless links are provided by SBSs, whereas computing tasks are processed by edge devices located at the vicinity of mobile users. Supported by the SDN-based controller, which can be located in one of the edge devices (master device) or further in the BBU pool, we assume that traffic is monitored and can be rescheduled among the edge devices efficiently and easily. Sufficient orthogonal wireless channels are considered to be available with constant link speed since interference is neglected in this work.

Let $\mathcal{E} = \{1, 2, ..., N_e\}$, $\mathcal{V} = \{1, 2, ..., N_v\}$, and $\mathcal{U} = \{1, 2, ..., N_u\}$ denote the sets of edge devices, VMs per edge device, and users, respectively. Each edge device $n \in \mathcal{E}$ is modeled with the well-known M/M/k queueing system with k servers or VMs. For each edge device

n, let k_n, λ_n, and μ_n be the number of VMs, arrival rate, and departure (service) rate, respectively, then the utilization of the nth edge device ρ_n is expressed as

$$\rho_n = \frac{\lambda_n}{k_n \mu_n}. \tag{4.1}$$

The performance metric of the proposed system is the queue delay; therefore, we consider the queueing system steady state analysis based on the CTMC of M/M/k systems [142]. Here, the probability that a user will have to queue at edge device n (all servers are occupied) can be calculated as:

$$\begin{aligned} P_Q^n &= \sum_{i=k_n}^{\infty} \pi_i \\ &= \pi_0 \frac{k_n^{k_n}}{k_n!} \frac{\rho_n^{k_n}}{1-\rho_n}, \end{aligned} \tag{4.2}$$

where π_i represents the steady state probability that i servers are occupied. The steady state probability that zero task exists in edge device n, can be written as

$$\pi_0 = \left[\sum_{i=0}^{k_n-1} \frac{(k_n\rho_n)^i}{i!} + \frac{k_n^{k_n}}{k_n!} \frac{\rho_n^{k_n}}{1-\rho_n} \right]^{-1}. \tag{4.3}$$

Therefore, the queue delay of the nth edge device can be obtained by

$$E[T_Q^n] = \frac{1}{\lambda_n} \cdot \frac{\rho_n}{1-\rho_n} \cdot P_Q^n. \tag{4.4}$$

From the user perspective, the energy consumption of user u associated with edge device n depends on the end-to-end delay experienced by that user; that is to say, the sum of queue delay, task transmission over the wireless link, and task completion time as follows

$$d_u = E[T_Q^n] + \frac{D_u}{s_u} + \frac{D_u}{c_e}, \tag{4.5}$$

where D_u, s_u and c_e denote the task data size, wireless link speed, and edge device processing speed, respectively. Thus, the energy required to complete a task is

$$e_u^n = p_s d_u, \tag{4.6}$$

where p_s is the user transmit power.

4.3 Problem Formulation and Solution Approach

The proposed on/off scheme for edge computing aims to save energy while maintaining satisfactory service levels for end-users. Herein, energy saving schemes are proposed, namely, the full sleep mode in which the entire edge device with all VM resources enters the sleep mode, and partial sleep mode where the necessary amount of VMs according to the QoS requirements remain active while the rest enter the sleep mode locally within edge devices.

4.3.1 Full Sleep Mode

To maximize energy saving in distributed computing networks, we aim to allow more edge devices to enter the 'off' mode taking into account the probability of queueing at each edge device that must remain below a pre-determined value. In addition, the queue system stability must be maintained throughout the entire process. Let p_e denote the power consumed by an individual edge device, then the total power consumption of edge devices can be obtained by the following

$$p_T = \sum_{n=1}^{N_e} x_n p_e, \tag{4.7}$$

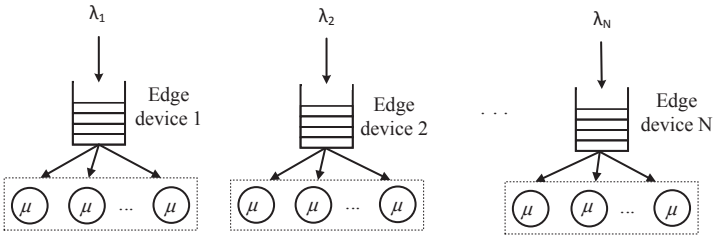
where $x_n \in \{0, 1\}$ is the on/off operator. From the aforementioned, the problem can be mathematically formulated as follows

$$\begin{aligned} \textbf{P1:} \quad & \underset{x_n}{\textbf{Minimize:}} \quad p_T \\ & \textbf{Subject to:} \quad C1: \ P_Q^n < \alpha, \quad \forall n \in \mathcal{E}, \\ & \qquad\qquad\quad C2: \ \rho_n < 1, \quad \forall n \in \mathcal{E}, \\ & \qquad\qquad\quad C3: \ x_n \in \{0,1\}, \quad \forall n \in \mathcal{E}. \end{aligned} \tag{4.8}$$

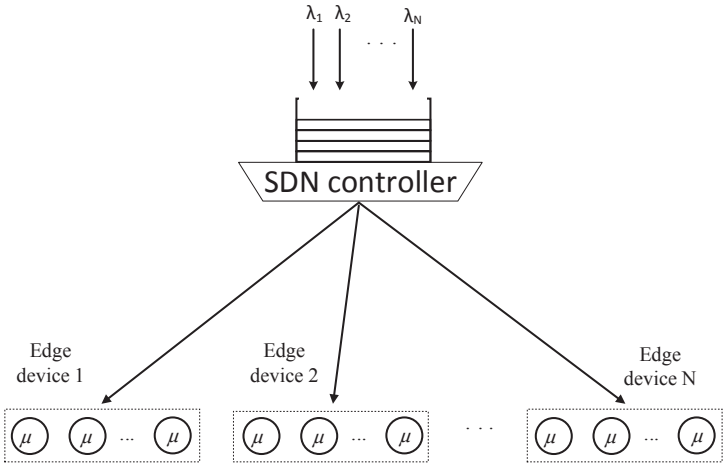
The constraint C1 ensures that the queueing probability at edge device n remains always below a predefined value α. C2 maintains system stability by ensuring that the utilization of each edge device remains below 1. In C3, the on/off operator x_n is presented, where $x_n = 1$ and $x_n = 0$ indicate that edge device n is set into the on and off modes, respectively. Since C1 and C2 in **P1** are non-linear on x_n, and C3 indicates that the decision variable x_n is a binary variable, the problem can be described as a binary integer non-linear programming and thus requires intensive computations to tackle. The challenge in this problem is that traffic offloaded from sleeping edge devices needs to be accommodated by active edge devices and that affects the queueing probability at the host edge device. Therefore, in order to solve the problem, we propose an SDN-assisted cooperative computing paradigm in which all VMs in all edge devices constitute a shared pool. Herein, all incoming traffic and edge resources are modeled as a single M/M/k queue. The goal of the proposed model is to achieve elastic control over the available computing resources among the cooperative edge devices. Fig. 4.2 depicts the SDN-assisted edge computing model.

In the SDN-assisted model, let the new variables $k_T = \sum_{n=1}^{E} k_n$, $\lambda_T = \sum_{n=1}^{E} \lambda_n$, and $\mu_T = \mu_n, \ \forall n \in \mathcal{E}$, denote the total number of VMs, total arrival rate, and the service rate of the controller queue, respectively. It should be noted that the computing capacity of the individual VM remains the same in the cooperative scheme and that explains the departure rate $\mu_T = \mu_n$.

To solve the power minimization problem, it is first required to find the optimal number of servers (i.e., VMs) that satisfies the queueing probability constraint. To this end, we propose the square-root staffing (SRS) rule [142] which requires the resource staffing in



(a) Traditional edge computing queue model.



(b) SDN-assisted queue model.

Figure 4.2: Proposed edge computing layout.

M/M/k queues to be greater than the resource requirement ($\frac{\lambda_T}{\mu_T}$) in order to satisfy both stability and delay requirements. Then, the total number of required VMs is subtracted from the total number of VMs to decide on the number of active and sleeping edge devices. The SRS rule in the proposed scheme is used to determine the number of edge devices that can enter the off mode without violating the desired service quality (i.e., low queueing probability). Let $R_T = \frac{\lambda_T}{\mu_T}$ denote the resource requirement at edge device n, then the optimal number of servers required to ensure a queueing probability less than α is [142]

$$k_T^* \approx R_T + c\sqrt{R_T}, \tag{4.9}$$

where c is the solution to the equation

$$\frac{c\Phi(c)}{\phi(c)} = \frac{1-\alpha}{\alpha}, \tag{4.10}$$

where $\Phi(.)$ and $\phi(.)$ denote the cumulative distribution and probability density functions of the standard normal distribution, respectively. From (4.10), the relation between α and c can be rewritten as follows:

$$\alpha = \left[1 + \frac{c\Phi(c)}{\phi(c)}\right]^{-1}. \tag{4.11}$$

The above function represents the Halfin-Whitt function which helps do determine the optimal number of servers required to maintain a queueing probability less than α in M/M/k queues [143]. Fig. 4.3 shows the Halfin-Whitt function according to (4.11). Hence, the number of edge devices that can be set into the off mode is obtained by:

$$n_s = \left\lfloor \frac{N_e k_n - k_T^*}{k_n} \right\rfloor, \tag{4.12}$$

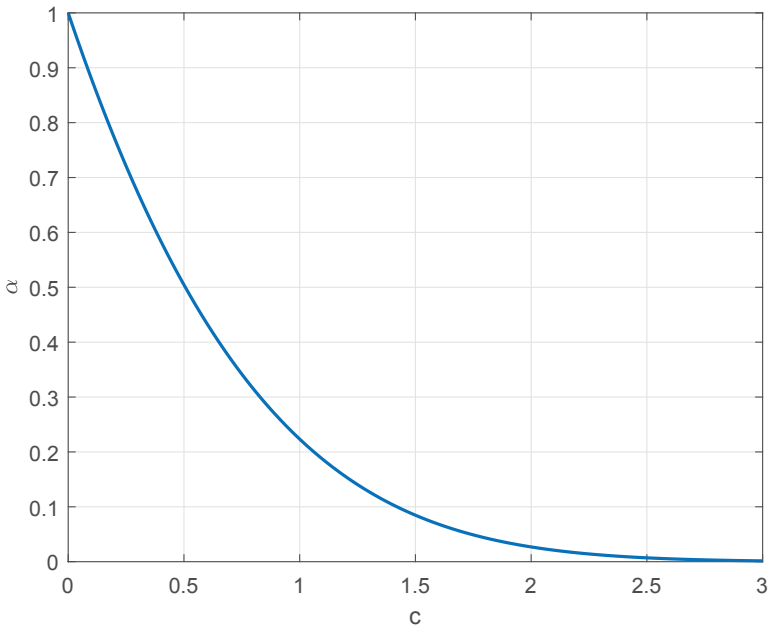


Figure 4.3: The Halfin-Whitt function showing the relationship between α and c.

Therefore, the problem **P1** is solved by finding c that satisfies α according to (4.10), then determining the required number of VMs using (4.9). Once n_s is obtained, edge devices

with the lightest traffic load will enter the off mode, while the traffic offloaded from the sleeping edge devices will be accommodated by active edge devices with the next lightest load. Algorithm 2 illustrates the proposed scheme.

Lemma 2. *Satisfying the queueing probability requirements also limits the queueing delay experienced by users.*

Proof. Since the SRS rule satisfies $P_Q^n < \alpha$, from (4.4) we can write :

$$\begin{aligned} E[T_Q^n]\frac{\lambda_n(1-\rho_n)}{\rho_n} &< \alpha, \\ E[T_Q^n] &< \tfrac{\alpha\rho_n}{\lambda_n(1-\rho_n)}, \\ \text{but} \quad \rho_n = \frac{\lambda_n}{k_n\mu_n}, E[T_Q^n] &< \tfrac{\alpha}{(k_n\mu_n-\lambda_n)}, \end{aligned}$$

where $k_n\mu_n > \lambda_n$ according to the SRS rule (refer to (4.9)). □

4.3.2 Partial Sleep Mode

Turning off the entire edge device can be a hard decision since large numbers of VMs are turned off at once; as a result, maintaining the QoS for users can be more challenging. Moreover, it increases the amount of migrated tasks that must be accommodated by other edge devices. Therefore, to improve the system performance in regard with energy saving, task migration, and flexibility, a partial sleep mode is proposed such that energy saving is achieved with fine granularity by considering the on/off operation of individual VMs instead of the entire edge device. In this model, VMs of all edge devices contribute to the SDN-controlled pool as in the previous model; however, the number of active VMs at each edge device can be different according to the traffic demands. In other words, each edge device takes part in energy saving, but all edge devices remain active. With the assistance of SDN, more agility can be achieved in regard with VM staffing and that helps reduce the delay

Algorithm 2: Proposed edge device sleeping scheme

Define λ_{off}: rate of tasks offloaded from sleeping edge devices;

Find k_T^* according to (4.9);

Find n_s according to (4.12);

Set $x_n = 1, \quad \forall n \in \mathcal{E}$;

$n \leftarrow 1$;

$\lambda_{off} \leftarrow 0$;

while $n \leq n_s$ **do**

  Find $\min\{\lambda_n\}, \quad \forall n \in \mathcal{E}$;

  $x_n \leftarrow 0$;

  $\mathcal{E} = \mathcal{E} \setminus n$;

  $\lambda_{off} \leftarrow \lambda_{off} + \lambda_n$;

  $n \leftarrow n + 1$;

end

Accommodating offloaded tasks:

Sort $\mathcal{E}$ in ascending order according to λ_n;

$i \leftarrow 1$;

while $\lambda_{off} > 0$ **do**

  Define $\lambda_q \subset \lambda_{off}$ such that $\frac{\lambda_i + \lambda_q}{k_n \mu_n} < 1$ to satisfy C2;

  $\lambda_i \leftarrow \lambda_i + \lambda_q$;

  $\lambda_{off} \leftarrow \lambda_{off} - \lambda_q$;

  $i \leftarrow i + 1$;

end

experienced by users and eliminates the need for additional load balancing processes. Let p_v denote the power consumed by an individual VM, then the power consumption per edge devices can be obtained by the following

$$p_n = \sum_{v=1}^{N_v} x_v^n p_v, \tag{4.13}$$

where $x_v^n \in \{0, 1\}$ is the on/off operator of VM v at edge device n. Thus, the total power consumed by edge devices is calculated as

$$p_T = \sum_{n=1}^{N_e} p_n. \tag{4.14}$$

From the aforementioned, the power minimization problem can be formulated as follows:

$$\begin{aligned} \textbf{P2:} \quad & \underset{x_v^n}{\textbf{Minimize:}} \quad p_T \\ & \textbf{Subject to:} \quad C1: \ P_Q^n < \alpha, \quad \forall n \in \mathcal{E}, \\ & \qquad\qquad\quad C2: \ \rho_n < 1, \quad \forall n \in \mathcal{E}, \\ & \qquad\qquad\quad C3: \ x_v^n \in \{0, 1\}, \quad \forall v \in \mathcal{V}, \forall n \in \mathcal{E}. \end{aligned} \tag{4.15}$$

It can be observed that **P2** is similar to **P1** except for C3 that provides control over VMs instead of edge devices. The solution to the problem is also pursued using the SRS and Halfin-Whitt function. Herein, a comparison between two computing paradigms is conducted, namely, the disjoint model where satisfying the condition $P_Q^n < \alpha$ depends on the VMs of an individual edge device. In other words, each edge device acts as a stand-alone entity such that $k_n^* \leq k_n$. The second is the SDN-assisted model where VM resources of edge devices form a shared pool and contribute towards the benefit of all edge devices such that $k_T^* \leq k_T$..

Lemma 3. *Carrying out partial (VM) sleeping can achieve more energy saving compared to the full edge device sleeping assuming the total edge device power is equally divided among VMs.*

Proof. To compare the obtainable energy saving at edge device n, consider a number u of users:

- In the disjoint model, when $0 < u < N_v$, energy saving is zero, whereas in the partial scheme $(N_v - u)p_v$ energy saving can be achieved, assuming on VM is required for each user.

- In the SDN-assisted model, the number of potential sleeping edge devices is $n_s = \left\lfloor \frac{N_e k_n - k_T^*}{k_n} \right\rfloor$, whereas in the partial model, the total number of sleeping VMs is $N_e k_n - k_T^*$.

