Evaluation of Power Consumption and Trade-offs in 5G Mobile Communications Networks

By

RAAD ALHUMAIMA

A Thesis Submitted in Partial Fulfilment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Department of Electronic and Computer Engineering

College of Engineering, Design and Physical Sciences

BRUNEL UNIVERSITY LONDON

June 2017

To my beloved family and friends.

ACKNOWLEDGMENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SYMBOLS

I_{BB}	Set of BBU function
P_{BBU}	PC of the BBU
$P_{i,BB}^{ref}$	BBU's sub component PC
x_{act}	Actual value of the power consumed
x_{ref}	Reference value of the parameter consumption
s_i	Scaling vector
$P_{DC,P}$	BBU pool DC-DC Conversion PC
B	Total number of active BBUs
R	Total number of active RRHs
$P_{BBU_b^r}$	PC of b-th BBU, connected to r-th RRH
$P_{opt,P_{b,r}}$	PC of optical device connects b-th BBU to r-th RRH
$l_{DC,P}$	Loss caused by DC-DC conversion
$\eta_{DC,P}$	DC conversion efficiency
ϖ	Exponential decay constant
l_o	Initial losses at time $(t = 0)$
η	Efficiency of the device
$P_{DC,P,out,max}$	Maximum DC conversion output power
$P_{MS,P}$	Power consumption of AC-DC-Conversion
$l_{MS,P}$	Losses of MS power conversion
$\eta_{MS,P}$	AC-DC conversion efficiency
$P_{MS,P,out,max}$	Maximum AC converter output power

$lcool$	Cooling loss
$Pcool$	Cooling power consumption in the BBU pool
P_{out}	Antenna output transmission power
P_{TX}	Output power of the power amplifier
RB	Number of the Physical Resource Blocks
P_{PA}	Power consumption of PA
η_{PA}	PA efficiency
$P_{DC,R}$	Power consumption of the RRH DC-DC conversion
$\eta_{DC,R}$	RRH DC-DC conversion efficiency
A	Total number of antenna or RF chains
$P_{RF_a}^{r,b}$	PC of the a-th RF of r-th RRH, connected to b-th BBU
$P_{PA_a}^{r,b}$	PC of the a-th PA of r-th RRH, connected to b-th BBU
$P_{opt,R}$	PC of the optical device inside the RRH
$\eta_{DC,R}$	DC-DC conversion efficiency
$P_{DC,R,out,max}$	Maximum output power of the DC-DC converter
$l_{MS,R}$	AC-DC losses
$\eta_{MS,R}$	AC-DC conversion efficiency
$P_{MS,R,out,max}$	Maximum output power of the AC-DC converter
P_{CRAN}	C-RAN total PC, single unit
P_{Pool}	Total PC of BBU pool
$P_{BBU_{b,r,a}}$	PC of b-th BBU; of r-th RRH; mounts a-th antenna
$P_{PA_{r,a}}^{b}$	PA PC of a-th antenna's, of r-th RRH; of b-th BBU
$P_{RF_{r,a}}^{b}$	RF PC of a-th antenna's, of r-th RRH; of b-th BBU
P_{supply}	C-RAN total PC, multi unit
P_{MME}	Power consumption of MME
P_{SGW}	Power consumption of SGW
P_{PGW}	Power consumption of PGW
P_{cl}	Power consumption of control plane
M_{MME}	Set of MME functions

SG_{SGW}	Set of SGW functions
PG_{PGW}	Set of PGW functions
Q	Total number of MME
D	Total number of SGW
G	Total number of PGW
v_q	Processing speed of MME (pps)
v_d	Processing speed of SGW (pps)
v_g	Processing speed of PGW (pps)
$E_{m,MME}^{ref}$	Energy consumption of m-th MME function
$E_{sg,SGW}^{ref}$	Energy consumption of sg-th SGW function
$E_{pg,Pgw}^{ref}$	Energy consumption of pg-th PGW function
y_m^A	Scaling exponent of the number of RF chains of MME
z_{sg}^A	Scaling exponent of the number of RF chains of SGW
j_{pg}^A	Scaling exponent of the number of RF chains of PGW
y_m^B	Scaling exponent of bandwidth of MME
z_{sg}^B	Scaling exponent of bandwidth of SGW
j_{pg}^B	Scaling exponent of bandwidth of PGW
P_{SDN}	SDN power consumption
P_{switch}	O.F switch power consumption
P_{port}	Port power consumption
$E_{prt,of,sw}$	Energy consumed at full speed by sw-th port of switch of
v_{sw}	Line speed corresponding the sw-th port
OF	Total number of O.F switches
SW	Total number of O.F ports
P_{flow}	O.F traffic power consumption
FL	Total number of flows
R_{pkt}	O.F flows packet rate (packets per sec)
MC	Total number of flows that matching
NC	Total number of flows that non-matching

ACT	Total number of flows that actioned
MT	Total number of matched packets per flow
DS	Total number of non-matched packets per flow
AK	Total number of actioned packets per flow
E_{mt}	Energy required to process mt-th match packet
E_{ds}	Energy required to process ds-th non-match packet
E_{ak}	Energy required to process ak-th actioned packet
P_{SDNctl}	SDN controller power consumption
$R^{O.F}_{ctl,of}$	Rate of O.F signalling from the ctl-th controller to of-th switch
$R^{MME}_{ctl,q}$	Rate of O.F signalling from the ctl-th controller to q-th MME
$R^{SGW}_{ctl,d}$	Rate of O.F signalling from the ctl-th controller to d-th SGW
$R^{PGW}_{ctl,g}$	Rate of O.F signalling from the ctl-th controller to g-th PGW
CTL	Total number of SDN controllers
$P^{sdn}_{DC,P}$	BBU pool DC-DC Conversion with SDN devices
$P_{SDC-RAN}$	Total power consumption of SDN based C-RAN
P_{CN}	Power consumption of core network
$\sigma_{fronthaul}$	Fronthaul loss factor
$\sigma_{backhaul}$	Backhaul loss factor
P^{pm}_{SDCRAN}	SDN based C-RAN power consumption in the parameters model
P^{pm}_{switch}	Power consumption of O.F switch in the parameters model
P^{pm}_{MME}	Power consumption of MME in the parameters model
P^{pm}_{SGW}	Power consumption of SGW in the parameters model
P^{pm}_{PGW}	Power consumption of PGW in the parameters model
P^{pm}_{pool}	PC of BBU pool in the parameters model
P^{pm}_{RRH}	Power consumption of RRH in the parameters model
P^{pm}_{BBU}	Power consumption of BBU in the parameters model
P^{pm}_{RF}	Power consumption of RF in the parameters model
N	Total number of virtual machines
Z_{RAM}	RAM size

dZ_{RAM}	Change in RAM size
dN	Change in number of virtual machines
Z_{int}	Initial RAM size
$P_{RAM}(Z_{RAM})$	PC of the RAM as a function of RAM size
α	Constant of RAM size
β	Constant RAM power consumption
dP_{RAM}	Change in RAM power consumption
P_{intRAM}	Initial PC of the RAM
Z	RAM size
i	RAM size index
P_{RAM}^i	PC of RAM holds i index
dP_{RAM}^i	Change in PC of RAM holds i index
ν	Constant of RAM power consumption change
$P_{RAM}^i(N)$	PC of RAM holds i index as function of N
P_{intram}^i	Initial PC of RAM size with index i
C	Total number of cores per CPU
N_C	Number of VMs in one core
P_{core}	CPU core PC
$P_{intcore}$	initial core PC
ε	Core PC constant
P_{CPU}	CPU power consumption
P_{server}^{CPU}	PC of CPUs per server
K	Total number of CPUs per server
L	Total number of NICs
P_{nic}	Power consumption of NIC
P_{intnic}	Initial PC of NIC
dP_{nic}	Change in NIC power consumption
γ	Constant of NIC power consumption
P_{server}^{NIC}	Total PC of NICs per server

t	Time of storage utilisation
$P_{storage}(t, N)$	Storage PC as a function of VMs and time
$P_{storage}^{i}$	Storage PC increases by time
δ	Increasing storage PC constant of time
P_o	Initial storage PC
$P_{storage}^{d}(t)$	Storage PC decreases by time
ξ	Storage PC —constant of N
AS	Number of accesses tasks
PAS	PC per task per VM
P_{HV}	Power consumption of Hypervisor
$PAS_{(as,n)}$	PC of as-th job allocated to n-th VM
P_{server}^{BBU}	vBBU server PC
$S1$	Total number of BBU servers
P_{server}^{cl}	Total PC of control plane servers
$S2$	Total number of servers that host the control planes
P_{cl}^{CPU}	Power consumption of control plane CPUs
P_{MME}^{CPU}	Power consumption of MME's CPU
P_{SGW}^{CPU}	Power consumption of SGW's CPU
P_{PGW}^{CPU}	Power consumption of PGW's CPU
P_{core}^{MME}	Core PC of MME
P_{core}^{SGW}	Core PC of SGW
P_{core}^{PGW}	Core PC of PGW
$C1$	Total number of BBU's cores
$C2$	Total number of MME's cores
$C3$	Total number of SGW's cores
$C4$	Total number of PGW's cores
$K1$	Total number of BBU's CPUs
$K2$	Total number of MME's CPUs
$K3$	Total number of SGW's CPUs

$K4$	Total number of PGW's CPUs
RB_n	Resource blocks of each VM
ϑ	Increment factor due to processing RB_n
P_{server}^{vCN}	Total PC of the virtualised core network
mcs	Modulation and coding scheme index
τ_{bare}	Execution time of a bare BBU server
τ_{init}	Initial BBU delay
τ_{HV}	HV delay
τ_v	Execution time of virtualised server with 1 VM
τ	Total execution time of all VMs
$\tau_v^{s,n}$	Total execution time of VM n of server s
P_{vCRAN}	Total PC of virtualised C-RAN
P_{cost}^{source}	Power cost of the source server
utl	Sever utilisation
$scof$	Source server constant factor
P_{cost}^{dest}	Power cost in the destination server
$rcof$	Destination server constant factor
S_{smig}	Total number of migration source servers
S_{rmig}	Total number of target servers
N_{smig}	Total number of migrated VMs
N_{rmig}	Total number of received VMs
P_{vCRAN}^{mig}	PC of live migration based virtualised C-RAN
τ_{mig}	Time cost of VM migrating
d_{mig}	Migration distance
v_{mig}	Migration speed
τ_T	Total delay of virtualisation and migration
$p1, p2, ...$	Source server coefficients

ABBREVIATIONS

ADC	analogue to digital converters
API	application programming interfaces
BBU	base band unit
BS	base station
CAPEX	capital expenditure
CA	carrier aggregation
CN	core network
CN	core network
CO-D	coherent detection
CPU	central processing unit
CPRI	common public radio interface
C-RAN	cloud radio access network
DAC	digital to analogue converters
D-RAN	distributed RAN
DD	direct detection
EE	energy efficiency
eNodeB	evolved NodeB
EPC	evolved packet core
EPS	evolved packet system
E-UTRAN	evolved universal telecommunication radio access network
FEC	forward error correction

FDM	frequency domain multiplexing
FFT	fast fourier transform
GOPS	giga operation per second
GSM	global system for mobile communications
HDD	hard disk Drive
HetNet	heterogeneous network
HSPA	high speed packet access
HSS	home subscriber server
HV	hypervisor
LTE-A	LTE Advanced
LTE	long term evolution
LNAs	low noise amplifiers
MIMO	multiple-input multiple-output
MME	mobility management entity
mmWave	millimeter wave
NAS	network attached storage
NFV	network function virtualisation
NIC	network interface controller
OF	open flow
OFDMA	orthogonal frequency division multiple access
OPEX	operational expenditure
ORM	optical receiver module
OSI	open systems interconnection
OTM	optical transmitter module
PC	power consumption
PCRF	policy and charging rules function
PGW	packet gateway
PM	power model
pm	parameters model

pps	packets per second
PtP	point-to-point
PtMP	point-to-multipoint
QoS	quality of service
RF	radio frequency
RAM	random access memory
RRH	remote radio head
SAE	system architecture evolution
SC-FDM	single carrier-frequency domain multiplexing
SDN	software defined network
SDC-RAN	software defined cloud radio access network
SE	spectral efficiency
SGW	serving gateway
SNR	signal-to-noise ratio
UE	user equipment
UMTS	universal mobile telecommunication system
VM	virtual machine
VCO	voltage controlled oscillators
vBBU pool	virtual BBU pool
3GPP	third mobile generation partnership project
5G	fifth generation
4G	fourth generation

Abstract

In this thesis, components and parameters based power models (PMs) are produced to measure the power consumption (PC) of cloud radio access network (C-RAN) architecture. In components PM, the power figure of each component within C-RAN is evaluated. After, this model is parametrised such that the computation complexity of each component is converted to a straightforward, but accurate method, called parameterised PM. This model compares cooling and total PC of traditional LTE architecture with C-RAN. This comparison considered different parameters such as, utilised bandwidth, number of antenna, base band units (BBUs) and remote radio heads (RRHs). This model draws about 33% reduction in power. Next, this PC model is updated to serve and exhibit the cost of integrating software defined networks (SDNs) with C-RAN. Alongside, modelling the power cost of the control plane units in the core network (CN), such as serving gateway (SGW), packet gateway (PGW) and mobility management entity (MME). Although there is power cost, the proposed model shows the directions to mitigate it. Consequently, a simplified PM is proposed for virtualisation based C-RAN. In this model, the power cost of server virtualisation by hosting several virtual machines (VMs) is shown, in a time and cost effective way. The total reduction in the PC was about 75%, due to short-cutting the number of active servers in the network. Alongside, the latency cost due to such technique is modelled. Finally, to enable efficient virtualisation technology, live migrating the VMs amongst the servers is vital. However, this advantageous situation is concurrent with VM's migration time and power cost. Therefore, a model is proposed to calculate the power cost of VM's live migration, and shows the effect of such decision upon the total PC of the network/C-RAN. The proposed work converts the complexity of other proposed PMs, to a simplified and costless method. Concurrently, the time cost is added to the imposed virtualisation's time cost to formulate the total delay expected prior to these techniques' execution.

Chapter 1:   Introduction

1.1   Overview

The increasing number of connected devices, each coupled with an abundance of new types of bandwidth-hungry applications and services, has resulted in high data rate demands [1]. According to Cisco and Ericson's recent forecast, mobile portable devices and connections will grow to 11.5 billion by 2019, and thereon to more than a ten-fold increase in mobile data traffic between 2013 and 2018 [1]. Furthermore, mobile device connections will grow to about 10.5 billion by 2018 compared to 7.2 billion in 2013 [1], [2]. Additionally, Ericsson reportedly forecasts that in 2021, 150 billion devices will be 5G connected, up from 4.100 billion connections using LTE technology [3]. To serve such demand, more base stations (BS), such as Macro, Pico and Femto are installed, which result in more consumed power.

Long term evolution (LTE), also called (3GPP, Release 8) represents the latest version of third mobile generation partnership project (3GPP). The idea behind this release is to cover the shortcoming of older generations, such as global system for mobile communications (GSM), universal mobile telecommunication system (UMTS), high speed packet access (HSPA), etc. These systems compete for at least 10 years before LTE emerges. The main differences amongst these systems are that LTE uses IP packet architecture and introduces new air interface which is orthogonal frequency division multiple access (OFDMA). Furthermore, it defines new quality of service (QoS) bearers to guarantee the user equipment (UE) requirements. LTE system is launched to compete for at least 10 years. The main objectives of this system were:

- Reduced latency.

- Packet oriented communication network architecture.

- Flexible frequency bands between 1.25-20 MHz.

- Higher data rates, up to 100 Mbps in downlink, and about 50 Mbps in uplink transmission.

- Using OFDMA multiple access technique for downlink and single carrier single carrier- frequency division multiple access (SC-FDMA) for uplink transmission.

1.2   LTE Architecture

In LTE, the evolution of legacy UMTS system has resulted what is called evolved universal telecommunication radio access network (E-UTRAN), together with system architecture evolution (SAE) which includes the evolved packet core (EPC), they comprise evolved packet system (EPS) [4]. Packet system means the data is divided into packets before passing it to the circuit to be received by the other host, each packet may take different route in the network via utilising routing algorithms to reach its final destination. This is in contrast to the circuit switching systems when the rout is static and pre-established prior to connection initialisation. Mainly, LTE architecture consists of three parts, these are the users (UEs), E-UTRAN and EPC, as shown in Fig 1.1. E-UTRAN consists of many eNodeBs, which offer the air interface to the UEs, precessing the received data and offers the necessary resources scheduling. When the call is established, the UEs' data plane using (S1-U) interface will be directed from the eNodeB to the serving gate way (SGW), then to the other eNodeB where the other UE resides. On the other hand, the mobility management entity (MME) is updated by the eNodeB using the control plane interface (S1-MME). However, the MME is responsible for updating the SGW about UE's position, its data base and necessary mobility functions. The SGW in turn is responsible mainly for forwarding/routing UEs' packets to the required destination. If the UE requires to

connect to the other networks or to the Internet, then the SGW forwards the packets to the packet gate way (PGW). Successively, PGW is responsible for many functions, for example, packets filtering, IP allocation and inter-operator charging.

1.2.1 Users (UEs)

The UE terminology can be interpreted as mobile device, laptops/computer, tablets, etc. The UE is the only device that can be seen and explored by the costumers, while the other network's parts are run by the network service operators.

1.2.2 Evolved universal telecom. radio access network (E-UTRAN)

The second part consists of many evolved NodeBs (eNodeBs), which are connected and communicated to each others through X2-interface for necessary handover operations, while S1 interface is used to connect these eNodeBs to the EPC. The eNodeB works as a bridge to connect the UE with the core network (CN) and provides the necessary protocols. The eNodeB is also responsible for scheduling the resources amongst the different UEs in both time and frequency dimensions while ensuring the required QoS for the UEs. In addition, eNodeB has the mobility management functionality such as handover signalling and radio link measurements.

1.2.3 Evolved packet core (EPC)

Also known as core network (CN), EPC included three main entities: MME, SGW and PGW. Furthermore, some logical entities can be considered as part of the CN, such as policy and charging rules function (PCRF), and home subscriber server (HSS). Each entity in turn is responsible for different functions. For example, MME is responsible for inter CN mobility signalling with other 3GPP networks, authentication, authorization, time zone signalling and bearer management, which

includes bearers activation/deactivation procedures. An EPS bearer means how the UE traffic will be treated when passing through the network.
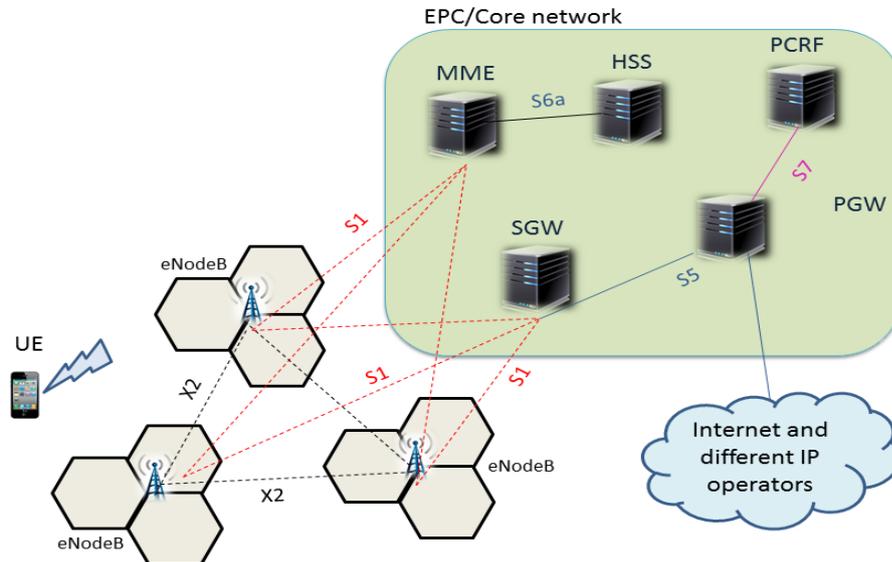
Fig. 1.1. LTE E-UTRAN network architecture [5]. It shows the three main parts of LTE: UE, E-UTRAN and EPC. E-UTRAN contains many eNodeBs, while EPC encompasses MME, SGW, PGW, HSS and PCRF units.

1.3 Beyond LTE

There have been several techniques to enhance LTE system to an advanced version, called LTE-Advanced (LTE-A), or (3GPP, Release 10). These techniques were utilised by LTE to provide higher data rates, wider coverage, higher throughput and lower latency, which result in higher UE satisfaction [6]. In addition, LTE-A is able to support heterogeneous networks, where the low power nodes containing Pico, Femto cells and relays are placed within the Macro cell layout. However, these technologies can be summarised as follows:

1. Carrier aggregation (CA)

   This technique is proposed to achieve about 1 Gbps peak rate by combining disparate spectrum bands into a wider bandwidth to serve a single UE [7]. For example, if a service provider (SP) own 10 MHz of bandwidth in the 800 MHz spectrum and 20 MHz in the 1900 MHz, then it is possible to aggregate both bands to construct a 30 MHz bandwidth. This number can grow to an upper limit, which is 100 MHz. There are two types of CA techniques [8]: continuous CA, when the carrier components are adjacent. The second type is called non-continuous CA, when the multiple carrier components are detached over the band width. Generally, the first type is easier to implement. It does not require changing in the LTE physical layer texture, therefore, a single module fast fourier transform (FFT) and single radio frequency (RF) component can be used while providing the backward compatibility to LTE system. Furthermore, this has impacted easier management algorithms and resource allocation in the first type compared to the second type. On the other hand, due to the scarce available resources in low bandwidth, practically implementing the continuous CA is more difficult. Therefore, the second type is more practical to enable the operators to utilize their available spectrum, including the scattered and unused bands, and those bands which are already assigned for the legacy systems.

2. Advanced multiple-input multiple-output (MIMO)

   Multiple-input multiple-output (MIMO) pointed to a communication systems when the transmitter and receiver included multiple antennas. Using MIMO technology was the key success of LTE and some other rival systems, LTE MIMO system was up to four layers. Now on, it continues with LTE-A. According to LTE-A requirements, the maximum required spectral efficiency urges using 8×8 spatial multiplexing. Unlike LTE, the uplink MIMO is adopted in LTE-A for a single UE with up to four layers. [9].

3. Coordinated multi point (CoMP) transmission and reception

Coordinated multi point (CoMP) (3GPP, Release 11) is a key LTE-A technology that effectively increases the average capacity of cells, improves the signal-to-noise ratio (SNR) and offers optimised utilization of the available system resources and spectral efficiency. CoMP enables the communication amongst multiple eNodeBs to dynamically avoid the interfering transmission signals. Accordingly, the downlink efficiency is dramatically enhanced as the transmitted signals by multiple eNodeBs coordinate to revoke the mutual inter-cell interference [10]. Since same spectrum is used by all BSs' sectors, this leads to increased interference at the cell edge UEs while signals are received from multiple BSs. By using COMP, the sectors of a single BS can cooperate using intra-site, while the inter-site type happens amongst multiple/adjacent eNodeBs. COMP is applied in both downlink and uplink. However, All comes with increased backhaul demand which includes low latency, high cell capacity, increased synchronization, higher complexity, increased channel estimation and power consumption (PC) [11].

4. Heterogeneous networks (HetNet) and relay nodes

Relay nodes are deployed at the cell edges to extend and provide better capacity and coverage. These low power stations operate as repeaters, it's purpose is to rebroadcast the received/transmitted signals and to improve signal quality [12]. The relays are connected to the eNodeBs via wireless links. Such existence offers magnificent savings in the cost when compared to deploying new eNodeBs. This concept ties into the idea of heterogeneous network (HetNet). Towards boosting the network capacity and coverage, HetNets enable varying sizes of cells, each with differentiated radio access technology and output power to work/cooperate together. The small size cells can be deployed within the coverage of large size cells to enable serving increased number of UEs at the same time and improve the service to the cell edge UEs.

Other than the aforementioned technologies, there are other several evolutionary proposals that are suggested for LTE-A, for example self-organizing networks, cognitive radio and enhanced inter-cell interference coordination [13].

1.4   Beyond 4G

Mobile operators continuously seek out innovative ideas, designs, protocols and advanced digital signal processing (DSP) techniques in order to effectively cope with the explosive high demand for data, while simultaneously providing scalable and faster connectivity [14].

It was promised by 2020, the 5G wireless and mobile communication systems have to provide about 1000 times more capacity than today, reducing up to 90 % of the consumed energy per service, offering 1000 Gbps per km^2 more spectral capacity in dense areas, 10 times more connected devices battery life, and 5 times less end-to-end latency in comparison with the current forth generation (4G) system [15], [16]. However, most of the futuristic algorithms tend to rely on over-provisioning of resources, ensuring that the desired demands are met, such as spatial densification and spectral aggregation [17]. This leads to high PC and thus, high costs for the network operators. To exceed such cost, new paradigms and network designs become a must [18]. These emerging technologies are variant in purpose, herein, some of these key technologies for 5G are briefly described as follows [19]:

1. Cloud radio access networks (C-RANs)

   C-RAN architecture has been suggested by both operators (e.g., NTT, KT, France -Telecom/Orange. Telefonica, SoftBank/Sprint, and China Mobile), as well as equipment vendors (e.g., Alcatel-Lucent, Light Radio, Nokia-Siemens,Liquid Radio).

   C-RAN is addressed as one of the prerequisite technologies to enable high performance 5G networks [20]. C-RAN architecture is an advanced version of the traditional network paradigms, which brings the concept of cloud computing

to the mobile systems [21], [22]. C-RAN consists of many remote radio heads (RRHs) and base band unit (BBU) pool, the low power RRHs are distributed and connected either via high-bandwidth optical fibers or wireless links to the BBU pool [23], [24]. The latter in turn represents the host/cloud/data centre for hundreds of BBU servers. This deployment guarantees a reduction in both operational (OPEX) and capital (CAPEX) expenditures, because of the reduced site visits, maintenance and leases, which eternally reduce the total cost required to operate the network [25]. Moreover, it allows the non static allocations amongst BBUs and RRHs, enabling the off-loading algorithms to share processing load through any neighbouring BBUs, so as some of the BBUs can be switched off to save energy. This collaboration in the pool enhances the system's throughput, spectral efficiency (SE) and energy efficiency (EE) [26]. Moreover, cooling requirement can be reduced, as in C-RAN there will be one or few distributed pools in very large geographical area. In each pool, there will be only few cooling units for many BBUs, while in the traditional BS sites, each BBU requires a separate cooling unit.

2. Heterogeneous C-RAN (H-CRAN)

C-RAN can also integrates the existing Macro BS networks deployment to yield Heterogeneous based Cloud Radio Access Networks (H-CRAN). In which, the deployed Macro BS cooperates and coexists with the proposed C-RAN scenario [23]. By doing so, the low-PC RRHs are deployed and cooperated with each other in the BBU pool, whilst the Macro BSs are interfaced with the BBU pool by the means of S1 and X2 protocols for data and control plane respectively.

H-CRAN aims delivering cooperative services in the dense areas. H-CRAN makes use of both high data rates provided by the RRHs to the UEs with high QoS. Besides, securing the coverage to the UEs with low QoS requirements [27]. Accordingly, the challenges faced the traditional HetNets are mitigated. In the latter, coordinating the transmission/reception amongst the small cells (Pico

and Femto) requires a huge signalling for handover processes and mitigating the inter-tier interferences with the large coverage Macro BSs; when the wireless connection capacity is always constrained. Furthermore, the densely deployed small cells improve the capacity, but with increased energy consumption, which degrades the EE performance. On the other hand, H-CRAN architecture combines and takes advantage of both C-RANs and HetNets. The RRHs collaborate in the pool to avoid the signalling cost with the large BS and gain high cooperation, whilst the pool is linked to the Large BS to coordinate the inter-tier interference [23].

3. Software defined network (SDN)

SDN is a promising step to modify the way networks' control plane is dealt with so as to offer enhanced communication [20]. Traditionally, each network device contains both data and control plane functions. SDNs aim to separate the control functions of each distributed forwarding device and place them in a centralized controllers [28], as shown in Fig. 1.2. This enables running the control plane by only software to relief the complexity of these forwarding devices and replace them with less intelligent open flow switches [29]. The controller will be responsible for installing the flow forwarding and processing rules in the data plane devices and collecting network status. As such implementation offers a global view for the network, it helps to dynamically manipulate and reduce the utilised network resources, innovate new policies and services, offering flexible network administration. Subsequently, it provides smarter scheduling to the network links which eventually enhances the throughput, network maintenance and hardware cost [30], [31]. Furthermore, the controller is able to update the switches intelligently while integrating other mobile control plane entities, such as MME, SGW and PGW.
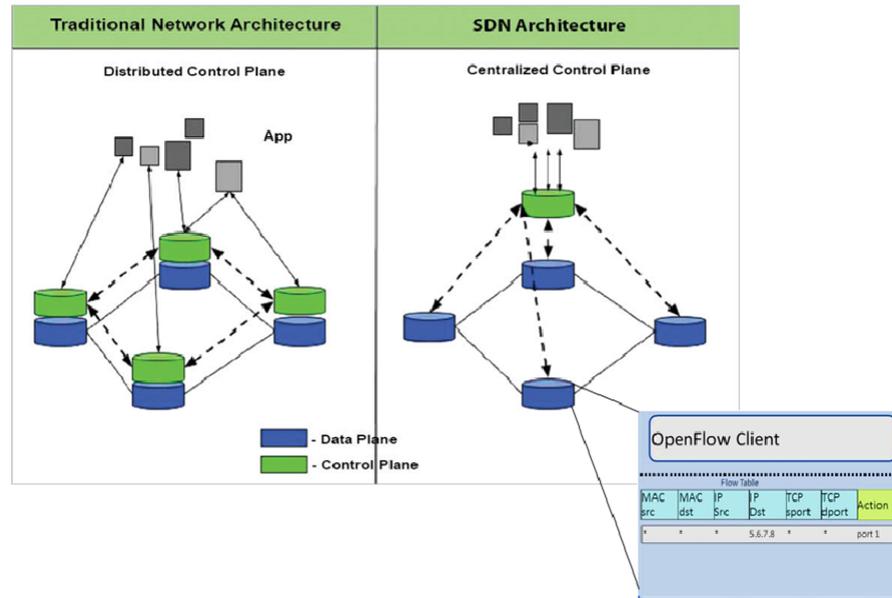
Fig. 1.2. Difference between the traditional and SDN network architecture.