This indicates that the partial energy saving scheme is more flexible and leads to better energy saving. □

4.4 Traffic Management in Overloaded Edge Devices

When the number of users exceeds that of VMs, the edge device is considered to be overloaded. In which case, some users have to be accommodated by another edge device. However, due to the different task sizes and deadline requirements of users, offloading tasks randomly without context awareness can result in a larger number of unsatisfied users. Moreover, with the dense deployment of small cells and edge devices in H-CRANs, the fronthaul traffic can be scaled up to multiple Gbps even under moderate mobile traffic, thus overwhelming the capacity fronthaul links [144], which is defined as the maximum sum data rate that can be allowed on the fronthaul link [145]. It should be noticed that fronthaul links connect the RRHs and SBSs with the SDN-based BBU pool, and can take the form of wired links such as fiber optic cables that provide large bandwidth but suffer inflexible and expensive installation, or wireless links that are less expensive but has smaller bandwidth [146]. Regardless of the fronthaul type, we consider the frothaul capacity in bits per second (bps) [147] as a constraint in transferring user tasks from their initial (host) edge device to other devices through the SDN-based controller. To this end, we aim to satisfy as many users as possible by optimizing the offloading decisions to meet the deadline requirements of users. Let d_u represent the delay experienced by a user u, which can be expressed as in (4.5) with the addition of the fronthaul link speed and the user association operator:

$$d_u = E[T_Q^n] + \frac{D_u}{s_u} + \frac{D_u}{c_e} + (1 - x_u^n)\frac{D_u}{F_u}, \tag{4.16}$$

where x_u^n is the user association operator such that $x_u^n = 1$ and $x_u^n = 0$ indicate respectively that user u is associated with the host edge device or another edge device (if the initial edge device is fully occupied). Now, let the variable l_u be the satisfaction indicator of a user u that is dependant on the delay experienced by the user compared to the task completion deadline as follows:

$$l_u = \begin{cases} 1, & \text{if } d_u \leq T_u, \\ 0, & \text{if } d_u > T_u, \end{cases} \tag{4.17}$$

where T_u is the task completion deadline of user u. Therefore, the problem can be formulated as follows

$$\begin{aligned} \textbf{P3:} \quad & \underset{x_u^n}{\textbf{Maximize:}} \quad \sum_{u=1}^{N_u} l_u \\ & \textbf{Subject to:} \quad C1: \ N_u^n \leq k_n, \quad \forall n \in \mathcal{E}, \\ & \qquad\qquad\quad C2: \ \sum_{u=1}^{N_u^n} F_u < F_n, \quad \forall n \in \mathcal{E}, \\ & \qquad\qquad\quad C3: \ x_u^n \in \{0, 1\}, \quad \forall v \in \mathcal{V}, \forall n \in \mathcal{E}. \end{aligned} \tag{4.18}$$

The objective function in **P3** aims to maximize the number of satisfied users within the cooperative edge group by associating users to either the initial (nearest) device or to another device through the SDN controller. Due to the limited amount of VMs, this problem is considered a multi-objective optimization problem since associating users without considering the distinct delay requirement of each mobile user, and the limited fronthaul capacity, can lead to larger number of unsatisfied users. The constraint C1 indicates that the number of users associated with an edge device n cannot exceed the number of VMs (k_n). In C2, the fronthaul bandwidth allocated to users associated with edge device n must not exceed the fronthaul capacity of that edge device. C3 presents the binary user association operator.

In the aforementioned problem, taking into account the fronthaul capacity and the different data size of tasks, allocating users randomly among edge devices can lead to some users exceeding their task completion deadline. When an edge device is overloaded, minimizing delay for one user by associating it with the local (overloaded) edge devices rather than transferring it to other devices can impact the delay experienced by other users due to C1 and C2. Hence, this multi-objective optimization problem requires optimizing each objective (user) while considering other users [20]. To solve the problem, a greedy-based heuristic algorithm is proposed following the Lexicographic method [148] where users are first sorted according to the $\frac{D_u}{T_u}$ ratio that ranks the importance of tasks according to their data size and deadline requirements. Accordingly, tasks with shorter deadlines and/or larger data sizes are ranked higher to avoid the task migration delay and to relieve fronthaul links. Afterwards, the user allocation process begins where users are allocated according to their ranking while the capacity constraints are updated after each step. It should also be noted that the problem only applies when the number of users outnumbers available VMs at the edge device. Algorithm 3 illustrates the process of handling overloaded edge devices.

4.5 Simulation and Results

To evaluate the performance of the proposed schemes, simulations were conducted to provide results regarding energy saving, queue delay, and per-user energy consumption. To ease tracking the simulation parameters, Table 4.2 provides a list of the used parameters.

The first comparison in Fig. 4.4 shows the percentage of power saving obtained using the full sleep mode. It can be observed that the smaller the value of α, the less energy saving is obtained since less queueing probability is enforced in the system. It can also be seen that when the arrival rate of users is increased, the overall energy saving is declined since more servers (VMs) are required to accommodate the incoming traffic. The queueing delay is depicted in Fig. 4.5 where the delay shows an increase when either or both the arrival rate and the value of α are increased.

Algorithm 3: Heuristic algorithm for **P3**

Define N_u^n: number of users hosted by edge device n;

Define F_n: fronthaul capacity at edge device n;

Find $\frac{D_u}{T_u} \quad \forall u \in \mathcal{U}$;

Sort $\frac{D_u}{T_u} \quad \forall u \in \mathcal{U}$ in ascending oredr;

Allocated top k_n users to the nth edge device;

Transfer the next $N_u^n - k_n$ to the SDN controller as follows:

$i \leftarrow k_n + 1$;

while $i < N_u^n$ **do**

> Allocate fronthaul resources according to user requirements (i.e., $F_u = \frac{D_u}{T_u}$);
>
> $F_n \leftarrow F_n - F_u$;
>
> $i \leftarrow i + 1$;

end

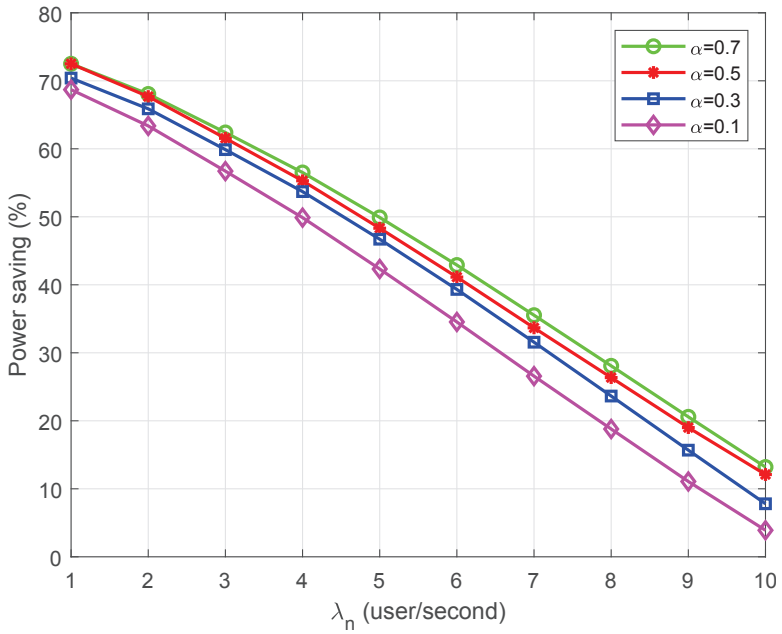


Figure 4.4: Power saving obtained by full edge device SDN-assisted sleeping scheme.

Table 4.2: Simulation Parameters

Description	**Value**
Number of edge devices (N_e)	10
Number of VMs per edge device (N_v)	10
Number of users associated with edge device (N_u)	$1 - 10$
Power consumption of an edge device (p_e)	50W
Power consumption of one VM (p_v)	5W
User transmit power (p_s)	0.05W
Arrival rate at each edge device (λ_n)	1-10 user/sec
Departure rate at each edge devices (μ_n)	1 user/sec
Data size of tasks (D_u)	1 MB
Wireless link speed for each user (s_u)	1 Mbps
Edge device processing speed (c_e)	1 Gbps
Fronthaul capacity of edge device n (F_n)	$0.5, 1, 2$ Gbps

As presented earlier, implementing the proposed partial energy saving scheme in a smaller granularity using VMs can achieve better energy saving, reduce traffic offloading, and improve system flexibility. Figs. 4.6 and 4.7 show respectively the amount of traffic offloaded with respect to power saving, and the comparison of power saving using the full and partial edge sleeping schemes.

The amount of energy saved using the partial sleep mode under different values of α is presented in Fig. 4.8. In the SDN-assisted model, more VMs can be set into the off mode since the SDN makes full use of all available VMs in the cooperative edge system. As a result more energy can be saved, and with higher values of α, more energy saving can be obtained at the cost of reduced service quality (i.e., higher queueing probability). On the other hand, the disjoint computing model, where VMs are provided to users from one edge device and

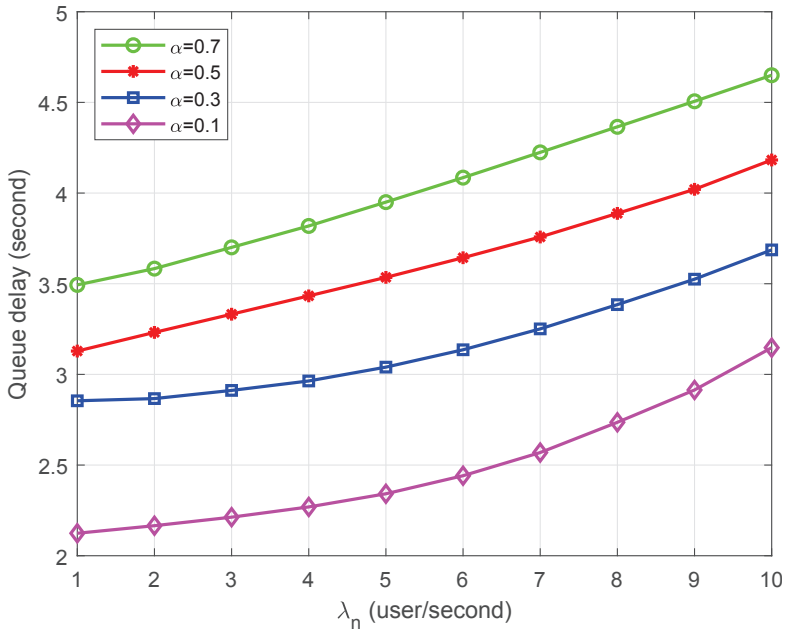


Figure 4.5: Queue delay experienced by users.

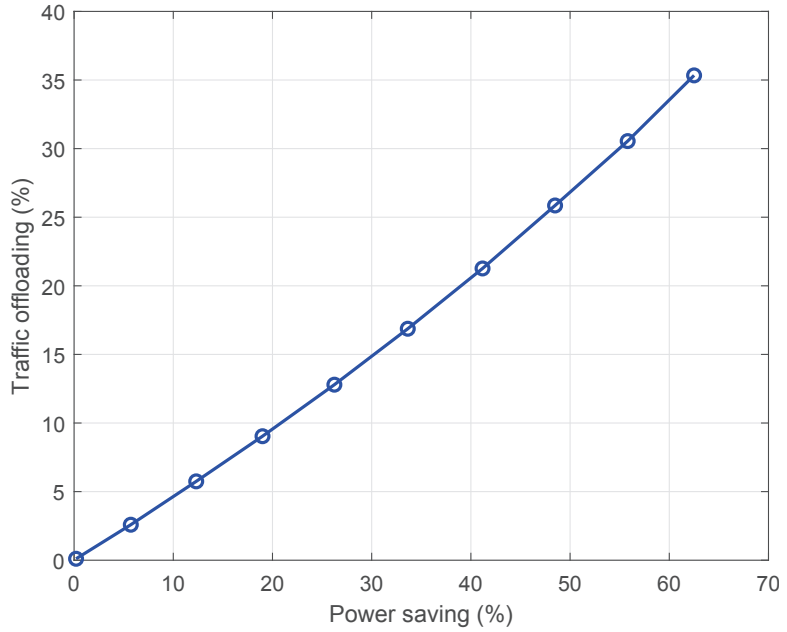


Figure 4.6: Traffic offloading due to full edge device sleep.

do not extend to other available resources in other devices, reduces the possibility of turning off VMs due to the limited amounts of VMs necessary to suffice the QoS requirements.

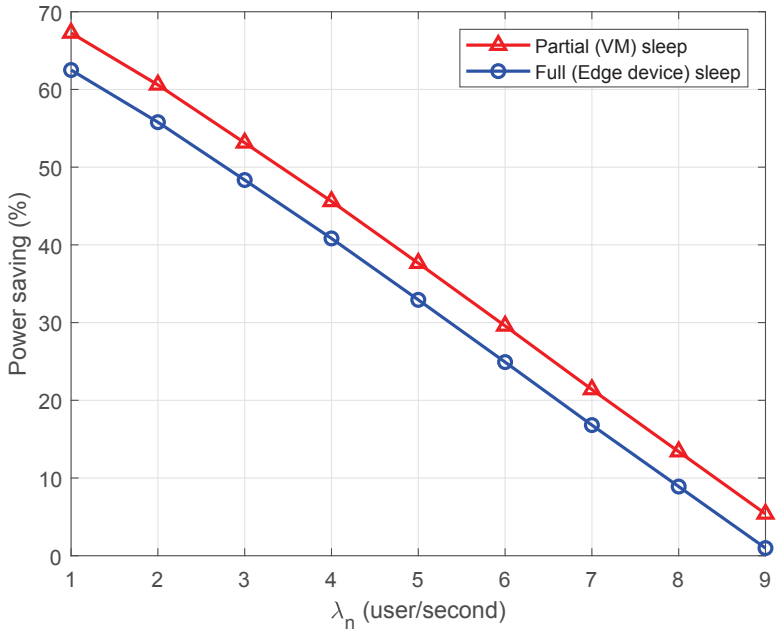


Figure 4.7: Energy saving comparison between the partial and full energy saving schemes.

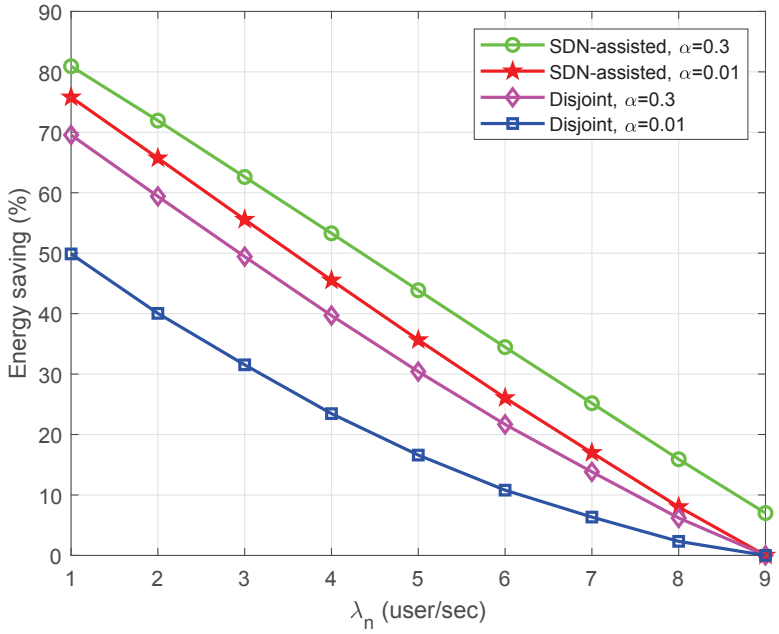


Figure 4.8: Partial energy saving using different schemes.

In regard with the queueing delay, the SDN-assisted model in Fig. 4.9 shows a relatively constant delay since the square-root staffing approach allocates adequate amount of VMs

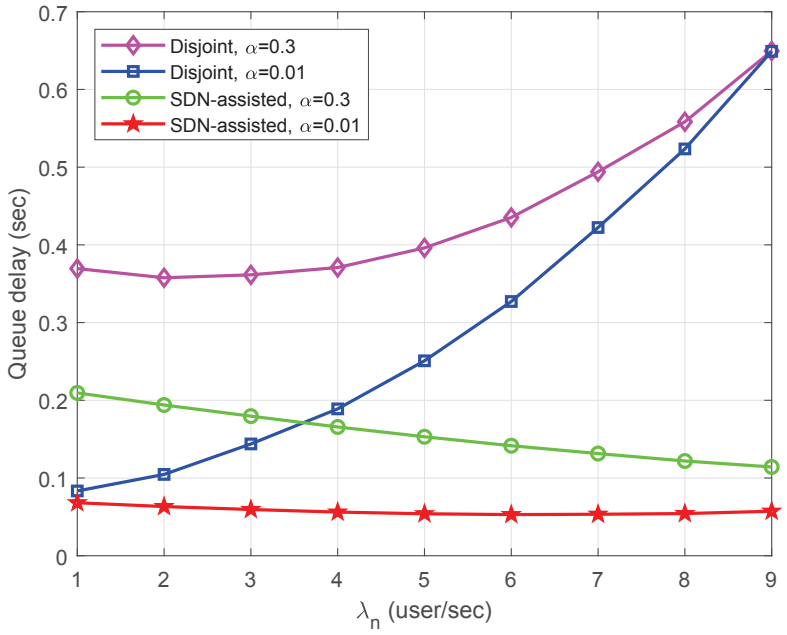


Figure 4.9: Queue delay using partial VM sleeping.

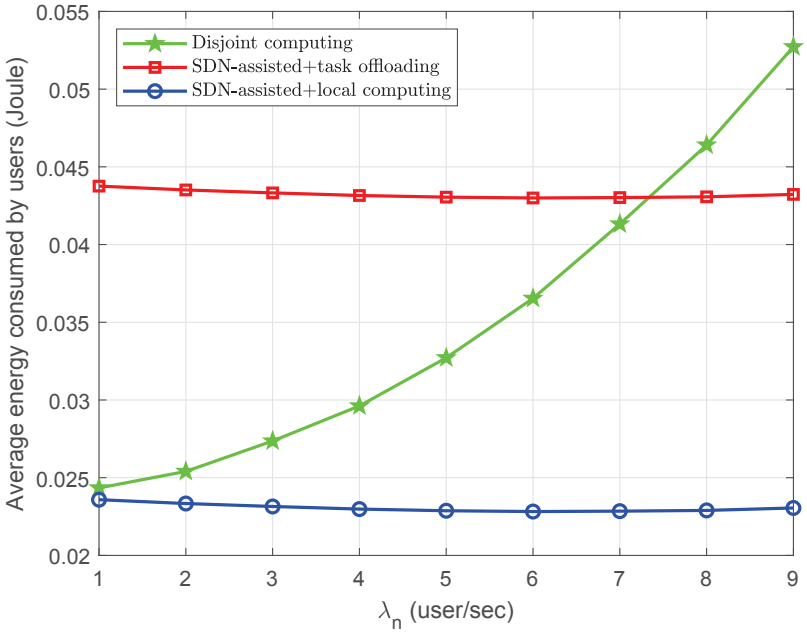


Figure 4.10: Average energy consumed by users.

with respect to the traffic load. On the other hand, the disjoint computing model which has a limited and fixed computing resources shows a sharp increase in the queueing delay when

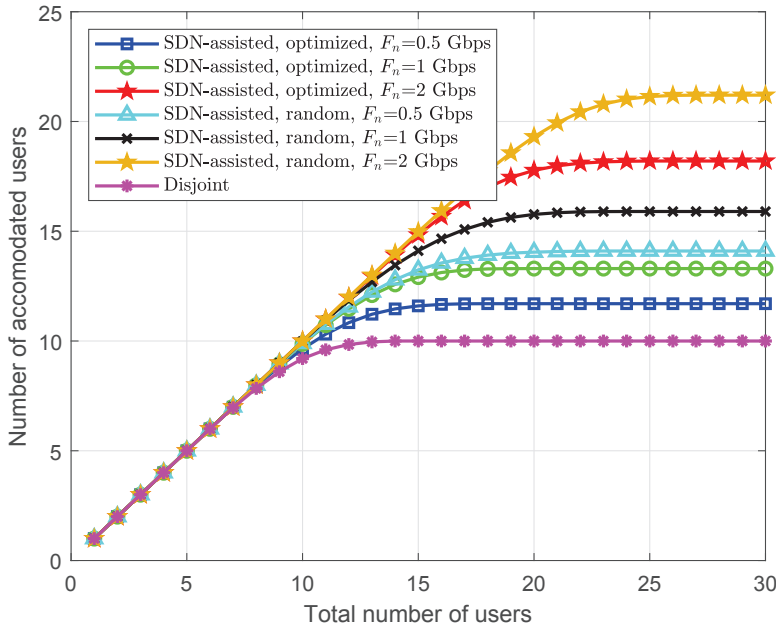


Figure 4.11: Accommodated users under different fronthaul capacity constraints.

the arrival rate increases towards full utilization of the queueing system [142].

The amount of energy consumed by users is directly affected by the queueing delay. Fig. 4.10 shows the user energy consumption using different schemes. In the SDN-assisted model, the queue delay is generally smaller, and with the local computing (i.e., tasks are processed within the nearest edge device without migrating through fronthaul links) the energy is lowest. On the other hand, the energy consumption is increased when task migration is involved when edge device are overloaded. It can also be noticed that serving users in the disjoint model incurs more queueing delay and as a result more energy consumption.

The last comparison in Fig. 4.11 shows the number of accommodated users under the SDN-assisted and the disjoint computing models. It can be observed that the maximum number of accommodated users in the disjoint model is limited to the number of VMs available at an edge device. Unlike the SDN-assisted model which can accommodate more users due to the better exploitation of resources, nevertheless, the number of accommodated users is limited by the fronthaul capacity. In other words, the fronthaul capacity bounds the number of tasks migrated among edge devices as seen in Fig. 4.11 where larger fronthaul

capacity helps to accommodate more users.

4.6 Chapter Summary

An SDN-assisted energy saving scheme is presented in this chapter aiming to turn off unnecessary computing resources in edge devices while maintaining the desired QoS for edge users. The proposed SDN architecture helps to reduce the queue delay experienced by edge users since all computing resources within a group of cooperative edge devices constitute a shared pool of VMs accessible by the SDN controller. To maintain the queueing probability below certain levels, the square-root staffing rule and the Halfin-Whitt function were used before deciding on whether to set edge devices into the on or off modes. Furthermore, a partial edge device sleep mode, where only portion of the available VMs are turned off locally within edge devices rather than the entire edge device, was introduced to enhance system flexibility and reduce the amount of task migration which aims to satisfy more users when some edge devices become overloaded. Moreover, the fronthaul link capacity was considered as a constraint that limits the amount of task migration among edge devices. Results showed that energy saving can be achieved with amounts that depend on the desired QoS requirements. Furthermore, with the partial sleep mode, more satisfied users can be obtained due to the improved system flexibility when dealing with VMs rather than the entire edge devices.

Besides the importance of energy saving on the large network scale, reducing the energy required by mobile users is crucial for the battery-enabled mobile devices. In Chapter 5, a NOMA-based resource allocation scheme will be presented to maximize the data rate provision for mobile users and to reduce the task completion time; as a result, reducing the energy consumed by mobile users. Moreover, such non-orthogonal frequency allocation schemes help to connect more mobile users, and that can tremendously help to enable future IoT systems which are featured with massive connectivity. A heterogeneous network with edge computing facilities will be considered to investigate the proposed scheme.

Chapter 5

Sparse Code Multiple Access-based Edge Computing for IoT Systems

5.1 Introduction

The Internet of Things (IoT) is expected to remarkably change the way we are living into a smarter, safer, and easier lifestyle. With the current trend towards IoT, it can be realized that IoT is confidently dominating the future of information and communication technologies. However, many challenges still exist regarding the establishment of efficient IoT systems, in particular, device connectivity and service latency. To meet the massive connectivity demands of IoT devices, the sparse code multiple access (SCMA) scheme, which is a NOMA-based scheme is envisioned as a promising solution to cope with the connectivity challenge and to fulfill the scalability needs of future networks [149], [150]. Unlike orthogonal frequency division techniques, NOMA-based schemes allow multiple users to share the same subcarriers to increase the number of users served. In contrast with other NOMA techniques, SCMA provides improved link-level performance and block error rate as introduced in [151] and [152]. Furthermore, comparing SCMA with code division multiple access (CDMA), which is a code-domain multiple access scheme, SCMA allows a multi-dimensional design of constellation points that in turn enhances system flexibility compared to the one-dimensional

constellation in CDMA [153].

Despite the aforementioned advantages, SCMA detection requires complicated algorithms to decode transmitted signals especially when the number of users sharing one subcarrier is increased. Along with the decoding process, implementing robust interference cancellation techniques is inevitable to maintain satisfactory signal quality at IoT receivers. In addition, the computing capability of IoT devices might be inadequate for fast SCMA detection, thus more delay will be experienced, and that is a serious issue for delay-sensitive applications such as e-health and vehicular communications that can only tolerate few milliseconds of delay [6]. Therefore, it is essential to consider the computing capabilities of IoT systems such as the microprocessor speed [154], when designing SCMA-based systems.

Rather than performing computing tasks using the on-device processors, IoT devices have the opportunity to offload their tasks to edge devices (fog nodes) in the vicinity, benefiting from the reduced end-to-end latency. The physical proximity of edge devices with end-users also supports IoT applications that require location awareness, low latency, and high QoS [155]. Instead of sending all data to the distant cloud, the operations of data aggregation, filtration, and analysis can be achieved by edge devices leaving only abstracted data to be further processed by the cloud. Moreover, edge devices can carry out machine learning techniques to harness the big IoT data for achieving accurate content caching and provide timely responses to end-users [156].

In [149], a NOMA-based radio and computing resource allocation scheme was proposed to reduce energy consumption in mobile edge computing. To enable an interactive communication among sensors and actuators, a power and channel allocation framework was proposed in [161] for 5G IoT networks. Furthermore, a hierarchical computing resource allocation scheme was proposed in [162] to optimally allocate the limited resources of fog nodes in IoT services. In [163], a comparison study showed that SCMA can provide better throughput in HetNets in contrast with other NOMA schemes at the cost of extra detection complexity.

In terms of SCMA encoding, several works in the literature considered optimal codebook design as in [153], where the system capacity and outage probability were derived for min-

Table 5.1: Related Works

Reference	Channel Allocation	Computing Model	Network Model	Objective
[149]	Non-orthogonal	Edge computing	Homogeneous	Energy consumption
[157]	Orthogonal	Edge computing	Homogeneous	Response time
[158]	Orthogonal	Edge computing	Homogeneous	System revenue
[159]	SCMA (non-orthogonal)	N/A	Homogeneous	Network utility
[160]	SCMA (non-orthogonal)	N/A	Homogeneous	Energy efficiency
[153]	SCMA (non-orthogonal)	N/A	Homogeneous	Outage probability and power allocation
[151]	SCMA (non-orthogonal)	N/A	Homogeneous	Energy efficiency and detection complexity
Our work	SCMA (non-orthogonal)	Edge computing	Heterogeneous	Device connectivity, sum rate, and task completion time

imizing outage probability for SCMA users using power allocation. In the same context, the work in [152] aimed to reduce the detection complexity in codebook design, namely the constellation design and codebook assignment. The detection complexity in SCMA has been investigated in [164], wherein the conventional message passing algorithm was enhanced using sphere decoding to reduce the number of superimposed constellation points in SCMA codebooks. Furthermore, decomposing high-order SCMA systems into smaller low-order systems using mapping modules was proposed in [165] to simplify the decoding process. In [150], a learning-based codebook generation and decoding strategy was proposed to adaptively construct codebooks with enhanced bit error rate.

From the computing perspective, different edge computing models have been used in the literature to investigate the computing performance in IoT systems. In [149], the computing capacity of edge devices are divided into resource blocks with certain CPU cycles. Each of these resource blocks is then allocated to a cluster of users that share the same frequency resources in a NOMA-based system. The work in [157] considered associating IoT devices with different fog nodes depending on application requirements and resource availability. Afterwards, each associated IoT device is allocated one VM with constant CPU speed. Likewise, the study in [158] considered associating IoT devices with suitable fog nodes; however, the computing resources of each fog node were considered to be shared equally among all IoT

devices within that node. Table 5.1 summarizes recent related works considering different computing models, multiple access schemes, and objectives. As seen in the table, some studies considered edge computing using non-orthogonal multiple access techniques (other than SCMA) in homogeneous networks, whereas other studies considered SCMA in homogeneous networks without incorporating edge computing. However, SCMA has not been considered neither for edge computing in homogeneous or heterogeneous networks in general, nor in the context of IoT device connectivity and time latency in particular.

In this work, we conduct a comprehensive investigation on the feasibility of SCMA for distributed IoT computing systems. By selecting different SCMA parameters, the system performance is significantly affected especially with regard to connectivity and computing delay. Scalable SCMA codebook configuration is also proposed, carried out through simulations, and is shown to improve system performance compared to the conventional OFDMA scheme.

The rest of the chapter is organized as follows. In Section II, the system model is introduced where network, computing, and SCMA models are presented. The problem formulation and solution approach are given in Section III. Simulation results and discussions are presented in Section IV. Finally, Section V presents the Chapter summary.

5.2 System Model

In this section, the network and computing models are presented. Fig. 5.1 depicts the proposed joint communication-computing system layout.

5.2.1 Network Model

We consider a SCMA-based heterogeneous network consisting of one MBS and a set of small base stations denoted by $\mathcal{N} = \{1, ..., N_p\}$. A set of IoT users $\mathcal{U} = \{1, ..., N_u\}$ are served using a bandwidth B that is divided into a set of subcarriers $\mathcal{S} = \{1, ..., N_{sc}\}$ which are later mapped into a set of codebooks denoted by $\mathcal{C} = \{1, ..., N_c\}$. The sum rate of user $k \in \mathcal{U}$ is

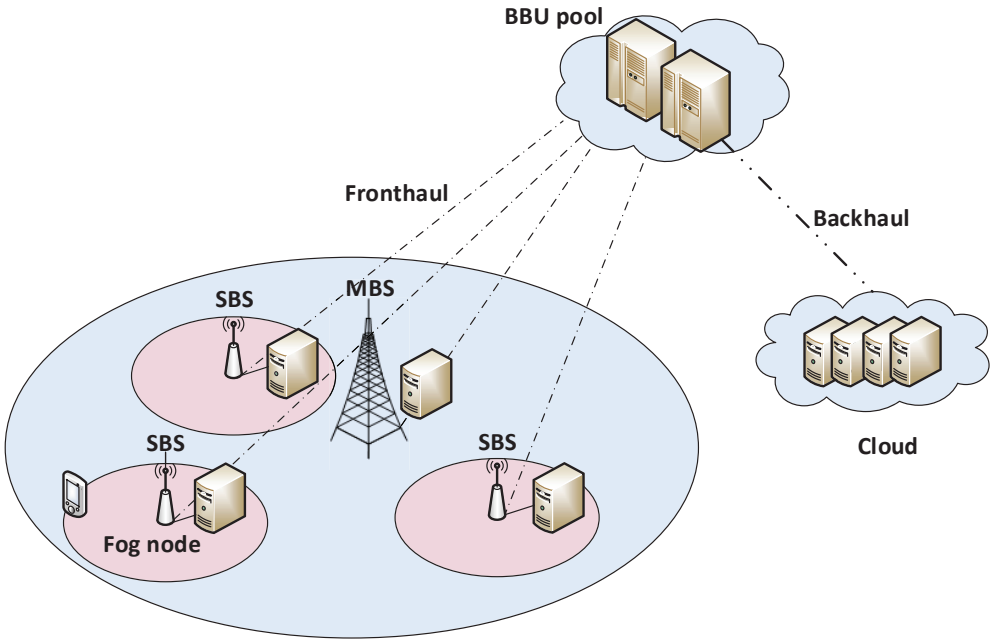


Figure 5.1: Proposed system layout.

given by

$$R_k = \sum_{n=1}^{N_p} \sum_{s=1}^{N_{sc}} a_{k,s}^n \log_2 \left(1 + \frac{p_{k,s}^n |h_{k,s}^n|^2}{I_{k,s}^n + N_0} \right), \tag{5.1}$$

where $a_{k,s}^n$, $p_{k,s}^n$, $h_{k,s}^n$, and $I_{k,s}^n$ denote respectively the user association, power allocation, channel gain, and inter-channel interference of user k over subcarrier s at base station n. N_0 is the noise power spectral density. It should be noted that the second term in (5.1) represents the SINR offered to user k from base station n over subcarrier s. In this model, users can be allocated subcarriers (codebooks) from different base stations in a CoMP fashion no matter whether those base stations are near or far. Nevertheless, to reduce the complexity of resource allocation, we first associate users with nearby base stations and then allocate codebooks according to their SINR values. Since several detection techniques such as multi-user detection based on MPA [166] and successive interference cancelation (SIC) [167] are carried out by SCMA users for signal detection, users treat the signals of users with lower

channel gains as noises [159]. Therefore, the inter-channel interference can be expressed as

$$I_{k,s}^{n} = \sum\nolimits_{\{i:|h_{i,s}^{n}|^2>|h_{k,s}^{n}|^2\}} p_{i,s}^{n}|h_{k,s}^{n}|^2, \quad \forall n \in \mathcal{N}. \tag{5.2}$$

Thus, the sum data rate obtained by all IoT devices can be calculated as

$$R_T = \sum_{k=1}^{N_u} R_k. \tag{5.3}$$

It should be noted that n refers to both the fog node and the SBS since they are considered functioning on the same site and serving the same users.