4. Network function virtualisation (NFV)

Further to control-data planes separation by SDN, NFV decouples the physical hardware from the network functions that are installed on them, which means that these functions are dispatched to the SPs in a form of plain software or software stack instance [32]. In this case, running fewer virtualised servers in the data center to run the whole network becomes possible while fulfilling the UEs' QoS requirements [15], [33], [34], [35]. Recently, the research community embraced the use of NFV techniques in the cloud for several reasons, such as flexible allocations for the network resources, enabling flexibility in the servers' operation and configuration, reducing the maintenance cost and total network cost, supporting multi-tenancy, and potential reduction in the energy cost. Therefore, NFV have been devoted to provide significant increase in the EE, which allows SPs to execute network's functions using software rather than

running proprietary built or dedicated appliances. Due to the latter, updating
and expressing new services and applications that are necessary to enable 5G
is increasingly intractable. NFV also enables the use of general off-the-shelf
servers to run these functions' software, also known as virtual machines (VMs).
Fig. 1.3 shows the concept of virtualisation when several VMs are sharing the
virtualised server's resources.
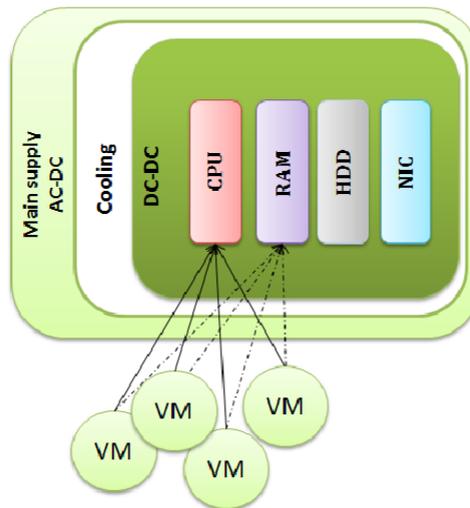
Fig. 1.3. The concept of virtualisation, several VMs sharing the server resources.

5. Live Migration of VMs

VM live migration technique is considered as a drive to improve the perfor-
mance of the networks. Simply, it is possible to migrate a VM from one server
to another, while keeping this VMs running/alive during the transfer process.
This idea offers a spatial oriented services. This also can offer on demand ser-
vices to the network providers [36]. Once a VM is seamlessly transferred, the
QoS of the UEs can be enhanced as it is been served by a near located ser-

vice instance/VM, rather than a far located and propriety based servers. This technique also provides resilience while performing hardware maintenance and software updates. The running applications and services can be migrated to other servers without lack of service availability [37].

6. Millimeter Wave(mmWave)

The wireless networks are greatly constrained to a narrow range of frequencies while operating, this range is around hundreds of MHz to few GHz. By now, this band is almost fully occupied during peak times and peak markets. Regardless how efficient are the optimisation methods nowadays, the need for more bandwidth is urgent [38]. Fortunately, mmWave spectrum range (lies between 30–300 GHz) can offer a wide range of frequencies to use and integrate with mobile and Wifi networks transmission. Nevertheless, until recently it is deemed difficult to implement such band for several reasons, such as its propagation quality including path loss, rain and atmospheric absorption, less diffraction and penetration while facing objects, this makes their application limited to the short range transmission. However, using these bands in radio over fiber technologies seems more promising than transmitting in pure wireless links. In radio over fiber, the light is modulated using a radio signal, then the modulated signals are transferred via an optical fiber and received at the receiver to enable wireless access. Further to providing the required bandwidth, another two advantages can be counted: the low cost of such technology [39], and the simplification in the design of BSs. This is facilitated by generating mmWave signals in the central office or the cloud and reuse the same wavelengths in the BS, so as no laser is required at the cell site [40].

1.5 Aim and Motivation

Amongst the different aspects, trends and challenges that are concurrent with 5G paradigms, such as improving the EE, optimising the radio planning, minimising the

latency, reducing the PC is a major. The electricity bills for the SPs and customers call charges, are both increasing linearly with the amount of provided data and type of service. The SPs are willing to provide an improved QoS to the UEs by innovating new technologies and algorithms, and increasing the number of deployed cells. However, this upgrades their PC and therefore their monthly bills, which results in harming the enviroment and increasing the global warming. This also reflects on increasing the customers charges to compensate such lost. Therefore, the PC is a crucial factor to be focused on in the future and hybrid/heterogeneous networks. As 80 % of the power required to operate the networks is consumed by the BS [41], C-RAN, SDN based C-RAN and virtualised networks represent the futuristic power reduction schemes. Within these networks, reducing the PC by using BBUs cooperation, PC reduction algorithms and offloading techniques is also paramount. However, these methods cannot be validated unless their PC and trade-offs are known.

1.6   Research Challenges and Objectives

The aim of any research is to improve/innovate over the state of the art (SotA) techniques so as the network performance is enhanced. Improving the EE represents one of the main affecting factors, which included a combination of enhancing both the bit rate of the system concurrent with reducing the PC. However, rather than proposing new algorithm to reduce the PC, offering a reliable method to measure the PC of the networks has been given a huge interest in the past few years. The following points briefly demonstrate why such evaluation is very important:

- In case of proposing to add new components to the SotA network design, this addition might enhance the network flexibility, programmability or system bit rate. At the same time, it might add considerable energy consumption. Therefore, an EE evaluation is needed to calculate the cost and judge the consumed power.

- The advanced proposals which facilitate coordination and cooperation in the network, for example, dynamically allocating the network resources based on traffic demand, enabling new services in the network, balancing the load amongst whatever servers found in the cloud, and sleep mode some of the units in the network. These methods require a specific device or server to run, this server then adds a considerable energy to the system and degrades the EE. Eventually such unpredicted energy cost might exceed the proposed method's gain, which urges to test such cost prior to implementation.

- To enhance the EE of any SotA system, a general platform for PC evaluation should be exited, so as the PC measurement of any other proposed systems is compared to it and the power gain is known.

- Some innovations enhance the EE but degrades the network performance. However, by evaluating the PC, a decision can be made to decide if it is worthwhile to compromise the network performance with the amount of reduced power.

- Understanding the way VMs consume power can help developing new methods to optimise their resources allocation and improve the HV scheduling amongst these VMs.

- When deciding to migrate a VM, a decision can be made according to the power cost. However, migrating a VM can cause the system to consume more power, this lost might not be justified by the migration benefits/gain. Nevertheless, by knowing the power consequences quickly through a power model (PM), an optimised decision can be made.

The studies concerning the PC usually attempt to describe such consumption in an easy, but accurate way, as SPs do not wish to spend much time and cost to only measure how much their network consumes, which can be easily estimated from the electricity bills they receive every month. Rather, they allocate more time and cost to develop new algorithms and optimisation techniques to reduce the PC and plan

the energy of their appliances and network. These techniques should be employed up on wherever within the network/device the enlarged consumption takes place. On the other hand, the power hungry components and devices cannot be known unless their power characteristics is evaluated, herein, it comes the effectiveness of the PM to show where the innovations should focus. Describing how the device consumes power is another point of conflict and PMs are variant regarding the complexity. In the literature, it is necessary to overcome the computational complexity of PC calculation to a simpler, accurate and parameterised models.

To exceed these challenges, the following objectives have been concentrated on:

- Modelling the PC of C-RAN paradigm, such modelling is important to estimate the amount of PC reduction in comparison to the SotA networks, and to judge the futuristic PC reduction techniques.

- Modelling the PC of software defined based cloud radio access networks (SDC-RAN) paradigm, this modelling is necessary to evaluate how much power is increased by adding SDN devices to C-RAN. Furthermore, it offers a general platform to evaluate the rising SDN concept and its innovated algorithms in the cloud.

- Modelling the PC of the virtualised C-RAN, this can estimate how much power can be reduced when the virtualisation techniques are implemented in the cloud. This is concurrent with modelling the trade-offs of virtalisation the servers, especially the latency. This is necessary to judge such technology which is a promising for low PC networks.

- Modelling the power and time cost of live migrating the VMs. This modelling can offer a judgement tool to decide about the EE of the networks while using this technology. Furthermore, the expected trade-offs can be known.

1.7   Thesis Contributions

1. In Chapter 3, components and parameterised models are proposed. The first model fulfils the desired level of fidelity which evaluates the PC of each part (component) of the C-RAN system. In the parameterised model, the complex representation of the network PC has been extensively reduced. Since only few parameters vary constantly in the model, and most of the parameters remain static, the parametrised model can be utilized to evaluate the futuristic power saving approaches in C-RAN. Such as the effects of the number of antennas on total PC, transmission power reduction, antennas' deactivation, bandwidth adaptations and PC reduction by activating sleep modes. Furthermore, the accuracy comparison of both models has been presented.

2. In Chapter 4, components and parameterised models are presented to evaluate the PC of SDC-RAN, this was based on the modelling of Chapter 3. Furthermore, the CN's control plane units, such as MME, SGW and PGW have been modelled. These components' consumption is added to the total network's PC, which in turn gives additional realistic visualisation to the PC of the entire network. Additionally, the accuracy comparison of both components and parameterised models of SDC-RAN has been presented.

3. In Chapter 5, modelling the way active VMs and processed resource blocks (RBs) by each VM increasingly affect the PC of the host server in terms of central processing unit (CPU), random access memory (RAM), network interface card (NIC) and hard drive (HDD). This modelling would provide a realistic, accurate and easy measurement to the PC compared to intrusive, software and utilisation level based models. Furthermore, the latency, which is concurrent with increasing the number of VMs and executed RBs at each VM in the virtualised server is also modelled.

4. In Chapter 6, modelling the cost of migrating the VMs from one server to another. This modelling has converted the available experimental and server utilisation or bit rate based, to a simpler expression. In addition, the migration time cost is also considered.

It is worth mentioning that this modelling is based on the concepts of existing measurements based works that dealt with evaluating the PC. Therefore, these concepts are converted to a mathematical models that reflect the actual behaviour of the network devices based on different parameters. Hence, such modelling exceeds the cost, complexity and proprietary of other works to produce simplified, costless, and adaptable models.

1.8 Thesis Organisation

In Chapter 2, a comprehensive literature review about energy consumption models of the various networks is presented. In 3, the PM of C-RAN is formulated. In Chapter 4, an SDN based C-RAN PM is derived based on C-RAN's PM, and compared with the latter. In Chapter 5, a virtualisation based networks/C-RAN PM is presented and compared with traditional C-RAN model. In Chapter 6, a PM to calculate the the cost of live migrating the VMs is introduced based simplified mathematical representation. In Chapter 7, we concluded the thesis and discussed the findings. Alongside, the forthcoming trends and research possibilities are discussed.

Chapter 2: Literature review

2.1 Introduction

Before discussing the literature, it is worth to define the energy E and power P, the former, in *Joules*, is the total work of a system/device performed by a unit time t, while the power, in *Watts*, is the rate at which this work is performed ($P = \frac{E}{t}$). Energy and power can be used interchangeably when the time has no effect on the system [42]. A large effort has been allocated to discussed the PC models in the literature, these models are variant regarding manipulating a single component or the entire network, or what is called system level modelling. These two types serves different purposes, the first type is when modelling the PC of a single unit device within the network can help understanding the way it works and the way it distribute/consumes the power. By knowing the core consumption area, it improves designing more efficient unit in terms of power. This type of modelling is usually more complicated than a system level modelling. In the latter, more generic assumptions can be made to cover the whole system consumption. For example, the work in [43] is proposed to investigate the radio frequency (RF) subsystem's PC. Moreover, [44] tackled the problem of RAM's PC. Although this thesis presents modelling of each network component, however, this type of modelling is not our main focus because of several reasons shown below. The server unit as a component is excluded from this assumption, as we will see in Chapters 5 and 6, as it is the workhorse equipment in the data centers, and represents the most important unit in terms of PC as a single server may costs more than 100 W even if it is in idle mode of operation. Moreover,

the demand of servers ranges from 50% to over than 90% of the electricity demand of IT devices [45].

Further to server PC model, modelling the PC of the entire network where the PC is larger represents another point of interest. Modelling a the entire network provides more understanding how networks plan the power. This also helps understanding which network devices that consume power more than others, which unit is more affected by the network resources and how much impact/weight that this unit put on the network. For example, reducing a unit with low PC does not impact as much as the high PC does in terms of the entire network EE. Therefore, modelling on system level clarifies such matter. In the other side of the network, the mobile device's PC is totally under control of the UEs. The impact of reducing its PC is not as important as reducing the PC in the data centre, where the power saving is larger. Although improving the life time of the battery is under investigation and measurement by many researchers, such as in [46]. However, such matter is out of our interest.

2.2 Traditional LTE Network/BS Oriented Power Models

On system level, a mathematical model is proposed in [47] to calculate the BS's PC. Specifically, the PA of the BS is switched off to put the BS on idle mode in case of no traffic has been sent, so as the energy consumption is reduced by using an algorithm. Subsequently, the gain of such method is being evaluated by the proposed power model. The BS's PC evaluation in the early phases design is established in [48]. The proposed model represents an end-to-end power measurement for heterogeneous LTE cellular systems, which includes Marco, Pico and Femto BSs. In [49], different SotA LTE BSs' PMs are investigated, this basic/simplified PM is presented relying on the data sheet of the BSs' components. In [50] and [51], complex LTE BS PMs are presented, providing estimation of the PC of different types of 3GPP LTE BSs. The presented models sweep through the functions complexity of each subcomponent to evaluate the total PC. For example, the RF component performs and affected by

several functions during operation such as bandwidth, type of modulation and number of antennas. To calculated its PC, the weight of each mentioned parameter up on the PC should be known. However, these model are time consuming and complex. These models are parameterised by the EARTH project PMs in [52], [41] and [53], which are used to simplify the BS's PM. To parameterise the PC, the complexity of the component's evaluation should be mitigated. This can be done by wrapping the inner functions' calculations into one parameter/value, while considering the accuracy and practical measurement considerations. The parameterised PM of [53] is compared to another parameterised model in [41], in terms of the number of radio chains/RFs and transmission bandwidth. These models are also used to linearise the PC of a the BS for further simplicity in evaluating the PC. However, these models are limited for only measuring the SotA BSs, such as Macro, Pico and Femto.

EARTH project has also targeted minimising the use of energy in mobile networks at least 2 times. Moreover, there has been several projects to observe and maximise the EE of the networks. The EE can be optimised by both offering higher data rates, and simultaneously, minimising the PC. For example, 'Green Radio' [54], aimed to reduce the energy requirement of delivering high data rates by 100x. 'Green Touch', has manipulated reducing the energy required per bit 1000 times by 2015 compared to 2010. 'OPERA-Net' [55], targeted improving the EE of the network by 20% in 2020 [56]. Contentiously, in [57], PC models of wireless BSs are proposed and compared regarding three mobile systems, these are HSPA, WiMAX and LTE. The SISO and MIMO systems are compared. In each case, the EE is evaluated and compared while using 10 Mbps data rate. In [58], an empirical and comprehensive PM of LTE is derived. The model compared the energy usage of LTE, 3G and WiFi networks using a data set of 20 smart phone UEs for a period of 5 months. These models are only valid for traditional BSs. In [59], a PM based on control plane, environmental units, data plane, the processing energy, storing and forwarding energy of each plane in the switches and routers within the network is proposed. In [60], a network PM is proposed based on link cost and switches cost. This model is based on energy

consumption of each bit, which requires additive measuring tools to evaluate how many bits are processed at a time period, so as the corresponding PC can be obtained, which adds additional complexity and power cost to the system. In [45], a model is proposed to estimate the electricity demands of the data center in a region within the United States. The model has shed light on the technical challenges associated with the various saving approaches of the network and IT devices.

2.2.1 Virtualisation Oriented Power Models

Generally, most of the power models describe the PC of a device/server as static plus dynamic. The first part is describing the PC when there is no traffic, this is fixed to a certain value no matter how much work load the device is countered with, this is also called idle or circuit consumption. The second part is more contradictory since it is confined to the dynamic load, which is affected by the traffic variation and amount of computations the device performs in a particular time. However, in virtualisation environment, there will be multiple dynamic loads which is linearly or exponentially affected by the installed VMs. Before discussing this, it is noteworthy to indicate that there are several proposals, algorithms and paradigms aimed adapting and implementing the NFV in the cloud based networks [61]. For example: NeFu-Cloud, PLayer, CloudNaaS, APLOMB, PACE, SIMPLE, CloudNFV and REALTIME CLOUD. These works investigate and define different mechanisms associated to the various modes of operations, and mostly they rely on open source tools to implement the suggested paradigms. Unfortunately, they lack for the mathematical representation, in-depth analysis and evaluation on the matter of power cost and allocations. Furthermore, there are different open source PC measurement platforms which use an online based algorithms, these platforms are totally hidden from the users. Rather, there have been enormous effort to mathematically and experimentally describe the way virtualised server consumes power.

In [62], the virtualisation effect upon the PC of a single server while running certain packages and applications is experimentally tested. This work however contemplates in association to a single server case study without providing a mathematical model based platform to measure the PC for components or system level, similarly in [63]. It was focused in [64] on the matter of VM's PC modelling, which was based on CPU , memory and storage utilization. The utilisation ratio can be easily measured from the operating system of the server it runs. However, such model provides mathematical based PM, but it was inherently based on monitoring measurements. Furthermore, the utilisation is taken as a rough value without any differentiation for the number of installed VMs, each with its allocated dynamic load, similarly to [65]. The authors in [66] have used the utilisation ratio based PM to optimise the trade-off problem between the performance and PC in the data centre. These models also require the utilisation ratio, specifically (CPU utilisation) to fulfil the calculations of the proposed PMs, otherwise, no prediction can be made. Such methods forces the researchers to posses a server so as the PC evaluation can be made. This conditional evaluation turn these models to a very sophisticated and expensive models. The authors in [67] have used the utilisation level and CPU temperature to calculate the server's generated heat. Based on this value, the PC of the server is calculated. In [68], the CPU utilisation, alongside the operating frequency have been used to model the PC of the server. In [69], a linear relationship is established between CPU utilisation and other components' PC in the server. This work describes the PC of a server as (CPU + others' PC), such assessment cannot be accurate as it ignores other server components; even though the CPU seize the major consumption within the server. In [70], a model is proposed based on speed of data and CPU utilization of data centre's servers. In addition to monitoring the CPU's resource allocated to the VMs, an algorithm has utilised the proposed model to minimise the operational cost in on-line services. In [71], a server PM is proposed based on CPU monitoring and utilisation ratio. In [72], [73], PMs are proposed to measure the PC of VMs based on evaluating the happening events, such as number and frequency/speed of memory

accesses and the number of active cores in a time period. This method is a complex one as it requires to measure the number of events and intrude the system. The intrusion device can be a reason for more PC within the network, which burden the network with unnecessary power cost.

However, intrusion based PMs means that the privacy of the server resources, such as CPU and RAM will be revealed by imposing an external (using hardware) or internal (using software) measuring tools to track the power usage of the server's units. This behaviour of tracking, monitoring and revealing is similar to the concept of hacking/intruding a communication channel/device without permission in the security systems. However, it is different when describing the PC as it means recording the PC with permission. This term is used to describe all the methods found in the literature that are based on using software, application or tacking devices, such as [74] and [75].

The models in [76], [77] and [78] are a single CPU level PC models. These performance counters based PMs are inaccuracy susceptible since most CPUs allow the measurement of a certain number of concurrent readings. This statement has been backed up in [42]. The process of tracking and converting to energy is a complex. It cannot guarantee accurate measurement for the happening events. This is because of the time response mismatch between some of the high frequency events and the software detection time window that is used to detect these events, similarly in [79], [80]. Although these models offer quick access to the power details. In fact, these models do not offer a simple or on the fly models for other researchers to rely on, where the PC measurement is not their main concern. For example, the investigators about improving the EE of the network are usually keen to improve the data rate by proposing advanced and optimised power allocation methods, rather than possessing a server and intrude it to measure its PC. On top of that, when the server is compact and sealed, it is difficult to intrude it as modern platforms do not allow measuring the PC of CPU, memory and disk of servers separately [42]. Subsequently, some PMs have proposed to track the resources utilised by each VM using software, then con-

verting the usage of these resources to energy units. This was called software based PMs. In which, it is required to install a monitoring software so as the usage of a particular unit or VM can be tracked or converted to energy figures. Such category is not different from the hardware based performance counter methods, but this one can be made by using special software, such as in [81] and [82]. The intruder software can be considered as one of the virtual machines, similarly to the hypervisor (HV) layer which contributes to the total consumption. Moreover, the tracking software is an expensive, error susceptible platform and requires to physically be available at the server site to record the power usage values, this weaken the importance of these methods.

Further to performance counters software based PMs, the second type of software based PMs is the machine learning or heuristic based PMs. [83], [84], [85], [86], [87] and [88] have used the expected I/O usage of server's CPU to calculate the PC using different categories of machine learning algorithms. The different set of workloads are collected by using software or hardware intruder, these information can be fed into and train the algorithm. However, learning algorithms consume more time to run, usually sub-optimal solution is given, the system is intrusive and decision-making algorithms add further PC as they require a separate device to run on. Even if the algorithm is installed on the same tracked server; that its PC needed to be calculated, it will require to compete the existing applications or VMs to use the hardware resources. This will add additional PC to the measured power. In addition, heuristics always provide sub-optimal solutions based on random distribution of the solution candidates. This requires repeating the process of optimisation several times, this process is a time consuming, and urges to repeat the learning process each time the network behaviour is changed. Clearly, a much simpler and costless PM to perform against software based models is vital. Such solution should offer an easy to evaluate, but accurate tool and on the fly model. It also be used by any researcher without complexity, and do not require to own a server or specific measuring device. It is also required that the proposed solution can translate the hardware utilisation ratio, machine learning

or tracking/counter based evaluation to a more understandable and effective network parameter, such as bandwidth or resource blocks utilisation, as this factor is much easier to manipulate.

2.2.2 Live Migration Oriented Power Models

The different types of models discussed above are generally applicable to whatever scenario that the server is operating on. For example, the utilisation ratio based model are valid in the scenario of bare, virtualised or live migration established servers. On the same basis, the performance or machine learning based models are also valid methods to quantify the PC of such situations. Predominately, the models found in the literature specify the scenario and application/type of service they run. For example, the model in [89] is dedicated to predict the cost of live migrating the VMs in terms of energy and time. By intruding the HV, the model reads the collected data of the memory usage and other work load based parameters in the server to simulate the process of migration, then the cost is obtained. The work handles only the cost of migration and no attention has been given to the inherent delay due to the virtualisation. In [90], the PC costs of the VMs' live migration for both the source and destination servers have been experimentally evaluated. This model was based on measuring the utilisation ratio of the CPU. Furthermore, the linear relationship between CPU utilisation and server PC is presented. In [91], the cost of VM's live migration has been experimentally measured, where several observations have been made. These are related to the amount of overhead expected during migration and concurrent latency. In [92], a model is proposed to evaluate the cost of VMs' migration quantitatively. This model was based on several experiments conducted to profile the cost of live migration process in terms of power and time.

2.2.3 Data Centre/C-RAN Oriented Power Models

In C-RAN system, various researches have been focused on the throughput/capacity enhancing based optimization problems, as well as resource sharing and allocation amongst the BBUs and RRHs. It was without a certain attention given to answer the questions: what/how C-RANs consume the power, and how will the power reduction techniques within this 5G field be examined?. As the deployment of any new communication technology/architecture/innovative design is excessively expensive, the pre-evaluation of the power budget shaping is crucial. When modelling a group of servers, the way of modelling will be different, as many parameters will be changed, such as cooling requirements, accounting for load balancing techniques and OFF/idle mode for some of the BBUs. On this basis, we present some of the contributions found in the literature that aim to model the system PC. In [93], a PM is proposed for the C-RAN's optical transport system of the fronthaul. The transport layer is designed based on wavelength division multiplexing using direct detection and coherent transmissions methods. This work compared the transport system of traditional LTE, partial and fully centralised C-RAN architectures using parametrised model. A group of servers' PC have been modelled in [94] and [95] based on average utilisation level of the CPUs of servers. The aim of this work is to reduce the electricity demands of a data centre by using renewable supply. This eventually has the effect of mitigating the environmental impact. The authors in [96], [97], [98] and [99] have modelled a complex group of servers' PC based on the number and size of the jobs arrival and queuing rate. However, it is required to measure these parameters by intruding the servers, which falls into the intruder based models, similarly in [100]. In [101], [102] and [103], a group of servers PC has been modelled based on measuring the PC of each process within the server. These models encountered major challenge, which is the difficulty of capturing the number and energy consumed at each process.

All the mentioned PMs in this Chapter are generally used to draw and shape the SotA BSs' PC when the PC of a single BS can be linearised, but they unfit to

address the upcoming hybrid network's PC such as SDN based LTE system, heterogeneous based C-RAN (H-CRAN) and the standalone C-RAN deployment. The above described models cannot serve a holistic network PC evaluation, which makes difficult to speculate the amount of power consumed in the proposed 5G networks. Furthermore, these models are restricted with the approach being used to enable the evaluation. This adds another difficulty when combining two models that use different approach for their PC evaluation. For example, to evaluate the total PC of the network, a BS PM that is based on machine learning should be combined with another PM that uses additive or utilisation based method to evaluate the cooling or RRH's PC. This makes evaluating the total PC of the network intractable.

Furthermore, the proposed models are able to overcome the validation process even if they are not experimentally proven. This is because of two reasons:

1. The assumptions of the proposed models are based on real time experiments, for example [53] and [41].

2. The parameters and PC values are extracted from real time measurements, as mentioned in Chapters 3, 4, 5 and 6, .

Therefore, these proposed models can defeat the need to prove its concept and output results by an experiment or extensive simulations, which results in cost effective models.

2.3 Summary

This Chapter has presented the available PC models that are used for different type of network architectures, units and applications. However, these PMs lack providing one or more of the below aspect, which are overcome in the proposed models, these aspects are:

1. Holistic PC evaluation, which means most of the available PMs are dealt with single unit PC and ignores the other part of the network.

2. Simplicity and flexibility, this means that the method used to evaluate the PC is a complex method that is based on enlarged number of variables and computations. Such issue can be overcome by modelling the network with least number of variables that are changing within the network.

3. Cheap or economical, this means that no extra cost is required to either buy the measuring device or consume more power by the measuring device.

4. Rapid PC evaluation, the model should fulfil the required calculations of the PC with least possible time.

5. Easy to fit, this means the PM is not proprietary based to serve a limited number/type of applications, devices or traffic.

Subsequently, C-RAN PM is presented in the following Chapter that fulfils the above requirements.

Chapter 3:   Power Model for Cloud Radio Access Network

3.1   Introduction

In contrast to the eNodeB, where the entire communication layers processing has been implemented within the cell site [22]. The main baseband physical procedures and processing of upper layers in C-RAN are executed in the BBU pool. Whilst the simple RF front and symbol processing functions are tackled by the RRH.

The RRHs are densely deployed in the network with minimum cost, which distinguishes C-RAN from traditional systems, and presents major functionalities of C-RAN. Potential benefits of C-RAN architecture include: (i) Using advanced signal processing and co-ordination techniques to process signals through any neighbouring BBU(s) in the cloud. (ii) Allowing cognitive radio to enhance the efficiency of the network spectrum-utilization [104]. (iii) Reduction in OPEX of the network operators, due to fewer site visits, easy upgrading and maintenance and lower site lease. (iv) As 80% of the PC results from the BS site, the ability to exploit processing load variations across BSs becomes a must, this can be done by pooling the BBUs in a cloud. This allows the operators to reduce the PC by turning off the unwanted processors [105]. (v) Separation of the control and data plane to enable controlled and intelligent self-organised networks [106].

Fig. 3.1 shows the breakdown structure of the simplified complete traditional BS [41], which can be generalised to all BS types i.e. Macro, Micro, Pico and Femto BSs.
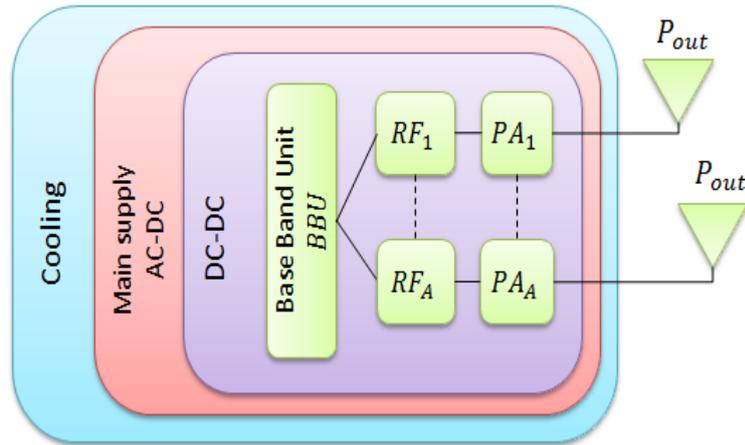
Fig. 3.1. Base station transceiver block diagram.