5.2.2 Computing Model

The diversity of IoT devices and applications such as e-health, smart transportation, and smart homes imposes different computing and delay requirements. Therefore, it is essential to address the specific needs of each particular device to maintain satisfactory QoS in terms of both computing and radio resource allocation. In the proposed system, each incoming user (IoT device) k is assumed to have a specific data size D_k bits, and a task completion deadline T_k seconds. Unlike the centralized cloud computing model where all computing tasks are processed in the distant cloud servers, fog nodes in the proposed edge computing model are responsible for handling computing tasks at the vicinity of IoT users within the small-cell tier. The computing (CPU) resources in cycle/sec allocated by fog node n to user k can be expressed as

$$c_k^n = \frac{\frac{D_k}{T_k}}{\sum_{k\in\mathcal{U}_n} \frac{D_k}{T_k}} \times C_n, \tag{5.4}$$

where C_n and $\mathcal{U}_n$ denote the total computing capacity in cycles/sec and the set of users being served by fog node n, respectively. As shown in (5.4), the computing resources of a fog node n are shared among all associated users ($k \in \mathcal{U}_n$). Moreover, the amount of resources allocated to a user k depends on the ratio $\frac{D_k}{T_k}$ such that a user with larger data size or more strict deadline will be allocated more CPU resources. Assuming that each bit of data requires

one cycle for processing (i.e., 1 cycle/sec is equivalent to 1 bit/sec), the task completion time of user k can be calculated as follows

$$t_k = \frac{D_k}{R_k} + \frac{D_k}{c_k^n}, \tag{5.5}$$

where $\frac{D_k}{R_k}$ and $\frac{D_k}{c_k^n}$ denote the delay incurred by wireless transmission and fog node processing, respectively. Each task k is considered satisfied if the task completion time t_k remains below the task completion deadline T_k (i.e., $t_k < T_k$).

5.2.3 SCMA Model

Codebook Structure

We consider a SCMA system that allows N_{sc} subcarriers to be shared by N_c codebooks which are later allocated to N_u IoT devices. Each individual subcarrier can be used simultaneously by d_s codebooks; whereas each codebook is assigned d_c subcarriers. Fig. 5.2 demonstrates the mapping relationship of subcarriers, codebooks and users for a SCMA system with $N_{sc} = 4$, $N_c = 6$, $N_u = 6$, $d_s = 3$ and $d_c = 2$.

Codebook design has been investigated in several studies including [152], [153], [164], and is considered beyond the scope of this work. Nevertheless, the design process implicates that $\log_2 M$ binary information bits are first mapped by the SCMA encoder into a d_c-dimensional constellation points, these constellation points are then zero-padded to spread over N_c codebooks. In this work, the conventional user-subcarrier (or codebook-subcarrier) association matrix is followed, where this sparse association matrix is also referred to as the factor graph matrix $\mathbf{F}$ [167].

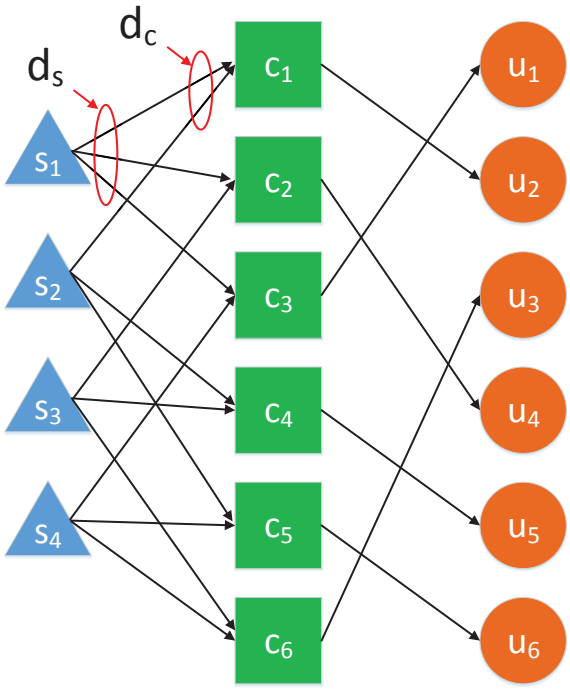


Figure 5.2: Factor graph of SCMA with $N_{sc} = 4$, $N_c = 6$, $N_u = 6$, $d_s = 3$ and $d_c = 2$.

$$\mathbf{F} = \begin{array}{c} \begin{matrix} c_1 & c_2 & c_3 & c_4 & c_5 & c_6 \end{matrix} \\ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \end{array} \begin{matrix} s_1 \\ s_2 \\ s_3 \\ s_4 \end{matrix}$$

Lemma 4. *To scale the SCMA system without violating the d_s and d_c constraints when using larger number of subcarriers, the factor graph matrix $\mathbf{F}$ can be used as a block in a diagonal matrix. For instance, the conventional factor graph $\mathbf{F}$ which is a 4×6 matrix can*

be expanded to a $4m \times 6m$ *matrix as*

$$\mathbb{F} = \begin{bmatrix} \mathbf{F} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{F} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{F} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{F} \end{bmatrix}$$

where $\mathbf{0}$ *is a* 4×6 *zero matrix and* $m \in \mathbb{Z}$ *indicates the number of diagonal blocks in* $\mathbb{F}$*.*

Proof. Since $\mathbb{F}_{4m \times 6m}$ is a diagonal matrix, this implies that $\sum_j \mathbb{F}_{i,j} = \mathbf{F}, \ \forall i$, and $\sum_i \mathbb{F}_{i,j} = \mathbf{F}, \ \forall j$. Since $\sum_j \mathbf{F}_{i,j} = d_s, \ \forall i$, and $\sum_i \mathbf{F}_{i,j} = d_c, \ \forall j$, then $\mathbb{F}$ maintains the same properties of $\mathbf{F}$ regarding d_s and d_c. □

SCMA Capacity

The motivation behind using SCMA for IoT systems originates from the scarcity of frequency resources to accommodate the massive numbers of IoT devices. Allowing one subcarrier to be shared by multiple users helps improving device connectivity. The capacity of SCMA system is determined by three factors, namely, the number of subcarriers N_{sc}, the number of users (codebooks) sharing one subcarrier (d_s), and the number of subcarriers per codebook (d_c).

Lemma 5. *The total number of obtainable codebooks can be expressed by* $N_c = \left\lfloor N_{sc} \frac{d_s}{d_c} \right\rfloor$.

Proof. Since $N_c \sum_j \mathbf{F}_{i,j} = N_{sc} \sum_i \mathbf{F}_{i,j} \equiv$ total number of ones in $\mathbf{F}$, which can also be expressed as $N_c d_c = N_{sc} d_s$, it implies that $N_c = N_{sc} \frac{d_s}{d_c}$. For non-integer values of $N_{sc} \frac{d_s}{d_c}$ the latter formula can be expressed as $N_c = \left\lfloor N_{sc} \frac{d_s}{d_c} \right\rfloor$. □

Detection Complexity

The detection complexity in SCMA receivers increases substantially with increasing d_s [165]. As a consequence, more delay will be incurred especially when IoT devices have relatively

low computing capabilities. Since IoT devices such as sensors, actuators, and wearable body sensors are inherently heterogeneous in their computing capabilities, allowing the same sub-carrier to be reused by large number of devices could be impractical for IoT systems. Thus, the detection complexity in IoT receivers using SCMA transmission needs to be investigated. For instance, the conventional message passing algorithm (MPA), which is a common low-complexity decoding technique for SCMA devices based on iterative propagation of messages between resource and user nodes, imposes exponential increase in complexity when the number of users and the codebook size increase. As introduced in [164], the complexity of the addition and multiplication operations required to decode SCMA signals can be expressed as

$$C_{Add} = d_s N_{sc} M^{d_s} + l_{max} d_s (N_{sc} M^{d_s} - N_{sc} M) \tag{5.6}$$

$$\begin{aligned} C_{Mult} = \quad & N_{sc} M^{d_s} (d_s + 4) + l_{max} d_s N_{sc} M^{d_s} (d_s - 1) \\ & + l_{max} N_u d_c M (d_c - 2), \end{aligned} \tag{5.7}$$

where M denotes the cardinality of the multi-dimensional constellation points. l_{max} is the maximum number of message passing update iterations. Assuming that each IoT device has a particular processing capability ζ_k, then t_k in (5.5) can be rewritten by adding an extra term related to MPA detection time as follows

$$t_k = \frac{D_k}{R_k} + \frac{D_k}{c_k^n} + \frac{C_{Add} + C_{Mult}}{\zeta_k}. \tag{5.8}$$

5.3 Problem Formulation

Abiding by the aim of the work, which is improving data rate provision to satisfy the delay requirements of IoT devices, the problem is formulated as a data rate maximization problem as follows.

$$
\begin{aligned}
\textbf{P1:}\quad & \underset{a^n_{k,s},p^n_{k,s}}{\textbf{Maximize:}} \quad R_T \\
& \textbf{Subject to:}\quad C1:\ \sum_{k=1}^{N_u} a^n_{k,s} = d_s, \forall n \in \mathcal{N}, \forall s \in \mathcal{S}, \\
& \qquad\qquad\quad C2:\ \sum_{s=1}^{N_{sc}} a^n_{k,s} = d_c, \forall n \in \mathcal{N}, \forall k \in \mathcal{U}, \\
& \qquad\qquad\quad C3:\ \sum_{k=1}^{N_u}\sum_{s=1}^{N_{sc}} p^n_{k,s} \leq P^n_{max}, \forall n \in \mathcal{N}, \\
& \qquad\qquad\quad C4:\ p^n_{k,s} > 0, \forall n \in \mathcal{N}, \forall k \in \mathcal{U}, \forall s \in \mathcal{S}, \\
& \qquad\qquad\quad C5:\ a^n_{k,s} \in \{0,1\}.
\end{aligned}
\tag{5.9}
$$

The constraint C1 ensures that each subcarrier is allocated to d_s codebooks (users), whereas C2 ensures that each user is allocated d_c subcarriers. C3 sets the upper limit for transmit power at each base station. C4 indicates that each subcarrier associated with a user is allocated a non-zero power, while C5 is the binary association variable of user k over subcarrier s. It is worth to mention that the task completion deadline T_k is not considered as a constraint but is used to evaluate the system performance by comparing the number of satisfied users under both SCMA and OFDMA schemes as will be seen in the results section.

The aforementioned problem involves both power allocation at each base station, and subcarrier allocation that is dependent on the availability of codebooks among all base stations under consideration. It is notable that **P1** involves real, integer, and binary variables which turn the problem into a mixed-integer nonlinear programming (MINLP) problem that is computationally intractable [149], [158], [2]. However, the problem can be solved with less difficulty when subdivided into two consecutive subproblems; (i) codebook allocation in which every user is allocated a codebook that provides the highest data rate (i.e., highest SINR) considering the combined effect of all subcarriers within that codebook, and (ii) power allocation whereby each base station undertakes power optimization for associated users. It should be noted that equal power allocation is carried out in the first subproblem (codebook allocation) since the subcarrier power is one of the parameters required to calculate the data

rate according to (5.1).

5.3.1 Codebook Allocation

The codebook allocation process has two phases: First, subcarriers are mapped onto N_c codebooks taking into account the d_s and d_c constraints. Second, each codebook is matched with an IoT device aiming at providing the highest SINR to IoT devices on a first-come-first-serve basis as shown in Fig. 5.2. The following algorithm illustrates the codebook allocation mechanism.

Algorithm 4: Codebook allocation

Define:

$\mathcal{C}$: Set of codebooks;

Initialize:

$k \leftarrow 1$;

Set c_k and R_k to zero $\forall k \in \mathcal{U}$;

Codebook allocation:

while $k \leq N_u$ **do**

  Find c^* satisfying R_k^* (highest SINR), $\forall k \in \mathcal{U}$;

  $c_k \leftarrow c^*$;

  $\mathcal{C} = \mathcal{C} \backslash c_k$;

  Update R_k;

  $k \leftarrow k + 1$;

end

Lemma 6. *The proposed one-to-one matching mechanism is a set-wise stable matching; that is, all users are guaranteed to be associated with a codebook.*

Proof. A matching function is considered stable if two conditions hold true. First, no individual element in both sets prefers being single (i.e., with no peer from the other set). Second,

no pair prefers other elements on their current outcome (i.e., each of the pair elements does not prefer the matched element). The proposed codebook-user matching algorithm guarantees stability due to the following:

1. All ($\mathcal{C}$,$\mathcal{U}$) elements are considered rational; that is to say, no user prefers being without a codebook and vice versa.

2. Since codebooks are allocated based on a first-come-first-serve basis, codebooks always prefer their associated users. On the other hand, users might prefer other subcarriers that are already allocated to other users. Nevertheless, that would not violate the second condition. Moreover, codebooks are considered strongly substitutable meaning that a user requesting an already occupied codebook can be allocated the next best codebook (with the next highest SINR). Accordingly, the codebook allocation mechanism achieves stability. □

5.3.2 Power Allocation

For a given codebook allocation, **P1** can be reduced to the power allocation problem that aims to maximize the system data rate (or minimize its negative) under power constraints as follows

$$\begin{aligned} \textbf{P2:} \quad \underset{p^n_{k,s}}{\textbf{Minimize:}} \quad & -\sum_{n=1}^{N_p}\sum_{k=1}^{N_u}\sum_{s=1}^{N_{sc}} a^n_{k,s}\log_2\left(1+\frac{p^n_{k,s}|h^n_{k,s}|^2}{I^n_{k,s}+N_0}\right) \\ \textbf{Subject to:} \quad & C1: \ \sum_{k=1}^{N_u}\sum_{s=1}^{N_{sc}} p^n_{k,s} \le P^n_{max}, \forall n \in \mathcal{N}, \\ & C2: \ p^n_{k,s} \ge 0, \forall n \in \mathcal{N}, k \in \mathcal{U}, s \in \mathcal{S}. \end{aligned} \tag{5.10}$$

Assuming that subcarriers have good channel conditions (i.e., SINR >> 1), then the logarithmic function $\log_2$(1+SINR) can provide an accurate approximation of $\log_2$(1+SINR) in **P2** [168]. As a result, the objective function in **P2** is a negative concave (convex) function since logarithmic functions are concave on positive real numbers [169]. Therefore, the optimization problem is convex and can be solved using the Lagrange multipliers method:

$$L(p_{k,s}^n, \lambda_n, v_n) = -\sum_{n=1}^{N_p}\sum_{k=1}^{N_u}\sum_{s=1}^{N_{sc}} a_{k,s}^n \log_2\left(1+\frac{p_{k,s}^n|h_{k,s}^n|^2}{I_{k,s}^n+N_0}\right) - \lambda_n p_{k,s}^n + v_n\left[\sum_{s=1}^{N_{sc}} p_s^n - P_{max}^n\right] \quad (5.11)$$

where λ_n and v_n are the optimal Lagrange multipliers related to base station n. Note that the interference is considered as additive white Gaussian noise for simplicity. By finding $\frac{\partial L}{\partial p_{k,s}^n} = 0$ and fulfilling the KKT conditions [169], the optimal power allocation can be calculated by

$$p_{k,s}^n = a_{k,s}^n\left(\frac{1}{\lambda_n} - \frac{N_0}{|h_{k,s}^n|^2}\right)^+, \quad (5.12)$$

where $(x)^+ = \max\{0, x\}$ and λ_n satisfies the following power constraint

$$\sum_{k=1}^{N_u}\sum_{s=1}^{N_{sc}} a_{k,s}^n\left(\frac{1}{\lambda_n} - \frac{N_0}{|h_{k,s}^n|^2}\right)^+ = P_{max}^n, \quad (5.13)$$

which can be rewritten as

$$\frac{1}{\lambda_n} = \frac{1}{\sum_{k=1}^{N_u}\sum_{s=1}^{N_{sc}} a_{k,s}^n}\left(P_{max}^n + \sum_{k=1}^{N_u}\sum_{s=1}^{N_{sc}}\frac{a_{k,s}^n N_0}{|h_{k,s}^n|^2}\right), \quad (5.14)$$

where $\frac{1}{\lambda_n}$ represents the power level at base station n. Therefore, the power allocated to each subcarrier is determined by $\frac{1}{\lambda_n}$ and the subcarrier's channel gain as seen in (5.12) where higher power is allocated to subcarriers with higher channel gain to maximize the data rate. It is worth mentioning that implementing water-filling in both OFDMA and SCMA systems is fundamentally the same; however, since SCMA allows subcarriers to be shared by multiple users and within different base stations, users that have a common subcarrier in their codebooks can be allocated different amounts of power over that subcarrier depending on the channel conditions. In other words, one subcarrier can have different amounts of power when assigned to different codebooks, unlike OFDMA where each subcarrier can be allocated to at most one user in a timeslot [159], [2]. Algorithm 5 illustrates the power allocation process in the SCMA scheme.