On the other hand, the proposed PM in this Chapter identifies C-RAN consumption while considering the fundamental BS site's blocks. The main objectives of the model are:

- Mapping RF radiated output power at the antenna to the total BS site PC, the component and system level interfaces.

- Load fluctuations description over the day, considering long and short term traffic model.

- Large-scale model's deployment that covers small-scale scenarios and can be extended to larger geographical areas.

It is worth noting this SotA PMs are unable to address the holistic PC of the futuristic network paradigms, for example C-RAN. To the best of our knowledge, an advanced PC model for a components or system level of the C-RAN has not been addressed in previous work. Particularly, the BBU pool/RRH combination. In this chapter, the system model of C-RAN, the proposed component PM and its correspondent PC components are introduced. The total PC of the network is described, concurrent with introducing the parameterised PM. Parameterised model means that the system has been represented with the least complex way possible by assuming the fewer number of sub-systems which yield a reasonable framework. Furthermore in

this chapter, the simulation results to verify the effectiveness of the proposed C-RAN architecture are provided. Finally, a summary has been given.

3.2 Components PM

The two main parts involved within this PM, these are:

- BBU pool, which contains several BBUs, each capable to serve one or many RRHs.

- RRH, with MIMO system supporting.

Both BBU pool and RRH are divided into sub components, which contribute to the PC model. Each component has its own power budget calculation. Fig. 3.2 shows the block diagram and the PC components of the C-RAN.
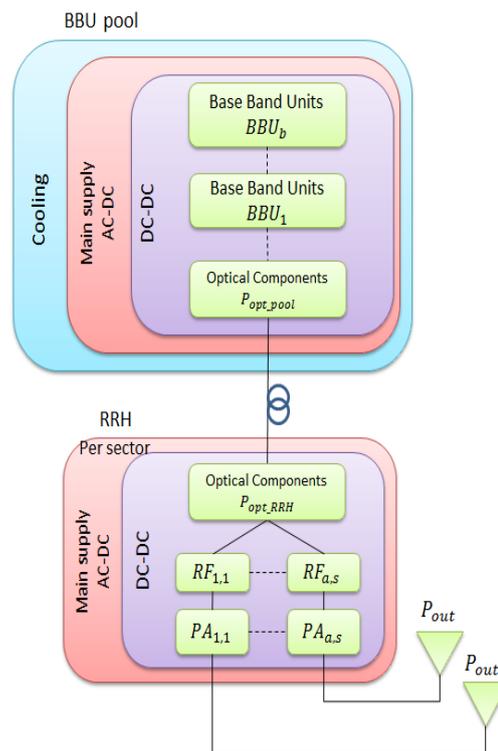
Fig. 3.2. Block diagram of C-RAN transceiver power model.

3.2.1   Base Band Unit Pool (BBU pool)

The BBU pool is a consolidation of multiple BBUs in a form of a cloud or cluster. All the RRHs connected to the BBU pool have a unified baseband processing and resources pool. The PC model for the BBU pool is considered to be the sum of the active BBUs in the BBU pool and other important components which are briefly discussed as follows:

3.2.1.1   Base Band Unit (BBU)

The BBU is responsible for digital signal processing and signals generation before passing it to the RF transceiver. The digital computation and processing of the BBU can be measured in giga operation per second (GOPS) and translated into power figures. This can be achieved by multiplication of the GOPS with the technology scaling factor revealing the operations performed per second per Watt (W) [51]. A set of different BBU functions (I_{BB}) such as the time and frequency domain processing, forward error correction (FEC), central processing units (CPU) and processing related to common public radio interface (CPRI) can be associated with GOPS. 40 GOPS/W is estimated to be the power cost of a large BBU [51]. The PM of the BBU can be expressed as

$$P_{BBU} = \sum_{i \in I_{BBU}} P_{i,BBU}^{ref} \times (\frac{x_{act}}{x_{ref}})_{s_i,x} \tag{3.1}$$

Where P_{BBU} denotes the PC of the BBU, $P_{i,BB}^{ref}$, is the BBU's sub component PC, i.e. BBU functions. x_{act} represent the actual value of the power consumed for parameter x, x_{ref} is the reference value of the parameter consumption, s_i denotes the scaling vector, which is either 0 or 1. This factor indicates whether the parameters are processed within the function or not. For example, to calculate the power consumption of frequency domain P_{fd} function within the BBU, there will be different

parameters involved, i.e. x=(Bandwidth, Antenna, Modulation, Code rate, Time domain, Frequency occupation), while the scaling vector is $[s_1, ..., s_6]$=[1,0,0,1,1,1].

If the frequency domain function is encountering 10MHz,2×2 MIMO, 16 QAM modulation, 3/4 code rate, 100% duty cycle, and 30 % frequency occupation, hence, the PC will be calculated as follows $P_{fd} = 1.5 \times (\frac{10}{20})^1 \times (\frac{4}{6})^0 \times (\frac{3/4}{1})^0 \times (\frac{2}{1})^1 \times (\frac{100}{100})^1 \times (\frac{30}{100})^1$. This process will be repeated for different function (i) of the BBU, which results the BBU total PC P_{BBU}.

3.2.1.2 DC-DC Conversion (DC,P)

The various components of the model require proper DC voltages for operation; therefore DC-DC converters are placed wherever needed. DC-DC converters have efficiency less than 100 %, thus it can be expressed as losses. It was mentioned in [41] and [52] that the losses incurred by the overhead (i.e. power conversions and active cooling) scale linearly with the PC of other components which require such DC conversion, i.e. (BBUs, Optical devices, etc.). Therefore; the PC caused by DC conversion ($P_{DC,P}$) inside the BBU pool is given as:

$$P_{DC,P} = \sum_{b=1}^{B} \sum_{r=1}^{R} l_{DC,P}(\eta_{DC,P}) \times (P_{BBU_b^r} + P_{opt,P_b^r}) \qquad (3.2)$$

Where B denotes the total number of active BBUs in the BBU pool, R denotes the total number of active RRHs. $P_{BBU_b^r}$ is the PC of b-th BBU, which is connected to r-th RRH, P_{opt,P_b^r} is the PC by the optical device in the BBU pool which connects the b-th BBU to r-th RRH, $l_{DC,P}$ is the loss caused by DC-DC conversion; as a function of DC conversion efficiency ($\eta_{DC,P}$), the loss function can be expressed as a decay function with an exponential decay constant (ϖ), this constant might differ amongst the devices when subjected to the manufacturer design quality, which eventually affects the value of the loss function. However, the losses can be expressed by

$$l(\eta) = l_o e^{-\eta \varpi} \qquad (3.3)$$

where l_o is the initial value of the losses at time $(t = 0)$, and η is the efficiency of the device. As known that each converter hold an efficiency of conversion, which can be known from the device's data sheet. This efficiency can be expressed as losses in this modelling to show that such suboptimal efficiency causes more PC to the network. Therefore, (3.3) converts the efficiency of the device to losses, as as to be used in (3.2). For example, if the device hold an efficiency equal to 90%, (3.3) will result about 0.15 losses $(l(\eta))$ if the initial losses of the device (l_o) is equal to 0.009. The more efficient the device, less incurred losses. Furthermore, there are two main reasons why the efficiency (η) is not used directly with (3.2) instead of $l(\eta)$:

- Each device holds an initial losses.

- Each device hold different efficiency, this can be tuned by the constant $(\varpi = 0.023)$.

After, the obtained losses can be fed into (3.2) to obtain the DC-DC PC. If the BBU PC (P_{BBU}) is 29.4W and the optical device PC $(P_{opt,P})$ is equal to 1W and $R = B = 1$. Hence, $P_{DC,P}$ can be calculated as $0.15 \times (29.4 + 1) = 4.56W$

3.2.1.3 Mains Supply, AC-DC-Conversion (MS,P)

The power form of the main supply grid has to be converted from AC to DC. This is done by the mains supply unit. The architecture of AC-DC converters varies across vendors. The PC of this unit $(P_{MS,P})$ is generally modelled the same way as the DC-DC power conversion, and is given as:

$$P_{MS,P} = l_{MS,P}(\eta_{MS,P}) \times \left(P_{DC,P} + \sum_{b=1}^{B} \sum_{r=1}^{R} P_{BBU_b}^r + P_{opt,P_{b,r}}\right) \tag{3.4}$$

$l_{MS,P}$ denotes the measured losses of the MS power conversion as a function of AC-DC conversion efficiency $(\eta_{MS,P})$.

3.2.1.4 Cooling

Cooling is responsible for the greatest energy waste [107]. BSs require active cooling and a large portion of the energy is wasted into BSs cooling. Cooling requirement may change depending on the geographical location, the position of BSs and the size of the cooling cabinet they are housed in. Cooling PC can be modeled as a fixed power loss, as it is a very slow operation, compared to the fast timing variations of the BBUs. The cooling unit is responsible for cooling each of the BBU pool's components, i.e. the BBUs, DC-DC, AC-DC conversion and the optical transceivers. Due to hosting several BBUs in the BBU pool; it requires more cooling. However, cooling PC in the BBU pool with B number of BBUs, is much lower than the amount of PC by the same number of Macro BSs. This is because the shifting of PA and RF from the BBU pool to the RRHs, as we know the RRH requires no cooling. This is in contrast to the eNodeB where such units contribute to the cost of cooling PC. In this modelling, cooling PC is proportional to the PC of all other components. This means in the BBU pool, the cooling is required only for BBUs, while in the eNodeBs, it was required for BBUs, Rf and PAs. This enforces a reduction in the cooling PC. If *lcool* is the cooling loss, then the cooling PC by the BBU pool (*Pcool*) can be calculated as:

$$P_{cool} = l_{cool} \times (P_{MS,P} + P_{DC,P} + \sum_{b=1}^{B} \sum_{r=1}^{R} P_{BBU_b^r} + P_{opt,P_{b,r}})$$ (3.5)

3.2.2 Remote Radio Head (RRH)

The RRH consists of components with much lower operational complexity. It is equipped with a radio receiver chains (RFs) and power amplifiers (PAs) that scales linearly with the number of antenna used. The RRH also consists of necessary voltage suppliers. The RRH PM can be broken down to the following sub sections:

3.2.2.1   Antenna (A)

The antenna itself does not influence the PM, as their purpose is only to transmit/receive the signals to/from end users. The number of antennas used in the system affects the PM, as each antenna requires RF and PA to provide the necessary signal operations and amplification.

3.2.2.2   RF transceiver (RF)

RF transceiver unit consists of an intermediate frequency and baseband interface. RF unit is responsible of, but not limited to the following functions:

1. Modulation/Demodulation the signals.

2. Voltage controlled oscillators (VCO) and Mixers.

3. Digital to analogue (DAC) and analogue to digital (ADC) convertors.

4. Low Noise Amplifiers (LNAs), gain amplifiers, Clocks, etc..

3.2.2.3   Power Amplifier (PA)

The PA is a prime element of consideration in the PM, as it consumes most of the power within the RRH. The PA amplifies the electrical signals received from the O/E converter before passing it to the RF circuit, it also amplifies the signals coming from the RF before transmitting to the air interface by the antenna or passing it to the E/O to send it via the optical fibre. Generally, PAs have low efficiencies at low antenna transmission power (P_{out}), however in rare cases its efficiency can reach up to 54% if high transmission powers are intended at the antenna(s) [50]. Due to the strong fluctuation in transmission powers of the orthogonal frequency division multiple access (OFDM) signals, PA usually operates with low efficiencies.

This low efficiency is originated from high peak to average power ratio (PAPR) of OFDM signals, PAPR is defined as the ratio of the maximum power transmitted

to its average. The large peaks impose a degradation to the performance of the non linear PA. This non linear behaviour produces out of band radiation and in band distortion, which causes adjacent interference in the channel and increased bit error rate, respectively. Practically, such high power fluctuation in OFDM sub carriers requires high response time of PA to cope with this variation in the power, such facility is difficult to be designed, which results in low efficiency. Modelling the PC of a PA requires the following important parameters to be considered:

1. Output transmitted power (P_{out}) of the antenna.

2. Output power of the PA (P_{TX}).

3. The share of maximum bandwidth (BW), that an antenna uses, i.e. the actual number of the physical resource blocks (RB) that occupies a certain bandwidth for transmission.

The PA's PC (P_{PA}) is affected by its efficiency (η_{PA}), which is a function of P_{TX}. The PA's PC, in W can be modelled as:

$$P_{PA} = \frac{P_{TX}}{\eta_{PA}(P_{TX})} \tag{3.6}$$

3.2.2.4 RRH's Power Conversion

An AC-DC and DC-DC voltage converters are required to provide the necessary voltages supplies to the RRH's components such as the PA, RF and the optical components. The PC of the RRH's DC conversion ($P_{DC,R}$) is modeled by considering its losses ($l_{DC,R}$) as a function of the efficiency ($\eta_{DC,R}$) along with the power requirement of all other components;

$$P_{DC,R} = l_{DC,R}(\eta_{DC,R}) \times (P_{opt,R} + \sum_{b=1}^{B}\sum_{r=1}^{R}\sum_{a=1}^{A}(P_{PA} + P_{RF_a}^{r,b})) \tag{3.7}$$

Where, $P_{RF_a}^{r,b}$, $P_{PA_a}^{r,b}$, in W denote the PC of the a-th RF and PA respectively, within r-th RRH which is connected to b-th BBU, $P_{opt,R}$ is the PC of the optical device inside the RRH.

The MS or AC-DC converter for RRH can be modelled in the same fashion as the BBU pool's MS, therefore its PM is configured

$$P_{MS,R} = l_{MS,R}(\eta_{MS,R}) \times (P_{DC,R} + P_{opt,R} + \sum_{b=1}^{B}\sum_{r=1}^{R}\sum_{a=1}^{A}(P_{RF} + P_{PA})_a^{r,b}) \qquad (3.8)$$

Where $l_{MS,R}$, denotes the MS losses as a function of the conversion efficiency $(\eta_{MS,R})$.

3.2.2.5 RRH Cooling

It was states by Nokia that the BSs with total PC of 500 W or less (except the output power of the BS (P_{out}), do not require cooling system [47]. This is applicable for the RRH, which is composed of components i.e. (PA, RF and optical components) having overall power less than 500 W. Therefore the only overhead for the RRH is the supply power, while its cooling PC is negligible [41].

3.2.3 Optical Transceiver PC

The access line that connects the multiple RRHs to the BBU pool is often known as the "front-haul". Optical fiber is the most promising infrastructure in the front-haul due to its large capacity support and scalability. The optical transceiver is responsible for the conversion of the signals from electrical to optical light with a certain wavelength and vice versa. Several factors influence the optical transceiver operation such as the technology used, the required output power and the operating condition [108]. These in turn affect the PM. Optical transceivers can be divided into two modules from the PC perspective i.e. optical transmitter module (OTM) in which, OFDM electrical signals are modulated over optical carrier using an external

or direct modulated laser. While the optical OFDM signals are detected by the optical receiver module (ORM) either by direct detection (DD) or coherent detection (CO-D).

Typical PCs of commercial point-to-point (PtP) and point-to-multipoint (PtMP) optical transceivers are 1 W and 1.5 W, respectively. The model uses PtP transceivers, as this type does not have passive optical power splitter and offers relaxed link budget, i.e. PtP link loss is governed by only the distance and the used operating wavelength. In contrast to the PtMP in which; constrained link budget about (20-35) dB is required; due to wavelength sharing nature in the same fiber. Whilst the link loss of a PtP is as low as 6 dB with 20 km network reach [109].

3.3 Total PC

The total PC (P_{CRAN}) of the C-RAN network; is the sum of the BBU pool and RRHs PC, it is formulated as follows:

$$P_{CRAN} = P_{Pool} + P_{RRH} \tag{3.9}$$

P_{Pool} is calculated by aggregating the PC of all active components, and calculated as:

$$P_{pool} = P_{cool} + P_{MS,P} + P_{DC,P} + \sum_{b=1}^{B} \sum_{r=1}^{R} \sum_{a=1}^{A} P_{BBU\,b,r,a} + P_{opt,P_{b,r}} \tag{3.10}$$

$P_{BBU\,b,r,a}$ is the PC by b-th BBU; attached to r-th RRH; which mounts a-th antenna. RRHs can be served by any active BBU within the BBU pool. This means that BBU-RRH mapping can be dynamic depending on the traffic conditions [110]. This service diversity is considered in the model. The total PC by RRHs is formulated

$$P_{RRH} = P_{MS,R} + P_{DC,R} + \sum_{b=1}^{B} \sum_{r=1}^{R} \sum_{a=1}^{A} (P_{PA} + P_{RF})_{r,a}^{b} + P_{opt,R_{r}} \tag{3.11}$$

$P_{PA}{}_{r,a}^{b}$ and $P_{RF}{}_{r,a}^{b}$ denote the PC of PA and RF respectively; by a-th antenna's within r-th RRH; which is attached to b-th BBU.

3.4   Parametrised PM

The parameterised model encompasses the architectural details and highlights parameters which are either assumed to be constant, or having negligible effects such as lookup tables, GOPS, manufacturing details. To this extent, the following approximations are made further to the SotA BSs' PMs:

1. The BBU and RFs' PC, both scale linearly with the number of antennas (A) and the bandwidth (BW), i.e., $P_{BBU} = A \times BW \times (P_{BBU}^{pm})$ and $P_{RF} = A \times BW \times (P_{RF}^{pm})$, where P_{BBU}^{pm}, P_{RF}^{pm} denote the PC by the BBU and RF in the parameterised model (pm), respectively.

2. The PC of PA depends on the maximum power transmitted per antenna (P_{max}/A) and its efficiency (η_{PA}). The feeder loss between PA and the antenna can be ignored for the RRH [41]; i.e., $\sigma_{feeder} = 0$ since the PA is placed close to the antenna. The PA's PC can be represented as $P_{PA} = P_{max}/A\eta_{PA}$. The PA efficiency is varying at different transmission powers, and assumed to decrease by a factor of γ for each halving of the transmitted power. Thus it is maximum (per single antenna), when the PA's maximum transmission power during operation (P_{max}) is equal to the maximum PA's data sheet transmission power ($P_{PA,limit}$), heuristically the efficiency is described by

$$\eta_{PA} = \eta_{PA,max}[1 - \gamma \log(P_{PA,limit}/\frac{P_{max}}{A})]$$

3. DC-DC, AC-DC conversions as well as cooling PC, scale linearly with other components PC and are approximated by the loss factors $\sigma_{DC,P}$, $\sigma_{MS,P}$, σ_{cool}, $\sigma_{DC,R}$ and $\sigma_{MS,R}$ for BBU pool DC, BBU pool MS, cooling, RRH's DC, and RRH's MS loss factors respectively.

4. The optical transceivers PC $P_{opt,P}$ and $P_{opt,R}$ scale linearly with the number of BBUs and RRHs in the network.

5. The losses incurred by the optical fiber between BBU pool and the RRH is approximated by the loss factor ($\sigma_{optical}$), it can be adapted to meet the fiber length and number of connectors and splices used.

The maximum PC (P_{CRAN}^{pm}) can be formulated by aggregation of the PC of single BBU in the BBU pool (P_{pool}^{pm}) serving a single RRH (P_{RRH}^{pm}) and is given as:

$$P_{CRAN}^{pm} = P_{pool}^{pm} + P_{RRH}^{pm} = \frac{A \times BW \times (P_{BBU}^{pm}) + P_{opt,P}}{(1 - \sigma_{DC,P})(1 - \sigma_{MS,P})(1 - \sigma_{cool})}$$
$$+ \frac{A \times BW \times (P_{RF}^{pm}) + \frac{P_{max}}{A\eta_{PA}} + P_{opt,R}}{(1 - \sigma_{DC,R})(1 - \sigma_{MS,R})(1 - \sigma_{optical})} \tag{3.12}$$

Then the total number of R RRHs and B BBUs is considered to obtain the total PC of the network (P_{supply}^{CRAN}).

$$P_{supply}^{CRAN} = B.P_{pool}^{pm} + R.P_{RRH}^{pm} \tag{3.13}$$

It is worth mentioning that the proposed model above is a modification of the traditional work presented in both [53] and [41], which is part of the EARTH project that aimed to evaluate the PC of the SotA BSs. According to their measurement, they have concluded and summarised the evaluation of the PC is affected by few main parameter, such as bandwidth, transmission power and number of antenna. This understanding for the network's units operation has resulted in a PM to describe their conducted real-time evaluation. The resulting PM can be expressed in (3.14) as follows [111]:

$$P_{BS} = \frac{A \times BW \times (P_{BBU} + P_{RF}) + P_{PA}}{(1 - \sigma_{DC,P})(1 - \sigma_{MS,P})(1 - \sigma_{cool})} \tag{3.14}$$

Where P_{BS} denotes the traditional BS PC. This architecture represnt the modelling of SotA BSs, i.e. Macro, Pico, Femto, etc, which is described in Fig. 3.1. There are many characteristics can be extracted from this modelling, these are:

- The PC of BBU and RF units are linearly proportional with the number of antennas and bandwidth used.

- The PA PC is based on the transmission power of the antenna as described earlier.

- The losses incurred within the network are linearly scaled with the PC of other components within the network, and are approximated by losses factors.

The splitting of the eNodeB's units to RRH and BBU pool has enriched the network with many advantages that are mentioned in the C-RAN concept. Although realising C-RAN architecture in hardware can be adapted from the traditional networks regarding BBU server, optical transceivers or RRHs, there are many differences between the architectures of traditional BS, i.e. Fig 3.1 and C-RAN, i.e. Fig. 3.2. These differences can be summarised as follows:

- The total PC is divided into RRH and BBU pool.

- The PA and RF units do not contribute to the total PC of the BBU pool.

- The PA and RF unit do not contribute to the cooling PC of the RRH as it requires no cooling.

- No feeder cable looses in the RRH as the antenna is placed closer to the transceivers.

- An optical transceivers are found in C-RAN architecture in both BBU pool and RRHs.

- Additional DC and AC converters are found in the RRHs.

- Optical fiber links and losses are found in C-RAN networks.

- A single PA is sufficient to each RRH, while a PA is used in each sector of the traditional BSs.

3.5 Results and Discussion

The proposed parameterised model is used to approximate the PC, and compared with the parameterised PM in [53], [111], which breaks down and measures the PC of the traditional BSs, in terms of bandwidth used, varying antenna numbers, varying Macro BS sectors, and varying RRHs. Parameters are chosen according to [41] and [53] when possible. The parameters have been adjusted in some cases aiming fair comparison, especially the RRH's transmitted power (Average P_{max}), i.e. it is assumed $20W$ in [41], while it is assumed $40W$ in this work to match the parameters settings of the compared work [53]. The resulting parameters are provided in Table 3.1, which summarises the measured SotA PC of LTE Macro BS and RRH.

The PM is simple to adapt, and to approximate different vendors' configurations, by changing the individual parameters in order to observe the variation in the resulting PC. The model is verified for one, two, and four antennas. The number of RRHs and Macro BS's sectors are up to 60. This number can be increased according to vendor' architectural demands. Each Macro BS is considered to mount 3 sectors. The increased band width above 10 MHz is expected to increase the RF and BBU's PC, while other parameters such as losses and transmission power are expected to remain unaffected by the system bandwidth. As the number of Macro BS's sectors increases, the cooling PC is reduced as shown in Fig. 3.3, compared to the increased number of BBUs in the BBU pool. Fig. 3.3 shows the cooling PC of C-RAN power model compared to traditional Macro BS for one, two and four antenna configurations. It shows a dramatic decrease in the PC. The result shows that for any antenna configuration, cooling PC has been reduced to about 87.4 %. Cooling PC reduction can be calculated

Table 3.1
Parameters breakdown

Component	Unit	Macro BS	RRH	Reference
P_{max}	W	40	40	[41]
P_{PA}	W	80	40	[41]
$P_{PA,limit}$	W	80	80	[41]
P_{TX}	W	6.8	6.8	[41]
P_{RX}	W	6.1	6.1	[41]
P_{RF}^{pm}	W	12.9	12.9	[41]
P_{BBU}^{pm}	W	29.4	29.4	[41]
η_{PA}	-	0.36	0.31	[41]
σ_{DC}	-	0.075	0.075	[41]
σ_{MS}	-	0.09	0.09	[41]
σ_{cool}	-	0.1	0	[41]
σ_{feeder}	dB	-3	0	[41]
P_{MME}^{pm}	W	65	40	[89]
P_{SGW}^{pm}	W	65	40	[89]
P_{PGW}^{pm}	W	68	40	[89]
P_{switch}^{pm}	W		58	[112]
$P_{SDNctrl}^{pm}$	W		20	[112]
$P_{opt,P}$	W	-	1	[108]
$P_{opt,R}$	W	-	1	[108]
$\sigma_{backhaul}$	-	-	0.085	[-]
$\sigma_{fronthaul}$	-	-	0.085	[-]
$sectors$	-	3	1	[-]

by using percentage change rule, i.e., $\frac{(v_1 - v_2)}{|v_1|} \times 100\% = \frac{(4000 - 750)}{|4000|} \times 100\% \cong 81\%$. However, more reduction is expected when increasing the number of BBUs.

It is worth mentioning that the cooling PC of the traditional networks [53] is described in (3.15) below, if M represent the total number of BS sectors, the cooling PC can be calculated as follows:

$$P_{cool} = l_{cool} \times (P_{MS,P} + P_{DC,P} + \sum_{m=1}^{M} P_{BBU} + P_{RF} + P_{PA}) \qquad (3.15)$$

The difference between (3.15) and (3.3) regarding cooling PC is described in Subsection (3.2.1.4). However, this is due to the fact that the PA and RF components are no longer contributing to the cooling PC in the BBU pool of (3.3). This is in contrast to the Macro BS, where the PA and RF units increases the cooling PC because they scale linearly with the required cooling consumption, as shown in (3.15).
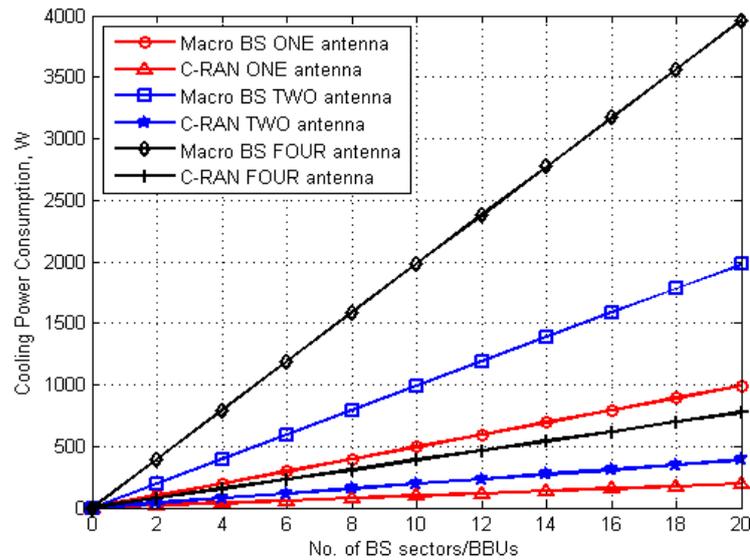

Fig. 3.3. Comparison of C-RAN cooling PC and PM of Macro BSs [53] with one, two and four antennas configurations.

Table 3.2 shows the differences in the PC between Macro and C-RAN models for 4 antenna configuration.

While Fig. 3.4, shows the effect of reduced cooling PC on the total network PC, with varying numbers of sectors/RRHs. By using the percentage rule for the values

Table 3.2

Comparison cooling PC between C-RAN and Macro BSs, 4 antennas configuration.

No. BBUs/sectors	Macro BS(W)	C-RAN(W)
2	2000	75
6	2000	260
10	4000	400
14	6000	550
18	8000	740
20	10000	760

circled in Fig. 3.4, it was found that the percentage change reduction in the total PC is about 33.3%. For the sake of comparison, it is assumed that the number of sectors in Macro BS is equal to the number of RRHs, i.e. each BBU is assumed to serve 1 RRH/sectors, and each BS considered supporting 3 Sectors/RRHs. Despite the addition of optical components PC to the proposed PM, considerable lower total PC is achieved rather than in [53], as shown in Fig. 3.4.

The proposed model shows extra PC reduction compared to Fig. 3.4, considering the case of switching off some of the BBUs during off-peak hours. It is assumed that 60 RRHs attached to only 20 BBUs. The PC is shown in (3.5) and (3.6).

Fig. 3.3, Fig. 3.4, Fig. 3.5 and Fig. 3.6 show a reduction in PC based on the parameterised model and compared to [53]. Fig. 3.7 shows the total PC as a function of the system's bandwidth share and varying numbers of antennas. It is observed that with four transmitting antennas, the PC by C-RAN approaches the Macro BSs PC, only when it uses nearly the entire bandwidth share and much lower PC otherwise.