Algorithm 5: Power allocation in SCMA

Calculate $\frac{1}{\lambda_n}$ using (5.14), $\forall n \in \mathcal{N}$;

For each user $k \in \mathcal{U}$ within base station n:

$s \leftarrow 1$;

while $s \leq d_c$ **do**

    Calculate $p_{k,s}^n$ using (5.12);

    $s \leftarrow s + 1$;

end

5.4 Simulation Setup and Results

In this section, different metrics are used to investigate the SCMA performance for edge IoT computing in terms of data rate, connectivity, detection complexity, and computing performance. To ease tracking simulation parameters, Table 5.2 presents the list of variables and corresponding values used in simulations.

5.4.1 Investigating system performance using different SCMA settings

The advantage of SCMA over OFDMA stems from the capability of SCMA to accommodate more IoT devices in order to enhance the system capacity. On one hand, allowing more users to share the same subcarrier (increasing d_s) helps to accommodate more users at the expense of higher incurred interference. On the other hand, allocating more subcarriers to users (increasing d_c) leads to higher data rate provisioning for users but degrades the system connectivity. Hence, choosing d_s and d_c in SCMA depends on the particular demands of IoT systems. Figs. 5.3 and 5.4 demonstrate the effect of d_s and d_c on the number of obtainable codebooks and sum data rate, respectively. The number of obtainable codebooks in Fig. 5.3 exhibits an increasing trend with both increasing d_s and decreasing d_c, and with 16

Table 5.2: Simulation Parameters

Description	**Value**
Bandwidth (B)	10 MHz
Number of subcarriers (N_{sc})	6
Maximum number of users per subcarrier (d_s)	3
Maximum number of subcarriers per user (d_c)	2
Cardinality of the constellation points (M)	4
Processing speed at each fog node (C)	10 GHz
Processing speed of IoT devices (ζ_k)	Uniform distribution $[20 - 1000]$ MHz
Data size of user k (D_k)	Uniform distribution $[2 - 8]$ Mb
Coverage of macro BS	1 km
Coverage of small BS	0.1 km
Maximum transmit power (macro BS)	40 W
Maximum transmit power (small BS)	1 W
Path loss (macro BS)	$131.1 + 42.8 \log_{10}(D)$ dB, D in km
Path loss (small BS)	$145.4 + 37.5 \log_{10}(D)$ dB, D in km
Shadowing standard deviation (macro BS)	10 dB
Shadowing standard deviation (small BS)	6 dB
Multipath fading (both macro and small BS)	Rayleigh distribution with unit variance
Noise power spectral density (N_0)	-173 dBm/Hz

subcarriers available in the system, up to 48 codebooks can be obtained. Fig. 5.4 shows that increasing d_s leads to better system throughput; however, the incurred interference could deteriorate the system performance when d_c is not carefully chosen. One downside of SCMA is the interference encountering IoT devices, which leads to large throughput gaps among devices due to the variable channel conditions. It is thus motivating to statistically

investigate the per-user throughput difference in SCMA systems. To this end, the sample standard deviation of the per-user rate can be calculated as follows

$$\sigma = \sqrt{\frac{1}{N_u - 1} \sum_{k=1}^{N_u} |R_k - \bar{R}|} \tag{5.15}$$

where $\bar{R}$ is the mean data rate of all $k \in \mathcal{U}$. As seen in Fig. 5.5, the variations in per-user throughput show remarkable increase when the values of d_s and d_c increase.

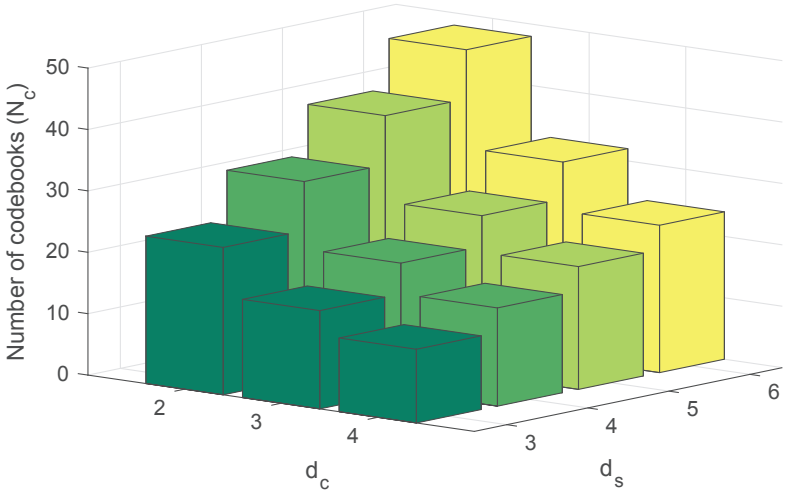


Figure 5.3: Obtainable codebooks using different values of d_s and d_c.

To support the massive connectivity of IoT devices, the ratio $\frac{N_c}{N_{sc}}$ (also referred to as the overloading factor) is desired to be much greater than one [170]. The latter condition can be satisfied by increasing d_s (refer to Fig. 3) which also leads to higher data rate provisioning as depicted in Fig. 4. However, the interference incurred by increasing d_s can also lead to higher variations among users in regard with the experienced data rate as shown in Fig. 5. On the other hand, increasing d_c has a negative impact on IoT connectivity; that is to say, less codebooks will be obtained when d_c is increased. Nevertheless, increasing d_c still shows an increase in the sum data rate since more subcarriers (higher data rates) are allocated to fewer number of users. Furthermore, since increasing d_c allows users to utilize more subcarriers, more interference will be encountered especially at high values of d_s, and that bounds the

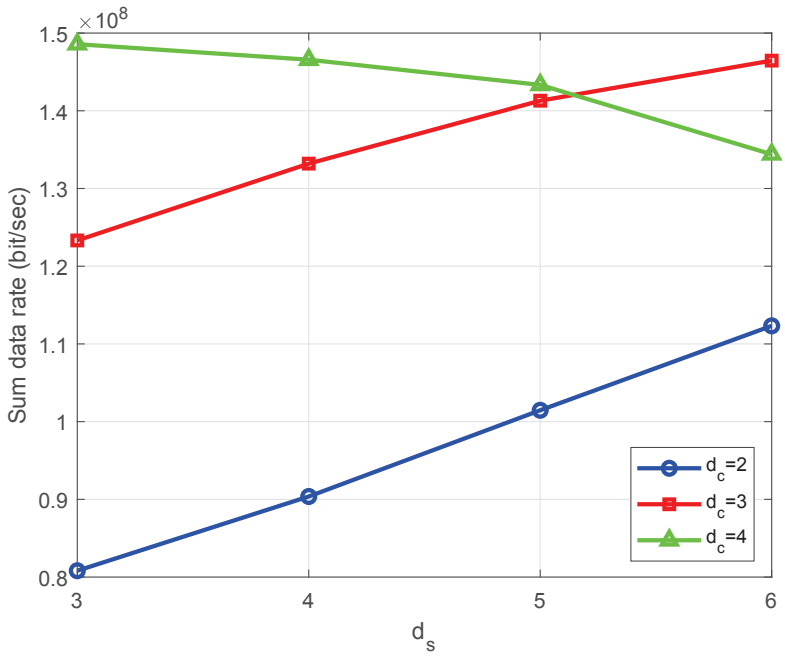


Figure 5.4: Effect of d_s and d_c on the sum data rate.

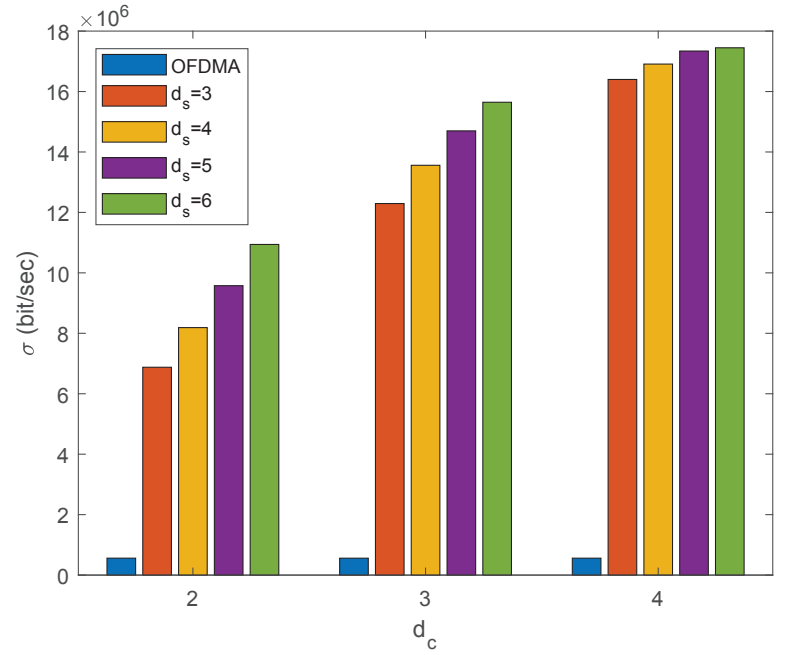


Figure 5.5: Effect of d_s and d_c on the per-user data rate.

increase in the sum data rate and incurs more variations in data rate provisioning as seen in Figs. 4 and 5, respectively.

Since SCMA allows each individual user to transmit/receive over multiple (d_c) subcarriers simultaneously, and allow the same subcarrier to be reused by d_s users, the sum throughput obtained shows an obvious increase compared to OFDMA. Nevertheless, the per-user data rate provision shows a considerable variation among SCMA users due to interference especially when the number of users increases. Fig. 5.6 demonstrates that feature of SCMA, where the central line indicates the median data rate, the box edges indicate the 25th and 75th percentiles, while the bottom- and top-most lines indicate the extreme data rate provided to IoT devices. From this figure, it can be seen that the per-user data rate of SCMA is higher than that of OFDMA, but declines with increasing number of users due to the incurred interference. In contrast, OFDMA users experience lower data rates that are relatively fixed with respect to number of users due to channel orthogonality. In addition to their effects on

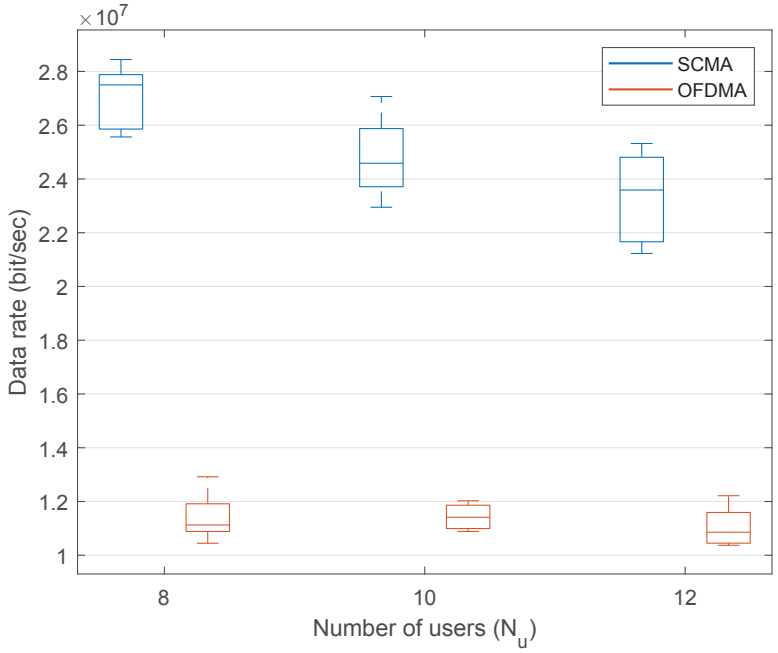


Figure 5.6: Comparison of the per-user data rate between SCMA and OFDMA systems at $N_u = 8, 10$, and 12.

connectivity and throughput, d_s and d_c have significant impact on the detection complexity at the receiver side. From (5.6) and (5.7), the total number of operations required to achieve signal detection is depicted in Fig. 5.7, whereas Fig. 5.8 shows the time required to execute these operations by an IoT device with a 20MHz processor assuming that each multiplication operation requires 3 clock cycles for processing, and each addition requires one cycle, while the total number of users in the system (N_u) is 12. It is worth mentioning that the effect of d_c on detection complexity is minor according to (5.6) and (5.7) compared to d_s which acts as an exponent and thus has significant impact. For this reason, Figs. 5.7 and 5.8 present results using different values of d_s while d_c is considered constant.

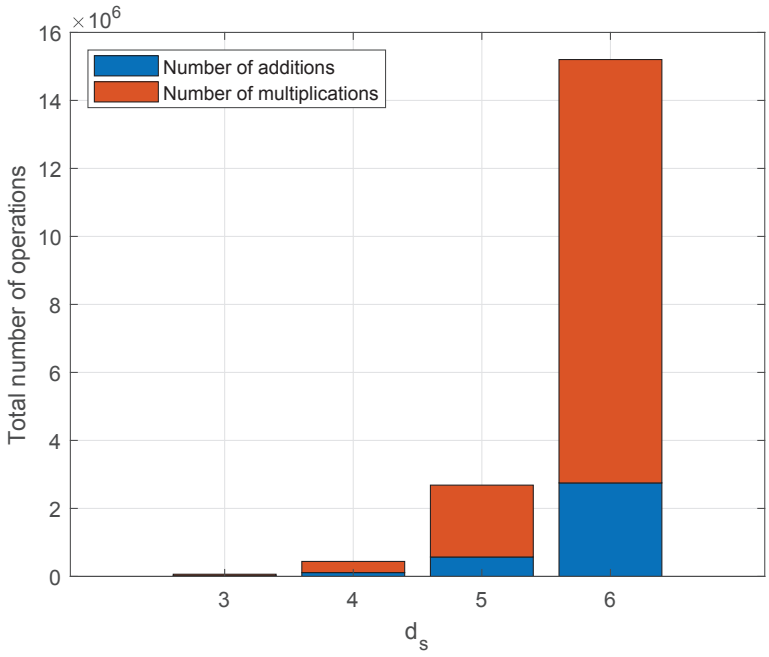


Figure 5.7: Total operations required for SCMA detection with d_c =2.

5.4.2 Proposed SCMA system performance

To quantify the performance gain of SCMA over OFDMA, we consider a SCMA system with $d_s = 3$ and $d_c = 2$ in order to improve connectivity and to satisfy more IoT devices while maintaining the detection complexity within satisfactory limits. The improved system

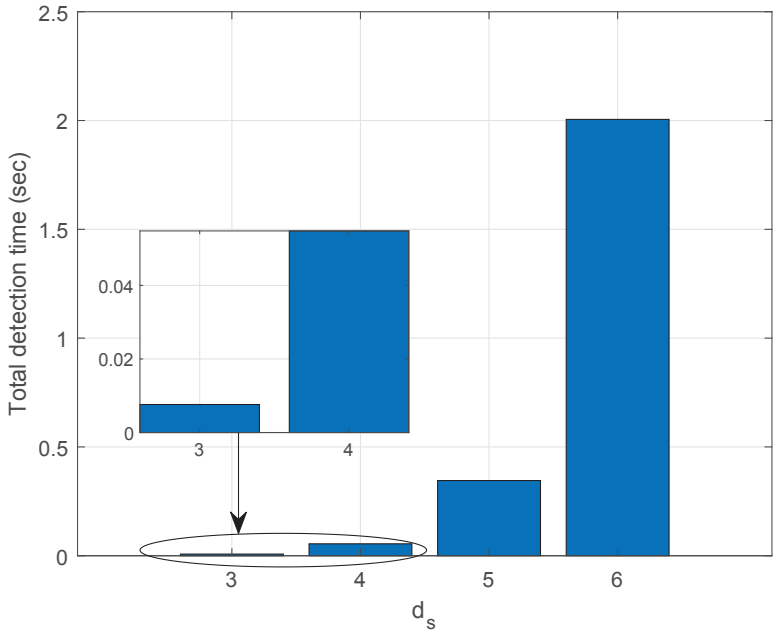


Figure 5.8: Time required by an IoT device with a 20MHz processing capacity, d_c =2.

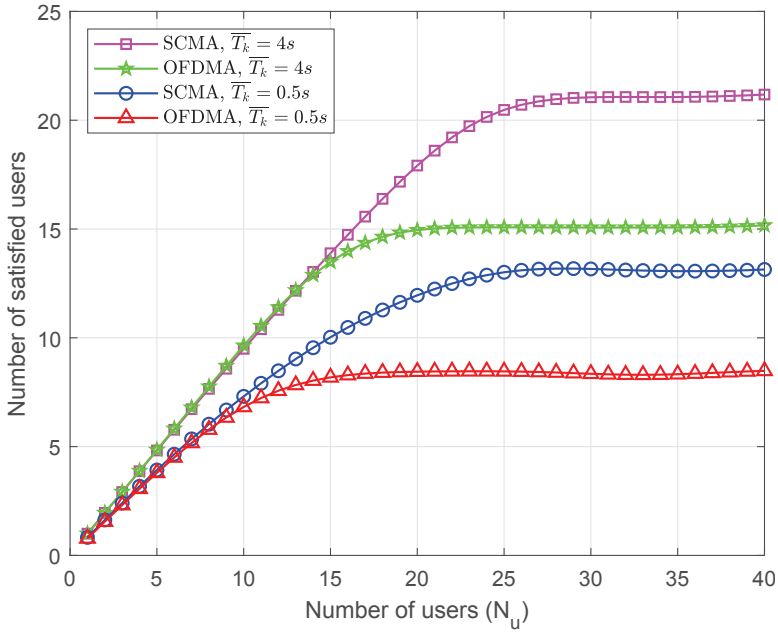


Figure 5.9: Computing performance comparison between SCMA and OFDMA under different deadline requirements.