The accuracy of the simplified parameterized model and the components model can be evaluated using the decay function 3.3. The main key comparison can be done by knowing the initial losses l_o of each sub components and the exponential decay constant (ϖ), which have been subjected to the manufacturer design. However, to
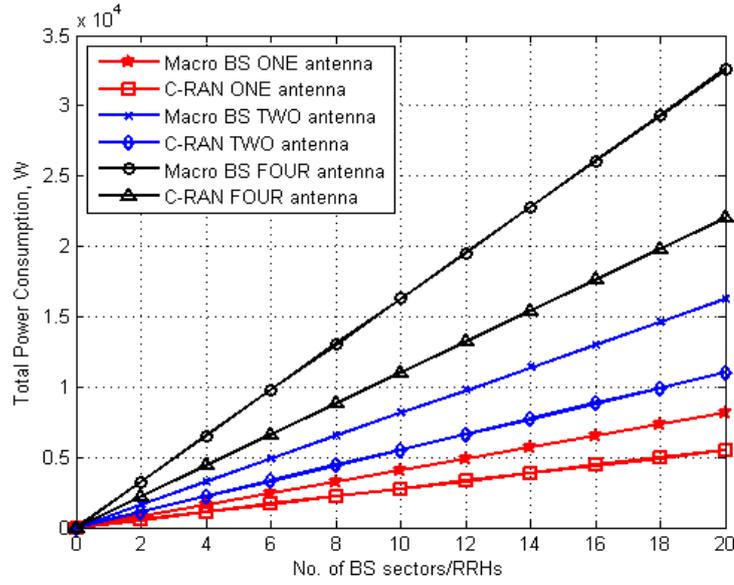
Fig. 3.4. Comparison of C-RAN total PC model with PM [53] of Macro BSs. 20 Macro BSs (each have 3 sectors) serving 60 Sectors/RRHs and 60 BBUs serving 60 RRHs/sectors with one, two, and four antennas configurations.

compare both the components and the parameterised models, some assumption has been made to the component PM:

1. It is assumed that the efficiency η is 90% for all the sub components (i.e. DC, AC convertors), cooling loss (l_{cool}) is 0.1 and the initial loss l_o is 0.009.

2. Using (3.3) to evaluate the losses values which are required in (3.2),(3.4),(3.7), and (3.8).

3. Calculate the total components PM using (3.10).

4. Compare with the parameterized model of (3.12).

Based on the subcomponent data sheet, and for different values of a given exponential decay constant (ϖ), the results show identical accuracy and flexibility performance between the parameterised and component PMs. Fig. 3.8 shows the total PC

Fig. 3.5. 3 dimensional view of C-RAN total PC with varying number of BBUs i-e up to 20, and up to 60 RRHS, having one, two and four antennas configurations.

of C-RAN with different ϖ values and different number of RRHs, using one antenna configuration.

3.6 Summary

Densifying the network might provide a solution for higher data rates demands. However this will lead to tremendous network infrastructure and increased PC. Consequently, C-RAN has emerged as a solution. Together, components and a parameterised PM are provided in this paper, which allows the PC calculation of C-RAN according to operational parameters and varying vendor configurations. The model is applicable and simple, to meet the new generation communication systems requirements. A comparison of the model performance has been made with previous work, regarding cooling and total PC, with varying parameters such as antenna and bandwidth sweep. The model has shown that the C-RAN network reduced the total
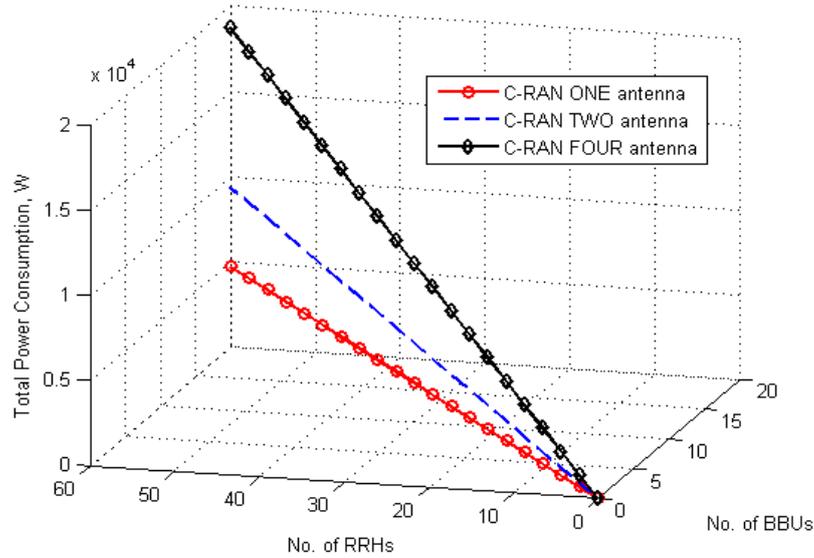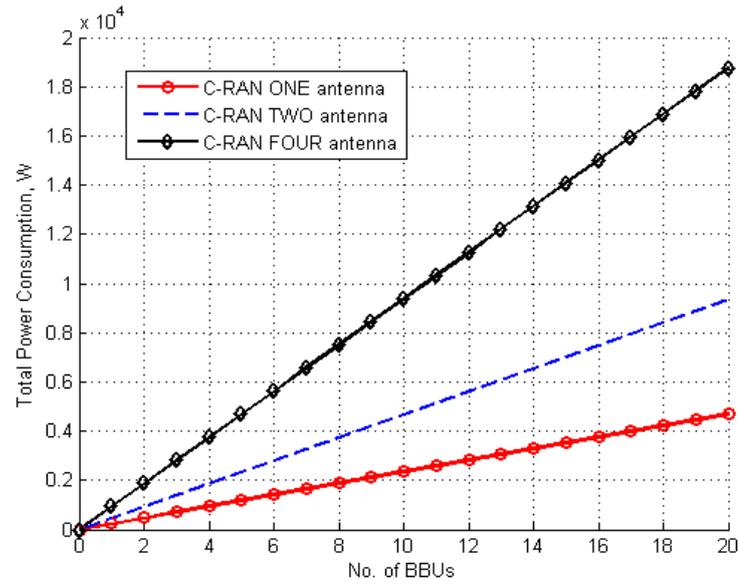
Fig. 3.6. Two dimensional view of C-RAN total PC with varying number of BBUs i-e up to 20, and up to 60 RRHS, having one, two and four antennas configurations.
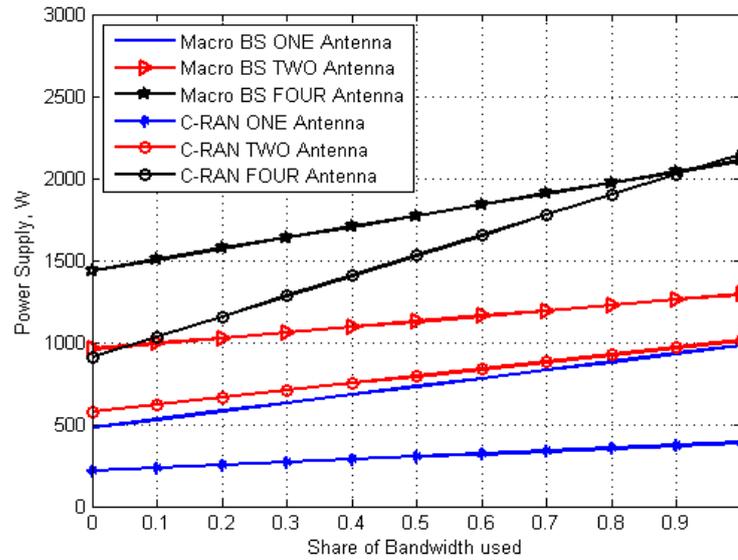
Fig. 3.7. Comparison of C-RAN total PC with [53], for 10 MHz band width.
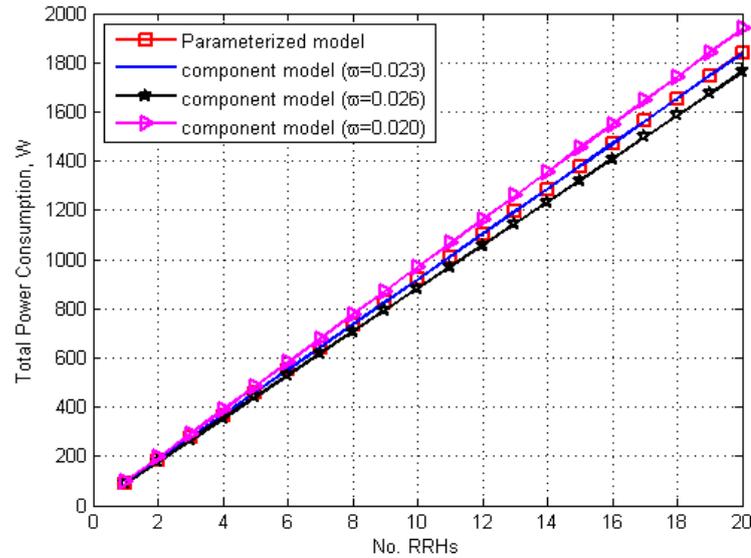
Fig. 3.8. Accuracy comparison of the total PC of the components and parameterized model with different number of (ϖ) values and different number of RRHs, using only one antenna configuration.

PC to about 33.3% as it reduced the cooling PC to about 87.4% when compared to traditional MBS architecture. The results also show that the parameterised PM is as accurate as the component PM. As for the future work, gathering the two parts i.e. C-RAN and Macro BS, has yielded H-CRAN as a promising architecture suggested, to enhance the EE of the upcoming generation. Using the proposed model to evaluate H-CRAN PC might give a wider vision, about the power figures of such extended architecture, when the Macro BS and C-RAN collaborate in the area of interest. Furthermore, this proposed modelling has been used to evaluate more integrating techniques within the race to build 5G, one of these rising techniques is SDN. The following Chapter will discuss the PC, advantages and disadvantages of SDN network.

Chapter 4: Energy Efficiency of Software Defined C-RAN

4.1 Introduction

Because of the large number of deployed RRHs, this increases the signalling cost in the CN in relation to the UEs' handover set ups and authentications. Therefore, techniques to improve the network scalability need to be extensively investigated. Recently, the research community has embraced SDN-LTE integration, which offers centralised administration for the underlying devices in the BBU pool. Eventually, this will promote a scalable, easier to configure, more efficient and faster network design [113], [114].

Essentially, SDN architecture consists of three main components: Open Flow (O.F) switch, a controller and a channel (O.F protocol). The principle is to extract/decouple the control plane from/and data plane in the network's devices. Consequently, these two planes interact using O.F interfaces. As a result, the functions of control planes can be combined or centralised in a unified controller rather than in a set of distributed devices bound with a stringent control plane. The controller in turn sustains the software abstract and presents application programming interfaces (APIs) to the network providers [115]. The O.F switch consists of flow tables, which are installed, modified and deleted by the controller using the O.F protocol, to enable the switch evaluating the received traffic in terms of the packet's contents [116]. The flow tables comprise: i) flow entries that hold a group of match fields (such as QoS

type, IP address, MAC address, packet priority, etc), and ii) actions, those executed by the switch to process the received packets. Once the switch receives a packet, it examines its flow tables to search for a match. Matching means that the incoming packets will be compared against the entry's fields to know if the packets are eligible to be handled or not. If the packet's content match the installed entry fields, the packets will be processed over an action correlated to that entry. These actions might include forwarding the packets to a specific port, dropping or flooding the packet to all ports [117]. The SDN layer is generally built on top of SDN controller, and should not obstruct or impede the legacy open systems interconnection (OSI) network model's layers. Therefore, the signalling burden on core network elements such as MME, SGW and PGW can be relieved, resulting in the recovery of their upgrade cost. However, bringing this idea to the C-RAN requires a comprehensive comparison for C-RAN with and without the use of SDN [114]:

1. Decentralising or separating the control and data plane by using SDN simplifies the design of the CN's entities (i.e. SGW, PGW and MME, etc.) by transferring part of the control plane's functions to the controller. This in turn decreases CN entities's overhead and therefore reduces their functional complexities.

2. Implementing SDN can benefit C-RAN more than distributed RAN (D-RAN) or traditional architecture. As C-RAN exhibits a unified infrastructure and management capabilities, the signalling delay amongst the BBUs and the controller can be significantly reduced when connecting near located BBUs rather than distributed BBUs. However, when pooling the BBUs, new protocols, smart SON systems and programming based management systems will be necessary to ensure the full benefits of C-RAN, SDN offers a solution by dominating the BBU pool's administration.

3. In contrast to D-RAN, C-RAN deploys small coverage RRHs to bring the cell site closer to the UEs, which shortens the transmission distance. Equivalently, SDN while deploying O.F switches, it also brings the CN's contents closer to the

UEs. The flow tables of the switches allow direct communication with the UEs, which means forwarding UEs' packets to their destination before approaching the BBU pool. This procedure further relaxes the latency bottleneck.

4. C-RAN is considered green architecture compared to D-RAN. In addition, there is a demand to develop and launch new services, policies and applications to unleash the maximum potential of C-RAN. This requires flexible and a programmable based paradigm such as SDN [118].

5. SDN is designed for wired and not for wireless networks. This consequently introduces complexity and causes inherent weakness to its effective deployment, since it requires an isolated and non-interfering wireless channel amongst the controller-switches and switches-RRHs. Eventually, power allocation and optimisation methods will be required amongst the mentioned parties. Nevertheless, the literature provides solutions for resources management and scheduling for an SDN based mobile communications, as in [119], [120] and [121].

While gaining the benefits of SDN, the conjectural increase in PC due to the addition of SDN devices to C-RAN can be neutralised by the following trends: (i) the legacy X-2 interface signalling cost amongst the BBUs is now partially or fully relieved, as the controller administrates the signalling amongst the BBUs, (ii) by means of virtualization, which completes SDN to relocate the functions of network from dedicated devices to general servers, resulting in the potential provision of fewer hardware and computational devices that will reduce the imposed PC, and (iii) replacing the currently deployed switches and routers with O.F switches will cost the network only the PC of the controller, not the O.F switches. Therefore, the overall PC can be further reduced. Accordingly, based on the aforementioned demonstrations and the benefits of integrating SDN and C-RAN, it is required to assess the power cost and the overall price of such investment by comparing the PC of C-RAN and SDC-RAN. At this extent, an experiment to measure the PC of SDN devices has been performed in [112]. Based on this measurement, a PC model is derived for both

major SDN enablers devices, these are O.F switch and SDN controller. The final formulation of this model is generally based on number of packets transmitted/received amongst the two parties. By introducing SDN to C-RAN, speculating on the EE of such architecture is essential to evaluate the power cost due to these additions. Due to failure to consider this matter in the literature, it has shaped the motivation for our investigations.

The structure of this Chapter is as follows. In Section 4.2, the component's PM of SDC-RANs and its corresponding PC components are introduced. The total PC of the network is described in Section 4.3. Section 4.4 presents the parameterised PM. Section 4.5 provides selected simulation results. Finally, the conclusion is given in Section 4.6.

4.2 Component's Power Model

The component's PM can be defined by two main parts:

- The CN's PC, which consists of three parts: (i) the BBU pool (ii) control plane units and (iii) SDN units.

- RRHs, with MIMO consideration.

Both the CN and the RRH PMs have been divided into several sub components, each component contributes to the total PM.

Fig. 4.1 shows the block diagram and the PC modules of the SDC-RAN.

CN's PC encompasses mainly the BBU pool, the control plane components (i.e. MME, SGW and PGW) and SDN components (SDN controller and SDN switch). The detailed CN's PC is described in the following subsections:

4.2.1 BBU power consumption (P_{BBU})

The PC for the BBU pool is considered to be the sum of the active BBUs in the BBU pool. This unit's PC has been previousley modelled in Chapter 3.
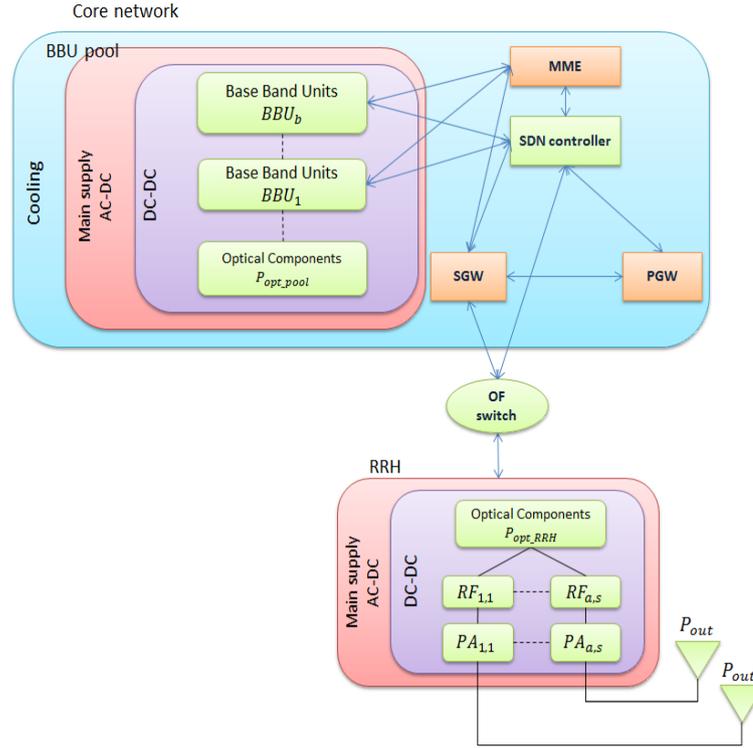
Fig. 4.1. Block diagram of SDC-RAN transceiver.

4.2.2 Control plane PC (P_{cl})

Similarly to BBU mode of operation, MME, SGW and PGW, each is accountable for several functions within the network. For example, MME is responsible for roaming, authentication, authorisation, bearer management functions, UE's handover, etc. SGW and PGW in turn, are in control of many other functions. Eventually, these functions require digital computation and processing power which is then translated to GOPS. Although the type and number of functions in each subcomponent is different from the other, the way power is modelled is not affected. The control plane's PC can be expressed as the combination of the PC of the three main parts, i.e., MME (P_{MME}), SGW (P_{SGW}) and PGW (P_{PGW}), in W, each with its relative speed. By considering the set of functions associated with each unit, P_{cl} in W, then can be expressed as:

$$P_{cl} = P_{MME} + P_{SGW} + P_{PGW}$$

$$= \sum_{q=1}^{Q} \sum_{m \in M_{MME}} (E_{m,MME}^{ref} A^{y_m^A} B^{y_m^B}) \, v_q + \sum_{d=1}^{D} \sum_{sg \in SG_{SGW}} (E_{sg,SGW}^{ref} A^{z_{sg}^A} B^{z_{sg}^B}) \, v_d$$

$$+ \sum_{g=1}^{G} \sum_{pg \in PG_{PGW}} (E_{pg,PGW}^{ref} A^{j_{pg}^A} B^{j_{pg}^B}) \, v_g \tag{4.1}$$

Where M_{MME}, SG_{SGW} and PG_{PGW} designates the set of various functions in regards to MME, SGW and PGW respectively. Q, D and G symbolise the total number of MME, SGW and PGW respectively. v_q, v_d and v_g, in packets per second (pps), denote the relative speed of each unit (the rate at which the packets are processed). $E_{m,MME}^{ref}$, $E_{sg,SGW}^{ref}$ and $E_{pg,PGW}^{ref}$, in Joule (J), denote the MME, SGW and PGW energy consumption of the m-th MME, sg-th SGW and pg-th PGW function corresponding to a reference value, respectively. y_m^A, z_{sg}^A and j_{pg}^A represent the scaling exponent of the number of RF chains that the MME, SGW and PGW serve, respectively. y_m^B, z_{sg}^B and j_{pg}^B stand for the scaling exponent of the bandwidth that MME, SGW and PGW exploit, respectively.

4.2.3   SDN power consumption (P_{SDN})

This part can be further distributed into two main parts:

4.2.3.1   O.F switch power consumption (P_{switch})

The switch PC also can be divided into two sub components:

4.2.3.1.1   Port power consumption (P_{port}) Each switch may contain several electronic chips (ports), each port with an associated relative line speed, this portion represents the static PC of the switch.

$$P_{port} = \sum_{of=1}^{OF} \sum_{sw=1}^{SW} v_{sw} \times (E_{prt,of,sw}) \tag{4.2}$$

$E_{prt,of,sw}$, in J, indicates the energy consumed at full speed by sw-th port located in switch of. v_{sw}, in pps, is the line speed corresponding the sw-th port (the rate at which the packets are processed). OF is the total number of O.F switches.

4.2.3.1.2 O.F Traffic power consumption (P_{flow}) The traffic or the flow represents the number of packets received by a particular O.F switch to be matched and accordingly actioned to the selected destination, this portion of PC represents the dynamic load in the switch. P_{flow}, in W, can be modelled as:

$$P_{flow} = \sum_{of}^{OF} \left(\sum_{fl}^{FL} R_{pkt} \left[\sum_{mt}^{MT} MC(fl, mt) E_{mt} + \sum_{ds}^{DS} NC(fl, ds) E_{ds} + \sum_{ak}^{AK} ACT(fl, ak) E_{ak} \right] \right)$$

$$(4.3)$$

Where FL, symbolises the total number of flows received with an associated packet rate (R_{pkt}), in pps. For each processed flow, there are three possible categories, these are matching, non-matching and actions; MC, NC and ACT, respectively. On the other hand, MT, DS and AK denote the total number of matched, non-matched/discarded packets (the packets without matching in the corresponding flow table) and actions in each flow, respectively. E_{mt}, E_{ds} and E_{ak}, in J, refers to the energy required to process mt-th match, ds-th non-match packet and ak-th action, correspondingly. At last, the total O.F switch PC P_{switch}, in W, can be modelled:

$$P_{switch} = P_{port} + P_{flow}$$

$$(4.4)$$

4.2.3.2 SDN controller PC (P_{SDNctl})

The controller PC can be expressed as the rates $R_{ctl,of}^{O.F}$, $R_{ctl,q}^{MME}$, $R_{ctl,d}^{SGW}$ and $R_{ctl,g}^{PGW}$, in pps, of outgoing O.F protocol control signalling from the ctl-th controller to the of-th O.F switch, q-th MME, d-th SGW and g-th PGW, correspondingly. Beside the corresponding energy $E_{ctl,of}^{O.F}$, $E_{ctl,q}^{MME}$, $E_{ctl,d}^{SGW}$ and $E_{ctl,g}^{PGW}$, in J/packet required to send the packet to the O.F, MME, SGW and PGW respectively. If CTL is the total

number of SDN controllers, and the connection is bidirectional amongst the C-RAN control plane units and the controller, the controller's PC can be modelled as:

$$P_{SDNctl} = \sum_{ctl=1}^{CTL} \left(\sum_{of}^{OF} R_{ctl,of}^{O.F} + 2\left[\sum_{q=1}^{Q} R_{ctl,q}^{MME} E_{ctl,q}^{MME} + \sum_{d=1}^{D} R_{ctl,d}^{SGW} E_{ctl,d}^{SGW} + \sum_{g=1}^{G} R_{ctl,g}^{PGW} E_{ctl,g}^{PGW} \right] \right)$$

(4.5)

Finally, the SDN's PC, in W, can be expressed as the aggregation of both the switch and the controller:

$$P_{SDN} = P_{switch} + P_{SDNctl}$$

(4.6)

4.2.4   BBU pool DC-DC Conversion $(P_{DC,P}^{sdn})$

The PC caused by the DC conversion $(P_{DC,P}^{sdn})$, in W, of the BBU pool containing the control plane units is modelled using the same style of Chapter 3 (i.e. 3.2.1.2):

$$P_{DC,P}^{sdn} = \sum_{b=1}^{B} \sum_{r=1}^{R} l_{DC,P}(\eta_{DC,P}) \times (P_{BBU_b^r} + P_{opt,P_{b,r}} + P_{SDN} + P_{cl})$$

(4.7)

Where, $b \in \{1, ..., B\}$ stands for the number of active BBUs in the BBU pool; $r \in \{1, ..., R\}$ indicates the number of active RRHs; $P_{BBU_b^r}$, in W, is the PC of b-th BBU, which is connected to r-th RRH; $P_{opt,P_{b,r}}$, in W, is the PC by the optical device in the BBU pool which connects the b-th BBU to r-th RRH; $l_{DC,P}$ is the loss caused by DC-DC conversion as a function of DC conversion efficiency $(\eta_{DC,P})$, which can be known from the device's data sheet. The loss function $l(\eta)$ can be modelled as in Chapter 3 (i.e. 3.3).

4.2.5   Mains supply (MS), AC-DC Conversion PC $(P_{MS,P})$

The PC of this unit, in W, is generally modelled the same way as the DC-DC power conversion, and it is given as:

$$P_{MS,P} = P_{DC,P} + \sum_{b=1}^{B}\sum_{r=1}^{R} l_{MS,P}(\eta_{MS,P}) \times (P_{BBU_b^r} + P_{opt,P_{b,r}} + P_{SDN} + P_{cl}) \quad (4.8)$$

$l_{MS,P}$ represents the measured losses of the MS power conversion as a function of the MS conversion efficiency $(\eta_{MS,P})$.

4.2.6 Cooling

The cooling unit is responsible for cooling the entire components in the CN. However, the PC is modelled to be proportional to the consumption of all other components in the BBU pool. If l_{cool} is the cooling loss, the cooling PC (P_{cool}), in W, can be modelled as:

$$P_{cool} = l_{cool} \times (P_{MS,P} + P_{DC,P} + \sum_{b=1}^{B}\sum_{r=1}^{R} P_{BBU_b^r} + P_{opt,P_{b,r}} + P_{SDN} + P_{cl}) \quad (4.9)$$

4.3 Total Power Consumption $(P_{SDC-RAN})$

The total PC of the SDC-RAN network $(P_{SDC-RAN})$, in W, is therefore the sum of the CN's PC (P_{CN}) and RRHs' PC (P_{RRH}), it is formulated as the following:

$$P_{SDC-RAN} = P_{CN} + P_{RRH} \quad (4.10)$$

P_{CN} , in W, is calculated by aggregating the PC of the corresponding components:

$$P_{CN} = P_{cool} + P_{MS,P} + P_{DC,P} + P_{SDN} + P_{cl} + \sum_{b=1}^{B}\sum_{r=1}^{R}\sum_{a=1}^{A} P_{BBU_{b,r,a}} + P_{opt,P_b^r} \quad (4.11)$$

$P_{BBU_{b,r,a}}$, in W, denotes the PC of b-th BBU attached to r-th RRH mounting a-th antenna. RRHs can be served by any active BBU within the BBU pool. This means that BBUs-RRHs mapping can be dynamic depending on the traffic conditions. This

service diversity is considered in the model. The total PC of the RRHs is formulated as:

$$P_{RRH} = P_{MS,R} + P_{DC,R} + \sum_{b=1}^{B} \sum_{r=1}^{R} \sum_{a=1}^{A} (P_{PA} + P_{RF})_{r,a}^{b} + P_{opt,R_r} \qquad (4.12)$$

$P_{PA}{}_{r,a}^{b}$, $P_{RF}{}_{r,a}^{b}$ in W, symbolise the PC of the PA and RF respectively; of a-th antenna served by r-th RRH that is attached to b-th BBU.

4.4 Parameterised PM

There are many assumptions have been made to enable the parameterised model, the following analysis and approximations are made in relevance to the SotA BS's PM:

- The BBU and RFs' PC, both scale linearly with the number of antennas (A) and the bandwidth (BW), i.e., $P_{BBU} = A \times BW \times (P_{BBU}^{pm})$ and $P_{RF} = A \times BW \times (P_{RF}^{pm})$. P_{BBU}^{pm}, P_{RF}^{pm} denote the PC by the BBU and RF in the parameterised model (pm), respectively.

- The control plane components' PC i.e. (MME, SGW and PGW) scale linearly with the number of antenna and the bandwidth i.e. $P_{MME} = A \times BW \times (P_{MME}^{pm})$, $P_{SGW} = A \times BW \times (P_{SGW}^{pm})$ and $P_{PGW} = A \times BW \times (P_{PGW}^{pm})$. The similar holds true for the O.F switch and the SDN controller PC, i.e. $P_{switch} = A \times BW \times (P_{switch}^{pm})$ and $P_{SDNctrl} = A \times BW \times (P_{SDNctrl}^{pm})$.

- The PC of PA depends on the maximum power transmitted per antenna (P_{out}/A) and its efficiency (η_{PA}). The feeder loss between PA and the antenna can be ignored for the RRH as the PA is placed close to the antenna. The PA's PC can be represented as $P_{PA} = \dfrac{P_{out}}{A \eta_{PA}}$.

- DC-DC, AC-DC conversions as well as cooling's PC, scale linearly with the other components PC and are approximated by the loss factors $\sigma_{DC,P}$, $\sigma_{MS,P}$,

σ_{cool}, $\sigma_{DC,R}$ and $\sigma_{MS,R}$ for BBU pool DC, BBU pool MS, cooling, RRH's DC, and RRH's MS loss factors respectively.

- The optical transceivers' PC $P_{opt,P}$ and $P_{opt,R}$, scale linearly with the number of BBUs and RRHs in the network.

- The losses incurred by the cable's connections amongst the entities within the CN or the backhaul can be expressed by the loss factor ($\sigma_{backhaul}$), moreover, it scales linearly with the numbers of BBUs.

- The losses incurred by the fronthaul optical fiber are estimated by the loss factor ($\sigma_{fronthaul}$). Theses losses scale linearly with the numbers of RRHs and can be adapted to meet the fiber lengths and number of connectors and splices used.