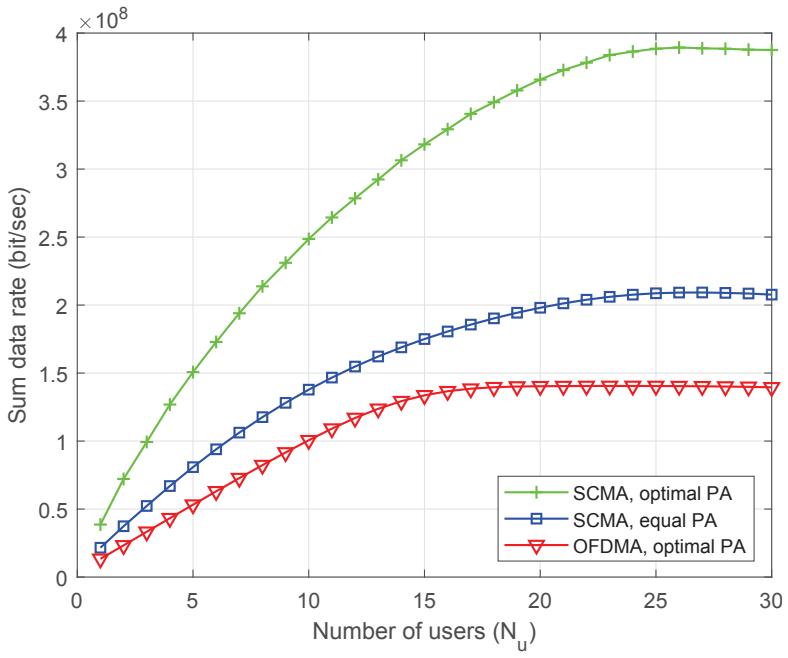


Figure 5.10: Sum data rate using different schemes.

connectivity of SCMA is evident compared to that of OFDMA as manifested by the number of satisfied users in Fig. 5.9 where users are assumed to have different task completion deadlines that follow a uniform distribution with mean $\overline{T_k}$. It can be noticed that the number of satisfied users saturates when the connectivity limit (i.e., 24 users in SCMA and 16 users in OFDMA) is reached; however, enforcing more stringent delay requirements by IoT devices results in less satisfied users since the task completion time will exceed the task completion deadline more easily.

Adopting optimized power allocation (PA) techniques such as water-filling leads to significant performance improvement compared to other strategies as shown in Fig. 5.10. It can also be observed that the sum data rate shows a continuous increase in SCMA until reaching the maximum connectivity limit which is 24 users at about 3.8×10^8 bits/sec, and that surpasses OFDMA which saturates at 1.4×10^8 bits/sec when the maximum connectivity limit is attained at 16 users.

A comparison between SCMA- and OFDMA-based edge computing is presented in Fig.

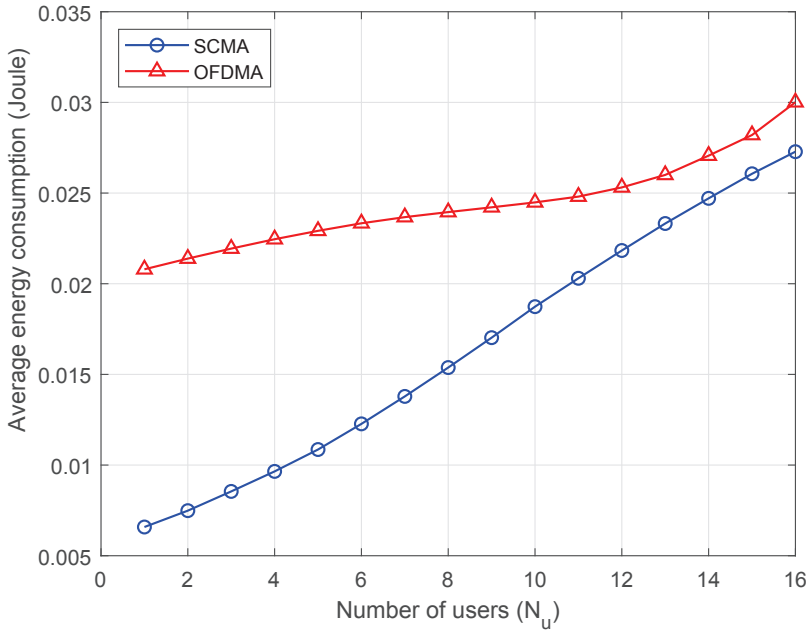


Figure 5.11: Average energy consumption of mobile devices.

5.11, where SCMA shows a clear advantage over OFDMA regarding energy consumption. This is due to the higher data rate offered by SCMA which reduces the transmission delay of computing tasks. However, when the number of users increases, the advantage of SCMA declines due to the encountered interference. In addition, the competition among users on the limited computing resources of fog nodes tends to increase when more users coexist in the system; as a result, the task completion time of both SCMA and OFDMA schemes increases with the number of users.

5.5 Chapter Summary

In this chapter, an SCMA-based edge computing scheme for IoT systems was proposed. Different SCMA parameters have been investigated to showcase the applicability of SCMA for IoT systems in comparison with traditional OFDMA-based schemes. The effects of these SCMA parameters, namely, the number of subcarriers allocated to one user, and the number of users sharing the same subcarrier have been extensively studied and their effects

on the system performance have been presented in detail. An optimization problem was also formulated to maximize system throughput under the power constraint and solved using the water filling approach. Results show the significance of implementing SCMA in improving network connectivity and maximizing data rate provision for better QoS performance in IoT systems.

Chapter 6

Conclusions and Future Work

The growing number of connected devices has urged the implementation of distributed multi-tiered networks from both cellular and computing perspectives. This concept is evident in the dense deployment of small base stations and edge devices to push radio and computing resources closer to mobile users benefiting the reduced delay and improved scalability. Nevertheless, due to the small coverage zones of these distributed nodes, more energy can be wasted imposing extra costs and greenhouse gas emissions. Moreover, frequency resources in cellular networks can be more efficient when shared by multiple users simultaneously; however, the detection complexity on mobile users can be increased. To this end, NOMA-based resource allocation schemes can help not only to improve network scalability but also to increase data rate provision, thus reducing the transmission time and energy consumption for mobile users.

6.1 Conclusions

In this thesis, we focused on implementing energy saving schemes in both C-RANs and edge computing networks. In addition, we aimed to improve the network connectivity by implementing a NOMA-based radio resource allocation, namely, the SCMA by which multiple users can use more than subcarrier to increase the data rate, and as a result, reduce the

delay that affects the per-user energy consumption.

In Chapter 2, a comprehensive literature review about energy efficiency in C-RANs, cloud-edge, and the cooperation between both C-RANs and cloud-edge networks was presented. The chapter first introduced the most recent advances in technology such as SDN and NFV that can significantly help coordinate the heterogenous network structure which involves multiple tiers, RATs, and operators. Then, we presented the most related works in the literature providing details about problem objectives, types, and solution methodologies. The chapter also elaborated the general C-RAN architecture that forms the basis for the system models in the subsequent chapters.

In Chapter 3, a base station sleeping mechanism was proposed and tested in a joint C-RAN-Cloud-Edge networks in a sense that allows small base stations to enter a sleep mode taking into account the cloud response time and the effect of task offloading from sleeping base stations. The network has been modeled using the M/M/k queueing system where the cloud queue response time is considered a constraint that restricts the decision on whether to set base stations into on or off modes. The cloud response time, transmission delay, and task processing time are all considered when finding the total time delay required by a user to complete the desired task. Adding such constraints in the energy saving mechanism showed a decrease the overall amount of energy saving but maintains the desired QoE requirements. The problem was initially formulated as a 0-1 knapsack optimization where the utilization and the amount of computing tasks at each small base stations were considered the weight and the value of that base station, respectively, then the problem was solved using the dynamic programming approach. Moreover, a shared-computing paradigm was proposed whereby edge and cloud servers constitute a joint queue. Herein, an exhaustive search algorithm was applied to find the optimal set of sleeping base stations. Results should that energy can be saved by turning off base stations while maintaining the desired QoE by satisfying the desired delay constraints.

The energy saving work was extended to Chapter 4, where the goal was to achieve energy saving in edge devices while maintaining the queueing probability below a desired threshold.

To this goal, an SDN-assisted controller similar to the BBU pool in C-RANs was presented to coordinate the resources of the cooperative group of edge devices. In this problem, the square-root staffing rule and the Halfin-Whitt function were used to satisfy the queueing threshold requirement by determining the number of required VMs. In addition, a partial sleep mode was proposed such that the on/off operation can be applied partially on VMs unlike the full sleep mode where all VMs in an edge device are turned off. A comparison regarding task migration due to sleeping edge devices is conducted taking into account the capacity of fronthaul links among edge devices. Results showed that the proposed schemes achieved successful energy saving performance without violating the delay (queueing) constraints. Also, the fronthaul capacity has been shown to act as a constraint in energy saving since it limits the amount of potentially migrated tasks among edge devices (users), and as a result, limits the flexibility of resource utilization and hence the amount of energy saving.

In order to accommodate larger numbers of users, and to maximize data rates for mobile users, an SCMA-based resource allocation scheme for IoT systems was presented in Chapter 5. In this chapter, the delay experienced by mobile users due to the detection complexity of SCMA subcarriers was considered as a satisfaction indicator for the SCMA scheme. Moreover, a power optimization and codebook allocation problems were formulated and solved using the water-filling approach and heuristic algorithm, respectively. Results showed the effectiveness of SCMA on improving the network scalability by accommodating larger number of mobile users with higher sum data rate provisioning. However, due to the higher incurred interference, users experience large variations in the provided data rates compared to OFDMA. Nevertheless, the per-user data rate provided by SCMA is generally higher than OFDMA, and that help to reduce the energy consumed by mobile users.

6.2 Future Work

The work in this thesis was focused on various schemes and approaches to achieve energy saving in both radio and computing networks. However, future work can still be done in

the direction of energy cost and service level agreement between users and mobile operators such that the tradeoff between user-level and network-level energy savings is considered in the energy saving scheme. Also, context-awareness needs to be investigated with the help of artificial intelligence and information security techniques in order to verify and predict the QoS requirements of mobile users. Furthermore, the algorithms used in this work can be further investigated and simplified to reduce the computation complexity and improve system performance.

Bibliography

[1] K. M. S. Huq, S. Mumtaz, J. Bachmatiuk, J. Rodriguez, X. Wang, and R. L. Aguiar, "Green HetNet CoMP: Energy Efficiency Analysis and Optimization," *IEEE Transactions on Vehicular Technology*, vol. 64, pp. 4670 – 4683, Oct. 2015.

[2] X. Sun and S. Wang, "Resource Allocation Scheme for Energy Saving in Heterogeneous Networks," *IEEE Transactions on Wireless Communications*, vol. 14, pp. 4407 – 4416, Aug. 2015.

[3] E. Chavarria-Reyes, I. F. Akyildiz, and E. Fadel, "Energy Consumption Analysis and Minimization in Multi-Layer Heterogeneous Wireless Systems," *IEEE Transactions on Mobile Computing*, vol. 14, pp. 2474 – 2487, Dec. 2015.

[4] J. Son, A. Dastjerdi, R. Calheiros, and R. Buyya, "SLA-Aware and Energy-Efficient Dynamic Overbooking in SDN-Based Cloud Data Centers," *IEEE Transactions on Sustainable Computing*, vol. 2, pp. 76 – 89, Jun. 2017.

[5] A. Alexiou, "Wireless World 2020: Radio Interface Challenges and Technology Enablers," *IEEE Vehicular Technology Magazine*, vol. 9, pp. 46 – 53, Mar. 2014.

[6] Y. Shih, W. Chung, A. Pang, T. Chiu, and H. Wei, "Enabling Low-Latency Applications in Fog-Radio Access Networks," *IEEE Network*, vol. 31, pp. 52 – 58, Dec. 2017.

[7] W. Wu, Q. Yang, B. Li, and K. S. Kwak, "Adaptive Cross-Layer Resource Optimization in Heterogeneous Wireless Networks with Multi-Homing User Equipments," *Journal of Communications and Networks*, vol. 18, pp. 784 – 795, Oct. 2016.

[8] C. Liu, B. Natarajan, and H. Xia, "Small Cell Base Station Sleep Strategies for Energy Efficiency," *IEEE Transactions on Vehicular Technology*, vol. 65, pp. 1652 – 1661, Mar. 2016.

[9] J. Wu, Y. Zhang, M. Zukerman, and E. Yung, "Energy-Efficient Base-Stations Sleep-Mode Techniques in Green Cellular Networks: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 17, pp. 803 – 826, 2nd Quart. 2015.

[10] X. Guo, Z. Niu, S. Zhou, and P. Kumar, "Delay-Constrained Energy-Optimal Base Station Sleeping Control," *IEEE Journal on Selected Areas in Communications*, vol. 34, pp. 1073 – 1085, May 2016.

[11] C. Jia and T. Lim, "Resource Partitioning and User Association With Sleep-Mode Base Stations in Heterogeneous Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 14, pp. 3780 – 3793, July 2015.

[12] B. Zhuang, D. Guo, and M. Honig, "Energy-Efficient Cell Activation, User Association, and Spectrum Allocation in Heterogeneous Networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, pp. 823 – 831, Apr. 2016.

[13] J. Rao and A. Fapojuwo, "Analysis of Spectrum Efficiency and Energy Efficiency of Heterogeneous Wireless Networks With Intra-/Inter-RAT Offloading," *IEEE Transactions on Vehicular Technology*, vol. 64, pp. 3120 – 3139, July 2015.

[14] J. Tang, D. K. C. So, E. Alsusa, K. A. Hamdi, and A. Shojaeifard, "Resource Allocation for Energy Efficiency Optimization in Heterogeneous Networks," *IEEE Journal on Selected Areas in Communications*, vol. 33, pp. 2104 – 2117, Oct. 2015.

[15] 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 Journal on Selected Areas in Communications*, vol. 34, pp. 1250 – 1266, May 2016.

[16] 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 Journal on Selected Areas in Communications*, vol. 34, pp. 3339 – 3353, Dec. 2016.

[17] Y. Chiang and W. Liao, "Green Multicell Cooperation in Heterogeneous Networks With Hybrid Energy Sources," *IEEE Transactions on Wireless Communications*, vol. 15, Dec. 2016.

[18] Y. Kwon, T. Hwang, and X. Wang, "Energy-Efficient Transmit Power Control for Multi-tier MIMO HetNets," *IEEE Journal on Selected Areas in Communications*, vol. 33, pp. 2070 – 2086, Oct. 2015.

[19] 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 *IEEE International Workshop on Computer Aided Modelling and Design of Communication Links and Networks (CAMAD)*, pp. 69 – 74, Sep. 2015.

[20] G. Yu, Y. Jiang, L. Xu, and G. Li, "Multi-Objective Energy-Efficient Resource Allocation for Multi-RAT Heterogeneous Networks," *IEEE Journal on Selected Areas in Communications*, vol. 33, pp. 2118 – 2127, Oct. 2015.

[21] W. Nie, F. 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 Journal on Selected Areas in Communications*, vol. 34, pp. 1192 – 1206, May 2016.

[22] M. Lin, S. Silvestri, N. Bartolini, and T. L. Porta, "On Selective Activation in Dense Femtocell Networks," *IEEE Transactions on Wireless Communications*, vol. 15, pp. 7018 – 7029, Oct. 2016.

[23] B. Niu and V. Wong, "Network Configuration for Two-Tier Macro–Femto Systems With Hybrid Access," *IEEE Transactions on Vehicular Technology*, vol. 65, pp. 2528 – 2543, Apr. 2016.

[24] M. Adedoyin and O. Falowo, "Self-Organizing Radio Resource Management for Next Generation Heterogeneous Wireless Networks," in *IEEE International Conference on Communications (ICC)*, pp. 1 – 6, May 2016.

[25] S. Zhang *et al.*, "Energy-Aware Traffic Offloading for Green Heterogeneous Networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, pp. 1116 – 1129, May 2016.