The maximum PC (P_{SDCRAN}^{pm}), in W, can be formulated by aggregation of the PC of the CN, which constitute of (BBUs, control plane, O.F switch and SDN controller) serving a single RRH:

$$
\begin{aligned}
P_{SDCRAN}^{pm} &= P_{CN}^{pm} + P_{RRH}^{pm} \\
&= \frac{A \times BW \times (P_{BBU}^{pm} + P_{MME}^{pm} + P_{SGW}^{pm} + P_{PGW}^{pm} + P_{switch}^{pm} + P_{SDNctrl}^{pm}) + P_{opt,P}}{(1 - \sigma_{DC,P})(1 - \sigma_{MS,P})(1 - \sigma_{cool})(1 - \sigma_{backhaul})} \\
&+ \frac{A \times BW \times (P_{RF}^{pm}) + \dfrac{P_{max}}{A\eta_{PA}} + P_{opt,R}}{(1 - \sigma_{DC,R})(1 - \sigma_{MS,R})(1 - \sigma_{fronthaul})}
\end{aligned}
$$

$$(4.13)$$

Then the total number of R RRHs and B BBUs is considered to obtain the total PC of the network (P_{supply}^{SDCRAN}), in W:

$$
P_{supply}^{SDCRAN} = B.P_{CN}^{pm} + R.P_{RRH}^{pm}
$$

$$(4.14)$$

4.5   Results and Discussion

The proposed model is used to effectively analyse and identify the PC in terms of bandwidth, varying antenna numbers, and varying number of BBUs, RRHs and SDN devices. The parameters of the study were selected according to [41] and [112] when possible. The resulting parameters are provided in Table 3.1, which summarises the SotA PC measurements based of the RRH and SDN devices.

The projected model is verified for one, two, and four antennas. R is up to 60, while B equals 20. Nevertheless, these numbers can be adjusted according to the network operator's architectural or configuration demands; this allows us to observe the variations in the individual parameters. The allocated bandwidth is 10 MHz; if increased, the PC is expected to increase as well. Fig. 4.2 shows the cooling PC of C-RAN and SDC-RAN of one, two and four antenna configurations. Due to the fact that the PA and RF components are no longer contributing to the cooling PC in the BBU pool, C-RAN reduces the cooling PC compared to the SotA BSs.
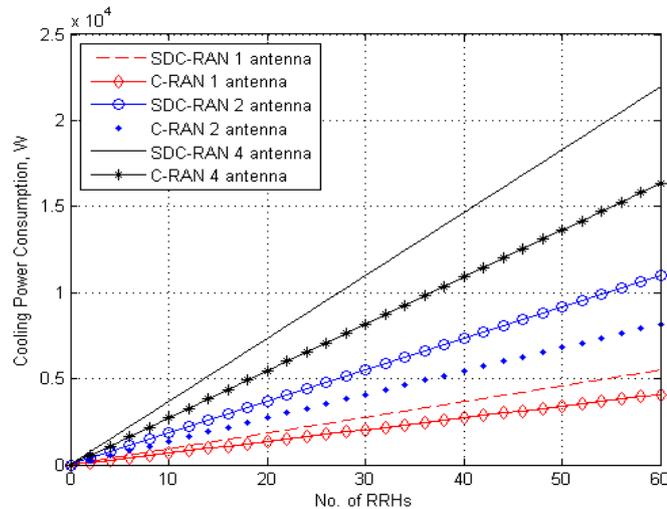

Fig. 4.2. Comparison the cooling PC of C-RAN and SDC-RAN, with one, two and four antennas configurations.

Table 4.1 shows the differences in the PC between SDC-RAN and C-RAN models for 4 antenna configuration.

Table 4.1

Comparison cooling PC of C-RAN and SDC-RAN for four antennas configurations.

No. RRHs	SDC-RAN(W)	C-RAN(W)
10	3000	2500
20	7200	5500
30	10200	8000
40	14500	10100
50	18200	14000
60	22500	16200

Fig. 4.3 shows the total network PC. By using the percentage change rule, i.e., $\dfrac{(v_1 - v_2)}{|v_1|} \times 100\%$ to compare the values of this Figure, it was found that the total PC increasing percentage by adding SDN to the C-RAN architecture is about 20%, for all antenna configurations. However, this value can be endangered to constraints such as operators' equipment PC and quality.

Fig. 4.4 indicates the total PC as a function of the system's bandwidth share with varying numbers of antennas.

The accuracy of the simplified parameterised model and the components model can be evaluated and compared using the loss function (3.3). The main key comparison can be done by acknowledging the initial losses l_o of each subcomponent and the corresponding exponential decay constant (ϖ); the latter have been subjected to the manufacturer design. However, to compare both the components and the parameterised models in terms of accuracy, some assumptions have been made to the component PM:

1. The efficiency η is 90% for all the sub components (i.e. AC, DC converters), and the initial loss l_o is 0.009.

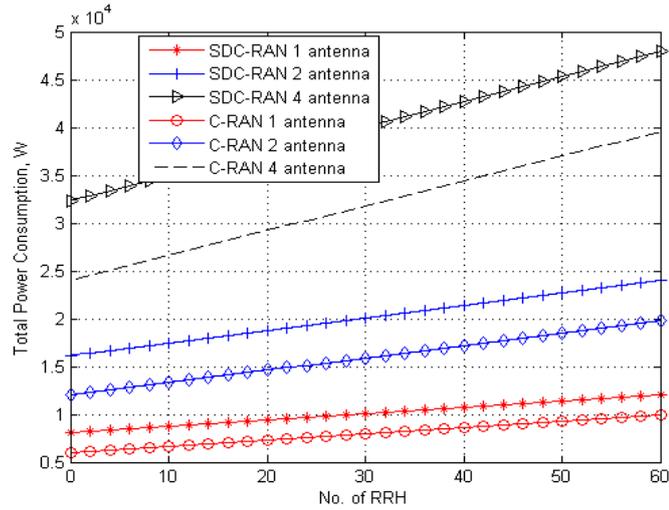Fig. 4.3. Comparison of C-RAN and SDC-RAN total PC, 20 BBUs are used to serve 60 RRHs, with one, two, and four antennas configurations.
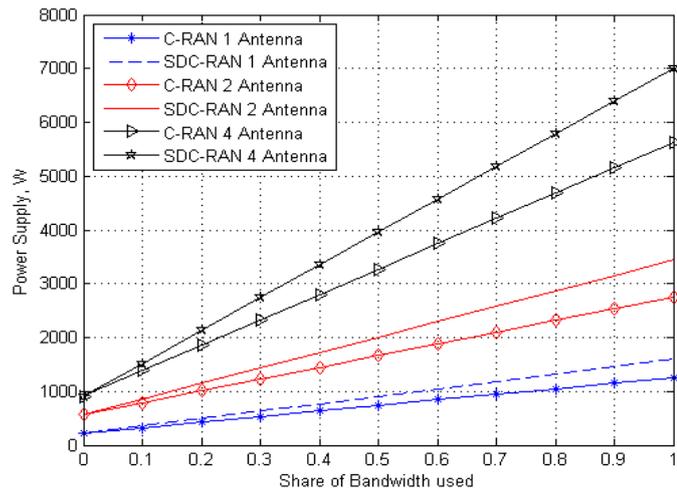

Fig. 4.4. Comparison of C-RAN and SDC-RAN total PC as a function of the used bandwidth (10 MHz).

2. (3.3) was used to evaluate the losses values which are required in (4.7), (4.8), (3.7), and (3.8).

3. The total component's PM were calculated using (4.10).

4. It was compared with the parameterised model of (4.13).

Based on the subcomponents' data sheet and by using different ϖ values, the results show as accurate parameterised as components PM. Fig. 4.5 shows the total PC of SDC-RAN with different ϖ values and different number of RRHs.
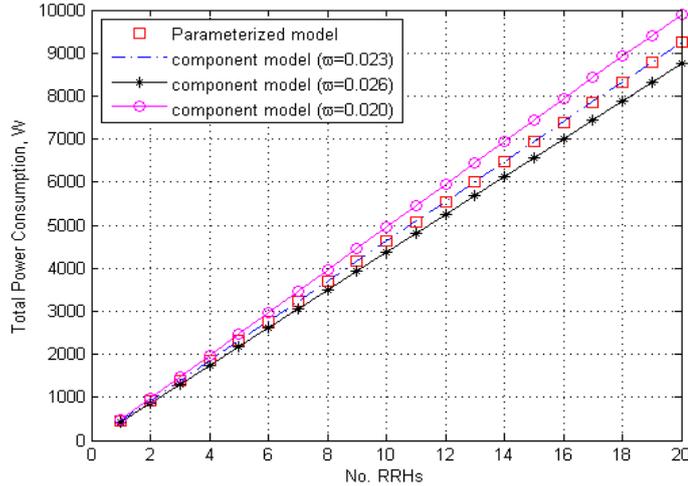
Fig. 4.5. Accuracy comparison of the total PC of the components and parameterised PMs with different values of (ϖ), using only one antenna configuration.

4.6 Summary

Cloudification the networks increases the signalling cost and degrades scalability. The latter can be mitigated by using SDN, while C-RAN reduces the PC. The components in conjunction with a parameterised PM are presented within this paper to demonstrate PC calculation of the extended network architecture SDC-RAN, according to operational parameters and the varying vendor configurations. The model is considered to visualise the PC cost, along with a comparison C-RAN and SDC-RAN models' performance regarding cooling and total PC, with varying parameters such as antenna and bandwidth sweep.

Intuitively, C-RAN model has less PC than SDC-RAN. However, the benefits demonstrated and the characteristics gained by the latter justifies the power cost, especially, when SDN units participate and cooperate the functional computations and the signalling process of the legacy control units to take over the administration and unleash the network potentials. Furthermore, SDN is advocated because of the capability and the flexibility within its architecture to introduce new services to the network and integrate with the new power reduction methods, such as NFV. However, the amount of power cost due to SDN deployment cannot be realised unless a reliable PM is introduced, this research resolves this ambiguity. Furthermore, the results show that the parameterised SDC-RAN is accurate as the component's PM subject to equipment design quality. Finally, the following Chapter will discuss the way to improve the EE of the network, this can be done using virtualisation. Virtualisation is a rising technique in the coming generations of communication. Therefore, providing a PC model that aim to reduce the PC of the network is necessary.

Chapter 5: Energy Efficiency of Virtualised C-RAN

5.1 Introduction

Intensifying the number of deployed RRHs and active BBUs can maintain a considerable amount of PC. Recently, NFV concept benches itself on the evolutionary route that leads to fully authorise 5G networks. It is emerged with the promise of reduction in the cost concurrent with deploying and operating the large networks. Once extracting the network functions of the dedicated appliances and employ them as software instances, it is now easier to run multiple instances (called VMs) using only one general purpose/off-the-shelf hardware. This shall enhance the mobile network flexibility, scalability, and enable Network-as-a-Service business, while deploying new services and applications will be faster [122], [123], [124].

However, running multiple VMs on a single hardware requires a supervisor or manager. This manager is also called a hyper-visor (HV). The HV is a software that runs on the host's hardware to dynamically control and allow the host server to be shared by guest VMs. Each VM then appears to hold or utilise server's RAM, CPU, NIC and HDD all to itself. But in fact, each VM shares these resources with other VMs. The HV then assures that the hosted VMs cannot obstruct each other while accessing these resources. However, the presence of the HV within the host server increases RAM accesses, CPU computations and storage usage. This increment happens as a consequence to the increased interruptions and orchestration sessions

between the VMs and HV. Therefore, the VMs have yielded extra overhead added to the server PC. Such matter urges to evaluate this increment in the PC and other trade-offs, such as the execution time delay within each host compared to the non virtualised servers. Furthermore, it is necessary to compare the entire PC of the CN or BBU pool with and without using virtualisation [62]. In general, virtualisation greatly reduces the overall network's PC by using fewer servers. However, such gain requires awareness to the consequences facing each virtualised server regarding the PC and latency, as follows:

5.1.1 Advantages and Disadvantages of NFV

1. NFV appears as a compensator to the increased PC due to integrating new network services and enablers such as SDN and load balancers appliances with C-RAN [125]. This can be achieved by provisioning and sharing for the available servers' resources while cascading multiple VMs to be run by fewer servers. Each server then holds several VMs, and each VM performs different network's function.

2. A virtualised server with 1 VM may take about 5 times more delay to process a packet compared to bare metal counterparts [62]. This delay comes from the fact that each VM owns a small share of the existing server resources to compute its load. This means when there are no resources available at any time, the VM must queue its load and wait for a window to be opened again by the HV. This matter urges to provide optimisation techniques to enhance the HV's cycle scheduling, By which, the fast and dynamic operation can relief such constraint.

3. The virtualised server itself gains a PC overhead due to fully utilising its resources by the VMs. In addition, the PC of each VM is directly proportional to the allocated bandwidth or RBs, which increases the dynamic consumption of the server to maximum.

4. While the number of VMs increases in one server, this reduces the resources' shares allocated to each VM, which more increases the delay to process UEs traffic [126]. Therefore, optimising the number of the VMs installed in one host server is prerequisite in regards to the traffic volume demand and available server resources. Consequently, the VMs can constantly meet their real time requirements.

The presence of the HV in the host server increases the RAM accesses, CPU's functional processes and HDD's usage as a consequence of the increased VMs interruptions and orchestration. The VMs therefore exhibits a measurable extra overhead added to the server PC. Such matter urges to evaluate this increment in PC and other trade-offs such as the execution time delay within each host compared to the non virtualised servers prior to implementing. Furthermore, it is necessary to compare the entire PC of the CN or BBU pool, that is reduced by using Virtualisation technology alongside the non virtualised networks. Hardware virtualization adds substantial overhead, as a busy web-server consumes about 40% more power, synchronised with less efficiency when compared to a non-virtualized server [62]. Therefore, to decide whether or not it is worthwhile to compromise the network performance with total PC reduction, this work has been accomplished.

The structure of this chapter is as follows: in Section 5.2, the server PC is described, the main parts of the server are modelled in terms of PC. In Section 5.3, the components PM is presented. The modelling expected losses are found in Section 5.4. The total PC is calculated in Section 5.5. The penultimate Section 5.6 shows the simulation results. Finally, the summery is given in Section 5.7.

5.2 Server Power Consumption

Server PC comes from the higher computation levels, generated I/O instructions executions and compound accessing for the device resources by the aggregated VMs' applications [126]. Generally, there are four major participants involved within the

constituency of a server PC: RAM, CPU, NIC and storage or HDD, the PC model of each virtualised part can be expressed as follows:

5.2.1 RAM Power Model

Usually, each VM requires a share from the RAM to utilise during operation. However, when the number of installed VMs (N) increases, the size of the RAM as well as its corresponding PC intensifies as a result in order to handle an amplified amount of RAM usage requests [127]. Therefore, it is compulsory to identify the appropriate RAM size (Z_{RAM}) to place in a server that contains N VMs. The proposed method initially describes the change in the RAM size corresponding to the change in N, i.e., $\frac{dZ_{RAM}}{dN} = \alpha Z_{RAM}$. Solving this equation yields $Z_{RAM}(N) = Z_{int}e^{\alpha N}$, where Z_{int} is the initial RAM size, α denotes the increment constant. Afterwards, the corresponding PC ($P_{RAM}(Z_{RAM})$) of the RAM can be modelled based on recognizing the maximum and initial PC of the RAM size. In this case, another constant ($\beta <<$ 0.1) is introduced to shape such change in the PC:

$$\frac{dP_{RAM}}{dZ_{RAM}} = \beta P_{RAM} \tag{5.1}$$

When solving equation (5.1), it yields:

$$P_{RAM}(Z_{RAM}) = P_{intRAM} \; e^{\beta Z_{RAM}} \tag{5.2}$$

Where P_{intRAM} is the initial PC of the RAM.

It is worth noting that the constants α and β may vary based on type of the RAM used and its initial and maximum PC.

Alternatively, a straightforward relation between N and the RAM PC is modelled by representing RAM size as $Z = 2^{i+1}GigaByte$, where i is an index ($i = 0, 1, 2, 3...$) to shape the commercial RAM sizes $i.e., Z = (2, 4, 8, 16, 32GB, ...)$. After, if P_{RAM}^{i} is the PC of the RAM which holds i index, then the change in PC of this specific RAM

dP^i_{RAM} is correlated to the change of N (dN). To estimate this change, the constant ν is presented as follows:

$$\frac{dP^i_{RAM}}{dN} = \nu P^i_{RAM} \tag{5.3}$$

When solving equation (5.3), it produces:

$$P^i_{RAM}(N) = P^i_{intram}\ e^{\nu N} \tag{5.4}$$

Where P^i_{intram} is the initial PC of RAM size with index i.

5.2.2 CPU Power Model

The initial assumption for modelling the PC of CPU is based on practical considerations that each CPU core can hold at least one VM. In other words, the number of cores per CPU (C) is always larger than the number of hosted VMs in one core (N_C), i.e., ($C \geq N_C + 1$). Subsequently, any additional VMs installed in each core means greater amount of power the core will consume until it reaches its maximum. Such affiliation is recognised as being exponential in a core PC level. Logically, the core PC (P_{core}) upsurges from an initial ($P_{intcore}$) while increasing N_C. The above assumptions can be translated into the following model:

$$\frac{dP_{core}}{dN_C} = \varepsilon P_{core} \tag{5.5}$$

Where ($\varepsilon < 0.1$) is a positive constant used to describe how the power is increased. Such linkage tends to be linear when ε approaches 0. However, when solving equation (5.5), it yields:

$$P_{core}(N_C) = P_{intcore}\ e^{\varepsilon N_C} \tag{5.6}$$

If P_{core} is calculated, the CPU's PC (P_{CPU}) can also be known by gathering the PCs of all cores:

$$P_{CPU} = \sum_{c=1}^{C} P_{core(c)} \tag{5.7}$$

Additionally, the total PC, in W, of CPUs per server (P_{server}^{CPU}) can be obtained as follows:

$$P_{server}^{CPU} = \sum_{k=1}^{K} P_{CPU(k)} \tag{5.8}$$

Where K denotes the total number of CPUs per server.

5.2.3 NIC Power Model

As each NIC is shared amongst multiple VMs simultaneously, the higher amount of VMs found per server, the more NICs are required to serve them. The PC of each NIC is obligated to an augmentation in the PC [128]. It was assumed that the maximum number of NICs (L) that can be placed in one server is shared amongst all the VMs, i.e., ($L \propto N$), where ($L \leq N$). For practical consideration, L is equivalent to 8, then each NIC (l) is assigned (N/L) VMs. Moreover, the PC of each NIC rises when the N/L is increased; as more packets will be received and transmitted via a particular NIC. Hence, the model has to outline an increment in the PC so as the virtualised NIC is driven to reach its PC (P_{nic}), up from an initial value (P_{intnic}). Therefore, P_{nic} can be expressed as ($P_{nic}(\frac{N}{L})$), a function of (N/L). This linkage refers to the upsurge of P_{nic} in correspondence to the change of $\frac{N}{L}$, such behaviour can be modelled as follows:

$$\frac{dP_{nic}}{d\frac{N}{L}} = \gamma P_{nic} \tag{5.9}$$

When solving (5.9), the following solution can be obtained:

$$P_{nic}(\frac{N}{L}) = P_{intnic}e^{\gamma \frac{N}{L}} \tag{5.10}$$

Where γ is a constant factor. Next, the total PC, in W, of all NICs per server (P_{server}^{NIC}) can be calculated:

$$P_{server}^{NIC} = \sum_{l}^{L} P_{nic(l)} \tag{5.11}$$

5.2.4 Storage (HDD) Power Model

As a matter of fact, over-utilising the storage by hosting more VMs increases server's PC [129]. There are two main behaviours the model has to reflect on: (i) the storage is shared amongst other servers in the CN, and (ii) the VMs per server boost the storage PC; as they increase its data accesses rate. In other words, there will be two influence parameters synchronised with such matter: the time (t) at which the storage is being utilised, and the number of VMs (N). Both variables then will indicate how much storage PC $(P_{storage}(t, N))$ is drawn. As far as the first variable (t) is concerned, it was considered that other servers can add, access and delete data from the tagged server. In this case, it was assumed that the storage capacity varies during time (t) to represent the sharing capability amongst CN's servers. The second variable in turn clarifies that N VMs increases the storage PC as a consequence to increasing the rate of accessing the stored data. When considering the time, two scenarios are determined: the first is when the storage PC $(P_{storage}^{i})$ increases by time, following the exponential method, the PC can be formulated as:

$$\frac{dP_{storage}^{i}}{dt} = \delta P_{storage}^{i} \tag{5.12}$$

Where $(\delta < 1)$ is a positive constant, which can control the maximum value of the storage PC, when solving (5.12), it yields:

$$P_{storage}^{i}(t) = P_{o}e^{\delta t} \tag{5.13}$$

Where P_o is the initial storage PC. The second case is when the storage PC $(P_{storage}^d(t))$ decreases by time, following the same procedure as in (5.12), the PC model can be expressed as follows:

$$P_{storage}^d(t) = 2P_o - P_o e^{\delta t} = P_o(2 - e^{\delta t}) \tag{5.14}$$

The amount $(2P_o)$ is added to uplift the initial consumption of $(P_{storage}^d(t))$ to start from (P_o) at time $(t = 0)$, so as $P_{storage}^d(t)$ and $P_{storage}^i(t)$ start from the same point. The second variable of the storage is modelled when the PC is proportionally increased with the number of VMs. The procedure of (5.12) is followed and the resulting PC model can be obtained as follows:

$$P_{storage}(N) = P_o e^{\xi N} \tag{5.15}$$

Where ξ is a positive constant. Note that the lesser value of the constants ξ and δ are assigned, the more the model approaches to be linear. As the variables t and N are independent of each other, the model separates their PC, yet the result of each variable is aggregated, i.e., $P_{storage}(t)$ will be added to $P_{storage}(N)$. The total virtualised storage PC, in W, can be finalised as:

$$P_{storage}(t, N) = \begin{cases} P_o(e^{\delta t} + e^{\xi N}) & \text{if } \delta \text{ increases} \\ P_o(2 - e^{\delta t} + e^{\xi N}) & \text{if } \delta \text{ decreases} \end{cases} \tag{5.16}$$

5.2.5 HyperVisor (HV) Power model

The HV assigns a number of accesses tasks (AS) to enable the VMs to compute their load within the server's resources, if the PC per task per VM is PAS, then the PC of the HV can be modelled as follows:

$$P_{HV} = \sum_{n=1}^{N} \sum_{as=1}^{AS} PAS_{(as,n)} \tag{5.17}$$

Where $PAS_{(as,n)}$ is the PC of the as-th job allocated to n-th VM .

5.3   Components Power Model

The participating PC modules of vC-RAN paradigm encompasses mainly two parts: virtual BBU pool (vBBU pool) and virtual control plane units (i.e. vMME, vSGW and vPGW).

The vBBU server PC (P_{server}^{BBU}) consists of whatever virtualised components found in the server, this can be evaluated in W, as follows:

$$P_{server}^{BBU} = \sum_{s1=1}^{S1} (P_{RAM} + P_{server}^{NIC} + P_{BBU}^{CPU} + P_{storage} + P_{HV})_{s1} \tag{5.18}$$

Where

$$P_{BBU}^{CPU} = \sum_{k1=1}^{K1} \sum_{c1=1}^{C1} P_{core(k1,c1)}^{BBU} \tag{5.19}$$

and $S1$, is the total number of BBU servers.

Comparably to BBU functions, PGW, MME, and SGW cores PC can be modelled following the same style by taking into consideration the functions set of each unit [125]. Therefore, the PC of control plane server (P_{server}^{cl}) is introduced as:

$$P_{server}^{cl} = \sum_{s2=1}^{S2} (P_{RAM} + P_{server}^{NIC} + P_{cl}^{CPU} + P_{storage} + P_{HV})_{s2} \tag{5.20}$$

Where $S2$ denotes the total number of servers that host the control plane's, P_{cl}^{CPU} is the control plane CPUs PC, which can be modelled as:

$$P_{cl}^{CPU} = P_{MME}^{CPU} + P_{SGW}^{CPU} + P_{PGW}^{CPU} \tag{5.21}$$

The PCs P_{MME}^{CPU}, P_{SGW}^{CPU} and P_{PGW}^{CPU} of the CPUs belong to MME, SGW and PGW, respectively, are equivalent to the sum of their corresponding cores' PCs:

$$P_{MME}^{CPU} = \sum_{k2=1}^{K2} \sum_{c2=1}^{C2} P_{core(k2,c2)}^{MME} \tag{5.22}$$

$$P_{SGW}^{CPU} = \sum_{k3=1}^{K3} \sum_{c3=1}^{C3} P_{core(k3,c3)}^{SGW} \tag{5.23}$$

$$P_{PGW}^{CPU} = \sum_{k4=1}^{K4} \sum_{c4=1}^{C4} P_{core(k4,c4)}^{PGW} \tag{5.24}$$

Where P_{core}^{MME}, P_{core}^{SGW} and P_{core}^{PGW} denote the core PC of MME, SGW and PGW, these are responsible of MME, SGW and PGW functions, respectively. $C1$, $C2$, $C3$ and $C4$ denote the total number of BBUs, MMEs, SGWs and PGWs cores, correspondingly. $K1$, $K2$, $K3$ and $K4$ are the total number of CPUs belong to BBUs, MMEs, SGWs and PGWs, respectively.

5.4   System Performance Loss

The previous description presents how the number of VMs (N) affect server's PC. Herein, the modelling can go further to a single VM level. The degraded performance mentioned in Subsection (5.1.1) can be modelled regarding the increased server delay and processed RBs within each VM by time. The reason why the exponential function is chosen in the previous and following formulation, as it was mentioned in [26] that the consumption of computing resources scales non-linearly with the number of UEs. To fulfil this, it was assumed that the dynamic PC of each VM is linearly or convexly proportional with the number of physical resource blocks (RB) or bandwidth processed at each VM. In total, the PC of all VMs found in the server draw exponential (non-linear) consumption with the processed RBs. The RB is an easy parameter to obtain [105]. The higher processed RBs of each VM, the more share from the limited server resources is demanded and the more PC. In OFDMA systems, the bandwidth is divided into RBs, each RB is 180 kHz wide in frequency and one slot (0.5 ms) in time domain. The RB is the smallest time-frequency unit which can be allocated to the network user. By doing so, we exceeded the cost of possessing a server and the complexity of intruding the server. Concurrently, we have produced a PM based on best-known network optimisation parameter.

To model the dynamic or linear PC of each VM at each server component, the number of RBs linearly influences the constant base load PCs P_{intRAM}, $P_{intcore}$, P_{intnic}

or P_o, these are independent of N or the number of RBs allocated to each VM n, where $n \in N$. These initial values are increased by the amount $(\sum_n^N e^{\vartheta * RB_n})$, where ϑ is the increment factor due to processing RB_n by n-th VM in any of the server resources. This granularity in the PC is jointly added to the models of Section (5.2) to produce total PC of the virtualised CN (P_{server}^{vCN}), as follows:

$$P_{server}^{vCN} = (P_{server}^{BBU} + \sum_{s_1}^{S_1} \sum_{n}^{N} e^{\vartheta * RB_n}) + (P_{server}^{cl} + \sum_{s_2}^{S_2} \sum_{n}^{N} e^{\vartheta * RB_n}) \tag{5.25}$$

The above equation draws about 40% gain in the PC of each server, this is originated from increasing both VMs and RBs allocation, which fits but not limited to the real time server measurements presented in [62]. Figure 5.1 compares the initial PC of a single server and the power cost imported by increasing the number of RBs, VMs and both.
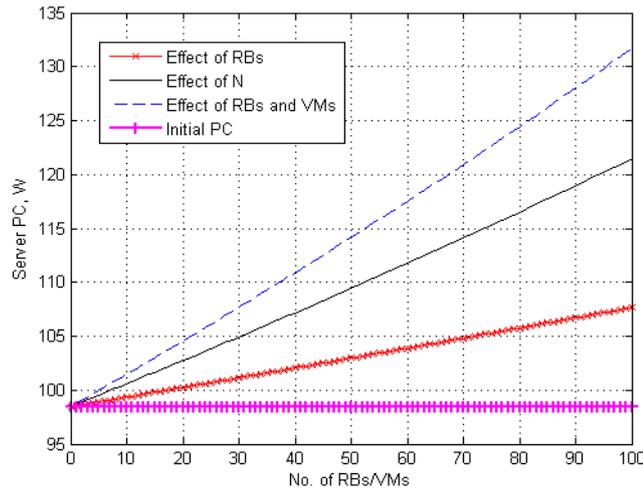

Fig. 5.1. PC of a BBU server at 100 VMs or RBs

Another performance factor is the time it takes the VM to process these RBs. The execution time of the traditional load in a BBU increases linearly with the number of RBs and the modulation coding scheme ($MCS \in \{9, 16, 25\}$) used to transmit/receive these RBs [105], and a single VM may require 5 times more delay to processed a packet

compared to a bare metal LTE BBU; due to increased accessing calls and interrupts between VM-HV and HV-server resources. Modelling this concept requires introducing a factor called MCS index (mcs) to describe the degree of linearity between the RBs and execution time in a bare BBU server (τ_{bare}), as shown in Fig. 5.2, where $\tau_{bare} = \tau_{init} + (mcs * RB)$, and τ_{init} is the initial BBU delay due to other BBU functions rather than MCS. Subsequently, the HV delay (τ_{HV}) is added to the above description, i.e. $\tau_v = \tau_{bare} + \tau_{HV}$, where τ_v is the execution time of virtualised server when 1 VM is installed, as shown in Fig. 5.3 for different MCS values. After, the total execution time (τ) of all VMs is expressed as $\tau = \sum_s^S \sum_n^N \tau_v^{s,n}$, where $\tau_v^{s,n}$ denotes the execution time of VM n located in server s.
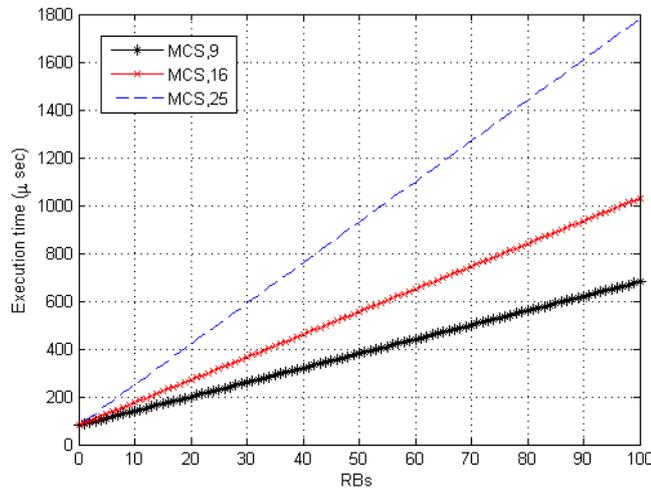

Fig. 5.2. Execution time of a bare server at different MCS

5.5 Total Power Consumption

Total PC of the CN is conformed to the effects of other losses such as AC-DC, DC-DC and cooling loss in a straightforward manner. This offers an easy but accurate way to calculate their PC without undergoing the computations of each unit. Therefore, AC-DC, DC-DC and cooling PC are linearly scaled with other components' PC and
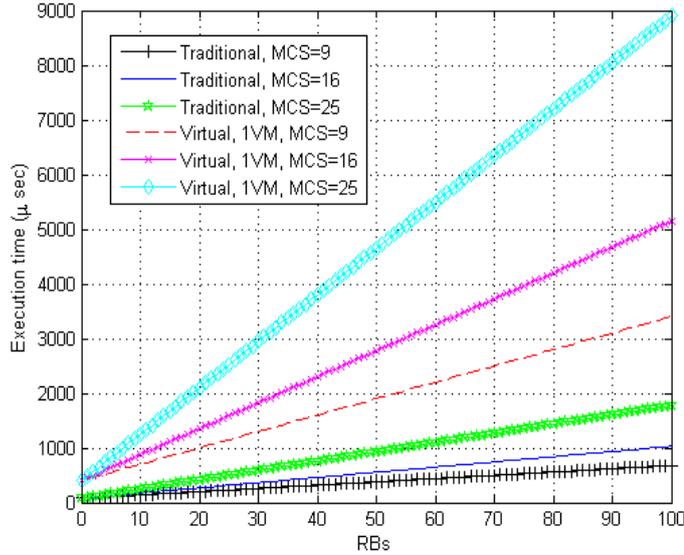
Fig. 5.3. Comparing the execution time of a bare metal and virtualised server at 1 VM and different MCSs.

approximated by using loss factors (σ_{DC}, σ_{AC}, σ_{cool}) to represent AC-DC, DC-DC and cooling loss factors, respectively. Successively, the total PC of virtualised C-RAN (P_{vCRAN}), in W, is formulated as follows:

$$P_{vCRAN} = \frac{P_{server}^{vCN}}{(1-\sigma_{DC})(1-\sigma_{MS})(1-\sigma_{cool})} \tag{5.26}$$

5.6 Results

To correlate the findings of this work with a real time measurement, the resulting parameters were selected from [41], [127], [128], [129], as shown in Table 5.2. The experimental data related to the PC of each component in the server demonstrates that the initial PC of the CPU is 29.6W, RAM is 4W, NIC is 2W and HDD is 25W, while rest of PC in the server is resulted from the overhead. While emphasising on the context of showing how the VMs increasingly affect the PC, only $P_{storage}^{i}(t)$ of the storage PC is considered. Furthermore, the second RAM PC model is adapted

in the results. To compare the model with particular experimental data without losing the generality of this work, the parameters used in Table 5.2 have resulted about 40 % PC increment at each virtualised server which fits what is measured in [62]. This increment also synchronised with degraded performance in each server as mentioned in Section (5.4). Fig. 5.4 shows a comparison of BBU pool's PC with and without virtualisation for different number of virtualised host servers (i.e., 5, 10, and 20 servers). This Figure compares the cost of processing the maximum LTE bandwidth allowed (100 RBs) by each one of the 100 bare metal BBU servers, and also shows the effect of processing the same amount of RBs by each of the 100 VMs installed in the virtualised servers.
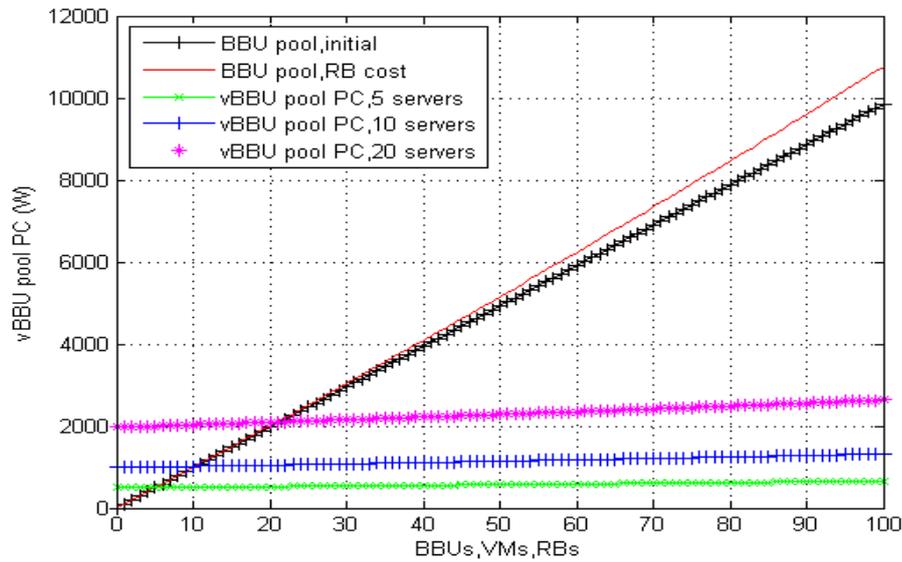

Fig. 5.4. Comparison of the BBU pool PC with and without virtualisation of 100 BBUs or VMs.

By using the rule of percentage change ($\frac{v_1 - v_2}{|v_1|} \times 100\%$), it was found that the maximum reductions in the PCs are about 93, 88, and 74 % corresponding to running 5, 10 and 20 virtualised servers compared with the non virtualised servers. On the same basis, Fig. 5.5 shows a comparison of the entire CN's PC with and without virtualisation. Three control plane servers for MME, SGW and PGW were added

to the system while following the previous procedure. This case has resulted a total PC reduction of about 91, 83 and 73 % when respectively running 8, 13, and 23 virtualised server compared to bare metal servers. Continuously, Fig. 5.6 shows cooling PC comparison of the CN. Cooling PC has been reduced to about 93, 88, and 74 % when compared to the bare metal servers' PC in [130]. Table 5.1 shows the results comparisons of Fig 5.4 between the initial BBU pool PC and virtualised case with 20 servers.

Table 5.1

Comparison of BBU pool PC with and without virtualisation for 20 servers.

(BBUs,VMs,RBs)	BBU pool(W)	vBBU,20(W)
20	2000	2000
30	3000	2100
40	4000	2200
50	5000	2300
60	6000	2400
70	7000	2500
80	8000	2600
90	9000	2700
100	10000	2800

It is worth noting that if another type of server with different specifications was used in this comparison, the eventual results of these comparisons will slightly change. This is because such matter instantly affects the initial and maximum consumption of each component within the virtualised servers, which affects the final outcomes. On the other hand, the more number of bare servers or VMs to be involved in this comparison, the more PC reduction; synchronised with more loss of performance. The reason is that any large number of bare servers will be multiplied by each server's PC. However, in virtualisation method, the main factor (N) is always less than the
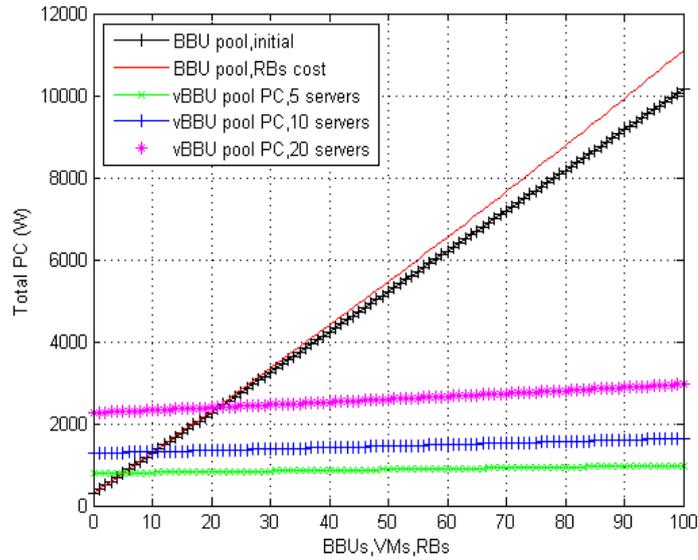
Fig. 5.5. Comparison the total PC of the CN with and without virtualisation of 100 BBUs, VMs or RBs.
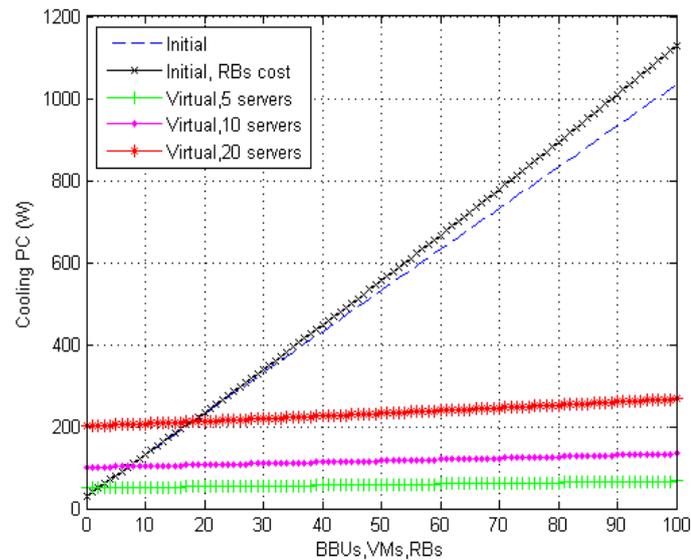

Fig. 5.6. Comparison of cooling PC of the CN with and without virtualisation.

total number of bare servers. This number only influences the number of virtualised servers (S_1 and S_2) while bearing RB cost.

Table 5.2
Model parameters

Component	Unit	Value	Reference
P_{server}^{BBU}	W	29.6	[53]
P_{server}^{cl}	W	29.6	[53]
$P_{intcore}$	W	3.7	[53]
σ_{DC}	-	0.075	[53]
σ_{MS}	-	0.09	[53]
σ_{cool}	-	0.1	[53]
P_{intram}^{i}	W	4	[127]
P_{intnic}	W	2	[128]
P_o	W	10	[129]
τ_{init}	μ sec	80	[105]
τ_{HV}	μ sec	500	[62]
β	-	0.003	[-]
μ	-	0.1	[-]
γ	-	0.08	[-]
δ	-	-0.001	[-]
ξ	-	0.007	[-]
ν	-	0.005	[-]
ϑ	-	0.001	[-]
ε	-	0.009	[-]
C	-	4	[-]
$mcs, 9$	-	6	[-]
$mcs, 16$	-	9.5	[-]
$mcs, 25$	-	17	[-]

Although the HV's PC is modelled separately than other server's components, its PC considerations are embedded within the model and not jointly added to the

total PC value. Once the HV shares host server's units, these chips however must not exceed their maximum power. Accordingly, if the HV PC is separately added to the server's RAM, CPU, etc., this will overtake the real assumptions and the physical realisation regarding the PC of host server's components. This conduct is already considered while modelling each component's PC. Nevertheless, the HV can be considered as one of the VMs that contributes to the PC calculations and inherently included within N. Rather, the effect of the HV up on the execution time a VM consumes to process 100 RBs is shown in Fig. 5.7. The execution time of a the HV ($\tau_{HV} = 500\mu sec$), which is responsible for 5 times more delay to process the same amount of RBs when the server holds 1 VM, is added to the bare metal server delay $\tau_{bare} = 100\mu sec$. After, the resulting τ_{HV} is multiplied by the number of VMs in the server to obtain the execution time of the virtualised server (τ_v) with 10 VMs, as shown in Fig.5.7. Furthermore, to extend the delay performance presented in Fig. 5.7 to the whole system, Fig. 5.8 compares the total execution time (τ) of three cases regarding the number of bare metal and virtualised servers. These are 5, 10 and 20 virtualised servers, each with 5 VMs, which means there will be 25, 50 and 100 VMs respectively; compared to the same number of bare metal counterparts, all are subjected to ($MCS = 9$). This case increases the delay to about 77, 79 and 80 % accordingly.

In terms of accuracy, the model mainly relies on the manufacturer specifications and design of each component (i.e., components' data sheets). As each equipment holds different operating conditions, such as initial and maximum PC, cooling requirement and efficiency, the outcomes of the model will be accordingly affected. On account of such variation is required to be adjusted through the tuning factors mentioned such as ϑ, β,γ, etc., the power tolerance of each component can be precisely found and added to the total PC of virtual C-RAN. In is worth noticing that if evaluating systems with different number of VMs and RBs, the tuning parameters can be easily adjusted so as the maximum consumption of the each server's component is reached. Subsequently, total PC (P_{vCRAN}) can be predicted. At present, the model
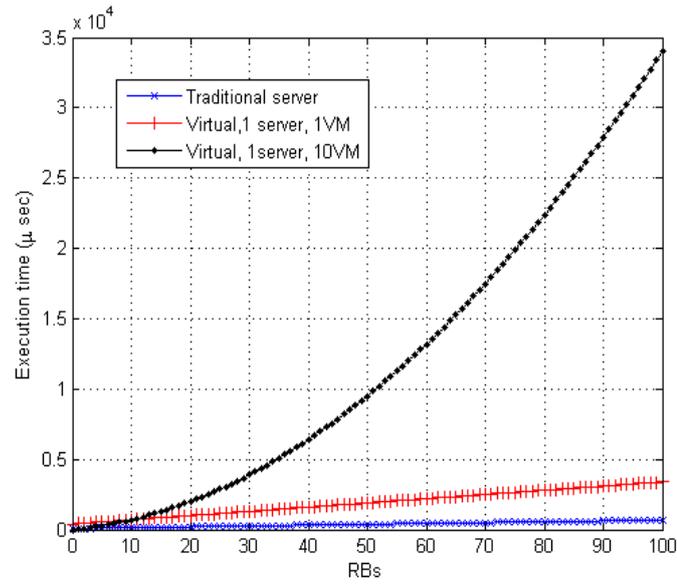
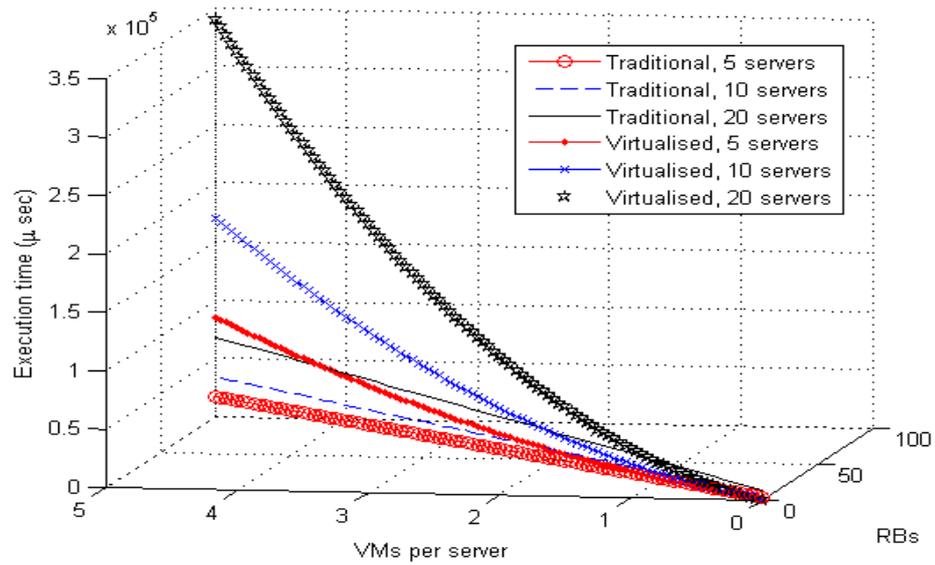Fig. 5.7. VMs and RB effect up on the execution time of a single server

Fig. 5.8. VMs and RB effect up on the execution time of the entire system

is reliable to judge the virtualised network as it relies upon initials and assumptions

come from real data measurements and experimental background as found in [62] and [105].

5.7 Summary

In large-scale computing centres, the advancement of knowledge in regards to the predicted PC and concerns of host servers that run VMs could improve the capacity planning and networks' EE. A parameters based PC model has been presented to demonstrate the PC calculations of virtualised C-RAN architecture. By using the proposed PC model, the power and execution time cost of NFV can be assessed. The model also enables the network providers to distinguish the PC reduction of the entire network while bearing the predicaments of NFV. The cost of increasing the VMs in a virtualised server and processed RBs of each VM is presented by providing a comparison for cooling, execution time performance and total PC with a non virtualised counterparts. The model is adaptable to any varying values related to any server resource, as the factors (γ, β, etc.) and initial PC values used in the model are changeable to describe the increasing/decreasing in PC of each resource found in the server. These latter are subjected to varying manufacturing specification of each electronic chip/device. Intuitively, NFV dramatically reduces network PC. At the same time, it degrades both the EE and execution time efficiency of the virtualised servers while preforming network's functions. To justify such issues, there have been recently several works proposed to optimise the problems of the HVs' scheduling procedure. These problems are related to synchronization, real-time constraints, security, VMs placement and performance enhancement. However, these researches are proposed to unleash and extend the HV capability to an optimum and enhance the lack of performance in a vitrualised server/network in real time services, such as in [131], [132] and [133]. The investigations also promise new futuristic techniques, protocols, algorithms and designs to be innovated. However, the benefits and characteristics gained by reducing the operational cost and total PC of the network compensate such performance loss in case of advanced techniques are used to mitigate the HV's constraints. Therefore, this work advocates the use of virtualisation in the coming generations as a way to greatly reduce the PC.

Chapter 6: Power Consumption of VMs' Live Migration

6.1 Introduction

Live migration is a agile concept in nowadays data centres [134]. In virtualisation environment, the interruption free live migrating of the VMs form one host server to another is an important issue to sustain running services to the UEs while gaining many benefits. Simply, live migrating the VM is the movement of one or many VMs for the original host server to another server, this is done when the VM is still running even after it resides in its target server. It is called 'Live' since the VM stayed running during the process of migration [135], [136].

Live migration comprises copying memory data on which the VM resides and CPU contents. Practically, an image file is stored in what is called network attached storage (NAS) rather than the local disk. NAS is accessible by all VMs and operates as HDD drive [137]. This means physically transferring the local disk is not required. More information about the process of copying a VM can be found in [134]. However, this technique has privileged the date centres with many advantages, as mentioned below:

- Maintenance: this technique represents a solution in case of the source server is required to be decommissioned due to its type promotion. Alongside, urgent required operating system or hardware maintenance.

- Reachability: the VM is usually reside on a host server which is physically located in a certain area. This VM might be serving UEs who are located far away from it. Alternatively, there might be host servers located closer to the UEs. Therefore, such migration will definitely reduce the link delay, channel losses, and help improving system administration.

- Load balancing: there might be servers experiencing heavy load due to their position in a dense area, or because of the service type they run. In this case, it is beneficial to distribute the load amongst other servers in the network via migrating the VMs. This is while proportionally considering their processing usage without degrading the performance of the participants [138].

- Off loading: when the traffic in the network is low, some servers can be selected to switch off so as the network EE is increased. In this case, live migrating the VMs from the chosen servers to another active ones is the solution.

On the other hand, the performance of migration process depends on many factors, such as the memory allocated to the VM, the size of work load it serves, the transmission rate at which the migration is occurring. Eventually, these factors affect the latency of migration and network traffic flow [89]. However, there are rising disadvantages symbolised by two major aspects:

- Time: the time it takes the process of migration degrades the network performance. The transferred VMs increase latency factor as it means more imposed link delay [139], this delay means degraded coverage and lower network capacity. In real time services this factor is essential and crucial, in contrast to the off-line services where the latency is relieved.

- Energy: the overhead cost of live migration is considerable. Up to 10 W is withdrawn from the destination, and this value increases when the server is source. This is on the basis of more computations within the tagged server will be performed in one unit of time [90].

In this chapter, we will try to model these costs which are synchronised with this technology, aiming to speculate the cost of such process prior to migration. The structure of this chapter is as follows: in Section 6.2, the process of live migration is modelled. In Section 6.3, the results are presented. Finally, the summery is given in Section 6.4.

6.2 Live Migration Power Model

Since the PM presented in Chapter 5 is a virtualisation based power model, it has been used to complete the live migration consumption model in this Chapter. Further to evaluating the virtualised server consumption in terms of increasing/decreasing the number of the hosted VMs, the increased PC cost due to migrating a single or group of VMs to the destination server is also counted. It has been noted that the power cost of the source server (P_{cost}^{source}) is changeable according to the utilisation of the CPU. We have translated this practical value to more understandable data to avoid the need to measurements' server. First, the extracted cost is re-drawn as a function of sever utilisation (utl) of [90] in stead of function of downtime (latency) of migration. Fig 6.1 shows the utilisation against the source server power cost.

The original data extracted is curve-fitted to a flexible and simple quadratic equation (6.2), with coefficients: cof1 = 0.0011189, cof2 = -0.25916, cof3 = 16.315.

$$P_{cost}^{source} = cof1 * utl^2 + cof2 * utl + cof3 \tag{6.1}$$

However, this description is valid for a server with specific characteristics, there might be slight change in this curve when the type of server is changed. To cover this issue, (6.2) is generalised by adding a constant called ($scof$), which is a real number. The latter will scale up and down the output of Fig. 6.1, so as it matches all possible costs. Therefore, the model in (6.2) has been updated, as follows:
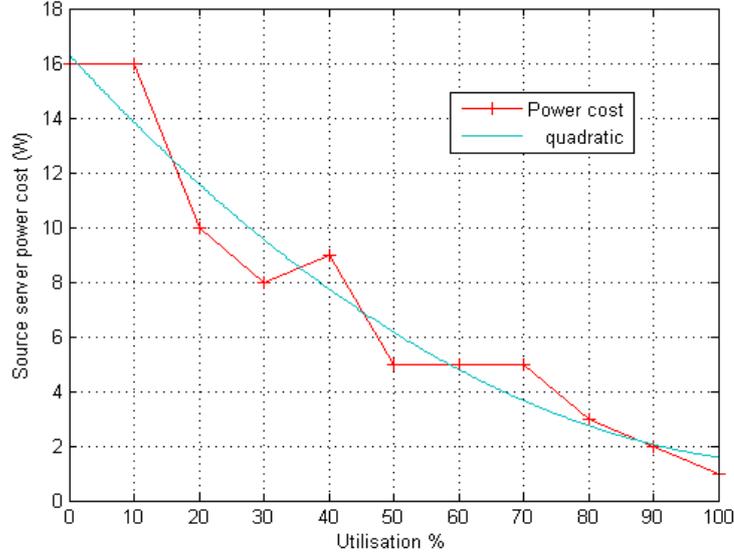
Fig. 6.1. Utilisation against the source server power cost.

$$P_{cost}^{source} = (cof1 * utl^2 + cof2 * utl + cof3) + scof \qquad (6.2)$$

There has been an objection with measuring the PC using utilisation ratio of a server, the reasons have been mentioned in Chapter 2. Briefly, this method does not offer simplicity for many reasons: it requires real experiment or measurement, it is expensive because it requires to physically be available at the data center to observe the utilisation values. On top of that, the SP propriety devices are not available to be tested. Therefore, this value is considered, but as a function of maximum and minimum PC of the virtualised server. This means the real time measurement power costs which are measured as a function of utilisation ratio are now transferred to a function of PC. To do so, we have used the formula in (6.3) to convert the utilisation of a server to a PC. In this case, the power cost of migration can be known directly from the server PC. For example, Fig. 6.2 shows the utilisation and PC conversions of a server. :

$$utl = 100 * (P_{server}^{BBU} - min(P_{server}^{BBU}))/(max(P_{server}^{BBU}) - min(P_{server}^{BBU})) \qquad (6.3)$$
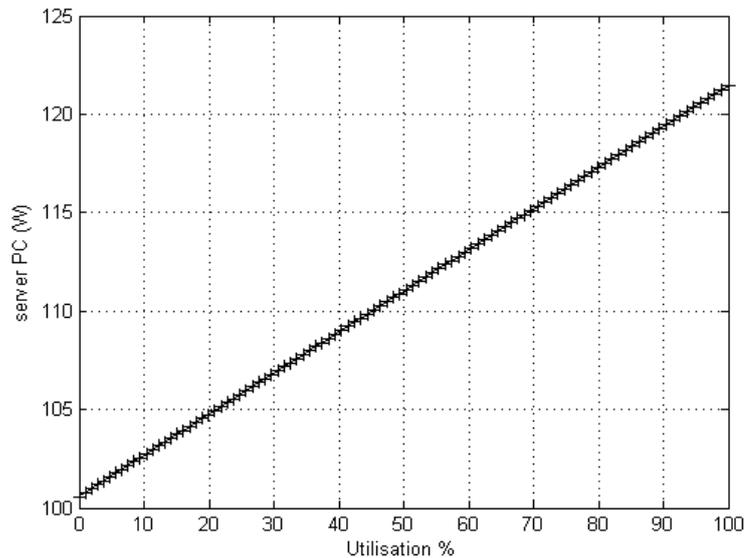
Where P_{server}^{BBU} is the BBU server.

Fig. 6.2. Utilisation ratio and its PC equivalent.

The 100% utilisation means the server is fully utilised and experience maximum PC ($max(P_{server}^{BBU})$), while 0 % utilisation means the server is load free ($min(P_{server}^{BBU})$) or in idle mode of operation. The resulting PC and its equivalent power cost is shown in Fig. 6.3.

In the receiving side, the power cost in the destination server (P_{cost}^{dest}) during the migration is experimentally measured in [90], which was reported as 10 W, regardless the percentage of utilisation in the server. This assumption is also generalised by adding another constant ($rcof$), which is a real number, so as all possibilities of servers' specifications are considered, as follows:

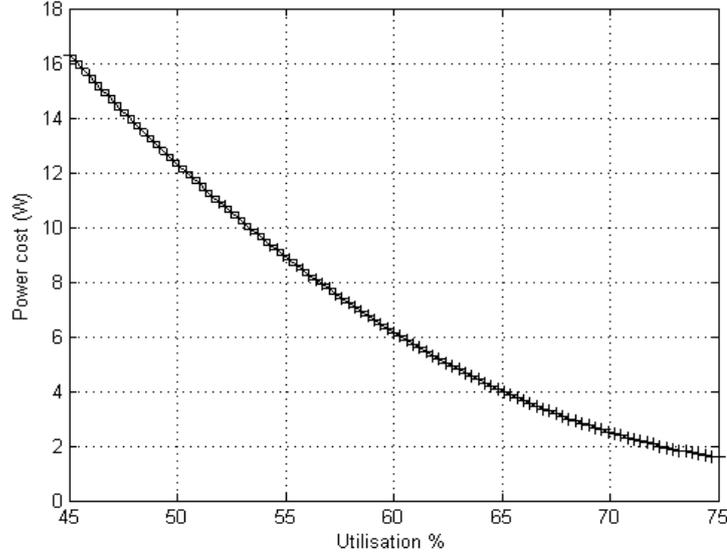$$P_{cost}^{dest} = 10 + rcof \qquad (6.4)$$

Fig. 6.3. Power cost of live migration and its PC equivalent.

If the total number of source servers undergoing live migration is (S_{smig}), the total number of target servers is (S_{rmig}), the number of migrated VMs is (N_{smig}), the total number of received VMs is (N_{rmig}). The overall PC of live migration based virtualised C-RAN data centre (P^{mig}_{vCRAN}) is formulated by adding these costs to the total PC of (5.26), which yields:

$$P^{mig}_{vCRAN} = P_{vCRAN} + \sum_{s_{smig}}^{S_{smig}} \sum_{n_{smig}}^{N_{smig}} (P^{source}_{cost})_{s_{smig},n_{smig}} + \sum_{s_{rmig}}^{S_{rmig}} \sum_{n_{rmig}}^{N_{rmig}} (P^{dest}_{cost})_{s_{rmig},n_{rmig}} \quad (6.5)$$

In terms of time, the live migrated downtime is relatively large, this process can happen within orders of milliseconds [140], or even orders of seconds. In any case, the higher data rate in which the VM is migrated, the less latency [90], [89]. The latter represents the time it takes the VM to be transferred to other server seamlessly. As the UEs are still connected while migration, this value does not mean that it is an effective latency to be added to the virtualisation process latency. However, when migrating the VM, the time cost can be established from the channel in which the

VM is transferred, which might be wireless or wired [90]. This cost eventually can be added to the modelling of Chapter 5.

If the time cost due to migrating the VM is (τ_{mig}), which is equivalent to the ($\frac{d_{mig}}{v_{mig}}$), where d_{mig} and v_{mig} denote the distance and speed at which the VM is transferred from the source to target server. v_{mig} is based on the channel type, it can be as fast as the speed light if the channel is wireless, experiencing cable losses if the channel is Ethernet via coaxial cable, or can be experiencing a refractive index in case of using an optical fiber channel. However, the effect of τ_{mig} can be major if the VM is moved between distant centres. Therefore, a virtualised data center experiences a total delay (τ_T) due to virtualisation delay (τ_v) and migration delay τ_{mig}, where

$$\tau_T = \tau_v + \tau_{mig} \tag{6.6}$$

Approximately, the experimental results in [90] is similar to [89] in terms of power cost and migration time. The latter has measured the cost with respect to the transmission bit rate used to send the VM. While in the former, the power cost is associated with server utilisation. These two measurements are correlated as the highly utilised server means it is sending using higher bit rate, and vice versa. Fig. 6.4 shows the VM migration bit rate relation with the power cost and migration time.