[26] J. Gong, J. Thompson, S. Zhou, and Z. Niu, "Base Station Sleeping and Resource Allocation in Renewable Energy Powered Cellular Networks," *IEEE Transactions on Communications*, vol. 62, pp. 3801 – 3813, Nov. 2014.

[27] D. Liu, Y. Chen, K. Chai, T. Zhang, and M. Elkashlan, "Two-Dimensional Optimization on User Association and Green Energy Allocation for HetNets With Hybrid Energy Sources," *IEEE Transactions on Communications*, vol. 63, pp. 4111 – 4124, Nov. 2015.

[28] W. Nie, Y. Zhong, F. Zheng, W. Zhang, and T. O'Farrell, "HetNets With Random DTX Scheme: Local Delay and Energy Efficiency," *IEEE Transactions on Vehicular Technology*, vol. 65, pp. 6601 – 6613, Aug. 2016.

[29] J. Liu, H. Guo, Z. Fadlullah, and N. Kato, "Energy Consumption Minimization for FiWi Enhanced LTE-A HetNets with UE Connection Constraint," *IEEE Communications Magazine*, vol. 54, pp. 56–62, Nov. 2016.

[30] H. Klessig *et al.*, "From Immune Cells to Self-Organizing Ultra-Dense Small Cell Networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, pp. 800 – 811, Apr. 2016.

[31] Z. Niu, Y. Wu, J. Gong, and Z. Yang, "Cell Zooming for Cost-Efficient Green Cellular Networks," *IEEE Communications Magazine*, vol. 48, pp. 74 – 79, Nov. 2010.

[32] H. Lateef, M. Shakir, M. Ismail, A. Mohamed, and K. Qaraqe, "Towards Energy Efficient and Quality of Service Aware Cell Zooming in 5G Wireless Networks," in *IEEE 82nd Vehicular Technology Conference (VTC)*, pp. 1 – 5, Sep. 2015.

[33] G. Wu, C. Yang, S. Li, and G. Y. Li, "Recent Advances in Energy-Efficient Networks and Their Application in 5G Systems," *IEEE Wireless Communications*, vol. 22, pp. 145 – 151, Apr. 2015.

[34] M. Peng, Y. Li, Z. Zhao, and C. Wang, "System Architecture and Key Technologies for 5G Heterogeneous Cloud Radio Access Networks," *IEEE Network*, vol. 29, pp. 6 – 14, Apr. 2015.

[35] A. Li, Y. Sun, X. Xu, and C. Yuan, "An Energy-Effective Network Deployment Scheme for 5G Cloud Radio Access Networks," in *IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS*, pp. 684 – 689, Apr. 2016.

[36] 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 Communications*, vol. 21, pp. 126 – 135, Dec. 2014.

[37] C. Fan, Y. J. Zhang, and X. Yuan, "Advances and Challenges Toward a Scalable Cloud Radio Access Network," *IEEE Communications Magazine*, vol. 54, pp. 29 – 35, June 2016.

[38] 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 Transactions on Vehicular Technology*, vol. 64, pp. 5275 – 5287, Nov. 2015.

[39] A. Douik, H. Dahrouj, T. Al-Naffouri, and M. Alouini, "Coordinated Scheduling and Power Control in Cloud-Radio Access Networks," *IEEE Transactions on Wireless Communications*, vol. 15, pp. 2523 – 2536, Apr. 2016.

[40] M. Peng, Y. Yu, H. Xiang, and H. V. Poor, "Energy-Efficient Resource Allocation Optimization for Multimedia Heterogeneous Cloud Radio Access Networks," *IEEE Transactions on Multimedia*, vol. 18, pp. 879 – 892, May 2016.

[41] S. Lien, S. Cheng, K. Chen, and D. Kim, "Resource-Optimal Licensed-Assisted Access in Heterogeneous Cloud Radio Access Networks With Heterogeneous Carrier Communications," *IEEE Transactions on Vehicular Technology*, vol. 65, pp. 9915 – 9930, Dec. 2016.

[42] S. Lien, S. Hung, K. Chen, and Y. Liang, "Ultra-Low-Latency Ubiquitous Connections in Heterogeneous Cloud Radio Access Networks," *IEEE Wireless Communications*, vol. 22, pp. 22 – 31, June 2015.

[43] Y. Zhang, Y. Wang, and W. Zhang, "Energy Efficient Resource Allocation for Heterogeneous Cloud Radio Access Networks With User Cooperation and QoS Guarantees," in *IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1 – 6, Apr. 2016.

[44] Y. Zhang and Y. Wang, "A Framework for Energy Efficient Control in Heterogeneous Cloud Radio Access Networks," in *IEEE/CIC International Conference on Communications in China (ICCC Workshops)*, pp. 1 – 5, July 2016.

[45] J. Li, M. Peng, Y. Yu, and Z. Ding, "Energy-Efficient Joint Congestion Control and Resource Optimization in Heterogeneous Cloud Radio Access Networks," *IEEE Transactions on Vehicular Technology*, vol. 65, pp. 9873 – 9887, Dec. 2016.

[46] C. Ran and S. Wang, "Resource Allocation in Heterogeneous Cloud Radio Access Networks: A Workload Balancing Perspective," in *IEEE Global Communications Conference (GLOBECOM)*, pp. 1 – 6, Dec. 2015.

[47] Z. Yu, K. Wang, H. Ji, X. Li, and H. Zhang, "Dynamic Resource Allocation in TDD-Based Heterogeneous Cloud Radio Access Networks," *China Communications*, vol. 13, pp. 1 – 11, June 2016.

[48] A. Li, Y. Sun, X. Xu, and C. Yuan, "Joint Remote Radio Head Selection and User Association in Cloud Radio Access Networks," in *IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 1 – 6, Sep. 2016.

[49] P. Huang, H. Kao, and W. Liao, "Hierarchical Cooperation in Heterogeneous Cloud Radio Access Networks," in *IEEE International Conference on Communications (ICC)*, pp. 1 – 6, May 2016.

[50] S. Kuang and N. Liu, "Energy Minimization via BS Selection and Beamforming for Cloud-RAN under Finite Fronthaul Capacity Constraints," in *IEEE 83rd Vehicular Technology Conference*, pp. 1 – 6, May 2016.

[51] V. Ha and L. Le, "Joint Coordinated Beamforming and Admission Control for Fronthaul Constrained Cloud-RANs," in *IEEE Global Communications Conference*, pp. 4054 – 4059, Dec. 2014.

[52] Y. Cao, T. Jiang, and C. Wang, "Optimal Radio Resource Allocation for Mobile Task Offloading in Cellular Networks," *IEEE Network*, vol. 28, pp. 68 – 73, Oct. 2014.

[53] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint Optimization of Radio and Computational Resources for Multicell Mobile-Edge Computing," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 1, pp. 89 – 103, June 2015.

[54] D. Mazza, D. Tarchi, and G. Corazza, "A Unified Urban Mobile Cloud Computing Offloading Mechanism for Smart Cities," *IEEE Communications Magazine*, vol. 55, pp. 30 – 37, Mar. 2017.

[55] D. Pompili, A. Hajisami, and T. X. Tran, "Elastic Resource Utilization Framework for High Capacity and Energy Efficiency in Cloud RAN," *IEEE Communications Magazine*, vol. 54, pp. 26 – 32, Jan. 2016.

[56] M. A. Marotta *et al.*, "Resource Sharing in Heterogeneous Cloud Radio Access Networks," *IEEE Wireless Communications*, vol. 22, pp. 74 – 82, June 2015.

[57] H. Dahrouj, A. Douik, O. Dhifallah, T. Y. Al-Naffouri, and M. Alouini, "Resource Allocation in Heterogeneous Cloud Radio Access Networks: Advances and Challenges," *IEEE Wireless Communications*, vol. 22, pp. 66 – 73, June 2015.

[58] D. Mishra, P. Amogh, A. Ramamurthy, A. Franklin, and B. Tamma, "Load-aware dynamic RRH assignment in Cloud Radio Access Networks," in *IEEE Wireless Communications and Networking Conference*, pp. 1 – 6, Apr. 2016.

[59] K. Wang, M. Zhao, and W. Zhou, "Traffic-Aware Graph-Based Dynamic Frequency Reuse for Heterogeneous Cloud-RAN," in *IEEE Global Communications Conference*, pp. 2308 – 2313, Dec. 2014.

[60] X. Zhang *et al.*, "Macro-Assisted Data-Only Carrier for 5G Green Cellular Systems," *IEEE Communications Magazine*, vol. 53, pp. 223 – 231, May 2015.

[61] M. Peng, C. Wang, V. Lau, and H. Poor, "Fronthaul-Constrained Cloud Radio Access Networks: Insights and Challenges," *IEEE Wireless Communications*, vol. 22, pp. 152 – 160, Apr. 2015.

[62] M. Wang, H. Xia, and C. Feng, "Joint Dynamic Point Blanking and ABS for ICIC in Cloud Cooperated Heterogeneous Network," in *IEEE/CIC International Conference on Communications in China (ICCC)*, pp. 1 – 5, Nov. 2015.

[63] B. Niu, Y. Z. H. Shah-Mansouri, and V. Wong, "A Dynamic Resource Sharing Mechanism for Cloud Radio Access Networks," *IEEE Transactions on Wireless Communications*, vol. 15, pp. 8325 – 8338, Dec. 2016.

[64] M. Peng, X. Xie, Q. Hu, J. Zhang, and H. V. Poor, "Contract-Based Interference Coordination in Heterogeneous Cloud Radio Access Networks," *IEEE Journal on Selected Areas in Communications*, vol. 33, pp. 1140 – 1153, June 2015.

[65] 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.

[66] A. Abdelnasser and E. Hossain, "Resource Allocation for an OFDMA Cloud-RAN of Small Cells Underlaying a Macrocell," *IEEE Transactions on Mobile Computing*, vol. 15, pp. 2837 – 2850, Nov. 2016.

[67] N. Abuzainab and W. Saad, "Cloud Radio Access Meets Heterogeneous Small Cell Networks: A Cognitive Hierarchy Perspective," in *IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 1 – 5, July 2016.

[68] X. Huang, G. Xue, R. Yu, and S. Leng, "Joint Scheduling and Beamforming Coordination in Cloud Radio Access Networks With QoS Guarantees," *IEEE Transactions on Vehicular Technology*, vol. 65, pp. 5449 – 5460, July 2016.

[69] K. Zhang, M. Peng, C. Wang, and S. Yan, "Perron-Frobenius Theory Based Power Allocation in Heterogeneous Cloud Radio Access Networks," in *IEEE 82nd Vehicular Technology Conference (VTC)*, pp. 1 – 5, Sep. 2015.

[70] S. Luo, R. Zhang, and T. Lim, "Downlink and Uplink Energy Minimization Through User Association and Beamforming in C-RAN," *IEEE Transactions on Wireless Communications*, vol. 14, pp. 494 – 508, Jan. 2015.

[71] 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 Communications*, vol. 22, pp. 92 – 99, June 2015.

[72] A. Adhikary, H. S. Dhillon, and G. Caire, "Massive-MIMO Meets HetNet: Interference Coordination Through Spatial Blanking," *IEEE Journal on Selected Areas in Communications*, vol. 33, pp. 1171 – 1186, June 2015.

[73] W. Lin, C. Lee, and H. Su, "Downlink-to-Uplink Interference Cancellation in Cloud Radio Access Networks," in *IEEE 79th Vehicular Technology Conference*, pp. 1–5, May 2014.

[74] 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.

[75] K. A. Meerja, A. Shami, and A. Refaey, "Hailing Cloud Empowered Radio Access Networks," *IEEE Wireless Communications*, vol. 22, pp. 122 – 129, Feb. 2015.

[76] O. Dhifallah, H. Dahrouj, T. Al-Naffouri, and M. Alouini, "Decentralized Group Sparse Beamforming for Multi-Cloud Radio Access Networks," in *IEEE Global Communications Conference (GLOBECOM)*, pp. 1 – 6, Dec. 2015.

[77] R. Brandt, R. Mochaourab, and M. Bengtsson, "Distributed Long-Term Base Station Clustering in Cellular Networks Using Coalition Formation," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 2, pp. 362 – 375, Sep. 2016.

[78] L. Z. A. U. Quddus, E. Katranaras, D. Wübben, Y. Qi, and R. Tafazolli, "Performance Analysis and Optimal Cooperative Cluster Size for Randomly Distributed Small Cells Under Cloud RAN," *IEEE Access*, vol. 4, pp. 1925 – 1939, Apr. 2016.

[79] M. M. U. Rahman, H. Ghauch, S. Imtiaz, and J. Gross, "RRH Clustering and Transmit Precoding for Interference-Limited 5G CRAN Downlink," in *IEEE Globecom Workshops (GC Wkshps)*, pp. 1 – 7, Dec. 2015.

[80] B. Dai and W. Yu, "Energy Efficiency of Downlink Transmission Strategies for Cloud Radio Access Networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, pp. 1037 – 1050, Apr. 2016.

[81] A. Hajisami and D. Pompili, "DJP: Dynamic Joint Processing for Interference Cancellation in Cloud Radio Access Networks," in *IEEE 82nd Vehicular Technology Conference*, pp. 1 – 5, Sep. 2015.

[82] R. Brandt, R. Mochaourab, and M. Bengtsson, "Globally Optimal Base Station Clustering in Interference Alignment-Based Multicell Networks," *IEEE Signal Processing Letters*, vol. 23, pp. 512 – 516, Apr. 2016.

[83] E. Chen and M. Tao, "User-Centric Base Station Clustering and Sparse Beamforming for Cache-Enabled Cloud RAN," in *IEEE/CIC International Conference on Communications in China (ICCC)*, pp. 1 – 6, Nov. 2015.

[84] H. Soliman and A. Leon-Garcia, "QoS-Aware Joint RRH Activation and Clustering in Cloud-RANs," in *IEEE Wireless Communications and Networking Conference*, pp. 1 – 6, Apr. 2016.

[85] L. Lei, D. Yuan, C. Ho, and S. Sun, "Optimal Cell Clustering and Activation for Energy Saving in Load-Coupled Wireless Networks," *IEEE Transactions on Wireless Communications*, vol. 14, pp. 6150 – 6163, Nov. 2015.

[86] C. Ran, S. Wang, and C. Wang, "Balancing Backhaul Load in Heterogeneous Cloud Radio Access Networks," *IEEE Wireless Communications*, vol. 22, pp. 42 – 48, June 2015.

[87] H. Yang, G. Geraci, and T. Quek, "Energy-Efficient Design of MIMO Heterogeneous Networks With Wireless Backhaul," *IEEE Transactions on Wireless Communications*, vol. 15, pp. 4914 – 4927, July 2016.

[88] H. Yang, G. Geraci, and T. Quek, "MIMO HetNets with Wireless Backhaul: An Energy-Efficient Design," in *IEEE International Conference on Communications (ICC)*, pp. 1 – 6, May 2016.

[89] T. Nguyen, A. Yadav, W. Ajib, and C. Assi, "Achieving Energy-Efficiency in Two-Tiers Wireless Backhaul HetNets," in *IEEE International Conference on Communications (ICC)*, pp. 1 – 6, May 2016.

[90] D. Zhang, M. Chen, M. Guizani, H. Xiong, and D. Zhang, "Mobility Prediction in Telecom Cloud Using Mobile Calls," *IEEE Wireless Communications*, vol. 21, pp. 26 – 32, Feb. 2014.

[91] C. Yang, Z. Chen, B. Xia, and J. Wang, "When ICN meets C-RAN for HetNets: an SDN approach," *IEEE Communications Magazine*, vol. 53, pp. 118 – 125, Nov. 2015.

[92] O. Simeone, A. Maeder, M. Peng, O. Sahin, and W. Yu, "Cloud Radio Access Network: Virtualizing Wireless Access for Dense Heterogeneous Systems," *Journal of Communications and Networks*, vol. 18, pp. 135 – 149, Apr. 2016.

[93] S. Park, C. Chae, and S. Bahk, "Large-Scale Antenna Operation in Heterogeneous Cloud Radio Access Networks: A Partial Centralization Approach," *IEEE Wireless Communications*, vol. 22, pp. 32 – 40, June 2015.

[94] 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 Transactions on Networking*, vol. 24, pp. 915 – 928, Apr. 2016.

[95] Y. Zhou and W. Yu, "Optimized Backhaul Compression for Uplink Cloud Radio Access Network," *IEEE Journal on Selected Areas in Communications*, vol. 32, pp. 1295 – 1307, June 2014.

[96] Y. Zhou and W. Yu, "Fronthaul Compression and Transmit Beamforming Optimization for Multi-Antenna Uplink C-RAN," *IEEE Transactions on Signal Processing*, vol. 64, pp. 4138 – 4151, Aug 2016.

[97] H. Ma, B. Wang, Y. Chen, and K. Liu, "Time-Reversal Tunneling Effects for Cloud Radio Access Network," *IEEE Transactions on Wireless Communications*, vol. 15, pp. 3030 – 3043, Apr. 2016.

[98] M. Kamel, W. Hamouda, and A. Youssef, "Ultra-Dense Networks: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 18, pp. 2522 – 2545, 4th Quart. 2016.

[99] J. Kim, W. Jeon, and D. Jeong, "Effect of Base Station-Sleeping Ratio on Energy Efficiency in Densely Deployed Femtocell Networks," *IEEE Communications Letters*, vol. 19, pp. 641 – 644, Apr. 2015.

[100] G. Carvalho, I. Woungang, A. Anpalagan, and E. Hossain, "QoS-Aware Energy-Efficient Joint Radio Resource Management in Multi-RAT Heterogeneous Networks," *IEEE Transactions on Vehicular Technology*, vol. 65, pp. 6343 – 6365, Aug. 2016.

[101] 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 Transactions on Vehicular Technology*, vol. 65, pp. 3657 – 3675, May 2016.

[102] 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 Transactions on Wireless Communications*, vol. 14, pp. 4391 – 4406, Aug. 2015.

[103] J. Kim, W. Jeon, and D. Jeong, "Base-Station Sleep Management in Open-Access Femtocell Networks," *IEEE Transactions on Vehicular Technology*, vol. 65, pp. 3786 – 3791, May 2016.

[104] L. Li, M. Peng, C. Yang, and Y. Wu, "Optimization of Base-Station Density for High Energy-Efficient Cellular Networks With Sleeping Strategies," *IEEE Transactions on Vehicular Technology*, vol. 65, pp. 7501 – 7514, Sep. 2016.

[105] C. Chang, W. Liao, H. Hsieh, and D. Shiu, "On Optimal Cell Activation for Coverage Preservation in Green Cellular Networks," *IEEE Transactions on Mobile Computing*, vol. 13, pp. 2580 – 2591, Nov. 2014.

[106] P. Kong, "Optimal Probabilistic Policy for Dynamic Resource Activation Using Markov Decision Process in Green Wireless Networks," *IEEE Transactions on Mobile Computing*, vol. 13, pp. 2357 – 2368, Oct. 2014.

[107] Z. Niu, X. Guo, S. Zhou, and P. R. Kumar, "Characterizing Energy–Delay Tradeoff in Hyper-Cellular Networks With Base Station Sleeping Control," *IEEE Journal on Selected Areas in Communications*, vol. 33, pp. 641 – 650, Apr. 2015.

[108] Z. Niu, X. Guo, S. Zhou, and P. Kumar, "Characterizing Energy–Delay Tradeoff in Hyper-Cellular Networks With Base Station Sleeping Control," *IEEE Journal on Selected Areas in Communications*, vol. 33, pp. 641 – 650, Apr. 2015.

[109] A. Alnoman and A. Anpalagan, "Towards the Fulfillment of 5G Network Requirements: Technologies and Challenges," *Telecommunication Systems*, vol. 65, pp. 101–116, May 2017.

[110] A. Alnoman, G. H. Carvalho, A. Anpalagan, and I. Woungang, "Energy Efficiency on Fully Cloudified Mobile Networks: Survey, Challenges, and Open Issues," *IEEE Communications Surveys & Tutorials*, vol. 20, pp. 1271 – 1291, 2nd Quart. 2018.

[111] J. Li, L. Huang, Y. Zhou, S. He, and Z. Ming, "Computation Partitioning for Mobile Cloud Computing in a Big Data Environment," *IEEE Transactions on Industrial Informatics*, vol. 13, pp. 2009 – 2018, Aug. 2017.

[112] A. Brogi and S. Forti, "QoS-Aware Deployment of IoT Applications Through the Fog," *IEEE Internet of Things Journal*, vol. 4, pp. 1185 – 1192, Oct. 2017.

[113] J. Ren, H. Guo, C. Xu, and Y. Zhang, "Serving at the Edge: A Scalable IoT Architecture Based on Transparent Computing," *IEEE Network*, vol. 31, pp. 96 – 105, Oct. 2017.

[114] S. Shah, E. Ahmed, M. Imran, and S. Zeadally, "5G for Vehicular Communications," *IEEE Communications Magazine*, vol. 56, pp. 111 – 117, Jan. 2018.

[115] R. Deng, R. Lu, C. Lai, T. H. Luan, and H. Liang, "Optimal Workload Allocation in Fog-Cloud Computing Toward Balanced Delay and Power Consumption," *IEEE Internet of Things Journal*, vol. 3, pp. 1171 – 1181, Dec. 2016.

[116] S. Hung, H. Hsu, S. Lien, and K. Chen, "Architecture Harmonization Between Cloud Radio Access Networks and Fog Networks," *IEEE Access*, vol. 3.

[117] T. Oo *et al.*, "Offloading in HetNet: A Coordination of Interference Mitigation, User Association, and Resource Allocation," *IEEE Transactions on Mobile Computing*, vol. 16, pp. 2276 – 2291, Aug. 2017.

[118] X. Meng, W. Wang, and Z. Zhang, "Delay-Constrained Hybrid Computation Offloading With Cloud and Fog Computing," *IEEE Access*, vol. 5.

[119] Y. Cui *et al.*, "Software Defined Cooperative Offloading for Mobile Cloudlets," *IEEE/ACM Transactions on Networking*, vol. 25, pp. 1746 – 1760, June 2017.

[120] X. Lyu *et al.*, "Selective Offloading in Mobile Edge Computing for the Green Internet of Things," *IEEE Network*, vol. 32, pp. 54 – 60, Feb. 2018.

[121] L. Tong, Y. Li, and W. Gao, "A Hierarchical Edge Cloud Architecture for Mobile Computing," in *IEEE INFOCOM*, pp. 1–9, Apr. 2016.

[122] T. H. Szymanski, "An Ultra-Low-Latency Guaranteed-Rate Internet for Cloud Services," *IEEE/ACM Transactions on Networking*, vol. 24, pp. 123 – 136, Feb. 2016.

[123] D. Xu, H. Jin, C. Zhao, and D. Liang, "Joint Caching and Sleep-Active Scheduling for Energy-Harvesting Based Small Cells," in *IEEE International Conference on Wireless Communications and Signal Processing (WCSP)*, pp. 1–6, Oct. 2017.

[124] L. Saker, S. E. Elayoubi, R. Combes, and T. Chahed, "Optimal Control of Wake Up Mechanisms of Femtocells in Heterogeneous Networks," *IEEE Journal on Selected Areas in Communications*, vol. 30, pp. 664 – 672, Apr. 2012.

[125] M. Islam, M. Razzaque, M. Hassan, W. Ismail, and B. Song, "Mobile Cloud-Based Big Healthcare Data Processing in Smart Cities," *IEEE Access*, vol. 5.

[126] L. Gkatzikis and I. Koutsopoulos, "Migrate or Not? Exploiting Dynamic Task Migration in Mobile Cloud Computing Systems," *IEEE Wireless Communications*, vol. 20, pp. 24 – 32, June 2013.

[127] M. Chowdhury, E. Steinbach, W. Kellerer, and M. Maier, "Context-Aware Task Migration for HART-Centric Collaboration over FiWi Based Tactile Internet Infrastructures," *IEEE Transactions on Parallel and Distributed Systems*, vol. 29, pp. 1231 – 1246, June 2018.

[128] H. Wu and K. Wolter, "Stochastic Analysis of Delayed Mobile Offloading in Heterogeneous Networks," *IEEE Transactions on Mobile Computing*, vol. 17, pp. 461 – 474, Feb. 2018.

[129] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications," *IEEE Communications Surveys & Tutorials*, vol. 17, pp. 2347 – 2376, 4th quar. 2015.

[130] J. Zhang *et al.*, "Energy-Latency Tradeoff for Energy-Aware Offloading in Mobile Edge Computing Networks," *IEEE Internet of Things Journal*, vol. 5, pp. 2633 – 2645, Aug. 2018.

[131] D. Xu, Q. Li, and H. Zhu, "Energy-Saving Computation Offloading by Joint Data Compression and Resource Allocation for Mobile-Edge Computing," *IEEE Communications Letters*, vol. 23, pp. 704 – 707, Apr. 2019.

[132] A. Bozorgchenani, D. Tarchi, and G. Corazza, "Centralized and Distributed Architectures for Energy and Delay Efficient Fog Network-Based Edge Computing Services," *IEEE Transactions on Green Communications and Networking*, vol. 3, pp. 250 – 263, Mar. 2019.

[133] Z. Sheng, C. Mahapatra, V. Leung, M. Chen, and P. Sahu, "Energy Efficient Cooperative Computing in Mobile Wireless Sensor Networks," *IEEE Transactions on Cloud Computing*, vol. 6, pp. 114 – 126, Mar. 2018.

[134] Q. Xu, Z. Su, Q. Zheng, M. Luo, and B. Dong, "Secure Content Delivery With Edge Nodes to Save Caching Resources for Mobile Users in Green Cities," *IEEE Transactions on Industrial Informatics*, vol. 14, pp. 2550 – 2559, Jun. 2018.

[135] T. Mekonnen *et al.*, "Energy Consumption Analysis of Edge Orchestrated Virtualized Wireless Multimedia Sensor Networks," *IEEE Access*, vol. 6, pp. 5090 – 5100, Feb. 2018.

[136] H. Wang, Y. Li, D. Jin, P. Hui, and J. Wu, "Saving Energy in Partially Deployed Software Defined Networks," *IEEE Transactions on Computers*, vol. 65, pp. 1578 – 1592, May 2016.

[137] Y. Wei, X. Zhang, L. Xie, and S. Leng, "Energy-aware Traffic Engineering in Hybrid SDN/IP Backbone Networks," *Journal of Communications and Networks*, vol. 18, pp. 559 – 566, Aug. 2016.

[138] N. Huin *et al.*, "Bringing Energy Aware Routing Closer to Reality With SDN Hybrid Networks," *IEEE Transactions on Green Communications and Networking*, vol. 2, pp. 1128 – 1139, Dec. 2018.

[139] K. Xie, X. Huang, S. Hao, and M. Ma, "Distributed Power Saving for Large-Scale Software-Defined Data Center Networks," *IEEE Access*, vol. 6, pp. 5897 – 5909, Jan. 2018.

[140] A. Alnoman and A. Anpalagan, "Computing-Aware Base Station Sleeping Mechanism in H-CRAN-Cloud-Edge Networks," *IEEE Transactions on Cloud Computing*.

[141] A. Alnoman and A. Anpalagan, "On Base Station Sleeping for Heterogeneous Cloud-Fog Computing Networks," in *IEEE 29th Biennial Symposium on Communications (BSC)*, pp. 1–4, Jun. 2018.

[142] M. Harchol-Balter, *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*. NY, USA: Cambridge University Press, 2013.

[143] S. Borst, A. Mandelbaum, and M. Reiman, "Dimensioning Large Call Centers," *Operations Research*, vol. 52, p. 17–34, Feb. 2004.

[144] L. Liu, S. Bi, and R. Zhang, "Joint Power Control and Fronthaul Rate Allocation for Throughput Maximization in OFDMA-Based Cloud Radio Access Network," *IEEE Transactions on Communications*, vol. 63, pp. 4097 – 4110, Nov. 2015.

[145] J. Tang, W. Tay, T. Quek, and B. Liang, "System Cost Minimization in Cloud RAN With Limited Fronthaul Capacity," *IEEE Transactions on Wireless Communications*, vol. 16, pp. 3371 – 3384, May 2017.

[146] K. Ahmed and S. Hranilovic, "C-RAN Uplink Optimization Using Mixed Radio and FSO Fronthaul," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 10, pp. 603 – 612, Jun. 2018.

[147] Z. Yan, M. Peng, and M. Daneshmand, "Cost-Aware Resource Allocation for Optimization of Energy Efficiency in Fog Radio Access Networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, pp. 2581 – 2590, Nov. 2018.

[148] R. Marler and J. Arora, "Survey of Multi-objective Optimization Methods for Engineering," in *Struct Multidisc Optim*, pp. 369 – 395, Mar. 2004.

[149] A. Kiani and N. Ansari, "Edge Computing Aware NOMA for 5G Networks," *IEEE Internet of Things Journal*, vol. 5, pp. 1299 – 1306, Apr. 2018.

[150] M. Kim, N. Kim, W. Lee, and D. Cho, "Deep Learning-aided SCMA," *IEEE Communication Letters*, vol. 22, pp. 720 – 723, Apr. 2018.

[151] S. Zhang *et al.*, "Sparse Code Multiple Access: An Energy Efficient Uplink Approach for 5G Wireless Systems," in *IEEE Global Communication Conference*, pp. 4782 – 4787, Dec. 2014.

[152] D. Zhai, "Adaptive Codebook Design and Assignment for Energy Saving in SCMA Networks," *IEEE Access*, vol. 5, pp. 23550 – 23562, Oct. 2017.

[153] J. Chen, Z. Wang, W. Xiang, and S. Chen, "Outage Probability Region and Optimal Power Allocation for Uplink SCMA Systems," *IEEE Transactions on Communications*, vol. 66, pp. 4965 – 4980, Jun. 2018.

[154] T. Adegbija, A. Rogacs, C. Patel, and A. Gordon-Ross, "Microprocessor Optimizations for the Internet of Things: A Survey," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, pp. 7 – 20, Jan. 2018.

[155] N. Abbas *et al.*, "Mobile Edge Computing: A Survey," *IEEE Internet of Things Journal*, vol. 5, pp. 450–465, Feb. 2018.

[156] E. Zeydan *et al.*, "Big Data Caching for Networking: Moving rom Cloud to Edge," *IEEE Communications Magazine*, vol. 54, pp. 36–42, Sep. 2016.

[157] Q. Fan and N. Ansari, "Application Aware Workload Allocation for Edge Computing Based IoT," *IEEE Internet of Things Journal*, vol. 5, pp. 2146 – 2153, Jun. 2018.

[158] Y. Gu, Z. Chang, M. Pan, L. Song, and Z. Han, "Joint Radio and Computational Resource Allocation in IoT Fog Computing," *IEEE Transnactions on Vehicular Technology*, vol. 67, pp. 7475 – 7484, Mar. 2018.

[159] W. Zhu, L. Qiu, and Z. Chen, "Joint Subcarrier Assignment and Power Allocation in Downlink SCMA Systems," in *IEEE 86th Vehicular Technology Conference (VTC-Fall)*, pp. 1 – 5, Sep. 2017.

[160] Y. Li *et al.*, "Cost-efficient Codebook Assignment and Power Allocation for Energy Efficiency Maximization in SCMA Networks," in *IEEE 84th Vehicular Technology Conference (VTC-Fall)*, pp. 1–5, Sep. 2016.

[161] S. Li, Q. Ni, Y. Sun, G. Min, and S. Al-Rubaye, "Energy-efficient Resource Allocation for Industrial Cyber-physical IoT Systems in 5G Era," *IEEE Transactions on Industrial Informatics*, vol. 14, pp. 2618 – 2628, Jun. 2018.

[162] H. Zhang *et al.*, "Computing Resource Allocation in Three-tier IoT Fog Networks: A Joint Optimization Approach Combining Stackelberg Game and Matching," *IEEE Internet of Things Journal*, vol. 4, pp. 1204 – 1215, Oct. 2017.

[163] M. Moltafet, N. Yamchi, M. Javan, and P. Azmi, "Comparison Study Between PD-NOMA and SCMA," *IEEE Transactions on Vehicular Technology*, vol. 67, pp. 1830 – 1834, Feb. 2018.

[164] L. Yang, X. Ma, and Y. Siu, "Low Complexity MPA Detector Based on Sphere Decoding for SCMA," *IEEE Communication Letters*, vol. 21, pp. 1855 – 1858, Aug. 2017.

[165] Y. Han, W. Zhou, M. Zhao, and S. Zhou, "Enabling High Order SCMA Systems in Downlink Scenarios with a Serial Coding Scheme," *IEEE Access*, vol. 6, pp. 33796 – 33809, July 2018.

[166] B. Di, L. Song, and Y. Li, "Radio Resource Allocation for Uplink Sparse Code Multiple Access (SCMA) Networks Using Matching Game," in *IEEE International Conference on Communications (ICC)*, pp. 1–6, May 2016.

[167] H. Nikopour and H. Baligh, "Sparse Code Multiple Access," in *IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 332 – 336, Sep. 2013.

[168] G. Tychogiorgos, A. Gkelias, and K. Leung, "A Non-convex Distributed Optimization Framework and its Application to Wireless Ad-hoc Networks," *IEEE Transactions on Wireless Communications*, vol. 12, pp. 4286 – 4296, Sep. 2013.

[169] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[170] K. Au *et al.*, "Uplink Contention Based SCMA for 5G Radio Access," in *IEEE Globecom Workshops (GC Wkshps)*, pp. 900–905, Dec. 2014.