The following Fig. 6.5 shows the differences between the outcomes of [89] and [90] in terms of bit rate and utilisation, respectively.

While Fig. 6.6 shows the difference between the two works in terms of PC.

However, this difference is originated from the characteristics of the two operated servers. Nevertheless, the behaviour of the outcomes is identical. To extend these models to a general formulation, another two data sets have been added for servers that are commercially available and consume from (130-150 W) and (75-93W) [141]. Together, there will be four groups of data sets. These latter produce Fig. 6.7, which is then poly-fitted using linear regression to produce the following model, if E is equal to the virtualised server PC (P_{server}^{BBU}), the power cost of source server (P_{cost}^{source}) can be given as:
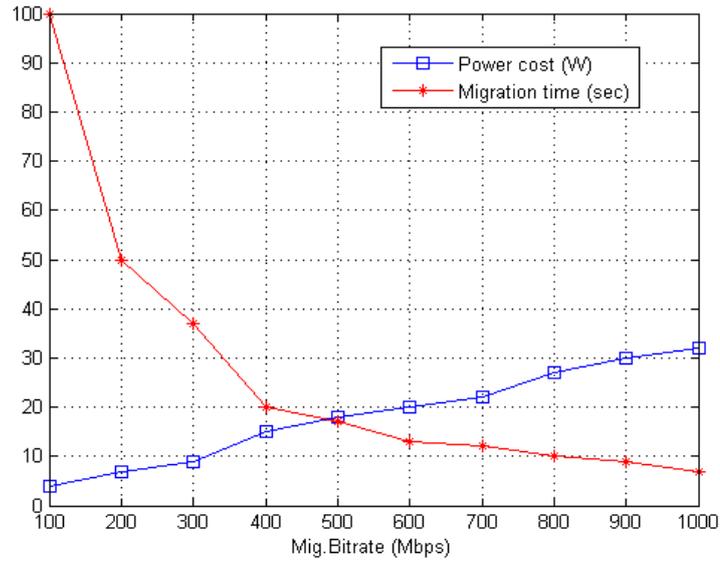
Fig. 6.4. Migration bit rate as a function of power cost and migration time of [89].
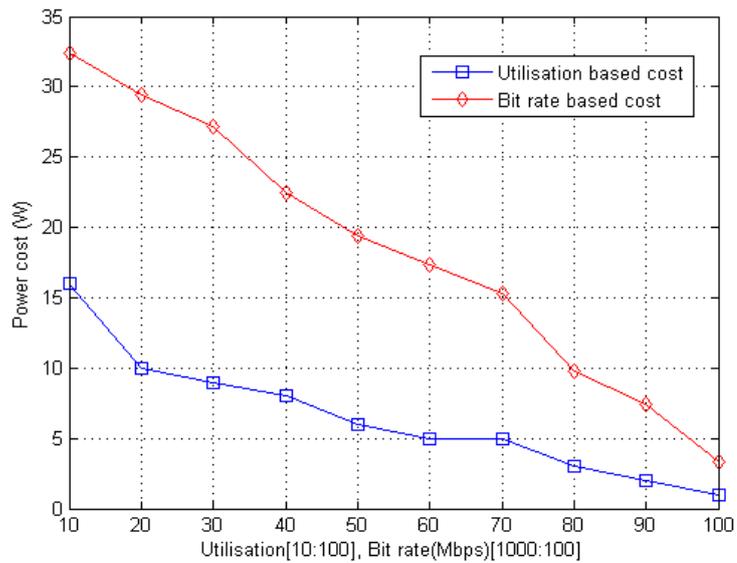

Fig. 6.5. Difference between the outcomes of two experiments in terms of utilisation and Bit rate.

$$P_{cost}^{source} = p1 \times E^{10} + p2 \times E^9 + p3 \times E^8 + p4 \times E^7 + p5 \times E^6 + p6 \times E^5$$
$$+ p7 \times E^4 + p8 \times E^3 + p9 \times E^2 + p10 \times E + p11$$
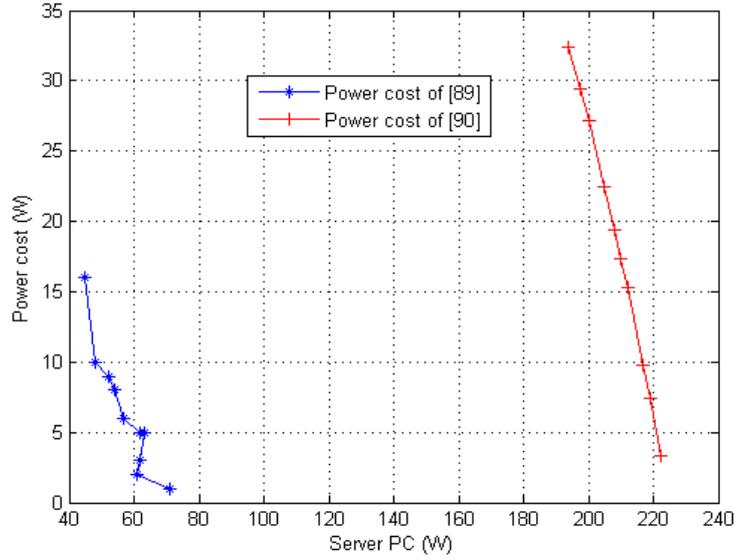
$$(6.7)$$

Fig. 6.6. Difference between the outcomes of two experiments in terms of PC.

The resulting coefficients are: $p1 = -8.6139e-17$, $p2 = 1.1482e-13$, $p3 = -6.6938e-11$, $p4 = 2.2424e-08$, $p5 = -4.7684e-06$, $p6 = 0.00067105$, $p7 = -0.063154$, $p8 = 3.9177$, $p9 = -153.09$, $p10 = 3399.7$, $p11 = -32561$. Subsequently, the value of P_{cost}^{source} is updated. It is worth noticing that the 10th order of (6.7) has been selected after repeating the fitting process several times to find the proper match using Matlab software. This process was necessary to find the most satisfaction and matching between the provided data and the fitting curve or output of (6.7), which yields no loss in the information of the original data set.

6.3   Results

Fig. 6.8 shows the power cost when a virtualised server holding 60 VMs is migrating 10 VMs, it also shows the power cost due to receiving the same number of VMs. Furthermore, it shows the total cost of both cases. This source cost was derived from the virtualised server consumption as mentioned above. At $N = 60$, the virtualised server PC is known. From this consumption, the power cost is obtained (about 4.9195
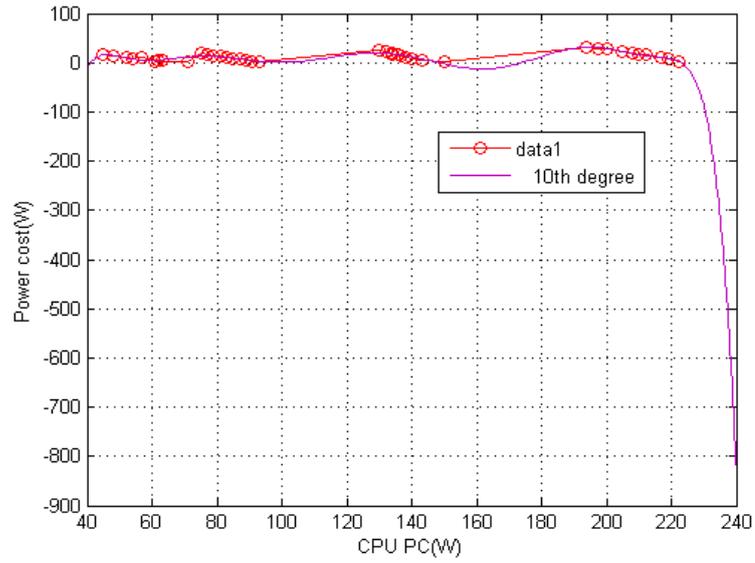
Fig. 6.7. Difference between the outcomes of two experiments.

W) and multiplied by the number of migrated VMs. On the other hand, 10 W times the number of received VMs is the power cost in the receiving side.
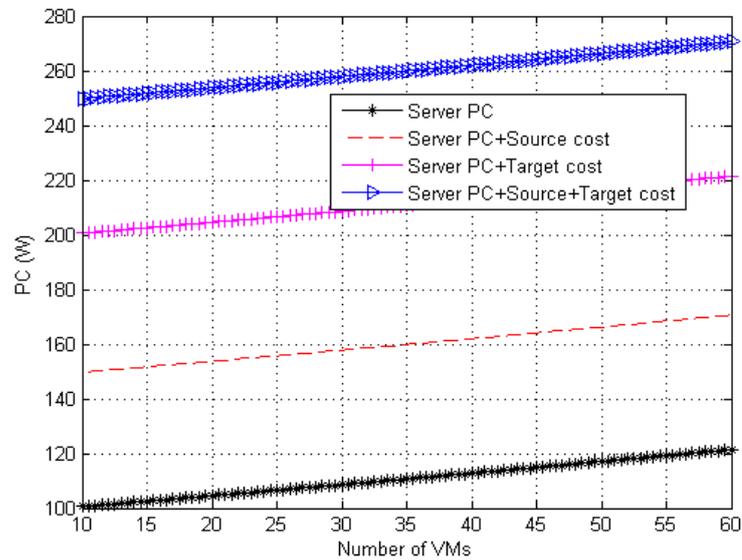

Fig. 6.8. PC of source server hosting 60 VMs, it shows the power cost of migrating 10 VMs.

Fig. 6.9 shows the increasing cost because of increasing the number of migrated VMs, up to 50 VMs. While the virtualised server already hosted 60 VMs.
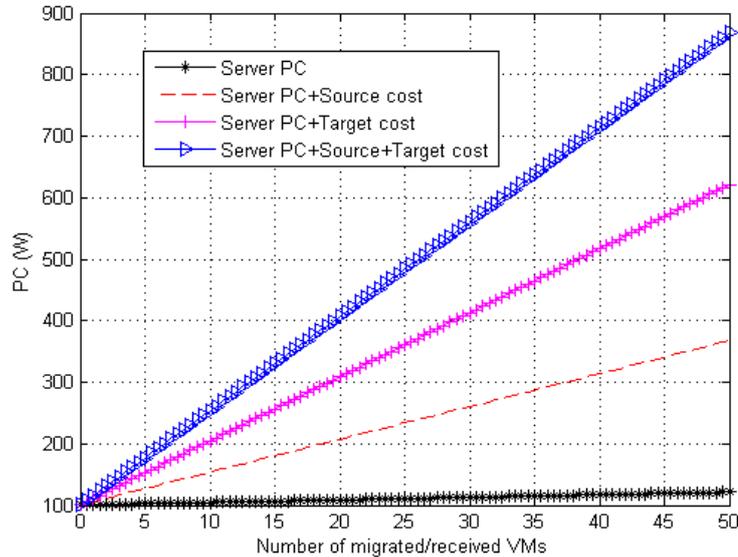

Fig. 6.9. PC of source server hosting 60 VMs, it shows the power cost of migrating 10 VMs.

As we can see from Fig. 6.8 and Fig. 6.9, migrating the VMs is not power effective process. To migrate and receive 10 VMs, the power cost is about 150 W, which is more than the original server PC. However, this cost is influenced by the period at which these VMs are transferred. Intuitively, the longer period these VMs are migrated, the less power cost and more efficient system. Therefore, algorithms/methods are needed to optimise at what time the VM is required to be moved. This can be based on several parameters, such as the number of UEs connected, load balancing requirement, position of the servers, etc. Fig. 6.9 exhibits the cost and total PC when the number of both migrated and received VMs is equal. However, sending and receiving a VM by more than one server at the same time represents another facility that can be added and offered by the model. This is when the number of migrated and received VMs is different. Fig. 6.10 then shows a 3 dimensional plot, x and y axes represent the number of migrated and received VMs, while z axis shows the PC.

This figure shows the cost of migrating or receiving 50 VMs, also it shows the cost of both cases. The virtualised server holds 60 VMs.
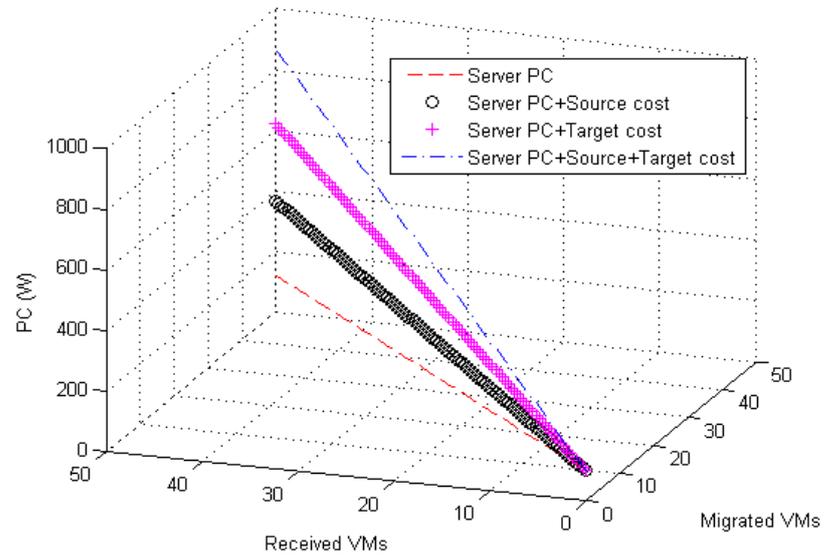
Fig. 6.10. PC of source server hosting 60 VMs, it shows the power cost of migrating 10 VMs.

Fig. 6.11 shows the channel delay experienced due to migrating a VM to a distant data center. This Figure compares wireless with lossy channel with refractive index = 1.4.

However, in both cases, wireless or delaying channel, this amount of delay is neglected compared to the virtualisation process delay. Fig. 6.12 shows the total delay of virtualisation and migration. This covers the wireless, up to an optical fiber channel with refractive index is equivalent to 1.4. This is when the virtualised server is hosting 10 VMs, each processing 200 RBs, and migrating 1 VM to variable distances (up to 100 Km).

Fig. 6.11. Delay comparison of wireless and delaying channel.
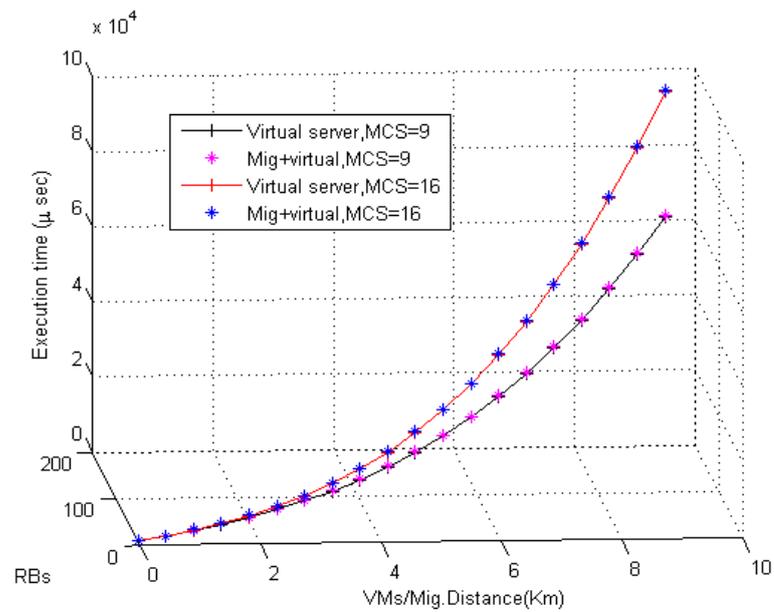
Fig. 6.12. Total delay of virtualised server while migrating a VM to different distances.

6.4   Summary

A power cost model has been presented to demonstrate the PC cost in a virtualised based data centres, specifically for live migration case. By using the proposed model,

the power and time cost can be assessed. Since there is no simple expression to describe such costs, this model has been proposed. It has converted the experimental results which are based on the server util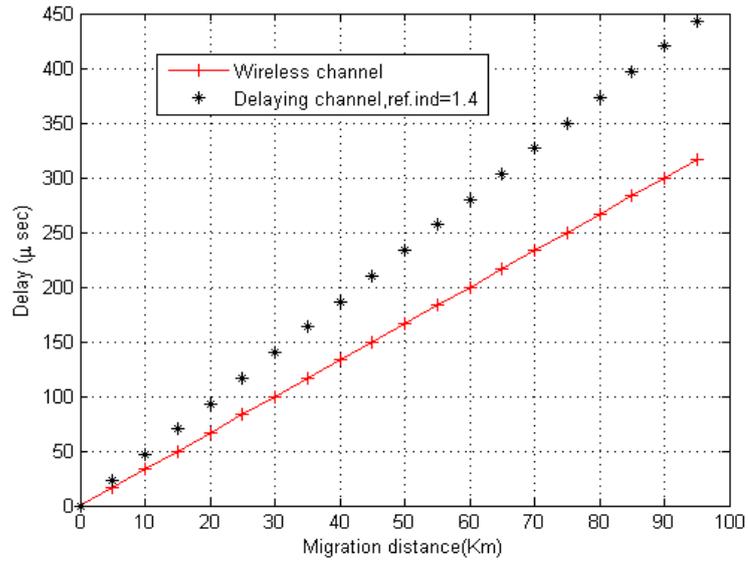isation and migration bit rate that is used to transmit the VM from one server to another, to a simplified mathematical formulation. The model enables the network providers to decide about the EE of such technology in the data centers and virtualised servers. For example the research conducted in [142] and [143] can benefit from the quick calculation of the proposed model to be implemented/adapted in their algorithms. The cost of repeated migration can cause a considerable amount of power lost. Nevertheless, by using a quick and simple calculation of the model, a decision about when and where the migration occurs can be easily made. On the other hand, the expected time cost due to migration has been shown. However, the time cost is negligible compared to the virtualisation. This time was the cost of the distance between the two cooperated servers, while the time of the migration process itself can reach more than 7 sec. This was not counted within this modelling since the VM is not terminated from serving the attached UEs and it is still live.

Chapter 7: Conclusions, Future Work and Recommendations

7.1 Conclusion

Four power models are proposed within this thesis, each manipulates different network design. The first paradigm is C-RAN. The model has offered a tool to measure the PC in an easy and accurate way. Based on the data sheets of the participating components and system bandwidth, the model can offer quick comprehensive PC calculation. This model also offers a test platform for the innovative PC reduction techniques within this field. Furthermore, the study showed the importance of C-RAN over the conventional architecture in terms of power and ability to adapt the futuristic algorithms and technologies, such as the evolution of SDN. The latter imposes enhanced administration for the network and network scalability when the signalling cost in the backhaul is increased. However, it adds considerable PC to the network through the foundation of the controller server and the deployed open flow switches. Accordingly, the second contribution has shed light up on this trade-off. The study showed the pros and cons of SDN and C-RAN integration in terms of power and the enhanced network characteristics brought by SDN. This increment in the PC due to SDN or any other innovative algorithms has triggered the speculation about how efficient can be such networks to realise the goals of 5G, especially the

consumed power or EE. Consequently, network virtualisation has played its role to overcome this consumption, thanks to the general off-the-shelf servers. Virtualisation dramatically reduced the PC of the data centres with a consequence of increased delay. The processed RBs/bandwidth has their inherent effect on the PC, the latter is driven up from an idle consumption due to the number of RBs processed within each VM. On the other hand, the model has tackled the imposed latency, this factor can heavily degrades the QoS of 5G UEs. Nevertheless, the literature showed major effort to relief such matter through optimised techniques and enhanced scheduling algorithms, these are shown within this study. However, the proposed model can judge how efficient are these methods. As the proposed PM within this study is based on simplified assumptions, it offered reduced complexity in comparison to the other models. In addition to the delay, another cost the network can endure if the facility of live migrating the VMs amongst servers is activated. This technique can fetch the network with enhanced data rates. However, the proposed model showed that the power cost is relatively high, up to $25W$ in the source server for each migrated VM. This is multiplied by the total number of VMs migrated in the data centre, which results in considerable amount of lost energy. In the source server, the power cost is linearly proportional with the server's PC. The proposed model has converted the experimental methods found in the literature to a simplified expression. The resulting is mathematical based model that can be easily adapted by any other algorithms, especially those carried out to advise about is it worth to migrate a VM or not, while concerning the gain, cost, and UEs' QoS compromisation.

7.2 Future Work

The core development tool while improving the communication and mobile systems is where the energy can be used with least cost. Accordingly, the concepts of green communications and sustainable energy has emerged as a complement to the idea of cloudification. There are some core comparisons can be raised amongst

the green sources of energy, such as solar cells and wind turbine with the traditional method. These comparisons include cost of deployment, amount of generated/required electricity, cost of maintenance, etc. This is followed by the average revenue brought by using both methods. Generally, the traditional method bear only the amount of electricity bill, paid each month. On the hand, the green sources are attached to maintenance issues. These include parts' replacement and regular equipment check. Furthermore, the effect of OPEX has to be added to the total cost of green energy. In the latter, the type of such source of energy should be chosen so as no high possibility of variation/reduction in the provided energy can happen. For example, choosing the solar cells based source in an area that is not highly exposed to the sun-light is critical. Such impact should be evaluated before deployment if the network is totally operated based on green energy, likewise other type of sources. Within the green area, another comparison can be included amongst these sources while considering the pros and cones of each type.

Cloudification alongside virtualisation, are a must to enhance the EE in the data centers. To achieve the required EE without compromising the performance, the matter of optimising the number of VMs that one server can hold is a vital to sustain the server from being overloaded while fulfilling the QoS requirements for the UEs. The idea behind this optimisation is to predict, according to the traffic demands, the number of VMs. By using Monte-Carlo simulation, the whole possibilities of resources assignments and power allocations can be observed in the area of interest. Subsequently, formulating the EE of the network by calculating the bit rate and the power consumed. The model in Chapter 5 can be used to seek for the missing number of VMs (N). The constraints of this problem can be the UE's bit rate, SINR or total UEs' sum rate. But most importantly, the constraints of time, as in virtualised servers the latency is enlarged. On the other hand, the latency of 5G has to be 5 times less than 4G networks. Based on these constraints, the solution of this optimisation can be obtained by using one of the heuristic algorithms, such as genetic algorithm or particle swarm optimisation.

Further, the Monte-Carlo simulation will convert the problem to a non time sensitive/based case. This can cancel the need to repeat the algorithm after a time, as Monte-Carlo considers huge number of possibilities or snapshots for the UEs and RRHs power allocations, resources assignment and locations.

Another important matter is a decision making algorithm can be set for the case of migrate/non migrate the VM in the data center from one server to another. This algorithm may rely on the calculation offered by Chapter 6, to easily and promptly decide on the gain and trade-off of migrating, rather than relying on time consuming and complex methods. Specifically, the algorithm compares the network EE gain with the non-migration decision. For example, when the migration is not enforced due to maintenance, is it efficient to migrate the VMs to switch off the server and save energy?. Another example is when the VM is migrated to a server which is closer to the UEs, this migration will decrease the latency. However, the migration energy cost might be a reason for enlarged PC within the network. Therefore, a comparison is required to evaluate the pros and cons of migration process. The position of the VM can affect the RRHs and their received powers, this will affect the UEs' power through reducing their path losses. Eventually it will affect the UEs data rates and then the EE. However, the algorithm can weight both the delay and EE, then the decision can be automated.

Finally, the placement of the virtualised BBU pool in the C-RANs based networks, is a crucial issue. The optimal position of the BBU pool can offer enhanced spectral efficiency, received data rates and QoS for the UEs. As the UEs resources assignments entities and set-ups functions are placed far away from the UEs in the pool, the problem of placing the BBU pool becomes a prerequisite to ensure that these functions' signals reach the RRHs without any extra delay. This eternally influences the network performance regarding the EE. Such optimization could also improve the transmission power to the RRHs to provide a higher coverage with less energy consumed. This in turn leads to an efficient network that is an important property of the future 5G networks.

7.3   Recommendations

Energy waste has been a major problem in today communication systems. The reasons of energy waste is the lack of efficient protocols and techniques to utilise all available network resources, mainly the physical hardware. An idle or non fully utilised device is a reason for energy waste. Another cause of energy waste is the inflexibility to apply the techniques that aimed to make the device fully utilised. For example, in distributed networks, the hardware is restricted in place and time. Therefore, it is difficult to hibernate this hardware even though few users are connected to the network. In contrast, the cloud networks are able to exceed this limitation with the possibility to off-load the UEs amongst servers to save energy. As such, virtualisation is another key stone to save more energy within the active servers in the cloud, as a single active server can operate on behalf of tens of servers. This can immediately reduce the CO_2 emissions which grow rapidly, and mitigate the risk of climate change.

Specifically, the presented work in this thesis is aimed to prove its impact up on the information and communication technology (ICT) section. ICT electricity use is about 1200TW in 2017, up to 1700TW expected in 2030. Parallel to such consumptions, globally, ICT section is responsible for 2% of the CO_2 emission, while mobile communication networks are responsible for 0.2% in 2007, up to 0.4% in 2020. C-RAN with its 30% power reduction is expected to reduce the latter by about 0.12 (30%), which yields 0.28% rather than 0.4% in 2020. However, this saving in the emission is reduced about 75% when the virtualisation technology is adapted. This reduced the emission from 0.4% to 0.1%. Furthermore, the average electricity demand of a base station is expected to be 1.2KW in 2020. C-RAN can reduce this amount by about 0.36KW, while virtualisation decreases such consumption by about 0.9KW. The power models proposed in this thesis hold the prove of efficiency and reliability for C-RAN and virtualised networks' to achieve such gain. According to such impact, this work recommends the use of network cloudifications and virtualisation as an efficient

way to reduce the PC in the next generations while considering the trade-offs of such techniques.

LIST OF REFERENCES

LIST OF REFERENCES

[1] T. Zhou, Y. Huang, L. Fan, and L. Yang, "Load-aware user association with quality of service support in heterogeneous cellular networks," *Communications, IET*, vol. 9, no. 4, pp. 494–500, 2015.

[2] S. Yang and L. Hanzo, "Fifty years of mimo detection: The road to large-scale mimos," *IEEE Communications Surveys Tutorials*, vol. 17, pp. 1941–1988, Fourthquarter 2015.

[3] P. Cerwall, P. Jonsson, R. Möller, S. Bävertoft, S. Carson, I. Godor, P. Kersch, A. Kälvemark, G. Lemne, and P. Lindberg, "Ericsson mobility report," *Ericsson.[Online]. Available.: www. ericsson. com/r es/docs/2015/ericsson-mobility-report-june-2015. pdf [June 2015]*, 2015.

[4] A. Lucent, "The lte network architecture—a comprehensive tutorial," *Strategic Whitepaper*, 2009.

[5] V. H. Muntean and M. Otesteanu, "Wimax versus lte - an overview of technical aspects for next generation networks technologies," in *Electronics and Telecommunications (ISETC), 2010 9th International Symposium on*, pp. 225–228, Nov 2010.

[6] A. Ghosh, R. Ratasuk, B. Mondal, N. Mangalvedhe, and T. Thomas, "Lte-advanced: next-generation wireless broadband technology [invited paper]," *IEEE Wireless Communications*, vol. 17, pp. 10–22, June 2010.

[7] G. Yu, Q. Chen, R. Yin, H. Zhang, and G. Y. Li, "Joint downlink and uplink resource allocation for energy-efficient carrier aggregation," *IEEE Transactions on Wireless Communications*, vol. 14, no. 6, pp. 3207–3218, 2015.

[8] X. Zhang and W. Wang, "Carrier aggregation for lte-advanced mobile communication systems," *IEEE Communications Magazine*, p. 89, 2010.

[9] T.-T. Tran, Y. Shin, and O.-S. Shin, "Overview of enabling technologies for 3gpp lte-advanced," *EURASIP Journal on Wireless Communications and Networking*, vol. 2012, no. 1, p. 1, 2012.

[10] Y. H. Chiang and W. Liao, "Green multicell cooperation in heterogeneous networks with hybrid energy sources," *IEEE Transactions on Wireless Communications*, vol. 15, pp. 7911–7925, Dec 2016.

[11] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H. P. Mayer, L. Thiele, and V. Jungnickel, "Coordinated multipoint: Concepts, performance, and field trial results," *IEEE Communications Magazine*, vol. 49, pp. 102–111, February 2011.

[12] C. Yang, Y. Yao, Z. Chen, and B. Xia, "Analysis on cache-enabled wireless heterogeneous networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 1, pp. 131–145, 2016.

[13] S. Deb, P. Monogioudis, J. Miernik, and J. P. Seymour, "Algorithms for enhanced inter-cell interference coordination (eicic) in lte hetnets," *IEEE/ACM Transactions on Networking*, vol. 22, pp. 137–150, Feb 2014.

[14] V. Sciancalepore, V. Mancuso, and A. Banchs, "Basics: Scheduling base stations to mitigate interferences in cellular networks," in *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2013 IEEE 14th International Symposium*, pp. 1–9, June 2013.

[15] 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, March 2015.

[16] 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.

[17] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. T. Sukhavasi, C. Patel, and S. Geirhofer, "Network densification: the dominant theme for wireless evolution into 5g," *IEEE Communications Magazine*, vol. 52, pp. 82–89, February 2014.

[18] M. Moltafet, P. Azmi, and N. Mokari, "Power minimization in 5g heterogeneous cellular networks," in *2016 24th Iranian Conference on Electrical Engineering (ICEE)*, pp. 234–238, May 2016.

[19] A. Gupta and R. K. Jha, "A survey of 5g network: Architecture and emerging technologies," *IEEE Access*, vol. 3, pp. 1206–1232, 2015.

[20] M. Liyanage, A. B. Abro, M. Ylianttila, and A. Gurtov, "Opportunities and challenges of software-defined mobile networks in network security," *IEEE Security Privacy*, vol. 14, pp. 34–44, July 2016.

[21] N. Xiong, W. Han, and A. Vandenberg, "Green cloud computing schemes based on networks: a survey," *IET Communications*, vol. 6, pp. 3294–3300, Dec 2012.

[22] 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, April 2016.

[23] 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.

[24] C. Liu, L. Zhang, M. Zhu, J. Wang, L. Cheng, and G. K. Chang, "A novel multi-service small-cell cloud radio access network for mobile backhaul and computing based on radio-over-fiber technologies," *Journal of Lightwave Technology*, vol. 31, pp. 2869–2875, Sept 2013.

[25] 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.

[26] Y. Liao, L. Song, Y. LI, and Y. Zhang, "How much computing capability is enough to run a cloud radio access network?," *IEEE Communications Letters*, vol. PP, no. 99, pp. 1–1, 2016.

[27] 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, December 2014.

[28] U. C. Kozat, G. Liang, K. Kökten, and J. Tapolcai, "On optimal topology verification and failure localization for software defined networks," *IEEE/ACM Transactions on Networking*, vol. 24, pp. 2899–2912, Oct 2016.

[29] D. S. Linthicum, "Software-defined networks meet cloud computing," *IEEE Cloud Computing*, vol. 3, pp. 8–10, May 2016.

[30] H. Kim and N. Feamster, "Improving network management with software defined networking," *IEEE Communications Magazine*, vol. 51, pp. 114–119, February 2013.

[31] B. Kar, E. H. K. Wu, and Y. D. Lin, "The budgeted maximum coverage problem in partially deployed software defined networks," *IEEE Transactions on Network and Service Management*, vol. 13, pp. 394–406, Sept 2016.

[32] R. Mijumbi, J. Serrat, J. L. Gorricho, N. Bouten, F. D. Turck, and R. Boutaba, "Network function virtualization: State-of-the-art and research challenges," *IEEE Communications Surveys Tutorials*, vol. 18, pp. 236–262, Firstquarter 2016.

[33] K. Katsalis, N. Nikaein, E. Schiller, R. Favraud, and T. I. Braun, "5g architectural design patterns," in *Communications Workshops (ICC), 2016*, pp. 32–37, IEEE, 2016.

[34] P. Demestichas, A. Georgakopoulos, D. Karvounas, K. Tsagkaris, V. Stavroulaki, J. Lu, C. Xiong, and J. Yao, "5g on the horizon: key challenges for the radio-access network," *IEEE Vehicular Technology Magazine*, vol. 8, no. 3, pp. 47–53, 2013.

[35] X. Costa-Pérez, J. Swetina, T. Guo, R. Mahindra, and S. Rangarajan, "Radio access network virtualization for future mobile carrier networks," *IEEE Communications Magazine*, vol. 51, no. 7, pp. 27–35, 2013.

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

[37] J. Zhang, F. Ren, R. Shu, T. Huang, and Y. Liu, "Guaranteeing delay of live virtual machine migration by determining and provisioning appropriate bandwidth," *IEEE Transactions on Computers*, vol. 65, pp. 2910–2917, Sept 2016.

[38] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5g be?," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, 2014.

[39] J. Yu, Z. Jia, T. Wang, and G. K. Chang, "Centralized lightwave radio-over-fiber system with photonic frequency quadrupling for high-frequency millimeter-wave generation," *IEEE Photonics Technology Letters*, vol. 19, pp. 1499–1501, Oct 2007.

[40] J. Yu, Z. Jia, T. Wang, and G. K. Chang, "A novel radio-over-fiber configuration using optical phase modulator to generate an optical mm-wave and centralized lightwave for uplink connection," *IEEE Photonics Technology Letters*, vol. 19, pp. 140–142, Feb 2007.

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

[42] M. Dayarathna, Y. Wen, and R. Fan, "Data center energy consumption modeling: A survey," *IEEE Communications Surveys Tutorials*, vol. 18, pp. 732–794, Firstquarter 2016.

[43] D. Musiige, L. Vincent, F. Anton, and D. Mioc, "Lte rf subsystem power consumption modeling," in *The 1st IEEE Global Conference on Consumer Electronics 2012*, pp. 645–649, Oct 2012.

[44] K. T. Malladi, B. C. Lee, F. A. Nothaft, C. Kozyrakis, K. Periyathambi, and M. Horowitz, "Towards energy-proportional datacenter memory with mobile dram," in *ACM SIGARCH Computer Architecture News*, vol. 40, pp. 37–48, IEEE Computer Society, 2012.

[45] E. R. Masanet, R. E. Brown, A. Shehabi, J. G. Koomey, and B. Nordman, "Estimating the energy use and efficiency potential of us data centers," *Proceedings of the IEEE*, vol. 99, no. 8, pp. 1440–1453, 2011.

[46] N. Balasubramanian, A. Balasubramanian, and A. Venkataramani, "Energy consumption in mobile phones: A measurement study and implications for network applications," in *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*, IMC '09, (New York, NY, USA), pp. 280–293, ACM, 2009.

[47] A. Chatzipapas, S. Alouf, and V. Mancuso, "On the minimization of power consumption in base stations using on/off power amplifiers," in *Online conference on green communications (GreenCom), 2011 IEEE*, pp. 18–23, IEEE, 2011.

[48] B. Wang, Y. Xu, R. Hasholzner, R. Rosales, M. Glaß, and J. Teich, "End-to-end power estimation for heterogeneous cellular lte socs in early design phases," in *2014 24th International Workshop on Power and Timing Modeling, Optimization and Simulation (PATMOS)*, pp. 1–8, Sept 2014.

[49] M. Deruyck, W. Vereecken, W. Joseph, B. Lannoo, M. Pickavet, and L. Martens, "Reducing the power consumption in wireless access networks: Overview and recommendations," *Progress In Electromagnetics Research*, vol. 132, pp. 255–274, 2012.

[50] K. Liu, J. He, J. Ding, Y. Zhu, and Z. Liu, "Base station power model and application for energy efficient lte," in *2013 15th IEEE International Conference on Communication Technology (ICCT)*, pp. 86–92, IEEE, 2013.

[51] C. Desset, B. Debaillie, V. Giannini, A. Fehske, G. Auer, H. Holtkamp, W. Wajda, D. Sabella, F. Richter, M. J. Gonzalez, *et al.*, "Flexible power modeling of lte base stations," in *Wireless Communications and Networking Conference (WCNC), 2012 IEEE*, pp. 2858–2862, IEEE, 2012.

[52] M. Imran and E. Katranaras, "Energy efficiency analysis of the reference systems, areas of improvements and target breakdown. ict-earth project, deliverable d2. 3, ec-ist office, brussels, belgium (january 2011)."

[53] H. Holtkamp, G. Auer, V. Giannini, and H. Haas, "A parameterized base station power model," *Communications Letters, IEEE*, vol. 17, no. 11, pp. 2033–2035, 2013.

[54] P. Grant, "Mcve core 5 programme, green radio-the case for more efficient cellular basestations," 2010.

[55] R. Esnault, "Optimising power efficiency in mobile radio networks (opera-net) project presentation," 2008.

[56] Y. Chen, S. Zhang, S. Xu, and G. Y. Li, "Fundamental trade-offs on green wireless networks," *IEEE Communications Magazine*, vol. 49, pp. 30–37, June 2011.

[57] M. Deruyck, W. Vereecken, E. Tanghe, W. Joseph, M. Pickavet, L. Martens, and P. Demeester, "Comparison of power consumption of mobile wimax, hspa and lte access networks," in *2010 9th Conference of Telecommunication, Media and Internet*, pp. 1–7, June 2010.

[58] J. Huang, F. Qian, A. Gerber, Z. M. Mao, S. Sen, and O. Spatscheck, "A close examination of performance and power characteristics of 4g lte networks," in *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*, MobiSys '12, (New York, NY, USA), pp. 225–238, ACM, 2012.

[59] Y. Zhang and N. Ansari, "Hero: Hierarchical energy optimization for data center networks," *IEEE Systems Journal*, vol. 9, no. 2, pp. 406–415, 2015.

[60] B. Heller, S. Seetharaman, P. Mahadevan, Y. Yiakoumis, P. Sharma, S. Banerjee, and N. McKeown, "Elastictree: Saving energy in data center networks.," in *Nsdi*, vol. 10, pp. 249–264, 2010.

[61] Y. Li and M. Chen, "Software-defined network function virtualization: A survey," *IEEE Access*, vol. 3, pp. 2542–2553, 2015.

[62] R. Shea, H. Wang, and J. Liu, "Power consumption of virtual machines with network transactions: Measurement and improvements," in *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*, pp. 1051–1059, April 2014.

[63] M. Marcu and D. Tudor, "Power consumption measurements of virtual machines," in *Applied Computational Intelligence and Informatics (SACI), 2011 6th IEEE International Symposium on*, pp. 445–449, May 2011.

[64] P. Bohrer, E. N. Elnozahy, T. Keller, M. Kistler, C. Lefurgy, C. McDowell, and R. Rajamony, "The case for power management in web servers," in *Power aware computing*, pp. 261–289, Springer, 2002.

[65] H. Li, G. Casale, and T. Ellahi, "Sla-driven planning and optimization of enterprise applications," in *Proceedings of the first joint WOSP/SIPEW international conference on Performance engineering*, pp. 117–128, ACM, 2010.

[66] Y. Tian, C. Lin, and K. Li, "Managing performance and power consumption tradeoff for multiple heterogeneous servers in cloud computing," *Cluster computing*, vol. 17, no. 3, pp. 943–955, 2014.

[67] S.-W. Ham, M.-H. Kim, B.-N. Choi, and J.-W. Jeong, "Simplified server model to simulate data center cooling energy consumption," *Energy and Buildings*, vol. 86, pp. 328–339, 2015.

[68] T. Horvath and K. Skadron, "Multi-mode energy management for multi-tier server clusters," in *Proceedings of the 17th international conference on Parallel architectures and compilation techniques*, pp. 270–279, ACM, 2008.

[69] V. Gupta, R. Nathuji, and K. Schwan, "An analysis of power reduction in datacenters using heterogeneous chip multiprocessors," *ACM SIGMETRICS Performance Evaluation Review*, vol. 39, no. 3, pp. 87–91, 2011.

[70] M. A. Islam, S. Ren, and G. Quan, "Online energy budgeting for virtualized data centers," in *2013 IEEE 21st International Symposium on Modeling, Analysis & Simulation of Computer and Telecommunication Systems (MASCOTS)*, pp. 424–433, IEEE, 2013.

[71] J. W. Smith, A. Khajeh-Hosseini, J. S. Ward, and I. Sommerville, "Cloudmonitor: profiling power usage," in *2012 IEEE 5th International Conference on Cloud Computing (CLOUD)*, pp. 947–948, IEEE, 2012.

[72] M. Gamell, I. Rodero, M. Parashar, and S. Poole, "Exploring energy and performance behaviors of data-intensive scientific workflows on systems with deep memory hierarchies," in *High Performance Computing (HiPC)*, pp. 226–235, IEEE, 2013.

[73] Q. Deng, D. Meisner, A. Bhattacharjee, T. F. Wenisch, and R. Bianchini, "Coscale: Coordinating cpu and memory system dvfs in server systems," in *Proceedings of the 2012 45th Annual IEEE/ACM International Symposium on Microarchitecture*, MICRO-45, (Washington, DC, USA), pp. 143–154, IEEE Computer Society, 2012.

[74] R. Joseph and M. Martonosi, "Run-time power estimation in high performance microprocessors," in *Proceedings of the 2001 international symposium on Low power electronics and design*, pp. 135–140, ACM, 2001.

[75] F. Bellosa, "The benefits of event: driven energy accounting in power-sensitive systems," in *Proceedings of the 9th workshop on ACM SIGOPS European workshop: beyond the PC: new challenges for the operating system*, pp. 37–42, ACM, 2000.

[76] C. Isci and M. Martonosi, "Runtime power monitoring in high-end processors: Methodology and empirical data," in *Proceedings of the 36th annual IEEE/ACM International Symposium on Microarchitecture*, p. 93, IEEE Computer Society, 2003.

[77] J. Chen, B. Li, Y. Zhang, L. Peng, and J.-k. Peir, "Tree structured analysis on gpu power study," in *2011 IEEE 29th International Conference on Computer Design (ICCD)*, pp. 57–64, IEEE, 2011.

[78] S. Wang, H. Chen, and W. Shi, "Span: A software power analyzer for multicore computer systems," *Sustainable Computing: Informatics and Systems*, vol. 1, no. 1, pp. 23–34, 2011.

[79] D. Economou, S. Rivoire, C. Kozyrakis, and P. Ranganathan, "Full-system power analysis and modeling for server environments," International Symposium on Computer Architecture-IEEE, 2006.

[80] G. Da Costa and H. Hlavacs, "Methodology of measurement for energy consumption of applications," in *Grid Computing (GRID), 2010 11th IEEE/ACM*, pp. 290–297, IEEE, 2010.

[81] R. Ge, X. Feng, S. Song, H.-C. Chang, D. Li, and K. W. Cameron, "Powerpack: Energy profiling and analysis of high-performance systems and applications," *IEEE Transactions on Parallel and Distributed Systems*, vol. 21, no. 5, pp. 658–671, 2010.

[82] X. Wu, H.-C. Chang, S. Moore, V. Taylor, C.-Y. Su, D. Terpstra, C. Lively, K. Cameron, and C. W. Lee, "Mummi: Multiple metrics modeling infrastructure for exploring performance and power modeling," in *Proceedings of the Conference on Extreme Science and Engineering Discovery Environment: Gateway to Discovery*, p. 36, ACM, 2013.

[83] J. L. Berral, R. Gavalda, and J. Torres, "Power-aware multi-data center management using machine learning," in *2013 42nd International Conference on Parallel Processing (ICPP)*, pp. 858–867, IEEE, 2013.

[84] J. L. Berral, Í. Goiri, R. Nou, F. Julià, J. Guitart, R. Gavaldà, and J. Torres, "Towards energy-aware scheduling in data centers using machine learning," in *Proceedings of the 1st International Conference on energy-Efficient Computing and Networking*, pp. 215–224, ACM, 2010.

[85] J. Dolado, D. Rodríguez, J. Riquelme, F. Ferrer-Troyano, and J. Cuadrado, "A two-stage zone regression method for global characterization of a project database," *Advances in Machine Learning Applications in Software Engineering*, vol. 1, 2007.

[86] F. Caglar and A. Gokhale, "ioverbook: intelligent resource-overbooking to support soft real-time applications in the cloud," in *2014 IEEE 7th international conference on Cloud computing (CLOUD)*, pp. 538–545, IEEE, 2014.

[87] G. Tesauro, R. Das, H. Chan, J. O. Kephart, D. Levine, F. L. Rawson III, and C. Lefurgy, "Managing power consumption and performance of computing systems using reinforcement learning.," in *NIPS*, vol. 7, pp. 1–8, 2007.

[88] J. L. Berral, R. Gavalda, and J. Torres, "Adaptive scheduling on power-aware managed data-centers using machine learning," in *2011 12th IEEE/ACM International Conference on Grid Computing (GRID)*, pp. 66–73, IEEE, 2011.

[89] H. Liu, H. Jin, C.-Z. Xu, and X. Liao, "Performance and energy modeling for live migration of virtual machines," *Cluster computing*, vol. 16, no. 2, pp. 249–264, 2013.

[90] Q. Huang, F. Gao, R. Wang, and Z. Qi, "Power consumption of virtual machine live migration in clouds," in *Third International Conference on Communications and Mobile Computing (CMC)*, pp. 122–125, April 2011.

[91] K. Rybina, W. Dargie, A. Strunk, and A. Schill, "Investigation into the energy cost of live migration of virtual machines," in *2013 Sustainable Internet and ICT for Sustainability (SustainIT)*, pp. 1–8, Oct 2013.

[92] A. Strunk, "A lightweight model for estimating energy cost of live migration of virtual machines," in *2013 IEEE Sixth International Conference on Cloud Computing*, pp. 510–517, June 2013.

[93] M. Fiorani, S. Tombaz, J. Martensson, B. Skubic, L. Wosinska, and P. Monti, "Modeling energy performance of c-ran with optical transport in 5g network scenarios," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 8, pp. B21–B34, Nov 2016.

[94] Z. Liu, Y. Chen, C. Bash, A. Wierman, D. Gmach, Z. Wang, M. Marwah, and C. Hyser, "Renewable and cooling aware workload management for sustainable data centers," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 40, pp. 175–186, ACM, 2012.

[95] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs, "Cutting the electric bill for internet-scale systems," in *ACM SIGCOMM computer communication review*, vol. 39, pp. 123–134, ACM, 2009.

[96] A. Gandhi, M. Harchol-Balter, and I. Adan, "Server farms with setup costs," *Performance Evaluation*, vol. 67, no. 11, pp. 1123–1138, 2010.

[97] M. Mazzucco, D. Dyachuk, and M. Dikaiakos, "Profit-aware server allocation for green internet services," in *Modeling, Analysis & Simulation of Computer and Telecommunication Systems (MASCOTS), 2010 IEEE International Symposium*, pp. 277–284, IEEE, 2010.

[98] I. Mitrani, "Trading power consumption against performance by reserving blocks of servers," in *European Workshop on Performance Engineering*, pp. 1–15, Springer, 2012.

[99] M. Mazzucco and D. Dyachuk, "Balancing electricity bill and performance in server farms with setup costs," *Future Generation Computer Systems*, vol. 28, no. 2, pp. 415–426, 2012.

[100] R. Lent, "Analysis of an energy proportional data center," *Ad Hoc Networks*, vol. 25, pp. 554–564, 2015.

[101] A. Aikebaier, Y. Yang, T. Enokido, and M. Takizawa, "Energy-efficient computation models for distributed systems," in *Network-Based Information Systems, 2009. NBIS'09*, pp. 424–431, IEEE, 2009.

[102] A. Aikebaier, T. Enokido, and M. Takizawa, "Distributed cluster architecture for increasing energy efficiency in cluster systems," in *Parallel Processing Workshops, 2009. ICPPW'09*, pp. 470–477, IEEE, 2009.

[103] T. Inoue, A. Aikebaier, T. Enokido, and M. Takizawa, "Evaluation of an energy-aware selection algorithm for computation and storage-based applications," in *15th International Conference on Network-Based Information Systems (NBiS)*, pp. 120–127, IEEE, 2012.

[104] M. G. Adian, H. Aghaeinia, and Y. Norouzi, "Spectrum sharing and power allocation in multi-input-multi-output multi-band underlay cognitive radio networks," *IET Communications*, vol. 7, pp. 1140–1150, July 2013.

[105] S. Bhaumik, S. P. Chandrabose, M. K. Jataprolu, G. Kumar, A. Muralidhar, P. Polakos, V. Srinivasan, and T. Woo, "Cloudiq: a framework for processing base stations in a data center," in *Proceedings of the 18th annual international conference on Mobile computing and networking*, pp. 125–136, ACM, 2012.

[106] M. Dirani and Z. Altman, "Self-organizing networks in next generation radio access networks: Application to fractional power control," *Computer Networks*, vol. 55, no. 2, pp. 431–438, 2011.

[107] R. Kumar and L. Mieritz, "Conceptualizing green it and data center power and cooling issues," *Gartner research paper*, no. G00150322, 2007.

[108] R. S. Tucker and K. Hinton, "Energy consumption and energy density in optical and electronic signal processing," *IEEE Photonics Journal*, vol. 3, pp. 821–833, Oct 2011.

[109] K. L. Lee, B. Sedighi, R. S. Tucker, H. Chow, and P. Vetter, "Energy efficiency of optical transceivers in fiber access networks [invited]," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 4, pp. A59–A68, Sept 2012.

[110] S. Namba, T. Warabino, and S. Kaneko, "Bbu-rrh switching schemes for centralized ran," in *Communications and Networking in China (CHINACOM), 2012 7th International Conference (ICST)*, pp. 762–766, Aug 2012.

[111] H. Holtkamp, "Enhancing the energy efficiency of radio base stations," *arXiv preprint arXiv:1311.3534*, 2013.

[112] F. Kaup, S. Melnikowitsch, and D. Hausheer, "Measuring and modeling the power consumption of openflow switches," in *10th International Conference on Network and Service Management (CNSM) and Workshop*, pp. 181–186, Nov 2014.

[113] M. Abolhasan, J. Lipman, W. Ni, and B. Hagelstein, "Software-defined wireless networking: centralized, distributed, or hybrid?," *IEEE Network*, vol. 29, pp. 32–38, July 2015.

[114] H. Wang, S. Chen, H. Xu, M. Ai, and Y. Shi, "Softnet: A software defined decentralized mobile network architecture toward 5g," *IEEE Network*, vol. 29, pp. 16–22, March 2015.

[115] M. Y. Arslan, K. Sundaresan, and S. Rangarajan, "Software-defined networking in cellular radio access networks: potential and challenges," *IEEE Communications Magazine*, vol. 53, pp. 150–156, January 2015.

[116] D. Wang, L. Zhang, Y. Qi, and A. U. Quddus, "Localized mobility management for sdn-integrated lte backhaul networks," in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, pp. 1–6, May 2015.

[117] D. Tuncer, M. Charalambides, S. Clayman, and G. Pavlou, "Adaptive resource management and control in software defined networks," *IEEE Transactions on Network and Service Management*, vol. 12, pp. 18–33, March 2015.

[118] 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, November 2015.

[119] X. Sun and N. Ansari, "Green cloudlet network: A distributed green mobile cloud network," *IEEE Network*, vol. 31, pp. 64–70, January 2017.

[120] T. Maksymyuk, M. Kyryk, and M. Jo, "Comprehensive spectrum management for heterogeneous networks in lte-u," *IEEE Wireless Communications*, vol. 23, pp. 8–15, December 2016.

[121] A. S. Thyagaturu, Y. Dashti, and M. Reisslein, "Sdn-based smart gateways (sm-gws) for multi-operator small cell network management," *IEEE Transactions on Network and Service Management*, vol. 13, pp. 740–753, Dec 2016.

[122] R. Riggio, A. Bradai, D. Harutyunyan, T. Rasheed, and T. Ahmed, "Scheduling wireless virtual networks functions," *IEEE Transactions on Network and Service Management*, vol. 13, pp. 240–252, June 2016.

[123] J. G. Herrera and J. F. Botero, "Resource allocation in nfv: A comprehensive survey," *IEEE Transactions on Network and Service Management*, vol. 13, pp. 518–532, Sept 2016.

[124] O. Rottenstreich, I. Keslassy, Y. Revah, and A. Kadosh, "Minimizing delay in network function virtualization with shared pipelines," *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, pp. 156–169, Jan 2017.

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

[126] R. Lent, "Evaluating the performance and power consumption of systems with virtual machines," in *Cloud Computing Technology and Science (CloudCom), 2011 IEEE Third International Conference on*, pp. 778–783, Nov 2011.

[127] L. Ye, C. Gniady, and J. H. Hartman, "Energy-efficient memory management in virtual machine environments," in *Green Computing Conference and Workshops (IGCC), 2011 International*, pp. 1–8, July 2011.

[128] M. Portolani and C. Elsen, "Network consolidation for virtualized servers," Sept. 27 2011. US Patent 8,027,354.

[129] L. Minas and B. Ellison, "The problem of power consumption in servers," *Intel Corporation. Dr. Dobb's*, 2009.

[130] R. S. Alhumaima, M. Khan, and H. S. Al-Raweshidy, "Component and parameterised power model for cloud radio access network," *IET Communications*, vol. 10, no. 7, pp. 745–752, 2016.

[131] S. Wu, L. Zhou, H. Sun, H. Jin, and X. Shi, "Poris: A scheduler for parallel soft real-time applications in virtualized environments," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, pp. 841–854, March 2016.

[132] C. Kim and K. H. Park, "Credit-based runtime placement of virtual machines on a single numa system for qos of data access performance," *IEEE Transactions on Computers*, vol. 64, pp. 1633–1646, June 2015.

[133] M. Zhang, Q. Yang, C. Wu, and M. Jiang, "Hierarchical virtual network mapping algorithm for large-scale network virtualisation," *IET Communications*, vol. 6, pp. 1969–1978, Sept 2012.

[134] A. Strunk and W. Dargie, "Does live migration of virtual machines cost energy?," in *2013 IEEE 27th International Conference on Advanced Information Networking and Applications (AINA)*, pp. 514–521, March 2013.

[135] J. Zhang, F. Ren, and C. Lin, "Delay guaranteed live migration of virtual machines," in *INFOCOM, 2014 Proceedings IEEE*, pp. 574–582, IEEE, 2014.

[136] J. Zheng, T. S. E. Ng, K. Sripanidkulchai, and Z. Liu, "Pacer: A progress management system for live virtual machine migration in cloud computing," *IEEE Transactions on Network and Service Management*, vol. 10, pp. 369–382, December 2013.

[137] Y. Ruan, Z. Cao, and Z. Cui, "Pre-filter-copy: Efficient and self-adaptive live migration of virtual machines," *IEEE Systems Journal*, vol. 10, pp. 1459–1469, Dec 2016.

[138] A. Agarwal and R. Shangruff, "Live migration of virtual machines in cloud," *International Journal of Scientific and Research Publications*, vol. 2, no. 6, pp. 1–5, 2012.

[139] A. Strunk, "Costs of virtual machine live migration: A survey," in *2012 IEEE Eighth World Congress on Services*, pp. 323–329, June 2012.

[140] C. Clark, K. Fraser, S. Hand, J. G. Hansen, E. Jul, C. Limpach, I. Pratt, and A. Warfield, "Live migration of virtual machines," in *Proceedings of the 2Nd Conference on Symposium on Networked Systems Design & Implementation - Volume 2*, NSDI'05, (Berkeley, CA, USA), pp. 273–286, USENIX Association, 2005.

[141] Buildcomputers, *Power Consumption of PC Components in Watts (http://www.buildcomputers.net/power-consumption-of-pc-components.html)*, 2017 (accessed March 19, 2017).

[142] R. Achar, P. S. Thilagam, N. Soans, P. V. Vikyath, S. Rao, and A. M. Vijeth, "Load balancing in cloud based on live migration of virtual machines," in *2013 Annual IEEE India Conference (INDICON)*, pp. 1–5, Dec 2013.

[143] Y. Zhao and W. Huang, "Adaptive distributed load balancing algorithm based on live migration of virtual machines in cloud," in *2009 Fifth International Joint Conference on INC, IMS and IDC*, pp. 170–175, Aug 2009.

list of publications

[**1**] Alhumaima, R. S., Khan, M. and Al-Raweshidy, H. S., "Component and parameterised power model for cloud radio access network," in *IET Communications*, vol. 10, no. 7, pp. 745-752, 2016.

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

[**3**] Alhumaima, R. S., and Al-Raweshidy, H. S., "Modelling the power consumption and trade-offs of virtualised cloud radio access networks," in *IET Communications*, vol. 11, no. 7, pp. 1158-1164, 2017.

[**4**] Khan, M., Alhumaima, R.S., Al-Raweshidy, H.S. "QoS-Aware Dynamic RRH Allocation in a Self-Optimised Cloud Radio Access Network with RRH Proximity Constraint," in *Network and Service Management*, IEEE Transactions, (accepted).

[**5**] Alhumaima, R. S., and Al-Raweshidy, H. S., "Energy Efficiency Maximisation of Virtualised Cloud Radio Access Network Via Optimising the Number of Virtual Machines," in *Green Communications and Networking*, IEEE Transactions, (Submitted).

[**6**] Alhumaima, R.S., Khan, M., Al-Raweshidy, H.S.: "Power Model for Heterogeneous Cloud Radio Access Networks", in *IEEE International Conference on Data Science and Data Intensive Systems*, Sydney, NSW, pp. 260–267, 2015.

[**7**] Khan, M., Alhumaima, R.S., Al-Raweshidy, H.S. "Reducing energy consumption by dynamic resource allocation in C-RAN," in *Networks and Communications (EuCNC), 2015 European Conference*, pp. 169–174, 2015.

[**8**] Khan, M., Alhumaima, R. S., H. S. Al-Raweshidy, "Quality of Service aware dynamic BBU-RRH mapping in Cloud Radio Access Network," *International Conference on Emerging Technologies (ICET)*, Peshawar, pp. 1-5, 2015.

[**9**] Alhumaima, R.S., Khan, M., Al-Raweshidy, H.S.: "Modelling the energy efficiency of Heterogeneous Cloud Radio Access Networks", in *Emerging Technologies (ICET) International Conference*, Peshawar, pp. 1–6, 2015.

[**10**] Alhumaima, R. S., and Al-Raweshidy, H. S., "Energy Consumption Reduction in Cloud Radio Access Network (C-RAN) Using SDN,", Poster presentation in *ResCon, Brunel University London*, 2014.

[**11**] Alhumaima, R. S., and Al-Raweshidy, H. S., "Optimising the BBU Pool Placement in Cloud Radio Access Networks based on Power Allocations," in *SAI Computing Conference*, London, UK, 2017, (Accepted).

[**12**] Alhumaima, R. S., J., Shireen and Al-Raweshidy, H. S., "Evaluating the Energy Efficiency of Virtualised 5G Networks, *Nova Science Publishers*, Book ch., Id 46182, 2017.